

LEVERAGING THE MICROBIOME IN HOST GENOME WIDE ASSOCIATION STUDIES

AWANY DENIS

Thesis Presented for the Degree of Doctor of Philosophy



Division of Human Genetics, Dept of Pathology
Faculty of Health Sciences
University of Cape Town

Supervisor: A/Prof. Emile R. Chimusa

Co-supervisor: Prof. Collet Dandara

July 2020

The copyright of this thesis vests in the author. No quotation from it or information derived from it is to be published without full acknowledgement of the source. The thesis is to be used for private study or non-commercial research purposes only.

Published by the University of Cape Town (UCT) in terms of the non-exclusive license granted to UCT by the author.

Declaration

I, Denis Awany, hereby declare that the work on which this thesis is based is my original work (except where acknowledgements indicate otherwise) and that neither the whole work nor any part of it has been, is being, or is to be submitted for another degree in this or any other university. I empower the university to reproduce for the purpose of research either the whole or any portion of the contents in any manner whatsoever.

Signature: Signed by candidate Date: May 12, 2021

I confirm that I have been granted permission by the University of Cape Town's Doctoral Degrees Board to include the following publications in my PhD thesis, and my co-authors have agreed that I may include the publications:

1. **Denis Awany**, Imane Allali, Shareefa Dalvie, Sian Hemmings, Kilaza S. Mwaikono, Nicholas E. Thomford, Andres Gomez, Nicola Mulder, and Emile R. Chimusa. (2019). Host and microbiome genome-wide association studies: current state and challenges. *Frontiers in Genetics* 9: 637.
2. **Denis Awany**, Imane Allali, and Emile R. Chimusa. (2019). Tantalizing dilemma in risk prediction from disease scoring statistics. *Briefings in Functional Genomics* 18, no. 4: 211-219.
3. **Denis Awany**, Imane Allali, and Emile R. Chimusa. (2020). Dissecting Genome-wide Studies for Microbiome-related Metabolic Diseases. *Human Molecular Genetics*.

Signature: Signed by candidate Date: May 12, 2021

Student Name: Denis Awany Student Number: AWNDEN001

"If I have seen further, it is by standing upon the shoulders of giants".

- Isaac Newton, *Letter to Robert Hooke*, 1675.

Acknowledgement

I could not have produced this this work without the help of many people. First I thank God for seeing me through to this point.

I am very grateful to the German Academic Exchange Service (DAAD) for funding this study.

I am indebted to my supervisor A/Prof. Emile R. Chimusa and co-supervisor Prof. Collet Dandara who have been instrumental in every stage of the project, right from planning to development and implementation of various parts of the project. I also thank them for their contributions to the quality of this thesis and the various research articles published during the course of this study. I am also very grateful to them for the encouragement and support throughout the time. My debt to them is beyond measure.

Heartfelt thanks, too, to my friend Dr. Delesa Damena for his unwavering support, outstanding kindness, perpetual good humour, and pragmatic advice that played no small part in this endeavour.

Special gratitude also goes to staff in the Division of Human Genetics. Special thanks to Prof. Raj Ramaser, Dr. Lisa Roberts, Dr. Gloudi Agenbag, Dr. George Rebello, Sr. Gamede Benefeld, and Heidi Pulse for the assistance and generous support that ensured I had the necessary resources. I also thank them for taking special interest in students' welfare, and for the care that extended well beyond the academic workplace. All of which made the Division a great working and family-like environment.

I would also like to thank my friends and colleagues in the Division of Human

Genetics, including Lerato Majara, Khuthala Mnika, Peter Opiyo, Wonderful Choga, Francis Agama, Prisca Thami, Joel Defo, Noluthando Manyisa, Mary Mufford, Chandre Oosterwyk, Caitlin McIntosh, Safiye Yildiz, Doreen Mhandire, Kagisho Montjane, and Oluwafemi Oluwole for the helpful discussions, and wonderful time spent together. Much thanks is also extended to Jacqueline Mugo, Epiphanis Geza, and Samar Elsheikh from the Computational Biology group for the collaboration and group discussions.

I should also express my gratitude to all those with whom I co-authored the different research articles. The list is long but I would especially like to thank Dr Imane Allali, Dr Delesa Damena, Prof. Nicola Mulder, Dr Shareefa Dalvie, and Dr Nicholas E. Thomford for their collaboration and invaluable contributions.

Finally, very special thanks go to my family and friends. That begins with my parents, Johnan Pule and Molly Pule, for their constant support and encouragement; my girlfriend Sarah Adong for her immense support and endurance during the challenging times; my brother and sisters: Josephine Abol, Okullo Solomon, Mike Obwala, Olem Ketty, Akello Beatrice, Pule Emmanuel, Ongura Stephen, and Obua Sadrack for always being there; and my friends, in particular, Emma Obua, Atwaru Antoni & Adongo Fenella Atwaru, Olwa Geoffrey, Kinene Tonny, and Egesa Joel for their support.

List of Publications

1. Publications from this thesis

The following papers are from the contributions of this thesis.

1. **Denis Awany**, Imane Allali, and Emile R. Chimusa. (2019). Tantalizing dilemma in risk prediction from disease scoring statistics. *Briefings in Functional Genomics* 18, no. 4: 211-219.
2. **Denis Awany**, Imane Allali, Shareefa Dalvie, Sian Hemmings, Kilaza S. Mwaikono, Nicholas E. Thomford, Andres Gomez, Nicola Mulder, and Emile R. Chimusa. (2019). Host and microbiome genome-wide association studies: current state and challenges. *Frontiers in Genetics* 9: 637.
3. **Denis Awany**, Imane Allali, and Emile R. Chimusa. (2020). Dissecting Genome-wide Studies for Microbiome-related Metabolic Diseases. *Human Molecular Genetics*.
4. **Denis Awany**, and Emile R. Chimusa. Associating the microbiome with host phenotype using Dirichlet process random effects model. Under Review.
5. **Denis Awany**, and Emile R. Chimusa. (2020). Heritability jointly explained by host genetics and microbiome. will improve trait prediction? *Briefings in Bioinformatics*.

2. Publications from other contributions

The following papers are from joint effort within the group during the course of this Ph.D study.

1. Shatha Alosaimi, Armand Bandiang, Noelle van Biljon, **Denis Awany**, Prisca K Thami, Milaine S Tchamga, Anmol Kiran, Olfa Messaoud, Radia Ismaeel

- Mohammed Hassan, Jacqueline Mugo, Azza Ahmed, Christian D Bope, Imane Allali, Gaston K Mazandu, Nicola J Mulder, Emile R Chimusa. (2020). A broad survey of DNA sequence data simulation tools. *Briefings in Functional Genomics* 19, no. 1: 49-59.
2. Damena Delesa, **Denis Awany**, Lemu Golassa, and Emile R. Chimusa. (2019). Genome-wide association studies of severe *P. falciparum* malaria susceptibility: progress, pitfalls and prospects. *BMC Medical Genomics* 12, no. 1: 1-14.
 3. Chimusa Emile R., Joel Defo, Prisca K. Thami, **Denis Awany**, Delesa D. Mulisa, Imane Allali, Hassan Ghaza, Ahmed Moussa, and Gaston K. Mazandu. (2018). Dating admixture events is unsolved problem in multi-way admixed populations. *Briefings in Bioinformatics*: 1-58.

Contents

Declaration	i
Acknowledgement	ii
List of Publications	iv
Abbreviations	xiii
Abstract	xiv
1 Introduction	1
1.1 Background	1
1.2 Motivation	5
1.3 Aims of the thesis	7
1.4 Thesis contributions	7
1.5 Thesis outline	8
2 Host Genome-wide Association Studies	10
2.1 Introduction	10
2.2 Mutation and Genetic Variation	11
2.3 Genetic Variation in Health and Disease	13
2.4 Statistical Methods in human GWAS	15
2.4.1 Linking Genomic Assay Results to Phenotypes	18
2.5 Heritability in GWAS	24
2.5.1 Broad-sense and narrow-sense heritability	26
2.5.2 Estimation of Heritability in GWAS	27
2.6 GWAS Findings and Implications for Disease Scoring Statistics	31
2.6.1 GWAS Findings	31
2.6.2 Implications for Disease Scoring Statistics	39
3 The Microbiome in Host Phenotype	50
3.1 Introduction	50

3.2	The role of the microbiome in host phenotype	54
3.3	Linking microbiome to the host phenotype	55
3.3.1	Modelling with microbiome as compositional predictor . .	57
3.3.2	Modelling with microbiome as non-compositional predictor	59
3.4	Interaction of host genome with microbiome	62
3.5	Dissecting GWAS for microbiome-associated metabolic diseases	80
4	Associating the Microbiome with Host Phenotype using Dirichlet Process Random Effects Model	109
4.1	Introduction	109
4.2	Materials and methods	112
4.2.1	Model specification	112
4.2.2	Modelling correlation structure among microbial taxa .	113
4.2.3	Dirichlet process for the random effects ψ	117
4.3	Model likelihood	120
4.4	Bayesian inference of the parameters	122
4.4.1	Estimation of model parameters, θ	122
4.4.2	Full conditional distributions	123
4.4.3	Generation of sub-clusters matrix A	128
4.5	Application	130
4.5.1	Simulation studies	130
4.5.2	Real data studies	138
4.6	Concluding remarks	150
5	Conclusions and Perspectives	155
5.1	Summary of contributions	157
5.2	Perspectives	158
5.2.1	Insights into the missing heritability mystery	158
5.2.2	Statistical power in genetic association mapping	162
5.2.3	Disease risk prediction	165
5.2.4	Extension to the model	166
	References	199

Appendices	200
A.1 Simulating genotypes, OTU scores, and phenotype	201
A.2 Scripts to simulate genotypes, OTU scores, and phenotypes and then perform GWAS analysis	202
B. R script to implement the phyloDPM model	209

List of Figures

2.1	Direct and indirect association of markers on a haplotype block with a disease phenotype. Direct association between the genotyped marker and the unobserved causal marker due to high LD between them makes the genotyped marker to have a statistical association with the phenotype, through indirect association. . . .	17
2.2	Schematic representation of the link between SNPs and phenotype. In (a) we assume there are no confounders. The null hypothesis of testing for testing for the effect of a given SNP on the phenotype, then is that the SNP and phenotype are independent, and the alternative hypothesis is that the SNP does affect the phenotype. In (b) there are unmodeled confounders, which include population substructure or other possible confounders. In this case, the null hypothesis that SNP and phenotype are independent is no longer valid since there is an a priori correlation between SNPs and phenotype by the unmodeled factors.	22
2.3	Comparison of genes discovered by GWAS with known drug targets for common complex traits/diseases.	34
4.2	Heatmap showing the OTU abundance distribution for (a) simulated microbial community of 200 OTUs from 500 samples, and (b) real microbial community of 856 OTUs obtained from the human upper respiratory tract from 60 samples. The scale indicates abundance level of the OTU counts with dark representing zero counts (see legends).	131

4.6	Plots to assess convergence of the posterior distribution for parameters of the model. Plots are from posterior draws of the model applied to the simulated OTU dataset: trace plots, estimated posterior densities, cumulative quantile plot and sample auto-correlation function plot for a representative sampling chain for the effects of the first 10 OTUs.	153
4.7	Plots to assess convergence of the posterior distribution for parameters of the model. Trace plots, estimated posterior densities, cumulative quantile plot and sample auto-correlation function plot for the first 4 species, using schizophrenia data.	154
5.1	Quantile-Quantile plot of observed p values versus the expected p values (on a $-\log_{10}$ scale) for the genome-wide association analysis of the joint (host's genotype and microbiome OTU scores) model and marginal (host's genotype only) model.	163

List of Tables

2.1	Results for enrichment analysis of Inflammatory Bowel Disease-associated genes: NGENES=number of genes, STD=standard deviation, SE=standard error, $p=p$ -value. Results presented for the top 20 significant associations.	36
2.2	Results for enrichment analysis of Type 2 diabetes-associated genes: NGENES=number of genes, STD=standard deviation, SE=standard error, $p=p$ -value. Results presented for the top 20 significant associations.	36
4.1	Results of estimation in the simulation study. Estimate of the parameter (Estim) and standard error of the estimate (SE) based on 100 replications.	136
4.2	Summary of parameter estimates for the schizophrenia data: posterior mean for the coefficient for species effect (Coef), standard error (SE) and 95% credible interval (CI) for estimated parameters for OTU effects.	141
4.3	Summary of parameter estimates for the HIV/AIDS data: posterior mean for the coefficient for species effect (Coef), standard error (SE) and 95% credible interval (CI) for estimated taxa effects. Also shown are the corresponding classes for the species. The (–) on the species column indicates the species information is unavailable.	147
4.4	Summary of parameter estimates for the atherosclerosis data: posterior mean for the coefficient for species effect (Coef), standard error (SE) and 95% credible interval (CI) for estimated taxa effects. Also shown are the corresponding classes for the species. The (–) on the species column indicates the species information is unavailable.	148

5.1 p values of SNPs from genome-wide association analysis with the joint (host's genotype and microbiome OTU scores) model and marginal (host's genotype only) model. p - the joint model; p^* - the marginal model. Results displayed for the first 20 SNPs. . . . 164

Abbreviations

ABC	Approximate Bayesian Computation
ASV	Amplicon Sequence variant
CI	Credible Interval
DP	Dirichlet Process
GWAS	Genome-Wide Association Study
LD	Linkage Disequilibrium
LMM	Linear Mixed Model
M-H	Metropolis-Hastings
MAF	Minor Allele Frequency
MCC	Matthews Correlation Coefficient
MCMC	Markov Chain Monte Carlo
mGWAS	microbiome Genome-Wide Association Study
MWAS	Microbiome-Wide Association Study
NGS	Next Generation Sequencing
OTU	Operational Taxonomic Unit
RBF	Radial Basis Function
SNP	Single Nucleotide Polymorphism
WGS	Whole-Genome Sequence

“If the Human Genome Project gave us a book, scientists are now learning how to read it ... and biologists are beginning to face up to the uncomfortable truth that they have only been looking at the nouns ... now we are reading the spaces in between—verbs, adverbs, adjectives, pronouns and the rest, and they are complicated indeed.”

- Roger Highfield, *The Telegraph*, 19 June 2007

Abstract

Genome-wide association study (GWAS) has emerged as an effective method for detecting genetic polymorphisms associated with expressed phenotypes. Over the past decade, GWAS of human traits and diseases has revolutionized the field of complex disease genetics, identifying hundreds of genetic variants associated with several different phenotypes, ranging from metabolic diseases to cardiovascular and neuropsychiatric conditions. These associations have provided fundamental insights into the genetic architecture of disease susceptibility and led to initial forays into clinical applications, particularly in creation of genetic risk scores for improved disease risk prediction and identification of new drug targets for novel drug development.

Despite this gratifying success, however, for almost all complex traits, the identified genetic loci explain only a small proportion, generally less than half, of the estimated heritability. A number of alternative explanations have been offered for this, including undetected genetic effects, unaccounted-for environmental factors, and gene–gene and gene–environment interaction effects. Although there is no consensus on these explanations, it is universally acknowledged that a substantial proportion of the trait heritability is attributable to existence of a large number of undetected genetic variants distributed across the entire allele frequency spectrum, each of which has very small to modest effect on the phenotype, and non host-DNA factors that contribute to phenotypic variation.

In parallel to host GWAS, the advent of next-generation sequencing technologies (NGS) that enable culture-independent profiling of microbial communities has led to the rediscovery of the microbiome - the collective genome of the microorganisms that inhabit the body - and the emergence of microbiome-wide association studies. These studies have linked the gut microbiome to a variety of human conditions, ranging from neurological conditions, such as Parkinson's disease and autism, to metabolic diseases, such as obesity, diabetes, and cardiovascular disease. Given the critical importance of the microbiome in host phenotype, it is clear that in order to more comprehensively understand the basis of host phenotypic status, both the host's genotype and microbiome information have to be examined.

This thesis explores the dissection of microbial taxa and host genetic polymorphisms associated with human complex traits and diseases, and the interaction of human host genetic polymorphisms with the microbiome. Then, a Bayesian statistical framework, based on the Dirichlet process random effects model, is proposed for identifying microbial species associated with host phenotype. The proposed method uses a weighted combination of phylogenetic and radial basis function kernels to model microbial taxa effects, and a non-parametrically defined latent variable to model latent heterogeneity among samples. Philosophically, the non-parametric specification amounts to the addition of an infinite amount of prior information about all fine details of the parameters being modelled; thus represents an attractive strategy. The utility of the method is demonstrated through simulation experiments and application to real microbiome datasets for schizophrenia, HIV/AIDS, and atherosclerosis diseases, where it is shown that the method is not only robust but also has high statistical power for association inference, resulting in a framework that can contribute to our understanding of the link between the microbiome and human diseases. Understanding the human genetic predisposing factors in concert with this link will make human GWAS fulfil its translational potential, from patient stratification and disease risk prediction to identification of new biology and drug discovery.

Introduction

“Any good poet, in our age at least, must begin with the scientific view of the world; and any scientist worth listening to, must be something of a poet, must possess the ability to communicate to the rest of us his sense of love and wonder at what his work is about”.

- Edward Abbey, *The Journey Home*

1.1 Background

The genetic profile of an individual influences his susceptibility to disease. Discovering the genetic modifiers of common traits and diseases that have a genetic component, shown by familial pattern of inheritance, and translating it into clinical and medical practice remains a key frontier in human genetics. The identification of genetic variants that influence Mendelian disorders - hereditary diseases resulting from mutation in a single gene ([Antonarakis and Beckmann, 2006](#)) - such as cystic fibrosis and Huntington disease have often been made through family-based linkage studies, assessing patterns of inheritance of a genetic marker in families with numerous infected individuals ([Ott, 1999](#)). This method, however, proved largely unsuccessful when applied to complex or multifactorial diseases - defined as diseases that are determined from the joint action of many genes and their interaction with the environment ([Mitchell, 2012](#)) - such as cancer, diabetes, hypertension, and cardiovascular disease. An alternative methodological advance, then, was the candidate gene association study, in which

one or a few more variants that are thought to belong to the same biological pathway or have similar biological function are selected and tested for association with the disease. Clearly, the success of this strategy depends highly on the correct choice of genes to study, and is unable to discover new genetic variants (Amos et al., 2010; Witte, 2010). Indeed, characterized by non-replication and inconsistency, the approach was not able to propel us forward in the search for common genetic variants underlying common complex diseases.

In the last two decades, prodded by the sequencing of the human genome and accelerated by advances in genotyping array technologies and the development of a high-resolution genetic map of the human genome, genome wide association study (GWAS) has become the contemporary approach to relate genetic variations to phenotypic diversity. In a GWAS, hundreds of thousands to millions of genetic variants spanning the entire genome are tested to identify genomic regions where true trait/disease-related genes may lie. By leveraging the strengths while overcoming weaknesses of linkage and candidate-based studies, GWAS provide a greater power and resolution to detect small to modest effects. Moreover, by casting a wide net of genetic markers across the entire genome, the approach provides a more comprehensive and unbiased approach to examine the biological underpinnings of common variation across the genome (Duncan et al., 2019).

A plethora of GWAS performed during this time have identified thousands of variants associated with many traits and diseases leading to a remarkable advance in our understanding of the genetic architecture underlying many complex traits and diseases. Despite the spectacular catches, however, for the majority of traits and diseases, the amount of trait variation explained by the genome wide-significant variants remain typically only a small fraction of that estimated from classical family correlation studies. This difference, often referred to as ‘missing heritability’, remain an important problem topic among geneticists, and to which many different explanations have been proposed. Among them, is that part of the missing heritability is due to underlying common genetic variants not yet identified because GWAS analyses are not powered enough to capture all the genetic variants involved (Génin, 2019). This hypothesis is corroborated by

the observation that (i) considering all SNPs, including those that do not reach genome wide significance threshold, substantially increases the proportion of trait variability explained (Yang et al., 2010; Lee et al., 2011), and (ii) studies of increased sample size continue to detect new variants (Robinson et al., 2014; Génin, 2019).

Besides the issue of missing heritability, the success of the GWAS technique is tempered by some critical limitation. Key among which is the simplifying assumption that environmental factors such as diet, lifestyle are homogeneous among the individuals under study (Sandoval-Motta et al., 2017b). This assumption is, however, usually not met as lifestyle and diet have all been shown to vary widely among groups of people and substantially impact the development and progression of complex diseases. The result of this is that individuals who are genetically distinct but have relatively similar diet and lifestyle may exhibit similar phenotypes. Thus, these unaccounted-for sources of phenotypic variation affects the statistical inference of causal loci, and, ultimately, leads to erroneous interpretation of the estimated heritability. Although why the missing heritability exists and how it may be addressed remains a contentious issue (Bourrat et al., 2017; Young, 2019; Pallares, 2019; Génin, 2019), the observations show that the variants so far identified represent only a minute fraction of those that influence phenotype, and there are likely to be many variants distributed across the entire allele frequency spectrum.

In order to enable further dissection of genetic variation, statistical power has to be improved. Statistical power is a function of sample size, effect size, allele frequency, and linkage disequilibrium (LD) between the markers. Given that the majority of the causal variants likely have low-frequency ($< 5\%$) (Robinson et al., 2014; Lettre, 2014; Bomba et al., 2017; Fournier et al., 2019), increasing sample size will have greatest effect on power. Apart from the usual desire of statistical power, commonly viewed in terms of sample size, it is clear from the aforementioned observations that in order for the basis of phenotypic variation to be broadly understood and accounted for, different sources of phenotypic plasticity have to be incorporated into the analysis. These include epigenetic and other vertically-

transmitted non-DNA factors; all of which contribute to phenotypic variation (Bourrat et al., 2017). Perhaps most crucial among these non-DNA factors is the microbiome - the collection of microscopic organisms, including bacteria, archaea, fungi and protozoa, that colonize our body surface together with their collective genome (Elinav et al., 2019).

While the importance of the human microbiome in conferring health benefits has been known for over a century (Gilbert et al., 2016), its role in pathogenesis of complex diseases, including immune, inflammatory, and metabolic phenotypes, has also been appreciated only during the past decade; thanks to advent of massively parallel sequencing enabling profiling of the microbiota without the need to culture the microorganisms. The gut microbiome, for instance, has been implicated in a diverse array of conditions, including colon cancer (Garrett, 2019), obesity (Tseng and Wu, 2019), irritable bowel syndrome (Tseng and Wu, 2019), cardiovascular disease (Peng et al., 2018), and depression (Winter et al., 2018). Given the role of the gut microbiome in modulating host's phenotypic expression, it is fairly intuitive that integrative analysis of host genotype and microbiome data represent an apt approach to understanding disease etiology. This concept is supported by several empirical observations. First, both host genetics and the microbiome are associated with the same phenotypes, and host genetic variants associated with various host traits, and variation in microbiome overlap. Second, the human genotype interacts with the microbiome, but cannot by itself explain microbial variation (Sandoval-Motta et al., 2017b). Third, microbiome-related phenotypes can be vertically transmitted between phylogenetically close humans (Wang et al., 2019a; Ferretti et al., 2018). Finally, the human microbiome co-evolved with its human host through several generations, and therefore is a potentially rich source of information on phenotypic plasticity and genetic variation (Sandoval-Motta et al., 2017a).

1.2 Motivation

Dissecting the basis of common complex diseases have increasingly become the focus of human genetic and microbiome studies as the burden of conditions such as cardiovascular disease, cancers, diabetes, and hypertension, and the like remains high. Yet, although many epoch-making discoveries have been made, elucidating the putative causative factors and providing insights into the potential underlying molecular mechanisms in disease pathogenesis, the causes, let alone the molecular pathways and mechanisms of these disease have not been fully delineated. In genetic association mapping, the GWAS-uncovered variants do not completely explain causality and heritability. For example, although the *LMTK2* genetic variant is significantly associated with prostate cancer, its presence does not increase the odds of a person developing prostate cancer (Sandoval-Motta et al., 2017b; Zuk et al., 2012). In human height, estimates of family studies indicate that about 80% of the variation in human height is due to genetic effects. However, the associated variants do explain only about 56% of variability in height across individuals (Yang et al., 2015). Furthermore, the implicated variants are less predictive of disease risk among individuals. Taken together, these observations, in part, imply that many of the genetic variants remain shielded in their genomic background, as evidenced by the fact that well-powered studies continue to reveal additional novel genetic variants.

In parallel, technological advances permitting culture-independent and rapid characterization of microbial communities in biological samples have facilitated microbiome wide association studies (MWAS) to detect and investigate the role of microbial species, represented by Operational Taxonomic Units (OTUs), in host traits. These studies continue to corroborate the intricate link between the microbiome and its host phenotype. Although the relationship between the human microbiome and phenotype is beginning to come into view, the specific microbial taxa involved in each phenotype are still not well known. There is therefore need for studies to pinpoint the phenotype-modifying community members. This in

turn would help to qualitatively and quantitatively define the role of the microbiome in health and disease. The need to detect complex associations between a large number of correlated microbial taxa and the host phenotypic state (clinical outcome) comes the need for statistical frameworks that are able to model the underlying relationships. The most common methodological approach towards this goal has been to apply a model that associates variation in the overall microbial community with the host phenotype. While useful in evaluating whether variability in the community-wide structure is associated with the particular host phenotype, this approach does not provide answers to the equally biologically interesting question of which microbial community members are involved in the association.

Besides the need to identify phenotype-associated microbial taxa, there is also an overarching need to understand the interaction of the microbiome with its host genetics. In view of this, inspired by the success of the GWAS methodology in host disease mapping, microbiome genome-wide association study (mGWAS) methodology has recently been devised. mGWAS seeks to identify host genetic polymorphisms associated with various microbiome attributes, such as relative abundance, alpha-diversity and beta-diversity measures (Lynch et al., 2017; Hua et al., 2015). While initial foray into mGWAS has revealed some exciting findings about this interaction, there is still paucity in knowledge regarding the host genes and pathways involved, the extent of the interaction, and the overlap of implicated genetic variants with those already established through host GWAS disease mapping. Furthermore, although the role of the microbiome has been more established for metabolic diseases, and the influence of individuals' genetic diversity on association inference from GWAS and MWAS is generally acknowledged, genome-wide studies for metabolic diseases have not been dissected.

Finally, another critical issue for trait association mapping is that of statistical epistemology. Complex trait association mapping is a many-to-many relationship of causes to trait, with many trait-modifying factors that are not necessarily independent. The deluge and stochasticity of these factors make association mapping epistemologically problematic, presenting a statistical as well as computational

challenge for association mapping frameworks. Thus, there is need to explore and highlight their strengths and inherent limitations to avoid potential pitfalls in using them, and to highlight areas where further methodological advances are needed.

These empirical and methodological issues motivate the work in this thesis, leading to the following thesis aims.

1.3 Aims of the thesis

1. To explore host and microbiome disease association mapping, and the interaction of host genetics with the microbiome.
2. To explore current statistical frameworks for host and microbiome association mapping.
3. To develop a statistical methodology for identifying microbial Operational Taxonomic Units (OTUs) associated with the host phenotype.

1.4 Thesis contributions

This thesis contributes to literary knowledge empirically and methodologically. In summary, these contributions are as follows:

1. A detailed investigation of the human microbiome and its interaction with human genetic polymorphisms to provide insights into the genetic and microbiome basis of complex human traits and diseases.
2. A comprehensive review of disease scoring models for genome-wide and microbiome-wide human trait/disease mapping, highlighting the strengths and limitations of contemporary disease association models, and proposing

methodological approaches to address some of the limitations to improve association inference.

3. A dissection of the genetics of microbiome-related metabolic diseases and the leveraging of whole genome sequence (WGS) samples from world-wide ethnic groups to examine genetic diversity in variants linked to metabolic diseases, demonstrating the marked differences in gene-specific proportion of pathogenic SNPs and gene-specific SNPs minor allele frequencies across ethnic groups.
4. The development of a statistical framework for studying association between microbial OTUs and the host's phenotype. The approach leveraging the phylogenetic tree structure and the radial basis function in modelling microbial OTU effects, and using a non-parametrically specified random variable to model heterogeneity across samples. The resulting model (Dirichlet process random effects model) offering considerable promise in addressing multicollinearity, and unobserved heterogeneity and over-dispersion in microbiome data.

1.5 Thesis outline

This thesis is organised in five chapters. The outline of the rest of the work is as follows:

Chapter 2: *Host Genome-wide Association Study*. This chapter presents and explores various aspects of host genome-wide disease association study, and the underpinning statistical methods.

Chapter 3: *The Microbiome in Host Phenotype*. This chapter gives a background of the microbiome, discusses the statistical models to link the microbiome with its host trait/disease, and explores the role of the microbiome on its host trait/disease and the interaction of host genetic polymorphisms with the microbiome.

Chapter 4: *Associating the Microbiome with Host Phenotype using Dirichlet process Random Effects Model*. This chapter presents a proposed Bayesian statistical model, the Dirichlet process random effects model: presenting its mathematical formulation, and evaluating its performance using simulation experiment and real microbiome data to show the merits of the approach.

Chapter 5: *Conclusions and Perspectives*. This chapter concludes the thesis with a summary of the work, and contributions of the thesis. It also points out the limitations of the work in the thesis and comments on some potential directions for future work.

Appendices: These sections present additional information on some of the materials discussed in the thesis.

Host Genome-wide Association Studies

“It is remarkable that a science which began with the consideration of games of chance should have become the most important object of human knowledge”

- Pierre Simon Laplace, *Théorie Analytique des Probabilités*

This chapter first introduces human genetic variation and discusses its role in human traits and diseases. Then, a discussion is presented of the statistical methods for genome-wide association study (GWAS), and heritability estimation from GWAS. The chapter concludes with a discussion of findings from GWAS of common human complex diseases, and the implications for disease scoring statistics.

2.1 Introduction

Deciphering the genetic basis of human health and disease has long been a subject of interest. Following the unveiling of complete human genome sequence, a number of large-scale studies were initiated to identify and catalogue nucleotides that differ among individuals, with studies focused on understanding the architecture of the different forms of genetic variants. In aggregate, these studies have increased our knowledge of the nature and pattern of genetic variation, including their correlation structure and the properties of mutational events that affects

the structural integrity of the genome. Given the relatively large frequency of most genetic polymorphisms, and the existence of linkage disequilibrium (LD) between DNA polymorphisms and disease mutations, association study provides a means to link genetic variation to phenotypic variation. In the pursuit to uncover the underlying predisposing genetic factors, however, association mapping is confronted by the characteristic feature of most human traits - that the majority of variants have relatively small to modest effects and act in concert with other genetic and environmental factors to shape the host's phenotype. These features complexity association studies.

In this regard, the genome-wide association study (GWAS) approach has arisen as a powerful association framework, becoming the contemporary approach to dissect the genetic basis of phenotypic variation. GWAS makes use of the dense array of SNPs that span the genome to look for allele-frequency differences between individuals with different phenotypic states, from where significant allele-frequency differences is taken to indicate that the corresponding genomic region contains DNA-sequence variants that have biological effect on the trait or disease under consideration. In the sequel, we explore human genetic variation and its connection to health and disease, and the underlying disease scoring approaches.

2.2 Mutation and Genetic Variation

The human genome contains approximately 3 billion distinct pieces of nucleic acid bases, namely Adenine (A), Cytosine (C), Thymine (T) and Guanine (G), which are arranged in a linear fashion forming the human genome sequence. The event of mutation changes the bases at a specific locations in the genome (e.g $A \rightarrow T$), resulting in a change in the nucleotide sequence (genetic variation). These mutations can be inherited from parents (germline mutations) or acquired over the course of life of an individual (somatic mutations). Inherited germline genetic variants and somatic alterations acquired during cell division are crucial

determinants in phenotypic variability and susceptibility to complex diseases. Whereas the vast majority of somatic mutations have been ascertained to be harmless, a small proportion of them do have critical functional effects and thus can influence disease susceptibility (Chen et al., 2020).

Mutational events continuously generate genetic variations in the genome, but their prevalence in the genome is determined by a diverse array of demographic and genomic factors. Population bottleneck, for example, in which the current population is descended from a few antecedents, can impact the pattern of genetic variability across the whole genome (Raynes et al., 2014; Pedersen et al., 2017). Similarly, population size determines the persistence of genetic variation as mutations are more likely to persist in a large population than in a small population, since it is less likely to be lost from a large population by sampling effects (Wright, 2001; Whitlock and Bürger, 2004). Mutation in humans vary across the genome from about 10^{-7} to 10^{-9} per nucleotide base per generation (Bush and Moore, 2012). Although the human genome has remained conserved throughout the course of human evolution, the accumulation of these mutations over many generations means that the genomes of any two individuals are never exactly identical. Indeed, humans are 99.5% identical at the genomic sequence level implying that the observed inter-individual phenotypic variation emanates from the contributory role of the remaining 0.5%.

A DNA sequence variation that is common in a population (that is, occurs in a population with a frequency of at least 1%) is termed a polymorphism (Karki et al., 2015). Just like mutations, polymorphisms can also be of one or more nucleotide changes. Single nucleotide polymorphisms (SNPs) are by far the most common type of genetic polymorphism in which a nucleotide changes from one purine base to another, for example $A \leftrightarrow G$ or $C \leftrightarrow T$, which is termed a transition, or $A \leftrightarrow C$ or T ; and $G \leftrightarrow C$ or T , which is referred to as a transversion. It is estimated that the human genome contains at least 10 million SNPs, arising approximately every 100–300 base pairs in the genome (Robert and Pelletier, 2018) and located in coding and non-coding regions, introns, and intergenic regions (Karki et al., 2015). Because of this, SNPs are used as genetic signatures in

populations to study how genotypic variation translates into phenotypic variation. If a genetic marker is associated to a particular phenotype, then the genotype or allele frequency of the marker (SNP) will be different among the case (affected) and control (healthy) subjects.

2.3 Genetic Variation in Health and Disease

Human genetic variation describes this genetic contrasts between individuals within and among populations. The spectrum of the genetic variation is broad, ranging from point mutations, where a single pair of nitrogenous bases is modified, and small deletions and insertions (indels), to gene deletions and gross chromosomal rearrangements. The polymorphisms affecting relatively large segments of the genome are generally referred to as structural variants (SVs), and include copy number variations (CNVs), inversions of DNA sequences, and block substitutions; albeit, SV has recently been redefined to refer to variations covering 50 or more base pairs (Liu et al., 2019a).

Single nucleotide polymorphisms (SNPs) and short indels are the most common type of genetic variation among people, constituting more than 99.9% of the genomic variants in the human genome (Consortium et al., 2015). Although SVs are only a minute fraction of genomic variants, their relatively large size (20 Mb of sequence, compared to 4-5 Mb for SNPs) (Vizmanos et al., 2020) means that an individual SV potentially has a higher phenotypic impact than single nucleotide variants, especially when affecting functional sequences (Lin and Gokcumen, 2019). Indeed, a recent analysis of data from the Human Gene Mutation Database showed that up to 34% of all disease-causing variants were made up of variants larger than single base-pair substitutions (Eichler, 2019), underscoring the important roles of this class of variation in human health and disease. That said, however, the many common diseases are likely to be effected by single site variants of small effects. These variants are not symmetrically distributed in the genome. With non-coding regions constituting the vast majority of the genome,

SNPs generally occur more frequently in non-coding regions than in coding regions (Barreiro et al., 2008), with the variant density being influenced by other factors such as genetic recombination and mutation rate (Mahmud et al., 2016).

A characteristic feature of a genomic variant is its population frequency, and the terms 'common', 'low frequency' and 'rare' are typically used to classify variants according to the frequency of their minor allele in the population. Common variants specifies variants with a minor allele frequency (MAF) of at least five percent in the population, whereas low frequency and rare variants have MAF greater than 1% but less than 5%, and MAF less than 1% respectively (Frazer et al., 2009). Advances in microarrays and next-generation sequencing have enabled large-scale studies to identify and catalogue nucleotides that differ among individuals. These studies - not withstanding the difficulty in detecting and characterizing SVs, as compared to SNVs - have provided insights into the range of patterns and frequencies of variations in the human genome. Using genome sequences from many different populations, it has recently been estimated that rare variants constitute the majority of human genetic variation (Consortium et al., 2015). Also, noteworthy, however, is that for a typical genome (as compared to the reference sequence), the majority of variants are common, with rare variants typically only about 200,000 per genome (Vizmanos et al., 2020). At a geographic and population level, there are also population-specific variations in genomic sequence and allele frequencies, that result from the effect of selection, genetic drift or both (Sirugo et al., 2019). Because the patterns of genetic variation among populations directly affect disease risk, knowledge of allelic variation and its distribution in a population is critical in the design and interpretation of genetic association studies. For instance, while genome-wide analysis at single base pair resolution may be sufficient to determine genetic risk factors for common variants, detecting association effects for rare variants would typically require large population-scale studies, due to its presence in only a few individuals. Besides the necessity in the design of association study, the allele frequency architecture informs on population history and the degree of natural selection the associated trait has undergone (Geschwind, 2018). Low frequency and rare

variants are enriched for alleles that have impact on fitness, and therefore are relatively young and under purifying selection; common variants are old and are either neutral or under positive selection (Gorlov et al., 2011; Geschwind, 2018; Platt et al., 2019). The implication of this is that different biologic pathways may be differentially enriched for different forms of variation.

