GENETICALLY INFORMATIVE DESIGNS FOR THE STUDY OF HEALTH ASSOCIATED BIOMARKERS

Finnicum, Casey Taylor

2021

document version
Publisher's PDF, also known as Version of record

GENETICALLY INFORMATIVE DESIGNS FOR THE STUDY OF HEALTH ASSOCIATED BIOMARKERS

Casey T. Finnicum

GENETICALLY INFORMATIVE DESIGNS FOR THE STUDY OF HEALTH ASSOCIATED BIOMARKERS

:

Casey T. Finnicum

ISBN: 978-94-6416-409-1
Lay-out and design: Daniëlle Balk | www.persoonlijkproefschrift.nl
Printing: Ridderprint | www.ridderprint.nl

VRIJE UNIVERSITEIT

GENETICALLY INFORMATIVE DESIGNS FOR THE STUDY OF HEALTH ASSOCIATED BIOMARKERS

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad Doctor
aan de Vrije Universiteit Amsterdam,
op gezag van de rector magnificus
prof.dr. V. Subramaniam,
in het openbaar te verdedigen
ten overstaan van de promotiecommissie
van de Faculteit der Gedrags- en Bewegingswetenschappen
op maandag 15 februari 2021 om 15.45 uur
in de aula van de Vrije Universiteit,
De Boelelaan 1105

door

Casey Taylor Finnicum

geboren te Edmonds, Washington, Verenigde Staten van Amerika

promotor:	prof.dr. J.C.N de Geus
copromotoren:	dr. E. A. Ehli prof.dr. C.V. Dolan

GENETICALLY INFORMATIVE DESIGNS FOR THE STUDY OF HEALTH ASSOCIATED BIOMARKERS

By

Casey T. Finnicum

B.S., South Dakota State University, 2014

A Dissertation Submitted in Partial Fulfillment of
the Requirements for the Degree of Doctor of Philosophy

Division of Basic Biomedical Sciences
Sanford School of Medicine
In the Graduate School
The University of South Dakota
February 2021

Reading committee:
prof.dr. Peter Beek (chair), FGB-VU
prof.dr. Harold Snieder, UMCG
prof.dr. Timothy Soundy, Avera
dr. Richard IJzerman, Amsterdam UMC
dr. Jenny van Dongen, FGB-VU

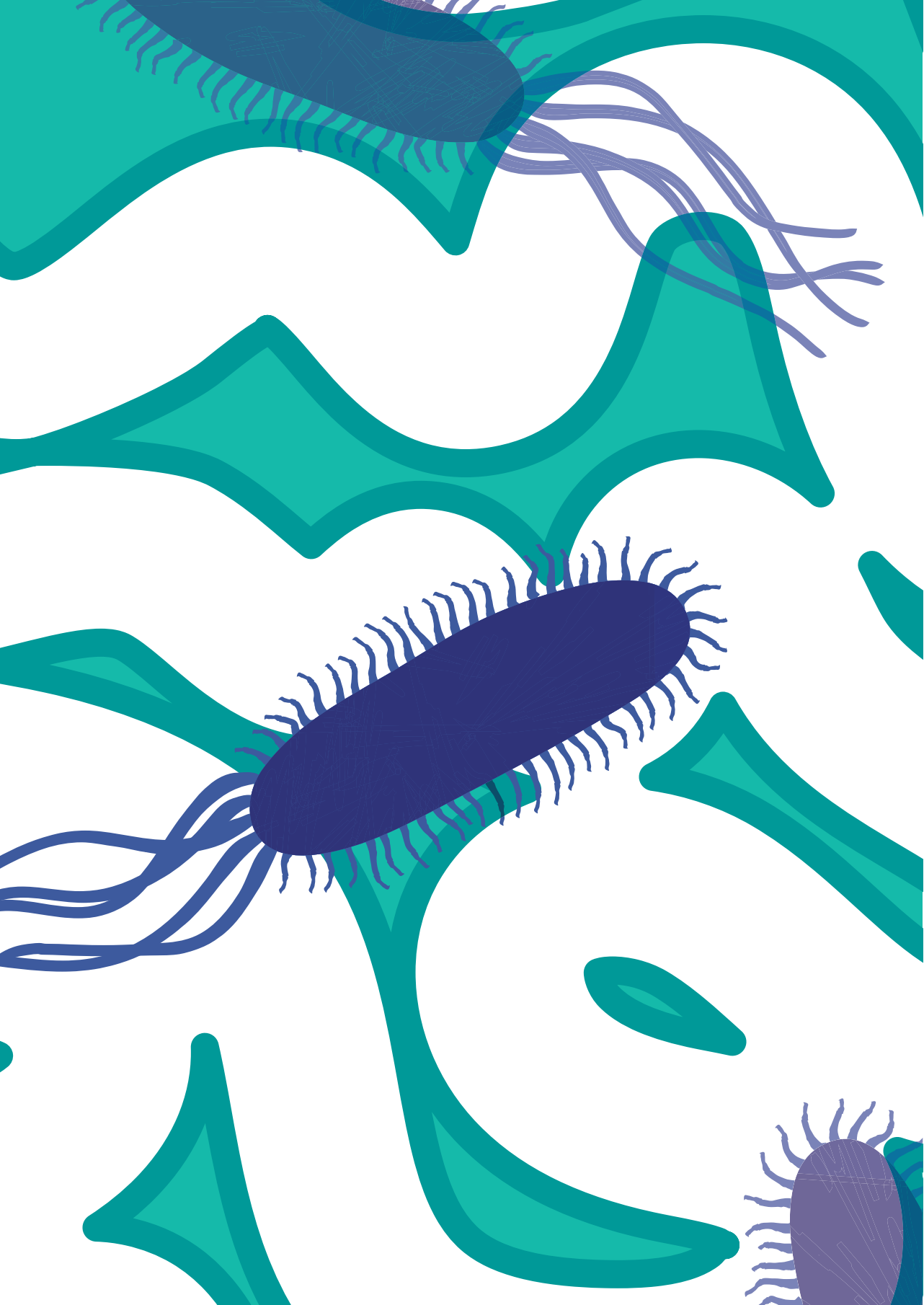
Paranymphs:
Jeffrey J. Beck
Matthijs van der Zee

Acknowledgment

First and foremost, I have to thank my wife, Megan, without her love, support and motivation, this thesis and my Ph.D. as a whole would not have been possible. My children, Luke and Ava, have been a source of continuous happiness and inspiration throughout the long endeavor that is graduate school. Additionally, there have been numerous family members and friends who have supported me, for which I am incredibly grateful.

Thank you to my promotor Eco de Geus, who has been invaluable in helping me to develop as a scientist. Furthermore, I am thankful that the Vrije Universiteit, the University of South Dakota and the Avera Institute for Human Genetics have all fostered a great atmosphere for learning and scientific development, all of which was necessary for the completion of this project.

TABLE OF CONTENTS



1

INTRODUCTION

TWIN DESIGN

Comparisons of twin pairs have long been a vital methodology to derive the degree to which a specific phenotype is influenced by genetic factors. The first studies to utilize the comparison of monozygotic and dizygotic twin pairs to deduce heritability of human traits were performed in the early 1920s, and focused on corneal refraction, melanocytic naevi, and IQ [1-3]. In the subsequent years, twin studies would be utilized to study a large number of measurable human phenotypes that were just as diverse as the phenotypes observed in the seminal studies. The "classical" twin design (CTD) generally involves both monozygotic (MZ) and dizygotic (DZ) twin pairs and notes the similarities or differences in a specific measurement within these groups. These observations are interpreted in the context that MZ twin pairs completely share their genome while DZ twins share, on average, half of their genome. Observing a higher degree of similarity in the MZ twins relative to the DZ pairs, for a particular trait, usually in the form of an increased correlation coefficient, is indicative of a proportion of the overall variance in that trait being attributable to heritable components (A). While these twin groups differ in the amount of shared genetic material, both co-twins, either MZ or DZ, are subjected to very similar environmental conditions throughout their development. Regardless of zygotic status, co-twins are generally reared in the same fashion, subjected to similar diets, and often several other critical environmental factors.

Given that both MZ and DZ co-twins share their environmental conditions to an equal degree, the correlations observed between sets of MZ and DZ twin pairs would be equally impacted by environmental factors. Thus, the relative correlations obtained from MZ and DZ pairs are also informative for determining the effect of shared environment (C). Although both MZ and DZ twin pairs share a large proportion of their environments, there are still aspects of each twin's environment that are unique to that twin relative to their co-twin. These factors could include a traumatic event or accident that only affected one twin or simply attending separate schools. The effect of these unique environmental influences, termed the unique environment (E), can be determined by observing the difference in correlations between MZ twin pairs for a particular trait. These three variation sources cumulatively influence the overall trait variation found in a population ($V_P = V_A + V_C + V_E$).

ACE MODELING

Through the use of data collected from twin participants, it is possible to partition the overall variance of a phenotype into the respective additive genetic, shared, and unique environmental components, using an ACE model [4]. The ACE model, efficiently represented with a path diagram, aims to visually represent a model of

the aforementioned genetic and environmental factors' influence over an observed phenotype. Figure 1.1 shows the visual representation of the ACE path diagram. In this path diagram, the observed phenotypes are depicted with boxes, whereas the latent or unobserved variables, A, C, and E, representing the additive genetic, common environment, and unique environment components respectively, are depicted with circles. The latent variables have an effect, denoted by a, c and e, on the variance of the observed phenotype of interest. Information regarding the covariance of latent variables between twin pairs is conveyed with a double-sided arrow. For MZ twin pairs, the covariance between the additive genetic components is always 1 due to completely sharing a genome. Similarly, the covariance between DZ twins' additive genetic components can be adequately estimated at 0.5 because DZ twin pairs share, on average, 50% of their genome. A covariance of 1 is observed between the common environmental components of both MZ and DZ twin pairs because both types of twins share certain aspects of their environments entirely. There is no arrow connecting the unique environmental components influencing a phenotype of co-twins as these factors are inherently unique to a single twin regardless of zygotic status. The model's parameters in Figure 1.1 are generally estimated using full information maximum likelihood methods implemented in programs such as OpenMx [5].

An alternative model is the ADE model, where the genetic influences are not restricted to additive genetic variance but also include non-additive variance (V_D) due to epistatic allele-allele interactions or allelic dominance. In the classical twin study, a choice must be made between ACE or ADE models based on the pattern of twin correlations. However, extended twin-family designs, which add data from parents and siblings, or second-degree relatives, to data from the twins themselves, can simultaneously estimate the ACDE components while relaxing the classical twin designs assumptions regarding mating and cultural transmission [6, 7]. The urgency of such complex models, however, has been questioned. In the most comprehensive analysis of the causes of individual differences in human traits important to medicine, psychology, and social sciences thus far, Polderman et al. reported on a meta-analysis of twin correlations and reported variance components for 17,804 traits from 2,748 publications including 14,558,903 partly dependent twin pairs [8]. The mean estimate of heritability across all traits was 49%. For a majority (69%) of traits, the observed twin correlations are consistent with a simple and parsimonious model where twin resemblance is solely due to additive genetic variation, with relatively modest influences from shared environment or non-additive genetic variation. Even so, the shared environment did show an impact on some of the traits, particularly those with low heritability. As this dissertation will show, the microbiome is a trait meeting this description.

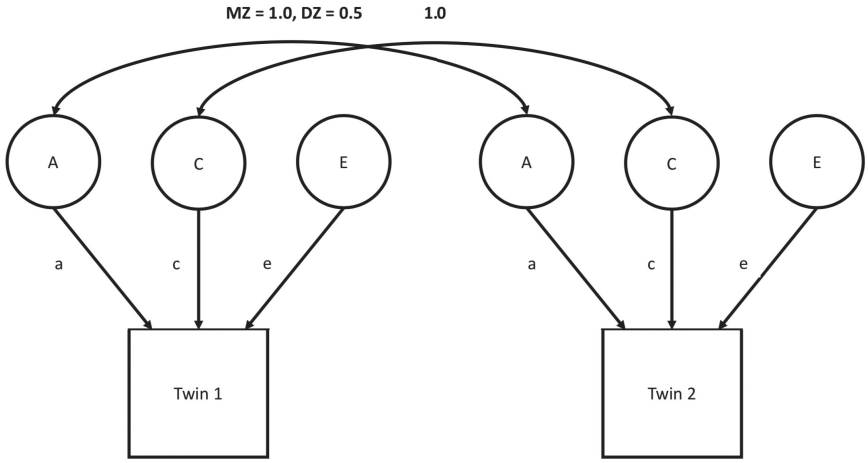


Figure 1.1 | Example of ACE path diagram. Depicts the latent components A, C and E and the effect these components have on the observed phenotype of interest.

THE NETHERLANDS TWIN REGISTER AT VU

Because of the position of authority that the twin design gained in the study of individual differences in human traits, specifically the long-standing issue of nature versus nurture, many countries have spawned national twin registries, with larger countries like the US even maintaining multiple region-specific registries [9]. Amongst these is the Netherlands Twin Register, maintained by the Department of Biological Psychology at the Vrije Universiteit (VU) Amsterdam, the data of which are at the core of the current dissertation. For more than three decades, the Netherlands Twin Register (NTR) has invited twins and their family members to participate in research studies [10], including siblings, spouses, children, parents, and even grandparents of twins. A primary driver has been to perform the type of analyses depicted in Figure 1.1 to detect the genetic and non-genetic contribution to cognitive and emotional development in childhood and adolescence, and adult behavioral and health and well-being outcomes. These outcomes were partly assessed by 2-3 yearly surveys and partly by experimental studies of, e.g., autonomic nervous system function, cardiovascular risk, brain structure and function, neurocognitive test performance and IQ, or (an)aerobic fitness and daily physical activity.

The largest of the experimental studies, the NTR biobank study of nearly 10,000 participants, set out to characterize NTR participants on a host of metabolic and immunological risk factors and create a resource for future “omics” biomarker studies. Examples of biomarkers assessed for all NTR Biobank participants include body mass index (BMI), waist-hip ratio, HDL-C, LDL-C, triglycerides, Hb1Ac, fasting

glucose and insulin, liver enzymes, C-reactive protein, fibrinogen, interleukin (IL)-6, TNF-α, and soluble IL-6 receptor. Together these biomarkers allow an assessment of cardiometabolic risk and chronic low-grade inflammation. The term cardiometabolic risk describes an individual's chances of damaging their heart and blood vessels when one or more of the following risk factors are present: obesity, high LDL ("bad") cholesterol, high blood fat (triglycerides), insulin resistance. Each of these risk factors is dangerous on its own, but a combination dramatically increases the risk of heart disease and stroke, particularly if blood pressure is also heightened. Metabolic syndrome, syndrome X, cardiometabolic syndrome, and insulin resistance syndrome are other terms for this cluster of risk factors. Low-grade inflammation is usually defined as "the chronic production, but at a low-grade state, of inflammatory factors." Low-grade inflammation does not stem from an overt infection but reflects a slight increase in the concentrations of multiple inflammatory factors compared to a healthy individual, even if each remains in the healthy range. Cytokines are often involved in low-grade inflammation [11, 12]. They have an important role in cell-to-cell signaling in the coordinated immune response to infection and can also act in the brain and induce behavioral changes like sickness behaviors.

DNA collection constituted an essential element of the NTR biobank project, setting the stage for a continuous and ongoing DNA collection effort in all NTR participants who have ever provided any phenotype information. DNA is collected from whole blood for most adult participants. Adult twins are typically asked to provide a buccal sample in addition. Buccal swabs are used for oral DNA collection in young twins and their siblings and parents. The motivation for such large-scale DNA collection is evident. Progress in technology for directly measuring genetic variants at the DNA level has made it possible to quantify the genetic similarity between individuals without the need for data on family relationships. Direct measurement of genetic variants, most often Single Nucleotide Polymorphisms (SNP), is possible through the use of technologies such as polymerase chain reaction (PCR), microarray genotyping, and full genome sequencing [13]. All of these technologies are capable of providing information about segregating alleles within the genome. The data obtained through these technologies, particularly those from microarray platforms, are now commonly used in genome-wide association studies (GWAS). These studies use large samples of human genomic data, coupled with phenotypic information of interest, to discover genetic loci significantly associated with that phenotype [14-16]. The success of the international GWAS consortia in terms of the number of confirmed loci is staggering, with the GWAS catalog containing more than 6000 GWAS comprising >75000+ variant-trait associations from nearly 4000 publications (https://www.ebi.ac.uk/gwas/).

THE NTR-AVERA COLLABORATION

The processes of measuring genetic variants and the subsequent data analyses require specialized skill sets, which often necessitates collaboration between research groups that focus on these areas to answer complex genetic questions. To this end, a collaboration was formed between the Avera Institute for Human Genetics (AIHG) and the VU/NTR. Within this collaboration, the molecular genetics laboratory at AIHG has brought forth the expertise and infrastructure capable of performing a vast array of molecular genetic experiments. The NTR research group, in turn, provided expertise concerning the many aspects of biobanking, phenotype collection, and advanced statistical methods vital to carrying out genetic research in longitudinal twin-family data.

From the genesis of the collaboration, AIHG has produced large amounts of genetic data on NTR participants through the use of DNA microarray technologies. Among the many microarray capabilities within the AIHG repertoire are the Axiom-NL array [17] and its successor, a customized version of the Illumina Global Screening (GSA) array [18], which are both themselves products of the NTR-Avera collaboration. These platforms optimize population-specific genotyping assays using an appropriate whole-genome sequence reference set and add SNPs known to influence pharmacogenomic responsivity, traits like cardiometabolic diseases and common psychiatric disorders, and traits of specific interest to NTR like fertility and twinning.

The AIHG genotyping and Avera-VU collaborative scientific input has been successfully employed in large-scale genetic discovery studies spanning numerous fields including but not limited to attention problems [19-24], aggression [25], substance use [26-28], personality [29], exercise behavior [30], depression [31-35], intelligence [36], cortical and subcortical brain structures [37-39], DNA methylation [40-42], twinning and female Fertility [43, 44] and many other traits [45-48].

Apart from its genotyping facilities, the AIHG lab provides NTR with many other sequencing-based technologies. An example is the large-scale epigenotyping possibilities using the Illumina Methylation arrays that investigate >450k or >850k (EPIC) DNA methylation sites to offer a broad view of methylation state, covering CpG islands, genes, and enhancers. Particularly when located in a gene promoter, DNA methylation typically acts to modify gene transcription, which may be as extreme as complete silencing of the gene. Studies in MZ twins have shown an age-related divergence of methylation patterns due to environmental rather than genetic influences [41, 49, 50]. Overall, there is a global loss of DNA methylation during aging. Differences in the speed of this loss act like a biological clock predicting disease-onset better than actual chronological age [51, 52]. In this dissertation, however, I used a second strategy to study aging and age-related disease based on an adaptive DNA trait, the measurement of which I implemented at AIHG for the NTR-Avera collaboration: the assessment of telomere length.

TELOMERE LENGTH

Aging is a process all humans inevitably face throughout their lifetime. At this point, it has been predicted that the number of individuals over the age of 60 will double from 11% to 22% between 2000 and 2050 [53]. Given the expected dramatic increase in the aging population, it is of the utmost importance that we understand the aging process to help individuals live longer, healthier lives. Furthermore, as we age, it is essential to understand the molecular nature of aging and discover biomarkers that lend information regarding an individual's aging process. One commonly used molecular test is to measure the telomere regions of an individual's DNA via molecular methods. Telomeres are genomic regions situated at the ends of chromosomes that provide chromosomal stability and many other functions critical to biological processes [54]. The telomere, which caps the end of each strand of DNA, is subjected to attrition throughout the life span due to the end replication problem, which results in the loss of approximately 50-100 base pairs per mitotic division [55]. Once a significant portion of the telomeric region has been degraded, the cell enters a state of replicative senescence characterized by a marked change in gene expression and the inability to divide further [56, 57]. Despite constant telomere degradation over the life span of a cell, mechanisms are available for telomere elongation, mainly through the use of the enzyme telomerase. The observation of telomere attrition in proliferating cells and the immortality conveyed via telomerase activation suggests that telomeres act as a central biological clock mechanism [58]. This association between telomere length and a so-called biological clock is supported by studies highlighting an association between telomere length and life span in humans [59-62].

To learn about telomere-associated dynamics across an individual's lifetime, studying telomere dynamics longitudinally, starting from a young age, is pivotal. One hindrance to this process is collecting DNA from individuals at a very young age. The standard method of telomere measurement utilizes blood-derived DNA. Intravenous blood draws on infants are considered by many to be an unnecessary burden on a young child. The ability to use a more easily collectible DNA sample would greatly aid researchers' ability to collect and perform telomere measurements at a young age. Specifically, buccal-derived DNA in place of leukocyte-derived DNA would greatly facilitate large-scale telomere measurement studies in child samples as it just needs a buccal swab and no blood draws.

In a large and overlapping set of participants, the NTR collected DNA from both blood and buccal samples. Using these DNA samples, an extensive characterization of the telomere length was undertaken for this dissertation exploiting two specific advantages of the NTR in full (multiple tissues sampling and the genetically informative design). These advantages make it feasible to understand how estimates of the cause of variation in telomere length (i.e., the ACE variance components in Figure 1.1) may

differ between various biological tissues. Additionally, the repeated measurement structure in part of the blood samples allows us to study if and how telomere length measurements can vary across different laboratories.

HUMAN MICROBIOME

Although data derived from genome-wide genomic marker platforms (e.g., SNPs or CpGs), especially when coupled with missing data imputation, are incredibly useful for many discoveries, these platforms do not come without limitations. Microarray technologies, for example, are only capable of measuring SNPs known to exist and are designed with a priori knowledge regarding the genetic architecture surrounding the SNP of interest, which is necessary for the design of nucleotide probes to query a particular genomic location of interest. In order to directly measure novel genetic variants, technologies such as DNA sequencing are available. AIHG has provided the means to directly sequence all nucleotides within the human genome, mainly through next-generation sequencing technologies [63]. One realm of research carried out by the joint NTR-Avera research collaboration that requires these measurement techniques is microbiome research.

The microbiome refers to the collective genomes of the commensal, symbiotic and pathogenic microorganisms found in and on all multicellular organisms. These prominently include bacteria - estimates have determined that there is at least one bacterial cell per human cell residing on the human body [64], but also archaea, protists, fungi, and viruses. These microorganisms have been found to be crucial for immunologic, hormonal, and metabolic homeostasis of their host and impact several aspects of human health [65]. The strongest direct empirical evidence that microbiomes can drive disease comes from experiments, mostly in animal models, in which the microbiome from diseased donors is “transplanted” into healthy germ-free hosts. By showing that the recipients of the obesity-associated microbiome themselves developed obesity, a strong case was made, e.g., the microbiome’s involvement in obesity [66]. Health effects of the microbiome composition have been particularly intensely researched for obesity and a range from autoimmune and cardiovascular diseases to mental disorders. Although the full human microbiome consists of many microorganisms that inhabit many parts of the human body, I will focus on the bacterial constituents of the gut microbiome in this dissertation. The necessity of sequencing techniques in gut microbiome research stems from the lack of a full genomic and taxonomic characterization of the many microorganisms that inhabit the microbiome-associated communities.

International projects such as MetaHIT and the Human Microbiome Project [67, 68] have used large scale sequencing to demonstrate a vast amount of variability in the

human gut microbiome both within and between subjects. Sequencing approaches fall into one of two categories. Shotgun metagenomics charts the collective genome of microorganisms from a sample by shearing all extracted DNA, sequencing the small fragments, and combining them into contigs for annotation of gene functions. Targeted amplicon studies focus on one or a few marker genes and use these markers to reveal the composition and diversity of the microbiome. Metagenomics approaches have the advantage of providing much richer data on the potential molecular functionality present in microbial communities. However, these experiments are quite resource intensive. Alternatively, targeted amplicon studies allow for taxonomic identification without the need for the large amount of resources necessary to carry out the metagenomic sequencing methods. These methods are commonly carried out through the sequencing of the 16S rRNA gene. This process involves amplifying particular variable regions within the 16 rRNA gene of a microbiome community's bacterial constituents. This gene region is amplified from the DNA of all bacterial members present within a sample and sequenced. 16S rRNA gene sequence data provide a relatively unbiased characterization of bacterial and archaeal diversity while sufficiently economical to allow large-scale epidemiological sampling. It is the approach used throughout this dissertation.

The most widely used software packages to handle16S rRNA gene sequence data from complex microbial communities are QIIME (http://www.qiime.org) and mothur (http://www.mothur.org). Both packages are open source and have online tutorials and forums. In this dissertation, I have predominantly used mothur [69]. A microbiome analyses end product is a frequency count for the number of sequence reads that cluster into so-called operational taxonomic units (OTU). OTU assignment of a read is based on a percentage of sequence identity (%ID). Various thresholds of sequence identity are used to represent different taxonomic levels (e.g., 97% ID for species, 95% for genera). These taxonomic thresholds are known to be very rough estimates: the degree of sequence variability depends on the region of the 16S rRNA gene sequenced, the length of the amplicon, and the specific taxa in question. OTUs picked at 97% sequence identity provide a naming convention for related bacterial species, while acknowledging that there is no rigorous "species" concept for bacteria.

A rather important choice with a significant impact on downstream findings is the OTU-picking algorithm chosen. OTU clustering algorithms fall into two main categories: de novo and reference-based methods. In de novo OTU picking, all sequences are used and clustered into OTUs, without any external reference sequences [70]. In contrast, reference-based OTU picking (also called phylotyping) uses a reference sequence database, such as the Ribosomal Database Project (RDP, http://rdp.cme.msu.edu/), greengenes (https://greengenes.secondgenome.com/), or SILVA (https://www.arb-silva.de/) to classify samples into known microbial taxa based on sequence similarity between the sample and the reference taxon. Sample sequences that fail

to sufficiently match the reference sequence database are discarded, or clustered separately in the case of open-reference OTU picking.

Although I appreciate that both methods have pros and cons (see http://qiime.org/tutorials/index.html), I have predominantly used de novo clustering in this dissertation. This choice was guided by the fact that defining OTUs using the reference-based approach can lead to a poor representation of the actual distances between sequences [70] leading to the lumping together of microbes into a single taxon, that are in fact quite distinct.

At the level of an individual, the microbiome can be qualitatively characterized by the OTU abundances for all OTUs encountered in that individual or by assessing the abundances of higher-level taxonomic groups, such as orders or families: summing the sequences for all OTUs belonging to the group of interest (collapsing taxonomies). In short, OTU-picking algorithms yield lists of OTUs with taxonomic labels. Note that a majority of the detected OTUs will not be shared at appreciable abundance levels by all individuals in a study [71]. In de novo clustering, many OTUs will lack a complete taxonomy label; for example, the classification might include a family level categorization but might lack genus or species categorization. At first glance, taxa that have associated genus/species information are more appealing and tend to get more weight in discussion of results. This is incorrect. OTUs without genus/species information are frequently both more abundant and more representative of total diversity than are OTUs with genus/species names.

After sequences have been assigned to OTUs, their phylogenetic relationships can be inferred, either by using an existing reference database with an associated phylogeny (such as RDP, greengenes, or SILVA) or by inferring the phylogeny using de novo sequence alignment tools. Even after sequences have been assigned to OTUs and related to one another using a phylogenetic tree, the scale of the data is still extensive. To avoid interpreting many spurious associations, care must be taken to correct association statistics for multiple comparisons, as there are generally hundreds of OTUs being tested for association with, e.g., cardiometabolic risk. Also, abundances of OTUs are seldom normally distributed because many samples will have zero counts for rare OTUs. Various approaches exist to interpret microbiome data to reveal meaningful patterns in microbial diversity. Typical methods include applying metrics of within-individual diversity (like the alpha-diversity) or between-individual diversity (like the beta-diversity), which can be visualized using ordination techniques, such as principal coordinates analysis (PCoA) that summarize beta diversity relationships in two- or three-dimensional scatterplots.

Building on this, individuals can be classified into distinct groups based on their microbiome composition. Classification methods can be supervised or unsupervised.

Supervised classification methods can be used to determine which taxa differ between predefined groups of samples (e.g., obese versus normal weight, cases versus controls, genotype A versus genotype B) and to build models that use these discriminatory taxa to predict the classification of a new sample [72]. Unsupervised classification, or clustering, does not use any prior knowledge about the samples and categorizes them into clusters based on the abundances of specific taxa. A between-individual distance metric, such as UniFrac or Bray-Curtis, is used to generate these clusters.

GENETIC AND ENVIRONMENTAL INFLUENCES ON THE HUMAN GUT MICROBIOME

Microorganisms within the gut microbiota have exposure to myriad conditions that help shape community membership. In animal studies, these conditions prominently include inoculation at birth (the maternal effect, breeding and raising conditions in the facility, co-caging, the water acidity, food, bedding) for which a number of experimental mitigation strategies have been devised [73]. In humans, the microbiome is impacted by several factors including, but not limited to, age, antibiotics, diet, lifestyles, pregnancy, mode of delivery, and ethnicity [74-88]. Strong evidence further exists for an impact of the shared living environment on the gut microbiome reflected amongst others in the resemblance seen in the microbiome composition of family members or people sharing a neighborhood [89-101].

Although similarities of the gut microbiota of family members have been partly attributed to a shared environment, the host genetic profile is also clearly capable of shaping this diverse ecosystem of microorganisms [102-110]. Recent extensive cohort studies have further strengthened our understanding of the role of human genetics in influencing the gut microbiota by unraveling associations between specific loci in the human genome and individual taxa of the gut microbiota [111].

Taken the evidence for influences of the host genetic profile and shared environmental influences on microbiota composition, genetically informative study designs can be particularly useful in studying the gut microbiota. In this thesis, I will use such designs to examine the mechanisms through which genetic and environmental variation impacts the gut microbiome. Understanding the mechanisms by which host genetic factors influence the gut microbiota and thus, in turn, impacts the development of disease is essential for developing biological therapies targeting the gut microbiota. In parallel, understanding the mechanisms by which modifiable environmental factors influence the gut microbiota may lead to new strategies for disease prevention.

HUMAN GUT MICROBIOME AND HEALTH ASSOCIATED BIOMARKERS

An area of research that has been of active interest to the scientific community is understanding how the gut microbiota can influence the development of obesity. Current hypotheses put forth that the gut microbiota of obese individuals may be more efficient at extracting energy from food the host consumes [66]. Animal studies have demonstrated that the gut microbiota of obese animals is capable of inducing an increase in fat accumulation in germ-free animals receiving microbiota transplantation from obese animals relative to animals receiving transplantation from lean donors [66]. It is worth noting that, along with the induction of increased fat accumulation, comorbidities such as changes in neuroinflammation and cognitive disruptions have also been induced through the transfer of the gut microbiota of obese hosts to recipient animals [112]. These findings suggest at least a partially causal effect of the gut microbiota on the development of obesity and other closely related health factors.

Obesity is known to be associated with changes in the cardiometabolic and inflammatory biomarker profile of humans [113] so by extension any microbiome effects on obesity may impact these two classes of health associated biomarkers. Indeed, various studies have found significant association between the microbiome on the one hand and inflammatory risk factors on the other [114-117]. However, the causality of these associations cannot readily be deduced. One of the most basic functions the microbiome plays is in the defense against pathogens through competition for resources. Furthermore, the human microbiome has been found to participate in a multitude of complex interactions with both the innate and adaptive immune responses of the human host [65]. Modulations in immune factors may, therefore, themselves affect the gut microbiota composition through changes in the immune-commensal interactions [118].

While the crosstalk and interactions between commensal microbes and the human host seem like a logical extension of the immune systems well-characterized role in interacting with microorganisms, the effect of microbially-derived metabolites on disease states such as metabolic syndrome and cardiovascular disease is currently less clear [119]. Even when associations are found, similar concerns apply regarding their causality. Behavioral factors commonly implicated in obesity, like eating behaviors and physical (in)activity are well-known to influence the inflammatory and metabolic profiles of the human hosts, while also having the capability of modulating the gut microbiota composition [120, 121]. Finally, an association between the microbiome on the one hand and obesity, metabolic and inflammatory risk factors on the other can be caused by independent effects of a common genetic vulnerability on the microbiome and the respective health associated biomarkers.

Simply put, does the composition of the gut microbiota increase the risk for disease? Do disease risk factors themselves change the composition of the gut microbiota? Or are they both reflective of common genetic and environmental factors? In this thesis, I will address this question of causality focusing on cardiometabolic and inflammatory variables as the main parameters of interest.

HYPOTHESES AND OBJECTIVES OF THE DISSERTATION

This dissertation examines the genetic and environmental determinants of telomere length and microbiome composition using several distinctive study designs allowed by having data available in whole-genome genotyped members of a twin family cohort. The work put forth within this dissertation takes advantage of unique study designs to answer several questions pertaining to human aging, obesity, and obesity related biomarkers of compromised health. In the second chapter, the classical twin design is used to determine the heritability of telomere repeat mass (TRM). This informative twin design, coupled with direct genomic measurement methods that utilize PCR to measure the TRM of individuals' genomic content, allowed for the estimation of the heritability of TRM. Earlier studies have attempted to establish the importance of genetic factors to variation in this measure of biological aging in blood [122, 123], but it is not clear if the heritability estimates hold across different tissue types. Availability of repeated samples from multiple biological tissues collected simultaneously from multiple participants allowed for the comparison of telomere measurements and subsequent heritability estimates between different tissues. These comparisons allow us to answer questions regarding the suitability of buccal DNA for telomere measurement. Such information is invaluable for individuals planning large-scale or longitudinal epidemiological studies that include telomere measurement. In samples repeatedly tested after significant handling and processing we could also test the effects of these procedures on heritability estimation. We hypothesized that buccal-derived DNA, even after repeated sample handling, would provide a suitable alternative to blood-derived DNA for TRM measurement purposes via qPCR.

Chapters three to five, using a series of different designs, all address the main question of the relative contribution of genetic and environmental influences on individual variation in the diversity and composition of the human gut microbiome. They capitalize on the stool samples collected in a subsample of over 400 NTR biobank participants in which an extensive characterization of the composition and diversity of the gut microbiome was undertaken for this dissertation. Microbiome composition was charted through the use of 16S rRNA sequencing to determine the taxonomy of the microorganisms present in the gut microbiota of the NTR participants. Chapters three and five are focused on the environmental effects, whereas chapter four has

the genetic effects as the central theme and describes my contribution to a large international genome-wide association consortium [124].

Chapter three also specifically addresses the question of causality underlying the well-known association with obesity. It does so by employing two different genetically informative designs, the discordant MZ twin design and a recall-by-genotype design. The latter examines the differences between individuals at varying degrees of genetic risk for obesity, determined through the use of polygenic risk scores (PRS) created with the results of GWAS aimed at identifying loci influencing BMI. The BMI-PRS was calculated for a large population of NTR participants, after which fecal samples were obtained from individuals at the ends of the population distribution for genetic predisposition for obesity and actual observed BMI. This selection included individuals who fell within the top or bottom 25% of the observed BMI distribution and either the top or bottom 20% of the distribution of genetic predisposition for obesity (see Figure 1.2). Fecal DNA from these participants was sequenced similarly to the BMI discordant MZ twin pairs. The use of this study design allowed us to examine the nature of the association between the gut microbiota and BMI in more detail.

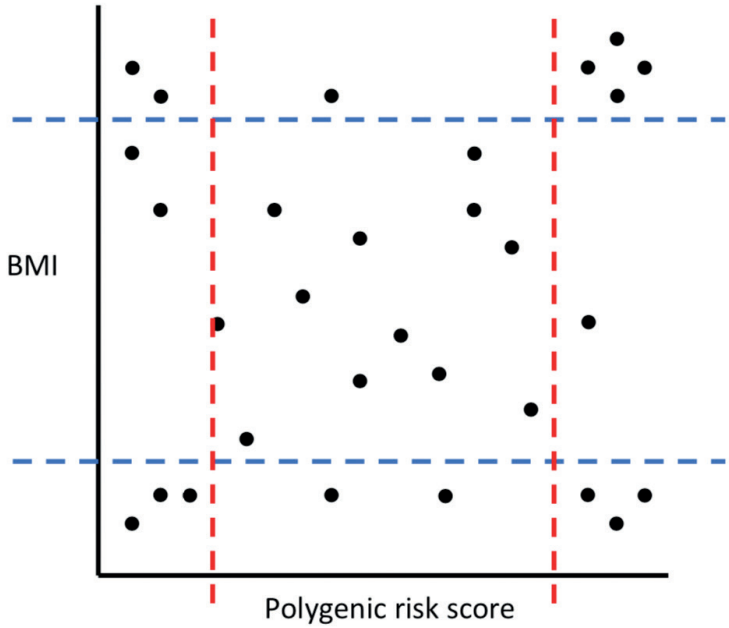


Figure 1.2 | Example plot representing the 4-corners design. Computed polygenic risk scores are on the X axis while BMI is on the Y axis. The blue lines represent the cutoffs of the top and bottom 25% of the BMI distribution and the red lines show the top and bottom 20% cutoff for the polygenic risk scores.

In testing the effect of BMI (high/low) and genetic risk (high/low) on the composition of the gut microbiota, we anticipated two outcomes. If the causal chain is high genetic risk → high BMI → gut microbiota composition, we expect a main effect of BMI (high/low) only. This expectation is based on the assumption that the relationship between genetic risk and composition is mediated by BMI. In contrast, if gut microbiota composition is a cause of high BMI, we expect a main effect of BMI independent of genetic risk on gut microbiota composition. The availability of MZ twin pairs in which one twin had a low and the co-twin had a high BMI further allowed us to discriminate between a direct causal effect of BMI on gut microbiota composition and an association brought about by genetic pleiotropy. The so-called discordant MZ design compares genetically identical individuals selected to be significantly different in a trait of interest. The discordant MZ twin design provides the ultimate case-control matching for genetic profile, pregnancy, age, sex, and childhood environment. If BMI is the causal agent, a comparison of MZ twins selected to be discordant for BMI should show a distinct composition of the gut microbiota in the lower and higher BMI individuals. We hypothesized that high BMI would be associated with quantitative (smaller species diversity) and qualitative effects (enrichment for different species) on the gut microbiome.