In parallel, holistic interrogation of the genome from many different human populations, have showed that rare variants constitute the majority of human genetic variation, with rare loss-of-function variants located in more than 90% of genes in the genome, despite only a small fraction of the human population sequenced (Geschwind, 2018). Despite the high prevalence of variants, however, the occurrence of functionally-impacting variation is low, and most variants - particularly those in intergenic regions or in nonfunctional gene segments such as introns - will likely be neutral (Vizmanos et al., 2020). In addition, deleterious mutation can be expected to be few as selective pressures eliminate the most deleterious mutations from the population (Andreoli et al., 2010). In effect, genetic variations between individuals can be neutral, normal, or deleterious; where neutral variations produce no effect on phenotype, normal causes benign phenotypic effects, and deleterious results in adverse effects on phenotypes. Deleterious mutations is of special interest in the search for variants that could be implicated in genetic diseases. A single mutation maybe able to produce a disease state - as in the case of *monogenic* diseases such as cystic fibrosis, sickle-cell anaemia, and Huntington's disease - or many mutations may act jointly to result in a disease condition - as in the case of *complex* diseases such as diabetes, asthma, and cancers. Genetic association study provides a means to link the putative causal loci to the phenotype under consideration.

2.4 Statistical Methods in human GWAS

Recent advances in genotyping technology have enabled fast and accurate high-throughput typing of genetic variations throughout the entire genome, setting the

stage for association studies. Genetic association studies seeks to identify genetic variants that vary systematically between individuals with the disease (cases) and individuals without the disease (controls), and which could therefore be playing a role in the pathogenesis of the disease. While other genetic association approaches do exist, for example ‘candidate gene’ studies and linkage studies, genome-wide association study (GWAS) has been widely adopted as a method of choice for agnostic evaluation of the association between genetic polymorphisms and disease risk owing to its higher sensitivity in detecting common variants (with large as well as small effects).

The predominant study design in GWAS is the case-control design, where unrelated (‘unrelated’ being used to mean that relatedness is unknown and presumed to be distant in generational time scale) affected cases and unaffected controls are genotyped at thousands or millions of loci distributed across the entire genome. Due to the size of the genome, however, we cannot genotype all markers (SNPs are typically used as *markers* of a genomic region). Luckily, owing to linkage disequilibrium (LD) - the non-random association of alleles at two or more loci (Sukhumsirichart, 2018) that arises because of the existence, in the current population, of small chromosomal segments that are descended from the same common ancestor (Hayes, 2013); the segment being so small that it can be assumed that no recombination will have arisen within the segment (Cordell and Clayton, 2005) - and the fact that some of the SNPs would provide redundant information, not all SNPs need to be genotyped. SNPs within close proximity on a chromosome may exist as a block that is in LD, and since this set of SNPs are inherited as a unit, knowing the base at any 1 SNP in the block can allow identification of all nearby SNPs in that block (Motsinger et al., 2007). Consequently, a subset of SNPs can be selected to effectively capture information about variations at nearby sites in the genome. These informative SNPs are usually referred to as *tagging* (Johnson et al., 2001; Motsinger et al., 2007), or simply *tag* SNPs because alleles for these SNPs “tag” other surrounding SNPs in the block. That said, we must note that, on the negative side, that strong long-range LD over large genomic regions can be problematic for GWAS as it can confound analysis; this is exemplified by the

human leukocyte antigen (HLA) locus which, as a result of a 3cM of high LD around it, was spuriously associated with susceptibility to immunological and infectious diseases in many early studies (Cardon and Bell, 2001; Grady et al., 2011). Because of this, LD informativeness for disease association studies should be made with regards to local proximity rather than only the magnitude of LD between pairs of markers.

The important significance of this leveraging of LD for a GWAS is that the effect of a causal polymorphism can still be detected even if that polymorphism is not genotyped in the study, as illustrated in Figure 2.1. This, however, results in

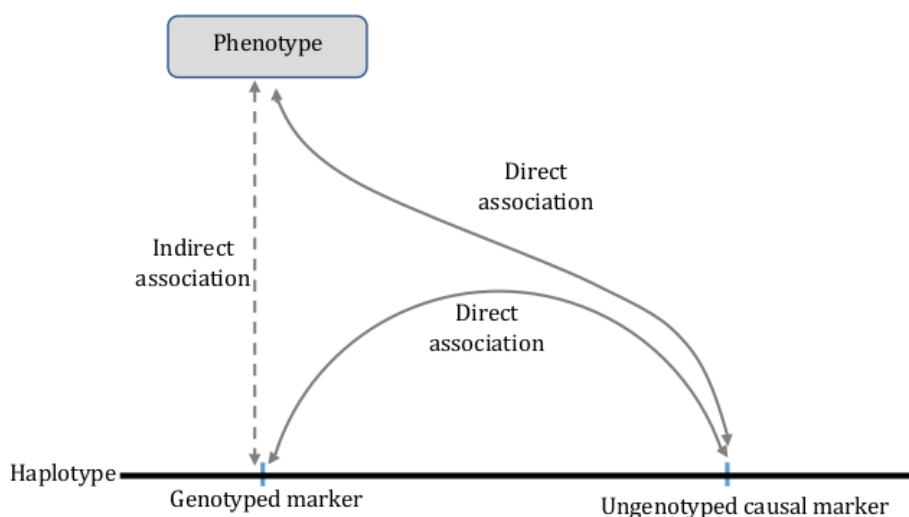


Figure 2.1: Direct and indirect association of markers on a haplotype block with a disease phenotype. Direct association between the genotyped marker and the unobserved causal marker due to high LD between them makes the genotyped marker to have a statistical association with the phenotype, through indirect association.

two possible associations being created from a genetic association study. First, is the direct association in which the SNP genotyped in the study is the ‘causal’ SNP (the SNP influencing the phenotype), and is statistically associated with the phenotype. Second, is the indirect association in which a tag SNP in high LD with the causal SNP, rather than the causal SNP itself, is the one genotyped and statistically associated with the phenotype. Because of these two possibilities, a variant found to be statistically significant in a GWAS should not be regarded as

causal, unless and until additional validation studies are carried out.

2.4.1 Linking Genomic Assay Results to Phenotypes

The GWAS approach, first proposed by [Risch and Merikangas \(1996\)](#) as a statistical method to detect common variants with modest genetic effects ([Scherer and Visscher, 2016](#)), is based on the fundamental theory that due to the large number of random recombination events occurring over many generations in the history of human population, only markers in tight linkage disequilibrium with loci responsible for the trait variation will exhibit significant statistical association with the trait ([Platt et al., 2010](#)). Based on this, statistical association in GWAS seeks to identify those variants that are associated with the phenotype of interest, whether through direct causality or indirect linkage disequilibrium.

Consider a randomly mating population and suppose that a sample of n individuals is taken from which measurements are made for m genotypes, denoted $x_{i1}, \dots, x_{im}$, and n phenotypes, denoted $y_1, \dots, y_n$. If we further suppose that the SNPs have additive effect on the phenotype and that there are no interactions (no polymorphisms whose presence can alter the effects of other polymorphisms), then, in the absence of population structure, the genotype can be linked to the phenotype using the linear model

$$y_i = \beta_0 + x_{i1}\beta_1 + \dots + x_{im}\beta_m + \epsilon_i \quad (2.1)$$

for a quantitative trait, and

$$P(y_i = 1) = \frac{\exp [\beta_0 + x_{i1}\beta_1 + \dots + x_{im}\beta_m]}{1 + \exp [\beta_0 + x_{i1}\beta_1 + \dots + x_{im}\beta_m]} \quad (2.2)$$

for a binary phenotype, where β_0 is an intercept, $\beta_1, \dots, \beta_m$ are SNP effects measuring the ‘strength’ of the association, and ϵ_i ’s are residuals assumed to be independent and identically distributed. The null hypothesis of interest then, is that the given SNP has no effect on the phenotype (that is, $H_0 : \beta_j = 0$, for

$j \in \{1, \dots, m\}$), whereas the alternate hypothesis is that marker has effect on the phenotype (that is, $H_1 : \beta_j \neq 0$, for $j \in \{1, \dots, m\}$) since it is in LD with the causal SNP.

We point out that association signals for binary case-control traits can alternatively be identified by testing each individual SNP using Pearson's χ^2 test or Fisher's exact test. The contingency table-based Pearson's χ^2 test and Fisher's exact test entails first determining the expected counts of each genotype under the assumption of no association between the phenotype and genotype, and then comparing these to the observed counts, to assess deviation from the null hypothesis that the phenotype and genotype are independent. One limitation of these contingency table-based methods is that there is no way of incorporating those factors that may be known to have important influence on the phenotype. Because of this, logistic regression model is often preferred. In contrast to Pearson's χ^2 test and Fisher's exact, application of logistic regression enables consideration of potential confounders, providing adjusted odds ratios as a measure of effect size. A confounding factor independently influences both with the phenotype and the genotype, but is not on the causal path between them.

To account for confounding factors, the linear and logistic models in (2.1) and (2.2) have to be extended by including such confounding factors as additional predictors in the model. We then have, in matrix form, the corresponding equations for the linear and logistic regression models as

$$y_i = W_i \boldsymbol{\alpha} + X_i \boldsymbol{\beta} + \epsilon_i, \quad \text{and} \\ P(y_i = 1) = \frac{\exp [W_i \boldsymbol{\alpha} + X_i \boldsymbol{\beta}]}{1 + \exp [W_i \boldsymbol{\alpha} + X_i \boldsymbol{\beta}]}, \quad (2.3)$$

where W is a $n \times q$ design matrix of covariates (for example, age, gender, etc), $\boldsymbol{\alpha}$ are covariate effects, X is a $n \times m$ design matrix of SNPs, $\boldsymbol{\beta}$ are SNP effects, and $\epsilon \sim N(\mathbf{0}, \sigma_\epsilon^2 I_n)$, I_n being a $n \times n$ identity matrix, is a vector of residuals.

The population-based design in GWAS provide attractive features for association mapping. In particular, the use of unrelated individuals makes it possible to obtain

large study sample sizes; large sample sizes are critical to attain the statistical power necessary to detect the underlying genetic effects, as the effects of single individual genetic loci that influence complex phenotypes are thought to be relatively small compared to the total phenotypic variability (O'Connor, 2019).

The recruitment of unrelated individuals, however, presents a pitfall. There may be population substructure in the study samples; the result of which is that an observed statistical association may be due to a true genetic effect but may also be due to confounding. Population substructure can broadly be used to refer to features of a population which changes the distribution of allele frequencies across individuals in a population, and comprise of population stratification, population admixture, and population inbreeding. Population stratification refers to the presence of multiple mutually exclusive subgroups in a population; the allele frequency being the same within each subgroup but varying across different subgroups. The subgroups are typically assumed to represent different racial, ethnic and/or geographic subgroups. Population admixture refers to the situation where individuals in a population have a mixture of different genetic ancestries due to the mixing of two or more populations at previous point(s) in time (Klein and Moeschberger, 1997). Generally, admixed populations results from a migration of one or more population groups from specific geographic regions into a different geographic location inhabited by a different population group. Population inbreeding, meanwhile, refers to the situation where individuals in a population mate among close relatives, either because of preference of the individuals to do so, or because geographic isolation restricts mating choices within the population subgroups.

Pertinent to GWAS, the presence of population substructure can influence both the distribution of genotypes in a population and the distribution of the test statistics, which is computed based on allele frequencies. In addition, it can influence the variance of the estimated effects. Furthermore, it may bias sample estimate if, for example, allele frequencies varies across population subgroups over which sample recruitment is done, and the sampling process does not take this information into consideration. The ultimate result of all these is that population substructure, of

whatever form, can lead to an apparent association between a loci and phenotype when in fact such an association does not exist.

The GWAS approach was originally intended and designed for application to homogeneous populations, that is, a population with unrelated (independent) individuals. Accordingly, genetic association frameworks based on the standard regression techniques, which rely on the assumption that the data have an independent and identically distributed property (i.i.d) would be optimal. In practice, however, GWAS samples is typically collected from heterogeneous populations with underlying population substructures. This problem becomes even more pronounced in today's era of big data where it becomes almost inevitable that the collected sample will contain individuals who are distantly related or from different population groups with different genetic background characteristics. Because of this, standard linear regression becomes unrealistic. Genetic relatedness, let alone non-independence of SNPs due to LD, means that the true effect of a single SNP is modulated by multiple other variants in the genome, and so the 'hypothetical' true genetic model must account for the effect of all SNPs. Thus, when testing for the effect of a particular SNP, say k , the true data generating model, from (2.3), should in fact be

$$y_i = W_i \boldsymbol{\alpha} + X_{ik} \beta_k + \sum_{j \neq k} X_{ij} \beta_j + \epsilon_i, \quad \text{and}$$

$$P(y_i = 1) = \frac{\exp \left(W_i \boldsymbol{\alpha} + X_{ik} \beta_k + \sum_{j \neq k} X_{ij} \beta_j \right)}{1 + \exp \left(W_i \boldsymbol{\alpha} + X_{ik} \beta_k + \sum_{j \neq k} X_{ij} \beta_j \right)}, \quad (2.4)$$

The term $\sum_{j \neq k} X_{ij} \beta_j$ in (2.4) accounts for the effects of SNPs in a genome other than the one being tested, and, in view of (2.3), can be regarded as a 'missing' or an 'unmodeled' factor. Accounting for the unmodeled factor, as done in (2.4), can mitigate bias and false positive in association results.

Unfortunately, population substructure is not the only unmodeled factor; there

are potentially many unmodeled factors that are unknown to us and, thus, cannot be accounted for (see [Figure 2.2](#)).

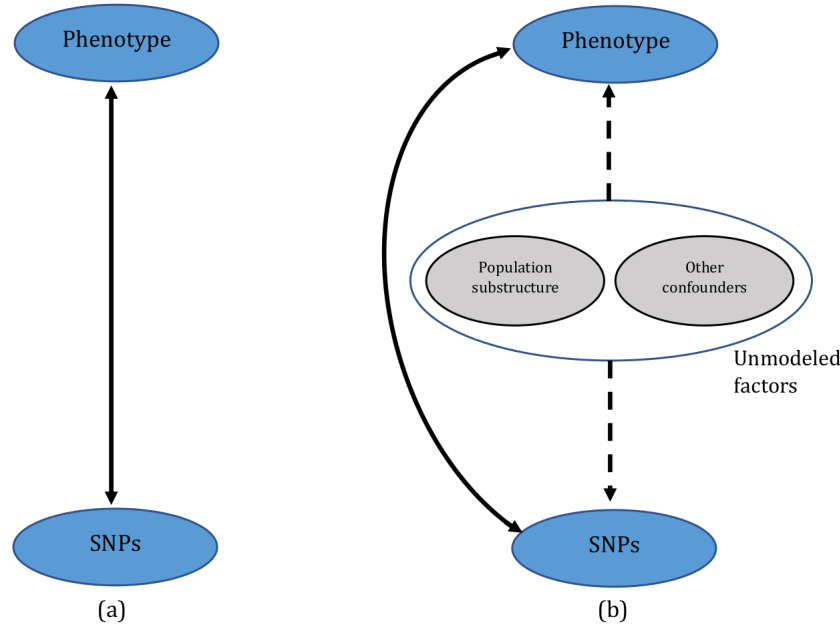


Figure 2.2: Schematic representation of the link between SNPs and phenotype. In (a) we assume there are no confounders. The null hypothesis of testing for the effect of a given SNP on the phenotype, then is that the SNP and phenotype are independent, and the alternative hypothesis is that the SNP does affect the phenotype. In (b) there are unmodeled confounders, which include population substructure or other possible confounders. In this case, the null hypothesis that SNP and phenotype are independent is no longer valid since there is an a priori correlation between SNPs and phenotype by the unmodeled factors.

Besides, the dependency structure among these factors, and the way they interact to influence host genetics and host phenotype is unknown. For the hypothesis-driven inference of association to be valid, the modelled factors have to be eliminated. To do this, we recognize the fact that the modelled factors act to influence specific SNPs, and therefore, while we cannot pinpoint the modelled factor on individual SNPs, we can use available knowledge about pairwise genetic similarities between individuals to perform estimation. This approach is valid since if two individuals have similar genomes then the causal variant in both individuals will share the same allele, and so the effect of the modelled factor will be identical in both individuals ([Sul et al., 2018](#)). The problem of confounding due to popula-

tion substructure, which includes its variants, family structure (the existence of related individuals with known family relationship (Cao et al., 2014)), and cryptic relatedness (where there are some closely related individuals but the common ancestry is unknown (Sul et al., 2018)) has been well recognised and appreciated in GWAS, and many statistical methods have been proposed to mitigate their deleterious effects.

One approach, for example, is to perform principal component analysis (PCA) to infer population substructure from datasets and use the inferred principal components, which captures the genetic ancestry of the individuals, as fixed effects-covariates in classical linear or logistic regression models. While appealing because of their genetic ancestry-informed correction and relative ease of implementation, this and related methods such as multidimensional scaling (MDS), are not capable of sufficiently correcting for the various forms of population substructure and genetic relatedness in empirical GWAS datasets. Perhaps the biggest limitations with all these methods based on the standard linear regression is their inability to account for correlation among the unmodeled factors. Along this line, the linear mixed modeling (LMM) approach has become the prevailing paradigm for association mapping due to its ability to effectively model nearly any form of genetic relationship among individuals, and doing so without the need for any prior explicit identification of existing relatedness or genetic ancestry.

A LMM, which is essentially an extension of the linear model to include random effects, models phenotypes using a mixture of fixed effects and random effects. The fixed effects in this case include the candidate SNP k and optional covariates such as age, gender and other relevant clinical covariates. The random effects, meanwhile, are based on the phenotypic covariance, and is modeled as a sum of genetic (heritable) and non-genetic (non-heritable) variations. Extending model (2.4) to the LMM framework, we have

$$\mathbf{y} = W\boldsymbol{\alpha} + X_k\beta_k + \mathbf{u} + \boldsymbol{\epsilon}, \quad (2.5)$$

where $Var(\mathbf{u}) = \sigma_g^2 K$. The term $\mathbf{u}$ is the random vector of genetic variations

across all SNPs, and is distributed according to the kinship matrix K . K is computed from the genotype data, where K_{ij} between individuals i and j is the product of the standardized genotypes for the two individuals, normalized by the number of SNPs. That is,

$$K_{ij} = \frac{1}{m} \sum_{k=1}^m X_{ik} X_{jk}, \quad (2.6)$$

where m is the number of SNPs. The kinship matrix K captures the effect of population substructure since population substructure act on genotypes, and K is defined according to pairwise genotypic similarity of the individuals in the study. Based on the form of the model in (2.5), the variance component σ_g^2 captures the extent to which individuals that are genetically similar are also phenotypically similar; thereby eliminating confounding effects (Price et al., 2010).

The estimate of β_k and the association are obtained by solving (2.5), which requires first estimating σ_g^2 and σ_e^2 . These variance components can be estimated using the maximum likelihood approach. The estimation is, however, a difficult optimization problem; an issue that has remained an impediment to the application of LMM in GWAS. Nevertheless, many efficient algorithms developed in recent years have made LMM remain the prevailing statistical method for GWAS.

2.5 Heritability in GWAS

Over a century ago, Weinberg (Weinberg, 1909), cognizant of the fact that phenotypic variation results from a combination of genetic and environmental factors, suggested methods of delineating genetic from environmental components of total phenotypic variability. This and subsequent works on statistical separation of environmental and genetic variation in general populations, culminated to the specification of the fraction of the total phenotypic variance due to genetic variance - a measure eventually termed ‘degree of heritability’ or simply ‘heritability’ in the genetic community (Lush, 1947). A distinction is, however, necessary be-

tween *total* (or *broad sense*) and *additive* (or *narrow sense*) heritability. The former measures the full contribution of genes, which includes additive, dominance and epistasis components, while the latter captures only the additive contribution of genes to phenotypic variance.

It is now known that many common human diseases and traits are complex, resulting from the joint effect of host genetic and environmental factors. Indeed, genome-wide association studies (GWAS), which assays hundreds to millions of genetic markers - commonly single nucleotide polymorphisms (SNPs) - in thousands of individuals, have uncovered hundreds of genetic variants associated with many common polygenic inherited diseases and traits; revealing scores of previously unknown key biological pathways, and providing valuable insights into the complexities of their genetic architecture (Hirschhorn et al., 2009; Zuk et al., 2012; Visscher et al., 2017). Despite this, however, GWAS has been puzzled by the apparent rather low proportion of the estimated heritability, which is far less than that obtained from familial studies - the difference being referred to as *missing* heritability. A classic, often cited, example is human height where whereas the estimated heritability is 80%, the (narrow sense) heritability estimate with tens of thousands of people is only about 45% (Yang et al., 2010). Many possible, and debated, explanations have been offered for this discrepancy, including sub-optimal sample size, poor detection of variants by genotyping arrays, dominance, epistasis, and shared environment [see references (Manolio et al., 2009; Zuk et al., 2012) for excellent reviews]. While many investigators have argued that a considerable part of the missing heritability may be attributed to non-additive effects such as dominance and epistasis, several recent empirical studies have found no strong effects from them (Zhu et al., 2015; Zaitlen et al., 2013; Nolte et al., 2017). A similar observation has been made for epigenetic effects. While epigenetic variation, including methylation has been suggested as another possible source of missing heritability (Slatkin, 2009), a recent study of body mass index found genetic predictors and methylation to be non-overlapping, suggesting the latter represented environmental effects on this phenotype; for human height, methylation profiles did not explain any variation (Nolte et al., 2017; Shah et al.,

2015).

The non-DNA factors inflates heritability estimates from twin-based studies that GWAS heritability estimates are compared to, and thereby inflating heritability estimates from the twin-based studies, or equivalently leading to inflation of the unexplained heritability gap. Likewise, the inherited non-DNA factors can also obscure estimation of true GWAS-heritability estimates (Bourrat et al., 2017). This corroborates the significance of non-DNA factors in understanding and estimating heritability in a GWAS.

2.5.1 Broad-sense and narrow-sense heritability

Heritability is a measure of the relative contribution of genetics to a phenotypic expression. Its estimation centers around the measure of variability, which makes sense only if the phenotype is quantitative. For categorical phenotypes, therefore, one typically postulates it to be resulting from some underlying quantitative (continuous) variable, often called *liability*, which has a *threshold* that defines the intervals corresponding to the different states of the categorical variable (Falconer, 1965). The basic idea of heritability estimation is simple, at least in theory: partition the variance of a phenotype into components attributable to the different factors that are known to affect the phenotype, and determine the ratio of the genetic variance component (assuming genetics modulates the phenotype) to the phenotypic variance.

Suppose a quantitative trait is modulated by its overall genotype G and the exposure environment E , where G can be partitioned into additive (σ_g^2), dominance (σ_D^2), and interaction (σ_I^2) components. That is,

$$\sigma_P^2 = \sigma_G^2 + \sigma_E^2 = \sigma_g^2 + \sigma_D^2 + \sigma_I^2 + \sigma_E^2,$$

where σ_I^2 may refer to additive-by-additive, dominance-by-dominance, additive-by-dominance as well as many higher-order interaction terms (Lynch et al.,

1998). Broad sense heritability H^2 is the ratio of the total genetic variance to the phenotypic variance: $H^2 = \frac{\sigma_G^2}{\sigma_P^2}$, and expresses the degree to which genotype determine phenotype of individuals. Narrow sense heritability h^2 , on the other hand, is the ratio of the additive genetic variance to the phenotypic variance: $h^2 = \frac{\sigma_g^2}{\sigma_P^2}$, and expresses the extent to which individual's phenotypes are determined by genes transmitted from parents. It therefore determines the degree of phenotypic resemblance between relatives, the observable genetic properties of a population and of the response of a population to external forces such as selection (Falconer et al., 1960). Broad sense heritability is of more theoretical interest than practical importance as it neither provide an understanding of the genetic properties of a population nor reveal the cause of phenotypic resemblance between relatives. Henceforth, as with all GWAS studies, we refer to narrow sense heritability (or simply heritability) in all subsequent discussions, unless stated otherwise.

2.5.2 Estimation of Heritability in GWAS

We are typically interested in the *explained* heritability - defined as the ratio of heritability explained by a set of variants known definitely to be associated with the trait to the heritability explained by all genetic variants [those known (discovered) plus those not yet discovered] that are associated with the trait: $h_{\text{explained}}^2 = h_{\text{known}}^2 / h_{\text{all}}^2$; missing heritability being defined as $h_{\text{missing}}^2 = 1 - h_{\text{explained}}^2$ (Zuk et al., 2012).

From a statistical genetics perspective, the additive variance σ_g^2 at a single locus is the genetic variance explained by the regression of the expected value of the phenotypic mean in each genotypic class on the genotype (Fisher, 1919; Zhu et al., 2015), or put differently, heritability is the coefficient of regression obtained from the regression of additive genetic effect on the the phenotype. Therefore, as we detail below, h_{known}^2 can be readily estimated from observed genotype-phenotype data using regression in a 'bottom-up' approach. The estimation of h_{all}^2 is, however, not straight forward because we do not know the complete repertoire of genetic variants associated with the trait; all we can do therefore is to infer

it from phenotypic correlations obtained from population data in a ‘top-down’ approach.

Given a GWAS, let $y_i, i \in \{1, \dots, n\}$ be the quantitative phenotypes measured on n individuals; $g_i = \{g_{i1}, \dots, g_{im}\}$ the genotype of the i^{th} individual for the m typed SNPs, with minor allele frequencies $p_j, j \in \{1, \dots, m\}$.

Employing the additive model, we have

$$y_i = \sum_{j=1}^m \beta_j z_{ij} + e_i,$$

where $z_{ij} = (g_{ij} - 2p_j) / \sqrt{2p_j(1 - p_j)}$ is the normalized genotype. and y is normalized to have mean 0 and variance 1.

If S is the subset of statistically-associated (assumed here to be causal) variants obtained from the GWAS, then, as the phenotype is normalized (i.e. $\sigma_P^2 = 1$), the additive variance is $\sigma_g^2 = \sum_{j \in S} \beta_j^2$ and the heritability is $h_{\text{known}}^2 = h_{\text{GWAS}}^2 = \sigma_g^2 / \sigma_P^2 = \sum_{j \in S} \beta_j^2$, the sum of squared effect sizes for the normalized genotypes over the causal variants (Zuk et al., 2012).

It is important to note, however, that the full set of causal variants are unknown; that is, the causal variants identified by the GWAS here is only a subset of causal variants. Consequently, h_{GWAS}^2 represents only the lower bound of the true heritability h_{all}^2 . The difference between h_{GWAS}^2 and h_{all}^2 is termed the *missing* heritability of the phenotype. This difference can be attributed to several factors. First, the non-additive genetic variance such as epistasis that is not included in estimation of heritability; the presence of such non-additive variations have been shown to inflate heritability estimates (Eichler et al., 2010; Zuk et al., 2012). Second, the exclusion of causal variants due to, say stringent GWAS significance threshold or low effect sizes, for example, can lead to underestimation of heritability. Likewise, false positive results would inflate observed estimates.

In practice, having estimated h_{GWAS}^2 , it is often of interest to know the proportion of explained heritability. In other words, we would like answer the question: *what proportion of heritability do all SNPs that contribute to the trait explain?* This question requires us to estimate h_{all}^2 .

The methodological estimation of h_{all}^2 is a ‘top-down’ approach that hinges on recognizing the equivalence between then classical definition of heritability [$h^2 = \sigma_g^2 / \sigma_P^2$] and the intuitive interpretation of the proportion of phenotypic variance explained by all causal variants [$\sigma_g^2 / (\sigma_P^2 = \sigma_g^2 + \sigma_\epsilon^2)$].

If

$$\mathbf{y} = w + g + \epsilon, \quad (2.7)$$

where w denote the fixed effects (including candidate SNP and optional covariates), g denote genetic effects assumed to subsume any genetic effects on the trait other than at candidate SNP, $\epsilon = N(0, I\sigma_\epsilon^2)$; I being the identity matrix, then by treating $g = X\beta$ as a random effect with $g \sim N(0, \sigma_g^2 A)$, the variance-covariance matrix of $\mathbf{y}$ can be expressed as

$$\text{Cov}(\mathbf{y}) = \frac{XX'}{m} \sigma_g^2 + I\sigma_\epsilon^2 = A\sigma_g^2 + I\sigma_\epsilon^2, \quad (2.8)$$

where A is the kinship (genetic relationship) matrix between pairs of individuals, defined over all causal loci (Yang et al., 2010), m is the number of causal variants, and σ_g^2 is the variance explained by the SNPs. Since A is defined over *all causal loci*, σ_g^2 denotes the *variance explained by all causal SNPs*.

The parameters to be estimated are σ_g^2 and σ_ϵ^2 , which can be done by using (2.8) in (2.7) and solving my optimization using the restricted maximum likelihood (REML) method. With these obtained, $\sigma_P^2 = \sigma_g^2 + \sigma_\epsilon^2$ can subsequently be obtained. In theory, this appears a trivial task. In practice, however, this is not the case because the estimation requires that we first obtain the matrix A , defined at the causal SNPs, but we do not know the causal SNPs. The traditional, and still used, approach involves using genetically related individuals from known pedigrees

(family/twins) to estimate a kinship coefficient Φ , where, for example, Φ_{ij} is taken as 0.25 for siblings and 0.5 for twins; A is then taken to be equal to 2Φ (Lange, 2003; Sandoval-Motta et al., 2017b). Clearly, this assumption does not necessarily hold since phenotypic resemblance may be influenced by other heritable factors, other than genotype; for example epigenetic modifications and the host's microbiome. Indeed, the true covariance has been observed to vary around this assumed value (Visscher et al., 2006). Consequently, this can lead to inflation of the corresponding estimated heritability.

With the unveiling of genotype data, prodded by the advent of next generation sequencing, methods have been devised to estimate A from genotype data of unrelated individuals. This is done (De Los Campos et al., 2010; Yang et al., 2011) by postulating that the ungenotyped causal SNPs are tagged by the genotyped ones, and therefore although the set of causal variants is unknown, one can use all the SNPs genotyped in GWAS to estimate A , and use it as proxy for A_{causal} . The key point to note here is that all genome-wide SNPs are used, not only the genome-wide significant SNPs. Indeed, this methodology proceeds without conducting any test of association between individual SNPs and the phenotype. The genetic relationship between individuals i and j is estimated using standardized genotype by

$$A_{ij} = \frac{1}{m} \sum_{k=1}^m \frac{(g_{ik} - 2p_k)(g_{jk} - 2p_k)}{2p_k(1 - p_k)},$$

where p_k is the allele frequency of the k^{th} SNP. With A , defined at all causal SNPs, now obtained, (2.7) can be solved to estimate σ_g^2 and σ_ϵ^2 , from where the variance heritability explained by all causal SNPs (h_{all}^2) can be determined from $h_{\text{all}}^2 = \sigma_g^2 / (\sigma_g^2 + \sigma_\epsilon^2)$.

In the estimation of both h_{all}^2 and h_{GWAS}^2 above, it is assumed that all causal SNPs have been genotyped or at least, are tagged by the genotyped SNPs. Accordingly, if the SNP array used does not fully cover the set of common genetic variants in the GWA study population, the resulting heritability estimates are likely to be

smaller than the actual value, however large the sample size maybe.

Finally, it must be noted because the set of causal variants are unknown and we have to rely on SNPs being tagged by LD, and yet LD strength declines with increasing difference in minor allele frequency (MAF) between SNPs (Wray, 2005; Yang et al., 2017), some causal variants in the low frequency spectrum may not be tagged; and, as a result will not play part in heritability estimates. Moreover, MAF of a disease allele can be population-specific (Klein and Moeschberger, 1997; Linck and Battey, 2019). Because of this, heritability estimate can be population specific. That said, the above model represents the general framework for heritability estimation from genotype data in a GWAS. Owing to the increasing availability of GWAS summary statistics (results from statistical tests of association in a GWAS), methods have been devised that extend the framework to estimate heritability from summary statistics data (Speed et al., 2020; Bulik-Sullivan et al., 2015; Speed et al., 2012).

2.6 GWAS Findings and Implications for Disease Scoring Statistics

2.6.1 GWAS Findings

The GWAS approach has been very successful in identifying genetic modifiers of common human traits and diseases. Since the first GWAS publications by Klein et al. (2005) and DeWan et al. (2006), both of which found common variants associated with age-related macular degeneration, the approach has uncovered 179,364 associations (as of 08 March, 2020) with a diverse array of complex phenotypes, ranging from diseases and syndromes (such as cancer, diabetes, cardiovascular disease, depression, and arthritis) to molecular phenotypes (such as total cholesterol, and glucose levels) (<https://www.ebi.ac.uk/gwas/>). While GWAS discovery rely on SNP tagging which means that statistical association do not nec-

essarily imply causal association requiring further functional studies to validate the signal, the discoveries so far made from GWAS has been nothing short of outstanding. This success has been driven by several joint efforts, but most notably (i) the development of micro-array technologies for massive parallel genotyping coupled with falling costs of genotyping, allowing studies of large sample sizes, (ii) the cataloging of human genetic variation by the International HapMap project (Consortium et al., 2003), that made it possible, owing to SNP tagging, to examine SNPs across the entire genome based on only a few million *representative* SNPs, making GWAS more efficient and cost effective, (iii) international collaborations, allowing high-powered discovery studies, and (iv) the development of robust and efficient statistical analysis tools. Altogether, these efforts have provided key advances in our understanding of the genetic underpinnings of diseases, offering insight and opportunity for drug development and stratification of patients for common diseases.

In drug discovery, for example, genomic discoveries have led to important new treatments. Genetic association studies have demonstrated that ‘knock-out’ mutations in the *PCSK9* gene results in lower levels of LDL cholesterol and a reduced risk of coronary heart disease (Breen et al., 2016). This has led to the development of a new class of lipid-lowering drugs for cardiovascular disease (Robinson et al., 2015). Similarly, loss-of-function mutations in *SLC30A8* gene has been confirmed to reduce the risk of type 2 diabetes (Pearson, 2014), leading to promising clinical therapies (Norouzi et al., 2017). Outside drug discovery and repositioning, GWA findings are now giving prospectus for stratified screening: identification of individuals at increased risk on the basis of their genotype. This can be exemplified by breast cancer (BC), where risk SNPs can be used to conduct risk discrimination to a degree of accuracy useful for population-based screening programmes (Sud et al., 2017). For example, in a recent study by Khera et al. (2018), a polygenic risk score (PRS) model for breast cancer achieved a discriminatory power of up to 69%, and would identify 1.5% of the population at ≥ 3 -fold risk.

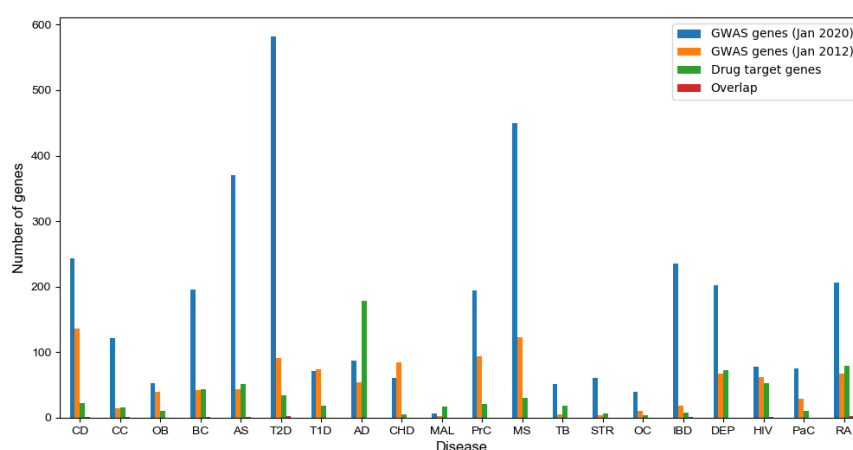
Where are we now?

Beyond these anecdotal examples, how far (or short) have we gone with GWAS of common complex human traits and diseases? To answer this question, we first assess, in the section below, the progress GWAS has so-far made in discovering drug targets for complex human traits and diseases. To this end, we extend a related earlier work by [Cao and Moulton \(2014\)](#), using updated data. We consider, as example, the common complex diseases: Crohn's disease, Colorectal cancer, Obesity, Breast cancer, Asthma, Type 2 diabetes, Type 1 diabetes, Alzheimer's disease, Coronary heart disease, Malaria, Prostate cancer, Multiple sclerosis, Tuberculosis, Stroke, Ovarian cancer, Inflammatory bowel disease, Depression, HIV/AIDS, Pancreatic cancer, and Rheumatoid arthritis, and examine the results of GWAS published to date, retrieved from the GWAS catalog (<http://www.genome.gov/gwastudies/>; accessed Jan 30 2020). The GWAS catalog is a comprehensive database of published GWA studies and results for different complex traits and disease. For each disease, we extract the full set of GWAS-reported genes.

Then, for each disease, we compare the GWAS-reported genes with the set of known drug targets (genes annotated as having known pharmacological action ([Nguyen et al., 2019a](#)) and thus expected to be found by GWAS). The drug targets were obtained from the Drugbank database (<http://www.drugbank.ca>; accessed Jan 30 2020). The database contains, for each disease, detailed drug information including chemical, pharmacological, and drug target information (biological sequence, structure, and pathway). To contextualize the GWAS findings viz a viz overlap with drug targets, we juxtapose the GWAS findings of 2012 with that of today (2020), for the same phenotypes.

The results are displayed in [Figure 2.3](#). It can be observed that while the number of genes discovered through GWAS have increased substantially over the years, the overlap of GWAS-discovered genes with known drug target genes is still very little. This may be attributed to several factors, including misalignment of

causal genes in the GWAS catalogue (Cao and Moul, 2014), and incorrectness of drug targets in Drugbank given that the targets for some drugs with no known mechanisms are probabilistically-determined (Cao and Moul, 2014; Overington et al., 2006; Zeng et al., 2020). Of course, this cannot be taken to mean that GWAS-discovered genes should necessarily converge to include the known drug targets, since some known drug targets may, in fact, be off-gene targets and their apparent effect being the result of location on off-target proteins contributing to the phenotypic expression. Therefore, GWAS results can help locate more disease-specific targets. That said, the results obtained here agree with those of previous studies that found a small level of overlap GWAS-reported disease genes and the corresponding drug targets (Cao and Moul, 2014; Sanseau et al., 2012; Yin et al., 2018).



CD=Crohn's disease, CC=Colorectal cancer, OB=Obesity, BC=Breast cancer, AS=Asthma, T2D=Type 2 diabetes, T1D=Type 1 diabetes, AD=Alzheimer's disease, CHD=Coronary heart disease, MAL=Malaria, PrC=Prostate cancer, MS=Multiple sclerosis, TB=Tuberculosis, STR=Stroke, OC=Ovarian cancer, IBD=Inflammatory bowel disease, DEP=Depression, HIV=HIV/AIDS, PaC=Pancreatic cancer, RA=Rheumatoid arthritis.

Figure 2.3: Comparison of genes discovered by GWAS with known drug targets for common complex traits/diseases.

To provide biological context to the results, we perform pathway and gene set enrichment analysis; testing whether biologically relevant pathways or sets of genes are enriched for genetic variants associated with these phenotypes.

Utilizing MAGMA (Multi-marker Analysis of GenoMic Annotation) and leveraging curated biological pathways (Molecular Signatures Database - MSigDB - GO biological processes) and the GWAS summary statistics data for Type 2 Diabetes (T2D) (from DIAGRAM-<https://www.diagram-consortium.org/downloads.html>, accessed April 12 2021 ;(Mahajan et al., 2014)), Inflammatory Bowel Disease (IBD) (from [xx-https://www.ibdgenetics.org/downloads.html](https://www.ibdgenetics.org/downloads.html), accessed April 12 2021 ; (Hong et al., 2018)), for example, we perform competitive gene-set analysis, which tests whether the genes in each gene set are more strongly associated with the phenotype than genes outside the gene set. For the MAGMA software, we used the data of 1000 Genome European panel as reference for the LD information between SNPs, and employed the Bonferroni correction to adjust the P values.

The primary enrichment identifies several gene-sets that are enriched for T2D- and IBD- associated SNPs (Table 2.1 & 2.2). The top-ranked significant pathways were *GOBP L LYSINE TRANSMEMBRANE TRANSPORT* ($P=5.661e-05$), *GOBP URATE METABOLIC PROCESS* ($P=3.281e-04$), and *GOBP EMBRYONIC SKELETAL SYSTEM MORPHOGENESIS* ($P=3.615e-04$) for IBD, and *GOBP NEGATIVE REGULATION OF RECEPTOR SIGNALING PATHWAY VIA STAT* ($P=1.422e-08$), *GOBP NEGATIVE REGULATION OF RECEPTOR SIGNALING PATHWAY VIA JAK STAT* ($P=1.983e-06$), and *GOBP REGULATION OF RECEPTOR SIGNALING PATHWAY VIA STAT* ($P=1.457e-05$) for T2D. Although the overlap of GWAS discovered genes with the drug targets is weak, the detected significant enrichment of T2D- and IBD- associated SNPs in biological processes suggests their potential widespread role in biological processes.

Table 2.1: Results for enrichment analysis of Inflammatory Bowel Disease-associated genes: NGENES=number of genes, STD=standard deviation, SE=standard error, $p=p$ -value. Results presented for the top 20 significant associations.

VARIABLE	TYPE	NGENES	BETA	BETA_STD	SE	P	FULL_NAME
GOBP_L_LYSINE_TRANSMEMBRANE...	SET	4	2.059	0.054	0.533	0.000	GOBP_L_LYSINE_TRANSMEMBRANE_TRANSPORT
GOBP_URATE_METABOLIC_PROCESS	SET	4	1.588	0.041	0.466	0.000	GOBP_URATE_METABOLIC_PROCESS
GOBP_EMBRYONIC_SKELETAL_SYST...	SET	35	0.602	0.046	0.178	0.000	GOBP_EMBRYONIC_SKELETAL_SYSTEM_MORPHOGENESIS
GOBP_NEGATIVE_REGULATION_OF_...	SET	2	2.163	0.040	0.646	0.000	GOBP_NEGATIVE_REGULATION_OF_MAST_CELL_DEGRANULATION
GOBP_REGULATION_OF_SYNAPTIC_...	SET	14	0.954	0.047	0.287	0.000	GOBP_REGULATION_OF_SYNAPTIC_VESICLE_RECYCLING
GOBP_POSITIVE_REGULATION_OF_...	SET	6	1.520	0.049	0.468	0.001	GOBP_POSITIVE_REGULATION_OF_ENDOPLASMIC_RETICULUM_UNFOLDED_PROTEIN_RESPONSE
GOBP_URATE_TRANSPORT	SET	3	1.676	0.038	0.517	0.001	GOBP_URATE_TRANSPORT
GOBP_CHONDROITIN_SULFATE_PRO...	SET	9	1.161	0.045	0.359	0.001	GOBP_CHONDROITIN_SULFATE_PROTEOGLYCAN_BIOSYNTHETIC_PROCESS
GOBP_ARGININE_TRANSPORT	SET	7	1.228	0.042	0.385	0.001	GOBP_ARGININE_TRANSPORT
GOBP_ESTROGEN_BIOSYNTHETIC_P...	SET	4	1.527	0.040	0.484	0.001	GOBP_ESTROGEN_BIOSYNTHETIC_PROCESS
GOBP_CHONDROITIN_SULFATE_BIO...	SET	8	1.143	0.042	0.365	0.001	GOBP_CHONDROITIN_SULFATE_BIOSYNTHETIC_PROCESS
GOBP_ESTROGEN_METABOLIC_PRO...	SET	12	0.957	0.043	0.306	0.001	GOBP_ESTROGEN_METABOLIC_PROCESS
GOBP_GLYCOXYLATE_METABOLIC_P...	SET	4	1.654	0.043	0.533	0.001	GOBP_GLYCOXYLATE_METABOLIC_PROCESS
GOBP_VIRAL_RNA_GENOME_REPLIC...	SET	10	1.021	0.042	0.345	0.002	GOBP_VIRAL_RNA_GENOME_REPLICATION
GOBP_NEGATIVE_REGULATION_OF_...	SET	2	2.187	0.040	0.742	0.002	GOBP_NEGATIVE_REGULATION_OF_MEMBRANE_DEPOLARIZATION
GOBP_GLAND_DEVELOPMENT	SET	148	0.237	0.037	0.081	0.002	GOBP_GLAND_DEVELOPMENT
GOBP_POSITIVE_REGULATION_OF_...	SET	7	1.188	0.041	0.413	0.002	GOBP_POSITIVE_REGULATION_OF_CALCIUM_ION_DEPENDENT_EXOCYTOSIS
GOBP_ARGININE_TRANSMEMBRANE...	SET	6	1.253	0.040	0.436	0.002	GOBP_ARGININE_TRANSMEMBRANE_TRANSPORT
GOBP_POSITIVE_REGULATION_OF_...	SET	4	1.572	0.041	0.554	0.002	GOBP_POSITIVE_REGULATION_OF_INTERLEUKIN_5_PRODUCTION
GOBP_REGULATION_OF_PROTEIN_L...	SET	3	1.728	0.039	0.616	0.003	GOBP_REGULATION_OF_PROTEIN_LOCALIZATION_TO_CILJUM

Table 2.2: Results for enrichment analysis of Type 2 diabetes-associated genes: NGENES=number of genes, STD=standard deviation, SE=standard error, $p=p$ -value. Results presented for the top 20 significant associations.

VARIABLE	TYPE	NGENES	BETA	BETA_STD	SE	P	FULL_NAME
GOBP_NEGATIVE_REGULATION_OF_...	SET	22	0.99266	0.03491	0.17875	0.00000	GOBP_NEGATIVE_REGULATION_OF_RECEPTOR_SIGNALING_PATHWAY_VIA_STAT
GOBP_NEGATIVE_REGULATION_OF_...	SET	21	0.83924	0.02883	0.18186	0.00000	GOBP_NEGATIVE_REGULATION_OF_RECEPTOR_SIGNALING_PATHWAY_VIA_JAK_STAT
GOBP_REGULATION_OF_RECEPTOR...	SET	91	0.35928	0.02564	0.08593	0.00001	GOBP_REGULATION_OF_RECEPTOR_SIGNALING_PATHWAY_VIA_STAT
GOBP_CARDIAC_CELL_FATE_COMMIT...	SET	12	1.23470	0.03207	0.29647	0.00002	GOBP_CARDIAC_CELL_FATE_COMMITMENT
GOBP_SNO_S_RNA_METABOLIC_PRO...	SET	12	0.77610	0.02016	0.19882	0.00005	GOBP_SNO_S_RNA_METABOLIC_PROCESS
GOBP_REGULATION_OF_MULTICELL...	SET	57	0.42109	0.02381	0.10988	0.00006	GOBP_REGULATION_OF_MULTICELLULAR_ORGANISM_GROWTH
GOBP_RECEPTOR_SIGNALING_PATH...	SET	151	0.24616	0.02260	0.06595	0.00010	GOBP_RECEPTOR_SIGNALING_PATHWAY_VIA_STAT
GOBP_ENERGY_RESERVE_METABOL...	SET	75	0.36215	0.02348	0.09911	0.00013	GOBP_ENERGY_RESERVE_METABOLIC_PROCESS
GOBP_REGULATION_OF_TELOMERAS...	SET	44	0.42178	0.02096	0.11747	0.00017	GOBP_REGULATION_OF_TELOMERASE_ACTIVITY
GOBP_POSITIVE_REGULATION_OF_...	SET	50	0.42762	0.02265	0.11972	0.00018	GOBP_POSITIVE_REGULATION_OF_CALCIUM_ION_TRANSPORT_INTO_CYTOSOL
GOBP_MULTICELLULAR_ORGANISM_...	SET	121	0.24918	0.02049	0.07147	0.00025	GOBP_MULTICELLULAR_ORGANISM_GROWTH
GOBP_CELLULAR_GLUCAN_METABOL...	SET	64	0.36276	0.02173	0.10620	0.00032	GOBP_CELLULAR_GLUCAN_METABOLIC_PROCESS
GOBP_POSITIVE_REGULATION_OF_...	SET	5	1.22190	0.02049	0.36124	0.00036	GOBP_POSITIVE_REGULATION_OF_INHIBITORY_POSTSYNAPTIC_POTENTIAL
GOBP_REGULATION_OF_CALCIUM_I...	SET	87	0.30514	0.02130	0.09101	0.00040	GOBP_REGULATION_OF_CALCIUM_ION_TRANSPORT_INTO_CYTOSOL
GOBP_OXIDATIVE_DNA_DEMETHYLA...	SET	6	1.11590	0.02050	0.33527	0.00044	GOBP_OXIDATIVE_DNA_DEMETHYLATION
GOBP_SINOATRIAL_NODE_DEVELOP...	SET	6	1.25800	0.02311	0.37852	0.00045	GOBP_SINOATRIAL_NODE_DEVELOPMENT
GOBP_INTESTINAL_HEXOSE_ABSOR...	SET	6	0.93798	0.01723	0.28373	0.00047	GOBP_INTESTINAL_HEXOSE_ABSORPTION
GOBP_CELLULAR_CARBOHYDRATE_M...	SET	253	0.17275	0.02047	0.05230	0.00048	GOBP_CELLULAR_CARBOHYDRATE_METABOLIC_PROCESS
GOBP_RESPONSE_TO_HEAT	SET	149	0.21060	0.01920	0.06421	0.00052	GOBP_RESPONSE_TO_HEAT
GOBP_REGULATION_OF_ANTIGEN_P...	SET	5	1.19480	0.02004	0.36439	0.00052	GOBP_REGULATION_OF_ANTIGEN_PROCESSING_AND_PRESENTATION_OF_PEPTIDE_ANTIGEN

In sum, the results here coupled with existing literature paint a clear picture of the current clinical utility of GWAS findings. While appreciable strides have been made, current progress seems to amount to only the ‘tip of the ice-berg’. Much remains to be understood about underlying aetiological processes before clinical significance can be gleaned from GWAS disease-implicated loci.

Disease risk prediction comes as another major application of GWAS. As GWAS progressively unveils disease susceptibility loci, the prospectus of using these findings to predict risk of disease at the individual level stimulates interest. Un-

fortunately, to date, beyond a few sporadic cases of population-level stratification, genetic variants-based risk prediction models do not achieve satisfactory levels of performance in identifying individuals at elevated risk of a disease; a scenario that is shared for GWAS of nearly all complex traits. While disappointing, these findings could also portend major untapped opportunities and directions for GWAS. In the sequel we summarize the reasons for the apparent low predictive power, and provide specific examples and justifications for how, instead of using host genetic information in isolation, leveraging different sources of information relevant to phenotypic variation may leapfrog the practical utility of GWAS findings in risk prediction at population and individual levels.

There are two fundamental approaches to disease risk prediction. The first approach, which is based on the polygenic and infinitesimal genetic model of allelic architecture - in which the phenotype is assumed to be influenced by numerous alleles of modest effects - proceeds by simultaneously fitting the genotypes at all causal variants. In essence, the risk prediction method utilizes similar SNPs data as standard genome-wide association studies, but differs from it in that while the chief aim of GWAS is detection of disease-associated variants, the chief aim of risk prediction is risk prediction in individuals, and, as a result, clarity of causality for the variants is not a prerequisite to obtaining a good predictive power (Fallin et al., 2016; Abraham and Inouye, 2015). Indeed, although the predictive power of a genomic risk prediction mainly comes from strongly associated SNPs, risk prediction models constructed using all variants (including the ones that have not reached genome-wide significance threshold) have overwhelmingly been shown to outperform those constructed using only genome-wide significant variants (Gola et al., 2020; Vassy et al., 2014). The reason for this should be apparent when you consider that the stringent genome-wide significance threshold filters out variants whose individual effects are modest but joint effects can be substantial. That said, this set of predictions models applies more rigorous statistical and machine learning frameworks, including l_1 -penalized logistic regression (Wei et al., 2013; Abraham et al., 2013), Support Vector Machines (Di Noia et al., 2020; Mittag et al., 2012), and Bayesian methods (Lloyd-Jones et al., 2019), offering the

potential to account for random effects and correlations among variants, albeit with increased computational cost.

The second approach, in contrast to the first one that requires individual-level genotype data, relies on summary statistics from discovery GWAS, and proceeds by aggregating the risk alleles (weighted by the effects size for association of the variants with the disease) in an individual into a ‘risk score’, commonly referred to as *polygenic risk score* (PRS) (Levin and Rader, 2020). The premise of the PRS method is that the risk SNPs are independent of each other, and their effects is additive; albeit recent applications correct for non-independence due to linkage disequilibrium by eliminating SNPs that in high LD with a candidate SNP (Bogdan et al., 2018). In recent years, PRS has emerged as a method of choice for genetic risk prediction because of its simplicity and computational efficiency (Song et al., 2020).

Both approaches have been utilized in disease risk predictions; these have shown that genotype-based risk predictors can achieve a substantial discriminatory power, providing proof-of-concept for its power as genetic-based risk prediction method. Mavaddat and colleagues (2019), using a large dataset comprised 94,075 case subjects and 75,017 control subjects, gleaned from 67 studies published on breast cancer, found that PRS with 313 SNPs achieved remarkable performance in predicting individual’s lifetime risk for overall breast cancer and ER-positive and ER-negative breast cancer sub-types; with those in the top 1% of risk having a 4.37- and 2.78-fold risks, while those in the bottom 1% of risk having a 0.16- and 0.27-fold risks, compared with those in the middle quantile (Mavaddat et al., 2019).

However, despite the gratifying promise, genotype-based risk predictors have generally still had modest success; an observation shared across many complex diseases. These generally show that despite the several novel connections made between genes and diseases, using genetic-derived information alone do not provide sufficient discriminatory power to predict disease susceptibility. For instance, a risk prediction model constructed using 62 known type 2 diabetes

(T2D)-associated loci showed no edge over the alternative model using only clinical risk factors (Vassy et al., 2014). Along this line, it has been asserted that in view of the highly polygenic nature of complex traits, genomic risk profile should not be used as a basis for risk prediction but rather as an addition information to clinical information (Chen et al., 2017; Floyd and Psaty, 2016).

The decoupling between mechanistic GWAS discovery and predictive performance of discovered variants springs from several factors that include (i) limitations in statistical models for prediction, as for example, modelling LD between markers (Vilhjálmsen et al., 2015), and, for the case phenotypes with overlapping clinical states e.g. psychiatric disorders, modelling genetic correlation among the traits (Turley et al., 2018), (ii) limitations in the discovery of causative loci by GWAS; an issue also determined by factors including sample size and allelic architecture, which also translate to incorrect estimates of marker effects (iii) the non-generalizability of estimated polygenic risk scores due to ancestry- and population-specific dependence of the risk scores.

Taken together, the foregoing examples and discussion shows that although GWAS has borne much fruit for complex trait genomics, it is still a long journey before we can be able to fully characterize the extent and mechanisms of genetic contribution to human complex traits, for clinical and epidemiological relevance.

2.6.2 Implications for Disease Scoring Statistics

As discussed above, our understanding of disease aetiology is far from complete. This is mainly because most heritable common human diseases have complex genetic architecture, involving locus and allelic heterogeneity (mutations in different genes, and different allele on the same gene), epistasis (genetic interactions between loci), and environmental modification, and other phenomena such as gene-environment interaction. This complexity means that the majority of disease-causing genetic variants will carry weak to subtle effect sizes (Weinberger, 2002); the modest effect sizes being consistent with the large-effects variants could be

subject to selection pressure (Hindorff et al., 2011; Pritchard and Cox, 2002). From the perspective of disease scoring statistics (any statistical model to link phenotypes to its modifying factors), the complexity makes it challenging to prioritize delineate causal genetic markers and estimate the true effect sizes. Indeed, some of the GWAS results have been deceptive because of potential confounding that may occur (Moore and Williams, 2009; Spitz et al., 2008).

In the face of these challenges, a number of strategies have been pursued. These efforts range from re-formulating the conceptual view and methods for studying variants at the high and low frequency spectra, and redesigning disease scoring statistics to be consistent with empirical observations of disease architecture. Different study designs have been utilized for identifying high frequency (common) and low frequency (rare) variants. Away from study design, the theoretical foundation coupled with the inherent assumptions under which a disease scoring statistics (DSS) - any application of statistical methods to any biological data to identify and characterize the factors underlying the disease - is designed is of astounding importance. For instance, while the use of linear model and the infinitesimal genetic architecture are ubiquitous in association mapping frameworks, the linear aspect means all, and potentially important, non-additive marker effects will be missed, whereas the infinitesimal model may not reflect biological reality (Zhang, 2015). Nonetheless, they provide a tractable yet satisfactory way to study genotype-phenotype relationships.

Since these plethora of factors do not act in isolation and can be stochastic, they do present a dilemma for DSS. In a previous article (in the sequel in this thesis), we examined the challenges faced by current DSS methods in the pursuit to understand the mechanism underlying complex traits, and proposed how leveraging other sources of information, particularly the host microbiome data, can provide a more complete picture of not only the basis of phenotypic variation but also disease/trait pathogenesis.

Tantalizing dilemma in risk prediction from disease scoring statistics

Denis Awany, Imane Allali and Emile R. Chimusa 

Corresponding author: Emile R. Chimusa, Division of Human Genetics Level 3, Wernher and Beit North Institute of Infectious Disease and Molecular Medicine Faculty of Health Sciences University of Cape Town Observatory, 7925 South Africa. Tel.: (+27 21)406-6425; Fax: (+27 21) 650-2010; E-mail: emile.chimusa@uct.ac.za

Abstract

Over the past decade, human host genome-wide association studies (GWASs) have contributed greatly to our understanding of the impact of host genetics on phenotypes. Recently, the microbiome has been recognized as a complex trait in host genetic variation, leading to microbiome GWAS (mGWASs). For these, many different statistical methods and software tools have been developed for association mapping. Applications of these methods and tools have revealed several important findings; however, the establishment of causal factors and the direction of causality in the interactive role between human genetic polymorphisms, the microbiome and the host phenotypes are still a huge challenge. Here, we review disease scoring approaches in host and mGWAS and their underlying statistical methods and tools. We highlight the challenges in pinpointing the genetic-associated causal factors in host and mGWAS and discuss the role of multi-omic approach in disease scoring statistics that may provide a better understanding of human phenotypic variation by enabling further system biological experiment to establish causality.

Key words: genome-wide association study; microbiome; multi-omics; microbiome GWAS

Introduction

Our understanding of the diversity of the human genome has improved considerably in the past decade with the advances in high-throughput sequencing technology [1]. These high-throughput technologies have also revolutionized disease scoring statistics (DSS) approaches. DSS refer to the application of statistical methods to any 'omic' data to identify and characterize the factors underlying human phenotypic variation (Figure 1). Leveraging on the huge technological developments, DSS have led to the identification of various factors that shape the diversity of the human genome at individual, family and population levels. DSS provided deeper insights into the basis of phenotypic variation, including human appearance, disease susceptibility/resistance, disease severity and response to treatment [2]. In recent years, studies have also corroborated

the role of the microbiome on human phenotypic variation, and the microbiome has emerged as a complex trait in human variation [3, 4]. These studies have shown that the microbiome is intimately involved in the interplay between health and disease [5, 6]. For example, alterations in the composition of the gut microbiome, also known as dysbiosis, are now known to be associated with many complex diseases such as inflammatory bowel disease, cancer and autoimmune disorders [7]. Altogether, these remarkable discoveries from human genome-wide association study (GWAS) and microbiome GWAS (mGWAS) have raised enthusiasm among researchers to conduct research to obtain a broader understanding of human genetic architecture, particularly as each new DSS approach continues to reveal novel biomarkers for disease phenotypes.

Although we know that the taxonomic composition and relative abundance of the microbiome is associated with host

Awany Denis received his MSc degrees from Makerere University and African Institute for Mathematical Sciences. He is a PhD student at the University of Cape Town, South Africa.

Imane Allali received her PhD degree in bioinformatics from Mohammed V University in Rabat, Morocco. She is a postdoctoral fellow at the Computational Biology Division, University of Cape Town, South Africa.

Emile R. Chimusa received his PhD degree in bioinformatics from University of Cape Town. He is a senior lecturer at the Division of Human Genetics, Department of Pathology, University of Cape Town, South Africa.

© The Author(s) 2019. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oup.com

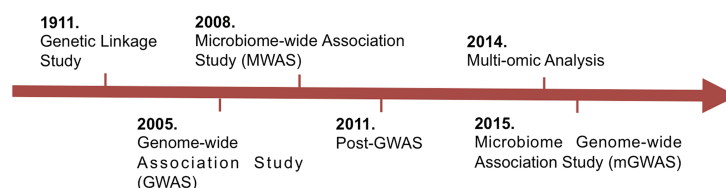


Figure 1. Development of various DSS.

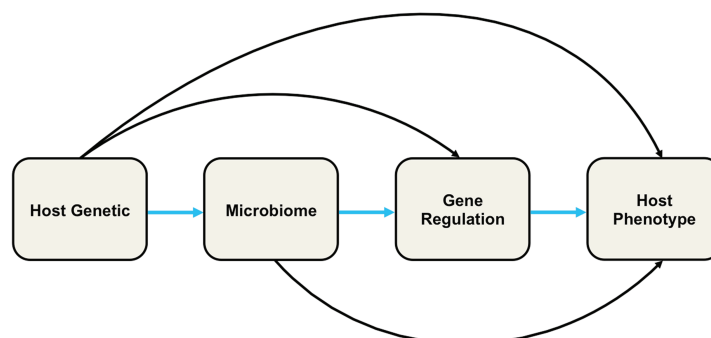


Figure 2. Illustrative representation of host GWAS, mGWAS and the integrative host-microbiome study approach.

genetics and host phenotypes, little is known about the associated causative genetic factor; the precise molecular mechanisms underlying the expression of a given phenotype, is at best, poorly understood [3, 5]. Recent studies have suggested that the microbiome impacts host phenotype through alteration of the host gene regulation in the interfacing host epithelial cells or modification of open chromatin status in the intraepithelial lymphocytes [8]. The genetic risk factors in the observed relationship between host genetic polymorphisms, the host's phenotypic expression and the microbiome composition are of fundamental biological and medical interest as a 1st step toward causality. DSS can play major roles toward dissecting associated genetic factors by helping to identify and prioritize likely associated variants in linkage disequilibrium (LD) with causal variant from the set of statistically associated genetic polymorphisms. This will enable further biological experiment to reveal their precise molecular mode of action. While DSS have enabled identification of novel genes, pathways and networks that harbor genetic variations responsible for a horde of phenotypes, current human host and mGWASs do not clearly provide a mechanistic understanding of how the consortium of host genetic variation, the microbiome and the environment cooperate to influence traits/disease. Furthermore, we even delve into the challenges host or mGWASs face in discerning true risk variants and the perplexity in understanding how these risk variants exert their effects [9]. However, these constitute an important step toward a global understanding of what and how human variation contributes to phenotypic differences ranging from development, physiology and behavior to pathogenesis of many human diseases. With this global view, it has today become critical to integrate different factors that potentially play roles in the human phenotypic variation. Consequently, the integration of multi-omic data within integrative DSS approaches (Figure 2) will greatly improve our understanding and unravel the complex interaction between host genetics, microbiome and environment that are pertinent to human health and disease. Such integrative DSS approaches will facilitate the understanding of causality and subsequently

translate to clinical and medical applications, including proper diagnosis, prevention and treatment. It is important to note that the integration of such high-dimensional and diverse multi-omic data is itself challenging [10].

In this paper, we (1) discuss genetic risk in the observed associations between host genetics, the microbiome and the complex diseases using DSS, (2) outline the role of integrative multi-omic approaches to unravel causality and (3) conclude by highlighting some research areas where further work on DSS is needed to establish integrative genetic risk factors of complex traits and diseases.

Human variation

Human populations differ in the distribution and frequency of their phenotypic expressions. These 'human variations' result from both their genetic components whose compositions are largely shaped by their genomic history and nongenetic components. The genomic history of human evolution is characterized by the exchange of genetic materials across individuals, resulting into individuals with unique genetic features. Consequently, the current pattern of human genetic diversity is hypothesized to be 'ancestral', most genetic variants having occurred once in human history and vertically spread across populations, rather than due to recurrent mutations [11]. This genetic mixing generated substantial genetic variation. It is estimated that the human genome contains between 4.1 and 5.0 million polymorphisms, at least 99% of them being Single Nucleotide Polymorphisms (SNPs) and short indels [12]. Despite the large number of genetic variants in the genome, only a minute fraction of them are hypothesized to be causal of disease [11], the majority being neutral (having no contribution to phenotypic variation) or near neutral.

In addition to genetic variation, the human microbiome, the collection of bacteria, archaea, fungi, protozoans and viruses that colonize our body surfaces and their respective genomes, plays an important role in health and disease variation [5, 8]. A

long-standing goal in human genetic studies is elucidating the fraction of heritability and environment that may contribute to the variation in health and disease [11]. This goal is, however, challenging owing to the myriad of confounding factors that likely mask true and causative associations. To this end, various DSS methods for host and mGWAS have been developed to investigate the association between genetic variability and/or microbiome composition and disease susceptibility/resistance [12, 13].

It is worth to note that current developed omic technologies, particularly genotyping arrays and imputation panels, have widely been designed for populations of European descent with long-range patterns of LD [13]. Additionally, current microbial databases are built with genomes of European descent populations. The insufficient capture of some haplotypes from non-European populations limits the power to detect important associations. In addition, many new genetic associations to diseases that have been identified from both host and mGWASs have primarily been applied to samples from population of European ancestry [12, 13]. This has an implication that a substantial proportion of functionally important variations in other populations are not captured. This may partly explain the challenge in replicating variants identified in some populations let alone drawing a strong mechanistic link between the associated variants and disease phenotypes. Variants associated with diseases found in populations of European descent do not always replicate in non-European, particularly African populations [12, 14] for several reasons, including differences in allelic architecture, LD and confounding of environmental factors across populations. In addition, genetic determinants of disease and their effect sizes have also been shown to vary significantly between European and African populations [13, 14].

The high levels of genetic diversity and the burden of complex diseases in non-European, particularly in African populations, may further introduce both challenges and opportunities not only for host and mGWAS but also for the general omic analyses [9, 14]. The high genetic diversity, environmental heterogeneity and high burden of diseases that characterize the African population [15] make it clear that leveraging the African genomic data is pertinent to a robust analysis. Notwithstanding, current genome-wide DSS that leverage a single level of omic data may be limited for adequately gaining insights into the basis of observed phenotypic expressions.

Whereas the importance of increasing host GWAS to include samples from non-European ancestry has been long appreciated, the observation that the composition and relative abundance of the human microbiome is influenced by diet, environment and host genetics has demonstrated how studies on diverse backgrounds may provide valuable insights. For example, Smits *et al.* compared the gut microbiome of the Hadza hunter-gatherers of Western Tanzania to 18 others in 16 different countries with varying lifestyles [16]. They found the gut microbiome composition to be clearly differentiated between traditional and industrialized populations. With the inclusion of multiple genetically diverse populations in the analysis, it is hoped that it will be possible to (1) elucidate the genetic architecture of many complex traits, (2) more accurately reconstruct ancestral haplotype which cannot occur in non-Africans and (3) shed light on the role of the human microbiome in disease susceptibility and resistance given differing environment and the burden of communicable and noncommunicable diseases in Africa, South America and Asia. The disparity in omic research, in terms of genetic diversity of the study population, and technology capacity across the globe,

may not favor the advancement of our understanding of the intricate pathogenesis of complex diseases. In addition, such disparity forms a major obstacle toward the full development of appropriate global prevention and treatment strategies particularly in light of precision medicine.

Methodologies underpinning host and microbiome DSS

To identify one or more genetic variants associated with a given disease, the most commonly used approach today is to perform a genetic association study. This goal is achieved by comparing the frequency of one or more genetic variants between cases and controls [17, 18]. This has led to the advent of the development of various host-based GWAS models, mostly based on linear mixed, random, mixed effects and Bayesian frameworks [19, 20]. **Box 1** provides a brief genesis of statistical methods for host-based GWAS approaches. Although many of these approaches and tools have been effective at uncovering the genetic basis of many complex traits, they have potentially missed out many novel genetic variants and/or failed to disentangle true signals from spurious associations owing to limitations in the underlying statistical models. Methodologically, these association frameworks (**Supplementary Table S1**) can be classified into two broad categories: linear regression-based and linear mixed model (LMM)-based frameworks. **Supplementary Table S1** displays the category of the tools and provides a brief description on each tool.

Box 1: statistical methods in host GWAS

The traditional statistical method for GWAS was the simple linear (1) or logistic (2) regression, which phenotype Y and the fixed effects X by the relation:

$$Y = X\beta + \varepsilon \quad (1)$$

$$\log \text{it}(p) \sim \alpha + \beta X \quad (2)$$

where β is the effect size, ε is random noise and $P = E(Y|X)$ is the expected value of phenotype given genotype. LMM modifies (1) by introducing a random effects U so that the model becomes

$$Y = X\beta + U + \varepsilon. \quad (3)$$

While LMM represented a powerful methodology, three key issues remained at the forefront of its implementation: (i) computational cost in evaluating the variance parameters, (ii) strategy of modeling SNP effects (fixed SNP effect vs random SNP effect) and confounders and (iii) method of modeling genetic architecture of the phenotype (infinitesimal versus non-infinitesimal), leading to development of various statistical tools. Under an infinitesimal model, the χ^2 [1 Degree of Freedom (d.o.f)] test statistic for testing association in equation (3), with the hypothesis $\beta_{\text{test}} = 0$ is

$$\chi^2 = \frac{(x'_{\text{test}} V^{-1} y)^2}{x'_{\text{test}} V^{-1} x_{\text{test}}}, \quad (4)$$

where, $V = \text{cov}(y) = \sigma_g^2 K + \sigma_e^2 I$; K is the genetic relationship matrix (kinship matrix) that models sample structure and

I is $n \times n$ identity matrix. The variance parameters σ_g^2 and σ_e^2 are typically unknown and to be computed. The computation of variance components in LMM is, however, expensive. Various approaches have been proposed to reduce computational cost. Kang et al. [22] proposed a direct estimation σ_g^2 and σ_e^2 by maximizing the REML function, and then applying spectral decomposition [23] obviating more computationally intensive approach of determining the best linear unbiased prediction via Henderson's iterative procedure [24].

Computational efficiency is improved from other approaches that replace the usual relatedness matrix computed from all genome-wide SNPs by low-rank relatedness matrix [23]. In addition, the effective sample size of a random effect is reduced by clustering subjects into genetically similar groups. Moreover, logistic mixed model has recently been advocated for, in place of LMM, analysis of binary traits. This is because LMM is based on the assumption that the trait has constant residual variance—an assumption that is usually violated by binary traits in the presence of covariates [24]. In recent years, in light of trait polygenicity, LMM methods that simultaneously test the effect of multiple SNPs have been developed, for example, GCTA [25]. In these, the model is set as

$$Y = X\beta + \sum_{i=1}^p g_i + \varepsilon \quad (6)$$

where g_i is a vector of genetic effects on whole genome. Bayesian modeling technique has since been adopted for joint analyses of multiple SNPs, because of their ability to increase power by leveraging known information on marker effects. Chen et al. [24] proposed Hierarchical Bayesian model, in which the SNP random effects are assumed to follow the mixture distribution:

$$Y = X\beta + U\alpha + \varepsilon \quad (7)$$

where α_j values are assumed to follow a Gaussian distribution.

This improved power over other LMM-based methods, such as GCTA modeled [25] by equation (6). Meanwhile, Loh et al. [33], in the current popular BOLT-LMM tool, undertook a different Bayesian approach to capture non-infinitesimal genetic architecture, modeling SNP effect sizes by fitting non-Gaussian mixture prior distribution that better models small- and large-effect sizes. In this framework, the test statistic in (4) is derived to be $\chi^2 = \frac{(X_{\text{test}} y^*)}{c}$, where y^* is a vector of residual phenotypes obtained by fitting a Gaussian mixture of priors to the standard LMM and c is calibration factor defined so that the LD score regression intercept of the χ^2 test statistic under the non-infinitesimal model matches with that under the infinitesimal model (4).

Linear regression-based approaches (Box 1) model the phenotype of an individual as function of fixed effects (which include genotype at the candidate marker, as well as optional covariates such as age, gender and other clinical information). In such models, the inflation in the test statistic can then be controlled

by using different methods. These include genomic control, multidimensional scaling, structured association and principal component analysis. These methods are implemented in the GWAS software tools including STRAT [21] and the currently popular PLINK [22] (Supplementary Table S1). Although these methods are effective at controlling inflation in test statistic when the population has structures, they do not suffice in the presence of population substructure [19]. In particular, they do not account for the complete genealogy of all study subjects. LMM (Box 2), an extension of the standard linear regression, partitions the explanatory variables into two groups: fixed effects, which are modeled as parameters that are fixed but unknown, and random effects, which are modeled as being drawn from a random distribution. This provides a powerful method to simultaneously account for various levels of sample structure, including population stratification, family structure and cryptic relatedness. Principally, this is achieved by fitting population structure as a fixed effect and incorporating marker-based kinship information via the variance-covariance structure of the random effect for the individuals [23, 24]. To date, LMM remains the workhorse for association mapping and nearly all-current GWAS tools are based on it. Importantly, not only does LMM provide a control for confounding due to sample structure, but it also increases the statistical power to detect causal variants and enables estimation of heritability explained by genotyped markers [25]. This is achieved by applying a correction that is specific to a given type of sample structure [26, 27].