Chapter four is based on a large meta-analysis of the genetic contribution to variation in diversity and composition of the microbiome by the MiBioGen consortium [124]. We used the results of this meta-analysis to test how well the results of the NTR were reflective of the meta-analysis, where a good correspondence would act as a general validation of the results used throughout this thesis. Specifically, we tested whether the top consortium mbQTL showed at least nominal significance in the NTR, and vice versa, whether the top NTR-specific mbQTL was among the genome-wide significant results of the consortium. In addition, we explored the idea that a priori knowledge from heritability estimates from the classical twin design could be used to reduce the multiple testing burden in microbiome genetics. This would be the case if the more heritable taxa tend to show much smaller p-values in the genome-wide association tests relative to less heritable taxa. By focusing the GWA effort on taxa that demonstrate a minimal MZ twin correlation, the p-value penalty for multiple testing could be relaxed.

Chapter five highlights the use of yet another genetically informative design feasible with twin-family data. It utilizes samples collected from MZ twin pairs that cohabitate and MZ twin pairs that no longer cohabitate. Additionally, this study utilizes samples collected from spouse pairs who cohabitate. The use of the MZ twin-spouse study design is unique in that it allowed us an opportunity to key in on the environmental influences of the gut microbiota by observing shared microbiota features within cohabitating but genetically unrelated individuals. Contrasting the co-twins and spouses' gut microbiotas allows for an exciting look at the role of environmental and

genetic influences on shaping the gut microbiota composition. In addition to this unique look at environmental impacts on the gut microbiota composition, the wealth of existing phenotypic data collected within the NTR enabled us to compare the microbiota composition to metabolic and inflammatory associated phenotypes. This comparison aids in understanding which aspects of the gut microbiota are associated with meaningful physiological changes with the host. We hypothesized that this genetically informative study design would help us show that cohabitation results in similarities in the gut microbiome composition of individuals. We further hypothesize that these gut microbiome similarities may influence health-associated phenotypes in the domains of cardiometabolic risk and chronic low-grade inflammation.

Chapter six first presents a summary of the main results in the various chapters and then discusses where we stand with regard to the relative contribution of genetic and environmental effects to variation in the composition of the gut microbiome and with regard to the progress in understanding the role of the gut microbiome in human health.

References

1. Jablonski, W., *A contribution to the heredity of refraction in human eyes.* Arch Augenheilk, 1922. **91**: p. 308-28.
2. Merriman, C., *The intellectual resemblance of twins.* Psychological Monographs, 1924. **33**(5): p. i.
3. Siemens, H.W., *Die Zwillingspathologie: Ihre Bedeutung· Ihre Methodik· Ihre Bisherigen Ergebnisse*. 2013: Springer-Verlag.
4. Neale, M. and L.R. Cardon, *Methodology for genetic studies of twins and families*. Vol. 67. 2013: Springer Science & Business Media.
5. Boker, S., et al., *OpenMx: an open source extended structural equation modeling framework.* Psychometrika, 2011. **76**(2): p. 306-317.
6. Keller, M.C., S.E. Medland, and L.E. Duncan, *Are extended twin family designs worth the trouble? A comparison of the bias, precision, and accuracy of parameters estimated in four twin family models.* Behavior genetics, 2010. **40**(3): p. 377-393.
7. Keller, M.C., et al., *Modeling extended twin family data I: description of the Cascade model.* Twin Research and Human Genetics, 2009. **12**(1): p. 8-18.
8. Polderman, T.J., et al., *Meta-analysis of the heritability of human traits based on fifty years of twin studies.* Nature genetics, 2015. **47**(7): p. 702-709.
9. Hur, Y.-M., et al., *Twin family registries worldwide: An important resource for scientific research.* Twin Research and Human Genetics, 2019. **22**(6): p. 427-437.
10. Ligthart, L., et al., *The Netherlands twin register: longitudinal research based on twin and twin-family designs.* Twin Research and Human Genetics, 2019. **22**(6): p. 623-636.
11. Danesh, J., et al., *C-reactive protein and other circulating markers of inflammation in the prediction of coronary heart disease.* New England Journal of Medicine, 2004. **350**(14): p. 1387-1397.
12. Howren, M.B., D.M. Lamkin, and J. Suls, *Associations of depression with C-reactive protein, IL-1, and IL-6: a meta-analysis.* Psychosomatic medicine, 2009. **71**(2): p. 171-186.
13. Strachan, T., A. Read, and T. Strachan, *Human molecular genetics. 4th.* New York: Garland Science, 2011.
14. Visscher, P.M., et al., *Five years of GWAS discovery.* The American Journal of Human Genetics, 2012. **90**(1): p. 7-24.
15. Visscher, P.M. and G.W. Montgomery, *Genome-wide association studies and human disease: from trickle to flood.* Jama, 2009. **302**(18): p. 2028-2029.
16. Visscher, P.M., et al., *10 years of GWAS discovery: biology, function, and translation.* The American Journal of Human Genetics, 2017. **101**(1): p. 5-22.
17. Ehli, E.A., et al., *A method to customize population-specific arrays for genome-wide association testing.* European Journal of Human Genetics, 2017. **25**(2): p. 267-270.
18. Beck, J.J., et al., *Genetic Similarity Assessment of Twin-Family Populations by Custom-Designed Genotyping Array.* Twin Res Hum Genet, 2019. **22**(4): p. 210-219.
19. Abdellaoui, A., et al., *CNV concordance in 1,097 MZ twin pairs.* Twin Research and Human Genetics, 2015. **18**(1): p. 1-12.

20. de Zeeuw, E.L., et al., *Polygenic scores associated with educational attainment in adults predict educational achievement and ADHD symptoms in children.* American Journal of Medical Genetics Part B: Neuropsychiatric Genetics, 2014. **165**(6): p. 510-520.
21. Demontis, D., et al., *Discovery of the first genome-wide significant risk loci for attention deficit/hyperactivity disorder.* Nature genetics, 2019. **51**(1): p. 63-75.
22. Groen-Blokhuis, M.M., et al., *A prospective study of the effects of breastfeeding and FADS2 polymorphisms on cognition and hyperactivity/attention problems.* American Journal of Medical Genetics Part B: Neuropsychiatric Genetics, 2013. **162**(5): p. 457-465.
23. Middeldorp, C.M., et al., *A genome-wide association meta-analysis of attention-deficit/hyperactivity disorder symptoms in population-based pediatric cohorts.* Journal of the American Academy of Child & Adolescent Psychiatry, 2016. **55**(10): p. 896-905. e6.
24. Ehli, E.A., et al., *De novo and inherited CNVs in MZ twin pairs selected for discordance and concordance on attention problems.* European Journal of Human Genetics, 2012. **20**(10): p. 1037-1043.
25. Pappa, I., et al., *A genome-wide approach to children's aggressive behavior: The EAGLE consortium.* American Journal of Medical Genetics Part B: Neuropsychiatric Genetics, 2016. **171**(5): p. 562-572.
26. Baumert, J., et al., *No evidence for genome-wide interactions on plasma fibrinogen by smoking, alcohol consumption and body mass index: results from meta-analyses of 80,607 subjects.* Plos one, 2014. **9**(12): p. e111156.
27. Liu, M., et al., *Association studies of up to 1.2 million individuals yield new insights into the genetic etiology of tobacco and alcohol use.* Nature genetics, 2019. **51**(2): p. 237-244.
28. Minică, C.C., et al., *Genome-wide association meta-analysis of age at first cannabis use.* Addiction, 2018. **113**(11): p. 2073-2086.
29. van den Berg, S.M., et al., *Meta-analysis of genome-wide association studies for extraversion: findings from the genetics of personality consortium.* Behavior genetics, 2016. **46**(2): p. 170-182.
30. Huppertz, C., et al., *A twin-sibling study on the relationship between exercise attitudes and exercise behavior.* Behavior Genetics, 2014. **44**(1): p. 45-55.
31. Benke, K.S., et al., *A genome-wide association meta-analysis of preschool internalizing problems.* Journal of the American Academy of Child & Adolescent Psychiatry, 2014. **53**(6): p. 667-676. e7.
32. Mbarek, H., et al., *Genome-wide significance for PCLO as a gene for major depressive disorder.* Twin Research and Human Genetics, 2017. **20**(4): p. 267-270.
33. Middeldorp, C., et al., *The genetic association between personality and major depression or bipolar disorder. A polygenic score analysis using genome-wide association data.* Translational psychiatry, 2011. **1**(10): p. e50-e50.
34. Okbay, A., et al., *Genetic variants associated with subjective well-being, depressive symptoms, and neuroticism identified through genome-wide analyses.* Nature genetics, 2016. **48**(6): p. 624-633.
35. Wray, N.R., et al., *Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression.* Nature genetics, 2018. **50**(5): p. 668-681.
36. Franić, S., et al., *Intelligence: shared genetic basis between Mendelian disorders and a polygenic trait.* European Journal of Human Genetics, 2015. **23**(10): p. 1378-1383.

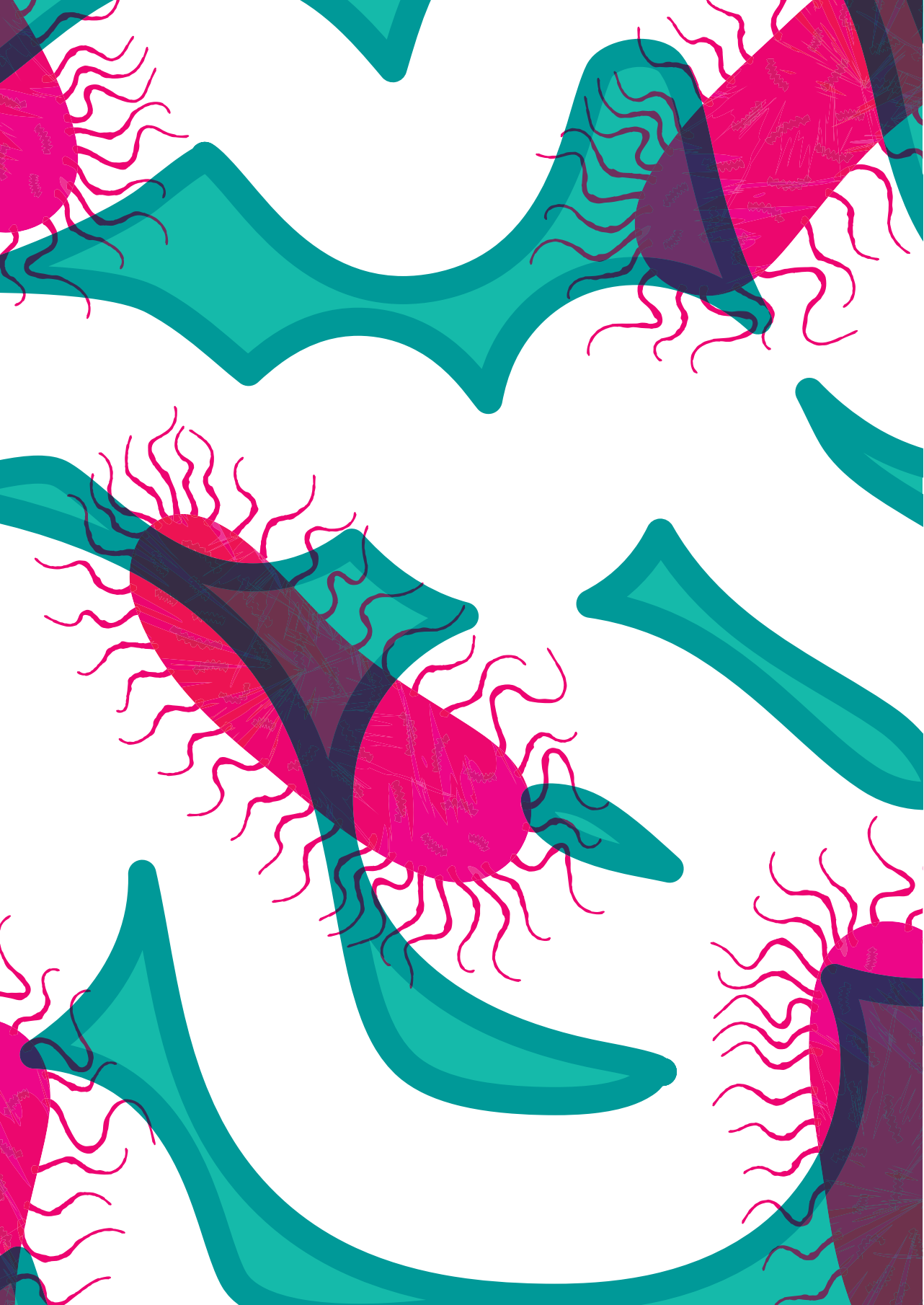
37. Adams, H.H., et al., *Novel genetic loci underlying human intracranial volume identified through genome-wide association.* Nature neuroscience, 2016. **19**(12): p. 1569-1582.
38. Hibar, D.P., et al., *Novel genetic loci associated with hippocampal volume.* Nature communications, 2017. **8**(1): p. 1-12.
39. Hibar, D.P., et al., *Common genetic variants influence human subcortical brain structures.* Nature, 2015. **520**(7546): p. 224-229.
40. van Dongen, J., et al., *DNA methylation signatures of educational attainment.* npj Science of Learning, 2018. **3**(1): p. 1-14.
41. Van Dongen, J., et al., *Genetic and environmental influences interact with age and sex in shaping the human methylome.* Nature communications, 2016. **7**(1): p. 1-13.
42. Van Dongen, J., et al., *Epigenetic variation in monozygotic twins: a genome-wide analysis of DNA methylation in buccal cells.* Genes, 2014. **5**(2): p. 347-365.
43. Barban, N., et al., *Genome-wide analysis identifies 12 loci influencing human reproductive behavior.* Nature genetics, 2016. **48**(12): p. 1462-1472.
44. Mbarek, H., et al., *Identification of common genetic variants influencing spontaneous dizygotic twinning and female fertility.* The American Journal of Human Genetics, 2016. **98**(5): p. 898-908.
45. Gieger, C., et al., *New gene functions in megakaryopoiesis and platelet formation.* Nature, 2011. **480**(7376): p. 201-208.
46. Ibrahim-Verbaas, C., et al., *GWAS for executive function and processing speed suggests involvement of the CADM2 gene.* Molecular psychiatry, 2016. **21**(2): p. 189-197.
47. Sabater-Lleal, M., et al., *Multiethnic meta-analysis of genome-wide association studies in> 100 000 subjects identifies 23 fibrinogen-associated Loci but no strong evidence of a causal association between circulating fibrinogen and cardiovascular disease.* Circulation, 2013. **128**(12): p. 1310-1324.
48. Wain, L.V., et al., *Genome-wide association study identifies six new loci influencing pulse pressure and mean arterial pressure.* Nature genetics, 2011. **43**(10): p. 1005-1011.
49. Fraga, M.F. and M. Esteller, *Epigenetics and aging: the targets and the marks.* Trends in Genetics, 2007. **23**(8): p. 413-418.
50. Wong, C.C.Y., et al., *A longitudinal study of epigenetic variation in twins.* Epigenetics, 2010. **5**(6): p. 516-526.
51. Chen, B.H., et al., *DNA methylation-based measures of biological age: meta-analysis predicting time to death.* Aging (Albany NY), 2016. **8**(9): p. 1844.
52. Horvath, S. and K. Raj, *DNA methylation-based biomarkers and the epigenetic clock theory of ageing.* Nature Reviews Genetics, 2018. **19**(6): p. 371.
53. Organization, W.H., *World report on ageing and health.* 2015: World Health Organization.
54. Turner, K.J., V. Vasu, and D.K. Griffin, *Telomere biology and human phenotype.* Cells, 2019. **8**(1): p. 73.
55. Allsopp, R.C., et al., *Telomere length predicts replicative capacity of human fibroblasts.* Proceedings of the National Academy of Sciences, 1992. **89**(21): p. 10114-10118.
56. Counter, C.M., et al., *Telomere shortening associated with chromosome instability is arrested in immortal cells which express telomerase activity.* The EMBO journal, 1992. **11**(5): p. 1921-1929.
57. Wagner, W., et al., *Aging and replicative senescence have related effects on human stem and progenitor cells.* PloS one, 2009. **4**(6): p. e5846.

58. Vaziri, H., et al., *Evidence for a mitotic clock in human hematopoietic stem cells: loss of telomeric DNA with age.* Proceedings of the National Academy of Sciences, 1994. **91**(21): p. 9857-9860.
59. Bakaysa, S.L., et al., *Telomere length predicts survival independent of genetic influences.* Aging cell, 2007. **6**(6): p. 769-774.
60. Deelen, J., et al., *Leukocyte telomere length associates with prospective mortality independent of immune-related parameters and known genetic markers.* International journal of epidemiology, 2014. **43**(3): p. 878-886.
61. Gomes, N.M., et al., *Comparative biology of mammalian telomeres: hypotheses on ancestral states and the roles of telomeres in longevity determination.* Aging cell, 2011. **10**(5): p. 761-768.
62. Sahin, E. and R.A. DePinho, *Linking functional decline of telomeres, mitochondria and stem cells during ageing.* nature, 2010. **464**(7288): p. 520-528.
63. Levy, S.E. and R.M. Myers, *Advancements in next-generation sequencing.* Annual review of genomics and human genetics, 2016. **17**: p. 95-115.
64. Sender, R., S. Fuchs, and R. Milo, *Revised estimates for the number of human and bacteria cells in the body.* PLoS biology, 2016. **14**(8): p. e1002533.
65. Shreiner, A.B., J.Y. Kao, and V.B. Young, *The gut microbiome in health and in disease.* Current opinion in gastroenterology, 2015. **31**(1): p. 69.
66. Turnbaugh, P.J., et al., *An obesity-associated gut microbiome with increased capacity for energy harvest.* nature, 2006. **444**(7122): p. 1027.
67. Blaser, M.J., *Harnessing the power of the human microbiome.* Proceedings of the National Academy of Sciences, 2010. **107**(14): p. 6125-6126.
68. Consortium, H.M.J.R.S., *A catalog of reference genomes from the human microbiome.* Science, 2010. **328**(5981): p. 994-999.
69. Schloss, P.D., et al., *Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities.* Applied and environmental microbiology, 2009. **75**(23): p. 7537-7541.
70. Westcott, S.L. and P.D. Schloss, *De novo clustering methods outperform reference-based methods for assigning 16S rRNA gene sequences to operational taxonomic units.* PeerJ, 2015. **3**: p. e1487.
71. Costello, E.K., et al., *Bacterial community variation in human body habitats across space and time.* Science, 2009. **326**(5960): p. 1694-1697.
72. Knights, D., E.K. Costello, and R. Knight, *Supervised classification of human microbiota.* FEMS microbiology reviews, 2011. **35**(2): p. 343-359.
73. Turner, P.V., *The role of the gut microbiota on animal model reproducibility.* Animal models and experimental medicine, 2018. **1**(2): p. 109-115.
74. Matamoros, S., et al., *Development of intestinal microbiota in infants and its impact on health.* Trends in microbiology, 2013. **21**(4): p. 167-173.
75. Koenig, J.E., et al., *Succession of microbial consortia in the developing infant gut microbiome.* Proceedings of the National Academy of Sciences, 2011. **108**(Supplement 1): p. 4578-4585.
76. van den Elsen, L.W., et al., *Shaping the gut microbiota by breastfeeding: the gateway to allergy prevention?* Frontiers in pediatrics, 2019. **7**: p. 47.

77. Cioffi, C.C., et al., *History of breastfeeding but not mode of delivery shapes the gut microbiome in childhood.* PloS one, 2020. **15**(7): p. e0235223.
78. Flaherman, V.J., et al., *The effect of early limited formula on breastfeeding, readmission, and intestinal microbiota: a randomized clinical trial.* The Journal of pediatrics, 2018. **196**: p. 84-90. e1.
79. Azad, M.B., et al., *Gut microbiota of healthy Canadian infants: profiles by mode of delivery and infant diet at 4 months.* Cmaj, 2013. **185**(5): p. 385-394.
80. Jakobsson, H.E., et al., *Decreased gut microbiota diversity, delayed Bacteroidetes colonisation and reduced Th1 responses in infants delivered by caesarean section.* Gut, 2014. **63**(4): p. 559-566.
81. Ho, N.T., et al., *Meta-analysis of effects of exclusive breastfeeding on infant gut microbiota across populations.* Nature communications, 2018. **9**(1): p. 1-13.
82. Forbes, J.D., et al., *Association of exposure to formula in the hospital and subsequent infant feeding practices with gut microbiota and risk of overweight in the first year of life.* JAMA pediatrics, 2018. **172**(7): p. e181161-e181161.
83. Borewicz, K., et al., *The effect of prebiotic fortified infant formulas on microbiota composition and dynamics in early life.* Scientific reports, 2019. **9**(1): p. 1-13.
84. McFarland, L.V., C.T. Evans, and E.J. Goldstein, *Strain-specificity and disease-specificity of probiotic efficacy: a systematic review and meta-analysis.* Frontiers in medicine, 2018. **5**: p. 124.
85. Esaiassen, E., et al., *Effects of probiotic supplementation on the gut microbiota and antibiotic resistome development in preterm infants.* Frontiers in pediatrics, 2018. **6**: p. 347.
86. Langdon, A., N. Crook, and G. Dantas, *The effects of antibiotics on the microbiome throughout development and alternative approaches for therapeutic modulation.* Genome medicine, 2016. **8**(1): p. 39.
87. Hermansson, H., et al., *Breast milk microbiota is shaped by mode of delivery and intrapartum antibiotic exposure.* Frontiers in nutrition, 2019. **6**: p. 4.
88. Mueller, N.T., et al., *The infant microbiome development: mom matters.* Trends in molecular medicine, 2015. **21**(2): p. 109-117.
89. Biedermann, L., et al., *Smoking cessation induces profound changes in the composition of the intestinal microbiota in humans.* PloS one, 2013. **8**(3): p. e59260.
90. Biedermann, L., et al., *Smoking cessation alters intestinal microbiota: insights from quantitative investigations on human fecal samples using FISH.* Inflammatory bowel diseases, 2014. **20**(9): p. 1496-1501.
91. Brito, I.L., et al., *Transmission of human-associated microbiota along family and social networks.* Nature microbiology, 2019. **4**(6): p. 964-971.
92. Dill-McFarland, K.A., et al., *Close social relationships correlate with human gut microbiota composition.* Scientific reports, 2019. **9**(1): p. 1-10.
93. Evans, C.C., et al., *Exercise prevents weight gain and alters the gut microbiota in a mouse model of high fat diet-induced obesity.* PloS one, 2014. **9**(3): p. e92193.
94. Finnicum, C.T., et al., *Cohabitation is associated with a greater resemblance in gut microbiota which can impact cardiometabolic and inflammatory risk.* BMC microbiology, 2019. **19**(1): p. 1-10.
95. Lee, S.H., et al., *Association between cigarette smoking status and composition of gut microbiota: population-based cross-sectional study.* Journal of clinical medicine, 2018. **7**(9): p. 282.

96. Ley, R.E., et al., *Human gut microbes associated with obesity.* nature, 2006. **444**(7122): p. 1022-1023.
97. Matsumoto, M., et al., *Voluntary running exercise alters microbiota composition and increases n-butyrate concentration in the rat cecum.* Bioscience, biotechnology, and biochemistry, 2008. **72**(2): p. 572-576.
98. Muegge, B.D., et al., *Diet drives convergence in gut microbiome functions across mammalian phylogeny and within humans.* Science, 2011. **332**(6032): p. 970-974.
99. Song, S.J., et al., *Cohabiting family members share microbiota with one another and with their dogs.* elife, 2013. **2**: p. e00458.
100. Walker, A.W., et al., *Dominant and diet-responsive groups of bacteria within the human colonic microbiota.* The ISME journal, 2011. **5**(2): p. 220-230.
101. Wu, G.D., et al., *Linking long-term dietary patterns with gut microbial enterotypes.* Science, 2011. **334**(6052): p. 105-108.
102. Benson, A.K., et al., *Individuality in gut microbiota composition is a complex polygenic trait shaped by multiple environmental and host genetic factors.* Proceedings of the National Academy of Sciences, 2010. **107**(44): p. 18933-18938.
103. Goodrich, J.K., et al., *Genetic determinants of the gut microbiome in UK twins.* Cell host & microbe, 2016. **19**(5): p. 731-743.
104. Goodrich, J.K., et al., *Human genetics shape the gut microbiome.* Cell, 2014. **159**(4): p. 789-799.
105. Korpela, K., et al., *Selective maternal seeding and environment shape the human gut microbiome.* Genome research, 2018. **28**(4): p. 561-568.
106. Lee, S., et al., *Comparison of the gut microbiotas of healthy adult twins living in South Korea and the United States.* Applied and environmental microbiology, 2011. **77**(20): p. 7433-7437.
107. Monda, V., et al., *Exercise modifies the gut microbiota with positive health effects.* Oxidative medicine and cellular longevity, 2017. **2017**.
108. Rothschild, D., et al., *Environment dominates over host genetics in shaping human gut microbiota.* Nature, 2018. **555**(7695): p. 210-215.
109. Scott, K.P., et al., *The influence of diet on the gut microbiota.* Pharmacological research, 2013. **69**(1): p. 52-60.
110. Turnbaugh, P.J., et al., *A core gut microbiome in obese and lean twins.* nature, 2009. **457**(7228): p. 480-484.
111. Bonder, M.J., et al., *The effect of host genetics on the gut microbiome.* Nature genetics, 2016. **48**(11): p. 1407-1412.
112. Bruce-Keller, A.J., et al., *Obese-type gut microbiota induce neurobehavioral changes in the absence of obesity.* Biological psychiatry, 2015. **77**(7): p. 607-615.
113. de Heredia, F.P., S. Gómez-Martínez, and A. Marcos, *Obesity, inflammation and the immune system.* Proceedings of the Nutrition Society, 2012. **71**(2): p. 332-338.
114. Francescone, R., V. Hou, and S.I. Grivennikov, *Microbiome, inflammation and cancer.* Cancer journal (Sudbury, Mass.), 2014. **20**(3): p. 181.
115. Jiao, N., et al., *Gut microbiome may contribute to insulin resistance and systemic inflammation in obese rodents: a meta-analysis.* Physiological genomics, 2018.
116. Martin-Ruiz, C.M., et al., *Reproducibility of telomere length assessment: an international collaborative study.* International journal of epidemiology, 2015. **44**(5): p. 1673-1683.

117. Omenetti, S. and T.T. Pizarro, *The Treg/Th17 axis: a dynamic balance regulated by the gut microbiome.* Frontiers in immunology, 2015. **6**: p. 639.
118. Xiao, E., et al., *Diabetes enhances IL-17 expression and alters the oral microbiome to increase its pathogenicity.* Cell host & microbe, 2017. **22**(1): p. 120-128. e4.
119. Perry, R.J., et al., *Acetate mediates a microbiome–brain–β-cell axis to promote metabolic syndrome.* Nature, 2016. **534**(7606): p. 213-217.
120. Monteiro, R. and I. Azevedo, *Chronic inflammation in obesity and the metabolic syndrome.* Mediators of inflammation, 2010. **2010**.
121. Rodríguez-Hernández, H., et al., *Obesity and inflammation: epidemiology, risk factors, and markers of inflammation.* International journal of endocrinology, 2013. **2013**.
122. Hjelmborg, J.B., et al., *The heritability of leucocyte telomere length dynamics.* Journal of medical genetics, 2015. **52**(5): p. 297-302.
123. Broer, L., et al., *Meta-analysis of telomere length in 19 713 subjects reveals high heritability, stronger maternal inheritance and a paternal age effect.* European journal of human genetics, 2013. **21**(10): p. 1163-1168.
124. Wang, J., et al., *Meta-analysis of human genome-microbiome association studies: the MiBioGen consortium initiative*. 2018, BioMed Central.



2

RELATIVE TELOMERE REPEAT MASS IN BUCCAL AND LEUKOCYTE-DERIVED DNA

This chapter was published as: Finnicum CT, Dolan CV, Willemsen G, et al. (2017) Relative Telomere Repeat Mass in Buccal and Leukocyte-Derived DNA. PLoS One

ABSTRACT

Telomere length has garnered interest due to the potential role it may play as a biomarker for the cellular aging process. Telomere measurements obtained from blood-derived DNA are often used in epidemiological studies. However, the invasive nature of blood draws severely limits sample collection, particularly with children. Buccal cells are commonly sampled for DNA isolation and thus may present a non-invasive alternative for telomere measurement. Buccal and leukocyte derived DNA samples obtained from participants collected during the same time period were analyzed for telomere repeat mass (TRM). TRM was measured in buccal-derived DNA samples from individuals for whom previous TRM data from blood samples existed. TRM measurement was performed by quantitative polymerase chain reaction (qPCR) and was normalized to the single copy 36B4 gene relative to a reference DNA sample (K562). Correlations between TRM from blood and buccal DNA were obtained and also between the same blood DNA samples measured in separate laboratories. Using the classical twin design, TRM heritability was estimated (N = 1892, MZ = 1044, DZ = 775). Buccal samples measured for TRM showed a significant correlation with the blood-1 (initial TRM measurement) ($R = 0.39$, $p < 0.01$) and blood-2 (TRM at AIHG) ($R = 0.36$, $p < 0.01$) samples. Sex and age effects were observed within the buccal samples as is the norm within blood-derived DNA. The buccal, blood-1, and blood-2 measurements generated heritability estimates of 23.3%, 47.6% and 22.2%, respectively. Buccal derived DNA provides a valid source for the determination of TRM, paving the way for non-invasive projects, such as longitudinal studies in children.

INTRODUCTION

Telomere measurements have been of great interest as a potential tool for assessment of the cellular aging process. The telomere, which caps the end of each strand of DNA, is subjected to attrition throughout the life span due to the end replication problem, which results in the loss of approximately 50-100 base pairs per mitotic division [1]. Once a significant portion of the telomeric region has been lost, the cell enters a state of replicative senescence characterized by a marked change in gene expression, as well as the inability to further divide [2, 3]. Telomere length has also been implicated in the development of many disorders, either as a marker or causal agent [4, 5].

The telomere region consists of hexanucleotide (TTAGGG) repeat sequences, which are associated with multiple protein factors such as the shelterin complex [6, 7]. Although no exons are contained within the telomeric region, it plays a vital role in genomic protection, stability, and can impact regulation of gene expression elsewhere in the genome [8]. In spite of constant telomere degradation over the life span, mechanisms are available for telomere elongation, mainly through the use of the enzyme telomerase. Telomerase is generally inactive in most somatic cells, but activation is a hallmark of immortal cells [9]. The observation of telomere attrition in proliferating cells, as well as the immortality conveyed via telomerase activation, suggests that telomeres act as a central biological clock mechanism [10]. This is compounded by studies highlighting an association between TL and life span in humans [11-15].

Several studies have addressed the genetic contribution to individual differences in TL in humans [16-20], and specific genomic loci associated with mean leukocyte TL have been identified [16]. In addition to genetic factors, multiple factors such as smoking, sedentary behavior, and periods of high stress, which themselves are partly genetic, may contribute to individual differences in TL [21].

In order to address the role of telomere dynamics in the development of both aging and specific disease pathologies, it is necessary to perform telomere measurements in a longitudinal manner. Investigations into telomere attrition across multiple time points would shed light on differences in telomere attrition over age. However, this presents a challenge as DNA derived from circulating leukocytes obtained by intravenous blood draw is currently the most widely used DNA source for telomere studies. It has been observed that telomere measurements are correlated among somatic tissues regardless of replicative capacities [22-24]. This presents the possibility of utilizing other cellular sources of DNA for telomere measurement studies. Buccal-derived DNA samples are easily collected and are commonly used in biomedical research [25-27]. The use of buccal-derived DNA in place of leukocyte-derived DNA would greatly facilitate large-scale telomere measurement studies. Buccal swab samples are generally composed of buccal epithelial cells but can also contain a small fraction of leukocytes [28].

The use of qPCR for quantifying relative telomere abundance does not provide an estimate of definitive telomere length due to telomere heterogeneity across chromosomes, rather this technique allows for the determination of the relative abundance of telomere repeat mass present within a sample. Due to this, qPCR-based telomere measures are referred to as TRM rather than TL. Here we compared telomere repeat mass measures based on buccal-derived DNA with TRM measures based on leukocyte-derived DNA. The blood and buccal samples were obtained in a sample of monozygotic and dizygotic twins. The twin data provide us with the unique opportunity to estimate the heritability of buccal and blood-based TRM, and to estimate the genetic correlation between the two measures. Our aim is to determine whether TRM measures in buccal-derived DNA are suitable for large scale studies of TRM.

The original leukocyte-derived DNA, which was previously measured for TRM, was subjected to a second TRM measurement to compare the effects of sample handling on TRM measurements. It has been documented that variations in the DNA extraction processes have an impact on telomere measurements, which may have implications for large epidemiological studies [29, 30] as repeated handling of genomic material may lead to changes in the telomere regions, thus altering TRM results. We addressed this issue here as well, as it is relevant to the design of epidemiological studies and to biobanking procedures.

MATERIALS AND METHODS

PARTICIPANTS

Blood and buccal samples for DNA extraction were obtained concurrently from individuals in the Netherlands Twin Register [31, 32], as previously described [33]. The telomeric DNA from blood samples was measured twice, once in Leicester (England) as a part of a previous study [20] (Blood-1), and once at the Avera Institute for Human Genetics (AIHG; Blood-2). The buccal DNA TRM measurement was performed only once at the AIHG (Buccal). The total sample size comprises 1892 individuals clustered in 1133 families. The individuals include 1809 twins (271 MZ male, 773 MZ female, 156 DZ male, 320 DZ female, and 299 DZ opposite sex), 77 siblings (of whom 12 are multiples, e.g., a member of a triplet), 5 mothers and 1 father. There were 618 MZ twin pairs and 501 DZ twin pairs. Zygosity was based on genome-wide single nucleotide polymorphism (SNP) data [34]. The 1892 individuals are distributed over the 1133 families as follows: 1 member (429 families), 2 members (652 families), 3 members (49 families) and 4 members (3 families). Of the 1892 individuals, 589 were males (31%) and 1309 (69%) were females. The mean age was 34.18 years (SD = 13.2, min = 12, max = 78). The Blood-1 TRM measures are available in all 1892 individuals. The measures were distributed over 33 batches (plates), with the mean number per

batch equal to 57.3 (SD = 15.3). The Blood-2 TRM measures are available in N = 1338, distributed over 40 batches (mean number per batch 34.3, SD = 12.3). The Buccal TRM measures are available in 1691, distributed over 40 batches (mean number per batch 42.2, SD = 10.5). The study was approved by the Central Ethics Committee on Research Involving Human Subjects of the VU Medical Center, Amsterdam, NL. All participants and/or parents provided informed consent for the study. Written informed consent was obtained from the appropriate next of kin, caretakers, or guardians for all minors present within the study.

MEASUREMENT OF TELOMERE REPEAT MASS

The TRM analysis performed in Leicester has been described previously [20]. At AIHG TRM measurements were generated by qPCR to amplify the telomeric region of the DNA, as well as amplification of a single copy gene for normalization purposes. It is important to note that previously generated telomere measurements were referred to as telomere length whereas in this manuscript the results of qPCR-based telomere analysis are referred to as telomere repeat mass. The primers used to amplify the telomeric region were tel1B (5'-CGGTTTGTTTGGGTTTGGGTTTGGGTTTGGGTTTGGGTT-3') (600nM) and tel2B (5'-GGCTTGCCTTACCCTTACCCTTACCCTTACCCTTACCCT-3') (600nM). The single copy reference gene was amplified using 36b4U (5'-CAGCAAGTGGGAAGGTGTAATCC-3') (300nM) and 36b4D (5'-CCCATTCTATCATCAACGGGTACAA-3') (500nM). Each qPCR reaction contained 20 ng of DNA as input, and all samples were run in triplicate (Applied Biosystems ViiA 7). The cycling conditions for both reactions were an initial 95°C for 15 minutes, followed by 30 cycles of 95°C for 15 seconds, and 58°C for 1 minute. Contained within each qPCR batch (telomere and 36b4) was a calibrator sample K562 (Promega, Madison, WI) used for comparison using Cawthon's calculation method [35]. A standard curve ranging from 100 ng to 6.25 ng was included in each telomere and 36B4 qPCR run. TRM was expressed as a ratio of the telomere region (T) to the single copy 36b4 (S), resulting in a T/S ratio. Each TRM measurement was compared to the K562 sample included with each plate, resulting in a relative TRM ratio.

STATISTICAL ANALYSIS

All TRM measures were first corrected for batch effects by means of an analysis of variance (ANOVA), with batch as the fixed factor. Unless stated otherwise, the residuals were used in the subsequent analyses. We studied the effects of age and sex on TRM and the linear association between the TRM measures with linear regression analysis, using generalized estimating equations (GEE) [36]. We corrected the standard errors for the dependency in the data due to family clustering [37]. We regressed Buccal TRM on Blood-2 TL TRM and on Blood-1 TRM separately. In addition, we regressed Blood-2 TRM on Blood-1 TRM to estimate the association between the blood-based measures of TRM (Table 2.1). In all analyses, we included sex and age as covariates.

Table 2.1 | Regression analyses of TRM measures (standard errors in parentheses). ** p<0.01; % variance is the variance explained by the predictor. The percentage in parentheses is due to the predictor + age + sex. The parameter b is the raw regression coefficient in the regression analyses including the covariates age and sex.

Dependent	Predictor	b (st err)	% variance
Buccal TRM	Blood-2	0.286 (0.0314)**	10.8% (13.9%)
Buccal TRM	Blood-1	0.208 (0.0183)**	12.2% (15.3%)
Blood-2 TRM	Blood-1	0.373 (0.0180)**	30.4% (39.4%)

The data from twins were used to estimate the contributions of genetic and environmental influences to the phenotypic (co)variance of the TRM measures. Twins raised together form the basis of the classical twin design, which exploits the fact that MZ twins are genetically (nearly) identical, while DZ twin share on average 50% of their alleles [38-40]. The CTD allows us to fit an ACE model, which includes additive genetic (A), shared environmental (i.e., shared by the twins; C), and unshared environmental effects (E) on TRM. The phenotypic TRM variance is modeled as $\sigma^2_{TRM} = \sigma^2_A + \sigma^2_C + \sigma^2_E$, and the twin covariances are modeled as $\sigma_{TRM1\text{-}TRM2} = \sigma^2_A + \sigma^2_C$ in MZs, and $\sigma_{TRM1\text{-}TRM2} = 0.5*\sigma^2_A + \sigma^2_C$ in DZs. We obtain standardized estimates, usually denoted h^2, c^2, and e^2, by calculating $h^2 = \sigma^2_A/\sigma^2_{TRM}$, $c^2 = \sigma^2_C/\sigma^2_{TRM}$, and $e^2 = \sigma^2_E/\sigma^2_{TRM}$. Note that h^2 is the heritability (hence h^2 rather than a^2, although the notation is arbitrary). Dropping C (or A) from the model reduces the model to an AE (CE) model. The statistical significance of variance components (e.g., σ_C^2 or σ_A^2) can be tested by means of a likelihood ratio test. It is well established that TRM decreases with age (see Table 2.2). It is also possible that contributions of A, C, and (or) E to the phenotypic variance changes with age. To investigate this, we fit a moderated ACE model, in which the A, C, and E variance components are free to vary in magnitude with age [41].

The univariate twin model can readily be extended to the multivariate case [42]. That is, we can decompose the phenotypic 3x3 covariance matrix of the 3 TRM measures, Σ_{TRM}, into genetic and environmental components analogously to the univariate case: $\Sigma_{TRM} = \Sigma_A + \Sigma_C + \Sigma_E$, where the Σ_A, Σ_C and Σ_E represent the additive genetic, shared and unshared environmental 3x3 covariance matrices, respectively. The twin 1—twin 2 covariance matrix is modeled $\Sigma_{TRM1\text{-}TRM2} = \Sigma_A + \Sigma_C$ in the MZs, and $\Sigma_{TRM1\text{-}TRM2} = 0.5*\Sigma_A + \Sigma_C$ in the DZs. This decomposition reveals the contributions of genetic and environmental effects to the phenotypic variances and covariances amongst the TRM measures. We used the full information maximum likelihood (FIML) estimation in the OpenMx R library to fit the twin models [43]. We first estimated the MZ and DZ covariance matrices. Note that these are 6x6 because we have 3 TRM measures, observed in two twin members. Table 2.3 contains the FIML estimates of the correlation and covariance matrices in the MZ and DZ twins estimated with age and sex as covariates. Table 2.3 also includes the number of observed values for each phenotype.

Table 2.2 | Age and sex effects on batch corrected TRM (standard errors in parentheses). ** p<0.01, *p<0.05, % variance is the variance explained by sex and age. The percentage in parentheses is due to age alone. The parameters b are the raw regression coefficients.

variable	N	b (sex)	b (age)	% variance
Blood-2 TRM	1338	.0678 (.0168)**	-.0067 (.00067)**	9.0 (7.5)
Buccal TRM	1691	.0306 (.0156)*	-.0036 (.00059)**	3.1 (2.8)
Blood-1 TRM	1892	.1130 (.0217)**	-.0105 (.00082)**	10.1 (8.5)

Table 2.3 | Full information maximum likelihood estimates of twin variances, covariances, and correlations, corrected for sex and age. The correlations are shown below the diagonal in italics. The within phenotypic correlations are underlined. N represents the number of observed values.

		MZ twin 1			**MZ Twin 2**	
	Blood-1	Blood-2	Buccal	Blood-1	Blood-2	Buccal
N	375	257	321	390	261	338
Blood-1	0.179	0.062	0.037	0.117	0.057	0.042
Blood-2	*0.518*	0.079	0.019	0.055	0.028	0.025
Buccal	*0.319*	*0.244*	0.076	0.046	0.023	0.036
Blood-1	*0.655*	*0.466*	*0.393*	0.177	0.074	0.044
Blood-2	*0.512*	*0.384*	*0.313*	*0.676*	0.068	0.026
Buccal	*0.356*	*0.32*	*0.467*	*0.38*	*0.354*	0.077
		DZ twin 1			**DZ twin 2**	
N	522	410	478	522	402	478
Blood-1	0.177	0.06	0.051	0.076	0.043	0.033
Blood-2	*0.565*	0.063	0.028	0.032	0.025	0.021
Buccal	*0.415*	*0.383*	0.086	0.034	0.017	0.03
Blood-1	*0.422*	*0.296*	*0.274*	0.182	0.07	0.038
Blood-2	*0.346*	*0.345*	*0.198*	*0.563*	0.085	0.028
Buccal	*0.301*	*0.311*	*0.373*	*0.324*	*0.356*	0.075

ACE MODELING

We investigated the contributions of genetic and environmental influences to the phenotypic covariance matrices by fitting an ACE model, with age and sex as covariates. As mentioned above, in fitting the ACE model, we modeled the 3x3 phenotypic covariance matrix (Σ_{TRM}) as $\Sigma_{TRM} = \Sigma_A + \Sigma_C + \Sigma_E$, where Σ_A is the 3x3 additive genetic, Σ_C is the 3x3 shared environmental, and Σ_E is the 3x3 unshared environmental covariance matrix. In Figure 2.1 and in Table 2.4, we express each covariance matrix as $\mathbf{\Sigma} = \mathbf{DRD}^t$, where **D** is a (3x3) diagonal matrix containing the standard deviations, and **R** is the 3x3 correlation matrix. By calculating Σ_A / Σ_{TRM}, we obtain the contribution of additive genetic effect to the phenotypic variances (i.e., diagonal elements of Σ_A / Σ_{TRM}) and covariances (off-diagonals elements of Σ_A / Σ_{TRM}). The diagonal elements are the heritabilities (h^2). The same applies to the environmental covariance matrices Σ_C / Σ_{TRM} and Σ_E / Σ_{TRM}, where the standardized diagonals are the c^2s and, e^2s, respectively.

Table 2.4 | Parameter estimates in the ACE model. The R_A is the additive genetic correlation matrix, and stdev A are the additive genetic standard deviations. Σ_A/Σ_{TRM} is the contribution of additive genetic effects to the TRM variances (h^2) and the covariances (the shared and unshared environmental results are defined analogously). Note that the h^2, c^2, e^2 estimates appear twice.

	Blood-1	**Blood-2**	**Buccal**		**Blood-1**	**Blood-2**	**Buccal**		**Blood-1**	**Blood-2**	**Buccal**
stdev A	0.291	0.128	0.135	**stdev C**	0.179	0.117	0.138	**stdev E**	0.248	0.209	0.202
R_A	1			R_C	1			R_E	1		
	0.983	1			0.895	1			0.219	1	
	0.663	0.789	1		0.788	0.678	1		-0.062	0.000	1
Σ_A/Σ_{TRM}	0.476			Σ_C/Σ_{TRM}	0.179			Σ_E/Σ_{TRM}	0.345		
	0.549	0.222			0.281	0.186			0.170	0.591	
	0.615	0.554	0.233		0.459	0.447	0.244		-0.074	-0.001	0.522
h^2	0.476	0.222	0.233	c^2	0.179	0.186	0.244	e^2	0.345	0.591	0.522
CI lower	0.276	0.070	0.037	**CI lower**	0.013	0.002	0.043	**CI lower**	0.299	0.519	0.456
CI upper	0.654	0.437	0.459	**CI upper**	0.363	0.347	0.420	**CI upper**	0.397	0.669	0.595

Finally, we tested whether the estimates of the variance components in univariate ACE models (σ^2_A, σ^2_C, and σ^2_E) varied with age. We did this by modeling σ_A as $\sigma_A = b_{0A} + b_{1A}$*age, $\sigma_C = b_{0C}+b_{1C}$*age, and $\sigma_E = b_{0E} + b_{1E}$*age [41]. That is, we considered the possibility that the genetic and environmental standard deviations linearly increase (or decrease) with age. The test of age moderation boils down to the test of the omnibus hypothesis: $b_{1A} = b_{1C} = b_{1E} = 0$.

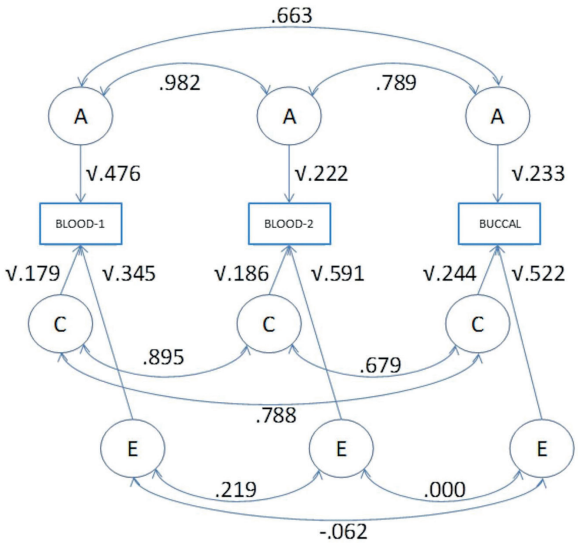


Figure 2.1 | Path diagram depicting the A, C, and E variance components calculated for each of the sample groups.

RESULTS

Age and sex together explained 9%, 3.1%, and 10.1% of the variance in Blood-2 TRM, Buccal TRM, and Blood-1 TRM, respectively. The percentages of variance explained by age alone equaled 7.5%, 2.8%, and 8.5%, respectively. We tested for the interaction (sex X age), but this interaction was consistently insignificant ($p > 0.1$). Batch effects accounted for 12.4% (TRM Blood-1), 23.1% (TRM Blood-2), and 23.3% (TRM Buccal) of the variance. The zero-order correlations among TRM measures were 0.36 (Blood-2—Buccal), 0.39 (Buccal—Blood-1), and 0.62 (Blood-2 –Blood-1).

Age and sex corrected twin correlations estimated through FIML were 0.655 for MZ twins and 0.422 for DZ twins in the Blood-1 measurements. The Blood-2 measurements performed at the AIHG had MZ and DZ correlations of 0.384 and 0.345, respectively. Similar analyses were performed on the Buccal samples yielding a MZ correlation of 0.467 and a DZ correlation of 0.373 (Table 2.5).

Table 2.4 contains the parameter estimates of the (trivariate) ACE model, including the estimates of the standardized variance components (h^2, c^2, and e^2), with their upper and lower 95% CIs (last three rows). The heritability of the Blood-1 TRM measure is 0.476, i.e., about 47.6% of the phenotypic variance is due to genetic effects. The C and E effects account for about 17.9% and 34.5% of the TRM variance, respectively. The standardized variance components of the Blood-2 TRM and Buccal TRM equal $h^2 = 0.222$ and $h^2 = 0.233$, respectively with C effects accounting for 18.6 and 24.4%, and E effects accounting for 51.9 and 45.6% of the Blood-2 and Buccal TRM variance, respectively. We note that the E influences contribute relatively little to the covariance among the TRM measures (see Σ_E/Σ_{TRM} in Table 2.5); whereas both the A and C effects contribute considerably to the covariances among the blood and buccal TRM measures. The genetic correlations between Buccal TRM and the Blood-1 and Blood-2 TRM are 0.663 and 0.789, respectively. The genetic correlation between Blood-1 and 2 TRM is 0.983.

We tested the significance of the A and C effect by dropping these from the model. Dropping C resulted in $chi^2(6) = 10.97$ ($p = 0.089$). In contrast, dropping A resulted in $chi^2(6) = 49.15$ ($p<0.001$). From this, we may conclude that the contributions of C to the phenotypic covariance matrix is not significant, but the contributions of A are. However, the fact that we cannot reject the hypothesis $\Sigma_C = 0$ is likely to be due to a lack of statistical power [45]. We therefore retained the ACE results.

Next, we tested whether variance components were moderated by age. For both blood DNA TRM measures and for buccal TRM measures, the results indicate that there is no evidence to support the hypothesis of age moderation of the variance components σ^2_A, σ^2_C, and σ^2_E, with the $chi^2(3)$ for Blood-1 being equal to 4.32 ($p = 0.23$); for Blood-2 1.46 ($p = 0.69$), and for Buccal 3.38 ($p = 0.37$). Based on these results we conclude that there is no evidence to support the hypothesis of age moderation of the variance components σ^2_A, σ^2_C, and σ^2_E.

Table 2.5 | FIML estimates of MZ and DZ twin correlations for the TRM measures (corrected for sex and age), with lower and upper 95% confidence intervals.

	95% lower	**correlation**	**95% upper**
MZ Blood-1	0.598	0.655	0.704
MZ Blood-2	0.289	0.384	0.469
MZ Buccal	0.388	0.467	0.537
DZ Blood-1	0.312	0.422	0.516
DZ Blood-2	0.209	0.345	0.464
DZ Buccal	0.268	0.373	0.468

DISCUSSION

The ability to utilize buccal samples in lieu of a blood sample would greatly increase the ability to perform longitudinal TRM studies. Due to the negligible invasiveness of buccal DNA sampling, future studies may be designed that may span a long period of time, including TRM measurement at birth. Utilizing blood and buccal-derived DNA collected from 1892 participants, mostly twins, we were able to investigate the relationship between DNA derived from different cellular sources, as well as investigate the genetic and environmental components associated with TRM.

The buccal-derived TRM measurements showed a significant association with both sex and age indicating that the TRM data is showing an expected result (telomere attrition). Similar observations have been widely observed in previous studies [46, 47]. This is evidence of similarities in telomere dynamics between the tissue types, which would allow for use of buccal-derived DNA samples for telomere measurement studies. Note that the effect of age on blood based TRM is appreciably greater than the effect on buccal based TRM. This may be due to greater error variance of buccal-derived TRM measurements. We address the contributions of genetic and environmental influences to these phenotypic associations in the analyses of the twin data.

Buccal samples showed a significant phenotypic correlation with both of the blood measurements performed on the same sample multiple years apart. This finding highlights the ability of buccal-derived DNA samples to characterize the cellular aging process in a similar manner as blood-derived DNA. The blood and buccal samples showed similarity compared to measurements regardless of the laboratory performing the assay.

Using twin data to observe phenotypic correlations between MZ and DZ as well as to fit ACE models was informative in yielding estimates of genetic and environmental influences on the TRM phenotype of the sample types under study. Given an AE model we would expect the DZ correlations to be about half the MZ correlations. However, the DZ correlations are clearly larger, which suggests the presence of shared environmental influences [19]. These correlations are consistent with an ACE model. We note that the blood based TRM measures correlate about 0.58 (0.518 and 0.565 in the MZs and 0.676 and 0.563 in the DZs). The correlation between the blood based and buccal based TRM measures are smaller, ranging from 0.244 to 0.415. The two blood measurements showed a difference in estimated heritability with the Blood-1 estimate at 46.7% of total phenotypic variance, whereas the repeated blood measurement and the buccal sample measurements both showed a heritability estimate of 22.2% and 23.3% respectively. Discrepancies within the TRM measurements replicated on the same blood samples may arise due to a combination of inter-lab variation and possible degradation of samples due to extended handling.

Measurement of the blood samples at different time points allowed for information to be derived concerning the effects of extended handling, as well as inter-lab variation. There have been questions raised regarding the reliability of the relative TRM measurements produced in different laboratories [48, 49]. This study showed a difference in the heritability estimates of TRM both produced in replicated blood samples. The blood samples were first measured in one laboratory, shipped elsewhere, utilized for genomic analysis, and then finally shipped for TRM analysis a second time. It is possible that the extended sample handling, as well as known inter-lab variation in TRM measurement, is responsible for the differences in heritability estimates observed.

Having the ability to easily sample buccal-derived DNA would open the doors to further large-scale longitudinal sample collections for TRM measurement. The negligible invasiveness of the collection process makes collection possible from an early age. Cohorts such as those included within the NTR can be followed over multiple time-points in order to investigate temporal effects on TRM throughout an individual's life span.

REFERENCES

1. Allsopp, R.C., et al., *Telomere length predicts replicative capacity of human fibroblasts.* Proceedings of the National Academy of Sciences, 1992. **89**(21): p. 10114-10118.
2. Counter, C.M., et al., *Telomere shortening associated with chromosome instability is arrested in immortal cells which express telomerase activity.* The EMBO journal, 1992. **11**(5): p. 1921-1929.
3. Wagner, W., et al., *Aging and replicative senescence have related effects on human stem and progenitor cells.* PloS one, 2009. **4**(6): p. e5846.
4. Oliveira, B.S., et al., *Systematic review of the association between chronic social stress and telomere length: A life course perspective.* Ageing research reviews, 2016. **26**: p. 37-52.
5. Blackburn, E.H., E.S. Epel, and J. Lin, *Human telomere biology: a contributory and interactive factor in aging, disease risks, and protection.* Science, 2015. **350**(6265): p. 1193-1198.
6. Moyzis, R.K., et al., *A highly conserved repetitive DNA sequence,(TTAGGG) n, present at the telomeres of human chromosomes.* Proceedings of the National Academy of Sciences, 1988. **85**(18): p. 6622-6626.
7. Palm, W. and T. de Lange, *How shelterin protects mammalian telomeres.* Annual review of genetics, 2008. **42**: p. 301-334.
8. Baur, J.A., et al., *Telomere position effect in human cells.* Science, 2001. **292**(5524): p. 2075-2077.
9. Collins, K. and J.R. Mitchell, *Telomerase in the human organism.* Oncogene, 2002. **21**(4): p. 564-579.
10. Vaziri, H., et al., *Evidence for a mitotic clock in human hematopoietic stem cells: loss of telomeric DNA with age.* Proceedings of the National Academy of Sciences, 1994. **91**(21): p. 9857-9860.
11. Bakaysa, S.L., et al., *Telomere length predicts survival independent of genetic influences.* Aging cell, 2007. **6**(6): p. 769-774.
12. Deelen, J., et al., *Leukocyte telomere length associates with prospective mortality independent of immune-related parameters and known genetic markers.* International journal of epidemiology, 2014. **43**(3): p. 878-886.
13. Gomes, N.M., et al., *Comparative biology of mammalian telomeres: hypotheses on ancestral states and the roles of telomeres in longevity determination.* Aging cell, 2011. **10**(5): p. 761-768.
14. Sahin, E. and R.A. DePinho, *Linking functional decline of telomeres, mitochondria and stem cells during ageing.* nature, 2010. **464**(7288): p. 520-528.
15. Heidinger, B.J., et al., *Telomere length in early life predicts lifespan.* Proceedings of the National Academy of Sciences, 2012. **109**(5): p. 1743-1748.
16. Andrew, T., et al., *Mapping genetic loci that determine leukocyte telomere length in a large sample of unselected female sibling pairs.* The American Journal of Human Genetics, 2006. **78**(3): p. 480-486.
17. Vasa-Nicotera, M., et al., *Mapping of a major locus that determines telomere length in humans.* The American Journal of Human Genetics, 2005. **76**(1): p. 147-151.
18. Slagboom, P.E., S. Droog, and D.I. Boomsma, *Genetic determination of telomere size in humans: a twin study of three age groups.* American journal of human genetics, 1994. **55**(5): p. 876.

19. Broer, L., et al., *Meta-analysis of telomere length in 19 713 subjects reveals high heritability, stronger maternal inheritance and a paternal age effect.* European journal of human genetics, 2013. **21**(10): p. 1163-1168.
20. Codd, V., et al., *Identification of seven loci affecting mean telomere length and their association with disease.* Nature genetics, 2013. **45**(4): p. 422-427.
21. Mathur, M.B., et al., *Perceived stress and telomere length: a systematic review, meta-analysis, and methodologic considerations for advancing the field.* Brain, behavior, and immunity, 2016. **54**: p. 158-169.
22. Daniali, L., et al., *Telomeres shorten at equivalent rates in somatic tissues of adults.* Nature communications, 2013. **4**(1): p. 1-7.
23. Gardner, J.P., et al., *Telomere dynamics in macaques and humans.* The Journals of Gerontology Series A: Biological Sciences and Medical Sciences, 2007. **62**(4): p. 367-374.
24. Friedrich, U., et al., *Telomere length in different tissues of elderly patients.* Mechanisms of ageing and development, 2000. **119**(3): p. 89-99.
25. Min, J.L., et al., *High microsatellite and SNP genotyping success rates established in a large number of genomic DNA samples extracted from mouth swabs and genotypes.* Twin Research and Human Genetics, 2006. **9**(4): p. 501-506.
26. Meulenbelt, I., et al., *High-yield noninvasive human genomic DNA isolation method for genetic studies in geographically dispersed families and populations.* American journal of human genetics, 1995. **57**(5): p. 1252.
27. Beekman, M., et al., *A powerful and rapid approach to human genome scanning using small quantities of genomic DNA.* Genetics Research, 2001. **77**(2): p. 129-134.
28. Slieker, R.C., et al., *Identification and systematic annotation of tissue-specific differentially methylated regions using the Illumina 450k array.* Epigenetics & chromatin, 2013. **6**(1): p. 1-12.
29. Cunningham, J.M., et al., *Telomere length varies by DNA extraction method: implications for epidemiologic research.* Cancer Epidemiology and Prevention Biomarkers, 2013. **22**(11): p. 2047-2054.
30. Hofmann, J.N., et al., *Telomere length varies by DNA extraction method: implications for epidemiologic research.* Cancer Epidemiology and Prevention Biomarkers, 2014. **23**(6): p. 1129-1130.
31. Boomsma, D.I., et al., *Netherlands Twin Register: from twins to twin families.* Twin Research and Human Genetics, 2006. **9**(6): p. 849-857.
32. Willemsen, G., et al., *The Adult Netherlands Twin Register: twenty-five years of survey and biological data collection.* Twin Research and Human Genetics, 2013. **16**(1): p. 271-281.
33. Willemsen, G., et al., *The Netherlands Twin Register biobank: a resource for genetic epidemiological studies.* Twin Research and Human Genetics, 2010. **13**(3): p. 231-245.
34. Scheet, P., et al., *Twins, tissue, and time: an assessment of SNPs and CNVs.* Twin Research and Human Genetics, 2012. **15**(6): p. 737-745.
35. Cawthon, R.M., *Telomere measurement by quantitative PCR.* Nucleic acids research, 2002. **30**(10): p. e47-e47.
36. Dobson, A.J. and A.G. Barnett, *An introduction to generalized linear models*. 2018: CRC press.
37. Minică, C.C., et al., *Sandwich corrected standard errors in family-based genome-wide association studies.* European journal of human genetics, 2015. **23**(3): p. 388-394.

38. Keller, M.C., S.E. Medland, and L.E. Duncan, *Are extended twin family designs worth the trouble? A comparison of the bias, precision, and accuracy of parameters estimated in four twin family models.* Behavior genetics, 2010. **40**(3): p. 377-393.
39. Falconer, D. and T. Mackay, *Introduction to quantitative genetics. Essex.* UK: Longman Group, 1996.
40. Plomin, R., J.C. DeFries, and G.E. McClearn, *Behavioral genetics*. 2008: Macmillan.
41. Purcell, S., *Variance components models for gene–environment interaction in twin analysis.* Twin Research and Human Genetics, 2002. **5**(6): p. 554-571.
42. Martin, N.G. and L.J. Eaves, *The genetical analysis of covariance structure.* Heredity, 1977. **38**(1): p. 79-95.
43. Boker, S., et al., *OpenMx: an open source extended structural equation modeling framework.* Psychometrika, 2011. **76**(2): p. 306-317.
44. Finnicum, C.T., et al., *Relative telomere repeat mass in buccal and leukocyte-derived DNA.* PLoS One, 2017. **12**(1): p. e0170765.
45. Martin, N.G., et al., *The power of the classical twin study.* Heredity, 1978. **40**(1): p. 97-116.
46. Rizvi, S., S.T. Raza, and F. Mahdi, *Telomere length variations in aging and age-related diseases.* Current aging science, 2014. **7**(3): p. 161-167.
47. Dalgård, C., et al., *Leukocyte telomere length dynamics in women and men: menopause vs age effects.* International journal of epidemiology, 2015. **44**(5): p. 1688-1695.
48. Martin-Ruiz, C.M., et al., *Reproducibility of telomere length assessment: an international collaborative study.* International journal of epidemiology, 2015. **44**(5): p. 1673-1683.
49. Aviv, A., A.M. Valdes, and T.D. Spector, *Human telomere biology: pitfalls of moving from the laboratory to epidemiology.* International journal of epidemiology, 2006. **35**(6): p. 1424-1429.

3

METATAXONOMIC ANALYSIS OF INDIVIDUALS AT BMI EXTREMES AND MONOZYGOTIC TWINS DISCORDANT FOR BMI

This chapter was published as: Finnicum CT, Doornweerd S, Dolan CV, Luningham JM, Beck JJ, Willemsen G, Ehli EA, Boomsma DI, Ijzerman RG, Davies GE, de Geus EJC. (2018) Metataxonomic Analysis of Individuals at BMI Extremes and Monozygotic Twins Discordant for BMI. Twin Research and Human Genetics

ABSTRACT

The human gut microbiota has been demonstrated to be associated with a number of host phenotypes including obesity and obesity-associated phenotypes. This study was aimed at further understanding and describing the relationship between the gut microbiota and obesity associated measurements obtained from human participants.

Here we utilize genetically informative study designs including a four-corners design (extremes of genetic risk for BMI and of observed BMI) (N = 50) and the BMI monozygotic discordant twin pair design (N = 30) in order to help delineate the role of host genetics and the gut microbiota in the development of obesity.

Our results highlight a negative association between BMI and alpha diversity of the gut microbiota. The low genetic risk / high BMI group of individuals had a lower gut microbiota alpha diversity when compared to the other three groups. Although the difference in alpha diversity between the lean and heavy groups of the BMI discordant MZ twin design did not achieve significance, this difference was observed to be in the expected direction with the heavier participants having a lower average alpha diversity. We have also identified 9 operational taxonomic units (OTUs) observed to be associated with either a leaner or heavier phenotype, with enrichment for OTUs classified to the *Ruminococcaceae* and *Oxalobacteraceae* taxonomic families.

Our study presents evidence of a relationship between BMI and alpha diversity of the gut microbiota. In addition to these findings, a number of OTUs were found to be significantly associated with host BMI. These findings may highlight separate sub-types of obesity, one driven by genetic factors, the other more heavily influenced by environmental factors.

INTRODUCTION

Microbial organisms are now understood to be important residents within the human host. This finding is strengthened through a multitude of studies implicating commensal microbes in many host biological processes such as nutrient metabolism [1, 2], developmental processes [3, 4], and predispositions to certain disease states [5]. One of the areas garnering particular interest is the association of the human microbiota, mainly that of the gastrointestinal tract, in the development of obesity, and obesity-associated phenotypes [6-8]. Recent work has demonstrated the ability of the gut microbiota of obese animals to induce obesity in non-obese animals [5]. It is worth noting that along with the induction of obesity, comorbidities such as changes in neuroinflammation and subsequent cognitive disruptions have also been induced through the transfer of the gut microbiota of obese mice to non-obese mice [9].

These findings suggest a causal effect of the gut microbiota on the development of obesity, but they do not rule out simultaneous but reverse causal effects of obesity on the gut microbiota. Obesity is associated with changes in the inflammatory profile in humans that may affect gut microbiota, as well as with eating behaviors and physical activity patterns that may also impact the microbiota [10, 11]. Therefore, associations between obesity and the composition of gut microbiota may also reflect reverse causal effects. A specific composition of the gut microbiota may increase the risk for obesity, whereas obesity, either directly or through the lifestyle behaviors of which obesity is a marker or a codeterminant, may also actively change the composition of the gut microbiota [12].

In order to understand the complex nature of the interactions occurring between the gut microbiota and the human host, it is necessary to have proper models to do so. Studies performed in animals have provided one necessary approach to study microbiota dynamics in a genetically controlled environment. We do not yet know the extent to which results derived from mice can be extrapolated to humans. There are large differences in the anatomy of the murine and human gastrointestinal (GI) tract, and up to an 85% difference is found in the bacterial genera observed within the mouse GI tract relative to that of a human [13, 14]. Experimental manipulation of the human gut microbiota is feasible [15], but difficult to do on the scale possible in animal models. A potential approach to examine causal effects of the gut microbiota in observational studies in humans is to exploit the fact that individual differences in the gut microbiota composition are partly caused by heritable variation [16, 17]. If we assume that the heritability of obesity reflects, in part, the heritable effects on the gut microbiota, genetically informative designs can be used to test the predictions from causal hypotheses in both directions [18, 19]. Here, we make use of two genetically informative designs: (1) unrelated individuals selected to be in four corners defined by low or high genetic risk for BMI and by observed high or low BMI, and (2) genetically

identical monozygotic twins discordant for current BMI. Genetic risk was defined on the basis of a multi-SNP genetic risk profile from the recent meta-analysis of the GIANT consortium [20].

The aim of this study was to elucidate the gut microbiota constituents and subsequent community structure that differentiates heavier from leaner human individuals. This is achieved through 16s rRNA analysis to identify microbial community members within the gut microbiota. We hypothesized that high genetic risk for increased BMI will be associated with quantitative (smaller species diversity) and qualitative effects (enrichment for different species) on the gut microbiota. Using the four-corners design, we tested whether this association is compatible with a causal effect of the gut microbiota on BMI [18]. In testing the effect of BMI (high/low) and genetic risk (high/low) on the composition of the gut microbiota, we anticipated two outcomes. If the causal chain is high genetic risk → high BMI → gut microbiota composition, we expect a main effect of BMI (high/low) only (Figure 3.1). This expectation is based on the assumption that the relationship between genetic risk and composition is mediated by BMI. In contrast, if gut microbiota composition is a cause of high BMI, we expect a main effect of BMI and genetic risk on gut microbiota composition (Figure 3.1). This expectation does not depend on the absence (or presence) of a direct relationship between genetic risk and composition. Furthermore, the availability of MZ twin data allowed us to use the co-twin control method to discriminate between a direct causal effect of BMI on gut microbiota composition and an association brought about by a 'third factor' such as shared environment or shared genes that influence both BMI and microbiota composition [21]. If BMI is the causal agent, a comparison of genetically identical twins selected to be discordant for BMI should show a distinct composition of the gut microbiota in the lower and higher BMI individuals.

MATERIALS AND METHODS

PARTICIPANTS

The first group of individuals (N = 50) was selected from a large population (N = 11,495) within the Netherlands Twin Register for which BMI and polygenic risk score scores for BMI were available (Table 3.1) [22]. This allows for the use of a four-corner design where the study participants are selected from the top and bottom 25% of the BMI distribution, and the top and bottom 20% of the distribution of BMI polygenic risk scores produced using genome-wide SNPs. The second group of individuals (n = 30) were MZ twins (15 pairs) discordant for BMI (mean BMI difference 4.2 ± 1.9 kg/m^2 (range 1.0-8.2) that have been previously described in detail elsewhere (Table 3.1) [23]. All study participants were female in order to decrease the possibility of sexual dimorphic confounding factors. Participants were excluded if they were not within 18-75 years of age, had experienced recent weight change, or had been currently diagnosed with

heart disease, liver or renal disease, diabetes mellitus, malignancies, uncontrolled thyroid disease, or psychiatric or neurological disorders. In addition, participants were also excluded if they were pregnant, breast feeding, currently taking psychoactive or glucose-lowering drugs, or had reported drug/alcohol abuse. Due to the fact that participants were initially selected for an MRI-associated study, participants were also excluded based on MRI contraindication. Body fat measurements for all individuals were obtained through the use of bio-electrical impedance. The study was approved by the ethics committee of the VU Medical Centre and was performed in accordance with the Helsinki Declaration. All subjects involved provided written informed consent.

SAMPLING METHODS

Fecal samples were collected from individuals and stored at 4°C until delivered to the laboratory within 36 hours. Anaerocult was used in order to preserve anaerobic species present within a sample. The samples were homogenized, aliquoted, and stored at -80°C until utilized for DNA extraction.

BMI POLYGENIC RISK SCORE

Polygenic risk scores were calculated based on 77 of the 97 SNPs previously identified as having a role in obesity. These 77 SNPs were the ones that reached genome-wide significance level (5×10^{-8}) within individuals of European ancestry. The scores were determined by summing the risk alleles weighted by their respective effect sizes.

Table 3.1 | Descriptive Statistics for the Study Participants

	MZ twins			**Four-corners**		
	Leaner twin	Heavier twin	High BMI / Low GR	Low BMI / Low GR	High BMI / High GR	Low BMI / High GR
N	15	15	9	14	14	13
Age	29 (9.9)	29 (9.9)	39.84 (5.8)	36.70 (6.8)	39.40 (5.6)	35.07 (8.0)
BMI	24.25 (3.2)	28.22 (3.6)	31.20 (2.1)	20.87 (0.97)	34.08 (5.1)	20.63 (1.9)
Body Fat (Kg)	22.03 (6.8)	30.51 (8.3)	35.79 (11.6)	17.20 (3.3)	39.27 (11.1)	17.25 (4.1)
Waist-hip ratio	0.80 (0.06)	0.84 (0.08)	0.88 (0.06)	0.78 (0.04)	0.89 (0.05)	0.78 (0.03)
Inverse Simpson	25.52 (11.3)	22.66 (11.4)	15.44 (5.3)	29.07 (8.4)	24.14 (5.6)	27.24 (6.9)

SEQUENCING METHODS

DNA was extracted using the MO Bio PowerSoil Kit with the addition of the heating steps from the Power Fecal Kit (Mo Bio, Carlsbad, CA). Sequencing library preparation and indexing was adapted from Kozich et al. to generate libraries for sequencing-by-synthesis on the Illumina MiSeq platform [24]. The V4 region of the 16S rRNA gene was chosen for amplification and sequencing [24]. Sequence data was generated on the MiSeq platform, using a 2 x 251 paired-end sequencing run with 20% Phix to increase base diversity during the run. Use of a mock community aided as a positive control, and a non-template negative control was also sequenced.

MICROBIOTA SEQUENCING QUALITY CONTROL AND DATA ANALYSIS

MiSeq reads were filtered based upon the work published by Kozich et al. describing a method for analysis of dual-indexed amplicon sequences resulting from the Illumina MiSeq platform [24]. The MiSeq sequencing run resulted in demultiplexed paired-end FASTQ files for each sample, which were then analyzed using the Mothur software package version 1.36.1. The forward and reverse reads were overlapped, producing contigs for each sample (Figure 3.2). The joining of reads resulted in 6,188,475 reads. Sequences were filtered to remove sequences with ambiguous bases, as well as sequences shorter than 275 bp. The SILVA v123 database was trimmed to cover the V4 region of the 16s rRNA gene, and unique sequences were subsequently aligned to the customized SILVA v123 database [25]. After alignment and filtering, the reads were preclustered to join sequences that are within two nucleotides of one another. UCHIME was used to identify and remove possible chimeric reads from the data [26]. After chimera removal, the sequences were classified using the naïve Bayesian classifier trained on the Ribosomal database project (RDP) training set [27]. Non-bacterial lineages were removed; these included eukaryotes, archaea, chloroplasts, mitochondria, as well as unknown lineages. Within the samples sequenced was a mock community of 20 known bacterial sequences. This mock community was used to calculate the error rate of the sequencing run after read filtering. The reads from this mock community were compared to the known sequences, and the error rate was determined to be 0.0053%. The mock community was removed from further processing. After the quality control process, 4,838,970 sequences remained, 45,057 of which were unique sequences. Unique sequences were then clustered into operational taxonomic units with a 0.03 cut-off using the average-neighbor clustering algorithm. Consensus taxonomies of the OTUs were determined using classify.OTU command within Mothur. The OTU clustering ultimately resulted in 4,236 unique OTUs. Of these 4,236 OTUs, 67.68% were unclassified at the genus level and 36.66% of all OTUs were unclassified at the family level. In order to achieve proper sampling depth for all samples, the reads for each sample were subsampled to the lowest read depth, which was 36,783 reads.

STATISTICAL ANALYSES

In order to investigate the alpha diversity associated with the sampled communities, inverse Simpson values were generated. Inverse Simpson values are a function of both the species richness (number of species present) and the relative abundances of species level organisms. Inverse Simpson values were generated using the Mothur software package [28]. In order to compare the inverse Simpson values for the four-corners individuals, a two-way ANOVA was employed. A paired samples t-test was utilized to determine whether there was a difference in mean inverse Simpson values between the BMI discordant MZ twins. Beta diversity calculations have been performed for the individuals within the four-corners design. We generated Bray-Curtis dissimilarity measures between all 50 individuals and then tested whether

the mean BC measures were significantly different between any of the groups. First, we tested whether or not there was a significant difference in BC distances between the high/low BMI groups and the high/low genetic risk groups by utilizing a t-test with 10,000 permutations. To expand upon this, we also split the 50 four-corners individuals into their respective groups (low BMI/low PRS, low BMI/high PRS, high BMI/low PRS, high BMI/high PRS) and tested for any differences using a one-way ANOVA with 10,000 permutations. In the same manner we tested whether there was a significant difference in BC distances between the group of leaner co-twins relative to the heavier co-twins. BMI was regressed on the inverse Simpson diversity values by utilizing the GEE package within R, accounting for the relatedness of the MZ twin pairs. To more fully capture obesity, this regression was repeated for two additional traits, waist-hip ratio and body fat percentage.

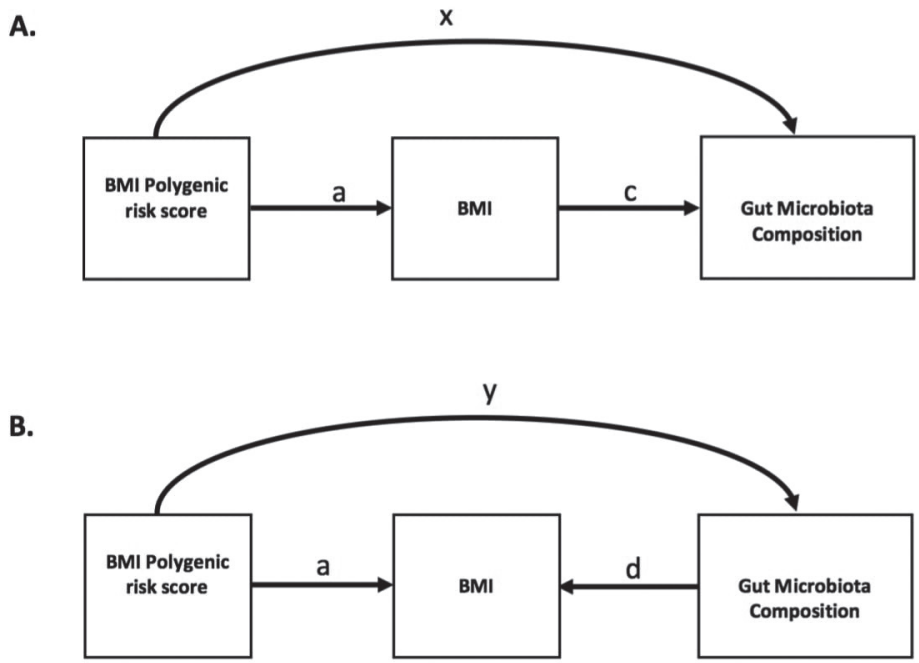


Figure 3.1 | (A). Given the causal model depicted in the top diagram, x is assumed to equal 0 and thus the two-way ANOVA employed is expected to yield a main effect of BMI, but no effect of BMI polygenic risk. This model would reflect a paradigm in which genetic risk for BMI influences BMI which subsequently influences the gut microbiota composition. (B). Under this causal model, the two-way ANOVA is expected to yield a main effect of both BMI and BMI polygenic risk score. This expectation does not depend on the absence or presence of a direct relationship between BMI genetic risk and gut microbiota composition (i.e., y may be 0 or greater than 0.)