Box 2: statistical methods in mGWAS

Diversity metrics (alpha diversity or beta diversity) can be leveraged as phenotype for mGWAS. In mGWAS [35] beta diversity is leveraged as phenotype in GWAS. The beta diversity analysis uses microbiome distance measures such as UniFrac and Bray-Curtis dissimilarity. Let $D = (d_{ij})$ be a beta diversity distance matrix between subjects i and j . Let $G_{ij} = |g_i - g_j|$ be the genetic distance. The main hypothesis here is that if a SNP is associated with the microbiome (through its distance matrix), then the microbiome distance measure should be smaller for pairs of subjects that have identical genotype at the SNP given SNP. Then, assuming a linear relationship,

$$d_{ij} = \beta_0 + \beta G_{ij} + \varepsilon_{ij}$$

where ε_{ij} is the environmental effect. If n is the number of individuals, then $\frac{n}{2}(n-1)$ pairs of subjects can be clustered into three groups with genetic distance 0, 1 and 2. The hypothesis $H_0: \beta = 0$ versus $\beta > 0$ is tested using a score statistic, S , derived by minimizing $\sum_{i < j} (d_{ij} - \beta_0 - \beta G_{ij})^2$:

$S = \sum_{i < j} d_{ij}^* G_{ij}$, where $d_{ij}^* = d_{ij} - \frac{2}{n(n-1)} \sum_{p < q} d_{pq}$. The conditional variance of S on D is $\text{Var}_D(S|D) = \sigma^2 = \sum_{i < j, p < q} d_{ij}^* d_{pq}^* \text{Cov}(G_{ij}, G_{pq})$, $\text{Cov}(G_{ij}, G_{pq}) = 0$ when i, j, p and q are distinct.

Then the variance-scaled score statistic for testing association of a microbiome distance with a genetic distance across pairs of subjects is then given by $Z_{sc} = S/\sigma \sim N(0, 1)$.

This model was subsequently extended to incorporate multiple distance matrices so as to improve statistical power.

Lynch *et al.* [36], proposed an mGWAS method that uses, as phenotype, relative abundance computed at a given taxonomic level. Owing to the high dimension of microbiome relative abundance data, their method applies regularization based on lasso regression, which results in a sparse solution. Host SNPs associated with microbiome are then determined via permutation test, while specific taxa correlated with host genetic variation identified using stability selection [37].

While LMMs has become the method of choice for GWAS, it presents a substantial computational challenge. Methods that compute the exact association test statistics from complete or imputed genotype data certainly provide 'accurate' *P*-values and thus maintain high power. Current software tools using exact computation of the test statistic include EMMA [28], FaST-LMM [29] and more recently GEMMA [23] (Supplementary Table S1). Although the *P*-values produced by GEMMA are identical to EMMA and FaST-LMM, the algorithm in GEMMA provides a higher efficiency, in terms of per-SNP computational time. It is important to note, however, that even with efficient implementation, the computation of variance component in these tools scale in time of $O(mn^3)$, where *m* is the number of markers and *n* is the number of individuals, and as such these tools are impractical for analysis of large samples (Box 2) [26]. To circumvent computational cost, several approximation methods have been proposed. These include (i) obviating repetitive estimation of variance components, as implemented in the software tools TASSEL [29] and EMMAX [30] (Supplementary Table S1) and (ii) step-wise implementation [31]: estimation of the residuals from the LMM under the null model and then using these residuals as phenotypes for analysis by the standard linear model (e.g. implemented in software GRAMMAR [32]. This substantially reduces the per-SNP computation time. Another important note is that these approximation-based tools show reduced power at SNPs with small effect sizes. Recently, Loh *et al.* [33] proposed an efficient approximation method that adapts the LMM by taking a Bayesian perspective and modeling non-infinitesimal genetic architectures via non-Gaussian mixture prior distributions, invoking the fast variational approximation to compute approximate phenotypic residuals. Their methods, implemented in BOLT-LMM and BOLT-REML (Supplementary Table S1), are the current state-of-the-art methods for host GWAS analysis not only in terms of computation time and memory requirement in large cohorts but also, importantly, in terms of the genetic architecture modeled. Apart from the methods proposed by Loh *et al.* [33], all currently existing tools are based on the infinitesimal model in which all variants are assumed causal with effect sizes following independent Gaussian distribution. In addition, all current tools, as far as we know (Supplementary Table S1), have computational times that scale with the square or cube of sample size, rendering them unfeasible in large data sets. Overall, all these state-of-the-art methods of host GWAS cannot distinguish confounding from polygenicity in the association test [19]. In contrast to host GWAS, methodological development for mGWAS has just begun to emerge. The past 3 years have seen the development of statistical methods for investigating host-microbiome interactions. These methods leverage microbiome features (relative abundance of microbial taxa, alpha diversity, beta diversity or microbial pathway) as a complex trait and determine their correlation with host genetics, by testing either multiple-distance matrices across pairs of subjects or taxa

relative abundance [34]. Principally, GWAS can be performed between any given set of genotypes and phenotypes, and thus, although most of mGWAS carried out to date have been limited to the metagenomic level, the framework can similarly be performed at the metatranscriptomic, metaproteomic and metabolomic levels.

The two currently existing mGWAS tools are based on linear regression model as illustrated in Box 3 and Supplementary Table S2 summarizes the functionalities of these mGWAS tools. The 1st mGWAS tool is mGWAS [35]. It is a statistical framework for identifying host genetic variants associated with microbiome beta diversity with or without interacting with environmental factors and corrects for skewness and kurtosis. The 2nd mGWAS tool, HOMINID [36], which is based on Lasso regression, identifies associations between host SNPs and microbiome taxa. Additionally, by using Lasso regression plus stability selection with randomized Lasso, this tool enables identification of microbial taxa that are correlated with specific host SNPs.

Box 3: some unsolved challenges in host and microbiome GWAS

- Statistical methodologies for host-based and mGWAS for identifying causal factors in the observed associations between the environment, host genetics, the microbiome, and complex phenotypes
- Determining the direction of causality in the interplay between host genetics and microbiome in complex traits
- Determining the impact of host epigenetics on microbiome features
- Accounting for interaction among microbial species, in identifying specific microbial taxa associated with host genetic variation
- Correcting for confounders arising from the multiple factors that modulate the microbiome composition
- Modelling the true genetic architecture of complex traits
- Accounting for the missing heritability in host GWAS

Limitations in current methods and tools for host and mGWAS

Although substantial progress in the development of methods and tools for host genome-wide association mapping has occurred over the past decade and many new tools continue to be unveiled, many significant challenges need to be addressed before the gap between statistical association and biological association can be narrowed. Similarly, initial foray into mGWAS using custom-made mGWAS tools or ported host GWAS tools have illuminated several important microbiome-associated host genetic polymorphisms. Even so, however, several methodological limitations and pitfalls exist in using these tools. Box 1 provides some challenges in host and mGWAS.

Most current host GWAS tools are built on the infinitesimal genetic architecture, which makes the implicit assumption that all variants are causal with small-effect sizes independently drawn from Gaussian distributions. For complex traits, it is now known, however, that only a small proportion of the genetic variants are actually causal [17]. Thus, employing this assumption clearly limits power. Bayesian techniques have been

invoked to incorporate non-infinitesimal genetic architecture [33, 34]. However, the prior on marker effect sizes is assumed to follow independent Gaussian distributions. The validity of the Gaussian assumption, according to the central limit theory, requires sufficiently large sample sizes that may not necessarily be the case in GWAS analyses. This concern can be even greater for the case of ascertained case-control studies because in such cases the distributions of genetic effects are no longer Gaussian [37] and the independence between genetic effects and environmental effects is lost due to disproportionate sampling of cases and controls [38]. In the same light, methods and tools so far developed to leverage the microbiome as a complex quantitative trait for association mapping assumes the infinitesimal genetic architecture and do not incorporate the effect of genetic interactions on the microbiome. Early insights gained from the mGWAS studies suggest that only a few host loci interact with the microbiome; thus, a more realistic non-infinitesimal genetic architecture could be modeled by taking a Bayesian approach and incorporating small and large effect size loci using suitable distributions. However, the question of how to choose the appropriate distribution and/or incorporate the (possibly) different distributions of marker-effect sizes (because not all markers were created equal) remains open for future research. In addition, existing methods [34] that leverages multiple microbiome beta-diversity distance matrices apply the strongest association, defined as the highest P-value from the set of P-values obtained from each distance matrix, to evaluate significance. Statistical techniques that allow to rescale the overall statistic threshold by effect of distance matrices should provide increased power. Here methods such as truncated product method [39] could be employed. Meanwhile, aptly approaches to tackle the complexity implied by the multi-dimensional interactions among genetic loci would perhaps include recursive partitioning method, multifactor dimensionality reduction or Bayesian technique.

It must be pointed out that another great limitation of current host and mGWAS methods alike is the absence of robust models for interactions of host genetics and the microbiome with the environment. While significant progress has been achieved in this direction for host GWAS, no appreciable stride has been made on the mGWAS side. The plasticity of microbiome data to a plethora of environmental factors makes realistic investigation of microbiome-environment interactions an extremely difficult problem. A 1st step toward addressing this issue will be to define a 'gold standard' for statistical methods that will be developed for such complex interactions. Of course, given that complexity of the microbiome, in terms of dimensionality and features (zero inflated, over dispersed and multiple outliers), any such models that incorporate environmental factors is likely to be computationally intractable. In such a case, likelihood free methods such as approximate Bayesian computation could be adopted. These limitations and challenges in current DSS frameworks consequently impact the progress toward a complete understanding of the nature association between host genetics, gene regulation and microbiome.

Dilemma in risk factors: host genetics, gene regulation and microbiome causality

The results of several host GWAS conducted over the years have provided several insights into the biological processes underlying many diseases [40]. Several genes, pathways and regulatory

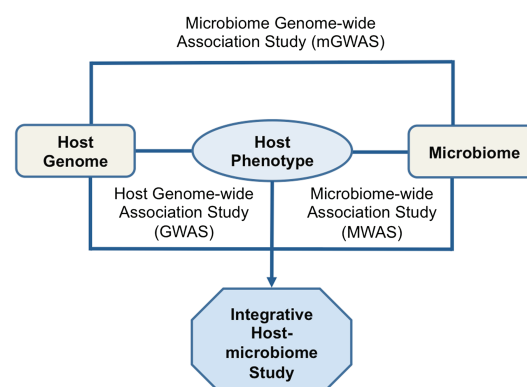


Figure 3. Possible direction of interaction between host genetic, microbiome and gene regulation on host phenotype.

networks have been identified for a number of complex diseases, including cardiovascular diseases, type 2 diabetes and cancer [41, 42]. While the pathogenesis and nature of the aberrant activities in these pathways and regulatory networks is coming into view for some complex diseases, the apparent molecular circuitry remains generally elusive for most traits. Meanwhile, initial forays into mGWAS have already demonstrated the intimate link between host genetic polymorphisms and microbiome attributes [43, 44]. For example, a pioneering study by Blekhan et al. using samples from 93 individuals identified multiple host loci associated with changes in abundance of microbial taxa [5]. Intriguingly, some of the loci detected in this and other mGWAS carried out to date overlap with several expression quantitative trait loci that have previously been identified across multiple tissues [38]. Moreover, this link is observed to be tissue specific [45] and enriched with specific human proteins [43]. These discoveries suggest the likely influence of host gene regulation on specific tissues that interact with the microbiome. However, little is known about the interplay between host genetic variation, microbiome composition and host gene expression and how this impacts host traits (Figure 3).

Several challenges remain toward establishing causality. These association results are often limited to correlations and may generally end up identifying consequential changes rather than the true risk factors that may lead to establish genetics causality [13, 2]. For example, the predictive power of disease risk remains poor because current identified variants account for only a small proportion of additive genetic variation [2]. Consequently, current findings from disease scoring approaches have not yet had major impact on therapeutic optimization for the majority of complex traits. Moreover, it has become apparent that changes in gene regulation are at the center stage of biological mechanisms underlying most associations, and yet our current knowledge of gene regulation is still limited [46]. Without a comprehensive knowledge of gene regulation and the paucity in available tools for studying regulation, the transition from statistical associations to biological insights (biological mechanisms, the particular genes involved and the direction of causality) remains a challenge. Nevertheless, significant advances in technology coupled with multi-omic approaches, which is able to simultaneously capture millions of data points, will enable system-wide examination of complex interactions in biological system [47].

Omic data integration

The multi-omic approach involves understanding the complexity of the living as a whole by incorporating data at multiple biological levels: from gene sequencing to protein expression and metabolic structures [48]. Therefore, these data cover wide range of the biological information involved in the variations that occur in the genes and cellular networks and influence the functioning of organic systems in their entirety [48]. This approach has mainly been driven by the surge in high-throughput sequencing efforts over the past two decades. Multi-omic approach will be able to dramatically improve our way of analyzing the biological system, which relies on a single-omic model, that offers limited insight into the complex and dynamic nature of biological networks and their association with environment factors. Integrating data from all these omic levels will be useful for identifying new biomarkers, generating new knowledge and/or developing new diagnostic tools. The integration of multi-omic data with GWAS will bring more insight in the causal factors and the interactions between the host, microbiome and environment. However, multi-omics is still in its infancy [47, 49] and requires large-scale studies that need state-of-the-art tools. So far, many integrative multi-omic tools have been developed [50]. Multi-omic method, however, poses significant challenges in terms of the analysis approach, the statistical methods and the interpretation of these numerous data [51]. One of the most important challenges is the difficulty of representing existing knowledge about the molecular processes involved in many complex diseases, given that biological systems are a myriad of complex connections and are closely linked to their evolutionary history. Another fundamental challenge is that of matching the heterogeneous data from different methods and platforms. It requires a synchronization of huge amount of data that vary in data format, thus potentially will add bias and noise to data integration processes. This step can use various strategies either merging the different data coming from the same subjects or trying to homogenize the data. However, it is very critical because matching the data must not only put the data in the same files but also bring a new meaningful knowledge. In addition, incorporating various omic data each of which typically contains many measurement errors raises the issue of quality control. This is compounded by the fact that the measured molecular data are prone to bias arising from samples preparation and processing. Another challenge is analyzing the data of huge size and of different classifications (molecular data, measurement features, technology used, biological samples, type of study, etc.). Therefore, bioinformatics approaches and pipelines and their associated mathematical models need to address the dimensionality and heterogeneity of the data. In doing so, machine-learning approaches may be more suitable for data integration and related modeling, as they provide a robust way of leveraging hidden knowledge from various omic data types to improve analyses. On the other hand, topological data analysis methods that examine the shape of data with geometric dimensional conversions [52] can also be suitable in finding hidden patterns compared to other standards methods such as correlation-based analysis [53, 54] and unsupervised data integration (matrix factorization methods, Bayesian methods, network-based methods, multiple kernel learning and multi-step analysis) [53, 55]. The other approach to explore is the complex network-based approaches, which may be worthy of exploration to efficiently handle the multi-omic data deluge [53].

Role of bioinformatics in the era of multi-omics

Considering the molecular variation of biomarkers, the rapid growth in large-scale omic technologies opens windows for global views of biological system in a holistic hypothesis-driven manner. Integrative approaches need to be designed and applied to various levels of biological information to comprehend the pathogenesis of complex diseases [49]. Bioinformatic approaches, resources and computational biology tools are at the center for the advancement in implementing real-time multi-omic integration and health care analytics. Bioinformatics is arguably a recent field but has substantially contributed to the advancement in the modernization of computational techniques and capacity to handle the amount of biological data generated by genome sequencing and variation studies [1, 10, 15]. Bioinformatics currently plays a critical role in deciphering omic data and organizing information in all aspects of independent omic layers. While bioinformatics and computational biology tackle challenges raised from each independent omic data type in the past decades [15, 49], today multi-omic era has raised further challenges in integrating various levels of biological information [47, 49]. Furthermore, analyzing and interpreting such integrative biological information demand outstrips supply and further bioinformatics and computational biology capacities are needed. This issue raises the need to strengthen the multidisciplinary nature of bioinformatics and education. In doing so, this will have a critical impact on the discovery of multi-omic diagnostics, biomarkers, clinical decision-making and data-driven medicine.

Conclusions and perspectives

The substantial role of host genetics on the microbiome and on host phenotypes has been identified using the wealth of available DSS tools. Although many of these methods and tools have been effective at uncovering the genetic basis of many complex traits, they have potentially missed out many novel genetic variants and/or failed to disentangle true signals from spurious associations owing to limitations in the underlying single-omic statistical models. Given the current experimental observations of the strong influence of host genetics on the microbiome and the role of the microbiome on host phenotypes, it has become increasingly apparent that integrating the microbiome in host GWAS will reveal many important insights and it will be a strong 1st step toward establishing causality. A critical challenge facing the host genetics and microbiome field is the identification of genetics risk factors with strong effect to allow the establishment of the direction of causality in the observed associations between the environment, host genetics, microbiome and complex diseases. A complete solution for this issue would be one of the major breakthroughs for a century-long problem on understanding human variation relevant.

The integrative analysis may be the key to better understand the role and mechanisms of host genetics, microbiome and environment in the manifestation of many complex diseases (Figure 2). This will ultimately revolutionize therapeutics, driving the era of precision medicine. The major challenge for this integration lies in the development of novel statistical techniques and multi-omic data integration. The integrative analyses will throw bioinformatics and human genetic studies into a brand-new era, quickening the pace of movement from population level to individual level understanding of complex human traits and diseases.

It will be interesting to perform simulations to examine the power of various current host and mGWAS methods and to investigate their performance under various important factors, such as sample size, marker effect size, host genetic correlations, population structure and microbial interactions. Such simulations will not only highlight current methodological limitations but also guide development and validation of future association tools. In brief, we discussed DSS approaches in host GWAS and in mGWAS, outlining their associated methods and tools. We further discussed the limitations of these methods and highlighted the dilemma in dissecting causality between host genetic, microbiome and environment. Finally, we underscored the importance of integrating multi-omic data with GWAS and outlined some of the challenges in this data integration. We believe that this paper may motivate the development of new methods for mGWAS.

Key Points

- Discussing issues related to host and mGWAS
- Outlining current methods and tools available for host and mGWAS
- Discussing the importance of integrative multi-omic approaches in understanding causal factors and the direction of their effects on host phenotypes

Supplementary Data

Supplementary data are available online at <https://academic.oup.com/bfg>.

Funding

This work was funded in part by DAAD, the German Academic Exchange Programme, under grant number 91653117. Some of the authors are also funded in part by the National Institutes of Health Common Fund under grant numbers 1U54HG009790-01 (IFGeneRA), U01HG009716 (HI Genes Africa), U24HG006941 (H3ABioNet) and U01HG007459-01 (SADaCC) and by Wellcome Trust/AESA under grant number H3A/18/001. The content of this publication is solely the responsibility of the authors and does not necessarily represent the official views of the funders.

Acknowledgements

We are grateful to Oyekanmi Nashiru and Hassan Ghazal for helpful discussions. Computations were performed using facilities provided by CHPC (<https://www.chpc.ac.za/>).

References

1. Zhang J, Chiodini R, Badr A, et al. The impact of next-generation sequencing on genomics. *J Genet Genomics* 2011; **38**(3):95–109.
2. Visscher PM, Brown MA, McCarthy MI, et al. Five years of GWAS discovery. *Am J Hum Genet* 2012; **90**(1):7–24.
3. Goodrich JK, Davenport ER, Clark AG, et al. The relationship between the human genome and microbiome comes into view. *Annu Rev Genet* 2017; **51**:413–33.
4. Sandoval-Motta S, Aldana M, Martínez-Romero E, et al. The human microbiome and the missing heritability problem. *Front Genet* 2017; **8**:80.
5. Blekhman R, Goodrich JK, Huang K, et al. Host genetic variation impacts microbiome composition across human body sites. *Genome Biol* 2015; **16**(1):191.
6. Cho, Blaser MJ. The human microbiome: at the interface of health and disease. *Nat Rev Genet* 2012; **13**(4):260.
7. Wang J, Jia, H. Metagenome-wide association studies: fine-tuning the microbiome. *Nat Rev Microbiol* 2016; **14**(8):508.
8. Richards A, Muehlbauer L, Alazizi A, et al. Gut microbiota composition impacts host gene expression by changing chromatin accessibility. 2018 bioRxiv. doi: <https://doi.org/10.1101/210294>.
9. Hall B, Tolonen AC, Xavier RJ. Human genetic variation and the gut microbiome in disease. *Nat Rev Genet* 2017; **18**(11):690.
10. Kohl M, Megger DA, Trippler M, et al. A practical data processing workflow for multi-OMICS projects. *Biochim Biophys Acta* 2014; **1844**(1):52–62.
11. Frazer KA, Murray SS, Schork NJ, et al. Human genetic variation and its contribution to complex traits. *Nat Rev Genet* 2009; **10**(4):241.
12. Visscher PM, Wray NR, Zhang Q, et al. 10 years of GWAS discovery: biology, function, and translation. *Am J Hum Genet* 2017; **101**(1):5–22.
13. Peprah E, Xu H, Tekola-Ayele F, et al. Genome-wide association studies in Africans and African Americans: expanding the framework of the genomics of human traits and disease. *Public Health Genomics* 2015; **18**(1):40–51.
14. Wang K, Li M, Hakonarson H. Analysing biological pathways in genome-wide association studies. *Nat Rev Genet* 2010; **11**(12):843.
15. Shameer K, Badgeley MA, Miotto R, et al. Translational bioinformatics in the era of real-time biomedical, health care and wellness data streams. *Brief Bioinform* 2016; **18**(1):125–4.
16. Smits SA, Leach J, Sonnenburg ED, et al. Seasonal cycling in the gut microbiome of the Hadza hunter–gatherers of Tanzania. *Science* 2017; **357**(6353):802–6.
17. Zuk O, Hechter E, Sunyaev SR, et al. The mystery of missing heritability: genetic interactions create phantom heritability. *Proc Natl Acad Sci U S A* 2012; **109**(4):1193–8.
18. Cantor RM, Lange K, Sinsheimer JS. Prioritizing GWAS results: a review of statistical methods and recommendations for their application. *Am J Hum Genet* 2010; **86**(1):6–22.
19. Yang J, Zaitlen NA, Goddard ME, et al. Advantages and pitfalls in the application of mixed-model association methods. *Nat Genet* 2014; **46**(2):100.
20. Pasaniuc, Price AL. Dissecting the genetics of complex traits using summary association statistics. *Nat Rev Genet* 2017; **18**(2):117.
21. Porras-Hurtado L, Ruiz Y, Santos C, et al. An overview of STRUCTURE: applications, parameter settings, and supporting software. *Front Genet* 2013; **4**:98.
22. S. Purcell, B. Neale, K. Todd-Brown, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 2007; **81**(3): 559–75.
23. Zhou X, Stephens M. Genome-wide efficient mixed-model analysis for association studies. *Nat Genet* 2012; **44**(7):821.
24. Chen H, Wang C, Conomos MP, et al. Control for population structure and relatedness for binary traits in genetic association studies via logistic mixed models. *Am J Hum Genet* 2016; **98**(4):653–66.

25. Yang J, Lee SH, Goddard ME, et al. GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet* 2011; **88**(1):76–82.
26. Price L, Zaitlen NA, Reich NA, et al. New approaches to population stratification in genome-wide association studies. *Nat Rev Genet* 2010; **11**(7):459.
27. Golan D, Rosset S. Mixed models for case-control genome-wide association studies: major challenges and partial solutions. In: Borgan Ø, Breslow N, Chatterjee N, et al. (1st edn). *Handbook of Statistical Methods for Case-Control Studies*. Boca Raton, FL: Chapman and Hall/CRC 2018:495–514.
28. Yu J, Pressoir G, Briggs WH, Bi IV, et al. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat Genet* 2006; **38**(2):203.
29. Zhou X, Stephens M. Genome-wide efficient mixed-model analysis for association studies. *Nat Genet* 2012; **44**(7):821.
30. Aulchenko YS, De Koning DJ, Haley C. Genomewide rapid association using mixed model and regression: a fast and simple method for genomewide pedigree-based quantitative trait loci association analysis. *Genetics* 2007; **177**(1): 577–85.
31. Kang HM, Sul JH, Service SK, et al. Variance component model to account for sample structure in genome-wide association studies. *Nat Genet* 2010; **42**(4):348.
32. Golan ES, Lander, Rosset S. Measuring missing heritability: inferring the contribution of common variants. *Proc Nat Acad Sci U S A* 2014; **111**(49):E5272–81.
33. Loh P-R, Tucker G, Bulik-Sullivan BK, et al. Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat Genet* 2015; **47**(3):284.
34. Hua X, Song L, Yu G, et al. MicrobiomeGWAS: a tool for identifying host genetic variants associated with microbiome composition. 2015 BioRxiv:031187.
35. Lynch J, Tang K, Priya S, et al. HOMINID: a framework for identifying associations between host genetic variation and microbiome composition. *GigaScience* 2017; **6**(12):1–7. doi: [10.1093/gigascience/gix107](https://doi.org/10.1093/gigascience/gix107).
36. Zhao N, Chen J, Carroll IM, et al. Testing in microbiome-profiling studies with MiRKAT, the microbiome regression-based kernel association test. *Am J Hum Genet* 2015; **96**(5): 797–807.
37. Günther F, Wawro N, Bammann K. Neural networks for modeling gene–gene interactions in association studies. *BMC Genet* 2009; **10**(1):87.
38. Koch L. Complex disease: a global view of regulatory networks. *Nature Rev Genet* 2016; **17**(5):252.
39. Shu L, Chan KHK, Zhang G, et al. Shared genetic regulatory networks for cardiovascular disease and type 2 diabetes in multiple populations of diverse ethnicities in the United States. *PLoS Genet* 2017; **13**(9):e1007040.
40. Gao L, Uzun Y, Gao P, et al. Identifying noncoding risk variants using disease-relevant gene regulatory networks. *Nat Commun*, 2018; **9**(1):702.
41. Goodrich JK, Davenport ER, Clark AG, et al. The relationship between the human genome and microbiome comes into view. *Annu Rev Genet* 2017; **51**:413–33.
42. Davenport ER, Cusanovich DA, Michelini K, et al. Genome-wide association studies of the human gut microbiota. *PLoS One* 2015; **10**(11):e0140301.
43. Huttenhower C, Gevers D, Knight R, et al. Structure, function and diversity of the healthy human microbiome. *Nature* 2012; **486**(7402):207.
44. Price AL, Spencer CC, Donnelly P. Progress and promise in understanding the genetic basis of common diseases. *Proc Biol Sci* 2015; **282**(1821):20151684.
45. Palsson B, Zengler K. The challenges of integrating multi-omic data sets. *Nat Chem Biol* 2010; **6**(11):787.
46. Ritchie MD, Holzinger ER, Li R, et al. Methods of integrating data to uncover genotype–phenotype interactions. *Nat Rev Genet* 2015; **16**(2):85.
47. Hasin Y, Seldin M and Lusis A. Multi-omics approaches to disease. *Genome Biol* 2017; **18**(1):83.
48. Huang S, Chaudhary K, Garmire LX. More is better: recent progress in multi-omics data integration methods. *Front Genet* 2017; **8**:84.
49. Palsson B, Zengler K. The challenges of integrating multi-omic data sets. *Nat Chem Biol* 2010; **6**(11):787.
50. Zhong S, Jiang D, McPeck MS. CERAMIC: case–control association testing in samples with related individuals, based on retrospective mixed model analysis with adjustment for covariates. *PLoS Genet* 2016; **12**(10):e1006329.
51. Yu G, Gail MH, Consonni D, et al. Characterizing human lung tissue microbiota and its relationship to epidemiological and clinical features. *Genome Biol* 2016; **17**(1):163.
52. Yoo S, Huang T, Campbell JD, et al. MODMatcher: multi-omics data matcher for integrative genomic analysis. *PLoS Comput Biology* 2014; **10**(8):e1003790.
53. Zhang Z, Ersoz E, Lai CQ, Todhunter RJ, et al. Mixed linear model approach adapted for genome-wide association studies. *Nat Genet* 2010; **42**(4):355.
54. Lippert C, Listgarten J, Liu, et al. FaST linear mixed models for genome-wide association studies. *Nat Methods* 2011; **8**(10):833.
55. Bradbury PJ, Zhang Z, Kroon DE, et al. TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics* 2007; **23**(19):2633–5.

The Microbiome in Host Phenotype

"It is in the nature of science that once a position becomes orthodox it should be subjected to criticism.... It does not follow that, because a position is orthodox, it is wrong".

- John Maynard Smith, *Annual review of Biology*

This chapter first describes the human microbiome and discusses its role on human traits and diseases. Then, a discussion is presented of the statistical methods to link the microbiome to its human host phenotypes. The chapter then concludes with a discussion on the interaction between the human microbiome and its host genotype.

3.1 Introduction

The human microbiome represents the complex assemblage of all microorganisms, including bacteria, archaea, viruses, and fungi, that colonize the human epithelial surfaces such as the mouth and gut. It is estimated to contain about the same number of microbial cells as human cells and 100 times more genes than the human genome (Kho and Lal, 2018; Ding et al., 2019). Most of these microorganisms reside in the gastrointestinal tract (gut) (Schmidt et al., 2018). These gut-associated microorganisms are referred to as the *gut microbiota*, and consists of mostly commensal bacteria which are predominantly from four major phyla: Firmicutes,

Bacteroidetes, Proteobacteria, and Actinobacteria (Clemente et al., 2012; Zitvogel et al., 2018; Khan et al., 2019). Notwithstanding this, there are significant inter-individual differences, with regards to microbial diversity, within these major phylotypes (Koboziev et al., 2014; Khan et al., 2019). Given the vast number of these microorganisms that have co-evolved with their human host, the human host has come to be described as a ‘superorganism’, whose genetic landscape encompasses that of the human host and the affiliated microbiome (Gill et al., 2006).

Until recent decades, our understanding of the microbiome, its functionality, and roles in human health and disease has been very limited due to technology; specifically given that knowledge about the human gut microbiome stemmed from culture-based methods and yet microbes are non-cultivable, let alone the labour intensiveness of the culturing (Thursby and Juge, 2017; Council et al., 2007). During the past few years, however, technological improvements by the advent of culture-independent high-throughput sequencing methods have made it feasible to examine these microbial communities.

Currently, two approaches are used to explore microbial communities, namely amplicon sequencing, and metagenomic shotgun sequencing. Amplicon sequencing that can allow study of bacteria, archaea as well as micro-eukaryotic composition (Falony et al., 2019), relies on sequencing of a specific (target) taxonomic marker gene (usually the 16S rRNA gene [also designated 16S rDNA, and the terms are often used interchangeably; current policy for *American Society for Microbiology* is that ‘16S rRNA gene’ be used (Clarridge, 2004)] for bacteria and ITS for fungi) (Regalado et al., 2019). This bacterial gene is chosen because it is ubiquitous in the bacterial kingdom and contains nine highly variable regions (V1–V9) that exhibits considerable sequence diversity among different bacteria, which provides species-specific signatures for identification of bacteria (Chakravorty et al., 2007). The basic idea is to isolate, from all bacteria, the DNA strand corresponding to a given variable region of the 16S rRNA gene and to sequence the resulting amplicons, after DNA amplification. Since up to 80% of the bacterial RNA is made up of rRNA, the 16S rRNA has high resolution to detect rare bacterial members (Balding

et al., 2008). In addition, through taxonomic assignment, the sequencing approach can also reveal the phylogenetic structure of the detected bacteria assemblage (Balding et al., 2008; Shah et al., 2011).

The shotgun metagenomic approach, meanwhile, sequences the entire genomic DNA of an microorganism, as opposed to a just a specific gene in the case of amplicon sequencing. In this, bacterial DNA, extracted from all cells, is shredded into fragments which are then independently sequenced. Since different sections of the genome harbour different information, by sequencing entire DNA the shotgun metagenomic sequencing can provide information on existing bacterial clades as well as insight into the functional roles of their genes (Ranjan et al., 2016). Clearly, the shotgun metagenomic sequencing would generate a greater amount of information and ultimately better assessment of microbial community assemblage than the targeted (16S rRNA) sequencing. That said, it is important to note that each method had its inherent strength and limitation. For example, shotgun metagenomic sequencing generally has limited ability to detect rare species in complex communities, because of its shallow sequencing depth (Kalyuzhnaya et al., 2008), whereas the 16S rRNA sequencing does not provide high sensitivity in identifying bacteria at the species or strain level (Balding et al., 2008). Nonetheless, the 16S rRNA approach is widely used due to its relatively cheap sample processing and well-established analytical pipelines (Falony et al., 2019); albeit they are usually very informative. Although the microbiome encompasses viruses and fungi that also inhabit the body, indigenous microbiome studies have focused on bacteria. This focus is partly because sequencing methods used to examine microbial communities were developed to study bacteria (Young, 2017).

Progress made with 16S rRNA (or 16S rDNA)-based sequencing has greatly expanded our repertoire of knowledge of the microbiome, and begun to unravel the underpinnings of the intricate link between humans and their microbial residents. The main development of an individual's microbiome occurs during the first few years of life, in as much as some developments are thought to occur before birth (Deering et al., 2020; Thursby and Juge, 2017). By around 2.5 years of age, the microbiota of resembles that of adult microbiota, in terms of composition,

diversity and functional capabilities (Thursby and Juge, 2017; Koenig et al., 2011; Rodríguez et al., 2015). In adulthood, the microbiota of an individual is fairly resilient, although it can be altered by both endogenous and exogenous factors. Variation in the composition of the microbiome is shaped by multiple factors, ranging from lifestyle factors (e.g. smoking and physical activity) to diet, host's genotype, antibiotic usage, pre- and pro-biotic treatment, social interactions, dietary composition, and environmental exposure to xenobiotics (Scotti et al., 2017).

It is increasingly becoming apparent that symbiosis between the microbiome and its human host's genome is a two-way interaction. Host genetics modulates the microenvironment of the gut microflora, with respect to both the chemical characteristics and activity of the immune system (Jeyakumar et al., 2019). Conversely, the composition and functional activity of the microbiome can influence the host genetics by different modalities, including interaction through the nervous system, mutual crosstalk with innate immune and the endocrine systems, and direct synthesis and management of neuro-chemicals (Niccolai et al., 2019). Although there is clear evidence of the former, recent studies, through the so-called *microbiome genome-wide association study* (see our previous article; Awany et al. (2019)), have only begun to shed light on the latter interaction.

Together, this plethora of microbiome-modulating factors makes the diversity of the microbiome between individuals much greater than the diversity of their genotypes. Indeed, while individual humans are estimated to be about 99.9% by comparison of their genotypes, they can be up to $\sim 80 - 90\%$ different by comparison of their gut microbiomes (Adams and Gutierrez, 2018). Given this, it is rational to expect that the variation contained within the microbiome can provide important insight into the phenotypic state of the host, just as host genotype. Indeed, as we expound in the next section, the composition and functional diversity of the microbiome has been associated with various host phenotypes.

3.2 The role of the microbiome in host phenotype

The gut microbiome provides a range of beneficial roles to the host, playing critical roles in host's physiology. These include metabolism (e.g. synthesis of micro-nutrients such as vitamins, metabolism of bile acids) (Wang et al., 2017), development of host immunity and regulation of immune response regulation (Erny et al., 2015), and resistance to colonization it confers protection to host against colonization by pathogens (Kho and Lal, 2018), as in *Clostridium difficile* infection where a decrease in dominant microbiota members due to perturbation reduces the colonization resistance capacity, permitting invasion by pathogens, leading to occurrence of the infection (Kho and Lal, 2018).

Besides these beneficial roles, however, the gut microbiome can induce inflammation in the host. As discussed above, microbiome composition is subject to shaping by external factors (such as dietary constituent, antibiotic consumption, xenobiotics including antibiotics and other drugs, psychological and physical stress) as well as host factors. These factors can cause dysbiosis, that is, microbial imbalance (in terms of microbial compositional or metabolic changes). Dysbiosis likely impairs the normal functioning of gut microbiota that maintains the host normal state, culminating to disease states in the host. In infectious diseases, the most common class of diseases that results from dysbiosis of the microbiota (Wang et al., 2017), the pathogenic microbe colonizes the intestinal mucosa, resulting in inflammatory response; for example in *Clostridium difficile* infection (Young, 2017), human immunodeficiency virus (HIV) (Kuss et al., 2011), urogenital infection (Schuijt et al., 2013), and Hepatitis B and Hepatitis C virus (Plottel and Blaser, 2011; Wei et al., 2020; Lv et al., 2019). Dysbiosis of microbiota has also been implicated in a diverse array of autoimmune and inflammatory disorders, including inflammatory bowel diseases (Halfvarson et al., 2017), obesity (Cani, 2019), diabetes (Singer-Englar et al., 2019), atherosclerosis (Ahmad et al., 2019), rheumatoid arthritis (Bergot et al., 2020), pulmonary disease (Dunlap et al., 2019), non-alcoholic fatty liver disease (Safari and Gérard, 2019),

and even mental illnesses (such as schizophrenia, depression, and Alzheimer's disease) (Taylor, 2019). However, for the majority of these diseases, the mechanisms involved are not well understood. Nonetheless, these microbiome-wide association studies have provided evidence for causal roles of specific bacterial taxa on host phenotypes. For example, bacteria such as *Streptococcus gallolyticus*, *Bacteroides fragilis*, *Escherichia coli*, *Enterococcus faecalis*, and *Fusobacterium nucleatum* have high relative abundance in patients with colorectal cancer (after disease onset) as compared with their normal counterparts (Saus et al., 2019). Several cross-sectional and longitudinal studies have also demonstrated that the relative abundance and diversity of bacteria within the *Bacteroidetes* (vs *Firmicutes*) phyla is reduced in obese (vs lean) individuals (Wade and Hall, 2019). It is important to note that for most of these associations, no single species has been found to be universally present among all individuals. Ostensibly, this suggests that the bacterial species do not act in isolation but rather as a consortium to effect disease onset and/or progression.