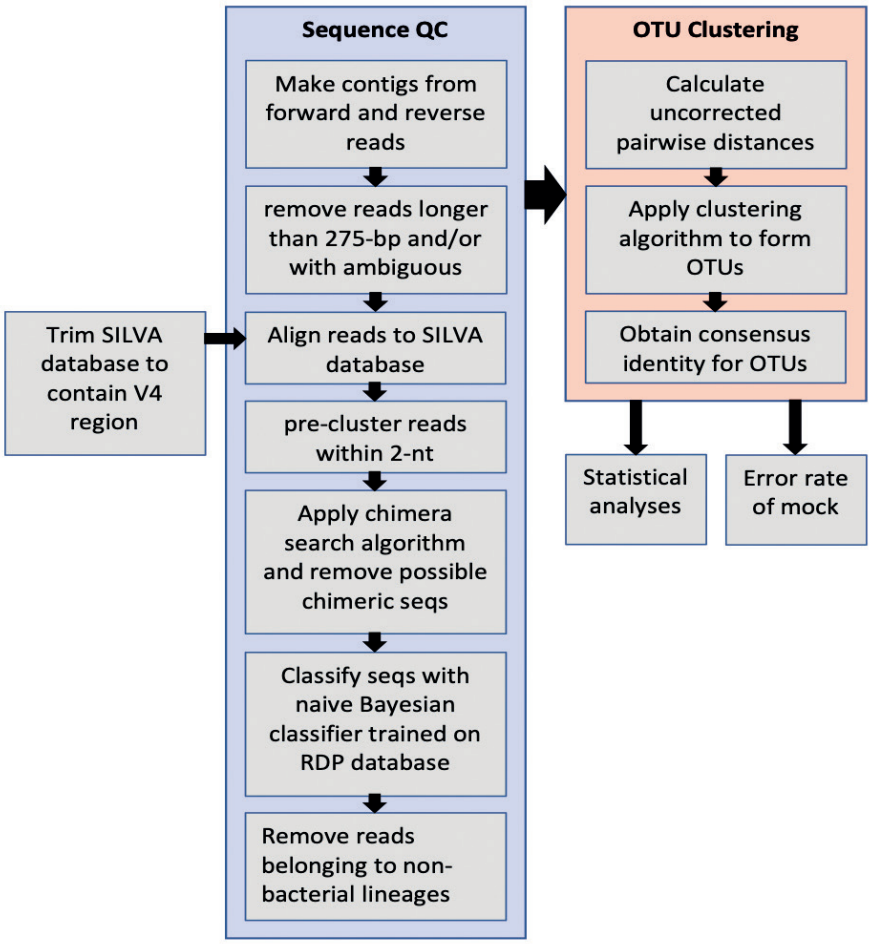


Figure 3.2 | Bioinformatics pipeline utilized to analyze 16S rRNA data.

To detect OTUs differentially enriched within leaner and heavier individuals two classification strategies were used: LEfSe analysis, and the random forest approach. The LEfSe analysis is aimed at determining a significantly different presence of OTUs in various subgroups (e.g., low vs. high BMI) with an alpha of 0.05 for both the Kruskal-Wallis and Wilcoxon tests within LEfSe [29]. The traditional LEfSe analysis was modified to include 100,000 permutations, from which an empirical p value was derived. The LDA threshold was set at 2.0.

Random forest classification was performed on the four-corners individuals within the Mothur software. The OTUs sampled per split were calculated based on the log2 of the total number of features. Each classification utilized 20,000 trees. Other parameters

included using tree pruning with a pruning aggressiveness of 0.9. Trees with an error rate above 0.4 were discarded. Any feature with a standard deviation less than 0.1 was also discarded. Of note, the random forest classifier allows for the identification of OTUs that do not necessarily have a linear relationship with the phenotype of interest.

Regression analysis was performed by utilizing the generalized estimating equations to account for the MZ twin pairs present. BMI, WHR, and body fat mass (kg) were regressed on the separate OTU abundances, while accounting for family structure. Because of the presence of outliers in the OTU data, points were removed that fell outside of four standard deviations from the mean for that specific OTU. The regression on the multiple OTU abundances was corrected for multiple testing via a false discovery rate correction.

For the purposes of the LEfSe, random forest and regression analyses, OTUs were discarded if they were not present within at least 40% of all individuals (32 people). This resulted in a total of 279 OTUs remaining, including an OTU that combines all excluded OTUs.

RESULTS

ALPHA DIVERSITY COMPARISONS

The two-way ANOVA, including main effects of PRS and BMI and their interaction on the mean inverse Simpson index values, showed a significant main effect of BMI, that is, a difference in alpha diversity between individuals with obesity and leaner participants ($p = 0.00009$), with a decreased alpha diversity within the gut microbiota of individuals with obesity. The main effect of genetic risk was not significant (Table 3.2).

There was an unanticipated significant interaction between genetic risk and BMI ($p = 0.0096$). Plotting the inverse Simpson values clearly showed the high BMI/low genetic risk individuals had a decreased alpha diversity (Figure 3.3).

Inverse Simpson values were also generated for the BMI discordant twin pairs. The gut microbiota of the heavier twin had a lower average inverse Simpson value relative to the mean values of the leaner twins. However, this difference failed to reach significance ($p = 0.298$; Table 3.3).

REGRESSING BMI, BODY FAT, AND WAIST-TO-HIP RATIO ON INVERSE SIMPSON VALUES

In order to investigate the relationship between alpha diversity and a number of different obesity-associated measures including BMI, kilograms of body fat, and WHR, we regressed these outcomes on the inverse Simpson values in the subjects from both the four-corners and BMI discordant twins ($n = 80$). The data from the two

study designs were combined in order to increase the sample sizes for the regression analyses. Significant negative relationships were observed between BMI and alpha diversity as well as body fat and alpha diversity (Table 3.4). WHR measures were not significantly associated with alpha diversity.

REGRESSING OTU ABUNDANCES ON POLYGENIC RISK SCORES

Using the four-corners design, each of the 279 OTUs were regressed on the polygenic risk scores to identify OTUs associated with individuals with varying degrees of genetic risk for obesity. After multiple testing correction, there were no significant associations between the polygenic risk scores and any of the OTUs.

BETA DIVERSITY MEASUREMENTS

Comparison of the Bray-Curtis (BC) distances between the four-corners participants showed that there was no significant difference between any of the four groups (Figure 3.4). Similarly, there was no statistical difference between the BC distances of high BMI individuals versus low BMI individuals or for the high genetic risk group versus low genetic risk group (Figure 3.5, Figure 3.6). Testing for a difference between the leaner co-twin groups relative to the group of heavier co-twins also failed to reach statistical significance (Figure 3.7).

Table 3.2 | Effects of Genetic Risk, BMI, and Their Interaction on the Mean Inverse Simpson Values

	F	p-value
Genetic Risk	3.108	0.085
BMI	18.439	0.00009
Genetic Risk * BMI	7.293	0.009658

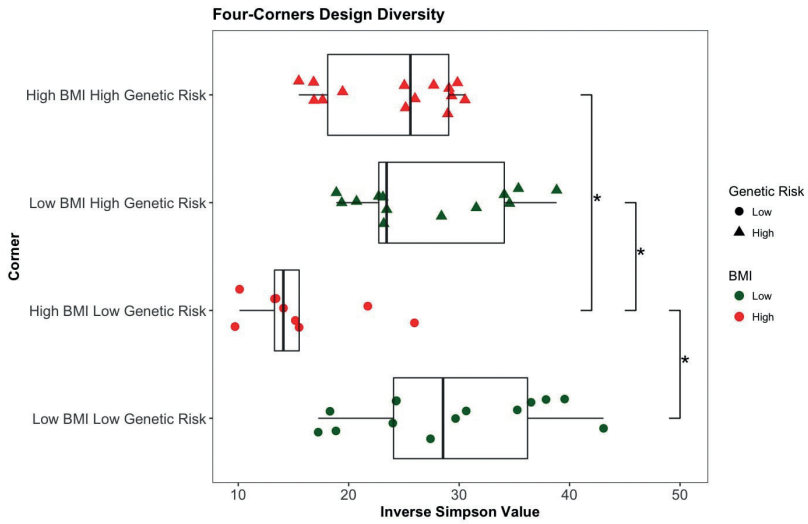


Figure 3.3 | Box plot of the mean inverse Simpson values from the four-corners design.

Table 3.3 | Results of the t-test between the inverse-Simpson values of the leaner and heavier co-twins

	MeanInverseSimpson	N	sd	t	p-value
LeanerTwin	25.517	15	11.258		
HeavierTwin	22.66	15	11.396	1.08	0.298

Table 3.4 | Regression of BMI, Body Fat, and Waist-Hip Ratio on the Mean Inverse Simpson Values

Measure	Beta	R2	p-value
BMI	-0.162	0.002	0.004
Body Fat (Kg)	-0.28	0.000402	0.015
Waist-hip ratio	-	-	0.156

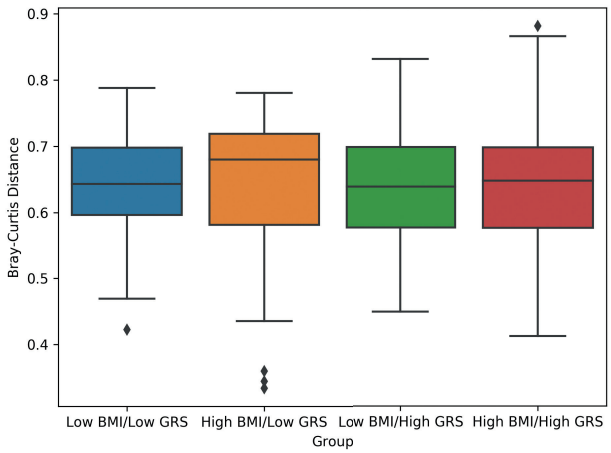


Figure 3.4 | Beta diversity comparison between the four-corners design.

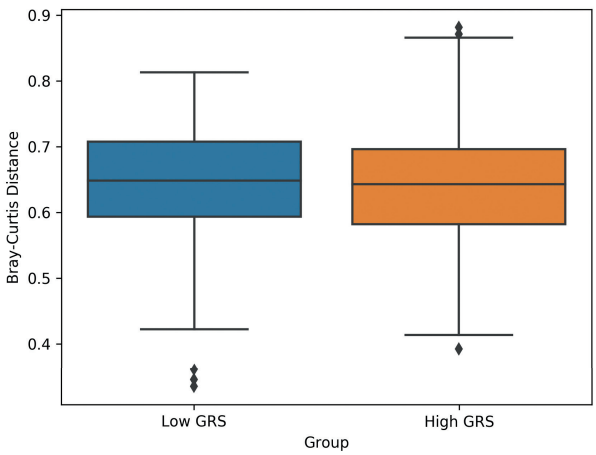


Figure 3.5 | Beta diversity test between the individuals in the four-corners design with a high/low genetic risk for obesity.

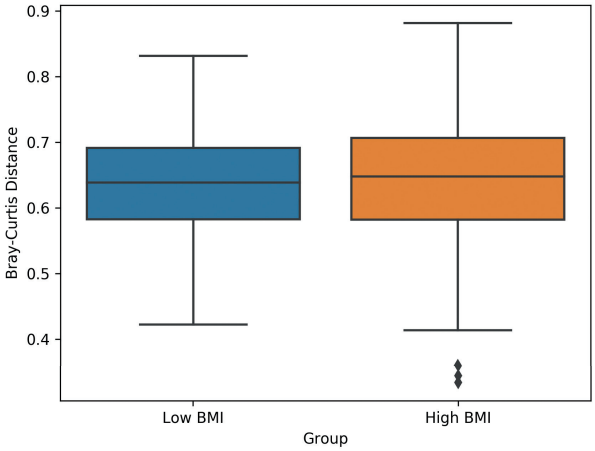


Figure 3.6 | Beta diversity test between the individuals in the four-corners design with a high/low observed BMI.

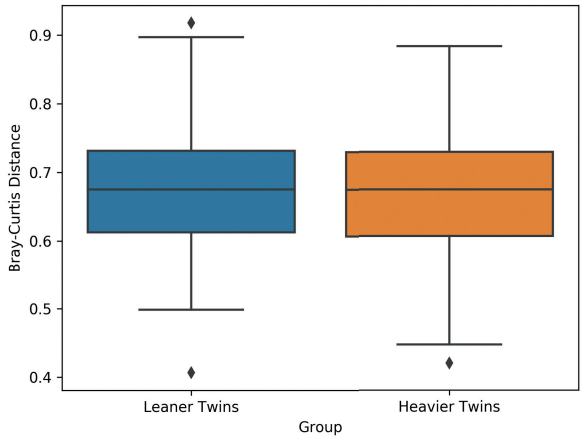


Figure 3.7 | Beta diversity test between the BMI discordant twin pairs.

REGRESSING BMI, BODY FAT, AND WAIST-TO-HIP RATIO ON OTU ABUNDANCES

A number of OTUs were identified as being significantly associated with the three measured obesity associated traits (Table 3.5). There were 9 OTUs that were associated with BMI in the host individual. Regressing body fat and WHRs on the individual OTU abundances identified four and seven significant OTUs respectively. There were varying degrees of overlap between the significant OTUs identified in the three separate analysis.

LEFSE IDENTIFICATION OF DIFFERENTIAL OTUS

Through the use of linear discriminant analysis (LDA), effect size (LEfSe) analysis, OTUs differentially enriched within leaner versus heavier individuals were analyzed.

First, the individuals within the four-corners design were analyzed for differentially enriched features. This yielded no OTUs significantly enriched in any of the four corners. Second, individuals within the four-corners groups were compared using high/low BMI as the class and high/low genetic risk as the subclass, resulting in seven OTUs enriched in the low BMI participants and three OTUs enriched in the high BMI participants Figure 3.8. We followed this up by LEfSe analysis performed between the BMI discordant MZ twins to see which OTUs were enriched within a model that controls for the host genetic profile. This analysis showed two OTUs enriched in the heavier co-twins and 17 OTUs enriched in the leaner co-twins, five of which were also found enriched in the leaner groups of the four-corner design (Figure 3.8).

RANDOM FOREST CLASSIFICATION

To examine the BMI association in more detail, random forest classifications were performed, again separately in the leaner and heavier individuals within both the four-corners individuals and the BMI discordant MZ twins based on the observed OTUs. The classification process was able to accurately classify 96% of the four-corners individuals into low and high BMI as well as 93.3% of the discordant MZ twins into the correct lean and obese category. Assuming that the overall classification is decent, the random forest classifier provides information on which OTUs yielded the most predictive information. Applying the random forest classification aimed at classifying high and low genetic risk was only able to accurately classify 66% of the individuals. The 50 first OTUs used by the random forest classifier can be observed in Table 3.6.

CONVERGENCE ACROSS DIFFERENT ANALYTIC STRATEGIES

Table 3.7 summarizes how the various OTUs were similar across the various analytic approaches used, substantiating their relevance for obesity. OTUs were included only if they were either observed in a significant manner in multiple LEfSe analyses, identified in a LEfSe analysis as well as through regressing the BMI associated measures on OTU abundances, or if they were significant in multiple regression analyses.

Table 3.5 | Regressing BMI, Body Fat, and Waist-to-Hip Ratio on OTU Abundances

	BMI			**Body Fat**			**Waist-hip ratio**		
OTUs	p-value	beta	R2	p-value	beta	R2	p-value	beta	R2
Otu00074	-	-	-	-	-	-	0.0098	-8.0715	0.9582
Otu00091	-	-	-	-	-	-	2.98E-05	-8.9895	0.9748
Otu00134	-	-	-	-	-	-	0.0010	-26.0822	0.9770
Otu00195	-	-	-	0.003	-2773.9526	0.0005	-	-	-
Otu00204	0.0073	-3430.4213	0.0046	-	-	-	0.0044	-38.1458	0.9650
Otu00220	0.0002	-5006.5632	0.0081	0.0015	-8519.5674	0.0015	-	-	-
Otu00243	0.0182	-3271.0966	0.0032	-	-	-	-	-	-
Otu00251	-	-	-	-	-	-	0.0004	-48.0467	0.9701
Otu00325	0.043	-3778.2293	0.0016	-	-	-	-	-	-
Otu00344	0.0027	-4511.9497	0.0018	0.0178	-7226.7798	0.0003	0.0442	-72.9685	0.9649
Otu00405	0.0397	-8080.2343	0.0019	-	-	-	-	-	-
Otu00443	0.0108	-7980.6241	0.002	-	-	-	-	-	-
Otu00462	0.0095	-16968.4413	0.0042	-	-	-	-	-	-
Otu00500	0.0002	-22438.2911	0.0072	9.60E-05	-39185.8192	0.0014	0.0113	-241.1956	0.9741

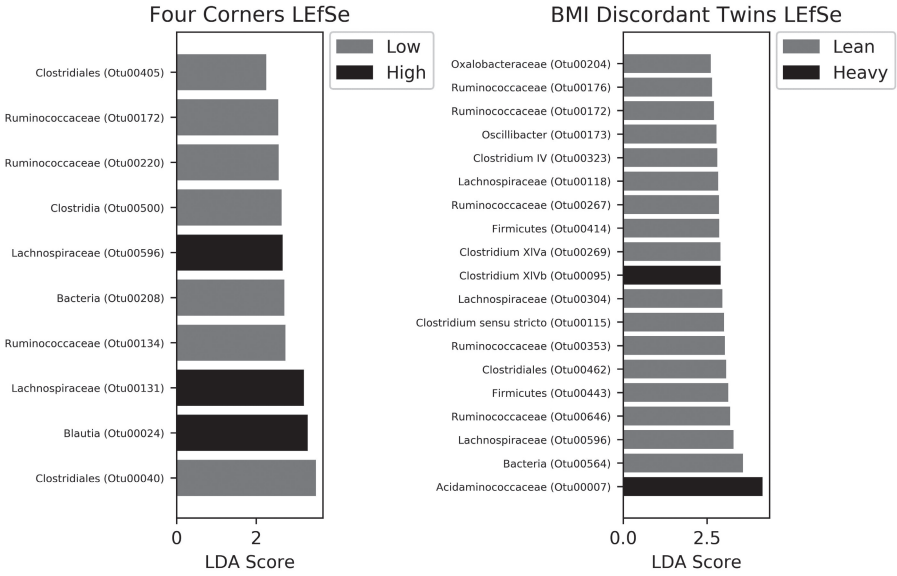


Figure 3.8 | Left: LEfSe analysis results indicating the OTUs differentially enriched in the high and low BMI groups of the four-corners design. Right: LEfSe analysis results indicating the OTUs differentially enriched in the heavy and lean co-twins of the BMI discordant MZ twin design.

Table 3.6 | Random Forest OTU rank

Rank	4-Corners RF		4-Corners GRS RF		Twins RF	
	OTU	Mean decrease accuracy	OTU	Mean decrease accuracy	OTU	Mean decrease accuracy
1	OTU00220	0.12755	OTU00071	0.0539	OTU00304	0.06025
2	OTU00060	0.0855	OTU00105	0.04025	OTU00323	0.05585
3	OTU00134	0.06065	OTU00529	0.032	OTU00564	0.0352
4	OTU00131	0.06015	OTU00195	0.0316	OTU00172	0.0323
5	OTU00204	0.0567	OTU00033	0.02185	OTU00269	0.0261
6	OTU00500	0.05565	OTU00160	0.01505	OTU00173	0.0215
7	OTU00447	0.0447	OTU00153	0.01305	OTU00176	0.02045
8	OTU00024	0.04075	OTU00140	0.01275	OTU00095	0.0192
9	OTU00118	0.0347	OTU00028	0.01215	OTU00267	0.0188
10	OTU00004	0.0341	OTU00688	0.01125	OTU00447	0.0188
11	OTU00044	0.02905	OTU00065	0.0105	OTU00286	0.01765
12	OTU00407	0.0284	OTU00711	0.0093	OTU00076	0.0173
13	OTU00172	0.0241	OTU00225	0.00905	OTU00007	0.0125
14	OTU00005	0.02375	OTU00093	0.00905	OTU00118	0.01195
15	OTU00136	0.0237	OTU00115	0.0087	OTU00042	0.01175
16	OTU00094	0.02145	OTU00019	0.00845	OTU00321	0.0111
17	OTU00462	0.0198	OTU00106	0.0082	OTU00074	0.01045
18	OTU00097	0.0181	OTU00234	0.00775	OTU00045	0.00855
19	OTU00062	0.01755	OTU00312	0.00765	OTU00199	0.00855
20	OTU00156	0.0174	OTU00390	0.0076	OTU00041	0.0084
21	OTU00063	0.01725	OTU00361	0.00755	OTU00022	0.00835
22	OTU00006	0.0169	OTU00001	0.00735	OTU00115	0.0082
23	OTU00042	0.0165	OTU00087	0.0073	OTU00414	0.00775
24	OTU00132	0.01595	OTU00108	0.0069	OTU00174	0.00705
25	OTU00077	0.01585	OTU00130	0.0069	OTU00109	0.0068
26	OTU00555	0.01485	OTU00004	0.0068	OTU00135	0.0068
27	OTU00016	0.0145	OTU00044	0.0066	OTU00146	0.00625
28	OTU00295	0.0139	OTU00145	0.0066	OTU00344	0.0061
29	OTU00353	0.0135	OTU00002	0.0064	OTU00136	0.00535
30	OTU00262	0.0133	OTU00539	0.0062	OTU00190	0.00535
31	OTU00208	0.01265	OTU00020	0.0062	OTU00008	0.00515
32	OTU00621	0.0118	OTU00095	0.0062	OTU00229	0.00505
33	OTU00068	0.0111	OTU00152	0.00615	OTU00353	0.005

Table 3.6 | Continued.

Rank	4-Corners RF		4-Corners GRS RF		Twins RF	
	OTU	Mean decrease accuracy	OTU	Mean decrease accuracy	OTU	Mean decrease accuracy
34	OTU00017	0.01105	OTU00413	0.00615	OTU00048	0.00485
35	OTU00033	0.01105	OTU00147	0.00585	OTU00062	0.0047
36	OTU00323	0.00995	OTU00010	0.0058	OTU00017	0.0047
37	rareOTUs	0.00975	OTU00084	0.00565	OTU00301	0.00465
38	OTU00458	0.0096	OTU00110	0.0056	OTU00134	0.0045
39	OTU00251	0.009	OTU00564	0.00545	OTU00087	0.0044
40	OTU00081	0.0086	OTU00260	0.00515	OTU00500	0.00415
41	OTU00002	0.00855	OTU00146	0.005	OTU00481	0.00415
42	OTU00014	0.0085	OTU00037	0.00495	OTU00462	0.0041
43	OTU00302	0.00835	OTU00040	0.00485	OTU00153	0.0041
44	OTU00269	0.0082	OTU00009	0.0048	OTU00204	0.004
45	OTU00064	0.00815	OTU00255	0.0048	OTU00064	0.004
46	OTU00051	0.0081	OTU00034	0.0048	OTU00084	0.00395
47	OTU00777	0.0073	OTU00174	0.00475	OTU00016	0.0038
48	OTU00053	0.0071	OTU00679	0.00475	OTU00341	0.0038
49	OTU00560	0.007	OTU00400	0.0047	OTU00555	0.00375
50	OTU00286	0.00695	OTU00102	0.00455	OTU00165	0.0037

Table 3.7 | Convergence across different analytic strategies.

		LEfSe		**OTU regressions**			**Random forest rank**	
OTUs	**Taxonomic classification**	Four-corners	MZ twins	BMI	Body Fat	Waist-hip ratio	Four-corners	MZ twins
Otu00134	Clostridiales, Ruminococcaceae	Low	-	-	-	X	3	38
Otu00172	Clostridiales, Ruminococcaceae	Low	Lean	-	-	-	13	4
Otu00204	Burkholderiales, Oxalobacteraceae	Low	Lean	X	-	X	5	44
Otu00220	Clostridiales, Ruminococcaceae	Low	-	X	X	-	1	-
Otu00344	Bacteria, Unclassified	-	-	X	X	X	-	28
Otu00443	Bacteria, Firmicutes	-	Lean	X	-	-	-	-
Otu00462	Clostridia, Clostridiales	-	Lean	X	X	-	17	42
Otu00500	Firmicutes, Clostridia	-	-	X	X	X	6	40
Otu00596	Clostridiales, Lachnospiraceae	Low	Lean	-	-	-	54	60

DISCUSSION

Through 16s rRNA analysis, we examined the gut microbiota constituents and subsequent community structure that differentiates heavier from leaner human individuals using two genetically informative designs: (1) unrelated individuals selected to be in one of four corners defined by low or high genetic risk for BMI based on a multi-SNP genetic risk profile and by observed high or low BMI, and (2) genetically identical MZ twins discordant for current BMI. Alpha diversity was significantly different between the leaner and heavier individuals within the four-corner design, that is, there was a main effect of BMI (high/low). However, there was no main effect of PRS (high/low). As such, the results are consistent with a causal effect of BMI on alpha diversity. However, the presence of an unanticipated significant interaction complicates this interpretation of the results. It is important to note that the average difference in BMI in the BMI discordant MZ twin pairs (mean BMI difference 4.2 ± 1.9 kg/m^2 (range 1.0 - 8.2)), was much lower than the average difference in BMI between the leaner and heavier individuals in four corners design (BMI ≤ 22 kg/m^2 and ≥ 27 kg/m^2) and thus this could possibly explain the smaller difference in alpha diversity observed between the co-twins.

The four-corner design allowed for the exploration of differences in gut microbiota alpha diversity. If low alpha diversity was a consequence of high BMI, it would be expected that both high BMI groups would have a decreased alpha diversity (i.e., the anticipated main effect of BMI). While this main effect was observed, its interpretation is complicated by the significant interaction between PRS and BMI. Specifically, the individuals with a low genetic risk for BMI (low PRS) and high BMI showed a lower alpha diversity when compared to each of the other three groups. These findings may highlight separate subtypes of obesity, one driven by genetic factors, the other more heavily influenced by environmental factors. The latter subtype of obesity may be influenced by a separate external cause that either increases BMI through an effect of decreased gut microbiota diversity, or independently causes increases in BMI and decreases in gut microbiota diversity. The latter leaves open that it may not necessarily be the gut microbiota causing the obese state itself. The observed decrease in alpha diversity of the gut microbiota of low genetic risk/high BMI individuals may be a consequence of the actual cause of obesity such as dietary intake or exercise activity.

Although previous studies have identified decreased alpha diversity associated with increased BMI as well as specific dietary patterns such as consumption of a Western diet, other studies have only identified a relatively weak association between gut microbiota alpha diversity and obese status [30-32]. For the obese status, it may be possible that the weak association with gut microbiota alpha diversity is due to the presence of the aforementioned sub-phenotypes of obesity present within the participants examined. In this case, the lack of decreased alpha diversity within the

individuals at a high genetic susceptibility would hinder the ability to detect the effect in the larger population.

In order to further understand the association between gut microbiota alpha diversity and obesity, BMI, body fat mass (kg), and WHR measurements were all regressed on alpha diversity using all 80 individuals. BMI and body fat mass showed significant negative associations with alpha diversity whereas WHR did not show a significant association. These findings are supported by a recent study that observed a negative association between gut microbiota alpha diversity and a number of adiposity associated measures [31]. Interestingly, this study also did not observe a significant association between alpha diversity and WHR. Together, these findings point towards a paradigm where gut microbiota composition is associated with general adiposity rather than fat distribution (i.e., gynoid vs. android obesity) as reflected in WHR. Gut microbiota involvement in development of adipose tissue has been previously explored in animal studies, where it was observed that the transfer of gut microbiota contents from a conventionally raised mouse to a mouse raised in a germ-free (GF) environment resulted in a 60% increase in body fat while consuming significantly less food [7].

In addition to comparisons between obesity-associated measures and alpha diversity, we explored whether there were OTUs significantly associated with either a leaner or heavier phenotype within the two separate study designs. Regressing three separate obesity associated measures (BMI, body fat, WHR) on the OTU relative abundances resulted in 14 OTUs significantly associated with one of these measures. There was a varying degree of overlap between the OTUs identified through the regression analyses, with only two OTUs significantly associated with all three obesity associated measurements. It may be possible that the subtle differences in the obesity-associated measures may be the cause of the slightly different results. As was previously noted, BMI and body fat may actually represent slightly different obesity-associated phenotypes regarding fat distribution in comparison to WHR (gynoid vs. android obesity), which could explain the lack of overlap between these measures. When comparing the overlap of OTUs between BMI and body fat, there is actually a fair amount of overlap given that three of the four OTUs identified in the body fat regression were also identified in the BMI regression. As BMI is a function of both weight and height of an individual and not necessarily purely body fat, it may be plausible that OTUs identified via the BMI regression may also be associated with other factors such as height or muscle mass, and not purely fat content. Taking a closer look at the taxonomic classification of the OTUs showed that 10 of the 14 significant OTUs belong to the *Firmicutes* phylum, all of which are related in an inverse manner to BMI. Although it may appear tempting to utilize a phylotype-based approach and test whether any of our study groups and/or BMI have an association with the *Firmicutes* phylum, it should be noted that the LEfSe analyses identified five OTUs enriched in heavy individuals, all of which corresponded to the *Firmicutes* phylum. All of the OTUs

identified in lean individuals through the use of LEfSe analysis, with the exception of three OTUs, two of which were only classified to the Bacteria domain, also belonged to the *Firmicutes* phylum. This clearly demonstrates that various members of the *Firmicutes* phylum may have varying contributions to the obese phenotype, with some OTUs associated with a lean phenotype and others associated with a heavier phenotype.

Utilization of these various analytical approaches ultimately converged on nine OTUs that showed an association to BMI or other obesity-associated measures in multiple analyses. These OTUs were able to be classified down to various taxonomic ranks (e.g., order and family). OTUs belonging to the *Ruminococcaceae* (OTU 220) and *Oxalobacteraceae* (OTU 204) families were found to be enriched within leaner individuals as well as generally negatively associated with obesity measures. Members of the *Ruminococcaceae* family have been observed in a similar fashion in separate studies outlining the relationship between the gut microbiota and body fat composition [31]. Although the *Oxalobacteraceae* family has less of a documented association with obesity, previous studies of human and animal gut microbiota contents have observed decreases of this family in response to administration of antibiotics as well as enriched within individuals with no previous contact with the Western world [33, 34]. Prenatal maternal and early-life antibiotic use have both been shown to impact the development of obesity [35, 36]. Understanding the relationship between *Oxalobacteraceae* microorganisms, antibiotic use, and subsequent host body composition could be of value for understanding how environmental influences may impact the susceptibility to obesity.

One of the hypotheses put forth to explain how the gut microbiota may impact the development of an obese phenotype revolves around the idea that the gut microbiota composition of heavier individuals may have an increased capacity to harvest energy from food consumed by the host [5]. This concept is supported by studies that have observed enrichment of genomic material encoding products involved in the breakdown of dietary substrates within the gut microbiota of heavier mice and humans relative to their leaner counterparts [5, 30]. Our current study does not address the gut microbiota genomic functional repertoire, thus representing a limitation of this study as well as an avenue of future research. In addition to the increased energy absorption hypothesis, recent work has demonstrated that individual microorganisms can manipulate appetite within the human host. *Salmonella typhimurium* has been shown to inhibit sickness-induced anorexia by way of the gut-brain axis [37]. Although this specific example of microbial modulation of host appetite arises within a state of host distress (*S. typhimurium* infection), it presents the existence of a molecular mechanism resulting in microbial influence of host eating behavior. OTUs significantly associated with lean and heavy phenotypes such as those observed across the analyses within this study would be logical candidates for future exploration into such mechanisms.

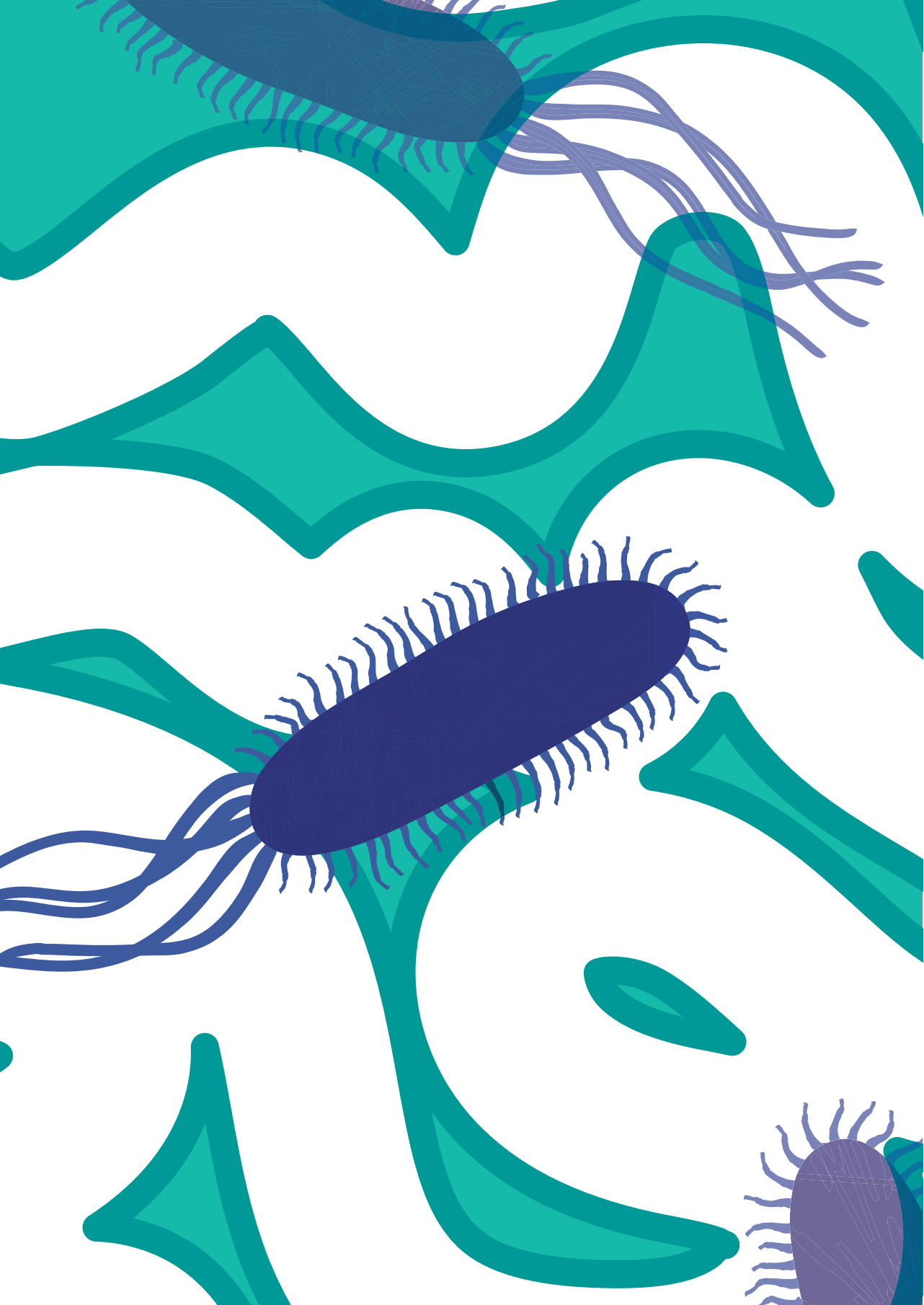
CONCLUSIONS

Our study demonstrates the utility of genetically informative study designs aimed at investigating the human gut microbiota. Through the use of such designs, we successfully highlighted a distinctly lower gut microbiota diversity in individuals with high BMI that were low in the genetic susceptibility to obesity. Additionally, we identified a number of OTUs that have a significant association with obesity-associated measures as well as being enriched in groups of lean or heavy individuals independent of genetic factors. These findings provide further support for the relationship between the human gut microbiota and the obese phenotype.

References

1. Donohoe, D.R., et al., *The microbiome and butyrate regulate energy metabolism and autophagy in the mammalian colon.* Cell metabolism, 2011. **13**(5): p. 517-526.
2. Koeth, R.A., et al., *Intestinal microbiota metabolism of L-carnitine, a nutrient in red meat, promotes atherosclerosis.* Nature medicine, 2013. **19**(5): p. 576-585.
3. Borre, Y.E., et al., *Microbiota and neurodevelopmental windows: implications for brain disorders.* Trends in molecular medicine, 2014. **20**(9): p. 509-518.
4. Heijtz, R.D., et al., *Normal gut microbiota modulates brain development and behavior.* Proceedings of the National Academy of Sciences, 2011. **108**(7): p. 3047-3052.
5. Turnbaugh, P.J., et al., *An obesity-associated gut microbiome with increased capacity for energy harvest.* nature, 2006. **444**(7122): p. 1027.
6. Ley, R.E., et al., *Human gut microbes associated with obesity.* nature, 2006. **444**(7122): p. 1022-1023.
7. Bäckhed, F., et al., *The gut microbiota as an environmental factor that regulates fat storage.* Proceedings of the national academy of sciences, 2004. **101**(44): p. 15718-15723.
8. Villanueva-Millán, M., P. Perez-Matute, and J. Oteo, *Gut microbiota: a key player in health and disease. A review focused on obesity.* Journal of physiology and biochemistry, 2015. **71**(3): p. 509-525.
9. Bruce-Keller, A.J., et al., *Obese-type gut microbiota induce neurobehavioral changes in the absence of obesity.* Biological psychiatry, 2015. **77**(7): p. 607-615.
10. Monteiro, R. and I. Azevedo, *Chronic inflammation in obesity and the metabolic syndrome.* Mediators of inflammation, 2010. **2010**.
11. Rodríguez-Hernández, H., et al., *Obesity and inflammation: epidemiology, risk factors, and markers of inflammation.* International journal of endocrinology, 2013. **2013**.
12. Richmond, R.C., et al., *Assessing causality in the association between child adiposity and physical activity levels: a Mendelian randomization analysis.* PLoS Med, 2014. **11**(3): p. e1001618.
13. Ley, R.E., et al., *Obesity alters gut microbial ecology.* Proceedings of the national academy of sciences, 2005. **102**(31): p. 11070-11075.
14. Nguyen, T.L.A., et al., *How informative is the mouse for human gut microbiota research?* Disease models & mechanisms, 2015. **8**(1): p. 1-16.
15. Smits, L.P., et al., *Therapeutic potential of fecal microbiota transplantation.* Gastroenterology, 2013. **145**(5): p. 946-953.
16. Goodrich, J.K., et al., *Human genetics shape the gut microbiome.* Cell, 2014. **159**(4): p. 789-799.
17. Lim, M.Y., et al., *The effect of heritability and host genetics on the gut microbiota and metabolic syndrome.* Gut, 2017. **66**(6): p. 1031-1038.
18. Noon, J.P., et al., *Impaired microvascular dilatation and capillary rarefaction in young adults with a predisposition to high blood pressure.* The Journal of clinical investigation, 1997. **99**(8): p. 1873-1879.
19. van Dongen, J., et al., *Longitudinal weight differences, gene expression and blood biomarkers in BMI-discordant identical twins.* International Journal of Obesity, 2015. **39**(6): p. 899-909.

20. Locke, A.E., et al., *Genetic studies of body mass index yield new insights for obesity biology.* Nature, 2015. **518**(7538): p. 197-206.
21. Stubbe, J.H., et al., *The association between exercise participation and well-being: a co-twin study.* Preventive medicine, 2007. **44**(2): p. 148-152.
22. Willemsen, G., et al., *The Netherlands Twin Register biobank: a resource for genetic epidemiological studies.* Twin Research and Human Genetics, 2010. **13**(3): p. 231-245.
23. Doornweerd, S., et al., *Physical activity and dietary intake in BMI discordant identical twins.* Obesity, 2016. **24**(6): p. 1349-1355.
24. Kozich, J.J., et al., *Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform.* Applied and environmental microbiology, 2013. **79**(17): p. 5112-5120.
25. Pruesse, E., et al., *SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB.* Nucleic acids research, 2007. **35**(21): p. 7188-7196.
26. Edgar, R.C., et al., *UCHIME improves sensitivity and speed of chimera detection.* Bioinformatics, 2011. **27**(16): p. 2194-2200.
27. Cole, J.R., et al., *The Ribosomal Database Project: improved alignments and new tools for rRNA analysis.* Nucleic acids research, 2009. **37**(suppl_1): p. D141-D145.
28. Schloss, P.D., et al., *Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities.* Applied and environmental microbiology, 2009. **75**(23): p. 7537-7541.
29. Segata, N., et al., *Metagenomic biomarker discovery and explanation.* Genome biology, 2011. **12**(6): p. 1-18.
30. Turnbaugh, P.J., et al., *A core gut microbiome in obese and lean twins.* nature, 2009. **457**(7228): p. 480-484.
31. Beaumont, M., et al., *Heritable components of the human fecal microbiome are associated with visceral fat.* Genome biology, 2016. **17**(1): p. 1-19.
32. Sze, M.A. and P.D. Schloss, *Looking for a signal in the noise: revisiting obesity and the microbiome.* MBio, 2016. **7**(4): p. e01018-16.
33. Raymond, F., et al., *The initial state of the human gut microbiome determines its reshaping by antibiotics.* The ISME journal, 2016. **10**(3): p. 707-720.
34. Torok, V.A., et al., *Influence of antimicrobial feed additives on broiler commensal posthatch gut microbiota development and performance.* Applied and environmental microbiology, 2011. **77**(10): p. 3380-3390.
35. Bailey, L.C., et al., *Association of antibiotics in infancy with early childhood obesity.* JAMA pediatrics, 2014. **168**(11): p. 1063-1069.
36. Mueller, N.T., et al., *Prenatal exposure to antibiotics, cesarean section and risk of childhood obesity.* International journal of obesity, 2015. **39**(4): p. 665-670.
37. Rao, S., et al., *Pathogen-mediated inhibition of anorexia promotes host survival and transmission.* Cell, 2017. **168**(3): p. 503-516. e12.



4

CORRESPONDENCE BETWEEN SINGLE COHORT AND FULL META-ANALYTIC RESULTS IN GENETIC ASSOCIATION ANALYSIS OF THE GUT MICROBIOME COMPOSITION

This chapter is based on my contribution to the main consortium paper, published as: Kurilshikov, A., C. Medina-Gomez, R. Bacigalupe, D. Radjabzadeh, J. Wang, A. Demirkan, C. I. Le Roy, J. A. R. Garay, C. T. Finnicum, X. Liu . . . Eco J.C. de Geus, Katie A. Meyer, Jakob Stokholm, Eran Segal, Elin Org, Cisca Wijmenga, Hyung-Lae Kim, Robert C. Kaplan, Tim D. Spector, Andre G. Uitterlinden, Fernando Rivadeneira, Andre Franke, Markus M. Lerch, Lude Franke, Serena Sanna, Mauro D'Amato, Oluf Pedersen, Andrew D. Paterson, Robert Kraaij, Jeroen Raes, Alexandra Zhernakova (2020). "Large-scale association analyses identify host factors influencing human gut microbiome composition" Nature Genetics.

ABSTRACT

Whereas environmental influences on the human microbiome have been characterized in substantial detail, less is known about the genetic contribution to variation in diversity and composition of the microbiome. To address this, the MiBioGen consortium performed a large-scale meta-analysis across multiple cohorts intending to identify genomic loci that influence the abundance of individual taxa within the gut microbiome. Here, we determine how well the results of the meta-analysis are reflected in the results of a single cohort, the Netherlands Twin Register (N = 267). We tested whether the top consortium mbQTL showed at least nominal significance in the NTR, and vice versa, whether the top NTR-specific mbQTL was among the genome-wide significant results of the consortium. To further test whether cohort-specific outcomes forecast consortium outcomes, we tested whether the pattern seen in the consortium analysis, where the more heritable taxa showed much smaller p-values in the genome-wide association tests relative to less heritable taxa, was also present in the NTR data. The consortium identified the association between the *Bifidobacterium* genus and rs182549 as the most significant mbQTL (8.6332×10^{-21}), whereas NTR-specific results for this mbQTL showed a p-value of 0.06, just failing to achieve the nominal significance for a candidate gene approach ($\alpha = 0.05$). The mbQTL that showed the most significant effect in the NTR cohort was between rs11755686 and the *Clostridium* sensu stricto 19 genus ($p = 1.05012 \times 10^{-9}$) with the MiBioGen results browser, containing all mbQTLs with a p-value less than 0.001, returning no hits associated with this NTR-specific mbQTL. With regard to the relationship between the heritability of a taxon and its p-value in the genetic association, we found that the results of the NTR already show the pattern seen in the larger MiBioGen study. Taxa that demonstrated heritable effects with MZ twin correlations greater than 0.3 showed a lower mean p-value at the family and genus levels. We confirm the notion that results deriving from a consortium analysis generalize reasonably well to a single cohort with a modest sample size, whereas the reverse is less likely to hold.

INTRODUCTION

Because research has continually demonstrated an association between differences in the gut microbiome and human health associated traits [1], it has become even more important to understand the factors that cause individuals to differ in the composition of the gut microbiome. When considering human traits, such as the human microbiome, the overall variation observed within a population can be broadly attributed to genetic and environmental influences. Based on evidence for the modifiability of the microbiome in animal studies, which show strong effects of changes in maternal fostering, housing conditions or other environmental exposures [2], numerous studies on the human microbiome have focused on the role of anti-inflammatory and antibiotic treatments, different lifestyles (most prominently diet, smoking, and regular exercise) and more generally the effects of sharing a family or neighborhood environment. Each of these has shown to capable of modifying the gut microbiome [3-15].

In contrast to this accumulated evidence for the ability of environmental lifestyle factors to modulate gut microbiome characteristics [16, 17], the impact of the host genetic profile on the microbiome has been harder to delineate. Twin studies are the first method of choice to establish the genetic contribution to trait variation, and a number of these have been carried out [18, 19]. Whereas these studies unanimously show compelling evidence for a genetic component in the determination of gut microbiota composition, only a relatively small proportion of bacterial taxa shows moderate to high heritability (e.g., h2 > 20%), for instance the *Christensenellaceae* [19]. Similar modest and taxa-specific genetic contribution is reflected by genome-wide association studies (GWAS) on microbiome composition [20-22].

Little cross-replication of the GWAS signals related to microbiome compositions, so-called microbiome quantitative trait loci (mbQTLs), has been observed so far [23]. This lack of reproducibility may be due to heterogeneity in the collection, processing, sequencing, and annotation of stool microbiota. Likely, it additionally reflects a well-known issue in GWAS studies. In general, these are notorious for requiring large sample sizes to increase statistical power and thus the likelihood of detecting real significant associations between a trait of interest and various genomic loci [24]. The existing studies of a "mere" thousands of participants are prone to non-replication due to low power, even if a single trait is studied. The added complication in microbiome genetics is that hundreds of microbes are being tested, each representing an individual trait to compare against the genome for significance. This requires stringent multiple testing correction at the independent variable side of the regression equation used to test association (millions of genetic variants) and at the dependent variable side (hundreds of bacterial taxa).

The above challenges for detecting mbQTLs create the necessity of large-scale collaboration amongst research groups that have collected fecal DNA in many cohorts. The pooling of that data in meta-analysis is critical to elucidate loci that may influence the abundance of gut microbiome-associated organisms. To address this, the international MiBioGen consortium recently curated and analyzed whole-genome genotypes and 16S fecal microbiome data from 18,473 individuals from 25 cohorts [25]. Apart from providing increased statistical power, this collaboration also increases the representation of global populations, allowing for bona fide population differences in genetic or environmental effects. The MiBioGen meta-analysis identified 30 microbiome Quantitative Trait Loci (mbQTL) at a genome-wide significant ($P < 5 \times 10^{-8}$) threshold. Just one, the lactase (LCT) gene region, reached study-wide significance (GWAS signal $P = 8.6 \times 10^{-21}$). Other associations were suggestive ($1.94 \times 10^{-10} < P < 5 \times 10^{-8}$) but enriched for taxa showing high heritability and genes expressed in the intestine and brain. Follow-up analyses indicate enrichment of mbQTL SNPs in metabolic, nutrition, and environment domains, suggesting that most of the heritable contribution to microbiome composition depends on genetic effects on food preference and dietary behaviors.

Although a significant breakthrough for the field, one of the alarming features laid bare by the analyses of the MiBioGen consortium is the lack of standardization in microbiome assessment methodology. Many technical differences existed between the cohorts in the collection methods and DNA extraction protocols and the microbial composition of different cohorts was profiled by targeting three distinct variable regions of the 16S rRNA gene. Despite strict protocols to at least maximally standardize post-collection and post-sequencing processes, microbial composition showed very high variability across cohorts: only 9 out of the 410 genus-level taxonomic groups with a relative abundance higher than 1% in at least one cohort could be detected in more than 95% of the cohorts [25]. This immediately raises the question of how well each of the single cohorts can recapture the results of this meta-analysis. Although we fully expect that a single cohort does not have the power to detect genome-wide associations, we also assume that consortium-derived associations do generalize to the original contributing cohorts. That this assumption holds is vital when these cohorts, for instance, want to use leave-cohort-out meta-analysis based polygenic risk scores or the significant mbQTLS as genetic instruments for the microbiome composition, for purposes such as Mendelian Randomization analysis.

To address the issue of back-translation of consortium results to the cohorts contributing to a GWA meta-analysis, this chapter compares the GWA performed within the NTR to the meta-analysis of the entire MiBioGen consortium. We first investigated the performance of the top mbQTL result from the MiBioGen meta-analysis, rs182549, within the NTR cohort. We tested the hypothesis that this candidate SNP from the meta-analysis produces a significant result in this single cohort at a

nominal p-value for a single test ($\alpha = 0.05$). In reverse, we tested whether the most significant hit within the NTR cohort would attain significance in the meta-analysis (or be a false positive). Overall, we hypothesized that the results of the NTR genome-wide association analysis, by itself under powered, would still be reflective of the overall consortium analysis.

To further establish that a single cohort could reproduce the pattern of findings in the meta-analysis, we conducted an additional test. As indicated above, the already stringent multiple testing correction needed for the genome-wide association (millions of genetic variants) is compounded in microbiome genetics because the dependent variable is not a singular value (e.g., LDL-C or blood pressure) but consists of hundreds of different bacterial taxa, each of which needs to be tested against the full genomic variation. An important pattern of results seen in the consortium analysis was that the more heritable taxa showed much smaller p-values in the genome-wide association tests than less heritable taxa. This could be employed in GWA studies of microbiomes to reduce the multiple testing burden by prioritizing heritable taxa for association analysis, a strategy also applied in other high-dimensional phenotypes like MRI [26].

To see if the relationship between heritability and association statistics held in the NTR used as a stand-alone cohort, we tested whether the most heritable taxa in our twin sample, as assessed by the MZ correlations, were also the ones with the lowest p-values in the genetic association testing.

METHODS

PARTICIPANTS

The NTR sample used in MiBioGen ($N = 267$) was reduced compared to the full NTR biobank sample with fecal DNA used elsewhere in this dissertation because only a single MZ twin was allowed for the meta-analysis. We choose a random twin from all twin pairs.

SAMPLE PROCESSING AND DNA SEQUENCING

DNA extraction was performed on all samples using the MoBio PowerSoil kit, using approximately 75 mg of feces. The samples were quantified using the Qubit picogreen dsDNA picogreen fluorometric method (Invitrogen, Carlsbad, CA, USA). DNA is considered to be of high quality if the absorbance ratio at 260 nm and 280 nm (A260/A280) is between 1.7 and 2.00. DNA quantity from fecal extractions is considered sufficient if there is a concentration of at least five ng/ul within the 50 ul eluted. The V4 region of the 16S gene was amplified [27]. The resulting amplicons were sequenced in both the forward and reverse orientation using a 2x251 paired-end Illumina MiSeq run.

16S DATA PROCESSING

The resulting forward and reverse 16S reads from the MiSeq platform overlapped to produce contig reads using the algorithm implement within Mothur [27, 28]. After overlaying the reads, resulting sequences were screened to discard reads with ambiguous bases or sequences that were longer than 275 base-pairs in length. While previous research efforts produced by our research group have utilized de novo OTU picking strategies, these studies were able to adequately control aspects such as which variable region of the 16S gene is amplified before sequencing. The main differences of these approaches are that de novo OTU picking strategies define OTU clusters based on sequence similarity alone before generating a consensus OTU classification. In contrast, the closed-reference OTU strategy defines membership in a specific taxon by database classification alone, without regard to the similarity of reads given the same classification. While reference-free OTU strategies have been demonstrated to be preferential to reference-based methods [29], the considerable heterogeneity of methods used to interrogate the 16S gene region of bacterial genomes within the MiBioGen consortium, necessitated a database classification process. Before classification using the RDP classifier v.1.12, all samples were rarefied to 10,000 reads. All NTR samples had greater than 10,000 reads. The RDP classifier was used to bin reads into taxonomic units at various levels using a posterior probability of 0.8. Alpha diversity metrics, in the form of Shannon, Simpson, and inverse Simpson indices, were calculated on non-manipulated taxa counts. Additional analyses were performed using natural log-transformed taxa counts. Taxa counts were adjusted for the covariates, sex, age, PC1, PC2, PC3, microbiome sequencing batch, and the genotype batch. Residuals were scaled to have a mean of 0 and a standard deviation of 1. Taxa present in at least 10% of samples within a cohort were included in the quantitative microbiome QTL (Q-mbQTL) testing. The binary mQTL (B-mbQTL) analysis included taxa present in more than 10% and less than 90% of samples.

GENETIC DATA PROCESSING

The DNA samples collected from the various NTR participants within this study were genotyped on numerous platforms and, as such, needed to be harmonized. After harmonization of the genotype information, samples were imputed using the Michigan Imputation Server with the HRC reference panel. After imputation, the GenotypeHarmonizer software was used to filter the data. SNPs with a minor allele frequency > 0.05, pointwise imputation QC > 0.4 or SNP-wise call rate filtering > 0.95 were discarded.

Three main association testing strategies were performed in the overall consortium analysis. The first of these tests was aimed at determining whether there were genomic loci associated with gut microbiome alpha diversity. The second association method was aimed at identifying loci that are significantly associated with the abundance of individual taxa (mbQTL). The third association strategy was focused on identifying

loci associated with the presence or absence of different taxa, termed the binary association test (mbBTL). This chapter focuses on the quantitative aspect of the overall analysis. Quantitative mbQTL analysis was carried out using the Spearman correlation test between the SNP data and the covariate-adjusted bacterial abundance data. Samples with zero abundance for individual taxa are ignored in the association for that taxa.

COMPARISONS OF THE TOP CONSORTIUM MBQTLS IN NTR, AND VICE VERSA

The primary association analyses from the MiBioGen consortium identified a single mbQTL that reached the predefined significance level ($P = 8.6 \times 10^{-21}$). This association was between rs182549 and the abundance of the *Bifidobacterium* genus. We sought to identify the significance of this association amongst the numerous association analyses ran in the NTR cohort. Vice versa, we tested whether the mbQTL that showed the most significant effect in the NTR cohort could be found in the top MiBioGen results, using the browser containing all mbQTLs with a p-value less than 0.001.

MBQTLS IN LOW HERITABLE VERSUS HIGH HERITABLE TAXA

Prior to heritability estimation, the taxonomic abundance was normalized using inverse rank-sum transformation. Since the vast majority of the NTR twins included in this study comprised only MZ twins, heritability was calculated as the intraclass correlation coefficient (ICC) for the MZ twins. We computed ICCs for taxa defined by the genus, family, order, class, and phylum level. To test the hypothesis that the more heritable taxa within each level of taxonomic classification would be over-represented in the top NTR mbQTLs, we divided all taxa into a low and high ICC group, using various cut-offs for the ICC between 0.1 and 0.4. Using the top million lowest p-values of the NTR genome-wide association analysis, we next computed the mean p-value of the most significant hit per taxon for all taxa in the low and high ICC groups and plotted these as a function of the ICC cut-off.

RESULTS

COMPARISON OF THE TOP CONSORTIUM HIT IN NTR

Association analyses were performed in the NTR to identify genetic loci that influence the abundance of microbiome-associated taxa following the procedures used by all cohorts contributing to the overall meta-analysis of the consortium. Where the consortium identified the association between the *Bifidobacterium* genus and rs182549 on chromosome 2 as the most significant mbQTL (8.6332×10^{-21}), the NTR analysis did not. Not only was the association between the *Bifidobacterium* genus and rs182549, not the top hit, but this mbQTL also was not even in the top 1,000,000 most significant mbQTL associations. However, a deeper delve in the NTR-specific results for this mbQTL showed a p-value of 0.06 in NTR, so just not achieving the

nominal significance for a candidate gene approach. Although this locus was not the most significant in the NTR data set, it should be noted that the Cochran's Q test showed an NTR effect size similar to other cohorts (Figure 4A in Kurilshikov et al., 2020; Appendix A1). Furthermore, the leave-one-out strategy did not identify the NTR cohort as a significant contributor to heterogeneity at this locus. The overall results for the *Bifidobacterium* genus are depicted in Figures 4.1 and 4.2. The most significant association for the *Bifidobacterium* genus in the NTR cohort was with the SNP rs8122116. This SNP is an intron variant associated with the ZFP64 gene on chromosome 20.

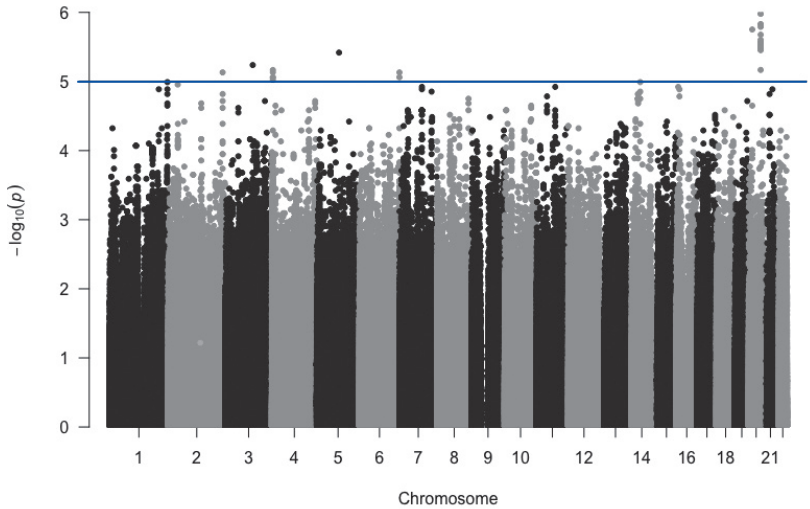


Figure 4.1 | Genome-wide association quantitative mbQTL results in the NTR cohort for the *Bifidobacterium* genus.

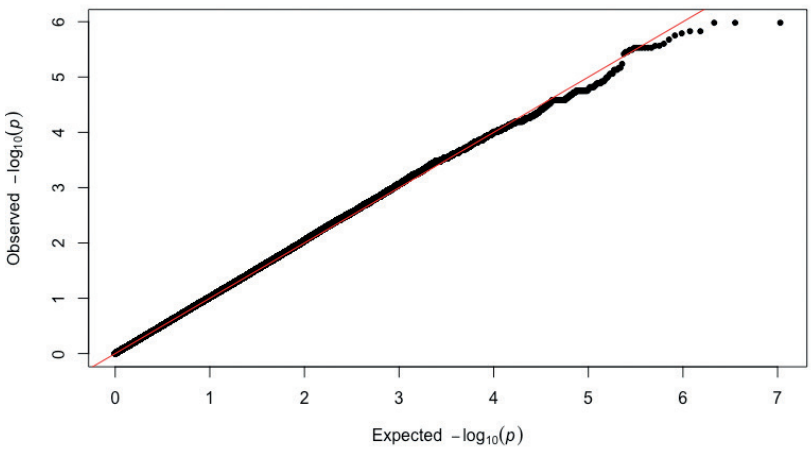


Figure 4.2 | QQ plot of the *Bifidobacterium* mbQTL results in the NTR cohort.

COMPARISON OF THE TOP NTR HIT IN THE CONSORTIUM

The mbQTL that showed the most significant effect in the NTR cohort was between rs11755686 and the *Clostridium sensu stricto 19* genus ($p = 1.05012 \times 10^{-9}$) as depicted in Figures 4.3 and 4.4. This identified SNP is an intron variant associated with the MAN1A1 gene on chromosome 6. Evaluating the top 1,174,837 results at $p < 0.001$ from the MiBioGen consortium meta-analysis showed no genome-wide significant associations for this taxon.

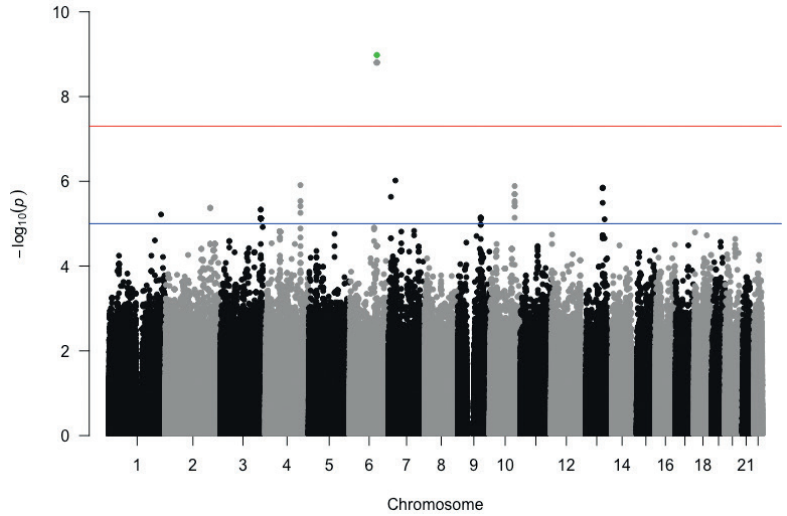


Figure 4.3 | Genome-wide association quantitative mbQTL results for the *Clostridium sensu stricto 19* genus (p = 1.05012e-9) in the NTR cohort. The most significant association on chromosome 6 (rs11755686) is labeled in green.

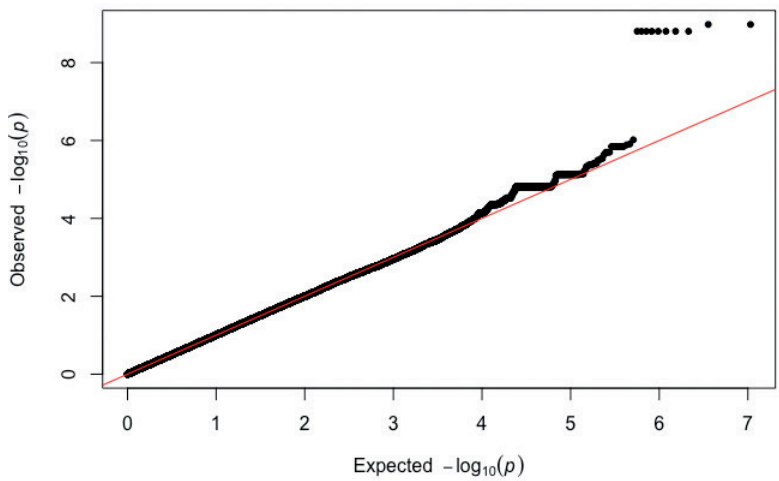


Figure 4.4 | QQ plot of the *Clostridium sensu stricto 19* mbQTL results in the NTR cohort.

ENRICHMENT OF HERITABLE MICROORGANISMS (ACCORDING TO THE MZ CORRELATIONS) IN THE NTR GWA RESULTS

The mean ICC in MZ twins for taxa at various levels of taxonomic classification is depicted in figure 4.5. The proportions of taxa with ICCs above 0.3 are described in Table 4.1. The phylum-level demonstrated the most substantial proportion of taxa with an ICC above 0.3, at 25%. The class-level showed the lowest proportion at 12.5%.

Table 4.1 | Proportion of taxa with MZ twin intraclass correlation coefficient larger than 0.3 for the various taxonomic levels tested.

Taxonomic Level	Proportion of taxa with ICC >0.3
Phylum	0.25
Class	0.125
Order	0.194
Family	0.246
Genus	0.178

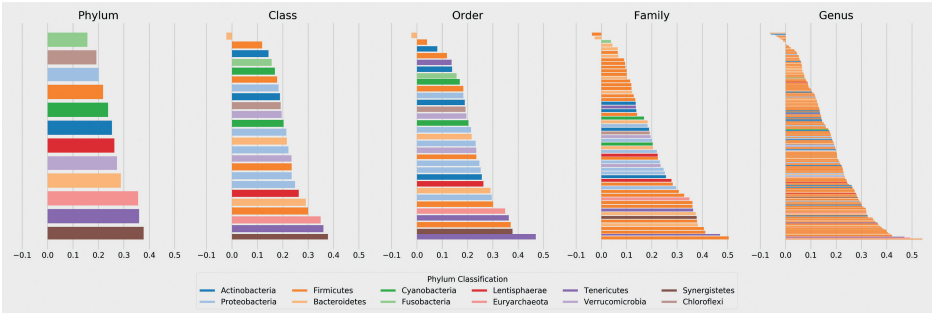


Figure 4.5 | Intraclass correlation coefficients for the individual taxa within the NTR cohort. Taxa are colored by the phylum classification.

Table 4.2 further depicts the taxa that showed the highest and lowest ICC values at the various taxonomic levels tested. The taxon that showed the largest ICC in the NTR sample was the *Ruminococcaceae UCG 10* genus group ($r = 0.54$, $p = 2.91 \times 10^{-13}$). The most significant association for this taxon was rs72712135 (p-value = 4.53521×10^{-7}) on chromosome 14, which is displayed in more detail in Figures 4.6 & 4.7. This SNP is a variant found 2KB upstream of the C14orf177 gene. Searching the mbQTLs meeting the 0.001 p-value cut-off in the overall consortium analysis returned 4,927 mbQTLs associated with the *Ruminococcaceae UCG 10* genus.

To test for an enrichment of more heritable microorganisms in the top NTR GWA results, we divided all taxa into a low and high ICC group for phylum, class, order, family and genus, using various cut-offs for the ICC. We next computed the mean p-value of

the most significant hit per taxon for all taxa within these low and high ICC groups at all levels of taxonomic classification (see Figure 4.8). We find that lower mean p-values for the taxa that showed higher heritability, but most convincingly for ICC cut-offs > 0.3 and at the family and genus level of classification.

Table 4.2 | Taxa identified as having the highest or lowest MZ-twin intraclass correlation coefficient within the numerous taxonomic levels tested.

Taxonomic Level		Taxa	ICC	ICC p-value
Phylum	High	*Synergistetes*	0.377	1.18861e-06
	Low	*Fusobacteria*	0.156	0.0515418
Class	High	*Synergistia*	0.377	1.18861e-06
	Low	*Cytophagia*	-0.022	0.785679
Order	High	*Anaeroplasmatales*	0.47	5.96735e-10
	Low	*Cytophagales*	-0.022	0.785679
Family	High	*Peptostreptococcaceae*	0.504	1.99325e-11
	Low	*Carnobacteriaceae*	-0.037	0.646196
Genus	High	*Ruminococcaceae UCG10*	0.541	2.90624e-13
	Low	*Parabacteroides*	-0.064	0.430724

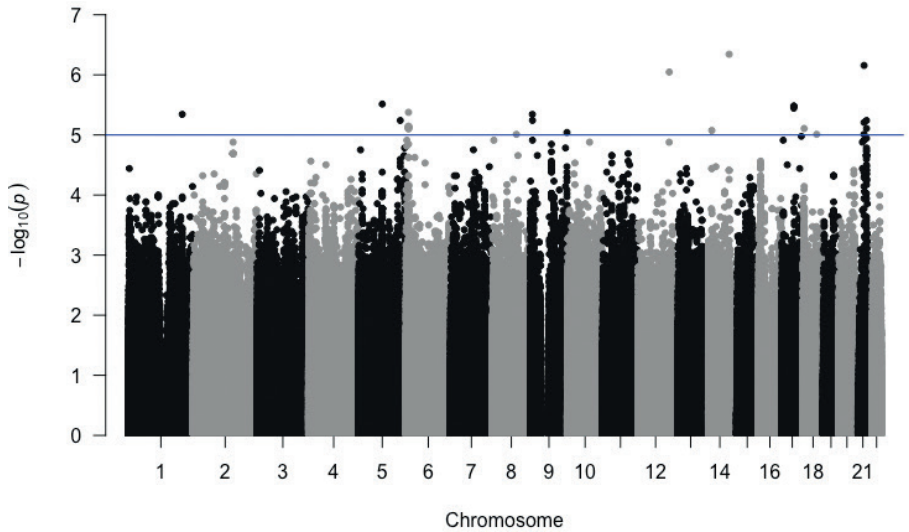


Figure 4.6 | Genome-wide association quantitative mbQTL results in the NTR cohort for the *Ruminococcaceae UCG 10 genus*.

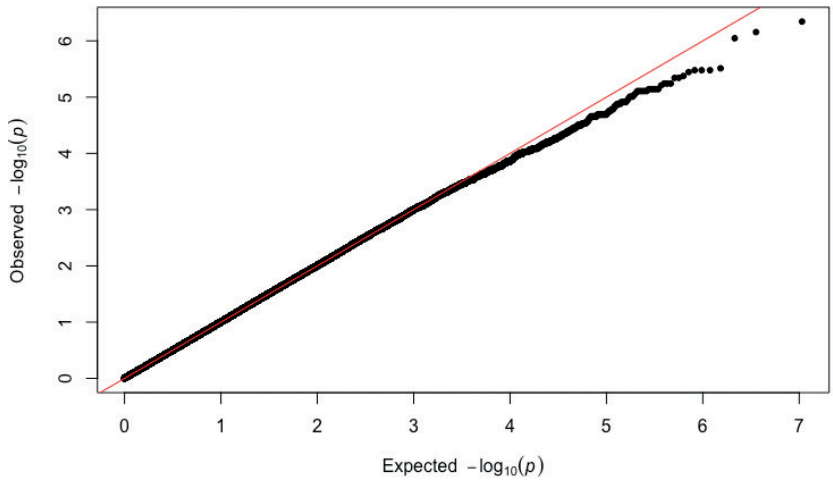


Figure 4.7 | QQ plot of the *Ruminococcaceae UCG 10* mbQTL results in the NTR cohort.

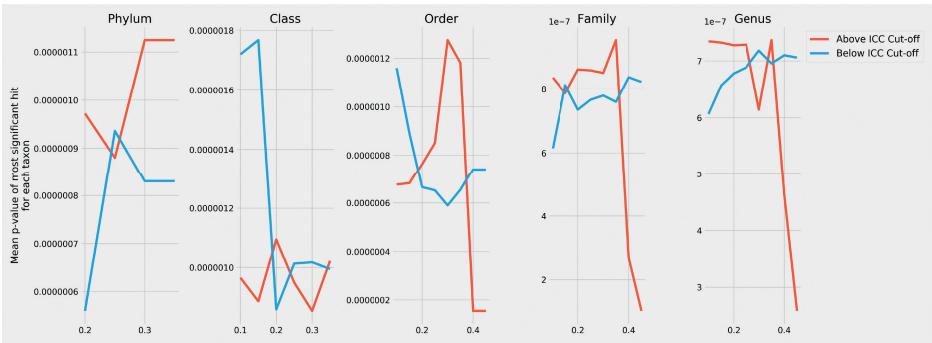


Figure 4.8 | Mean p-value of the most significant hit per taxon for all taxa in low and high ICC groups, plotted as a function of the ICC cut-off.

ENRICHMENT OF HERITABLE MICROORGANISMS (ACCORDING TO TWINSUK COHORT ACE MODELING) IN THE NTR GWA RESULTS

As a first ancillary analysis, we also looked at the most heritable taxa that were identified by the TwinsUK study that used both MZ and DZ twins to estimate the ACE variance components in full. The *Turicibacter* genus had the most significant genetic (A) component of the 19 significantly heritable taxa (A = 0.39875, p = 0.00512). We sought to investigate the GWAS performed within the NTR for this heritable genus. Within the NTR cohort, the most significant *Turicibacter* result pertained to rs35788049 (p-value = 1.52753×10^{-6}) (Figures 4.9 and 4.10). This SNP is an intron variant associated with the FOXG1-AS1 gene on chromosome 14.

As a second ancillary analysis, we looked at the *Peptostreptococcaceae* family for which the TwinsUK ACE modeling showed a significant genetic component (A = 0.388) and the NTR ICC for the *Peptostreptococcaceae* family was 0.50. The *Peptostreptococcaceae* GWAS results are depicted in figures 4.11 and 4.12, with the most significant association belonging to rs11963901 (p-value = 5.13417×10^{-8}). This SNP is an intron variant associated with the TIAM2 gene on chromosome 6.

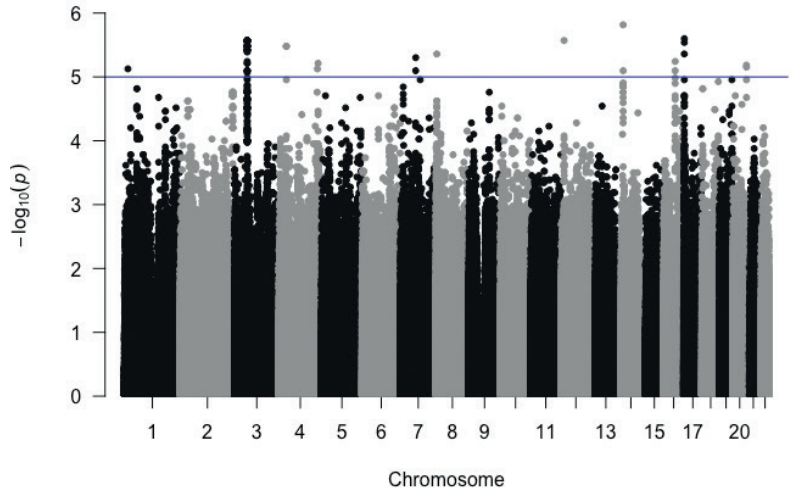


Figure 4.9 | Genome-wide association quantitative mbQTL results in NTR for the *Turicibacter* genus.

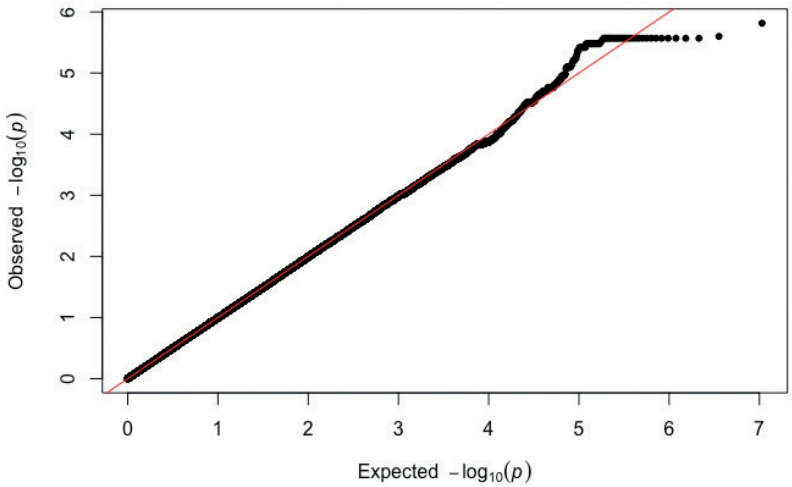


Figure 4.10 | QQ plot of the *Turicibacter* mbQTL results in the NTR cohort.

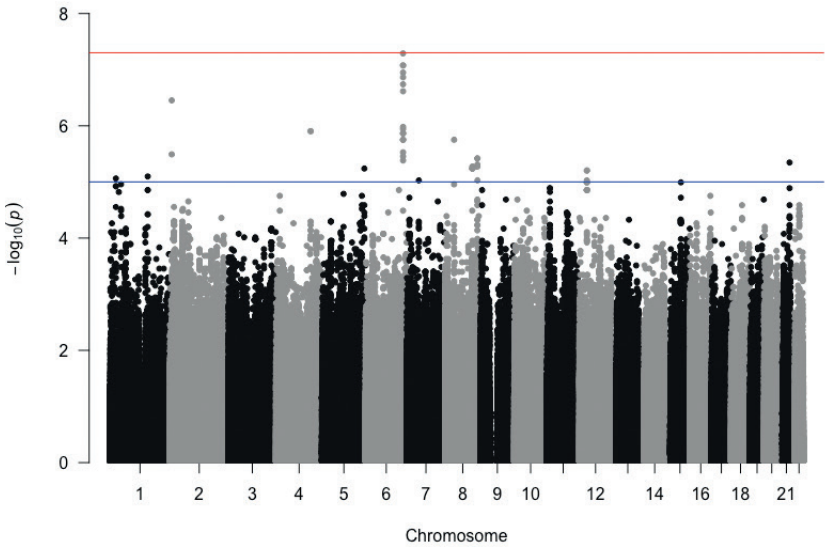


Figure 4.11 | Genome-wide association quantitative mbQTL results in the NTR cohort for the *Peptostreptococcaceae* family.