3.3 Linking microbiome to the host phenotype

The features derived from 16S rRNA sequencing include counts of operational taxonomic units (OTUs), which are sequence reads clustered into taxonomy units, based on sequence similarity. (usually 99% similarity threshold for species, and 97% for genera). A commonly used similarity threshold is 97%, which was established based on an empirical study that showed most strains had 97% 16S rRNA sequence similarity (Nguyen et al., 2016; Konstantinidis and Tiedje, 2005). At this similarity level, the OTUs are taken to approximate 'species', although they do not precisely represent bacterial species. As an aside, the arbitrariness of this similarity threshold and the consequential lack of fine resolution, has recently led to the introduction of similarity-independent finer resolution sequence clustering called amplicon sequence variants (ASVs) (Callahan et al., 2017; Eren et al., 2013; Tikhonov et al., 2015; Eren et al., 2015; Amir et al., 2017).

The central theme in microbiome-wide association study (MWAS) in humans is to determine whether the abundance of particular OTUs, or some microbiome feature in general, is associated with a biological or clinical state. For example, to identify bacterial species whose relative abundance is increased in patients with colorectal cancer as compared to the healthy individuals. The aim, therefore, is to mathematically model the relationship between the observed OTU abundance and the phenotype of interest. However, before this can be done, it is take into consideration the characteristic feature of this OTU counts or abundance (if normalized) dataset [reference will be made to OTUs for the rest of the discussion, as the same does apply to ASV], which are zero-inflation, over-dispersion, non-normality, high dimensionality ($p \gg n$), correlation imposed by the phylogenetic structure of the OTUs, and uneven sequencing depths. In light of uneven sequencing depths, an often first step is to normalize the OTU counts to correct/reduce the bias that would be introduced due to variability in sequencing depth (library size) across the samples (Chen et al., 2018b). These peculiar characteristics make analysis using classical statistical models such as standard linear or logistic regression models inappropriate. For example, classical linear models are known to be inappropriate for zero inflated data due to the violation of homoscedasticity (constancy of variance) and normality assumptions (Loeys et al., 2012). Furthermore, logistic regression where zero counts are treated as non-events that will result in the loss of valuable information and, in effect, reduce the power to detect covariate effects (Xu et al., 2015). Still, while non-parametric methods such as Wilcoxonrank sum (WRS) test may seem an alternative to circumvent the non-normality property (Li et al., 2014), they have the limitation that they cannot incorporate covariates into the analysis model (Loeys et al., 2012); a critical issue given that covariates can confound association. Thus, specialized analytical models are required to address these issues. Accordingly, a suite of statistical models have been proposed. In general, the model applied depends on whether or not the OTU data is normalized as a composition.

3.3.1 Modelling with microbiome as compositional predictor

Compositional microbiome data arise when the generated OTU counts are normalized by dividing the each read count for a given sample by the size factor - a calculated numerical value used as an estimate for the library size - for that sample. An example is a normalization by total sum scaling (TSS), in which the total read count for each sample is used as the size factor, and each OTU count in a sample is divided by the corresponding sample's size factor.

Suppose $\mathbf{y} = (y_1, y_2, \dots, y_n)'$ is the n -vector of responses for n samples from which measurements are made for p OTUs, and let $X = (x_{ij})$, $i = 1, \dots, n$; $j = 1, \dots, p$ denote $n \times p$ matrix of measured OTUs counts. The compositional $n \times p$ matrix for the OTUs is the matrix $Z = (z_{ij})$, where $z_{ij} = f(x_{ij})$ such that $\sum_{j=1}^{p-1} z_{ij} = 1$ for all i , and $z_{ij} \geq 0$. In other words, the composition makes each row of Z to lie in the $(p - 1)$ -dimensional positive simplex

$$\mathcal{S}^{p-1} = \left\{ (z_{i1}, z_{i2}, \dots, z_{ip}) \mid z_{ij} \geq 0, \sum_{j=1}^p z_{ij} = 1 \right\}.$$

The unit sum constraint (to the simplex space) makes it impossible to alter one component without altering at least one of the other components. This property renders classical statistical methods, where the coordinates are defined in real Euclidean space, inappropriate (Aitchison, 1982). There are two general methods to proceed. The first approach, pioneered by Aitchison (1982), is to apply a transformation (commonly log-ratio transformation) to the compositional data which then transforms the data points from the simplex space to the Euclidean space, from where classical methods such as linear regression can be applied. Based on this, Lin et al. (2014) developed a linear regression for analysis of microbiome compositional data. This model was later adopted and extended by Shi et al. (2016) to include compositional and sub-compositional analyses,

under a regularized regression. One severe limitation of methods based on log-transformation is that they are incompatible with zero values in the dataset, since logarithm of zero is undefined. Often zero values are assumed to be resulting from rounding or measurement errors, so that zeros are replaced by, say, the maximum rounding error of 0.5 (Aitchison, 2003; Kurtz et al., 2015; Shi et al., 2016; Mandal et al., 2015). Clearly, this can have undesirable impact on the association inference when the data has many zeros, and especially physical zeros, as is the case here.

The alternative method is based on recognizing that the support of a Dirichlet distribution is the simplex, so that a model based on the Dirichlet distribution can be directly applied to the data in its compositional form Wadsworth et al. (2017). An example for this is the zero-inflated generalized Dirichlet multinomial (ZIGDM) regression model developed by Zhang et al. (2017). Although appealing because of no requirement to transform the data, the Dirichlet distribution-based models intrinsically impose a negative correlation among the OTUs, yet the actual data may display positive and negative correlations (Mandal et al., 2015). Moreover, it may not sufficiently model zero-inflation (Zhang et al., 2017).

Alternatively, kernel-based regression method has been used to achieve this goal. In this, a distance measure is defined between any two microbiome samples. Commonly used distance measures are the UniFrac (weighted, unweighted, variance-adjusted or generalized UniFrac distances) and the Bray-Curtis distances (Bray and Curtis, 1957). The UniFrac is a phylogenetic-based distance that measures the phylogenetic distance between sets of taxa in a phylogenetic tree as the fraction of the branch length of the tree that leads to descendants from either one biological condition or the other, but not both (Lozupone and Knight, 2005; Chen et al., 2012a), whereas the Bray-Curtis is measure of degree of community similarity and dissimilarity based on the number and abundance of species shared by two communities and the number and abundance of species unique to each of them. The general idea for this kernel-based regression method is as follows. Suppose $\mathbf{y}$ and X are as defined above, and $\mathbf{w}_i = (w_{i1}, w_{i2}, \dots, w_{iq})'$ be a q -dimensional vector of covariates. Then, the OTUs are related to the phenotype by linear kernel

machine model

$$y_i = \alpha_0 + \mathbf{w}_i \boldsymbol{\alpha} + f(X_i) + \epsilon_i \quad (3.1)$$

for a quantitative phenotype, and by the logistic kernel machine model

$$\text{logit}(P(y_i = 1)) = \alpha_0 + \mathbf{w}_i \boldsymbol{\alpha} + f(X_i) \quad (3.2)$$

for a binary phenotype; where α_0 is intercept, ϵ_i is residual, and $f(X_i)$ is a function from a reproducing Kernel Hilbert space, such that $f(X_i) = \sum_{j=1}^n \lambda_j K(X_i, X_j)$ for some scalars $\lambda_1, \dots, \lambda_n$ and Kernel function $K(., .)$. The kernel-based methods currently implemented in the microbiome analysis include MiRKAT (Zhao et al., 2015) and the recent extension GLMM-MiRKAT (Koh et al., 2019). One key concern with the kernel-based methods is the fact that performance can be influenced by the choice of the distance measure Chen et al. (2012a). Moreover, it is unclear, if not impossible, how variable selection or variable weighting, which can alleviate noise accumulation especially in the high-dimensional microbiome setting, can be conducted (Wu et al., 2016a). Still, they are primarily designed to address the question of whether entire microbial community is associated with a phenotype, and thus cannot be used to answer the equally, if not more, interesting question of which specific taxa are associated with the phenotype.

3.3.2 Modelling with microbiome as non-compositional predictor

Although the majority of microbiome count data normalization methods lead to compositional data, that has the undesirable properties mentioned above, some recently proposed methods, the cumulative sum scaling (CSS) (Paulson et al., 2013) and the more recent geometric mean of pairwise ratios (GMPR) (Chen et al., 2018b) normalization, lead to normalized datasets where the data vectors are unconstrained (non-compositional). Since the coordinates of the

resulting data vector lie in the usual real Euclidean space, any suitable classical statistical model can be applied. Of course, this does not mean that the modelling complexity is henceforth simplified or addressed. Any applied statistical inference and estimation of the parameters for OTU effects ought to take into consideration the issue of zero-inflation, high dimensionality, over-dispersion, and structure among bacterial taxa at various phylogenetic depths. [Xiao et al. \(2018b\)](#) recently proposed a regularized regression framework, refereed to as ‘SICS’, that can be applied to microbiome abundance data, following GMPR normalization. A particularly appealing feature of their method is the ability to regularize the coefficients of bacterial taxa with respect to the phylogenetic tree, which allows incorporation of biologically-important information on linkage among OTUs into the analysis. Despite the gain in power, the reliance of the method phylogenetic tree smoothing makes it prone to loss in power when the phylogenetic tree is misspecified or uninformative with respect to OTU effects on the phenotype under study ([Xiao et al., 2018b](#)). Additionally, the current exposition of the phylogeny-based smoothness penalty maybe somewhat unfair since a single universal weight parameter is used to equally penalize all OTUs. This issue is especially important when the data contains many outliers, as is frequently the case in microbiome datasets ([Chen and Chen, 2017](#)). The consequence of which is that the estimation maybe negatively biased toward less abundant OTUs. However, we also reckon the mathematical complexity and computational burden introduced by defining OTU-specific penalties, and we are unsure whether the benefits gained by adopting such a strategy would not be over-weighed by the increased computational complexity, given its direct scaling with the number of OTUs as well as sample size.

Besides unconstrained normalized OTU data, statistical models have been devised to directly model OTU counts. This class of models primarily hinges on the fact that counts data follow a multinomial distribution. The basis ideas is as follows. Suppose $(n_1, n_2, \dots, n_p)$ are the observed OTU counts, as obtained from 16S rRNA sequencing, and let $n = \sum_{j=1}^p n_j$ be the total count of all OTUs. To specify the

distribution of these counts by a multinomial distribution, we have to posit that the observation of all the OTU counts $(n_1, n_2, \dots, n_p)$ is governed by a probability distribution. Let $\phi = \phi_1, \phi_2, \dots, \phi_p$ be probability density. This technically means that the pdf $f(\phi)$ is defined only in the $(p - 1)$ -dimensional space, and not p , since $\sum_{j=1}^p \phi_j = 1$. A corollary to this is that $\phi_j \neq 1 \ \forall j$. The probability mass function for the counts, which we denote by $f(n_1, n_2, \dots, n_p; \phi_1, \phi_2, \dots, \phi_p; n)$, is then given by

$$f(n_1, n_2, \dots, n_p; \phi_1, \phi_2, \dots, \phi_p; n) = \frac{\Gamma(N + 1)}{\prod_{j=1}^p \Gamma(n_j + 1)} \prod_{j=1}^p \phi_j^{n_j}.$$

However, because of the microbiome features mentioned above, particularly overdispersion, correlation and zero-inflation, extensions of the multinomial distribution such as negative binomial (Zhang et al., 2017), zero-inflated Poisson (Xu et al., 2019), beta-binomial (Hu et al., 2018), and Dirichlet multinomial (Chen and Li, 2013) have been adopted in lieu of the multinomial itself. For example, Chen and Li (2013), in a bid to account for overdispersion in observed OTU counts, extended the multinomial model to the Dirichlet multinomial model; which is essentially the multinomial distribution, where the vector of probabilities $(\phi_1, \dots, \phi_p)$ are modeled as realizations from a Dirichlet distribution.

As the preceding section makes clear, current strategies for modelling the link between bacterial OTUs and host's phenotype depends on whether the OTUs abundance data are counts, or compositional. The application of these models have led to several exciting discoveries, providing clear evidence for the role of bacterial taxa in the pathogenesis of complex traits/diseases. Despite this, however, there are some limitations with current suite of statistical frameworks used to explore such associations. The limitations mainly emanate from overdispersion, zero-inflation, non-normality, and high dimensionality features that characterizes microbiome datasets.

3.4 Interaction of host genome with microbiome

As discussed earlier, it has been demonstrated and established, through microbiome-wide association studies, that variation in the microbiome is involved in many host's traits and diseases. Interestingly, these same traits and diseases are also influenced by host genetics polymorphisms, as established by the over a decade-long genome-wide association study of complex human traits and diseases. This observation makes it apparent that the microbiome and host genetics are not mutually exclusive. Therefore, exploring the interaction of host genetics with the microbiome is a precursor to understanding host's phenotypic variation.

A number of studies have investigated the influence of human genetic variation on the microbiome. Earlier studies geared to uncovering this link applied the candidate gene study approach, focusing on specific genes of interest, specifically those implicated in disease susceptibility (Rausch et al., 2011; Khachatryan et al., 2008). The findings from these studies were convergent, with specific host genes being found to explain variation in microbiome's community composition, diversity, and structure. Using twin-based studies, similar findings have also been obtained. Zoetendal et al. (2001) and Stewart et al. (2005), in some of the early studies, using Denaturing gradient gel electrophoresis (DGGE) technique, found a correlation between overall genetic relatedness and gut microbiome community similarity, indicating that host genotype has a significant effect on determining the microbiome community structure. Later, subsequent studies performed using the more powerful 16S rRNA sequencing technique arrived at similar conclusions. For example, Turnbaugh et al. (2009), employing complementary phylogenetic and taxon-based methods, reported that human gut microbiome is shared among family members, but the degree of inter-person's variability in the specific bacterial lineage present is smaller in monozygotic than in dizygotic twins.

Inspired by the apparent success of the genome-wide association technique in human disease association studies, recent studies of host's genotype-microbiome interaction have been performed on a genome-wide scale, in what is now dubbed

‘microbiome genome-wide association study’ (mGWAS). For purposes of this analysis, the microbiome is then considered as a trait influenced by genetic variation in the host (Awany et al., 2019; Luca et al., 2018). Analogous to GWAS in the host, mGWAS seeks to find the host’s genetic polymorphisms that are associated with, and therefore could be causal to, microbiome ‘attributes’. The microbiome attribute generally refer to any measurable property of the microbiome, and thus typically encompass taxonomic composition, relative abundance, and functional diversity. In principle, at least four types of metrics derived from host and microbiome can be used for association modelling. First, the microbial abundance at a given taxonomic level can be used as a microbiome phenotype, and directly regressed with the host SNPs in the same manner as in classic GWAS for quantitative trait mapping. Second, beta-diversity metrics (which defines diversity between samples or communities) calculated for the overall bacterial community can be used as the phenotypic metric to be correlated with host SNPS. Third, alpha-diversity metrics (which defines diversity within a sample or community) can be computed, and directly regressed with host SNPS. Moreover, if insight is required into the biological significance of the host’s genotype-microbiome association, a correlation analysis can be performed between host’s SNPs and microbial functional units or pathways.

In a study of the impact of host genetic variation on microbiome composition, Blekhman et al. (2015) performed the first mGWAS using 93 individuals whose genetic and corresponding microbiome information were gleaned from the Human Microbiome Project (HMP) (Turnbaugh et al., 2007). Their study identified significant associations between host genes and microbiome composition across multiple body parts; the implicated SNPs being enriched for genes located mainly in immunity-related pathways. Interestingly, the genes found to be associated with bacterial abundance overlap those genes previously discovered through host GWAS to be associated with auto-immune diseases such as meningitis, and various forms of cancer. Although the sample size was relatively small (relative to typical sample sizes used host GWAS), the results of this ground-breaking study highlighted the critical role of host genetic polymorphisms in modulating

microbiome composition, further adding evidence to the accumulating body of literature that host genetics and microbiome are not independent, and suggesting that an understanding of the signatures of host genetic-microbiome interaction in disease-specific scenarios may provide crucial insights into the trajectory of disease pathogenesis.

Following this pioneering study, a number of other studies followed suit. For example, [Goodrich et al. \(2016\)](#) extending on the original study in ([Blekhman et al., 2015](#)) after increasing the sample size (to $n = 1,126$ twin-pairs), replicated some of the associations and identified other microbiome-associated genes. With the increased power, the authors obtained an association between *Bifidobacterium* and the lactase (*LCT*) gene locus, as well as the link between the host gene *ALDH1L1* and *SHA-98* bacteria. Additionally, the expanded study re-established that bacterial members of the family *Christensenellaceae* was the most heritable, whereas members of the phylum *Bacteroidetes* were generally not heritable. Meanwhile, [Hua et al. \(2015\)](#) using 147 samples, with microbiome data in the context of lung cancer found association of alpha-diversity, beta-diversity, and bacterial abundance with variation in microbiome composition; although no loci attained genome-wide significance ($P < 10^{-8}$) after correcting for skewness and kurtosis in the test statistics.

Based on these initial studies, it is apparent that host genetics can structure microbiome variation, just like it does on other human complex traits. However, so far the impact of host genetics on the microbiome is still poorly understood, particularly the strength of the effect, the functional significance of the genes and pathways involved with regards to host phenotype, and the context- and environmental-dependence of this association. This is demonstrated by observations that the study findings have been contrasting ([Rothschild et al., 2018](#)), and results across different cohorts have not had a high degree of overlap ([Burns and Blekhman, 2019](#)). These discrepancies might be due to methodological, statistical, or conceptual issues; entailing and including sample size, statistical power, and confounding effects. So far, most such studies have included far fewer participants (in the range $\sim 100 - 3000$). This is very low in comparison to the sample

sizes (typically in tens to hundreds of thousands) required to discover disease-associated variants in host GWAS of complex diseases. If a microbiome attribute is treated as a complex trait, the power to discover statistically significant genetic effects critically depends on the underlying genetic architecture of the trait, that encompasses the number of causal loci and the corresponding effect sizes. It is therefore possible that the current challenge with replication and/or identification of new variants might relate to the low effects sizes of microbiome traits. Indeed, the effect sizes (variation in beta-diversity explained) obtained in mGWAS have been low, for example $\sim 0.29 - 0.49\%$ in the study by [Wang et al. \(2016\)](#). Based on this architecture, exceptionally large sample sizes will be required to detecting significant microbiome-associated SNPs in future mGWAS.

The likely small effect size of microbiome traits also present yet another challenge for statistical modelling frameworks. The current methodological paradigm for host GWAS, though effective for genetic variations of large effect sizes, has relatively low sensitivity for small genetic effects ([Gundogdu and Nalbantoglu, 2017](#); [Bush and Moore, 2012](#)). Thus, although it makes pragmatic sense to port statistical frameworks designed for host GWAS directly to mGWAS analyses, such frameworks may not be ideally suited. Therefore, even though some of the mGWAS carried using conventional host GWAS frameworks, for example in the study by [Blekhman et al. \(2015\)](#), [Bonder et al. \(2016\)](#) and ([Davenport et al., 2015](#)), have been successful at detecting association signals, such associated-genetic variants might only be ‘low-hanging-fruits’. That being said, it is pleasing to note that initial forays into this methodological adaptation have been by the development of two tools custom-made for mGWAS ([Lynch et al., 2017](#); [Hua et al., 2015](#)).

Another issue is the presence of confounding effects. It is well known that the microbiome can be modulated by a diverse array of ‘external’ factors, including as age, diet, social interaction, and environmental conditions. The basis premise when conducting association analysis is that the distribution of the trait is the same in the sampled population. If the subjects in the study population are subject to the similar external factors, then any discovered association would be largely

due to genetic polymorphisms in the host. In practice, however, these external factors can differ markedly, meaning that the study population would differ with respect to their microbiome structure (bacterial trait). This is exemplified in a recent study of 1,046 healthy individuals with several distinct ancestral origins that found the environment to, in fact, dominate host genetics in shaping the microbiome (Rothschild et al., 2018). Heterogeneity in such factors can mask host genetic effects in shaping the gut microbiome, which can lead to spurious association. Of course, the way out is to either control for such confounding sources in the analysis or eliminate them through a better study design. However, irrespective of the strategy taken, it is nearly impossible to completely eliminate such confounding sources. Nonetheless, the use of better statistical models, such as random effects models, that can incorporate random variations into the analysis, would help reduce confounding effects.

To conclude, the advent of high-throughput sequencing technology, which allows rapid, high-throughput, and culture-independent profiling of microbial communities, has led to the re-discovery and appreciation of ‘our second genome’ - the microbiome. Microbiome-wide association studies have shown that the gut microbiome plays an important role in a plethora of traits and disease states ranging from neuro-developmental disorders to chronic gastro-intestinal diseases. At the same time, initial forays into microbiome genome-wide association studies has made it apparent that host genetic variation can shape the taxonomic composition and functional diversity of the microbiome. Of note from these disparate studies is that the microbiome and host’s genetics are not mutually exclusive relative to the host’s phenotype, microbial taxa are heritable, and the host’s SNPs associated with variability in the microbiome are enriched for genes in immunity-related pathways and gastro-intestinal tissues. This is exciting given that genes implicated in many of the common complex traits are also mainly in these same host pathways and tissues; thus pointing to their functional relevance and providing key insights into possible mechanism in complex diseases.

Despite success and optimistic outlook, however, the specific contribution of the gut microbiota to traits and diseases is largely unclear. Elucidating the details of

the gut microbiome's involvement in human complex traits and diseases holds promise for use in medical diagnosis and treatment based on an individual's genotype and microbiome medicine. However, to reach this optimistic outlook, research efforts must be directed towards delineating true associations from the deluge of potential confounding factors [these factors that vary both within individuals and across population factors, can directly modulate the composition and relative abundance of the microbiome (the predictors)], transitioning from association to causation, and establishing the biological mechanisms underlying the causative factors. In this pursuit, there is need for a paradigm shift towards integrative disease modelling, by which, in particular, host GWAS and microbiome studies are not treated as parallel fields. In our previous articles (in the sequel in this thesis), we discussed the current state and challenges in association mapping in the host and microbiome, and proposed how host genetic information may be combined with the microbiome to provide a more-complete picture of the factors and mechanisms underlying phenotypic variation.



Host and Microbiome Genome-Wide Association Studies: Current State and Challenges

Denis Awany^{1†}, Imane Allali^{2†}, Shareefa Dalvie³, Sian Hemmings⁴, Kilaza S. Mwaikono², Nicholas E. Thomford¹, Andres Gomez⁵, Nicola Mulder² and Emile R. Chimusa^{1*}

¹ Division of Human Genetics, Department of Pathology, Institute of Infectious Disease and Molecular Medicine, Faculty of Health Sciences, University of Cape Town, Cape Town, South Africa, ² Computational Biology Division, Department of Integrative Biomedical Sciences, Institute of Infectious Disease and Molecular Medicine, Faculty of Health Sciences, University of Cape Town, Cape Town, South Africa, ³ Department of Psychiatry and Mental Health, University of Cape Town, Cape Town, South Africa, ⁴ Department of Psychiatry, Faculty of Medicine and Health Sciences, Stellenbosch University, Cape Town, South Africa, ⁵ Department of Animal Science, University of Minnesota-Twin Cities, St. Paul, MN, United States

OPEN ACCESS

Edited by:

Sara Leslie Pulit,
University Medical Center Utrecht,
Netherlands

Reviewed by:

Omer Weissbrod,
Harvard T.H. Chan School of Public
Health, United States
Alexander Teumer,
University of Greifswald, Germany

*Correspondence:

Emile R. Chimusa
emile.chimusa@uct.ac.za

[†]These authors have contributed
equally to this work as first authors

Specialty section:

This article was submitted to
Human Genomics,
a section of the journal
Frontiers in Genetics

Received: 24 August 2018

Accepted: 27 November 2018

Published: 22 January 2019

Citation:

Awany D, Allali I, Dalvie S,
Hemmings S, Mwaikono KS,
Thomford NE, Gomez A, Mulder N
and Chimusa ER (2019) Host
and Microbiome Genome-Wide
Association Studies: Current State
and Challenges. *Front. Genet.* 9:637.
doi: 10.3389/fgene.2018.00637

The involvement of the microbiome in health and disease is well established. Microbiome genome-wide association studies (mGWAS) are used to elucidate the interaction of host genetic variation with the microbiome. The emergence of this relatively new field has been facilitated by the advent of next generation sequencing technologies that enable the investigation of the complex interaction between host genetics and microbial communities. In this paper, we review recent studies investigating host-microbiome interactions using mGWAS. Additionally, we highlight the marked disparity in the sampling population of mGWAS carried out to date and draw attention to the critical need for inclusion of diverse populations.

Keywords: genome-wide association study, microbiome, microbiome-GWAS, host-genetic, host-microbiome interaction

INTRODUCTION

The past two decades have seen tremendous advancement in our understanding of human genetic variation and its implication in health and disease. This has, in part, been facilitated by extensive scientific collaboration and the exponential increase of technical and methodological advancements (Figure 1). Examples of notable large scale scientific collaboration include the Human Genome Project (Sawicki et al., 1993) which published the DNA sequence of the entire human genome; the International Haplotype Map (HapMap) Project (Thorisson et al., 2005) which cataloged the patterns of common polymorphisms (typically minor allele frequency (MAF) larger than 1%) in the human genome and its linkage disequilibrium (LD) structure across multiple ancestral populations. Further, advances in genotyping made feasible, at a relatively low cost, the genotyping of hundreds of thousands (or even millions) of common variants across the human genome. Together these factors catapulted the genome-wide association study (GWAS) in humans (herein referred to as “host”) population.

Genome-wide association study in host populations (hGWAS) has identified hundreds of genetic variants associated with many complex human traits and diseases, novel biological

mechanisms and drug targets for infectious and non-infectious diseases (Relling and Evans, 2015). The microbiome, which is the collection of bacteria, archaea, fungi, protozoa, and viruses that colonize our body surface and their respective genome (Blum, 2017), has shown to play a major role in human health and disease. The success of hGWAS approach provided an optimistic outlook for and eventual implementation to the microbiome. The microbiome genome-wide association study (mGWAS) aims to identify the host's genetic polymorphisms that interact with its microbiome. Recently, mGWASs have identified and validated many heritable bacterial taxa, including the *Christensenellaceae* and *Methanogens* families (Goodrich et al., 2017). Moreover, mGWAS has linked host genotypes and identified pathways with inter-individual variability in microbiome composition in states of health and disease (Hall et al., 2017; Imhann et al., 2018). These findings corroborate the common view that the microbiome plays a significant role in a host's traits, disease susceptibility and resistance, and treatment response.

Even though multiple lines of evidence have indicated significant host-microbiome interactions (Goodrich et al., 2017; Kurilshikov et al., 2017; Weissbrod et al., 2018), the relative strength of these interactions is unclear, with studies yielding somewhat contrasting results (Rothschild et al., 2018). This is perhaps unsurprising given the plasticity of the microbiome to external factors. In light of this, a key, and yet challenging task, is the establishment of truly causative factors in the observed associations between the environment, host genetics and the microbiome when investigating complex traits and diseases. Including the various microbiome data types, that is, proteomic, metabolomic and transcriptomic, to complement the current mostly used genomic data may help illuminate these interactions. However, combining complex and high dimensional data is not straight forward, introducing yet another challenge. In addition, as in the case of hGWAS (Peprah et al., 2015), the existing disparity in microbiome research, in terms of genomic diversity of the sampling population, further thwarts insights into the complex host-microbiome-environment interaction. As depicted in **Figure 2** and elaborated in the section “disparity in host-microbiome GWAS,” there is a striking lack of genomic diversity of the study populations in mGWAS published to date.

Here, we review recent studies investigating host-microbiome interactions, through the concept of mGWAS. Then, we highlight the marked disparity in the sampling population of mGWAS carried out to date and draw attention to the critical need for inclusion of non-European populations. Finally, we explore some pertinent challenges in mGWAS.

HUMAN VARIATION

In the realm of genetics, human variation, the variation in allele and/or allele frequency, is inherent in all human populations and underlies population differences in many phenotypic expressions, including resistance and susceptibility to diseases. Genomic variation ranges from large microscopic rearrangements such as insertions and deletions, to smaller submicroscopic variations

such as single-nucleotide polymorphisms (SNPs) and copy number variation (CNV). Analysis of human genomic variation in the 1000 genomes project reported that a typical genome contains ~4.1–5.0 million variants, of which >99.9% are SNPs and indels (1000 Genomes Project Consortium et al., 2015). It is important to note, however, that despite being rare (MAF < 0.5%), structural variants affect more bases, have larger effect sizes (Chiang et al., 2017), and are also thought to be potentially involved in disease pathogenesis due to their enrichment for changes that alter protein sequence and function (Casals and Bertranpetit, 2012; Nelson et al., 2012). Thus, in pursuit of unraveling host genetic variants that interact with the microbiome, knowledge of the abundance and distribution of genetic variants along the genome is critical for characterizing the genetic architecture of common as well as rare traits/diseases, and discerning functionally important variations from the myriad of genomic polymorphisms.

With recent technological advances in genotyping arrays and next-generation sequencing (NGS), it is becoming increasingly feasible to conduct large-scale studies, presenting the opportunity for discovery of both rare and common genetic variation. Deep sequencing offers the opportunity to uncover the complete repertoire of these variations (Marth et al., 2011). However, performing deep sequencing on a genome-wide scale is currently limited due to relatively high cost. To this end, whole-exome sequencing and, particularly, genotype arrays have become methods of choice in the geneticist's arsenal. Intriguingly, these studies, particularly those targeted to protein-coding genes, have revealed the existence of multitude of variants at population and individual level that disrupt protein-coding genes in every human genome (Reijnders et al., 2018), some of which having different phenotypic effects (Blauwendraat et al., 2018; Grarup et al., 2018). These variants generally referred to as loss or gain-of-function variants, occur at low frequency in the genome, and have gene-disrupting ability (MacArthur and Tyler-Smith, 2010); which result in their implications for clinical interpretation of genomic sequences (MacArthur and Tyler-Smith, 2010). It is clear that with NGS technology, many novel genomic variants will be unveiled which, in effect, will facilitate the development of a comprehensive catalog of human genetic diversity.

HUMAN MICROBIOME DIVERSITY

The human microbiome exhibits both intra- and inter-individual variability (Lax et al., 2014; Gupta et al., 2017). Studies on twins have shown that the microbiota of identical twins are more similar compared to that of their siblings. Also, siblings have a more similar microbiota than that of unrelated individuals (Goodrich et al., 2014). Similar to genetic variation, the human microbiome plays an important role in health and disease (Blekhman et al., 2015; Hall et al., 2017; Gilbert et al., 2018). The microbiome has been associated with variants in host genes involved in immunity and metabolism (Blekhman et al., 2015). The human immune system has a complex

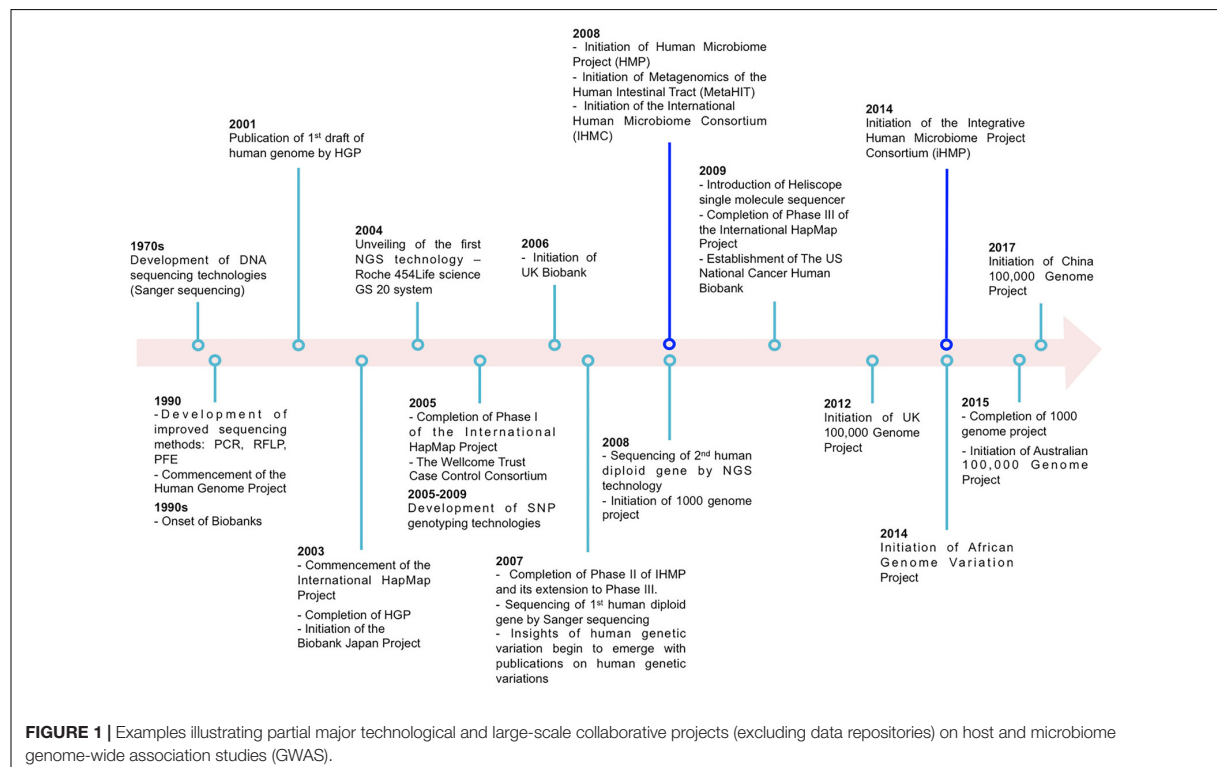


FIGURE 1 | Examples illustrating partial major technological and large-scale collaborative projects (excluding data repositories) on host and microbiome genome-wide association studies (GWAS).