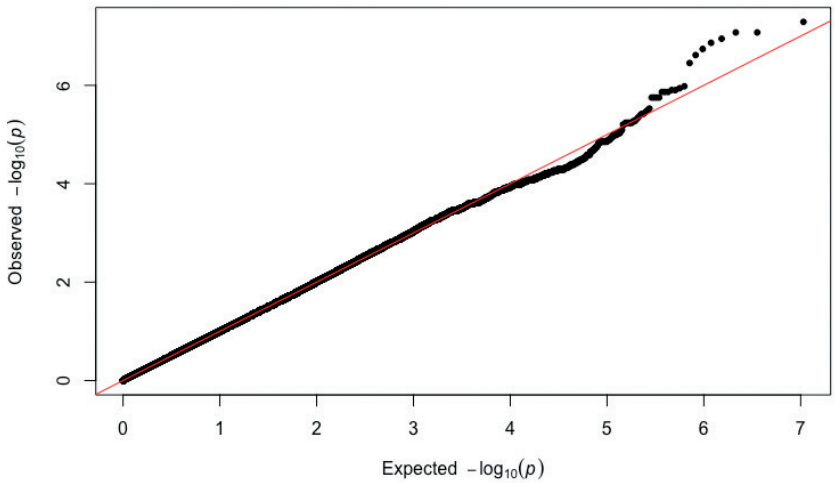


Figure 4.12 | QQ plot of the *Peptostreptococcaceae* family mbQTL results in the NTR cohort.

DISCUSSION

Within this investigation, we sought to determine how well the results generated in the individual NTR cohort reflected the results obtained in the larger consortium meta-analysis [23]. Our first method of comparison was to contrast the most significant mbQTL result from both the NTR cohort and the complete consortium. The consortium analysis had identified the *Bifidobacterium* genus and rs182549 as the most significant SNP-microbe association. While the effect of this mbQTL was not glaring in the NTR cohort, it was still in line with the expected effect size, generating a p-value of 0.06. This is a clear example of the increased power of the consortium affording the opportunity to identify this association. The NTR sample contained 267 individuals for the *Bifidobacterium* association test, whereas the overall meta-analysis contained 14,911 individuals.

We also set out to address the reverse question of how well the top NTR mbQTLs performed in the overall consortium meta-analysis. The most significant mbQTL in the NTR cohort was the *Clostridium sensu stricto 19* genus ($p = 1.05012 \times 10^{-9}$). Examining all consortium mbQTLs with a p-value lower than 0.001 returned no results for this taxon. In part, this could be caused by this specific taxon not being present in sufficient individuals across cohorts participating in MiBioGen, a common problem identified in this meta-analytic effort. Even in the NTR where *Clostridium sensu stricto 19* produced the most significant mbQTL, it was only present in 41 of the participants included in the association analysis. The most likely explanation, however, is that this mbQTL association was just a product of chance in NTR and once again demonstrates the necessity of including a large number of study participants, such as in the more considerable MiBioGen study.

One of the observations from the consortium meta-analysis was a relationship between heritable taxa with genome-wide significant mbQTLs and the number of suggestive GWAS hits ($P < 1 \times 10^{-5}$), indicating an agreement between traditional heritability estimation methods and GWAS results. Our results show this to hold even in a single cohort. We showed that the taxa demonstrating higher levels of heritability, in the form of ICCs greater than 0.3, tended to show a lower average p-value. This effect was most predominant at the Family and Genus levels. Examining the GWAS results for heritable taxa shows relatively strong genetic signals (*Turicibacter* p-value = 1.52753×10^{-6}, *Peptostreptococcaceae* p-value = 5.13417×10^{-8}), given the minimal sample sizes available for both of these taxa ($N = 170$ and $N = 270$). This confirms the notion that taxa showing genetic effects via the results of traditional heritability estimation methods, would also be expected to show genetic effects if studied with an appropriately powered GWAS.

This seemingly small connection between traditional heritability methods and expected GWAS results presents a positive opportunity for the microbiome field. As

indicated by the consortium meta-analysis, GWAS methods applied to microbiome-based phenotypes struggle due to the severe lack of overlap of gut microbiome-associate taxa across populations, inherently limiting possible sample sizes. This heterogeneity is likely to be more pronounced for gut microbiota that are sensitive to environmental factors, but perhaps less so for the more heritable species. Having an a priori idea that variation in a taxon is strongly genetic, as indicated through the family-based estimation, may allow for enhanced study design, perhaps in the form of a singular GWAS for this taxon with refined phenotype measurement.

A clear limitation of the heritability estimation in our study is that by using the MZ ICC as an estimate of genetic influences we are ignoring the effect of shared environment on the gut microbiome composition that could be driving the high intrapair resemblance. One method of accounting for this would be to incorporate ICCs generated from DZ twin pairs into our heritability estimation. Given the current lack of these data in the NTR cohort, another approach would be to identify taxa mainly influenced by environmental effects by considering the cohabitation status of the MZ twins surveyed and comparing these correlations to those of non-related spouses. Contrasting ICCs for these groups would effectively identify taxa with particular strong environmental effects, allowing us to separate those from taxa with strong genetic effects, and restricting genetic association testing to the latter.

In summary, our analyses confirm that the results from a meta-analysis performed in the large consortium consisting of data from 18,473 individuals in 25 cohorts is reasonably representative of the results obtained in 267 individuals in a single cohort. Accordingly, it would be appropriate to use the results of the MiBioGen consortium for extended analyses in the NTR cohort, such as the generation of polygenic risk scores or performing Mendelian Randomization.

REFERENCES

1. Gilbert, J.A., et al., *Current understanding of the human microbiome.* Nature medicine, 2018. **24**(4): p. 392-400.
2. Turner, P.V., *The role of the gut microbiota on animal model reproducibility.* Animal models and experimental medicine, 2018. **1**(2): p. 109-115.
3. Biedermann, L., et al., *Smoking cessation alters intestinal microbiota: insights from quantitative investigations on human fecal samples using FISH.* Inflammatory bowel diseases, 2014. **20**(9): p. 1496-1501.
4. Biedermann, L., et al., *Smoking cessation induces profound changes in the composition of the intestinal microbiota in humans.* PloS one, 2013. **8**(3): p. e59260.
5. Brito, I.L., et al., *Transmission of human-associated microbiota along family and social networks.* Nature microbiology, 2019. **4**(6): p. 964-971.
6. Dill-McFarland, K.A., et al., *Close social relationships correlate with human gut microbiota composition.* Scientific reports, 2019. **9**(1): p. 1-10.
7. Evans, C.C., et al., *Exercise prevents weight gain and alters the gut microbiota in a mouse model of high fat diet-induced obesity.* PloS one, 2014. **9**(3): p. e92193.
8. Finnicum, C.T., et al., *Cohabitation is associated with a greater resemblance in gut microbiota which can impact cardiometabolic and inflammatory risk.* BMC microbiology, 2019. **19**(1): p. 1-10.
9. Lee, S.H., et al., *Association between cigarette smoking status and composition of gut microbiota: population-based cross-sectional study.* Journal of clinical medicine, 2018. **7**(9): p. 282.
10. Ley, R.E., et al., *Human gut microbes associated with obesity.* nature, 2006. **444**(7122): p. 1022-1023.
11. Matsumoto, M., et al., *Voluntary running exercise alters microbiota composition and increases n-butyrate concentration in the rat cecum.* Bioscience, biotechnology, and biochemistry, 2008. **72**(2): p. 572-576.
12. Muegge, B.D., et al., *Diet drives convergence in gut microbiome functions across mammalian phylogeny and within humans.* Science, 2011. **332**(6032): p. 970-974.
13. Song, S.J., et al., *Cohabiting family members share microbiota with one another and with their dogs.* elife, 2013. **2**: p. e00458.
14. Walker, A.W., et al., *Dominant and diet-responsive groups of bacteria within the human colonic microbiota.* The ISME journal, 2011. **5**(2): p. 220-230.
15. Wu, G.D., et al., *Linking long-term dietary patterns with gut microbial enterotypes.* Science, 2011. **334**(6052): p. 105-108.
16. Rothschild, D., et al., *Environment dominates over host genetics in shaping human gut microbiota.* Nature, 2018. **555**(7695): p. 210-215.
17. Falony, G., et al., *Population-level analysis of gut microbiome variation.* Science, 2016. **352**(6285): p. 560-564.
18. Goodrich, J.K., et al., *Genetic determinants of the gut microbiome in UK twins.* Cell host & microbe, 2016. **19**(5): p. 731-743.
19. Goodrich, J.K., et al., *Human genetics shape the gut microbiome.* Cell, 2014. **159**(4): p. 789-799.
20. Bonder, M.J., et al., *The effect of host genetics on the gut microbiome.* Nature genetics, 2016. **48**(11): p. 1407-1412.

21. Turpin, W., et al., *Association of host genome with intestinal microbial composition in a large healthy cohort.* Nature genetics, 2016. **48**(11): p. 1413.
22. Wang, J., et al., *Genome-wide association analysis identifies variation in vitamin D receptor and other host factors influencing the gut microbiota.* Nature genetics, 2016. **48**(11): p. 1396-1406.
23. Kurilshikov, A., et al., *Host genetics and gut microbiome: challenges and perspectives.* Trends in immunology, 2017. **38**(9): p. 633-647.
24. Visscher, P.M., et al., *10 years of GWAS discovery: biology, function, and translation.* The American Journal of Human Genetics, 2017. **101**(1): p. 5-22.
25. Kurilshikov, A., et al., *Genetics of human gut microbiome composition.* BioRxiv, 2020.
26. Ge, T., et al., *Massively expedited genome-wide heritability analysis (MEGHA).* Proceedings of the National Academy of Sciences, 2015. **112**(8): p. 2479-2484.
27. Kozich, J.J., et al., *Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform.* Applied and environmental microbiology, 2013. **79**(17): p. 5112-5120.
28. Schloss, P.D., et al., *Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities.* Applied and environmental microbiology, 2009. **75**(23): p. 7537-7541.
29. Westcott, S.L. and P.D. Schloss, *De novo clustering methods outperform reference-based methods for assigning 16S rRNA gene sequences to operational taxonomic units.* PeerJ, 2015. **3**: p. e1487.

5

COHABITATION IS ASSOCIATED WITH A GREATER RESEMBLANCE IN GUT MICROBIOTA WHICH CAN IMPACT CARDIOMETABOLIC AND INFLAMMATORY RISK

This chapter was published as: Finnicum CT, Beck JJ, Dolan CV, et al. (2019) Cohabitation is associated with a greater resemblance in gut microbiota which can impact cardiometabolic and inflammatory risk. BMC Microbiology

ABSTRACT

The gut microbiota composition is known to be influenced by a number of factors including the host genetic profile and a number of environmental influences. Here, we focus on the environmental influence of cohabitation on the gut microbiota as well as whether these environmentally influenced microorganisms are associated with cardiometabolic and inflammatory burden. We perform this by investigating the gut microbiota composition of various groups of related individuals including cohabitating monozygotic (MZ) twin pairs, non-cohabitating MZ twin pairs and spouse pairs.

A stronger correlation between alpha diversity was found in cohabitating MZ twins (45 pairs, $r = 0.64$, $p = 2.21 \times 10^{-6}$) than in non-cohabitating MZ twin pairs (121 pairs, $r = 0.42$, $p = 1.35 \times 10^{-6}$). Although the correlation of alpha diversity did not attain significance between spouse pairs (42 pairs, $r= 0.23$, $p = 0.15$), the correlation was still higher than those in the 209 unrelated pairs ($r= -0.015$, $p = 0.832$). Bray-Curtis dissimilarity metrics showed cohabitating MZ twin pairs had the most similar gut microbiota communities which were more similar than the BC values of non-cohabitating MZ twins (empirical p-value = 0.0103), cohabitating spouses (empirical p-value = 0.0194), and pairs of unrelated non-cohabitating individuals (empirical p-value < 0.00001). There was also a significant difference between the BC measures from the spouse pairs and those from the unrelated non-cohabitating individuals (empirical p-value < 0.00001). Intraclass correlation coefficients were calculated between the various groups of interest and the results indicate the presence of OTUs with an environmental influence and one OTU that appeared to demonstrate genetic influences. One of the OTUs (OTU0190) was observed to have a significant association with both the cardiometabolic and inflammatory burden scores (p-values < 0.05).

Through the comparison of the microbiota contents of MZ twins with varying cohabitation status and spousal pairs, we showed evidence of environmentally influenced OTUs, one of which had a significant association with cardiometabolic and inflammatory burden scores.

INTRODUCTION

The gut microbiome plays an important role in health and disease, in a wide range of areas spanning outcomes in the cardiometabolic, immune and mental health domains [1-3]. The composition of the gut microbiota is now understood to be influenced by a number of factors including the host genetic profile [4, 5] and a myriad of environmental influences. These environmental factors can include the seeding at birth [6, 7], the composition of mothers milk or formula [8], exposure to pathogens [9] and health-associated behaviors like dietary intake and exercise activity [10-12]. Human studies comparing the microbiome of family members have been important in delineating the role of these various factors in influencing the human gut microbiome. Similarities in the gut microbiome of family members living together may be due to shared genetic factors, shared past, or current environmental exposures. Shared environmental factors may range from a shared womb (as applies to siblings and especially twins) passage through the birth canal to shared parenting rearing styles (breastfeeding), as well as all other exposures resulting from the actual sharing of a household (e.g. dietary habits, pet exposure, pollutant exposure). These common exposures may extend beyond the immediate household to an exposome shared by family members that includes neighborhood characteristics. Specific knowledge of the contribution of household effects to the abundance of specific microbial taxa could help delineate interventions to influence microbiome compositions associated with specific disease burden.

By comparing the resemblance of the gut microbiota of monozygotic and dizygotic twins, the relative contribution of genetic factors and shared environment lending to individual variation in the microbiome can be estimated, without the need to measure the genome or the environment [4, 5, 13, 14]. However, the shared environment is mainly reflective only of shared household effects up until late adolescence, at which point twins will typically move out of the family home.

When considering familial resemblances in adulthood, individuals typically will have started their own families and started to share their environment with others to whom they are not related (e.g. their spouses) or with whom they share genes and environment (e.g. their offspring). In the classical twin design, the effects of the current household of the twins tend to become subsumed under the non-shared or unique environment, to the extent that previous experience and genotype do not influence current household environment. Unique environment may also include many other environmental factors unrelated to the home environment such as exposures at work, as well as all measurement error. Therefore, estimation of the effects of sharing a household on the adult microbiome requires a unique design which recognizes current sharing in addition to carry over effects from previous household sharing.

Previous work has demonstrated that aspects of the gut microbiome composition are influenced by the individuals within a shared household, particularly between spouse pairs [15-17]. The impact that contact with others has on microbiota composition has even been found to extend to non-family members in a shared household as well as individuals within our social networks [13, 15]. Although these studies have provided evidence that cohabitation is capable of influencing the gut microbiota, it is necessary to confirm that microbes shared amongst cohabitating individuals are truly shared due to a common environment and not aspects of shared genetic ancestry. This is the case even for studies that focus on microbiome composition similarities amongst spouses and social networks as it has been previously demonstrated that spouse pairs and individuals within a similar social group resemble each other genetically more so than unrelated individuals [18, 19].

Here we detected shared household effects by two different strategies. First, we compared young adult MZ twins who still share a household (cohabitating) with each other to adult MZ twins who no longer share a household (non-cohabitating). Both types of MZ twin pairs are genetically identical. A larger resemblance in the abundance of specific (taxa of) microbes in younger MZ twins, who still share a household, compared to older MZ twin who do not, would reflect shared household effects. By observing OTUs strongly correlated between cohabitating MZ twin pairs but not non-cohabitating MZ twin pairs, we can be confident that the identified microbes are not influenced by genetic similarities. Second, we also tested whether similar cohabitation effects were found in unrelated persons sharing a household by comparing the resemblance in the gut microbiome of spouse pairs currently sharing a household to that of randomly matched pairs of the same age who never shared a household. For completeness, and to replicate previous findings that host genetic factors contribute to variation in the microbiome [4, 20-22], we further explored the microbiome resemblance in the MZ twins who do not share a household.

To explore the clinical relevance of the microbes detected to be sensitive to shared household effects, we computed the cardiometabolic and pro-inflammatory burden profiles based on a host of fasting blood-derived parameters [23]. Under the hypothesis that shared household can influence disease burden with the microbiome as a mediator, we expect the microbes that show a significant shared household effect to be associated with different metabolic and pro-inflammatory burden.

MATERIALS AND METHODS

PARTICIPANTS

Study participants consisted of individuals registered with the Netherlands Twin Register (N = 419, 272 females, 147 males). The majority of individuals assayed were MZ twins

and their spouses (MZ N = 332 (166 pairs), DZ N = 6 (3 pairs), spouses of MZ twins = 42, unrelated individuals = 39). Within the group of MZ twins, 45 pairs still cohabitated (mean age = 23.33, range 19-68), and 121 pairs no longer did so (mean age = 35.76, range 19-59, minimum live-apart time: 1 year, mean live-apart time: 17.77 years).

FECAL COLLECTION, DNA ISOLATION AND SEQUENCING

Fecal samples collected from participants were stored at 4°C until delivered to the laboratory within a 36-hour period. Anaerocult was used in order to preserve anaerobic species present within a sample. The samples were homogenized, aliquoted, and stored at -80 °C until used for microbial DNA extraction. DNA extraction was performed using the Qiagen Powersoil kit with the addition of the heating step from the Powerfecal kit (heating at 65 °C for 10 minutes). Resulting microbial DNA was subjected to PCR in order to amplify the V4 region of the 16S rRNA gene. The resulting library fragments were normalized using the SequalPrep normalization plates (Invitrogen, Carlsbad, CA). The libraries were pooled and analyzed via an Agilent Bioanalyzer trace. Samples were split into two sequencing runs in order to increase sample read depth. Samples were grouped together by family groups (twins, spouses) in order to make sure all samples from a family were sequenced in the same sequencing run. Sequencing data was generated on the MiSeq platform, using a 2x251 paired-end sequencing run with 20% Phix to increase base diversity during the run.

SEQUENCE QUALITY CONTROL

Sequence processing was carried out as previously described [24, 25]. Briefly, after the DNA sequencing process, demultiplexed forward and reverse reads were obtained after the DNA sequencing process using Mothur (version 1.39.5) [26]. Forward and reverse reads were overlapped in order to obtain contigs. We subsequently screened to discard reads longer than 275 base-pairs or reads that contained any ambiguous base calls. Unique reads were then aligned to a trimmed version of the SILVA (version 128) database containing the V4 region of the 16S rRNA gene. Reads that fell outside of this region were discarded. Performing the preclustering step, reads that only differed by up to two nucleotides were grouped. Chimera detection was performed using the VSEARCH algorithm (version 2.3.4) and probable chimeric reads were removed. Sequences were classified using a Bayesian classifier trained on the RDP database (version 16). Non-bacterial reads were removed from downstream analysis. After the aforementioned quality control process, sequence reads were clustered into species level operational taxonomic units (OTUs) at a 97% similarity cutoff through the use of the Opti-Clust algorithm [27], a *de novo* OTU clustering method implemented in Mothur (version 1.39.5). The formed OTUs were taxonomically labelled using the consensus taxonomy for each OTU. In order to explore higher taxonomic levels, phylotype binning was performed based on the classification of each sequence read. Phylotyping was performed at both the genus and family levels using Mothur (version 1.39.5). Total reads for each sample were subsampled to the depth of the sample with the lowest

sequence coverage (16,242 reads). After subsampling, alpha diversity was calculated for each sample in the form of Shannon and Chao1 indices. Beta diversity measures were also generated by computing Bray-Curtis dissimilarity measures between all individuals. A mock community was also sequenced along with the samples. Analysis of the mock community sequences after the sequence QC process determined the error rate to be 0.00253%.

CARDIO-METABOLIC AND INFLAMMATORY FACTOR MEASUREMENT

Data on a number of cardiometabolic and inflammatory factors were available for all of participants within the study. The measurement of these factors has been previously described elsewhere as well as the criteria for exclusion based on laboratory measurements, such as the measurements falling outside the limit of detection for a particular assay [23]. Cardiometabolic measures included body mass index, waist-hip ratio, LDL-cholesterol, HDL-cholesterol, triglycerides, glucose, insulin, systolic blood pressure (SBP), and diastolic blood pressure (DBP). Inflammatory traits included fibrinogen, interleukin-6 (IL-6), C-reactive protein (CRP), and tumor necrosis factor alpha (TNF-α). These data were used to generate disease burden scores separately for both inflammatory and cardiometabolic traits. Disease burden scores were standardized (i.e., mean of zero, standard deviation of one). To ensure that an increase in the variables assayed is associated with an increase in disease burden, the scale of some variables (e.g., HDL) were reversed by multiplying the standardized score by -1. Next, following the metabolic syndrome definition of the American Heart Association, we summed Z-scores for WHR, triglycerides, HDL-cholesterol, SBP and glucose to a single cardiometabolic burden score. There were 416 individuals that had valid measurements for all of the cardio-metabolic factors utilized. An additional pro-inflammatory burden score was computed by summing the Z-scores for fibrinogen, IL-6, CRP, and TNF-α. There were 401 individuals that had valid measurements for all of the inflammatory factors utilized.

STATISTICAL ANALYSES

After generation of Shannon and Chao1 indices for all participants, we sought to compare the resemblance of gut microbiota alpha diversity between individuals forming a twin or spouse pair and individuals sharing and not sharing a household. Pearson correlations in alpha diversity were computed in 1) cohabitating MZ twin pairs, 2) the non-cohabitating MZ twin pairs, 3) spouse pairs, and 4) pairs of randomly selected unrelated individuals who did not share a household. Selection of the latter pairs was performed in a manner that ensured the resulting unrelated pairs were not matched with the spouse of a co-twin and that both unrelated individuals were sequenced in the same sequencing batch. Matching a twin to the spouse of a co-twin could possibly inflate the level of similarity between the unrelated individuals whereas inclusion of unrelated pairs derived from multiple sequencing batches could artificially inflate the dissimilarity of the unrelated individuals relative to the various

family pairings (MZ twins or spouse pairs), which were always sequenced in the same sequencing batch. To confirm the effects of the household on adult gut microbiota composition, Bray-Curtis dissimilarity measures based on the species OTU counts were calculated. BC measures of cohabitating and non-cohabitating twins were compared using a t-test with 10,000 permutations. Likewise, the BC measures of cohabitating spouse pairs were compared to that of non-cohabitating unrelated pairs who were sequenced in the same sequencing batch (to account for the fact that all family members-twins and spouses were in the same sequencing batch).

Finally, restricting the analyses to OTUs present in 40% of individuals, we computed the intraclass correlations (ICCs) in the four different sets of pairings (cohabitating MZ pairs, non-cohabitating MZ twin pairs, unrelated spousal pairs sharing a household, and unrelated opposite sex pairs not sharing a household) for individual species level OTUs to detect household effects on specific OTUs. OTUs were restricted to those present in 40% of the individuals to ensure to restrict the range of sample sizes used to estimate the ICC's. Otherwise vastly different sample sizes. (i.e. ranging from 5 to 419) would cause strong sampling variation bias in the estimation of the ICCs. Unrelated opposite sex pairs were generated in a similar manner to the pairs of unrelated individuals described above with the addition of making sure the unrelated individuals were of the opposite sex and within 4 years of age. The threshold age difference was chosen because the mean difference in age between the spouses in the data was 3.4 years. We also tested the group differences in ICCs at the family, genus and species level using an F-test, with a Bonferroni correction on the subsequent p-values to account for multiple testing.

We considered OTUs to be affected by the shared household if they had significant intrapair similarity in cohabitating MZ pairs and significant intrapair similarity in spouse pairs but not in non-cohabitating MZ twin pairs or unrelated opposite sex pairs with a similar age distribution. For completeness we also identified OTUs with significant intrapair similarity in MZ pairs, whether cohabitating or non-cohabitating, and interpreted these as reflective of predominant genetic effects.

Given that mode of birth (cesarean section vs. vaginal birth) has been shown to at least influence the gut microbiota composition at least temporarily [28], we repeated all analyses after the exclusion of the individuals born via cesarean section (N = 43) to see how this impacted the results.

RESULTS

SHARED HOUSEHOLD EFFECTS ON ALPHA DIVERSITY

Figure 5.1 displays scatterplots of alpha diversity in the four parings of interest.

Non-cohabitating MZ twin pairs (121 pairs) showed a moderately strong correlation between alpha diversity measurements ($r = 0.42$, $p = 1.35 \times 10^{-6}$). The cohabitating MZ twin pairs (45 pairs) showed a stronger correlation relative to the non-cohabitating twins ($r = 0.64$, $p = 2.21 \times 10^{-6}$). The Pearson correlation of the alpha diversity in the unrelated, cohabitating spousal pairs (42 pairs) did not attain significance ($r = 0.23$, $p = 0.15$) but were still higher than those in the 209 unrelated pairs ($r = -0.015$, $p = 0.832$). This same pattern of results generally held true with the Pearson correlations performed on the Chao1 indices (cohabitating MZ: $r = 0.66$, $p = 6.77 \times 10^{-7}$, non-cohabitating MZ: $r = 0.58$, $p = 3.26 \times 10^{-12}$, spouse pairs: $r = 0.13$, $p = 0.40$, unrelated pairs: $r = -0.046$, $p = 0.51$). These comparisons were repeated after exclusion of the cesarean-born individuals. The pattern of the results did not change.

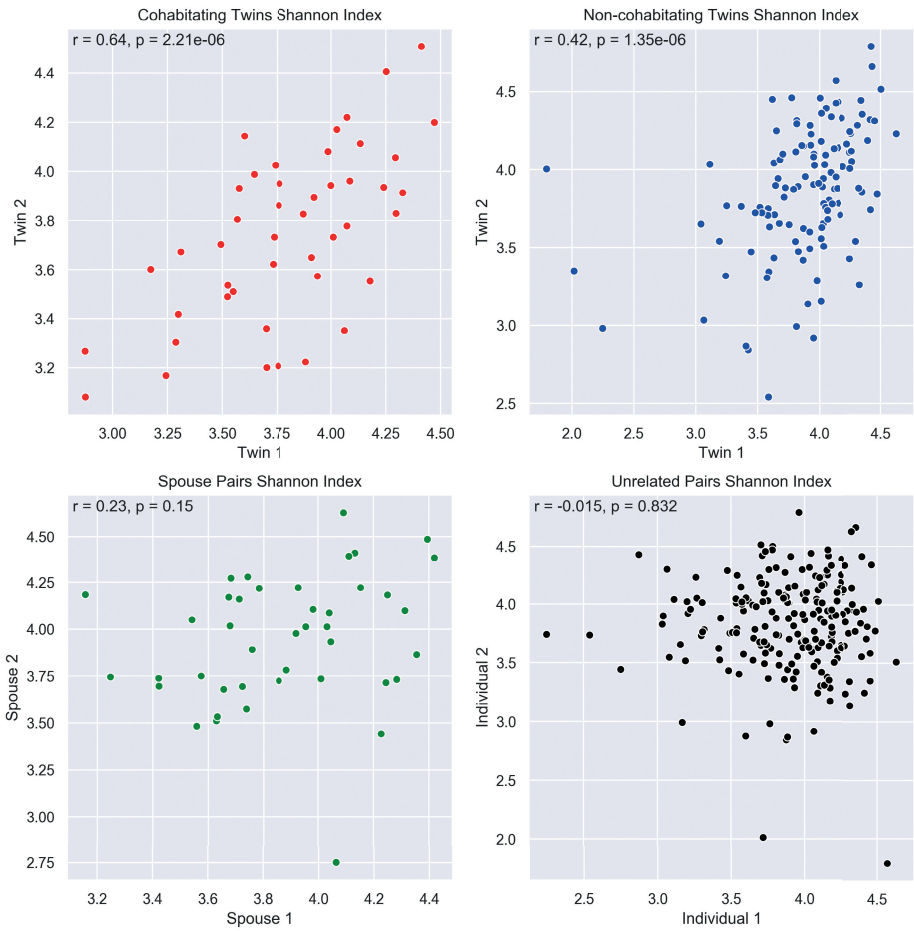


Figure 5.1 | Alpha diversity correlations between the different groups of individuals with varying degrees of relatedness.

SHARED HOUSEHOLD EFFECTS REFLECTED IN BETA DIVERSITY

Bray-Curtis dissimilarity metrics were generated between all individuals in the study. Figure 5.2 provides a boxplot of the BC values generated using all species level OTUs. A BC dissimilarity matrix was used as input for the principal coordinate analysis for visualization purposes (Figure 5.3). A series of t-tests including 10,000 permutations were used to determine differences in the mean BC metrics between any of the different groups with varying degrees of relatedness. Cohabitating MZ twin pairs had the most similar gut microbiota communities (lowest mean BC values) which was significantly lower than the BC of non-cohabitating MZ twins (empirical p-value = 0.0103), cohabitating spouses (empirical p-value = 0.0194), and pairs of unrelated non-cohabitating individuals (empirical p-value < 0.00001). There was also a significant difference between the spouse pairs and the unrelated non-cohabitating individuals (empirical p-value < 0.00001).

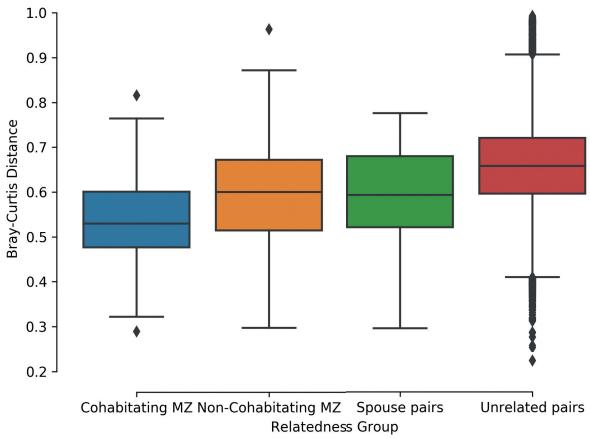


Figure 5.2 | Boxplot of Bray-Curtis dissimilarity corresponding to the various relatedness groups.

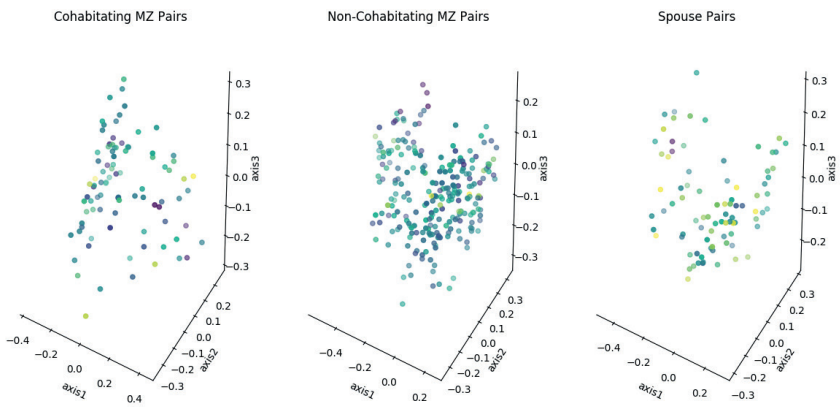


Figure 5.3 | PCoA plot generated from a Bray-Curtis dissimilarity matrix for visualization purposes.

SHARED HOUSEHOLD EFFECTS ON GENUS AND FAMILY LEVELS

Intraclass correlation coefficients were calculated between the aforementioned four sets of pairings for the genus and family taxons present in at least 40% of individuals. At the genus level there were 6 genera significantly correlated between non-cohabitating twins, 9 genera significantly shared between cohabitating MZ pairs, 3 shared genera amongst spouses and 1 genus that showed a significant correlation between the randomly generated opposite sex pairs (Table 5.1). The only overlapping genera was significantly correlated between cohabitating and non-cohabitating MZ pairs. This corresponded to an unclassified genus within the *Firmicutes* phylum.

At the family level there were 5 significantly correlated taxonomic families between cohabitating MZ pairs, 4 between non-cohabitating MZ pairs, 2 shared between spouses and no significantly correlated families between the randomly generated, opposite sex pairs. The family level showed greater overlap of the microorganisms shared between the sets of cohabitating and non-cohabitating MZ pairs with 3 families significantly correlated between both sets of MZ twins (Table 5.1). One of these families, an unclassified family within the *Firmicutes* phylum, was also significantly correlated between spouse pairs.

Table 5.1 | Genera and families identified as having a significant intraclass correlation coefficient (Bonferroni corrected p-value < 0.05)

	Cohabitating MZ	Non-Cohabitating MZ	Spouse
Genus	Firmicutes_unclassified*	Firmicutes_unclassified*	Barnesiella
	Senegalimassilia	Intestinimonas	Porphyromonadaceae_ unclassified
	Bacteria_unclassified	Dialister	Paraprevotella
	Veillonella	Akkermansia	
	Romboutsia	Terrisporobacter	
	Olsenella	Anaerostipes	
	Enterobacteriaceae_unclassified		
	Erysipelotrichaceae_unclassified		
	Flavonifractor		
Family	Firmicutes_unclassified*	Firmicutes_unclassified*	Firmicutes_unclassified*
	Bacteria_unclassified	Verrucomicrobiaceae	Porphyromonadaceae
	Enterobacteriaceae	Peptostreptococcaceae	
	Peptostreptococcaceae	Ruminococcaceae	
	Ruminococcaceae		

* = indicates same unclassified taxon.

SHARED HOUSEHOLD EFFECTS ON SPECIES LEVEL MICROBES

Figure 5.4 shows that the intraclass correlation coefficient (ICC) for cohabitating MZ pairs at the species level was generally higher than that for non-cohabitating pairs across all analyzed species level species level OTUs. Figure 5.5 similarly presents the difference between the intraclass correlation coefficients of the spouse pairs and the random opposite sex pairs. Species level OTUs were then separated based upon the consensus phylum classification of the OTU to determine the proportion of OTUs within a phylum that have a higher ICC in cohabitating MZ twins relative to non-cohabitating MZ twins (Table 5.2).

Maintaining strict correction for multiple testing (928 comparisons), the cohabitating MZ twin pairs had 13 OTUs with a corrected p-value meeting the predefined cutoff of 0.05, and the non-cohabitating MZ twin pairs also had 13 significant OTUs. OTU0095 was the only significant OTU that overlapped between the cohabitating (OTU0095 ICC: 0.625, F-stat: 4.34, p-value: 0.0013) and non-cohabitating MZ twins (OTU0095 ICC: 0.666, F-stat: 2.08, p-value: 0.032), suggesting a predominant genetic influence on intrapair resemblance.

ICC calculations between cohabitating spouse pairs resulted in 4 significant OTUs. In contrast, none of the randomly matched non-cohabitating pairs of random opposite sex pairs had significant OTUs. Of particular interest are OTU0081 and OTU0190 because these two species level OTUs showed a particularly strong relationship between both cohabitating MZ twin pairs (OTU0081 ICC: 0.676, F-stat: 5.18, p-value: 9.23×10^{-5}; OTU0190 ICC: 0.660, F-stat: 4.87, p-value: 2.31×10^{-4}) as well as the spousal pairs (OTU0081 ICC: 0.762, F-stat: 7.42, p-value: 9.26×10^{-7}; OTU0190 ICC: 0.666, F-stat: 5.00, p-value: 3.96×10^{-4}) but not in non-cohabitating MZ twin pairs.

Table 5.2 | Percentage of species level OTUs within a phylum that have a greater ICC value in the cohabitating MZ twins relative to the non-cohabitating twins. Phylum classification was based on the consensus taxonomic classification of the OTU.

PHYLUM OF OTU	# OF OTUS	PERCENTAGECOHAB ICC >NON-COHAB ICC
Actinobacteria	10	80%
Unclassified bacteria	4	75%
Bacteroidetes	26	61.5%
Firmicutes	182	54.4%
Proteobacteria	9	55.6%
Verrucomicrobia	1	0%

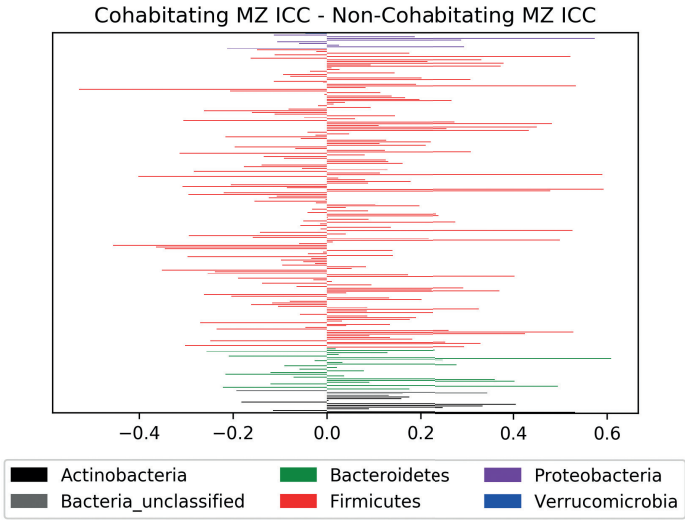


Figure 5.4 | Difference in the intraclass correlation coefficients (ICC) from the cohabitating and non-cohabitating twin pairs. Bars are labeled with the phylum classification of the species level OTU (0.03 cutoff). Bars that extend to the left indicate a larger intraclass correlation coefficient in the non-cohabitating MZ pairs for that particular OTU (Non-cohab. ICC > Cohab. ICC), whereas bars extending to the right indicate a larger intraclass correlation coefficient in the cohabitating MZ pairs relative to the non-cohabitating MZ pairs (Cohab. ICC > Non-cohab.).