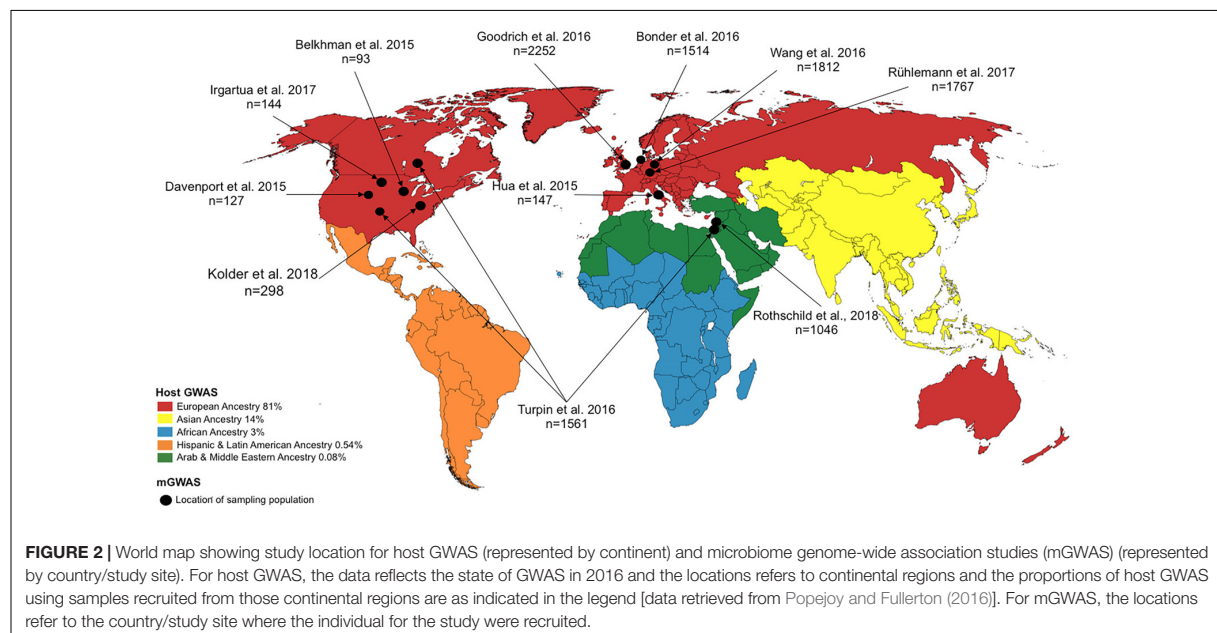


FIGURE 2 | World map showing study location for host GWAS (represented by continent) and microbiome genome-wide association studies (mGWAS) (represented by country/study site). For host GWAS, the data reflects the state of GWAS in 2016 and the locations refers to continental regions and the proportions of host GWAS using samples recruited from those continental regions are as indicated in the legend [data retrieved from Popejoy and Fullerton (2016)]. For mGWAS, the locations refer to the country/study site where the individual for the study were recruited.

bidirectional relationship with the microbiome. It has been shown that the microbiome is associated with the variability of the immune responses and these responses may also be

involved in modifying the microbiome itself (Belkaid and Hand, 2014; Schirmer et al., 2016; Takiishi et al., 2017). Furthermore, human genetic polymorphisms at various loci are

hypothesized to interact with each other and with an individual's microbiome to impact disease (Blekhman et al., 2015; Hall et al., 2017). In particular, mutations in host genes can influence its interaction with the compositional and functional diversity of the microbiome, potentially modulating an individual's susceptibility to disease (Sandoval-Motta et al., 2017b). In healthy individuals, the microbiome composition is balanced (Rajilic-Stojanovic et al., 2009; Lloyd-Price et al., 2016), and imbalance is now known to be associated with clinical conditions such as diabetes, and inflammatory bowel disease (van Tongeren et al., 2005; Upadhyaya and Banerjee, 2015). Besides host genetics, other factors have been associated with microbial community composition including diet and antibiotic consumption. Many studies have reported that both long- and short-term diet can influence the microbiome composition (Wu et al., 2011; Conlon and Bird, 2014). For example, high-carbohydrate diets have been associated with prevalence of *Prevotella*, while *Bacteroides* are associated with high-fat and high-protein diets (Singh et al., 2017). Additionally, consumption of antibiotics may shift the microbiome composition to a temporally quasi-stable state. This state can be either capable of reverting back to the initial state, or to an alternative irreversible post-antibiotic dysbiosis state (Lozupone et al., 2012). This dysbiosis state is characterized by a loss of taxonomic and functional diversity, which may shift the host's metabolic capacity and reduce the colonization resistance against invading pathogens (Langdon et al., 2016; Lange et al., 2016).

HOST AND MICROBIOME GWAS

Host Genome-Wide Association Studies

Genome-wide association study aims to determine the link between genotypic and phenotypic variabilities. This is achieved by obtaining genome-wide genotypic data and phenotypic measurements from a number of subjects, and comparing the frequency of these variants across phenotypic values. GWAS has undoubtedly had successes, identifying thousands of genetic variants associated with hundreds of traits (Visscher et al., 2017), providing valuable insights into the genetic basis of many common traits. Due to LD, any identified associated variant is not necessarily causal as it may simply be "tagging" the causal variant. In addition, most genomic variants are located outside protein-coding regions and are of unknown biological functions (Yang et al., 2017). Consequently, for most traits, little is known about the biological mechanism underlying the associations detected by GWAS.

A critical step toward the elucidation of the underlying biological mechanism is to discern the causal variants. Pinpointing the putative causal variant is, however, challenging for several reasons, including the fact that: (i) most risk regions encompass and implicate multiple variants in the case of complex traits, which without the functional information of the variants, makes it extremely difficult to pinpoint the true causative variant, and (ii) risk variants may reside outside risk regions, and their effects are propagated through regulatory elements (Lin et al., 2016). To this end, several post-GWAS approaches have been

introduced (Wang et al., 2010), driven by the need to leverage GWAS summary statistics to account for polygenicity at the SNP, gene or pathway levels to determine the functional role of the identified variants, uncover their biological mode of action and illuminate their regulatory mechanism (Chen et al., 2015; Chimusa et al., 2015).

Also pertinent to host GWAS is the "missing heritability" problem, which describes the observation that the proportion of heritability explained by the GWAS-associated variants is much less than calculated from familial studies. The reasons for this are still unknown and remains controversial (Marian, 2012; Zaitlen and Kraft, 2012; Sandoval-Motta et al., 2017b), with possible reasons being cited to include epistasis, epigenetics, small effect sizes of the variants, poor coverage of genetic variations on genotyping platforms (Hou and Zhao, 2013). Meanwhile, some researchers have attributed this discrepancy to the fact that GWAS only accounts for genetic variation in human cells and does not consider the effects of the microbiome on phenotype (Marian, 2012; Sandoval-Motta et al., 2017a,b). In light of this, incorporating the microbiome into host GWAS has been hypothesized to significantly reduce the missing heritability gap for microbiome-associated traits (Sandoval-Motta et al., 2017b). However, given that it is not yet known why microbiome is more similar in monozygotic than in dizygotic twins, incorporating other potential sources of variability such as diet, behavior – which are usually assumed to be homogenous across the subjects – may help to explain the missing heritability.

Microbiome Genome-Wide Association Studies: Approaches and Applications

The microbes that inhabit the human body exist in a synergistic relationship with the human host, performing several important roles in metabolism, detoxification, homeostasis, immunity and epithelial development (Tremaroli and Bäckhed, 2012; Belkaid and Hand, 2014). The human microbiome composition varies widely across different body sites and shows some stability at adulthood for the predominant bacterial communities (Blekhman et al., 2015; Lloyd-Price et al., 2016). Host-microbiome interactions soon establish an equilibrium, which determines the state of health of an individual (Tremaroli and Bäckhed, 2012; Blekhman et al., 2015). Thus, understanding the interactions of both the microbiome and host genetics may provide more insights on disease diagnosis, treatment, and prevention. A number of studies have linked the human microbiome at the various body sites to the development of a wide range of complex traits and diseases, including weight gain, obesity, inflammatory bowel disease, diabetes, cardiovascular disease, cancer, major depression, autism spectrum disorder, and asthma (Kostic et al., 2014; Jiang et al., 2015; Hall et al., 2017; Allali et al., 2018; Qiao et al., 2018). These phenotypic expressions are related to the changes in the overall taxonomic composition, as well as presence or absence of specific bacterial.

Owing to the role played by the microbiome in the pathogenesis of many diseases, there has been a surge of interest in understanding host DNA sequence variations that modulate

the human microbiome (Marchesi et al., 2016). Early insights into interaction between host genome and microbiome were obtained from animal based studies (Blekhman et al., 2015; Wang et al., 2016). For example, Rawls et al. (2006) showed that the observed difference between the microbiotas of zebrafish and mice is due to the underlying host genetics. Host genetic loci that shape diversity in skin microbiota and confer susceptibility to disease in mice were also identified (Srinivas et al., 2013). In humans, studies involving monozygotic and dizygotic twins have shown that the abundance of certain microbial taxa are more correlated amongst monozygotic than with dizygotic twin pairs indicating that host genetic factors are involved in modulating gut microbiome composition across human populations (Goodrich et al., 2014; Xie et al., 2016).

The observations above motivated the advent of microbiome genome-wide association study (mGWAS). Using microbiome attributes (such as alpha diversity, beta diversity or relative abundance of bacterial taxa) as the response variable and host's genotype data as the explanatory variable. mGWAS measures and analyses DNA sequence variations across the host's genome in order to identify genetic factors that modulate the composition and functional diversity of the microbiome. To date, studies published on mGWAS have raised interest and provided new insights (Blekhman et al., 2015; Davenport et al., 2015; Hua et al., 2015; Bonder et al., 2016; Goodrich et al., 2016; Turpin et al., 2016; Wang et al., 2016; Igartua et al., 2017; Rühlemann et al., 2018; Rothschild et al., 2018). The three first studies using mGWAS were conducted on a relatively small sample size. Blekhman et al. conducted the first mGWAS in 93 individuals using human microbiome data and host genetic information gleaned from the Human Microbiome Project (Gilbert et al., 2010); microbiome data and host DNA were from 15 body sites (Blekhman et al., 2015). The authors identified significant associations between several host genes and pathways with microbiome composition. Following this work, Davenport et al. (2015) reported the second mGWAS which investigated host genetic effects on the gut microbiome of 127 Hutteries (North America) and found host SNPs are associated with the abundance of several bacterial taxa. In the third study, Hua et al. have developed the microbiome-GWAS tool that has been tested on 16S rRNA microbiome data from 147 non-malignant lung tissue samples (Yu et al., 2016) to establish the microbiome composition in terms of cancer risk SNPs. The authors found significant associations between six previously established lung cancer risk SNPs and microbiome composition. Subsequent studies have used larger samples of ~300–2000 individuals and have reported significant (Bonder et al., 2016; Goodrich et al., 2016; Turpin et al., 2016; Wang et al., 2016). However, Kolde et al. (2018) did not identify significant associations in their untargeted genome-wide analysis in contrast with the findings of Blekhman et al. who used the same cohort and have reported 83 significant associations. Kolde et al. (2018) have reported that the main reason for this difference is the choice of significance thresholds; they used a more stringent Bonferroni correction while Blekhman et al. used false discovery rate (FDR) multiple hypothesis test correction. Additionally, a recent study of 1046 healthy Israeli individuals, with several

different ancestral origins and who share a relatively common environment, did not find any significant associations between (Rothschild et al., 2018). The results of the above studies suggest that some bacterial taxa are heritable but the results of one study cannot be replicated except for the bacterial taxa *Bifidobacterium* which were found to be significantly associated with the lactase *LCT* gene locus (Blekhman et al., 2015; Bonder et al., 2016; Goodrich et al., 2016; Rothschild et al., 2018). **Table 1** summarizes the mGWAS carried out to date.

mGWAS Approaches and Tools

There are many microbiome attributes that may be leveraged as phenotypes for a mGWAS. First, alpha diversity, that is, the diversity of species within community samples (Wilson and Shmida, 1984) may be used as phenotype and an association performed against host genotypes. Second, beta diversity, that is, the diversity between community samples (Wilson and Shmida, 1984), defined using phylogeny-informed or taxa abundance-informed pairwise distance measures, may be used as phenotype. It is important to note, however, that because the microbiota functions as a community, cross-sample analysis using beta diversity measure is more robust compared to alpha diversity (Hua et al., 2016). Third, the relative abundance of each taxon at a given taxonomic level (species, genus, family, order, class, and phylum) may be used as phenotype, and analysis performed to assess the association of each SNP with the taxon. Alternatively, with shotgun metagenomic sequencing that, unlike the 16S rRNA approach, provides functional information, bacterial pathways may be used as trait for mGWAS. The varied and peculiar features of microbiome phenotypes, particularly, the high dimension (Section "Challenges Underpinning mGWAS") limit the application of some host GWAS tools. Nonetheless, host GWAS (Zhou and Stephens, 2012) and (Lippert et al., 2011) have been applied in mGWAS using taxonomic abundance as bacterial phenotype; albeit, it cannot be used for association testing with microbiome distance metrics.

To this end, microbiome-specific tool and method have recently been developed: microbiomeGWAS (Hua et al., 2015), and microbiome-association index (Rothschild et al., 2018). microbiomeGWAS uses standard linear regression with beta diversity metrics and corrects for skewness and kurtosis. It identifies host genetic variants associated with microbiome beta diversity by testing both SNP-microbiome and SNP-environment interactions. Because the statistical power of these distance-based measures depends on the choice of the distance metric, this tool was subsequently improved to accommodate multiple distance matrices. Meanwhile, microbiome-association index (b^2) has been specifically developed to quantify the overall association of microbiome to host's phenotype, incorporating the contribution of host genetics. Using this association index, a measure similar to narrow sense heritability in hGWAS, the authors showed that several host phenotypes, including body mass index, fasting glucose levels, glycaemic status, and lactose consumption, exhibited substantial b^2 values in the range of 22–36%. In addition, different statistical methods have been

TABLE 1 | Summary list of microbiome genome-wide association studies (mGWAS) carried out to date.

Study	Year	Sequencing method	Analysis software	Sample size and location	Microbiome sampling site	Microbiome phenotype studied	N° of associations identified	Comment
Blekhman et al., 2015	2015	Shotgun metagenomic	PLINK	n = 93 United States of America (HMP)	Multiple sites (15)	Alpha diversity, beta diversity, and bacterial taxa	83 associations identified	Host genetic variants correlated with microbiome composition. Variants in the <i>LC7</i> gene correlated with abundance of <i>Bifidobacterium</i> ($P = 1.16 \times 10^{-5}$). Genes involved in Leptin signalling pathway significantly associated with microbiome composition; Leptin previously implicated in Obesity.
Davenport et al., 2015	2015	16S rRNA	GEMMA	n = 127 Hutterites (North America)	Gut	Bacterial taxa	≥8 bacterial taxa associated with SNPs in host genome in each season	SNPs in regions of the <i>PLD1</i> gene associated with abundance of genus <i>Akkermansia</i> ; the <i>PLD1</i> gene was previously implicated in GWAS of body mass index. one bacterial taxa (genus <i>Bifidobacterium</i>) correlated with age. ≥4 bacterial taxa differentially abundant by sex
Hua et al., 2015	2015	16S rRNA	microbiome GWAS	N = 147 Italy	Lung	alpha diversity, beta diversity	Six SNPs had suggestive association with beta-diversity	Analysis performed using both weighted and unweighted UniFrac distance matrices. No SNPs were significantly associated after correcting for skewness and kurtosis of beta-diversity distributions.
Goodrich et al., 2016	2016	16S rRNA	microbiome GWAS (for GWAS on the beta diversity measures)	n = 1,126 twin pairs United Kingdom	Gut	Bacterial taxa, beta diversity	31 associated host loci	<i>LC7</i> gene associated with <i>Bifidobacterium</i> . Also, SNPs in the region of the <i>R3HDM1</i> gene strongly associated with <i>Bifidobacterium</i> ($P = 4.38 \times 10^{-6}$). <i>ALDH1L1</i> gene associated with SHA-98 bacteria. SNPs associated with beta diversity metrics. alpha diversity found to be heritable. Heritable taxa found, including <i>Bifidobacterium</i> , <i>Turicibacter</i> , and <i>Blautia</i> .
Bonder et al., 2016	2016	Shotgun metagenomic	"base" in R	n = 1,514 Dutch	Gut	Bacterial taxa, bacterial pathways	42 associated host loci	Nine host loci associated with bacterial taxa, and 33 loci with bacterial pathways ($P < 5 \times 10^{-6}$). <i>LC7</i> SNP associated with <i>Bifidobacterium</i> ($P = 3.45 \times 10^{-6}$).
Turpin et al., 2016	2016	16S rRNA	Not sated	n = 1,098 (discovery cohort)- Canada and United States n = 463 (replication cohort)- Canada, United States and Israel	Gut	Bacterial taxa and alpha diversity	58 suggestive associations only six of which were significantly associated	Of these six SNPs, one was replicated in the replication cohort. Associated taxa included <i>Rikenellaceae</i> , <i>Faecalibacterium</i> , <i>Lachnospira</i> , and <i>Eubacterium</i> . Many taxa found to be heritable. No significant association with alpha diversity.
Wang et al., 2016	2016	16S rRNA	"envfit" in R	n = 1,812 Germany	Gut	Bacterial taxa, beta diversity	54 significant associations	42 loci (which included variants in <i>VDR</i> gene -encoding vitamin D receptor) associated with beta diversity ($P < 5 \times 10^{-8}$). Host loci associated with the <i>Firmicutes</i> , <i>Proteobacteria</i> , and <i>Bacteroidetes</i> .
Igartua et al., 2017	2017	16S rRNA	GEMMA	n = 144 Hutterites (North America)	Vestibule and Nasopharynx sites	Relative abundance (RA) of bacterial taxa, alpha, and beta-diversity	37 significant associations	Most significant association was between variant (rs117042385) upstream of the <i>TNCR</i> gene ($p = 1.61 \times 10^{-8}$; $q = 0.002$). Differences in RA, alpha and beta diversity were observed across sites and seasons. Significant association between host genetic (measured by kinship coefficient) and microbiome Euclidean distance.

(Continued)

TABLE 1 | Continued

Study	Year	Sequencing method	Analysis software	Sample size and location	Microbiome sampling site	Microbiome phenotype studied	N° of associations identified	Comment
Rühlmann et al., 2018	2017	16S rRNA	"snpsStats" in R	n (=1,767 Germany	Gut	beta diversity	4four significant associations	The four loci were significantly associated with variation in beta diversity. Reanalysis using permutation-based analysis were still identified all these loci as genome-wide significant.
Kolde et al., 2018	2018	Shotgun metagenomic	Matrix eQTL	n (=298 North America (HMP)	Multiple sites (6sx)	Relative abundance of bacterial taxa, bacterial pathway	five significant associations with bacterial taxa, and 82 with bacterial pathways	In stool, five species (<i>Lachnospiraceae bacterium</i> , <i>Roseburia intestinalis</i> , <i>Subdoligranulum</i> (unclassified), <i>Sutterella wadsworthensis</i>) out of 118 were significantly associated (false discovery rate (FDR)? < 70.05 by permutation test). In other body sites, <i>Propionibacterium propionicum</i> , <i>Porphyromonas catoniae</i> , and unclassified <i>Gemella</i> were significantly associated with host genomic variation in buccal mucosa.
Rothschild et al., 2018	2018	16S rRNA and Shotgun metagenomic	"envfit" and "ordiR2step" in R and FaST-LMM	n = 1,046 Israel	Gut	Relative abundance of bacterial taxa, alpha, and beta-diversity	seven suggestive associations	No significant association between host genetic variation and bacterial taxa or beta diversity, after correcting for multiple testing.

used in mGWAS such as the ordination and permutation-based envfit, ordi2step, snpStats, and Spearman's correlation statistical methods (Table 1).

DISPARITY IN HOST AND MICROBIOME GWAS

Despite the meteoric rise in GWASs over the last few years, the number of studies inclusive of genetically diverse populations is disproportionately low. A 2015 assessment of the number of NIH-funded GWAS focused on or utilizing non-European populations revealed great disparity; for example, of the 4,942 publications, African American, Hispanic, and Jewish ancestry constituted only ~3%, <1%, and <1%, respectively, of the sampling study population (Peprah et al., 2015). A recent analysis of a curated database of genomic variants associated with various traits/diseases, provided by the National Human Genome Research Institute (NHGRI) and the European Bioinformatics Institute (EMBL-EBI), revealed a bias in genomic diversity and disproportionate representation (in terms of ancestry, and physical and social environments of the study subjects) in published GWAS (Hindorff et al., 2018). As of August 2016, non-European ancestry represented only 19% of all individuals in GWAS. This becomes a pertinent issue given the observation that non-European individuals contribute a larger number of genotype-phenotype associations (Morales et al., 2018; Hindorff et al., 2018), and studies in other (non-European) population groups continue to identify novel genetic variants. This disparity in population representation is not confined to host GWAS. Although a relatively young field, mGWAS carried out since its inception in 2015, show marked disparity in terms of genetic diversity and population representation. As depicted in Figure 2, most studies involved individuals from North America and Europe. With mGWAS findings differing markedly across all these studies, better inclusivity of diverse populations will illuminate any underlying interactions since it is possible that the interaction of host genetic variant(s) with the microbiome may be population-, environment-, or even individual-specific owing to other yet-unknown clinical and environmental factors.

From a statistical genetics perspective, while the inclusion of non-European populations, in particular the African population, is a critical step toward discovery of important host-genetic interactions in traits/diseases, the implementation requires methodological and technological refinement. It is well known that variants associated with diseases found in populations of European descent do not always replicate in non-European, particularly African populations (Peprah et al., 2015). This discrepancy across populations are due to several possible reasons including differences in allelic architecture, LD, and environmental factors across populations (Popejoy and Fullerton, 2016; Bentley et al., 2017). Thus, there is a need to design appropriate novel statistical models that are tailored to leverage the characteristics of non-European subjects. Moreover, most of the current technologies for mining genetic data, for example genotyping arrays, have been designed for populations

of European descent with long-range patterns of LD (Peprah et al., 2015; Popejoy and Fullerton, 2016) or nearly homogeneous environments.

CHALLENGES UNDERPINNING mGWAS

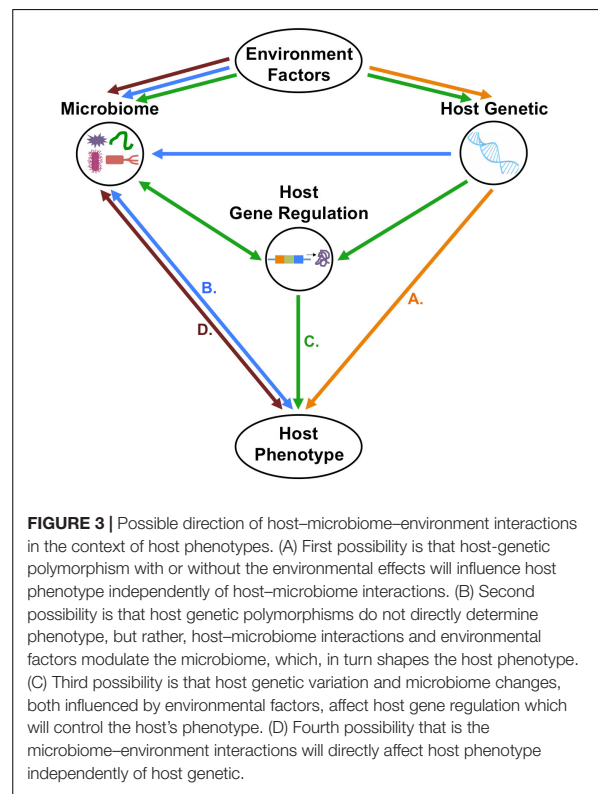
Microbiome genome-wide association studies findings have provided unprecedented views into the association of human host genes with microbes or microbial genes. However, there are several key challenges which also present new opportunities that need to be tackled if we are to assemble a global understanding of host-genetic association with the microbiome.

Demographic and Environmental Factors

The human microbiome is sensitive to a wide range of demographic and environmental factors. Factors as diverse as gender (Foster et al., 2017), age (Mäkituokko et al., 2010), and geography (Yatsunenko et al., 2012) have all been shown to influence the composition and functional diversity of the microbiome. These factors can introduce sampling artifacts or biases in mGWAS which can reduce statistical power. Therefore, the statistical models need to adjust for, the effect of these factors. This is particularly important if such factors do not have interaction effect with the genetic variants. Accounting for these factors remains a fundamental challenge for mGWAS. Although it is nearly impractical to adjust for all these factors in a typical setting, it is imperative that enumerable factors be considered as covariates in downstream analyses. Adjustment for potential endogenous and exogenous sources of variability will be key aspects for providing reliable and replicable results.

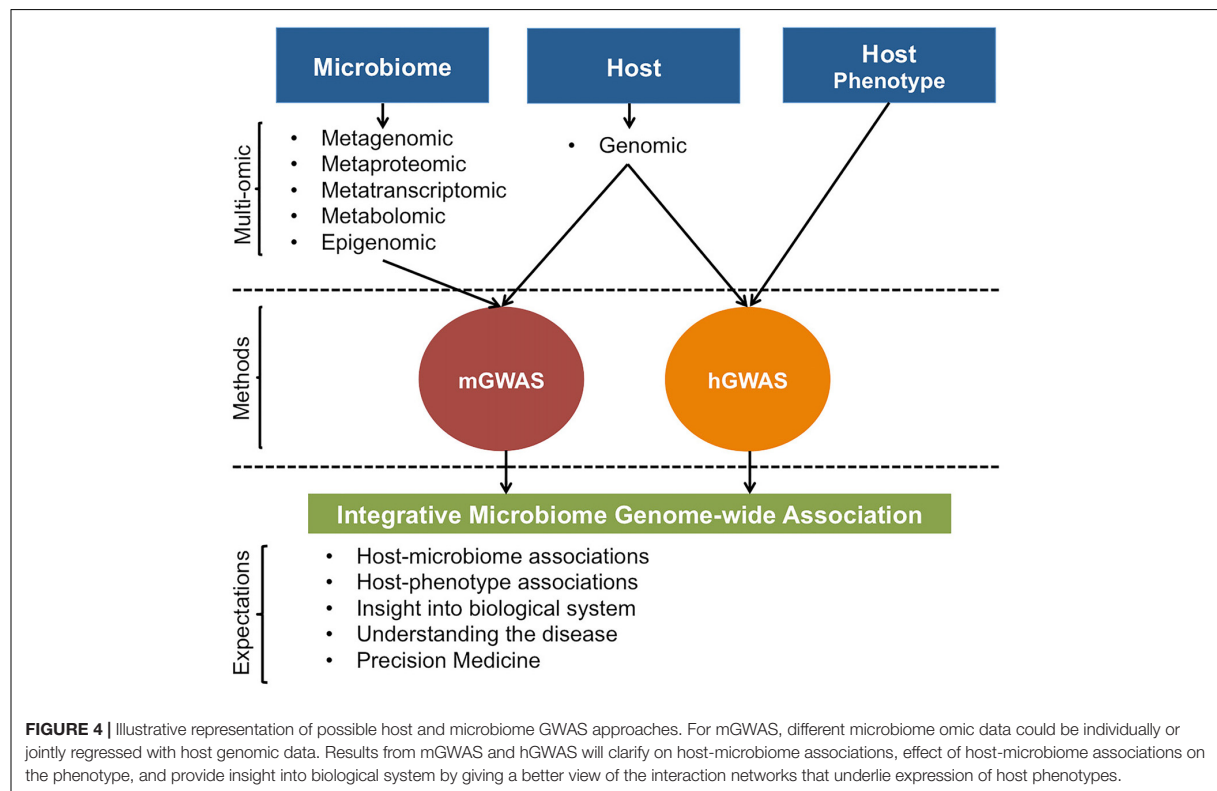
The Complexity of Microbiome Data

Pertinent to mGWAS, the complexity of microbiome data in terms of dimension, phenotype and correlation structure presents a challenge in the development of robust association frameworks. Microbiome data is highly dimensional, often consisting of hundreds of bacterial taxa. When searching for genetic variants associated with a bacterial trait, multiple tests are carried out, each time testing the null hypothesis of no difference in genotype distribution. With many taxa, this leads to not only high computational cost, but increases in the number of statistical tests. This requires correction for multiple testing to control for the occurrence of false positives. The correction (commonly genome-wide significance, permutation tests, FDR or Bonferroni correction), however, introduces yet another challenge – a potential reduction in statistical power – especially when an underpowered adjustment procedure or an inappropriate error rate is used. The inherent strengths and limitations of each of these correction methods influence association results. For example, (Blekhman et al., 2015) and (Kolde et al., 2018), using FDR and Bonferroni corrections, respectively, obtained contrasting results using the same research cohort. The current solution to circumvent the issue of multiple testing is to focus on only a subset of taxa or variants. For example, (Davenport et al., 2015) reduced the number of bacterial taxa by removing all taxa highly correlated with taxa at the same or lower taxonomic



level. Meanwhile, in another study by Bonder et al. (2016), the authors performed a targeted association analysis focusing only on SNPs in genes related to immunity and metabolism. While robust in detecting true associations, using only a selection of taxa leads to incomplete representation of microbiome composition and, consequently, limits the opportunity to discover novel associations. This issue is critical as there may be specific rare taxa that can interact with the host genetics. Likewise, although powerful at dissecting association at a biologically plausible region of interest, reducing the number of host genes variants examined can result in exclusion of relevant genes from the analysis.

In addition, the genetic architecture, the landscape of contributions of host genetics to a given microbiome phenotype, is at best poorly understood although studies of host–genetic interactions with microbiome suggests polygenicity. Moreover, the reported percentage of microbiome variation explained by the associated alleles are generally very small (for example 0.65–0.97% in Wang et al., 2016), and thus unable to explain much of the variability in microbiome phenotype. Given this effect size, large sample sizes will be required to detect modestly associated variants. Meanwhile, the high level of trait collinearity coupled with complex correlation structure (Kurilshikov et al., 2017) also make it challenging for statistical methods. Even though parametric linear models remain the cornerstone for genetic association studies and have played pivotal role in mGWAS



carried out to date, they are limited to detecting non-linear interaction patterns. In particular, when modeling complex structures such as varying effect and non-linear interactions, the exponential rise in the number of parameters increases computational cost and reduces statistical power (Moore et al., 2010). Moreover, linear models generally treat interaction effects as factors with independent marginal effects; a strategy that lowers its power in the presence of interaction effects (Millstein et al., 2006). Given these limitations, there is a need to develop mGWAS-adapted statistical methods to complement existing linear models.

Replicability of Results

Besides the challenges toward achieving reliable results, replication of mGWAS results has been poor. The first mGWAS, (Blekhman et al., 2015), using 93 individuals and bacterial taxa as phenotypes, detected 83 associations between genetic polymorphisms in host coding genes and the abundance of specific bacterial taxa. Of note, was the association of immune-related genes, *HLA-DRA* and *TLR1*, with abundance of *Selenomonas* and *Lautropia*, respectively, and the strong link between SNPs in the lactase persistence gene, *LCT*, and abundance of *Bifidobacterium*. In subsequent studies, for example, (Davenport et al., 2015; Bonder et al., 2016; Goodrich et al., 2016), many immune and metabolism-related host genes were found to be significantly associated with abundance of

bacterial taxa and beta diversity measures. However, there was little congruence between the results across these studies. This can possibly be attributed to factors including differences in statistical methods, multiple-testing corrections, lifestyle, diet, demographic and environmental conditions of the samples. Nonetheless, the enrichment of microbiome-associated variants with immunity and metabolism related genes, and the generally small effect size, the percentage of microbiome variation explained by the genes, (<1%) remain the most consistent. This suggests that a large proportion of heritability for most bacterial traits may be accounted for by many small effect genetic polymorphisms in immunity and metabolism-encoding genes. A corollary is that bacterial traits likely have an infinitesimal genetic architecture, requiring meta-analyses to detect the associated variants. A combination of larger sample sizes, unified robust analysis methods, and inter-cohort analyses facilitated by collaborations such as MiBioGen consortium (Wang et al., 2018) are crucial for both the attainment of power to detect small to moderate host genetic effects on microbiome traits and replicability of findings.

CONCLUSION AND PERSPECTIVES

Discovery of the role of the microbiome in normalcy and disease status spurred efforts to elucidate the interaction of host genetic variation with the microbiome, leading to the

development of mGWAS. Initial foray into mGWAS has led to the discovery of host genetic variants that contribute to variability in compositional and functionality diversity of the microbiome. Despite the significant strides made in this field, further developments are still needed to elucidate the extent, direction, and mechanism of host-microbiome association and how this association ultimately impact on host's phenotypic expression (Figure 3).

Moving forward, the importance of sufficient sample sizes cannot be overstated. This will require a collaborative approach, pooling samples from across different geographic regions of the world to generate sufficiently powered studies for discovery and replication. In doing so, the trade-off between sample size and between-sample heterogeneity must be carefully assessed, given the myriad of factors that can reduce the association power of mGWAS. Otherwise, a wide between-sample differences resulting from temporal and spatial heterogeneities will reduce the power to detect true association. In addition to sample size, maximizing bacterial trait information will potentially increase detection power. Current mGWAS has focused on independent analysis of various bacterial traits. This is probably due to the current lack of known software applications that can enable a joint analysis of microbial taxa/pathway and microbial diversity. Given the relative etiological similarity of these traits, it is likely that such joint multiple phenotype analysis will maximize discovery power.

In addition, even though mGWAS to date have primarily focused on the genomic level, regressing host's genetic variation with microbiome's transcriptomic, metabolomic, and proteomic data types or integrating them in a joint host genotype-microbiome association analysis (Figure 4) will be an exciting venture. mGWAS using these multi-level data types can potentially yield insights into whether host-microbiome interactions are, if any, universal or more pronounced to a specific microbial attribute; facilitating identification of particular host genetic polymorphisms that interacts with the microbiome on a molecular level. This is crucial if mGWAS results are to have any utility at level of understanding host's clinical and biological states.

REFERENCES

- 1000 Genomes Project Consortium, Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., et al. (2015). A global reference for human genetic variation. *Nature* 526, 68–74. doi: 10.1038/nature15393
- Allali, I., Boukhatem, N., Bouguenouch, L., Hardi, H., Boudouaya, H. A., Cadenas, M. B., et al. (2018). Gut microbiome of moroccan colorectal cancer patients. *Med. Microbiol. Immunol.* 207, 211–215. doi: 10.1007/s00430-018-0542-5
- Belkaid, Y., and Hand, T. W. (2014). Role of the microbiota in immunity and inflammation. *Cell* 157, 121–141. doi: 10.1016/j.cell.2014.03.011
- Bentley, A. R., Callier, S., and Rotimi, C. N. (2017). Diversity and inclusion in genomic research: why the uneven progress? *J. Commun. Genet.* 8, 255–266. doi: 10.1007/s12687-017-0316-6
- Blauwendraat, C., Reed, X., Kia, D. A., Gan-Or, Z., Lesage, S., Pihlström, L., et al. (2018). Frequency of loss of function variants in lrrk2 in parkinson disease. *JAMA Neurol.* 75, 1416–1422. doi: 10.1001/jamaneurol.2018.1885
- Blekhan, R., Goodrich, J. K., Huang, K., Sun, Q., Bukowski, R., Bell, J. T., et al. (2015). Host genetic variation impacts microbiome composition across human body sites. *Genome Biol.* 16, 1–12. doi: 10.1186/s13059-015-0759-1

Furthermore, the integration of environmental factors in mGWAS will lead to an exciting starting point and perspective for a more comprehensive and robust analysis of host-microbiome interaction in more powered studies. However, the current challenges include difficulty of adjusting for potential environmental factors, complexity of microbiome data, and lack of robust and unified analytical frameworks to handle the diverse and peculiar properties of microbial attributes as quantitative traits. Together, addressing these challenges coupled with increased sample size, independent replication, and meta-analysis in multiple populations will provide a more complete understanding of human variation and microbial diversity in connection with health and disease.

AUTHOR CONTRIBUTIONS

EC, DA, and IA conceived and structured the manuscript. DA, IA, SD, SH, KM, NT, AG, NM, and EC generated the contents and wrote the manuscript.

FUNDING

The authors would like to thank DAAD, the German Academic Exchange Programme, for the financial support under Reference number 91653117. Some of the authors are funded in part by the National Institutes of Health Common Fund under grant number U41HG006941, the H3ABioNet Project grant number U24HG006941, SAdACC grant number 1U01HG007459-01, and Wellcome Trust/AESA grant H3A/18/001.

ACKNOWLEDGMENTS

We are grateful to Oyekanm Nashiru and Hassan Ghazal for helpful discussions.

- Blum, H. E. (2017). The human microbiome. *Adv. Med. Sci.* 62, 414–420. doi: 10.1016/j.advms.2017.04.005
- Bonder, M. J., Kurilshikov, A., Tigchelaar, E. F., Muijagic, Z., Imhann, F., Vila, A. V., et al. (2016). The effect of host genetics on the gut microbiome. *Nature genetics* 48, 1407. doi: 10.1038/ng.3663
- Casals, F., and Bertranpetit, J. (2012). Human genetic variation, shared and private. *Science* 337, 39–40. doi: 10.1126/science.1224528
- Chen, H.-S., Hutter, C. M., Mechanic, L. E., Amos, C. I., Bafna, V., Hauser, E. R., et al. (2015). Genetic simulation tools for post-genome wide association studies of complex diseases. *Genet. Epidemiol.* 39, 11–19. doi: 10.1002/gepi.21870
- Chiang, C., Scott, A. J., Davis, J. R., Tsang, E. K., Li, X., Kim, Y., et al. (2017). The impact of structural variation on human gene expression. *Nat. Genet.* 49, 692. doi: 10.1038/ng.3834
- Chimusa, E. R., Mbiyavanga, M., Mazandu, G. K., and Mulder, N. J. (2015). ancgwas: a post genome-wide association study method for interaction, pathway and ancestry analysis in homogeneous and admixed populations. *Bioinformatics* 32, 549–556. doi: 10.1093/bioinformatics/btv619

- Conlon, M. A., and Bird, A. R. (2014). The impact of diet and lifestyle on gut microbiota and human health. *Nutrients* 7, 17–44. doi: 10.3390/nu7010017
- Davenport, E. R., Cusanovich, D. A., Michelini, K., Barreiro, L. B., Ober, C., and Gilad, Y. (2015). Genome-wide association studies of the human gut microbiota. *PLoS ONE* 10:e0140301. doi: 10.1371/journal.pone.0140301
- Foster, J. A., Rinaman, L., and Cryan, J. F. (2017). S*stress & the gut-brain axis: Regulation by the microbiome. *Neurobiology of stress*
- Gilbert, J. A., Blaser, M. J., Caporaso, J. G., Jansson, J. K., Lynch, S. V., and Knight, R. (2018). Current understanding of the human microbiome. *Nat. Med.* 24, 392. doi: 10.1038/nm.4517
- Gilbert, J. A., Meyer, F., Antonopoulos, D., Balaji, P., Brown, C. T., Brown, C. T., et al. (2010). Meeting report: the terabase metagenomics workshop and the vision of an earth microbiome project. *Standards in genomic sciences* 3, 243. doi: 10.4056/signs.1433550
- Goodrich, J. K., Davenport, E. R., Beaumont, M., Jackson, M. A., Knight, R., Ober, C., et al. (2016). Genetic determinants of the gut microbiome in uk twins. *Cell host & microbe* 19, 731–743. doi: 10.1016/j.chom.2016.04.017
- Goodrich, J. K., Davenport, E. R., Clark, A. G., and Ley, R. E. (2017). The relationship between the human genome and microbiome comes into view. *Annu. Rev. Genet.* 51, 413–433. doi: 10.1146/annurev-genet-110711-155532
- Goodrich, J. K., Waters, J. L., Poole, A. C., Sutter, J. L., Koren, O., Blekhman, R., et al. (2014). Human genetics shape the gut microbiome. *Cell* 159, 789–799. doi: 10.1016/j.cell.2014.09.053
- Grarup, N., Moltke, I., Andersen, M. K., Dalby, M., Vitting-Seerup, K., Kern, T., et al. (2018). Loss-of-function variants in *adcy3* increase risk of obesity and type 2 diabetes. *Nat. Genet.* 50, 172. doi: 10.1038/s41588-017-0022-7
- Gupta, V. K., Paul, S., and Dutta, C. (2017). Geography, ethnicity or subsistence-specific variations in human microbiome composition and diversity. *Frontiers in microbiology* 8:1162. doi: 10.3389/fmicb.2017.01162
- Hall, A. B., Tolonen, A. C., and Xavier, R. J. (2017). Human genetic variation and the gut microbiome in disease. *Nat. Rev. Genet.* 18, 690. doi: 10.1038/nrg.2017.63
- Hindorf, L. A., Bonham, V. L., Brody, L. C., Ginoza, M. E. C., Hutter, C. M., Manolio, T. A., et al. (2018). Prioritizing diversity in human genomics research. *Nat. Rev. Genet.* 19, 175–185. doi: 10.1038/nrg.2017.89
- Hou, L., and Zhao, H. (2013). A review of post-gwas prioritization approaches. *Frontiers in genetics* 4:280. doi: 10.3389/fgene.2013.00280
- Hua, X., Goedert, J. J., Landi, M. T., and Shi, J. (2016). Identifying host genetic variants associated with microbiome composition by testing multiple beta diversity matrices. *Hum. Hered.* 81, 117–126. doi: 10.1159/000448733
- Hua, X., Song, L., Yu, G., Goedert, J. J., Abnet, C. C., Landi, M. T., et al. (2015). Micro*biomegwas: a tool for identifying host genetic variants associated with microbiome composition. *Biorxiv* doi: 10.1159/000448733
- Igartua, C., Davenport, E. R., Gilad, Y., Nicolae, D. L., Pinto, J., and Ober, C. (2017). Host genetic variation in mucosal immunity pathways influences the upper airway microbiome. *Microbiome* 5, 16. doi: 10.1186/s40168-016-0227-5
- Imhann, F., Vila, A. V., Bonder, M. J., Fu, J., Gevers, D., Visschedijk, M. C., et al. (2018). Interplay of host genetics and gut microbiota underlying the onset and clinical presentation of inflammatory bowel disease. *Gut* 67, 108–119. doi: 10.1136/gutjnl-2016-312135
- Jiang, H., Ling, Z., Zhang, Y., Mao, H., Ma, Z., Yin, Y., et al. (2015). Altered fecal microbiota composition in patients with major depressive disorder. *Brain Behav. Immun.* 48, 186–194. doi: 10.1016/j.bbi.2015.03.016
- Kolde, R., Franzosa, E. A., Rahnavard, G., Hall, A. B., Vlamakis, H., Stevens, C., et al. (2018). Host genetic variation and its microbiome interactions within the human microbiome project. *Genome medicine* 10, 6. doi: 10.1186/s13073-018-0515-8
- Kostic, A. D., Xavier, R. J., and Gevers, D. (2014). The microbiome in inflammatory bowel disease: current status and the future ahead. *Gastroenterology* 146, 1489–1499. doi: 10.1053/j.gastro.2014.02.009
- Kurilshnikov, A., Wijmenga, C., Fu, J., and Zhernakova, A. (2017). Host genetics and gut microbiome: challenges and perspectives. *Trends in immunology* 38, 633–647. doi: 10.1016/j.it.2017.06.003
- Langdon, A., Crook, N., and Dantas, G. (2016). The effects of antibiotics on the microbiome throughout development and alternative approaches for therapeutic modulation. *Genome medicine* 8, 39. doi: 10.1186/s13073-016-0294-z
- Lange, K., Buerger, M., Stallmach, A., and Bruns, T. (2016). Effects of antibiotics on gut microbiota. *Dig. Dis.* 34, 260–268. doi: 10.1159/000443360
- Lax, S., Smith, D. P., Hampton-Marcell, J., Owens, S. M., Handley, K. M., Scott, N. M., et al. (2014). Longitudinal analysis of microbial interaction between humans and the indoor environment. *Science* 345, 1048–1052. doi: 10.1126/science.1254529
- Lin, J.-R., Cai, Y., Zhang, Q., Zhang, W., Nogales, R., and Zhang, Z. (2016). In*tegrated post-gwas analysis shed new light on the disease mechanisms of schizophrenia. *Genetics, genetics*
- Lippert, C., Listgarten, J., Liu, Y., Kadie, C. M., Davidson, R. I., and Heckerman, D. (2011). FaST linear mixed models for genome-wide association studies. *Nat. Methods* 8, 833–835. doi: 10.1038/nmeth.1681
- Lloyd-Price, J., Abu-Ali, G., and Huttenhower, C. (2016). The healthy human microbiome. *Genome medicine* 8, 51. doi: 10.1186/s13073-016-0307-y
- Lozupone, C. A., Stombaugh, J. I., Gordon, J. I., Jansson, J. K., and Knight, R. (2012). Diversity, stability and resilience of the human gut microbiota. *Nature* 489, 220. doi: 10.1038/nature11550
- MacArthur, D. G., and Tyler-Smith, C. (2010). Loss-of-function variants in the genomes of healthy humans. *Hum. Mol. Genet.* 19, R125–R130. doi: 10.1093/hmg/ddq365
- Mäkituokko, H., Tiihonen, K., Tynkkynen, S., Paulin, L., and Rautonen, N. (2010). The effect of age and non-steroidal anti-inflammatory drugs on human intestinal microbiota composition. *British journal of nutrition* 103, 227–234. doi: 10.1017/S0007114509991553
- Marchesi, J. R., Adams, D. H., Fava, F., Hermes, G. D., Hirschfield, G. M., Hold, G., et al. (2016). The gut microbiota and host health: a new clinical frontier. *Gut* 65, 330–339. doi: 10.1136/gutjnl-2015-309990
- Marian, A. J. (2012). Elements of “missing heritability”. *Curr. Opin. Cardiol* 27, 197–201. doi: 10.1097/HCO.0b013e328352707d
- Marth, G. T., Yu, F., Indap, A. R., Garimella, K., Gravel, S., Leong, W. F., et al. (2011). The functional spectrum of low-frequency coding variation. *Genome Biol.* 12, R84. doi: 10.1186/gb-2011-12-9-r84
- Millstein, J., Conti, D. V., Gilliland, F. D., and Gauderman, W. J. (2006). A testing framework for identifying susceptibility genes in the presence of epistasis. *The American Journal of Human Genetics* 78, 15–27. doi: 10.1086/498850
- Moore, J. H., Asselbergs, F. W., and Williams, S. M. (2010). Bioinformatics challenges for genome wide association studies. *Bioinformatics* 26, 445–455. doi: 10.1093/bioinformatics/btp713
- Morales, J., Welter, D., Bowler, E. H., Cerezo, M., Harris, L. W., McMahon, A. C., et al. (2018). A standardized framework for representation of ancestry data in genomics studies, with application to the nhgri-ebi gwas catalog. *Genome Biol.* 19, 21. doi: 10.1186/s13059-018-1396-2
- Nelson, M. R., Wegmann, D., Ehm, M. G., Kessner, D., Jean, P. S., Verzilli, C., et al. (2012). A*ⁿ abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. *Science* doi: 10.1126/science.1217876
- Peprah, E., Xu, H., Tekola-Ayele, F., and Royal, C. D. (2015). Genome-wide association studies in Africans and African americans: expanding the framework of the genomics of human traits and disease. *Public Health Genomics* 18, 40–51. doi: 10.1159/000367962
- Popejoy, A. B., and Fullerton, S. M. (2016). Genomics is failing on diversity. *Nature* 538, 161. doi: 10.1038/538161a
- Qiao, Y., Wu, M., Feng, Y., Zhou, Z., Chen, L., and Chen, F. (2018). Alterations of oral microbiota distinguish children with autism spectrum disorders from healthy controls. *Scientific reports* 8, 1597. doi: 10.1038/s41598-018-19982-y
- Rajilä-Stojanović, M., Heilig, H. G., Molenaar, D., Kajander, K., Surakka, A., Smidt, H., et al. (2009). Development and application of the human intestinal tract chip, a phylogenetic microarray: analysis of universally conserved phylotypes in the abundant microbiota of young and elderly adults. *Environ. Microbiol.* 11, 1736–1751. doi: 10.1111/j.1462-2920.2009.01900.x
- Rawls, J. F., Mahowald, M. A., Ley, R. E., and Gordon, J. I. (2006). Reciprocal gut microbiota transplants from zebrafish and mice to germ-free recipients reveal host habitat selection. *Cell* 127, 423–433. doi: 10.1016/j.cell.2006.08.043
- Reijnders, M. R., Miller, K. A., Alvi, M., Goos, J. A., Lees, M. M., De Burca, A., et al. (2018). De novo and inherited loss-of-function variants in *tlk2*: Clinical and genotype-phenotype evaluation of a distinct neurodevelopmental disorder.

- The American Journal of Human Genetics* 102, 1195–1203. doi: 10.1016/j.ajhg.2018.04.014
- Relling, M. V., and Evans, W. E. (2015). P^{*}harmacogenomics in the clinic. *Nature* 526. doi: 10.1038/nature15817
- Rothschild, D., Weissbrod, O., Barkan, E., Kurilshikov, A., Korem, T., Zeevi, D., et al. (2018). Environment dominates over host genetics in shaping human gut microbiota. *Nature* 555, 210. doi: 10.1038/nature25973
- Rühlemann, M. C., Degenhardt, F., Thingholm, L. B., Wang, J., Skieceviäinen, J., Rausch, P., et al. (2018). Application of the distance-based F test in an mgwas investigating beta diversity of intestinal microbiota identifies variants in *slc9a8* (*nhe8*) and 3 other loci. *Gut microbes* 9, 68–75. doi: 10.1080/19490976.2017.1356979
- Sandoval-Motta, S., Aldana, M., and Frank, A. (2017a). Evolving ecosystems: Inheritance and selection in the light of the microbiome. *Arch. Med. Res.* 48, 780–789. doi: 10.1016/j.arcmed.2018.01.002
- Sandoval-Motta, S., Aldana, M., Martínez-Romero, E., and Frank, A. (2017b). The human microbiome and the missing heritability problem. *Frontiers in genetics* 8:80. doi: 10.3389/fgene.2017.00080
- Sawicki, M. P., Samara, G., Hurwitz, M., and Passaro, E. (1993). Human genome project. *The American journal of surgery* 165, 258–264. doi: 10.1016/S0002-9610(05)80522-7
- Schirmer, M., Smeekens, S. P., Vlamakis, H., Jaeger, M., Oosting, M., Franzosa, E. A., et al. (2016). Linking the human gut microbiome to inflammatory cytokine production capacity. *Cell* 166, 1125–1136. doi: 10.1016/j.cell.2016.10.020
- Singh, R. K., Chang, H.-W., Yan, D., Lee, K. M., Ucmak, D., Wong, K., et al. (2017). Influence of diet on the gut microbiome and implications for human health. *Journal of translational medicine* 552, 73. doi: 10.1186/s12967-017-1175-y
- Srinivas, G., Möller, S., Wang, J., Künzel, S., Zillikens, D., Baines, J. F., et al. (2013). Genome wide mapping of gene-microbiota interactions in susceptibility to autoimmune skin blistering. *Nature communications* 4, 2462. doi: 10.1038/ncomms3462
- Takiishi, T., Fenero, C. I. M., and Cámara, N. O. S. (2017). Intestinal barrier and gut microbiota: shaping our immune responses throughout life. *Tissue Barriers* 5, e1373208. doi: 10.1080/21688370.2017.1373208
- Thorisson, G. A., Smith, A. V., Krishnan, L., and Stein, L. D. (2005). The international hapmap project web site. *Genome Res.* 15, 1592–1593. doi: 10.1101/gr.4413105
- Tremaroli, V., and Bäckhed, F. (2012). Functional interactions between the gut microbiota and host metabolism. *Nature* 489, 242. doi: 10.1038/nature11552
- Turpin, W., Espin-Garcia, O., Xu, W., Silverberg, M. S., Kevans, D., Smith, M. I., et al. (2016). Association of host genome with intestinal microbial composition in a large healthy cohort. *Nat. Genet.* 48, 1413. doi: 10.1038/ng.3693
- Upadhyaya, S., and Banerjee, G. (2015). Type 2 diabetes and gut microbiome: at the intersection of known and unknown. *Gut Microbes* 6, 85–92. doi: 10.1080/19490976.2015.1024918
- van Tongeren, S. P., Slaets, J. P., Harmsen, H., and Welling, G. W. (2005). Fecal microbiota composition and frailty. *Appl. Environ. Microbiol.* 71, 6438–6442. doi: 10.1128/AEM.71.10.6438-6442.2005
- Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A., et al. (2017). 10 years of gwas discovery: biology, function, and translation. *The American Journal of Human Genetics* 101, 5–22. doi: 10.1016/j.ajhg.2017.06.005
- Wang, J., Kurilshikov, A., Radjabzadeh, D., Turpin, W., Croitoru, K., Bonder, M. J., et al. (2018). Meta-analysis of human genome-microbiome association studies: the microbiome consortium initiative. *Microbiome* 6, 101. doi: 10.1186/s40168-018-0479-3
- Wang, J., Thingholm, L. B., Skieceviäinen, J., Rausch, P., Kummén, M., Hov, J. R., et al. (2016). Genome-wide association analysis identifies variation in vitamin d receptor and other host factors influencing the gut microbiota. *Nat. Genet.* 48, 1396. doi: 10.1038/ng.3695
- Wang, K., Li, M., and Hakonarson, H. (2010). Analysing biological pathways in genome-wide association studies. *Nat. Rev. Genet.* 11, 843. doi: 10.1038/nrg2884
- Weissbrod, O., Rothschild, D., Barkan, E., and Segal, E. (2018). Host genetics and microbiome associations through the lens of genome wide association studies. *Curr. Opin. Microbiol.* 44, 9–19. doi: 10.1016/j.mib.2018.05.003
- Wilson, M. V., and Shmida, A. (1984). Measuring beta diversity with presence-absence data. *The Journal of Ecology* 1055–1064. doi: 10.2307/2259551
- Wu, G. D., Chen, J., Hoffmann, C., Bittinger, K., Chen, Y.-Y., Keilbaugh, S. A., et al. (2011). Linking long-term dietary patterns with gut microbial enterotypes. *Science* 334, 105–108. doi: 10.1126/science.1208344
- Xie, H., Guo, R., Zhong, H., Feng, Q., Lan, Z., Qin, B., et al. (2016). Shotgun metagenomics of 250 adult twins reveals genetic and environmental impacts on the gut microbiome. *Cell systems* 3, 572–584. doi: 10.1016/j.cels.2016.10.004
- Yang, J., Fritsche, L. G., Zhou, X., Abecasis, G., and International Age-Related Macular Degeneration Genomics Consortium. (2017). A scalable bayesian method for integrating functional information in genome-wide association studies. *The American Journal of Human Genetics* 101, 404–416. doi: 10.1016/j.ajhg.2017.08.002
- Yatsunenko, T., Rey, F. E., Manary, M. J., Trehan, I., Dominguez-Bello, M. G., Contreras, M., et al. (2012). Human gut microbiome viewed across age and geography. *Nature* 486, 222. doi: 10.1038/nature11053
- Yu, G., Gail, M. H., Consonni, D., Carugno, M., Humphrys, M., Pesatori, A. C., et al. (2016). Characterizing human lung tissue microbiota and its relationship to epidemiological and clinical features. *Genome Biol.* 17, 163. doi: 10.1186/s13059-016-1021-1
- Zaitlen, N., and Kraft, P. (2012). Heritability in the genome-wide association era. *Hum. Genet.* 131, 1655–1664. doi: 10.1007/s00439-012-1199-6
- Zhou, X., and Stephens, M. (2012). Genome-wide efficient mixed-model analysis for association studies. *Nat. Genet.* 44, 821. doi: 10.1038/ng.2310

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2019 Awany, Allali, Dalvie, Hemmings, Mwaikono, Thomford, Gomez, Mulder and Chimusa. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

3.5 Dissecting GWAS for microbiome-associated metabolic diseases

As the preceding section details, the composition and functional diversity of the microbiome is shaped by its interaction with host genetic factors and the environment. This interaction potentially modulates the crosstalk between the gut and peripheral organs such as the liver, pancreas, muscle, adipose tissue, and brain, and thereby contributing to reshaping the balance between normalcy and disease states (Sanz et al., 2015). Intriguingly, many of these traits/diseases that are linked to the microbiome are also the same ones with established strong genetic component, as determined by the over-a-decade long host GWAS. Of note are metabolic diseases - diseases known to have several genetic predisposing factors - where several observational and experimental studies corroborate their intimate link with shifts in the composition and functional diversity of the microbiome. We therefore dissect genome-wide studies for these diseases to shed on the overlap and relative contributions of the constituent predisposing factors. Then, because host genetics, like the microbiome, is highly dependent on, among other factors, ethnicity (Hawkins et al., 2019; Li et al., 2019; Brooks et al., 2018; Deschasaux et al., 2018), we examine genetic diversity in variants reported to be associated with metabolic diseases. These are presented in the article provided in the sequel.

Dissecting Genome-wide Studies for Microbiome-related Metabolic Diseases

Denis Awany¹, Imane Allali², Emile R. Chimusa^{1,3*}

¹Division of Human Genetics, Department of Pathology, Faculty of Health Sciences, University of Cape Town, Cape Town, South Africa. ²Laboratory of Human Pathologies Biology, Department of Biology, Faculty of Sciences, and Genomic Center of Human Pathologies, Faculty of Medicine and Pharmacy, Mohammed V University in Rabat, Morocco. ³Institute of Infectious Disease and Molecular Medicine, Faculty of Health Sciences, University of Cape Town, Cape Town, South Africa.

1

¹ Denis Away, PhD Candidate at Division of Human Genetics, Dept. of Pathology, University of Cape Town.

Imane Allali, PhD in Bioinformatics. Assis. Prof. of Bioinformatics at the Department of Biology, Faculty of Sciences, Mohammed V University in Rabat, Morocco.

Emile R. Chimusa, PhD in Bioinformatic. Assoc/Prof. of medical population genetics and Genetics epidemiology at the Division of Human Genetics, Department of Pathology, University of Cape Town.

Abstract

Despite the meteoric rise in genome-wide association studies for metabolic diseases (MetD) over the last few years, our understanding of the pathogenesis of these diseases is still far from complete. Recent developments have established that MetD arises from complex interactions between host genetics, the gut microbiome, and the environment. However, our knowledge of the genetic and microbiome components involved, and the underlying molecular mechanisms remains limited. Here, we review and summarize recent studies investigating the genetic and microbiome basis of MetD. Then, given the critical importance of study-individual's ancestry in these studies, we leverage 4,932 whole genome sequence (WGS) samples from 18 world-wide ethnic groups to examine genetic diversity in currently reported variants associated with metabolic diseases. The analyses show marked differences in gene-specific proportion of pathogenic SNPs, and gene-specific SNPs minor allele frequencies across ethnic groups, highlighting the importance of population- and ethnic-specific investigations in pinpointing the causative factors for MetD. We conclude with a discussion of research areas where further investigation on interactions between host genetics, microbiome, and the environment is needed.

*Corresponding author: Associate Prof. Emile R Chimusa; e-mail: emile.chimusa@uct.ac.za

Keywords: Metabolic diseases, microbiome, heritability, risk prediction, Genome-Wide Association Studies

Introduction

Metabolic diseases (MetD) is the combination of several metabolic disturbances that encompasses abdominal obesity, dyslipidemia, high blood pressure, insulin resistance and/or plasma glucose intolerance [14]. These diseases are a precursor to cardiovascular disease and type 2 diabetes mellitus [14, 52]. MetD results from complex interactions between the host's DNA and non-DNA factors. Traditionally, host's non-DNA factors, including environmental factors such as diets, lifestyle, age, gender, ethnicity and socioeconomic status, have been known to be important risk factors for metabolic defects. However, it has become clear that these factors are only a partial component of the causative factors for MetD. During the past

few years, prompted by observation of familial aggregation of disease cases and modulation of endophenotypes belonging to the lipid, glucose or vascular tone systems [11], there has been heightened interest in the genetic determinants of MetD. Accordingly, employing the genome-wide association study (GWAS) methodology, several genetic variants that with modifying effects at population and individual levels have been unveiled [20].

Besides the environmental and host genetic components, recent studies have also implicated bacteria and bacterial products of the human gut microbiota with metabolic defects. The gut microbiota not only processes nutrients into absorbable compounds, but also produces essential vitamins and metabolites [1]. Consequently, gut microbiome dysbiosis disrupts the normal regulation of host biochemical pathways and metabolic networks [20], which has important implications in the propensity for the development of metabolic dysfunction. On the other hand, the environmental factors listed above have been shown to modulate the relative composition and functional diversity of the human gut microbiome, whereas, established by the relatively new microbiome genome-wide association study (mGWAS) approach, variations in several host genes, such as nucleotide binding oligomerization domain 2 and fucosyltransferase 2, can influence the architecture of the gut microbiome [34]. Therefore, it is apparent that both host genetics and the gut microbiome are important non-mutually exclusive contributors to MetD. The prevalence of MetD in different populations is highly dependent on the genetic ancestry of the individuals. Ancestry-related difference in susceptibility to MetD springs from not just population differences in factors such as lifestyle and environmental exposure, which, as mentioned above, are important risk factors, but also, crucially, from differences in genetic architecture across different populations. Genetic variants in the low frequency (here used loosely to mean those with minor allele frequency, MAF, less than 5%) spectrum, which are mostly of recent origin, tend to be sporadic in terms of their geographic and ethnic representation. In contrast, many high frequency variants (MAF at least 5%), which are old, tend to have a global representation [3]; albeit variants robustly detected in one population may not easily be detectable in other populations owing to ethnic-differences in allele frequency. For example, in type 2 diabetes, a MetD, the risk SNP rs790314, for example, at the *TCF7L2* locus is more frequent in the European population (30%) and less frequent in the East Asian population (2%) [19, 3]. This ethnic-disparity in genetic predisposition to MetD points to the possible existence of ancestry-specific drivers and unique mechanisms in disease and draws

attention to importance of ethnic-specific and trans-ethnic investigations of the MetD risk factors.

Here, we explore the dissection of human genetic polymorphisms in predisposition to MetD and its intricate link with the microbiome, focusing on the role and mechanism of host genetics, heritability, and microbiome contributions to the general MetD. Furthermore, in light of variability in allelic architecture across ethnic groups, we leverage 4,932 whole genome sequence (WGS) samples from 19 world-wide ethnic groups to examine genetic diversity and differentiation at currently reported variants associated with metabolic diseases. We conclude with a discussion of research areas where further investigation on interactions between host genetics, microbiome, and the environment is needed.

Genetics of Metabolic Diseases

As stated before, the development of MetD has a genetic component. Estimates of genetic heritability for MetD phenotype have ranged from 10% to 70%, indicating that these phenotypes are partly heritable [4, 10, 5, 42]. Knowledge of the genetic factors responsible for the onset, progression and inter-individual variation in MetD states is of direct importance in mainstream healthcare. Using each component of MetD at a time, several extensive efforts employing the genome-wide study approach have uncovered many disease-modifying genetic polymorphisms. The progress in identifying common genetic predisposition to MetD has been relatively rapid, driven primarily by advances in genotyping array technology and international collaborations. Together, these have led to the discovery of many disease-relevant loci and brought to light important insights into the genetic architecture of the different MetD components. Association mapping for MetD has been performed using either each component included in the definition of MetD at a time or to MetD as an entity. In the latter case, the MetD status is treated as a presence/absence dichotomous trait based on one of the available case-control definitions, most commonly using the harmonized NCEP criteria for MetD [27] and in which case the association with overall MetD status is typically due to one or two strongly-associated MetD components [31].

Large scale GWAS conducted during the past 2 years have increased from the previous 29 to 93, the number of independent loci associated with MetD (binary) at genome-wide significance

threshold ($P < 5 \times 10^{-8}$) [35]. The majority of these loci overlap with those detected by GWAS of individual MetD components. This pleiotropic feature suggests that a joint analysis of MetD components may be valuable in the search for susceptibility loci. That said, another particularly interesting finding from these studies is that the vast majority of associated loci are for lipid-related trait components of MetD [35]. To date, the list of implicated loci continues to grow for each MetD (**Figure 1**), validating previous discoveries and/ or uncovering novel associations. At the level of specific MetD components, concrete evidence has been provided for the role of variants, particularly for SNPs in or near the genes encoding lipoprotein lipase, apolipoprotein A (APOA)-cluster, and cholesteryl ester transfer protein. In type 2 diabetes, there are currently over 250 associated genetic risk loci, with ~400 independent signals of association [24]; an increase from about 243 loci in 2018 [37].

In BMI/obesity, a recent large-scale analysis of ~700,000 individuals of European ancestry identified 941 near-independent SNPs, which includes 751 SNPs located in previously undetected genomic loci [54]. Cumulatively, these near-independent SNPs explained ~6% of the variability of BMI in independent samples; an increase when compared to ~3% explained by 97 near-independent SNPs in a previously largest GWAS with ~250,000 subjects [36]. In concomitance with BMI, a crude measure of overall adiposity, genetic mapping has also been carried out for specific aspects of body fat content, such as trunk fat, and hip and waist circumference. The most recent, and largest GWAS meta-analysis of body fat distribution to date, Pulit et.al (2019), using a sample of 694,649 individuals, identified 346 loci (of which 300 were novel) containing 463 independent signals associated with waist-hip ratio, after adjustment for BMI [43]. Also interesting from this fully powered study, is that the top associated loci indicated patterns of sex-dimorphism, with heritability and variant effects being generally stronger in women than in men. The component-wise delineation of MeTD-genetic relationship can be of epidemiological relevance since particular components may be more informative than the overall, with respect to predicting disease risk. Moreover, it can inform on the relative importance of the components on overall trait variation.

In connection with blood pressure, up to 535 independent loci have now been identified [12]. This includes association with various blood pressure traits: systolic blood pressure, diastolic blood pressure, and pulse pressure. Importantly, is the strong biological importance of the frequently replicated associated genes; for example, the genes *PDE3A* (which encodes a cyclic AMP (cAMP)-regulating phosphodiesterase involved in activation in cells [25], *RXFP2*

(receptor for hormone *relaxin* which is involved in many molecular activities, including vasodilation [17] and *ADK* (a key enzyme in regulation of physiological processes in cells and tissues [29]. Previous studies have also identified genetic variants involved in dyslipidemia, the disease condition relating to abnormal amounts of lipids in the blood [9]. Over 95 genetic loci have been associated with lipid traits in aggregate. The most consistently replicated SNPs are generally those in or near the *APOE*, *PCSK9*, *C679X*, *LPL*, *APOA5*, and *CETP* genes.

Of note from the above genetic associations, while the underlying biological mechanisms remain generally unclear, as do the true functional variants, many of the detected genes are enriched for genes in or near genes involved in the brain and hypothalamus activities. For example, the *TMEM18*, *MC4R*, and *NPC1* genes that have been replicated in GWAS of obesity and related metabolic traits are also highly expressed in the hypothalamus [55]. In the brain, particularly in the hypothalamus where it is predominantly expressed, the *MC4R* gene plays a pivotal role in regulating energy intake [50]. This observation, coupled with the observed overlap of some genetic variants associated with different MetD components, suggests that genetic regulation of obesity and related MetD phenotypes may also be mediated through neuronal systems.

While the prevailing independent genetic mapping of individual MetD components have had gratifying success, genetic association with MetD combined as a single phenotypic entity may yet lead to discovery of new SNPs that would be undetectable when studying each phenotype separately. This premise is based on the important observation of genetic correlation among the different MetD components. The notion that a joint phenotypic mapping may increase discoveries is nicely illustrated by a previous GWAS on Crohn's disease and Celiac disease, where jointly analysing the two diseases for shared loci led to the detection of new additional risk loci [15]. Taking all together, results from GWAS Catalogue, previous literature up to date, and gene-diseases database such as DisGeNET, there are approximately 458 genes known to associate with MetD components, harbouring a total of 318,797 SNPs in current dbSNPs database (**Supplementary Data**). These SNPs show difference in ethnic-specific minor allele frequencies at bin of 1% to 2% (**Figure 2**) across 18 ethnic groups, suggesting MetD components shaped ethnic genomics variation at common variants. This, therefore, provides unprecedented opportunities to map MetD genes across ancestrally divergent continental ethnic groups.

Microbiome and Metabolic Diseases

One of the main advances of the past decade has been the improvement in our understanding of metabolic diseases (MetD), including obesity, type 2 diabetes, high blood pressure, dyslipidemia, and plasma glucose intolerance. Up to these past few years, it was intuitive to imagine that these pathologies are all strongly linked to a nutritional imbalance; however, it is only recently, with the advent of next generation sequencing (NGS), that associations between MetD and the modifications of the microbiota composition have been established [28, 1]. These associations are still not well understood because of the complexity and variability of the microbiome composition. The gut microbiota plays an active role in regulating the gene expression of the host and is a major regulator of immunity [6]. Many studies have shown the impact of the host's genes, lifestyle, diet, and antibiotics consumption on the microbiota which affects the health of the host by modifying his metabolism regulation and immune system which leads to many diseases, such as chronic diseases, metabolic diseases, autoimmune diseases, and increased exposure to infectious diseases [1,39].

The human gut microbiota is dominated by two groups of bacteria: Firmicutes and Bacteroidetes. These two phyla represent almost 90% of all intestinal bacteria. The Firmicutes/Bacteroidetes ratio has been suggested as a marker of some metabolic diseases mainly in obesity [8]. Many studies comparing obese and lean subjects have shown an increase in Firmicutes and a decrease in Bacteroidetes in individuals with obesity [32, 2]. The microbiome plays a crucial role in the processing of nutrients. Foods not directly metabolized by the host are processed in the gut. The production of food in the gut is linked to the presence of more Firmicutes and lower Bacteroidetes. The Firmicutes have a greater capacity to digest complex carbohydrates, extract more calories and allow for a greater energy absorption. When obese subjects lost weight over a year, the proportion of Firmicutes approached that of thin subjects which suggests that the intestinal microbiota are more efficient in obese people in extracting energy from food than those of lean people [26]. Indeed, it has been found in the intestine of a majority of highly obese patients a reduced number of bacterial diversities [4]. Subjects with obesity showed a different gut microbiota profile at the genera level with an increase of *Coprococcus*, *Parvimonas*, and *Fusobacterium*, and a decrease in *Akkermansia*, *Faecalibacterium*, *Bifidobacterium*, *Oscillospira*, and *Lactobacillus* [51, 7, 30].

Moreover, the gut microbiome has been associated with type 2 diabetes (T2D) [47]. Many studies have reported associations between specific taxa and T2D [18, 56, 44]. It has been found, at the genera level, that *Bifidobacterium*, *Bacteroides*, *Faecalibacterium*, *Akkermansia*, and *Roseburia* were decreased while *Ruminococcus*, *Fusobacterium*, and *Blautia* were increased in patients with T2D [23]. However, the Firmicutes/Bacteroidetes ratio did not show a significant correlation with T2D. Among the negatively correlated microbial community, *Bifidobacterium* is the most reported genus in human studies that investigated T2D with microbiota composition. It is known that *Bifidobacterium* is an important probiotic necessary to the intestine homeostasis, plays a protective role against T2D and the development of obesity, and it has anti-inflammatory properties as well as *Faecalibacterium prausnitzii* species [16, 40]. Moreover, *Akkermansia muciniphila* seems to have a beneficial effect in the context of metabolic disorders. This bacterium has been associated with energy and carbohydrate metabolism, it was demonstrated that the administration of *Akkermansia muciniphila* to obese and diabetic animals decreased the gain in body weight, improved the glycemia, and reinforced the intestinal barrier which led to decrease the metabolic inflammation [13, 41, 48, 53].

In the case of high blood pressure, few studies have investigated its association with microbiome composition compared to obesity and T2D. Nevertheless, specific taxa were reported to be overrepresented in patients with high blood pressure such as *Prevotella*, *Klebsiella*, *Porphyromonas*, and *Actinomyces*. However, *Faecalibacterium*, *Oscillibacter*, *Roseburia*, *Bifidobacterium*, *Coprococcus*, and *Butyrivibrio* were underrepresented in patients with high blood pressure [33]. Additionally, studies have investigated the microbiome in dyslipidemia. The genera, *Bacteroides*, *Alistipes*, *Coprococcus*, *Bilophila*, and *Butyricimonas* were positively associated with dyslipidemia and lipid metabolism while *Blautia*, *Collinsella*, and *Holdemania* were negatively associated [45, 46, 57]. **Figure 3** summarizes these potential associations for each MetD component.

Dissecting ancestry differentiation of loci implicated in MetD

As with other complex human traits, the same old mantra of lack of genetic diversity in the major proportion of GWAS carried out to date also applies to metabolic diseases. It is well-

known that ancestry-specific differences in genetic variation, for example, allele frequencies and linkage disequilibrium (LD) structures have important implications in the identification and replication of disease association across different populations [49]. Besides the direct impact of ancestry-induced genetic differences on disease mapping, the realization that the microbiome, which can be thought of as a proxy for various environmental factors and which has ethnic-specific patterns, has profound influence on, especially, MetD makes inclusivity of diverse ethnic populations in MetD studies a scientific imperative.

To dissect the genomics and pathogenic distribution, across ethnic groups, of the common genetic variants reported in association studies of MetD, a total of 458 MetD-associated genes were obtained from GWAS catalog <https://www.ebi.ac.uk/gwas/>, literature and gene-disease databases such DisGeNET <http://www.disgenet.org/>. A total of 318,797 SNPs were extracted out of the genes using dbSNPs database (**Supplementary Data**), to perform analyses across 18 world-wide ethnic groups: Afro-Asiatic Semiti, African-American, African-Caribbean, Afro-Asiatic Cushitic, Afro-Asiatic Omotic, KhoeSan, Niger-Congo Bantu South, Niger-Congo Volta Niger, Niger Congo West, Latin American, European center, European North, European South, European USA, East Asian, South Asian, UK Indian, and USA Indian (**Supplementary Table 1**). The full description of quality check and down-stream processing are provided in the **Supplementary Data**. Enrichment analysis (**Supplementary Data**) of these 458 genes confirms that these associated genes are enriched in metabolic syndrome (pvalue = 3.760e-09), cholesterol metabolism (pvalue = 1.346e-08) pathways and lipid and cholesterolemia-related (**Supplementary Figure 1**) trait components of MetD [35]. We observed that majority of associated MetD genes (**Supplementary Figure 2**) show ethnic difference in gene-specific in SNPs minor allele frequencies. This provides unprecedented opportunities to map MetD genes across ancestrally divergent continental ethnic groups by leveraging ethnic-specific local microbiomes in model of association. These observations, juxtaposed with the variabilities in SNPs minor allele frequencies between 1% to 2% (**Figure 2**) across ethnic groups, shed some light on the architecture of MetD, lending support to the overarching view that context- and population-specific analytical studies is crucial to detect putative disease causative factors.

From the analyses, displayed in **Supplementary Figures 3A-C**, there are several noteworthy findings. First, for the majority of genes, the proportion of pathogenic SNPs in each associated gene differ markedly across ethnic groups. Second, while there is disparity in gene-specific patterns, individuals of the African Caribbean and Latin American ancestry appear to generally harbour greater numbers of pathogenic variants. Third, the distribution of pathogenic variants across ethnic groups is similar for some genes, particularly those with few pathogenic variants.

Concluding Remarks

As genome-wide studies of metabolic diseases continue to qualitatively and quantitatively dissect their genetic basis, fundamental insights into the genetic complexity, the role of the environment, particularly the microbiome, and the potential mechanism of action continue to emerge. Various reports of MetD genome-wide association studies mostly of European-ancestry encompassing larger samples, have been performed and, these have enabled the prediction of the genetic liability of MetD. However, the clinical utility of predictive association and heritability of MetD is still in its infancy stage and its application in genetics testing and counselling in real-world clinical populations is limited.