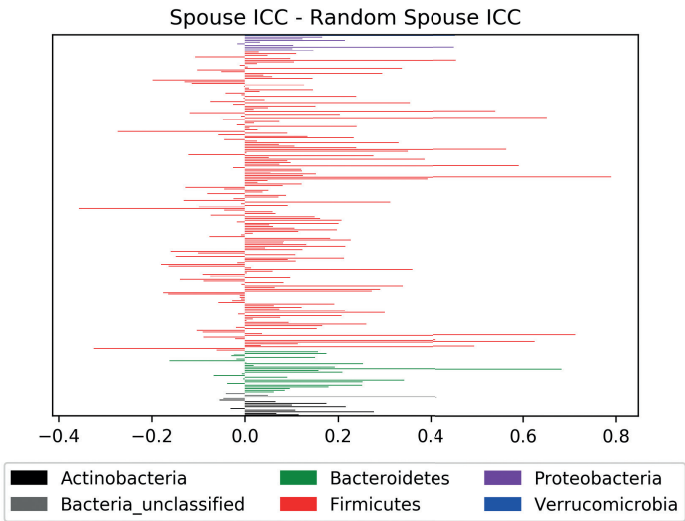


Figure 5.5 | Difference in the intraclass correlation coefficients (ICC) from the spouse pairs and the randomly generated spouse pairs. Bars are labeled with the phylum classification of the species level OTU (0.03 cutoff). Bars that extend to the left indicate a larger intraclass correlation coefficient in the unrelated spouse pairs for that particular OTU, whereas bars extending to the right indicate a larger intraclass correlation coefficient in the actual pairs relative to the unrelated spouse pairs.

ASSOCIATION OF THE SHARED HOUSEHOLD MICROBIOME EFFECTS WITH CARDIOMETABOLIC AND INFLAMMATORY BURDEN SCORES

Since the shared household affects alpha diversity, we first tested whether, across all participants, alpha diversity was associated with either cardiometabolic or inflammatory burden profiles, which did not yield significant results at a predefined alpha of 0.05.

Next, the species level OTUs previously observed as being particularly modulated by household effects (OTU0081 and OTU0190) were further explored in order to determine whether either of these OTUs were associated with cardiometabolic or inflammatory burden profiles. Burden scores were regressed on the OTU counts in the full sample using a Generalized Estimating Equation regression accounting for the relatedness of the MZ twins. Age and sex were also included in the GEE models. OTU0190 was observed to have a significant association with both the cardiometabolic and inflammatory burden scores (beta= -0.0072, $p < 0.05$; beta= -0.0085, $p < 0.05$, respectively). OTU0081 was not significantly associated with either the inflammatory or cardiometabolic burden scores.

Because cohabitating twins tend to be younger than non-cohabitating twins, our design assumes that there are no systematic age effects on intrapair resemblance confounding the comparison of habituating versus non-habituating resemblances. We tested this explicitly by regressing intrapair differences for the OTUs of interest on age, household status and the interaction of age and household status. We did not observe any significant associations between the predictors and the intrapair resemblance at a predefined alpha of 0.05.

DISCUSSION

Our results highlight a role for shared household effects on the adult gut microbiome. This held strongly for alpha diversity and beta diversity measures, with household effects on family, genus, or species levels harder to pinpoint. However, the species level OTUs, OTU190 and OTU0081, found to be significantly associated between cohabitating twin pairs and spouse pairs, but not within the non-cohabitating MZ twin pairs, provide direct evidence that specific members of the gut microbiota can be heavily influenced by environmental conditions. These findings support previous work that showed that beta diversity values as well as bacterial SNP variant similarity were positively correlated with the number of years that 32 MZ and 92 DZ twin pairs lived apart [5]. A number of other studies have also consistently demonstrated that twin pairs are more similar in gut microbiome composition from early life [29] as well as later in life [4] with lower beta-diversity measurements between twin pairs relative to unrelated individuals. Results are also congruent with the significant similarities in

the microbiomes found in 32 genetically unrelated individuals, who reported sharing a household and subsequently microbiome composition at the species level [13].

While alpha diversity appears to show a trend towards a household effect, it is worth noting the fairly strong correlation in alpha diversity measurements between MZ twins that are no longer cohabitating. This demonstrates that the shared household effect on gut microbiota alpha diversity is either so pronounced that similarities are still observed for long periods after cohabitation, or there are additional genetic effects on alpha diversity. Previous work performed in twin based microbiome studies has shown that the gut microbiome alpha diversity has a weak heritable component [21], while another recent study did not find evidence of similarities in alpha diversity between individuals of similar ancestry or genetic kinship [13]. Taken together, current evidence suggests a strong role for environmental influences on the gut microbiota composition.

Earlier studies in family members have often used the ongoing sharing of a household by spouses or siblings to detect its effects on the microbiome. However, these designs do not take into account that the resemblance of spouses can be decreased through sex effects on the microbiome as spouse pairs are (in majority) of opposite sex. Comparing resemblances in siblings or parent-offspring designs will confound the shared household effects with shared genetic effects on the microbiome, which are known to be non-zero [4, 21]. The classical twin design can estimate the relative contribution of genetic factors and shared environment to variation in the microbiome but then the shared environment is not specific for the actual current sharing of a household and includes sharing of pre- and perinatal and early life factors. By comparing adult MZ twin pairs that do or do not share a household, where both groups contain genetically identical individuals but differ in living status (whether living with twin or spouse), we obtain the least confounded view on the association of the current household with the microbiome composition. Our results show a reassuring convergence of the various designs. The systematic presence of more sharing of microorganisms between cohabitating compared to non-cohabitating MZ twin pairs and between spouse pairs compared to random male-female pairs and demonstrates the ability of a shared household to modulate specific microbiota members regardless of genetic similarity or sex.

OTU190 and OTU0081 were given consensus classifications as *Ruminococcaceae* and *Clostridiales* respectively. We attempted to further taxonomically characterize these species level OTUs by generating sequences that best represent these OTUs and performing a BLAST search with the subsequent FASTA file. OTU81 returned results indicating an uncultured *Oscillibacter* species, which resides within the *Clostridales* order. OTU190 returned more ambiguous results, which were largely classified as uncultured bacterium with an occasional hit reaffirming the consensus taxonomic

classification of *Ruminococcaceae*. It should be noted that because the OTUs are derived from *de novo* OTU clustering of pairwise sequence distances, the consensus classification process results in varying levels of classification with OTU0081 classified to the family level and OTU190 classified to the order level. In fact, *Ruminococcaceae* is a member of the *Clostridiales* order. Organisms belonging to the *Ruminococcaceae* family have been shown in previous studies to be impacted by both high fat diet and exercise activity [10]. A shared diet is an obvious component that may account for the found shared household effects here, both with regard to the alpha and beta diversity measures as the specific effects on the *Clostridiales* OTUs. Some evidence for clinical relevance of this shared household effects was found in the downstream effects of OTU0190 on both the inflammatory burden score and the cardiometabolic burden score.

OTU0095 showed a very strong resemblance within MZ pairs, independent of whether they were cohabitating or non-cohabitating, suggesting a predominant genetic influence on the intrapair resemblance for this OTU. Consensus classification of OTU0095 determined the OTU belonged to the *Lachnospiraceae* family. Previous work by Goodrich et al. [4] determined that the amounts of *Lachnospiraceae* were more similar between MZ twins relative to DZ twins. Furthermore, *Lachnospiraceae* was identified as one of the two taxonomic families that contained the majority of OTUs with the highest heritability estimate. We did not observe evidence of this OTU influencing either the cardiometabolic or inflammatory disease burden scores.

By exploiting the rapid progress in molecular genetic technology we have excellent strategies at our disposal to identify the elements in the microbiome that are influenced by genetic factors, and large scale international efforts for genome wide association studies are underway [30]. To detect environmental influences, including those of a shared household, the main strategy has been to select and measure a specific environmental factor and test its covariance with microbiome composition. A disadvantage of that strategy is that we do not uncover the effects of yet unknown environmental factors. The approach employed here, comparing cohabitating and non-cohabitating twins, and cohabitating spouse pairs to age-matched non-cohabitating pairs, provides a route to test the effects of unmeasured environmental factors on the microbiome, at least with regard to the shared household component.

CONCLUSIONS

Through the sampling of cohabiting MZ twins, non-cohabitating MZ twins and spouse pairs, we were afforded the unique opportunity to observe similarities and differences between these groups with regard to the gut microbiota, highlighting a role of cohabitation in shaping the gut microbiota composition. This study clearly

demonstrates that cohabitation results in a similar gut microbiota alpha diversity and lower BC distances relative to unrelated individuals. Furthermore, individual OTUs at varying taxonomic levels were found to be impacted by a shared household status. One species level OTU was found to be significantly associated with both cardiometabolic and inflammatory disease burden.

REFERENCES

1. Malan-Muller, S., et al., *The gut microbiome and mental health: implications for anxiety-and trauma-related disorders.* Omics: a journal of integrative biology, 2018. **22**(2): p. 90-107.
2. Belkaid, Y. and T.W. Hand, *Role of the microbiota in immunity and inflammation.* Cell, 2014. **157**(1): p. 121-141.
3. Hansen, T.H., et al., *The gut microbiome in cardio-metabolic health.* Genome medicine, 2015. **7**(1): p. 33.
4. Goodrich, J.K., et al., *Human genetics shape the gut microbiome.* Cell, 2014. **159**(4): p. 789-799.
5. Xie, H., et al., *Shotgun metagenomics of 250 adult twins reveals genetic and environmental impacts on the gut microbiome.* Cell systems, 2016. **3**(6): p. 572-584. e3.
6. Korpela, K., et al., *Selective maternal seeding and environment shape the human gut microbiome.* Genome research, 2018. **28**(4): p. 561-568.
7. Dominguez-Bello, M.G., et al., *Delivery mode shapes the acquisition and structure of the initial microbiota across multiple body habitats in newborns.* Proceedings of the National Academy of Sciences, 2010. **107**(26): p. 11971-11975.
8. Pannaraj, P.S., et al., *Association between breast milk bacterial communities and establishment and development of the infant gut microbiome.* JAMA pediatrics, 2017. **171**(7): p. 647-654.
9. Groves, H.T., et al., *Respiratory disease following viral lung infection alters the murine gut microbiota.* Frontiers in immunology, 2018. **9**: p. 182.
10. Evans, C.C., et al., *Exercise prevents weight gain and alters the gut microbiota in a mouse model of high fat diet-induced obesity.* PloS one, 2014. **9**(3): p. e92193.
11. Clarke, S.F., et al., *Exercise and associated dietary extremes impact on gut microbial diversity.* Gut, 2014. **63**(12): p. 1913-1920.
12. De Filippo, C., et al., *Impact of diet in shaping gut microbiota revealed by a comparative study in children from Europe and rural Africa.* Proceedings of the National Academy of Sciences, 2010. **107**(33): p. 14691-14696.
13. Rothschild, D., et al., *Environment dominates over host genetics in shaping human gut microbiota.* Nature, 2018. **555**(7695): p. 210-215.
14. Reyes, A., et al., *Viruses in the faecal microbiota of monozygotic twins and their mothers.* Nature, 2010. **466**(7304): p. 334-338.
15. Brito, I.L., et al., *Transmission of human-associated microbiota along family and social networks.* Nature microbiology, 2019. **4**(6): p. 964-971.
16. Dill-McFarland, K.A., et al., *Close social relationships correlate with human gut microbiota composition.* Scientific reports, 2019. **9**(1): p. 1-10.
17. Song, S.J., et al., *Cohabiting family members share microbiota with one another and with their dogs.* elife, 2013. **2**: p. e00458.
18. Domingue, B.W., et al., *The social genome of friends and schoolmates in the National Longitudinal Study of Adolescent to Adult Health.* Proceedings of the National Academy of Sciences, 2018. **115**(4): p. 702-707.
19. Domingue, B.W., et al., *Genetic and educational assortative mating among US adults.* Proceedings of the National Academy of Sciences, 2014. **111**(22): p. 7996-8000.

20. Benson, A.K., et al., *Individuality in gut microbiota composition is a complex polygenic trait shaped by multiple environmental and host genetic factors.* Proceedings of the National Academy of Sciences, 2010. **107**(44): p. 18933-18938.
21. Goodrich, J.K., et al., *Genetic determinants of the gut microbiome in UK twins.* Cell host & microbe, 2016. **19**(5): p. 731-743.
22. Bonder, M.J., et al., *The effect of host genetics on the gut microbiome.* Nature genetics, 2016. **48**(11): p. 1407-1412.
23. Sirota, M., et al., *Effect of genome and environment on metabolic and inflammatory profiles.* PLoS One, 2015. **10**(4): p. e0120898.
24. Kozich, J.J., et al., *Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform.* Applied and environmental microbiology, 2013. **79**(17): p. 5112-5120.
25. Finnicum, C.T., et al., *Metataxonomic analysis of individuals at BMI extremes and monozygotic twins discordant for BMI.* Twin Research and Human Genetics, 2018. **21**(3): p. 203-213.
26. Schloss, P.D., et al., *Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities.* Applied and environmental microbiology, 2009. **75**(23): p. 7537-7541.
27. Westcott, S.L. and P.D. Schloss, *OptiClust, an improved method for assigning amplicon-based sequence data to operational taxonomic units.* MSphere, 2017. **2**(2).
28. Stinson, L.F., M.S. Payne, and J.A. Keelan, *A critical review of the bacterial baptism hypothesis and the impact of cesarean delivery on the infant microbiome.* Frontiers in medicine, 2018. **5**: p. 135.
29. Hill, C.J., et al., *Evolution of gut microbiota composition from birth to 24 weeks in the INFANTMET Cohort.* Microbiome, 2017. **5**(1): p. 4.
30. Wang, J., et al., *Meta-analysis of human genome-microbiome association studies: the MiBioGen consortium initiative*. 2018, BioMed Central.

5

6

SUMMARY AND DISCUSSION

In the past decades, the exponential increase in the scale and richness of information extracted from DNA in molecular genetic laboratories, including the one we run at the Avera Institute of Human Genetics, has transformed the field of genetic epidemiology beyond recognition [1, 2]. A major contribution has been made by international GWAS consortia that have uncovered more than 75000 variant-trait associations leading to a deeper understanding of biological mechanisms of disease, new therapeutic leads, and better risk prediction. However, other than linking genetic variation in the DNA code to health-related traits, DNA itself can be a source of relevant phenotypic information that may be present at conception or arise de novo across the lifespan. Three examples of such "DNA-based phenotypes" are telomere length (TL), epigenetic modification of the DNA through methylation, and the composition of the genomic content of the many microorganisms that each of us carries in our body - the microbiome.

As part of my academic training, I mastered the assessments of these techniques, and spent considerable time refining the telomere length assay. But already in the early phase of my Ph.D. trajectory, my interest and focus turned to the gut microbiome, which became a central theme in my thesis. My thesis's second central theme is the use of several genetically informative study designs enabled by the Netherlands Twin Register and its vast Biobank resource. With the experiments outlined within this dissertation, we have been able to not only determine the presence of genetic contribution to DNA-based phenotypes but additionally, in parallel, explore the impact of environmental components on these traits. Below I summarize the main findings in this thesis per chapter before concluding with a more general discussion of the current state of microbiome research and future developments.

TELOMERE LENGTH

In the second chapter, we studied a well-established aging-associated biomarker, telomere repeat mass (TRM), as an index of telomere length. This chapter's main goal was to see whether we can safely use buccal-cell derived TRM as an alternative to blood leukocyte-based TRM to learn about telomere-associated dynamics across an individual's lifetime. The ability to utilize the more easily collectible buccal DNA sample would greatly aid researchers' ability to collect and perform telomere measurements from a very young age onwards. These longitudinal studies will be valuable in understanding how this aging-associated biomarker varies throughout one's lifetime.

In a broad and overlapping set of NTR participants, DNA was collected from both blood and buccal samples. Using these DNA samples, we established:

1. the buccal to blood correspondence in TRM,
2. the variation in TRM due to different laboratories using different molecular methods,
3. the genetic and environmental contribution to TRM in blood and buccal samples.

Ad 1) Blood and buccal derived DNA samples collected from the same individual were subjected to the identical telomere mass measurement and compared. Specifically, one assessment of TRM in blood was done in Leicester, England (Blood-1, N = 1892) and one at the Avera Institute for Human Genetics (AIHG; Blood-2, N = 1338). All buccal assessments were also done at AIHG (N = 1338). The comparisons showed that, when performed by the same laboratory (AIHG), 11% to 12% of the variance in TRM in blood samples was predicted by TRM in buccal cells, i.e., the correlation was significant but modest (r = 0.244 - 0.415). The buccal associated DNA did display the same age and sex associated effects commonly documented in measurements based on blood-derived DNA, indicating that the TRM data is showing the expected male disadvantage and telomere attrition with age.

Ad 2) In addition to the validation of buccal DNA for TRM measurement, our study allowed us to test the similarity in TRM measurements performed on the same sample in different laboratories. The blood derived TRM measured across different laboratories correlated about 0.58, so also not perfect. This non-perfect correlation highlighted differences in the molecular measurements performed in different laboratories, even when using the same blood sample. In part, this may reflect "sample aging." By the time the blood DNA samples were analyzed for TRM the second time at AIHG, several genomic analyses had been performed on the samples. This increased sample handling may also have caused the modest buccal-blood correlation in the TRM measurement. We note that batch effects indeed accounted for a less substantial part of the TRM variance in the first analysis (12.4%) than in the second analysis of the same blood sample (23.1%).

Ad 3) Performing the aforementioned telomere measurement in a cohort containing MZ and DZ twin pairs allowed us to determine the contribution of genetic and environmental factors to each of the two blood and the buccal TRM phenotypes. The correlation coefficients in MZ and DZ twins were sufficiently similar to adopt an ACE model. Effects of the shared environment were indeed found for TRM in blood in the first (17.9%) and second analysis (18.6%), and for TRM in the buccal analysis (24.4%). The two analyses in the blood samples showed a difference in estimated heritability, with genetic factors explaining 47.6% and 22.2% of total phenotypic variance in TRM in the first and second analysis of the same blood sample. Heritability of buccal TRM was 23.3%. We speculated that the differences in heritability estimates for the two blood samples are due in part to technical variation (i.e., batch effects accounting for 12.4% of TRM variance in Blood-1 vs. 23.1% Blood-2) and to the effect of increased sample handling. When comparing the "best" blood sample to the buccal sample, heritability was twice as high for blood than buccal TRM. Nonetheless, computation of the genetic ($0.663 < r < 0.983$) and shared environmental ($0.678 < r < 0.895$) correlations suggest that the A and C factors influencing buccal and blood TRM largely overlap. In contrast, no significant overlap was seen for the unique environmental factors - while these explain the largest part

of the variance in all three TRM measures. This suggests that they may largely reflect idiosyncratic measurement error rather than a truly unique environmental effect.

In conclusion, chapter 2 provides support, albeit modest, for the use of buccal derived DNA in large scale biobank studies focused on the study of TRM, which will allow for less invasive longitudinal studies. It is clear that repeated handling is ill-advised for TRM for either buccal or blood samples. Finally, future telomere studies should employ new telomere measurement techniques such as the telomere shortest-length assay (TeSLA), which measures the telomere length of individual chromosomes, as it has been established that the shortest telomeres in a cell are responsible for triggering the aging-associated senescence [3].

OBESITY AND THE MICROBIOME

In chapter three, we utilized two different genetically informed designs to better understand the association between obesity and the gut microbiome. Current hypotheses put forth that the gut microbiota of obese individuals may be more efficient at extracting energy from food the host consumes [4] or that they interfere with vagal afferents to the brain controlling feeding behavior and satiety signals generated in the gut [5]. However, these causal explanations ignore possible genetic confounding, such that the same genetic variants that influence the gut microbiome also (independently) influence obesity. In that case, the high genetic risk for increased BMI could be associated with quantitative (smaller species diversity) and qualitative effects (enrichment for different species) on the gut microbiota without a direct causal effect of obesity on the microbiome. Also, BMI and related shifts in cardiometabolic and inflammatory profiles may itself be affecting the composition of the gut microbiome.

We addressed the BMI-microbiome relationship taking into account genetic confounding by two different strategies. First, we employed the "four-corners" design, an extension of the Recall-by-Genotype design [6]. In this design, participants were selected for extremes in both the polygenetic risk for BMI and actual observed BMI. This design again allowed us to address the question of genetic confounding in the obesity - microbiome association. If the genetic risk for BMI drives the microbiome composition independent of actual BMI, we would expect differences between the low and high genetic risk groups independent of BMI. If actual BMI drives the microbiome composition independent of genetic risk for BMI, we would expect microbiome differences in the lean and obese groups independent of genetic risk for BMI. In a second genetically informed study design we analyzed the gut microbiome of MZ twins discordant for BMI. The use of BMI discordant MZ pairs allows for a unique view of the gut microbiome's differences that may be associated with an environmentally driven obese state rather than variation in genetics.

We first focused on microbiota diversity which is operationalized by the alpha diversity metric. High microbiota diversity seems to be associated with health and temporal stability with loss of diversity prognostic of increased disease risk [7]. The key finding was the full replication of the previously observed negative relationship between actual BMI (or other obesity measures like body fat percentage) and alpha diversity after balancing for genetic risk. In contrast, the genetic risk for BMI was not associated with alpha diversity, after balancing for actual BMI. This result was partly confirmed by the MZ discordant twin analysis. Although we found no significant difference in alpha diversity between the BMI discordant twin pairs, the a priori statistical power was modest at best. The difference observed did go in the expected direction, with the heavier participants having a lower average alpha diversity.

The above pattern of results is compatible with a causal effect of obesity on the gut microbiome. However, obesity may come in different flavors. An unanticipated significant interaction between genetic risk and BMI showed a lower gut microbiome alpha diversity in the low genetic risk / high BMI group relative to all other observed groups, including the high genetic risk/ high BMI participants. This finding points to a separate sub-type of obesity that is more heavily influenced by environmental factors, and which may be more detrimental for the gut microbiome than obesity caused by genetic factors. For instance, this could reflect microbiome-effects associated with correlated risk factors present in lower socioeconomic status, including dietary composition, that may characterize the former group of obese persons.

In addition to comparisons between obesity-associated measures and alpha diversity, we explored whether there were specific OTUs significantly associated with either a leaner or heavier phenotype within the two separate study designs. Using the regression of BMI, body fat, and WHR on OTU abundances resulted in 14 OTUs significantly associated with one of these measures. Taking a closer look at the taxonomic classification of the OTUs showed that 10 of the 14 significant OTUs belong to the *Firmicutes* phylum, all related inversely to BMI. However, the relationship between BMI and microbes from the *Firmicutes* phylum is complicated as LEfSe analyses identified five OTUs from the *Firmicutes* phylum enriched in heavy individuals. This clearly demonstrates that various *Firmicutes* phylum members may have varying contributions to the obese phenotype, with some OTUs associated with a lean phenotype and others associated with a heavier phenotype.

Using triangulation across three analytic strategies (OLS, LEfSe, Random Forest), we identified 9 OTUs associated with either a leaner or heavier phenotype. These OTUs were classified down to various taxonomic ranks (e.g., order and family). OTUs belonging to the *Ruminococcaceae* and *Oxalobacteraceae* families were found to be enriched within leaner individuals and generally negatively associated with obesity measures. Enrichment for the *Ruminococcaceae* family members has been associated

with lower body fat before [8]. Microbes in the *Oxalobacteraceae* family also appear to be specifically responsive to antibiotics administration, a well-known causal determinant of reduced alpha diversity [9]. They are also enriched in individuals with no previous contact with the Western world [10, 11], who are documented to have higher alpha diversity.

In conclusion, chapter 3 provides support for a distinctly lower gut microbiota diversity in individuals with high BMI, particularly if they are low in the genetic susceptibility to obesity. Additionally, we identified a number of OTUs in the *Ruminococcaceae* and *Oxalobacteraceae* families that have a significant association with obesity-associated measures and their depletion may be a source of the lower microbiota diversity seen in more obese individuals.

GENETIC VARIANTS INFLUENCING THE MICROBIOME COMPOSITION

The differential microbiome composition in obese and non-obese participants detected in chapter 3, even when accounting for genetic confounding, only partly answers the question of whether this difference in the microbiome reflects a true causal effect of BMI. An often-employed strategy in recent epidemiology to detect the presence and direction of causal effects is Mendelian Randomization or MR [12]. MR uses selected genetic variants robustly predictive of an "exposure" variable as instrumental variables to test causal effects on an "outcome" variable. Because alleles are randomly transmitted from parents to offspring, and a future biological outcome cannot alter a person's allelic variants, MR suffers much less from confounding and reverse causality than conventional observational research. For MR to work, a genetic instrument must be available that is robustly predictive of the exposure. Typically, this is a genetic variant (e.g., a SNP) that was significantly associated with the exposure variable, even when correcting for the multiple testing inherent in GWAS. However, a single SNP is typically not sufficient for robust MR, as a series of sensitivity analyses need to be performed to test MR's critical assumption of no pleiotropy, i.e., that the genetic instrument is not associated directly with the outcome variable, other than through its potential causal effect on the exposure variable. This means that MR requires a large set of genome-wide significant SNPs for the exposure, which can typically only be derived from meta-analysis of many cohorts.

To perform a bidirectional test of causality between BMI and a target measure of microbiome composition, e.g., alpha diversity or abundance of a specific OTU, MR requires genetic instruments for both types of traits. This means that a set of genome-wide significant SNPs for BMI-as-the-exposure has to be available as well as a set of genome-wide significant SNPs for microbiome-as-the-exposure. When considering the

causal effect of BMI-as-exposure on the microbiome, we are in good shape because the GIANT consortium is producing increasingly large numbers of genome-wide SNPs for BMI [13-15]. When I started my thesis, the reverse was not true. There were no good genetic instruments for the microbiome-as-exposure available because the few GWA studies performed on gut microbiota were grossly underpowered. To my good fortune, the MiBioGen consortium, intent on providing a collaborative effort on genome-wide association analyses on the gut microbiome, had started to take shape, and I was among the group of analysts who could simultaneously contribute to and learn a lot from this unique collaboration. As a consortium, we recently produced the largest GWA on human gut microbiome composition too date. It identified 30 loci affecting ten microbiome taxa at a genome-wide significant ($P < 5x10^{-8}$) threshold, with one locus, the lactase (LCT) gene region, remaining study-wide significant after additional correcting for the number of taxa tested (GWAS signal $P = 8.6x10^{-21}$). Other associations did not meet this stern criterion and were suggestive only ($1.94x10^{-10} < P_A < 5x10^{-8}$) but enriched for genes expressed in the intestine and brain. Phenome-wide association and MR analyses indicated food preferences and diseases as mediators of the genetic effects.

Chapter 4 describes results generated during the work needed to contribute to this GWA meta-analysis by the international consortium. While our main aim was to provide the NTR results for the meta-analysis we in parallel tested a few specific hypotheses using the genome-wide data in the NTR sample. In view of the potential for MR based analysis to address causality in e.g., the microbiome-BMI association, we first wanted to confirm that the meta-analytic results would be sufficiently applicable to our own NTR cohort. Two findings somewhat increased our confidence in this idea. The consortium identified the association between the *Bifidobacterium* genus and rs182549 as the most significant mbQTL ($p = 8.6332^{-21}$) and the NTR-specific results for this mbQTL showed a p-value of 0.06, only just failing to achieve nominal significance. A more convincing resemblance between the NTR cohort and the consortium findings was found with regards to the link between heritability of a taxon and its ability to generate significant associations. Taxa defined at the family and genus levels that had the highest MZ twin correlations (>0.30) showed much lower p-values in their associations with genome-wide SNPs than less heritable taxa.

Taken together, it seems appropriate to use the results of the MiBioGen consortium for extended analyses in smaller cohorts, such as performing MR with genetic instruments for the more heritable microbiota as predictors of inflammatory and cardiometabolic (including obesity) outcomes.

COHABITATION AND THE MICROBIOME

Based on evidence for the modifiability of the microbiome in animal studies, which show strong effects of changes in maternal fostering, housing conditions, or other external exposures, numerous studies on the human microbiome have focused on the role of environmental factors on the human gut microbiome. One of the most easily recognized factors capable of modifying the gut microbiome composition is dietary intake. Human diets consist of numerous macronutrients that are processed by the host and the microbiome constituents. It has been well-established that the long-term dietary intake is a considerable factor in determining the gut microbiota composition [16-19]. Similarly, it has been demonstrated that the gut microbiome is capable of responding to short-term dietary changes with changes in the overall composition [20]. Individuals exposed to diets consisting of entirely animal-based products were compared to those exposed to an entirely plant-derived diet, demonstrating unique taxonomic shifts associated with both exposure groups [20]. The shift in microbes inhabiting the gut microbiome reflected changes in the functional gene repertoire of the communities, equipping the microbiome to more efficiently process the available macronutrients.

In addition to macronutrient influx into the gut microbiome system, additional consumption-associated behaviors such as smoking have been shown to alter the gut microbiome composition. Smoking tobacco introduces a milieu of compounds that not only expose the airway but are also found widely in circulation. Smoking has indeed demonstrated an ability to induce differences in the gut microbiome. Again, some studies have shown similar trends with regard to the effect of smoking on the gut microbiome composition, such as an increase in *Bacteroidetes* and a decrease in *Firmicutes* [21-23], while also showing differences such as differing direction of enrichment for the *Proteobacteria* phylum. This lack of clear pattern across studies associated with smoking induced shifts in the gut microbiota, hints at additional confounding factors such as host genetic profile. Many current studies are aiming to untangle the complex web of host-genotype, environmental exposure (smoking), gut microbiome composition, and subsequent disease pathology.

Just as the association between smoking, the gut microbiome, and host intricacies presents a complex web of interactions, the effects of exercise on the gut microbiome composition are still beginning to be understood. Animal-based studies have demonstrated increases in butyrate production by gut microbiome taxa in response to exercise [24, 25]. Numerous studies have identified an ability for the exercise activity to alter the taxonomic composition of the gut microbiome [26, 27]. These studies showing the effects of exercise on the gut microbiome composition extend to both human and animal studies. However, specific patterns in gut microbiome alterations have been harder to pinpoint.

A key feature of the above health behaviors (diet, smoking, and regular exercise habits) is that they are often shared by members of the same household and even the broader neighborhood environment. Numerous studies have identified that aspects of the gut microbiome are influenced by cohabitation amongst individuals [28-31] and this influence extends to individuals within a shared social network, in the absence of cohabitation [28, 32]. Because members of a shared household also share part of their genome and genomic correlation even extends to the wider neighborhood [33], again a genetically informative study design was needed to separate environmental from genetic effects.

Chapter 5 used such a design. Samples collected from 42 spouse pairs and 166 MZ twin pairs with varying cohabitation status were subjected to 16s rRNA sequencing and analyzed concerning the overlap of microbiome features. These spousal and twin comparisons showed that cohabitation is an important factor influencing the composition of the gut microbiome. Stronger correlations in the gut microbiome alpha diversity were observed between the 45 cohabitating twin pairs ($r = 0.64$) relative to the 121 non-cohabitating MZ twin pairs ($r = 0.42$). Although the spouse pair correlation was insignificant at a predefined alpha ($r = 0.23$), it was larger than the true zero observed in unrelated individuals ($r = -0.015$). Beta diversity analysis using Bray-Curtis (BC) dissimilarity metrics showed cohabitating MZ twin pairs had the most similar gut microbiota communities, which were more similar than the BC values of non-cohabitating MZ twins (empirical p-value = 0.0103), cohabitating spouses (empirical p-value = 0.0194), and pairs of unrelated non-cohabitating individuals (empirical p-value < 0.00001). There was also a significant difference between the BC measures from the spouse pairs and those from the unrelated non-cohabitating individuals (empirical p-value < 0.00001). Taken together, these findings suggest a strong role for environmental influences on the diversity of the gut microbiota composition. This is in line with the relative importance of environmental factors highlighted by many previous works [32, 34].

Regarding individual microbes, the work in chapter five identified two species-level OTUs (Otu0081 and Otu0190) that were significantly shared between cohabitating MZ twin pairs and spouse pairs but not between non-cohabitating MZ twins. The lack of sharing between the non-cohabitating MZ twins rules out genetic influences on these microbes. OTU190 and OTU0081 were given consensus classifications as *Ruminococcaceae* and *Clostridiales* respectively. Further taxonomic characterization indicated OTU81 to be an uncultured *Oscillibacter* species, which resides within the *Clostridales* order. OTU190 gave ambiguous results, and the best taxonomic classification remained at order level, i.e., *Ruminococcaceae*. Organisms belonging to the *Ruminococcaceae* family have been shown in previous studies to be impacted by both high-fat diet and exercise activity. These are therefore obvious components that may account for the found shared household effects on these microbes.

In order to test the clinical relevance of shared household factors on the gut microbiome, cardiometabolic and immune-associated disease burden scores were calculated for all study participants using extensive data obtained during the NTR Biobank project [35]. We first tested whether alpha diversity was associated with either inflammatory or cardiometabolic burden profiles, which was not found to be true. Next, the disease burden scores were regressed on the OTU counts of the two species-level OTUs shared due to cohabitation. One of these OTUs (OTU190) was significantly associated with lower inflammatory and cardiometabolic disease burden. Whereas this demonstrates proof-of-principle for an effect of the shared household on health through the microbiome, it is unlikely that such effects hinge on a single microbe. More likely we were underpowered to detect many more such effects. Future studies that similarly link environmentally driven differences in the gut microbiome and aspects of human health continue to be necessary.

GENERAL DISCUSSION

Given the emphasis in this thesis on the genetic and environmental factors shaping the individual differences in the human gut microbiome obtained from stool samples, the discussion below also uses this as the main theme. I will first review the current state-of-the-art with regard to the nature and nurture of microbiome variation. Next, I will turn to the link between the microbiome and health: with the current knowledge can we define microbiome traits that constitute a risk factor for disease, or at least can be considered solid biomarkers? In closing, I discuss methodological developments in microbiome research and some of its promising future directions.

NATURE VERSUS NURTURE

On the whole, the question of the contribution of host genetics and numerous environmental influences on the diversity of the gut microbiome composition is one that has captivated the research community. What became readily apparent is that the gut microbiome composition does tend to show associations with a large set of environmental factors. The literature is inundated with a wealth of manuscripts that successfully identify the ability of the environmental factors such as antibiotics, diet, smoking, exercise, and countless others, to cause changes in the gut microbiome [21, 22, 36-50]. When looking at alpha diversity as the main outcome, our twin studies also make a strong case for the influence of environmental factors, including those accounted for by a shared household, on diversity. In contrast, the genetic contribution to alpha diversity was nearly absent. This appears to draw a consensus with the rest of the microbiome field, which does not find robust evidence for genetic control of gut microbiome alpha diversity [51]. It was also the environmentally driven sub-type of

obesity that appeared to be more detrimental for the gut microbiome diversity than the sub-type of obesity caused by genetic factors. In short, when it comes to diversity of our gut microbiome, nurture trumps nature.

With regard to genetic and environmental control over individual taxa, again it appears that environmental factors play a larger role than genetic factors [8] when we consider the full spectrum of taxa. Even so, while environmental influences more commonly shape most of the gut microbiome, research did highlight strong relationships between specific taxa and the genetic profile of the human host. Both GWAS [52, 53] and twin studies [54, 55] have demonstrated significant relationships between individual taxa and the host genome. One genomic region that has appeared multiple times associated with microbiome constituents is associated with the LCT region, which produces the well-known lactase gene [51, 55]. Not only does the genomic region show up continuously, but it also appears to be associated with somewhat similar microbes, namely those belonging to the *Bifidobacterium* genus [54, 55].

When considering the above summary, it is important to note that universal questions of the relative role of nature and nurture and their interplay that readily apply to other traits, do not similarly apply to the microbiome. The microbiome is special. Consider what is exactly meant by the research community when a particular microbe is referred to as "heritable". In the traditional sense, a trait would be considered heritable if the genetic makeup of an individual results in a difference in that particular trait. But what does it mean for a microbiome-associated constituent to be heritable? While some debate the presence of a very small amount of biomass present in utero, we are largely devoid of our resident microbial communities upon birth [56]. In short, we are conceived without our microbes. Surely it is not possible for the fetal host DNA to produce these microorganisms, by means of biochemical activity, from absolutely nothing.

What we actually mean by heritability is that, once a particular microbe is encountered, the likelihood of that microbe to become a residing member of the microbiome ecosystem is due to genetic factors. One can imagine a situation where a particular host genomic landscape is ripe for colonization by a microbe of interest. It may be that this host environment has the perfect combination of metabolic substrates, friendly microbial neighbors to coexist with, a plethora of host receptors to interact with and reside upon, and a non-hostile immune system that has developed in a manner that would readily welcome colonization. Now imagine that for whatever reason, the host that contains this fertile landscape never encounters this microbe. Would it then be correct to say that this microbe is not heritable, although the conditions were particularly ideal for colonization if given the chance?

No. If the host provides, by means of its genetic architecture, conditions amenable to colonization by a particular microbe upon encounter, the microbe can be considered to be heritable. On the other hand, we must concede that it would be quite difficult to think of an organism to demonstrate large heritability estimates regardless of environmental conditions. If an individual has a particular genetic makeup it is only guaranteed to have a particular microbe abundance when the environmental conditions allow it, i.e., when some initial exposure to the microbe occurred. Significant heritability estimates or genetic associations are feasible only if (1) the majority of hosts encounter the microbe at some point, allowing for colonization to virtually always occur if the proper genetic repertoire is available, or (2) the genetic repertoire of the host modifies the host behavior in a way that influences exposure to the microbe across individuals and thus the chance for colonization.

These complexities in interpreting the nature and nurture interplay are further compounded by a potentially huge contribution of stochastic processes to the gut microbiome. While the gut microbiome composition is impacted by genetic and non-genetic host factors, some studies have estimated that each of these factors only account for approximately 10% of the microbiome composition [57]. The other 80% may be attributable to sheer coincidence, such as the first microbe to colonize a particular microbiome ecosystem. Such a first microbe can have large and exponential effects on the subsequent community that develops, which is in line with ecological theory [58].

To understand the relationship between the genetic and environmental forces that dictate the gut microbiome composition, it would be prudent to add the ecological and evolutionary perspectives. The human host is an environmental landscape, inhabited by numerous organisms from various domains of life, competing with one another to reside within the ecosystem and make use of the available nutrients [59, 60]. The actions of the organisms residing within this ecosystem are capable of impacting both the environment, in this case the human host, as well as the other organisms residing within the environment in an ecological manner [59, 60]. The gut microbiome represents a dynamic metabolic biomass, the specific metabolic functionality of which is comprised of the cumulative metabolic capabilities of the microbiome associated organisms, encoded in their genomes. The influx of nutrients into the environment through human consumption provide the nutrients necessary for their growth and subsequent colonization. The microbes compete for these resources through efficient utilization of the nutrients available as well as direct targeting of competing organisms [60]. The microbes that win out through this fierce competition and accumulation of numerous random processes, constitute the gut microbiome composition. Organisms that can most readily utilize the nutrient sources available within the ecosystem and fend off targeting from the host and competing organisms, will be the ones that thrive and propagate.