Here, we have highlighted the genetics of MetD and its different components and showed the set of genes currently implicated in MetD; demonstrating the substantial overlap between the genes associated with various MetD components, which suggests an intricate mechanistic link between them. We have shown bacterial taxa that are associated with increased or lowered risk of MetD. Furthermore, we demonstrated the impact of ethnicity on genetic mapping for MetD, highlighting variability in gene-content proportion of pathogenic SNPs and gene-specific in SNPs minor allele frequencies across different world-wide ethnic groups. Consequently, genetic studies with different population groups and ethnicities are needed to not only improve the detection power for new genes and modest effects but also provide more reliable effect sizes estimates for the discovered variants.

Given the multifactorial nature of MetD, systematically leveraging different information on phenotypic plasticity, especially the local ethnic-specific host-microbiome owing to its profound link with MetD, will be pivotal in the pursuit of unravelling the basis of metabolic diseases. Moving forward, we anticipate that the genetic community, capitalizing on advances

in genotyping and sequencing technologies, will unlock, in the not-so-distant future, the treasure trove of the basis of metabolic diseases.

Moreover, considering the invaluable potential of the microbiome, and changes in individual or community lifestyle, it is crucial that current association approaches such GWAS leverage the deterministic and stochastic factors that have known contributory role to MetD variability. To this end, new risk prediction and association approaches that leverage population-specific characteristic including epigenetics, genetics ancestry, host-genetics interaction with microbiomes may improve the predictive power of the heritability and risk prediction of MetD and contribute in advancing clinical utility of MetD GWAS in global clinical populations.

Key Points

- Majority of current Metabolic disease-associated genes show ethnic differences in gene-specific SNPs minor allele frequencies and proportion of pathogenic variants.
- Metabolic diseases arise from complex interactions between host genetics, the gut microbiome, and the environment.
- Discovery of further disease-relevant loci from non-European ethnic groups will bring more insights into the genetic architecture of the different metabolic diseases, and the relative strength of each disease-predisposing factor.

Funding

The authors are supported in part by the German Academic Exchange Service (DAAD) [funding #91653117], and National Research Foundation of South Africa for funding (NRF) [grant #RA171111285157/119056]. This research was funded (partially) by the South African government through the National Research Foundation. Responsibility for the information and views expressed in the article lies entirely with the authors.

Acknowledgements

We are grateful to CHPC (<https://www.chpc.ac.za/>) for providing computing facility.

Conflict of Interest

The authors declare that they have no competing interests.

Figure Legends

Figure 1: List of genes associated with metabolic disease components; A: Obesity, body mass index, body fat, and waist circumference; B: diabetes, blood pressure, and hypertension; C: tryglyceride level, High-density lipoprotein (HLD) cholesterol, Glycated haemoglobin (HbA1c) level, and glucose level.

Figure 2: The distribution of the minor allele frequency categorised into 6 bins (0-0.05, >0.05-0.1, >0.1-0.2, >0.2-0.3, >0.3-0.4, >0.4-0.5) with respect to genotypes data of each ethnic group based on 318,797 SNPs from 458 genes known to associate to MetD.

Figure 3: Reported associations of bacterial taxa with metabolic diseases. (+) is increased in abundance, while (-) is decreased in abundance in metabolic disease state.

Supplementary Figures

Supplementary Figure 1: A) Human phenotype sub-network of 458 genes known to associate with MetD, and B) top associate pathways, biological processes and molecular functions.

Supplementary Figure 2: Gene-specific SNPs minor allele frequencies across 18 ethnic groups (Supplementary Data).

Supplementary Figure 3A: Gene-specific proportion of pathogenic SNPs across 18 ethnic groups (Supplementary Data).

Supplementary Figure 3B: Gene-specific proportion of pathogenic SNPs across 18 ethnic groups (Supplementary Data).

Supplementary Figure 3C: Gene-specific proportion of pathogenic SNPs across 18 ethnic groups (Supplementary Data).

Supplementary Table

Supplementary Table 1: Data obtained from 1000 Genomes Project (1KGP) (Consortium et al.,2012) and the African Genome Variation Project (AGVP) (Gurdasani et al.,2015) and used for analysis.

References

- [1] Safaa ALTVES, Hatice Kübra YILDIZ, and Hasibe CINGILLI VURAL. Interaction of the microbiota with the human body in health and diseases. *Bioscience of Microbiota, Food and Health*, pages 19–023, 2019.
- [2] Emmanouil Angelakis, Fabrice Armougom, Matthieu Million, and Didier Raoult. The relationship between gut microbiota and weight gain in humans. *Future microbiology*, 7(1):91–109, 2012.
- [3] Inês Barroso and Mark I McCarthy. The genetic basis of metabolic disease. *Cell*, 177(1):146–161, 2019.
- [4] Jonathan P Bradfield, Suzanne Vogelezang, Janine F Felix, Alessandra Chesi, Øyvind Helgeland, Momoko Horikoshi, Ville Karhunen, Estelle Lowry, Diana L Cousminer, Tarunveer S Ahluwalia, et al. A trans-ancestral meta-analysis of genome-wide association studies reveals loci associated with childhood obesity. *Human molecular genetics*, 28(19):3327–3338, 2019.
- [5] Gemma Cadby, Phillip E Melton, Nina S McCarthy, Corey Giles, Natalie A Mellett, Kevin Huynh, Joseph Hung, John Beilby, Marie-Pierre Dubé, Gerald F Watts, et al. Heritability of 596 lipid species and genetic correlation with cardiovascular traits in the busselton family heart study. *Journal of Lipid Research*, 61(4):537–545, 2020.
- [6] Tunahan Çakr, Gianni Panagiotou, Reaz Uddin, and Saliha Durmus. Novel approaches for systems biology of metabolism-oriented pathogen-human interactions: A mini-review. *Frontiers in Cellular and Infection Microbiology*, 10:52, 2020.

- [7] Olga Castaner, Albert Goday, Yong-Moon Park, Seung-Hwan Lee, Faidon Magkos, Sue-Anne Toh Ee Shiow, and Helmut Schröder. The gut microbiome profile in obesity: a systematic review. *International journal of endocrinology*, 2018, 2018.
- [8] Louise Crovesy, Daniele Masterson, and Eliane Lopes Rosado. Profile of the gut microbiota of adults with obesity: A systematic review. *European Journal of Clinical Nutrition*, pages 1–12, 2020.
- [9] Kruttika Dabke, Gustaf Hendrick, and Suzanne Devkota. The gut microbiome and metabolic syndrome. *Journal of Clinical Investigation*, 129(10):4050–4057, 2019.
- [10] Christine Dalgård, Søren Möller, and Kirsten O Kyvik. Heritability of curve patterns in oral glucose tolerance test. *Twin Research and Human Genetics*, pages 1–6, 2020.
- [11] Maria Benedetta Donati and Licia Iacoviello. Coronary heart disease, genetics, nutrition and physical activity. In *Nutrition and Fitness: Obesity, the Metabolic Syndrome, Cardiovascular Disease, and Cancer*, volume 94, pages 96–106. Karger Publishers, 2005.
- [12] Evangelos Evangelou, Helen R Warren, David Mosen-Ansorena, Borbala Mifsud, Raha Pazoki, He Gao, Georgios Ntritsos, Niki Dimou, Claudia P Cabrera, Ibrahim Karaman, et al. Genetic analysis of over 1 million people identifies 535 new loci associated with blood pressure traits. *Nature genetics*, 50(10):1412–1425, 2018.
- [13] Amandine Everard, Clara Belzer, Lucie Geurts, Janneke P Ouwerkerk, Céline Druart, Laure B Bindels, Yves Guiot, Muriel Derrien, Giulio G Muccioli, Nathalie M Delzenne, et al. Cross-talk between *akkermansia muciniphila* and intestinal epithelium controls diet-induced obesity. *Proceedings of the National Academy of Sciences*, 110(22):9066–9071, 2013.
- [14] Tove Fall and Erik Ingelsson. Genome-wide association studies of obesity and metabolic syndrome. *Molecular and cellular endocrinology*, 382(1):740–757, 2014.
- [15] Eleonora AM Festen, Philippe Goyette, Todd Green, Gabrielle Boucher, Claudine Beauchamp, Gosia Trynka, Patrick C Dubois, Caroline Lagace, Pieter CF Stokkers, Daan W Hommes, et al. A meta-analysis of genome-wide association scans identifies *il18rap*, *ptpn2*, *tagap*, and *pus10* as shared risk loci for crohn’s disease and celiac disease. *PLoS genetics*, 7(1), 2011.

- [16] Jean-Pierre Furet, Ling-Chun Kong, Julien Tap, Christine Poitou, Arnaud Basdevant, Jean-Luc Bouillot, Denis Mariat, Gérard Corthier, Joël Doré, Corneliu Henegar, et al. Differential adaptation of human gut microbiota to bariatric surgery–induced weight loss: links with metabolic and low-grade inflammation markers. *Diabetes*, 59(12):3049–3057, 2010.
- [17] Ayush Giri, Jacklyn N Hellwege, Jacob M Keaton, Jihwan Park, Chengxiang Qiu, Helen R Warren, Eric S Torstenson, Csaba P Kovesdy, Yan V Sun, Otis D Wilson, et al. Trans-ethnic association study of blood pressure determinants in over 750,000 individuals. *Nature genetics*, 51(1):51–62, 2019.
- [18] Aline Corado Gomes, Allain Amador Bueno, Rávila Graziany Machado de Souza, and João Felipe Mota. Gut microbiota, probiotics and diabetes. *Nutrition journal*, 13(1):60, 2014.
- [19] Struan FA Grant, Gudmar Thorleifsson, Inga Reynisdottir, Rafn Benediktsson, Andrei Manolescu, Jesus Sainz, Agnar Helgason, Hreinn Stefansson, Valur Emilsson, Anna Helgadóttir, et al. Variant of transcription factor 7-like 2 (tcf7l2) gene confers risk of type 2 diabetes. *Nature genetics*, 38(3):320–323, 2006.
- [20] Niels Grarup, Camilla H Sandholt, Torben Hansen, and Oluf Pedersen. Genetic susceptibility to type 2 diabetes and obesity: from genome-wide association studies to rare variants and beyond. *Diabetologia*, 57(8):1528–1541, 2014.
- [21] William B Gudykunst and Karen L Schmidt. Language and ethnic identity: An overview and prologue. *Journal of Language and Social Psychology*, 6(3-4):157–170, 1987.
- [22] Deepti Gurdasani, Tommy Carstensen, Fasil Tekola-Ayele, Luca Pagani, Ioanna Tachmazidou, Konstantinos Hatzikotoulas, Savita Karthikeyan, Louise Iles, Martin O Pollard, Ananyo Choudhury, et al. The african genome variation project shapes medical genetics in africa. *Nature*, 517(7534):327–332, 2015.
- [23] Manoj Gurung, Zhipeng Li, Hannah You, Richard Rodrigues, Donald B Jump, Andrey Morgun, and Natalia Shulzhenko. Role of gut microbiota in type 2 diabetes pathophysiology. *EBioMedicine*, 51:102590, 2020.
- [24] Prashantha Hebbar, Jehad Ahmed Abubaker, Mohamed Abu-Farha, Jaakko Tuomilehto, Fahd Al-Mulla, and Thangavel Alphonse Thanaraj. A perception on genome-wide genetic analysis of metabolic traits in arab populations. *Frontiers in endocrinology*, 10:8, 2019.

- [25] Miles Houslay. Hypertension linked to pde3a activation. *Nature genetics*, 47(6):562–563, 2015.
- [26] Zhekai Hu, Yu Zhang, Zhiyu Li, Yuedi Yu, Wenyan Kang, Yingnan Han, Xiwen Geng, Shaohua Ge, and Yundong Sun. Effect of helicobacter pylori infection on chronic periodontitis by the change of microecology and inflammation. *Oncotarget*, 7(41):66700, 2016.
- [27] Paul L Huang. A comprehensive definition for metabolic syndrome. *Disease models & mechanisms*, 2(5-6):231–237, 2009.
- [28] Kyu Yeon Hur and Myung-Shik Lee. Gut microbiota and metabolic disorders. *Diabetes & metabolism journal*, 39(3):198–203, 2015.
- [29] Katharina Kiese, Janos Jablonski, Detlev Boison, and Katja Kobow. Dynamic regulation of the adenosine kinase gene during early postnatal brain development and maturation. *Frontiers in molecular neuroscience*, 9:99, 2016.
- [30] Bohkyung Kim, Ha-Neul Choi, and Jung-Eun Yim. Effect of diet on the gut microbiota associated with obesity. *Journal of Obesity & Metabolic Syndrome*, 28(4):216, 2019.
- [31] Kati Kristiansson, Markus Perola, Emmi Tikkanen, Johannes Kettunen, Ida Surakka, Aki S Havulinna, Alena Stancáková, Chris Barnes, Elisabeth Widen, Eero Kajantie, et al. Genome-wide screen for metabolic syndrome susceptibility loci reveals strong lipid gene contribution but no evidence for common genetic basis for clustering of metabolic syndrome traits. *Circulation: Cardiovascular Genetics*, 5(2):242–249, 2012.
- [32] Ruth E Ley, Peter J Turnbaugh, Samuel Klein, and Jeffrey I Gordon. Human gut microbes associated with obesity. *nature*, 444(7122):1022–1023, 2006.
- [33] Jing Li, Fangqing Zhao, Yidan Wang, Junru Chen, Jie Tao, Gang Tian, Shouling Wu, Wenbin Liu, Qinghua Cui, Bin Geng, et al. Gut microbiota dysbiosis contributes to the development of hypertension. *Microbiome*, 5(1):14, 2017.
- [34] Mi Young Lim, Hyun Ju You, Hyo Shin Yoon, Bomi Kwon, Jae Yoon Lee, Sunghee Lee, Yun-Mi Song, Kayoung Lee, Joohon Sung, and GwangPyo Ko. The effect of heritability and host genetics on the gut microbiota and metabolic syndrome. *Gut*, 66(6):1031–1038, 2017.

- [35] Lars Lind. Genome-wide association study of the metabolic syndrome in uk biobank. *Metabolic syndrome and related disorders*, 17(10):505–511, 2019.
- [36] Adam E Locke, Bratati Kahali, Sonja I Berndt, Anne E Justice, Tune H Pers, Felix R Day, Corey Powell, Sailaja Vedantam, Martin L Buchkovich, Jian Yang, et al. Genetic studies of body mass index yield new insights for obesity biology. *Nature*, 518(7538):197–206, 2015.
- [37] Anubha Mahajan, Daniel Taliun, Matthias Thurner, Neil R Robertson, Jason M Torres, N William Rayner, Anthony J Payne, Valgerdur Steinthorsdottir, Robert A Scott, Niels Grarup, et al. Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. *Nature genetics*, 50(11):1505–1513, 2018.
- [38] Stelios Michalopoulos. The origins of ethnolinguistic diversity. *American Economic Review*, 102(4):1508–39, 2012.
- [39] Ivana Milosevic, Ankica Vujovic, Aleksandra Barac, Marina Djelic, Milos Korac, Aleksandra Radovanovic Spurnic, Ivana Gmizic, Olja Stevanovic, Vladimir Djordjevic, Nebojsa Lekic, et al. Gut-liver axis, gut microbiota, and its modulation in the management of liver diseases: a review of the literature. *International journal of molecular sciences*, 20(2):395, 2019.
- [40] David O’Mahony, Sharon Murphy, Thomas Boileau, JeanSoon Park, Frances O’Brien, David Groeger, Patrycja Konieczna, Mario Ziegler, Paul Scully, Fergus Shanahan, et al. *Bifidobacterium animalis* ahc7 protects against pathogen-induced nf-kb activation in vivo. *BMC immunology*, 11(1):63, 2010.
- [41] Elin Org, Brian W Parks, Jong Wha J Joo, Benjamin Emert, William Schwartzman, Eun Yong Kang, Margarete Mehrabian, Calvin Pan, Rob Knight, Robert Gunsalus, et al. Genetic and environmental control of host-gut microbiota interactions. *Genome research*, 25(10):1558–1569, 2015.
- [42] CM1 Povel, JMA Boer, E Reiling, and EJM Feskens. Genetic variants and the metabolic syndrome: a systematic review. *Obesity Reviews*, 12(11):952–967, 2011.
- [43] Sara L Pulit, Charli Stoneman, Andrew P Morris, Andrew R Wood, Craig A Glastonbury, Jessica Tyrrell, Loc Yengo, Teresa Ferreira, Eirini Marouli, Yingjie Ji, et al.

Meta-analysis of genome-wide association studies for body fat distribution in 694 649 individuals of european ancestry. *Human molecular genetics*, 28(1):166–174, 2019.

[44] Junjie Qin, Yingrui Li, Zhiming Cai, Shenghui Li, Jianfeng Zhu, Fan Zhang, Suisha Liang, Wenwei Zhang, Yuanlin Guan, Dongqian Shen, et al. A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature*, 490(7418):55–60, 2012.

[45] Pukar KC Remy Villette, Sophie Beliard, Maria Fernanda Salas Tapia, Dominique Rainteau, Maryse Guerin, and Philippe Lesnik. Unraveling host-gut microbiota dialogue and its impact on cholesterol levels. *Frontiers in Pharmacology*, 11, 2020.

[46] Marc Schoeler and Robert Caesar. Dietary lipids, gut microbiota and lipid metabolism. *Reviews in Endocrine and Metabolic Disorders*, 20(4):461–472, 2019.

[47] Manvi Sharma, Yuanyuan Li, Matthew L Stoll, and Trygve O Tollefsbol. The epigenetic connection between the gut microbiome in obesity and diabetes. *Frontiers in Genetics*, 10, 2019.

[48] Na-Ri Shin, June-Chul Lee, Hae-Youn Lee, Min-Soo Kim, Tae Woong Whon, Myung-Shik Lee, and Jin-Woo Bae. An increase in the *akkermansia* spp. population induced by metformin treatment improves glucose homeostasis in diet-induced obese mice. *Gut*, 63(5):727–735, 2014.

[49] Giorgio Sirugo, Scott M Williams, and Sarah A Tishkoff. The missing diversity in human genetic studies. *Cell*, 177(1):26–31, 2019.

[50] Katharina Timper and Jens C Brüning. Hypothalamic circuits regulating appetite and energy homeostasis: pathways to obesity. *Disease models & mechanisms*, 10(6):679–689, 2017.

[51] Ching-Hung Tseng and Chun-Ying Wu. The gut microbiome in obesity. *Journal of the Formosan Medical Association*, 118:S3–S9, 2019.

[52] Peter WF Wilson, Ralph B Dâ€™Agostino, Helen Parise, Lisa Sullivan, and James B Meigs. Metabolic syndrome as a precursor of cardiovascular disease and type 2 diabetes mellitus. *Circulation*, 112(20):3066–3072, 2005.

- [53] Xiaokang Wu, Chaofeng Ma, Lei Han, Muhammad Nawaz, Fei Gao, Xuyan Zhang, Pengbo Yu, Lianchuan Li, Aiping Zhou, Juan Wang, et al. Molecular characterisation of the faecal microbiota in patients with type ii diabetes. *Current microbiology*, 61(1):69–78, 2010.
- [54] Loic Yengo, Julia Sidorenko, Kathryn E Kemper, Zhili Zheng, Andrew R Wood, Michael N Weedon, Timothy M Frayling, Joel Hirschhorn, Jian Yang, Peter M Visscher, et al. Meta-analysis of genome-wide association studies for height and body mass index in 700000 individuals of european ancestry. *Human molecular genetics*, 27(20):3641–3649, 2018.
- [55] Piriya Yoganathan, Subashini Karunakaran, Maggie M Ho, and Susanne M Clee. Nutritional regulation of genome-wide association obesity genes in a tissue-dependent manner. *Nutrition & metabolism*, 9(1):65, 2012.
- [56] Yong Zhang and Heping Zhang. Microbiota associated with type 2 diabetes and its related complications. *Food Science and Human Wellness*, 2(3-4):167–172, 2013.
- [57] Alexandra Zhernakova, Alexander Kurilshikov, Marc Jan Bonder, Ettje F Tigchelaar, Melanie Schirmer, Tommi Vatanen, Zlatan Mujagic, Arnau Vich Vila, Gwen Falony, Sara Vieira-Silva, et al. Population-based metagenomics analysis reveals markers for gut microbiome composition and diversity. *Science*, 352(6285):565–569, 2016.

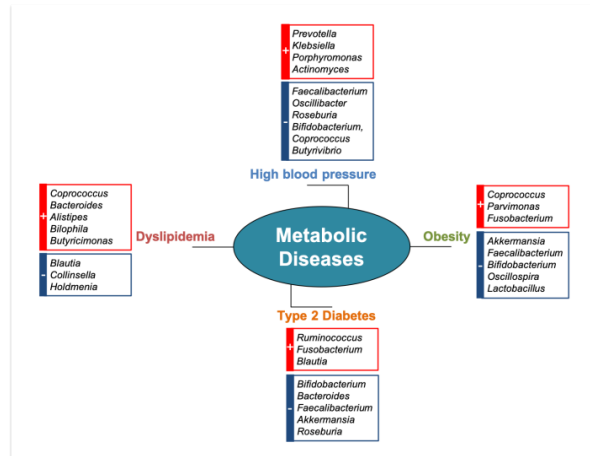
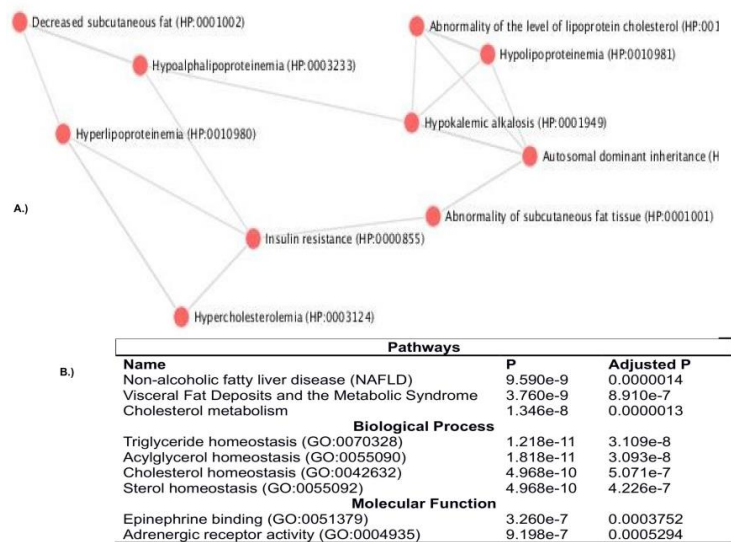
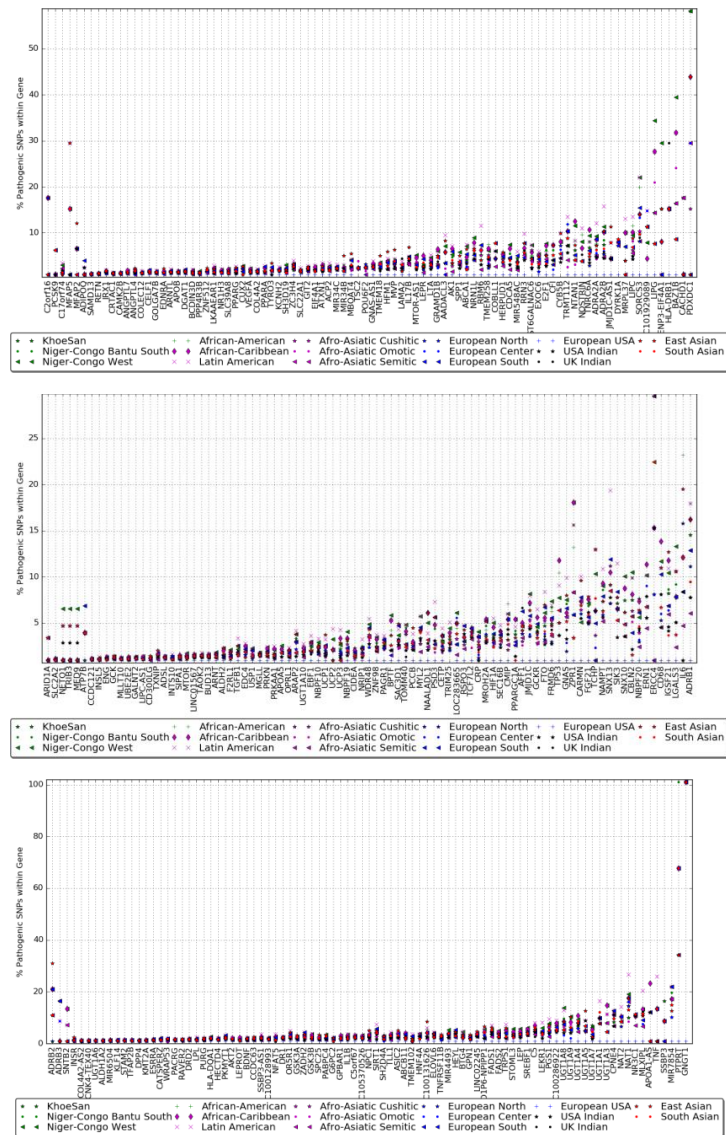


Figure 3: Reported associations of bacterial taxa with metabolic diseases. (+) is increased in abundance, while (-) is decreased in abundance in metabolic diseases.



Supplementary Figure 1: A) Human phenotype sub-network of 458 genes known to associate with metabolic diseases, and; B) top associate pathways, biological processes and molecular functions.



Supplementary Figure 3: Gene-specific proportion of pathogenic SNPs across 18 ethnic groups.

Supplementary Table 1: Data obtained from 1000 Genomes Project (1KGP) (Consortium et al.,2012) and the African Genome Variation Project (AGVP) (Gurdasani et al.,2015) and used for analysis.

Population label	Ethnic group	Population description	Total Samples
AFR	Afro-Asiatic_Semitic	Amhara:Ethiopia	22
	African-American	Americans of African Ancestry in SW USA (ASW)	60
	African-Caribbean	African Caribbeans in Barbado (ACB)	96
	Afro-Asiatic	Al-Gharbiyah, NA, Monufia, Kafrel-Sheikh, Mansoura, Alexandria, Dakahlia, Samanoud, Al-Buhayrah, Minya, AlSharqia, El-Mahalla all from Egypt	99
	Afro-Asiatic_Cushitic	Oromo, Somali from Ethiopia	47
	Afro-Asiatic_Omotic	Wolayta from Ethiopia	24
	Khoe-San	Khoe-San:Khoesan	84
	Niger_Congo_Bantu	Baganda, Banyarwanda, Barundi, Rwandese, Ugandan, Banyankole:Uganda Bakiga, Mutanzania, Basoga, other uganda gwas unknown, Mutooro, Batooro, Nyanjiro (Tanzania) from Uganda and Luhya in Webuye, Kenya (LWK)	2158
	Niger-Congo_Bantu_South	Zulu	98
	Niger_Congo_Volta_Niger	Esan in Nigeria (ESN), Yoruba in Ibadan, Nigeria (YRI)	205
	Niger_Congo_West	Gambian in Western Divisions in the Gambia (GWD), Mende in Sierra Leone (MSL)	198

AMR	Latin_American	Puerto Ricans from Puerto Rico (PUR), Colombians from Medellin, Colombia (CLM), Peruvians from Lima, Peru (PEL), Mexican Ancestry from Los Angeles USA (MXL)	347
EUR	European_center	British in England and Scotland (GBR)	91
	European_North	Finnish in Finland (FIN)	99
	European_South	Iberian Population in Spain (IBS), Toscani in Italia (TSI)	214
	European_USA	Utah Residents with Northern and Western European Ancestry (CEU)	99
EAS	East_Asian	Southern Han Chinese (CHS), Chinese Daiin Xishuangbanna, China (CDX), Kinh in Ho Chi Minh City, Vietnam (KHV), Han Chinese in Beijing, China (CHB), Japanese in Tokyo, Japan (JPT)	504
	South_Asian	Punjabi from Lahore, Pakistan (PJT), Bengali from Bangladesh (BEB)	180
SAS	UK_Indian	Sri Lankan Tamil from the UK (STU), Indian Telugu from the UK (ITU)	204
	USA_Indian	Gujarati Indian from Houston, Texas (GIH)	103
Total			4,932

Supplementary Material

Data Description and Quality Check

We collect from GWAS Catalogue <https://www.ebi.ac.uk/gwas/>, literature and gene-diseases database such DisGeNET <http://www.disgenet.org/> list of genes associated with metabolic diseases. We obtained 458 genes known to associate with metabolic disease. We leveraged the dbSNP database to extract SNPs associated with these genes per diseases and obtained a total of 318,797 SNPs. We accessed vcf files from 1000 Genomes Project¹ Consortium, 2015 and the African Genome Variation Project (AGVP) which has recently characterized the admixture across 18 ethno-linguistic groups from sub-Saharan Africa². A quality control check was conducted on these vcf files using plink and vcftools, we finally retained 2,504 samples from 1000 Genomes Project and 2,428 samples from AGVP². Based on initial sample description (population or country labels), we used the population ethno-linguistic information^{3,4} to categorize the obtained data per ethnic group as described in (**Supplementary Table 1**) and retained only these 318,797 SNPs per ethnic group.

Variant Annotation

The resulting VCF file contained 4,932 samples were split into 18 VCF files per ethnic group (**Supplementary Table 1**) and we used ANNOVAR⁵ to independently perform gene-based annotation in each final VCF data set to determine whether SNPs cause protein coding changes and produce a list of the amino acids that are affected. We used ANNOVAR setting, where the population frequency, pathogenicity for each variant was obtained from 1000 Genomes exome¹, Exome Aggregation Consortium⁶ (ExAC), targeted exon datasets and COSMIC⁷. Gene functions were obtained from RefGene⁸ and different functional predictions were obtained from ANNOVAR's library, which contains up to 21 different functional scores including SIFT^{9,10}, LRT¹¹, MutationTaster¹², MutationAssessor^{13,14}, FATHMM¹⁵, fathmm-MKL¹⁶, RadialSVM¹⁷, LR¹⁷, PROVEAN¹⁷, MetaSVM¹⁷, MetaLR¹⁷, CADD¹⁸, GERP++¹⁹, DANN⁵, M-CAP⁵, Eigen5, GenoCanyon⁵, Polyphen2 HVAR²⁰, Polyphen2 HDIV²¹, PhyloP²² and SiPhy²³. We additionally included conservative and segmental duplication sites, dbSNP code and clinical relevance reported in dbSNP²¹. From each resulting functional annotated data set, we independently filtered for predicted functional status (of which each predicted functional status is of "deleterious" (D), "probably damaging" (D), "disease_causing_automatic" (A) or "disease_causing" (D).^{23,24}) from SIFT, LRT,

MutationTaster, MutationAssessor, FATHMM, fathmm-MKL, RadialSVM, LR, PROVEAN, MetaSVM, MetaLR, CADD, GERP++, DANN, M-CAP, Eigen, GenoCanyon, Polyphen2 HVAR, Polyphen2 HDIV, PhyloP, and SiPhy.

We used a casting vote approach, by retaining only a variant if it had at least 17 predicted functional status “D” or “A” out of 21. Second, the retained variants from each data set were further filtered for rarity, exonic variants, and nonsynonymous mutations and with high quality call as described above, yielding a final candidate list of predicted mutant variants in each subject group, including the replication group. We report on the aggregated SiPhy score from all identified mutant SNPs within the gene. Section below provides details on how SNPs were mapped to genes.

Enrichment of genes

We examined these 458 candidate genes on how they can be associated with human phenotypes and what are their pathways, biological processes and molecular functions. To address this, we used a custom script and adopted to compare our results with different tools including DAVID²⁶ and PANTHER²⁷. We additionally conducted an enrichment analysis using the Enrichr software²⁸. The most significant pathway enriched for genes in the networks were selected from KEGG²⁹, Panther²⁷, Biocarta³⁰ and Reactome³¹, and gene ontologies from the Gene Ontology Consortium² were defined for cellular component, biological process and molecular function.

Distribution of Minor Allele Frequency and Gene-specific in SNP Frequencies

To examine the extent to how common variants across these 18 ethnic groups within metabolic disease, we, therefore, investigated the distribution of the minor allele frequency. To this end, the proportion of minor alleles were categorized into 6 bins (0-0.05, >0.05-0.1, >0.1-0.2, >0.2-0.3, >0.3-0.4, >0.4-0.5) with respect to each ethnic group with a disease. The minor allele frequency (MAF) per SNP for each category was computed using Plink software.

Furthermore, the fraction of genes-specific in SNPs frequency for each gene was computed. To this end, the fraction of gene-specific SNPs frequency was computed, assuming SNPs in upstream and downstream within a gene region are close and possibly in Linkage Disequilibrium (LD). Minor allele frequencies per SNP were aggregated at a gene level (see section Aggregating SNPs Summary Statistics at the Gene level).

Aggregating SNPs Summary Statistics at the Gene level

From each ethnic group, we aggregate SNPs summary statistics to gene level. To do so, we aggregated SNP-specific allele frequencies or proportion of ancestral/derived allele from SNPs 40kb downstream and upstream within gene region as per dbSNPs database. Under the null hypothesis, frequency/proportion p_k ($k = 1, \dots, L$) with a continuous distribution, are uniformly distributed in the interval $[0,1]$. It follows that a parametric cumulative distribution function F can be chosen and p_k can be transformed into quantiles according to $q_k = F(p_k)$. The combined frequency/proportion $C^p = \frac{\sum_{k=1}^L q_k}{\sqrt{L}}$ is a sum of independent and identically distributed random variables q_k . To account for the independence assumption given correlation among neighbouring genomic markers, we implement the Stouffer-Liptak method accounting for spatial correlations among SNPs within a gene or SNPs within a given sub-network. The overall statistic can be obtained by $p = \Phi(C^p)$, in which Φ is the cumulative distribution function of the standard normal distribution.

Associating the Microbiome with Host Phenotype using Dirichlet Process Random Effects Model

"Far better an approximate answer to the right question, which is often vague, than an exact answer to the wrong question, which can always be made precise".

- John Tukey, *The future of data analysis*.

This chapter first provides a brief background describing current statistical methods for associating the human microbiome with its host trait/diseases, highlighting the strengths and limitations of the different approaches. Following this, a description is presented of a proposed statistical method for associating the microbiome with its host's phenotype. The chapter concludes with application of the method to simulated datasets as well as to real microbiome datasets of schizophrenia, HIV/AIDS, and atherosclerosis diseases, and a discussion thereof.

4.1 Introduction

The discovery of the association between the microbiome and multiple human phenotypes has spurred interest in microbiome research to identify biological and environmental factors that are associated with variation in microbiome composition, and to establish the link between microbiome and human clinical

outcome (phenotypes).

To investigate the association between microbiome and host's clinical outcome, microbiome samples are collected, genomic DNA are isolated, and then sequenced using next-generation sequencing (NGS) machines. Two general approaches are followed: shotgun metagenomic sequencing and gene-targeted sequencing. Shotgun metagenomic approach involves an unrestricted sequencing of the genome of all microorganisms present in a sample (Jovel et al., 2016). Meanwhile, the gene-targeted approach targets and sequences a specific gene, usually the 16S rRNA gene, and is restricted to the domain of bacteria and archaea (Janda and Abbott, 2007). There are inherent strengths and limitations to each of these methods. For example, 16S rRNA sequencing may be biased because of unequal amplification of species' 16S rRNA genes, whereas shotgun metagenomic sequencing may not be deep enough to detect the 16S rRNA genes of rare species in a community (Shah et al., 2011). That said, the 16S rRNA gene method routinely used due to its relatively low cost. With this method, after sequencing and preprocessing, the 16S sequences are usually clustered into operational taxonomic units (OTUs) based on similarity, such as 97% similarity level, which are then considered to fairly represent species members.

The resulting species-level dataset of OTUs has several intrinsic properties. First, it is over-dispersed with a large proportion of the OTUs having zero or extremely low abundance across the samples; possibly due to sequencing error, the limit of detection of the sequencing machine or under-sampling for rare taxa (Weiss et al., 2017). Second, it has a complex correlation structure due to the phylogenetic tree structure imposed on the OTUs. Third, it is characterised by variability in sequencing depth (library size) across samples. Fourth, it is high-dimensional, usually consisting of hundreds to thousands of OTUs at a given taxonomic level. Furthermore, it has been observed that phylogenetically-close taxa tend to have similar biological functions; the implication of which is that a given phenotype (response) may result from the joint effect of multiple 'clusters' of taxa at a given taxonomic rank. Accordingly, leveraging the phylogenetic tree structure in analysis is critical for improved inference.

Against this background of microbiome features, a suite of model-based statistical frameworks, which involve direct modelling of the data while allowing the incorporation of key data properties, have been developed. For example, the inflated Gaussian (Paulson et al., 2013), negative binomial (Sohn et al., 2015) and zero-inflated beta (Peng et al., 2016) models