Let's reconsider how the genetic repertoire of the host modifies the host behavior in a way that influences the chance for colonization of the gut by certain microbes from an ecological perspective. For the most supported environmental influence of the gut microbiome, dietary preference, current evidence points to a causal chain of genetic background → dietary preference → microbiome composition [51]. The genetic makeup of an individual predisposes them to a particular dietary preference, which in turn leads to the formation of a community of microorganisms capable of processing the specific milieu of dietary-associated molecules present. The microbial community that is formed would be based off numerous factors including but not limited to which microbes were readily available from the environment (and their subsequent metabolic activities encoded in their genomes), the nutrients available for processing by the microbiome constituents, and the receptiveness of the local immune system to colonization. In this manner the host genetic architecture and the numerous factors modulated by genetics, in concert with environmental influences, and stochastic processes, manifest in a resident gut microbiome composition that has achieved this ecosystem through competition and evolution towards the optimal utilization of available resources and overall survival within the host.

While the similarities of the formation of the gut microbiome composition reflects ecological processes, it should not be forgotten that although the human host provides the habitat for the microbiota, the microbes and the human host are ultimately separate organisms, all vying for optimal fitness, sometimes at the expense of each other. While the gut microbiota is capable of producing many positive effects for the human host, the community can also exhibit negative effects on the host. Because of the ability of the gut microbiome to elicit both positive and negative effects on host health, it is necessary for the host to attempt to exert control over the gut microbiome. Human hosts attempt to exert control over their gut microbiome composition by means of immune surveillance, antimicrobial peptide production, and even production of specific compounds that may increase the odds of successful colonization by non-pathogenic organisms, among many other mechanisms [60]. Similarly, the microbes within the microbiome are acting in a manner that serves to increase their fitness in the environment, regardless of the effect on the human host. This competition between the gut microbiome organisms and the host organism results in an evolutionary process that can result in both mutualistic interactions as well as interactions that benefit one organism at the expense of others.

The main point to extract from these examples is that host genetics, host environmental factors, the gut microbiome composition and host health are deeply intertwined. This explains why the relationship between the gut microbiome and health will almost by necessity prove to be highly complex.

HUMAN GUT MICROBIOME AND HEALTH

In a manner similar to the hunt for understanding the relationship between how genetic and environmental factors shape the gut microbiome, the search for a relationship between the gut microbiome and host health has shown great initial promise. A large number of observational studies report differences in the gut microbiome of healthy individuals relative to an extremely wide number of disease states [61, 62]. One of the findings that does seem to hold, is the idea that higher diversity in the gut ecosystem is generally associated with a healthier state of the host [7]. As discussed, this theme is commonly observed in the case of BMI and was also confirmed in chapter 3. However, detecting consistent patterns between the microbiome and health, or demonstrating causality beyond-reasonable-doubt, have proven to be largely elusive [51, 55, 61].

The strongest evidence regarding the ability of the gut microbiome to play a role in disease development does not come from observational studies, but from directly testing the ability of the gut microbiota to induce disease states upon transplantation to animal recipients [4, 63, 64]. These experiments provide powerful evidence that something within the transplanted samples, either microbe, metabolite, or otherwise, is capable of impacting host physiology in a manner that results in a disease state.

One of the most prominent mechanisms identified through which the microbiota influences human health, is by producing biologically active metabolites. This metabolic activity could be in the form of production or degradation of biologically important compounds [65-67]. Not only can the microbiome produce numerous biologically active compounds that impact the local gastrointestinal environment, but these microbiome-derived compounds have also been implicated in the modulation of behavior and brain function, through microbiome metabolite action within areas of the brain. Using a murine model, Chu et al. demonstrated that the presence of the gut microbiota derived metabolites during a critical developmental period was essential for the development of normal extinction learning [68]. Additionally, microbiome-derived metabolites such as short-chain fatty acids (SCFA) and trimethylamine N-oxide (TMAO) have both shown interesting links to gut-brain communication and heart disease respectively [65, 69]. These studies make it clear that gut microbiome derived metabolites can influence a number of human associated phenotypes by exerting an effect in local and distal regions.

As discussed, there are numerous forces at work forming the gut microbiome ecosystem through evolutionary processes, with the gut microbiome evolving to adapt to available nutrients and host pressures [60]. With the microbial organisms and the host organism all striving for superiority, at times this results in winners and losers, and sometimes the loser is the host. One of the most obvious microbiome-

associated disease states which results in negative effects on the host is *Clostridium difficile* infection, where a shift in the gut microbiome, usually after administration of antibiotics, is associated with significant *Clostridium difficile* colonization. *Clostridium difficile* infection manifests in a gastrointestinal disease state in the host [70]. Treatment of this disease in humans has seen relative success with the administration of a fecal microbiota transfer (FMT) from healthy donors to reestablish a "healthy" gut microbiome and ameliorate the disease state [62]. In addition to the introduction of actual microbial communities in the case of FMT, recent studies have demonstrated that administering a sterile filtrate from the FMT donor sample to *Clostridium difficile* infection patients was capable of positively impacting the effects of *Clostridium difficile* infection. This provides evidence that microbiome-derived metabolites are associated with the microbiomes effects on health [71].

While the case of *Clostridium difficile* provides an elegant example of the ability of microbiome associated elements to play a significant role in the pathogenesis of human health, many investigations into the role of the gut microbiome in various disease states have not been so clear cut. Fecal transplants have had mixed success with alternative diseases believed to be associated to the gut microbiome [72, 73]. Similarly, probiotic therapies have demonstrated mixed success across studies and populations. Many probiotic studies show initial presence of the species of interest only for it to be washed out shortly after stopping treatment with the probiotic [61]. With both FMT and probiotic therapy, care should be given to consider how the newly introduced microbes will fare within the gut microbiome ecosystem, and how the resident community reacts to the introduction of the organism. One study demonstrated that introduction of a probiotic species, known to be a core microbiome-associated species, was able to achieve stable long-term persistence in the gut microbiome of 30% of individuals [74]. This demonstrates that introduction of new species to the gut microbiome may have a higher success rate if organisms are chosen based on a priori knowledge of their suitability within the ecosystem. This study also concluded that addition of species to the ecosystem was easier if species functionally related to the *Bifidobacterium longum* AH1206, were absent prior to administering the probiotic, which would presumably leave an open niche in the community for colonization by the microbe of interest.

When looking at the success or failure of studies aimed at inducing a "healthy" gut microbiome, we must consider that a "healthy" microbiome may look different for different individuals, based on their genetic profile, environment and history of stochastic microbiome associated effects such as which organism arrived first. Additionally, introduction of "healthy" organisms to microbiome ecosystems may have varying success for different individuals. If we further examine the case of observing a single phenotype or disease trait, which is believed to be dependent on either the microbial production or degradation of a particular biologically active molecule,

it would theoretically be possible for two different microbes, containing the same pertinent biochemical machinery necessary to complete a biochemical task, to be functionally interchangeable. In other words, when observing the functional capability of these two bacteria with regard to the production or degradation of the molecule of interest, it is possible that these two organisms are functionally identical. While two organisms may be functionally interchangeable with regard to one particular phenotype, these two organisms can still influence additional host phenotypes in other unrelated ways. Another complexity associated with this system is that different combinations of microorganisms may cumulatively have the same metabolic capability as a singular microorganism. This functional redundancy amongst the genomic repertoires of the microbiome-associated organisms may be the reason it has been difficult to ascertain the taxonomic identity of individual members of the microbiome capable of influencing a host phenotype.

A major driver of the desire to identify the gut microbes that are most heritable and the genetic variants that are driving this heritability, also reflected in this thesis, is that such variants allow MR techniques to explicitly test causal effects of the microbiome on human health. This approach has already begun to show promise for parsing out causality regarding the gut microbiome and human disease. For example, a recent study from Qin et al. [55] leveraged a single large-scale population-based cohort of 5,959 genotyped individuals with matched gut microbial shotgun metagenomes, dietary information and health records up to 16 years post-sampling, to characterize human genetic variations associated with microbial abundances, and predict possible causal links with various diseases using MR. This successfully identified individual SNPs, particularly in the LCT, ABO and MED13L regions that were significantly associated with microbial abundances in the gut microbiome. More interestingly, they demonstrated how different combinations of genetic and environmental influences, in the form of dietary input, resulted in compositional changes in the gut microbiome. Individuals with persistent production of lactase throughout adulthood (rs4988235:TT) showed no difference in abundances of *Bifidobacterium* in the gut microbiome regardless of dairy intake. Alternatively, lactose-intolerant individuals (rs4988235:CC) who reported regular dairy consumption showed an increase in *Bifidobacterium* relative to lactose-intolerant individuals with lower dairy consumption. It can be speculated that in individuals without lactase production, the available lactose becomes a prevalent dietary substance that is exposed to the gut microbiome constituents for possible metabolism. Exposure of the gut microbiome to this energy source may cause a shift in the microbiome to allow for efficient utilization of the nutrients available to the community. Likewise, secondary metabolites from lactose metabolism would possibly be capable of inducing shifts in the gut microbiome ecosystem as well.

This elegant finding demonstrates how the host genotype may provide an innate metabolic capability which influences the mixture of numerous food-derived

substrates that are made available to the gut microbiome for further metabolism. Variation in this innate metabolic capability encoded through the host genome, coupled with environmental exposure to a particular diet can induce a pressure on the gut microbiome ecosystem that results in taxonomic and functional shifts within the gut microbiome. This example provides a simple model that attempts to explain the complex connection of environmental factors, host genetics, the gut microbiome, and human health. In the same way that this complex web of interactions between host metabolic capability, bacterial metabolic capability, and substrate availability exists for the handling of lactose, similar interactions and systems likely exist for the vast number of molecules we are exposed to through diet, medication or otherwise.

Currently approaches as demonstrated in Qin et al. are beginning to attempt to disentangle this web of interactions starting with individual taxa, loci and metabolites. In the future, given appropriate data sets, machine learning, and neural network algorithms could possibly aid in tackling these problems while considering the multitude of possible complex interactions between the forces host genetics, environmental factors, the gut microbiome composition, and human health.

FUTURE DIRECTIONS

Throughout the period in which this thesis work was done, several exciting methodological developments took place in the fast-moving field of microbiome research. One of the currently debated topics in microbiome research is the most appropriate manner to define clusters of sequence reads that most accurately represent the correct taxonomic classification of the bacteria to which the reads belong. This process of binning sequences for taxonomic classification is known as operational taxonomic unit (OTU) picking. In the most simplistic form, closed reference OTU picking strategies cluster an unknown sequence against a collection of reference sequences with known taxonomy to determine the reference sequence with the highest degree of similarity to the unknown sequence. Reads that do not meet the predefined similarity cut off to any reference sequences are discarded from further analysis. Similarly, direct classification of reads using tools such as the RDP classifier is related to the closed reference approach through a naive Bayesian classifier directly classifying sequence reads, rather than clustering methods [75]. An improved version of this strategy, open reference OTU picking, involves the same initial clustering against a collection of reference sequences but deviates in that sequences that do not sufficiently match any sequence within the sequence collection are not thrown away but rather clustered using a de novo clustering approach. The de novo clustering process aims to cluster sequences against one another without an external database of reference sequences. The basis of the resulting clusters of sequences is sequence similarity alone. Although the taxonomic classification of these de novo OTUs is more

complicated, this approach allows for identifying and studying novel OTUs that may have yet to be taxonomically elucidated. These data would be thrown away otherwise. The final method is the use of purely de novo OTU clustering for all sequence reads. This process aims to cluster all reads against each other, and form clusters based on a predefined similarity cutoff. After the formation of de novo OTUs, the classification of the OTUs can be obtained by finding consensus amongst the individual classifications of the unique reads within the OTU.

One of the main criticisms of reference-based clustering strategies is the reliance on similarity to the database sequences without regard to the sequence similarity of the rest of the non-reference sequence reads binned within the same OTU. This can result in OTUs where the reads all share a sufficient degree of similarity with the reference sequence to be included in the OTU, but the similarity of the non-reference sequences may be below the defined similarity cut off. Similarly, a sequence read may have a tie regarding the best database match, resulting in the classification of the read belonging to whichever taxon is encountered first during the database matching process. Additionally, the resulting OTU table contains OTUs of two different "types". The first type contains OTUs formed by similarity to the reference database and the second type contains OTUs formed based off of sequence similarity alone. Using de novo clustering methods deals with this problem by first forming OTUs based on sequence similarity alone before attempting to attribute taxonomic classification. De novo clustering methods can have drawbacks as well. The generation and subsequent clustering of a distance matrix containing all unique, quality-controlled sequences within a data set necessitates considerable computational resources. Furthermore, the resulting taxonomic classification can leave room for desire due to the need to obtain a consensus classification from the individual classifications of the reads contained within an OTU.

The optimal OTU picking strategy to use in microbiome-associated studies is still an open debate, with large proportions of the research community using different approaches. Given the advantages and disadvantages of de novo and closed reference based OTU picking strategies, our research group gravitates towards the de novo methods, ensuring a minimum similarity amongst all the reads within the various OTUs formed [71]. Even so, we remain practical. As demonstrated by the MiBioGen consortium, the nature of the data available may necessitate one strategy over another. Due to the significant heterogeneity of molecular methods used to interrogate the 16S gene region of bacterial genomes within the MiBioGen consortium, a closed-reference strategy utilizing the RDP database had to be used to exploit the huge advantage of sharing data across many cohorts worldwide.

Although 16s rRNA gene sequencing has been critical for advancing the microbiome field, shotgun sequencing of all non-human DNA can provide much richer information

regarding the multitudes of Prokaryotes, Eukaryotes, Archaea, and viruses present within the microbiome. In addition to taxonomic information, shotgun sequencing gives information on the entire genomic repertoire of all microorganisms present within a particular microbiome ecosystem. This information is essential for understanding the biochemical capacity of the ecosystem and the metabolome it can produce.

Complementing shotgun sequencing, metatranscriptomic studies will help determine which of these non-human genes and their subsequent biochemical pathways are being expressed in the form of RNA transcripts. Finally, coupling with metabolomic studies will help determine the presence of the actual set of microbially- derived metabolites that these transcripts give rise to. Studies coupling metatranscriptomics with metabolomics will help to determine how the individual activities of the microbes, represented by the active transcriptional pathways and the profiling of the subsequently generated metabolites, will allow researchers to view the functional characteristics of the gut microbiome which will prove more informative than just using the genomic content to establish the taxonomic identity of the microbes present.

Clearly, combining the many sources of 'omics' information faces many of the similar challenges encountered by researchers of other traits, and these datasets will be even more high-dimensional when generated for each of the many gut microbiome constituents. Thus, statistical comparisons of individual microbiome-associated variables present a substantial multiple-testing burden due to a large number of resulting comparisons. Accordingly, adopting the approach of the various genomics consortia, namely cooperation, and consolidation of data, will be necessary to achieve sufficient statistical power to accurately represent the effects of the numerous of the microbiome constituents on various host-associated phenotypes. To this end, team science and collaborations in international consortia, as illustrated by the MiBioGen consortium in chapter 4, will be increasingly important in the microbiome field.

While large consortia such as MiBioGen will be increasingly important, applying combined metagenomics, metatranscriptomics, and metabolomics across many different samples to study the gut microbiome may face inherent difficulties. The larger MiBioGen analysis demonstrated that there were only 9 out of 410 genera that were present in 95% of samples. It is possible that this lack of overlap in gut microbiome organisms may be reflective of a lack of environmental exposure to the organisms, not necessarily a lack of heritable components influencing colonization success. As discussed previously, an individual may have the ideal landscape for colonization by a particular microbe, but never encounter the microbe. This demonstrates an increased need for family and twin studies, to aid in the understanding of the role of genetics and environment in shaping the gut microbiome. While individuals that live in geographical locations separated by great distances may have a higher likelihood of being exposed to different environmental communities of microbes, family units would

have a much higher likelihood of being exposed to similar environmental sources of microbes. Given the increased likelihood of exposure to similar microbes, as well as varying degrees of genetic similarity amongst family members, this presents an ideal situation to parse the genetic and environmental effects shaping the gut microbiome.

REFERENCES

1. Visscher, P.M., et al., *10 years of GWAS discovery: biology, function, and translation.* The American Journal of Human Genetics, 2017. **101**(1): p. 5-22.
2. Green, E.D., J.D. Watson, and F.S. Collins, *Human Genome Project: Twenty-five years of big biology.* Nature, 2015. **526**(7571): p. 29-31.
3. Shay, J.W. and W.E. Wright, *Telomeres and telomerase: three decades of progress.* Nature Reviews Genetics, 2019. **20**(5): p. 299-309.
4. Turnbaugh, P.J., et al., *An obesity-associated gut microbiome with increased capacity for energy harvest.* nature, 2006. **444**(7122): p. 1027.
5. Cawthon, C.R. and B. Claire, *Gut bacteria interaction with vagal afferents.* Brain research, 2018. **1693**: p. 134-139.
6. Corbin, L.J., et al., *Formalising recall by genotype as an efficient approach to detailed phenotyping and causal inference.* Nature communications, 2018. **9**(1): p. 1-11.
7. Lloyd-Price, J., G. Abu-Ali, and C. Huttenhower, *The healthy human microbiome.* Genome medicine, 2016. **8**(1): p. 1-11.
8. Beaumont, M., et al., *Heritable components of the human fecal microbiome are associated with visceral fat.* Genome biology, 2016. **17**(1): p. 1-19.
9. Raju, S.C., et al., *Antimicrobial drug use in the first decade of life influences saliva microbiota diversity and composition.* Microbiome, 2020. **8**(1): p. 1-13.
10. Raymond, F., et al., *The initial state of the human gut microbiome determines its reshaping by antibiotics.* The ISME journal, 2016. **10**(3): p. 707-720.
11. Torok, V.A., et al., *Influence of antimicrobial feed additives on broiler commensal posthatch gut microbiota development and performance.* Applied and environmental microbiology, 2011. **77**(10): p. 3380-3390.
12. Lawlor, D.A., et al., *Mendelian randomization: using genes as instruments for making causal inferences in epidemiology.* Statistics in medicine, 2008. **27**(8): p. 1133-1163.
13. Locke, A.E., et al., *Genetic studies of body mass index yield new insights for obesity biology.* Nature, 2015. **518**(7538): p. 197-206.
14. Speliotes, E.K., et al., *Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index.* Nature genetics, 2010. **42**(11): p. 937-948.
15. Graff, M., et al., *Genome-wide physical activity interactions in adiposity—A meta-analysis of 200,452 adults.* PLoS genetics, 2017. **13**(4): p. e1006528.
16. Ley, R.E., et al., *Human gut microbes associated with obesity.* nature, 2006. **444**(7122): p. 1022-1023.
17. Muegge, B.D., et al., *Diet drives convergence in gut microbiome functions across mammalian phylogeny and within humans.* Science, 2011. **332**(6032): p. 970-974.
18. Walker, A.W., et al., *Dominant and diet-responsive groups of bacteria within the human colonic microbiota.* The ISME journal, 2011. **5**(2): p. 220-230.
19. Wu, G.D., et al., *Linking long-term dietary patterns with gut microbial enterotypes.* Science, 2011. **334**(6052): p. 105-108.
20. David, L.A., et al., *Diet rapidly and reproducibly alters the human gut microbiome.* Nature, 2014. **505**(7484): p. 559-563.

21. Biedermann, L., et al., *Smoking cessation alters intestinal microbiota: insights from quantitative investigations on human fecal samples using FISH.* Inflammatory bowel diseases, 2014. **20**(9): p. 1496-1501.
22. Biedermann, L., et al., *Smoking cessation induces profound changes in the composition of the intestinal microbiota in humans.* PloS one, 2013. **8**(3): p. e59260.
23. Lee, S.H., et al., *Association between cigarette smoking status and composition of gut microbiota: population-based cross-sectional study.* Journal of clinical medicine, 2018. **7**(9): p. 282.
24. Evans, C.C., et al., *Exercise prevents weight gain and alters the gut microbiota in a mouse model of high fat diet-induced obesity.* PloS one, 2014. **9**(3): p. e92193.
25. Matsumoto, M., et al., *Voluntary running exercise alters microbiota composition and increases n-butyrate concentration in the rat cecum.* Bioscience, biotechnology, and biochemistry, 2008. **72**(2): p. 572-576.
26. Ortiz-Alvarez, L., H. Xu, and B. Martinez-Tellez, *Influence of Exercise on the Human Gut Microbiota of Healthy Adults: A Systematic Review.* Clinical and Translational Gastroenterology, 2020. **11**(2).
27. Campbell, S.C. and P.J. Wisniewski, *Exercise is a novel promoter of intestinal health and microbial diversity.* Exercise and sport sciences reviews, 2017. **45**(1): p. 41-47.
28. Brito, I.L., et al., *Transmission of human-associated microbiota along family and social networks.* Nature microbiology, 2019. **4**(6): p. 964-971.
29. Dill-McFarland, K.A., et al., *Close social relationships correlate with human gut microbiota composition.* Scientific reports, 2019. **9**(1): p. 1-10.
30. Finnicum, C.T., et al., *Cohabitation is associated with a greater resemblance in gut microbiota which can impact cardiometabolic and inflammatory risk.* BMC microbiology, 2019. **19**(1): p. 1-10.
31. Song, S.J., et al., *Cohabiting family members share microbiota with one another and with their dogs.* elife, 2013. **2**: p. e00458.
32. Rothschild, D., et al., *Environment dominates over host genetics in shaping human gut microbiota.* Nature, 2018. **555**(7695): p. 210-215.
33. Abdellaoui, A., et al., *Genetic correlates of social stratification in Great Britain.* Nature human behaviour, 2019. **3**(12): p. 1332-1342.
34. Falony, G., et al., *Population-level analysis of gut microbiome variation.* Science, 2016. **352**(6285): p. 560-564.
35. Willemsen, G., et al., *The Netherlands Twin Register biobank: a resource for genetic epidemiological studies.* Twin Research and Human Genetics, 2010. **13**(3): p. 231-245.
36. Matamoros, S., et al., *Development of intestinal microbiota in infants and its impact on health.* Trends in microbiology, 2013. **21**(4): p. 167-173.
37. Koenig, J.E., et al., *Succession of microbial consortia in the developing infant gut microbiome.* Proceedings of the National Academy of Sciences, 2011. **108**(Supplement 1): p. 4578-4585.
38. van den Elsen, L.W., et al., *Shaping the gut microbiota by breastfeeding: the gateway to allergy prevention?* Frontiers in pediatrics, 2019. **7**: p. 47.
39. Cioffi, C.C., et al., *History of breastfeeding but not mode of delivery shapes the gut microbiome in childhood.* PloS one, 2020. **15**(7): p. e0235223.

40. Flaherman, V.J., et al., *The effect of early limited formula on breastfeeding, readmission, and intestinal microbiota: a randomized clinical trial.* The Journal of pediatrics, 2018. **196**: p. 84-90. e1.
41. Azad, M.B., et al., *Gut microbiota of healthy Canadian infants: profiles by mode of delivery and infant diet at 4 months.* Cmaj, 2013. **185**(5): p. 385-394.
42. Jakobsson, H.E., et al., *Decreased gut microbiota diversity, delayed Bacteroidetes colonisation and reduced Th1 responses in infants delivered by caesarean section.* Gut, 2014. **63**(4): p. 559-566.
43. Ho, N.T., et al., *Meta-analysis of effects of exclusive breastfeeding on infant gut microbiota across populations.* Nature communications, 2018. **9**(1): p. 1-13.
44. Forbes, J.D., et al., *Association of exposure to formula in the hospital and subsequent infant feeding practices with gut microbiota and risk of overweight in the first year of life.* JAMA pediatrics, 2018. **172**(7): p. e181161-e181161.
45. Borewicz, K., et al., *The effect of prebiotic fortified infant formulas on microbiota composition and dynamics in early life.* Scientific reports, 2019. **9**(1): p. 1-13.
46. McFarland, L.V., C.T. Evans, and E.J. Goldstein, *Strain-specificity and disease-specificity of probiotic efficacy: a systematic review and meta-analysis.* Frontiers in medicine, 2018. **5**: p. 124.
47. Esaiassen, E., et al., *Effects of probiotic supplementation on the gut microbiota and antibiotic resistome development in preterm infants.* Frontiers in pediatrics, 2018. **6**: p. 347.
48. Langdon, A., N. Crook, and G. Dantas, *The effects of antibiotics on the microbiome throughout development and alternative approaches for therapeutic modulation.* Genome medicine, 2016. **8**(1): p. 39.
49. Hermansson, H., et al., *Breast milk microbiota is shaped by mode of delivery and intrapartum antibiotic exposure.* Frontiers in nutrition, 2019. **6**: p. 4.
50. Mueller, N.T., et al., *The infant microbiome development: mom matters.* Trends in molecular medicine, 2015. **21**(2): p. 109-117.
51. Kurilshikov, A., et al., *Genetics of human gut microbiome composition.* BioRxiv, 2020.
52. Goodrich, J.K., et al., *Genetic determinants of the gut microbiome in UK twins.* Cell host & microbe, 2016. **19**(5): p. 731-743.
53. Goodrich, J.K., et al., *Human genetics shape the gut microbiome.* Cell, 2014. **159**(4): p. 789-799.
54. Kurilshikov, A., et al., *Host genetics and gut microbiome: challenges and perspectives.* Trends in immunology, 2017. **38**(9): p. 633-647.
55. Qin, Y., et al., *Combined effects of host genetics and diet on human gut microbiota and incident disease in a single population cohort.* medRxiv, 2020.
56. Perez-Muñoz, M.E., et al., *A critical assessment of the "sterile womb" and "in utero colonization" hypotheses: implications for research on the pioneer infant microbiome.* Microbiome, 2017. **5**(1): p. 48.
57. Wang, J., et al., *Genome-wide association analysis identifies variation in vitamin D receptor and other host factors influencing the gut microbiota.* Nature genetics, 2016. **48**(11): p. 1396-1406.
58. Martínez, I., et al., *Experimental evaluation of the importance of colonization history in early-life gut microbiota assembly.* Elife, 2018. **7**: p. e36521.
59. Hibbing, M.E., et al., *Bacterial competition: surviving and thriving in the microbial jungle.* Nature reviews microbiology, 2010. **8**(1): p. 15-25.
60. Foster, K.R., et al., *The evolution of the host microbiome as an ecosystem on a leash.* Nature, 2017. **548**(7665): p. 43-51.

61. McBurney, M.I., et al., *Establishing what constitutes a healthy human gut microbiome: state of the science, regulatory considerations, and future directions.* The Journal of nutrition, 2019. **149**(11): p. 1882-1895.
62. Young, V.B., *The role of the microbiome in human health and disease: an introduction for clinicians.* Bmj, 2017. **356**: p. j831.
63. Bruce-Keller, A.J., et al., *Obese-type gut microbiota induce neurobehavioral changes in the absence of obesity.* Biological psychiatry, 2015. **77**(7): p. 607-615.
64. Berer, K., et al., *Gut microbiota from multiple sclerosis patients enables spontaneous autoimmune encephalomyelitis in mice.* Proceedings of the National Academy of Sciences, 2017. **114**(40): p. 10719-10724.
65. Canyelles, M., et al., *Trimethylamine N-oxide: a link among diet, gut microbiota, gene regulation of liver and intestine cholesterol homeostasis and HDL function.* International journal of molecular sciences, 2018. **19**(10): p. 3228.
66. Dannenberg, L., et al., *Targeting the human microbiome and its metabolite TMAO in cardiovascular prevention and therapy.* Pharmacology & Therapeutics, 2020: p. 107584.
67. Sanna, S., et al., *Causal relationships among the gut microbiome, short-chain fatty acids and metabolic diseases.* Nature genetics, 2019. **51**(4): p. 600-605.
68. Chu, C., et al., *The microbiota regulate neuronal function and fear extinction learning.* Nature, 2019. **574**(7779): p. 543-548.
69. Dalile, B., et al., *The role of short-chain fatty acids in microbiota–gut–brain communication.* Nature Reviews Gastroenterology & Hepatology, 2019: p. 1.
70. Van Nood, E., et al., *Duodenal infusion of donor feces for recurrent Clostridium difficile.* New England Journal of Medicine, 2013. **368**(5): p. 407-415.
71. Ott, S.J., et al., *Efficacy of sterile fecal filtrate transfer for treating patients with Clostridium difficile infection.* Gastroenterology, 2017. **152**(4): p. 799-811. e7.
72. Moayyedi, P., et al., *Fecal microbiota transplantation induces remission in patients with active ulcerative colitis in a randomized controlled trial.* Gastroenterology, 2015. **149**(1): p. 102-109. e6.
73. Rossen, N.G., et al., *Findings from a randomized controlled trial of fecal transplantation for patients with ulcerative colitis.* Gastroenterology, 2015. **149**(1): p. 110-118. e4.
74. Maldonado-Gómez, M.X., et al., *Stable engraftment of Bifidobacterium longum AH1206 in the human gut depends on individualized features of the resident microbiome.* Cell host & microbe, 2016. **20**(4): p. 515-526.
75. Wang, Q., et al., *Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy.* Applied and environmental microbiology, 2007. **73**(16): p. 5261-5267.

6

7

GENERAL SUMMARY

Genetically informative study designs have long been of great importance for helping to determine the role the environment and genetic susceptibility play in several human phenotypes. Of such designs, one of the most well-known is the classical twin design, including both monozygotic (MZ) and dizygotic (DZ) participants. In addition to the comparison of MZ and DZ twin pairs, studies focused on MZ twin pairs discordant for a specific phenotype allow for researchers to better understand how environmental influences impact a trait while controlling for the host genomic profile. Beyond the use of twin derived samples, with information on the genomic content of unrelated individuals, it is possible to create genetically informative designs aimed at understanding many biological phenomena.

In **Chapter 2** we study blood and buccal derived DNA samples collected simultaneously in order to investigate telomere repeat mass (TRM) in both tissue types. Performing the telomere measurement in a cohort containing MZ and DZ twin pairs allowed us to determine the contribution of genetic and environmental factors to each of the two blood and the buccal TRM phenotypes. Furthermore, the unique nature of the dataset allowed us to speak to the overall suitability of buccal derived DNA for TRM measurement. The two analyses in the blood samples showed a difference in estimated heritability, with genetic factors explaining 47.6% and 22.2% of total phenotypic variance in TRM in the first and second analysis of the same blood sample. Heritability of buccal TRM was 23.3%. The comparisons showed that, when performed by the same laboratory (AIHG), 11% to 12% of the variance in TRM in blood samples was predicted by TRM in buccal cells, i.e., the correlation was significant but modest ($r = 0.244 - 0.415$). The buccal associated DNA did display the same age and sex associated effects commonly documented in measurements based on blood-derived DNA, indicating that the TRM data is showing the expected male disadvantage and telomere attrition with age. **Chapter 2** provides support, albeit modest, for the use of buccal derived DNA in large scale biobank studies focused on the study of TRM, which will allow for less invasive longitudinal studies. It is also clear that repeated handling is ill-advised for TRM measurement for either buccal or blood samples.

Chapter 3 focuses on using multiple genetically informative designs to understand the relationship between the gut microbiota and host obesity-associated measures. These studies include the use of samples collected from body mass index (BMI) discordant MZ twin pairs and the four-corners study design, with unrelated individuals at high/low extremes for both genetic risk for obesity and actual BMI. **Chapter 3** provides support for a distinctly lower gut microbiota diversity in individuals with high BMI, particularly if they are low in the genetic susceptibility to obesity. Additionally, we identified a number of OTUs in the *Ruminococcaceae* and *Oxalobacteraceae* families that have a significant association with obesity-associated measures and their depletion may be a source of the lower microbiota diversity seen in more obese individuals.

Chapter 4 explored the relationship between variation in the host genome and the subsequent gut microbiome composition. As a consortium, we recently produced the largest GWA on human gut microbiome composition too date. It identified 30 loci affecting ten microbiome taxa at a genome-wide significant ($P<5x10^{-8}$) threshold, with one locus, the lactase (LCT) gene region, remaining study-wide significant after additional correcting for the number of taxa tested (GWAS signal $P= 8.6x10^{-21}$). **Chapter 4** describes results generated during the work needed to contribute to this GWA meta-analysis by the international consortium. While our main aim was to provide the NTR results for the meta-analysis, in parallel, we tested a few specific hypotheses using the genome-wide data in the NTR sample. The consortium identified the association between the *Bifidobacterium* genus and rs182549 as the most significant microbiome quantitative trait locus (mbQTL) ($p = 8.6332 \times 10^{-21}$) and the NTR-specific results for this mbQTL showed a p-value of 0.06, only just failing to achieve nominal significance. A more convincing resemblance between the NTR cohort and the consortium findings was found with regards to the link between heritability of a taxon and its ability to generate significant associations. Taxa defined at the family and genus levels that had the highest MZ twin correlations (>0.30) showed much lower p-values in their associations with genome-wide SNPs than less heritable taxa. Taken together, it seems appropriate to use the results of the MiBioGen consortium for extended analyses in smaller cohorts, such as performing MR with genetic instruments for the more heritable microbiota as predictors of inflammatory and cardiometabolic (including obesity) outcomes.

Chapter 5 was concerned with understanding how environmental factors, in the form of cohabitation, impact the gut microbiome composition. Samples collected from MZ twin pairs who cohabitate and those that do not, in addition to spouse pairs, allows a unique view into the role of cohabitation as an important environmental modulator of the gut microbiota. The work in **Chapter 5** identified two species-level OTUs (Otu0081 and Otu0190) that were significantly shared between cohabitating MZ twin pairs and spouse pairs but not between non-cohabitating MZ twins. The lack of sharing between the non-cohabitating MZ twins rules out genetic influences on these microbes. Microbes that are particularly influenced by cohabitation were further explored to determine relationships between host cardiometabolic and disease burden. Disease burden scores were regressed on the OTU counts of the two species-level OTUs shared due to cohabitation. One of these OTUs (OTU190) was significantly associated with lower inflammatory and cardiometabolic disease burden.

Considering the ultimate contributions of nature and nurture in dictating the gut microbiome composition, it is clear that both of these factors play a non-trivial role, but meaningful patterns are still largely yet to be deduced. During our discussion of these factors in **Chapter 6**, we aimed to reorient the microbiome field with regard to the discussion surrounding heritability of individual microbiome components. As

such, we define a microbiome constituent as heritable if the host provides, by means of its genetic architecture, conditions amenable to colonization by a particular microbe upon encounter. The key part of this definition relies upon environmental exposure to a microbe of interest in order for the heritable nature of the organism to be observed. What has only become clearer throughout the development of this thesis is that host genetics, host environmental factors, the gut microbiome composition and host health are deeply intertwined factors.

7

A

APPENDIX

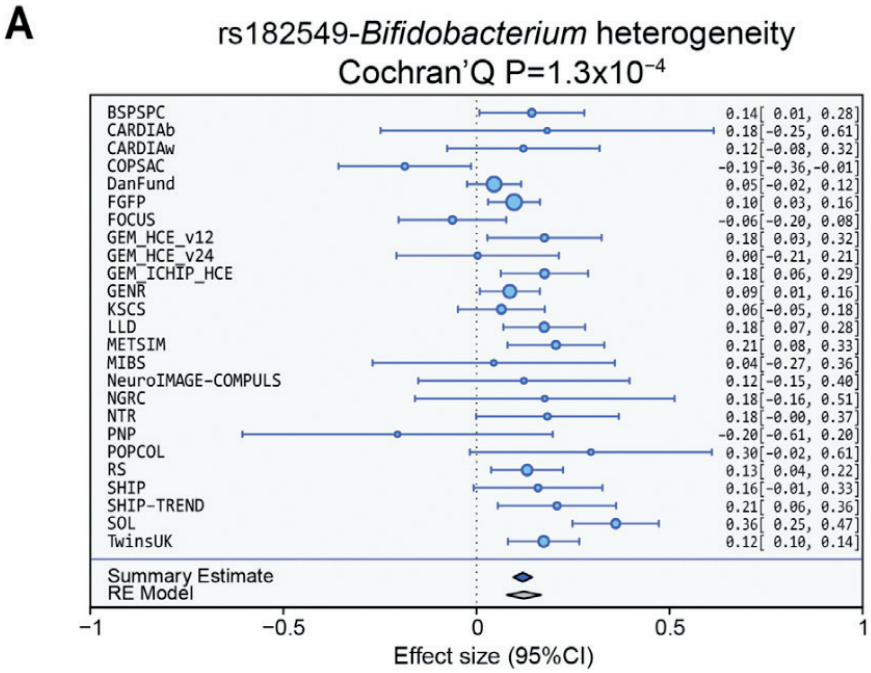


Figure A1 | Cochran's Q test demonstrating an NTR effect size similar to other cohorts. This figure is present in Kurilshikov et al., 2020.

LIST OF PUBLICATIONS

Finnicum, C.T., et al., *Relative telomere repeat mass in buccal and leukocyte-derived DNA.* PLoS One, 2017. **12**(1): p. e0170765.

Finnicum, C.T., et al., *Metataxonomic analysis of individuals at BMI extremes and monozygotic twins discordant for BMI.* Twin Research and Human Genetics, 2018. **21**(3): p. 203-213.

Finnicum, C.T., et al., *Cohabitation is associated with a greater resemblance in gut microbiota which can impact cardiometabolic and inflammatory risk.* BMC microbiology, 2019. **19**(1): p. 1-10.

Kurilshikov, A., et al., *Large-scale association analyses identify host factors influencing human gut microbiome composition.* Nature Genetics, 2020.

Wang, J., et al., *Meta-analysis of human genome-microbiome association studies: the MiBioGen consortium initiative*. 2018, BioMed Central.

Hur, Y.-M., et al., *The Nigerian Twin and Sibling Registry: An Update.* Twin Research and Human Genetics, 2019. **22**(6): p. 637-640.

Beck, J.J., et al., *Genetic similarity assessment of twin-family populations by custom-designed genotyping array.* Twin Research and Human Genetics, 2019. **22**(4): p. 210-219.

PUBMED REFERENCES: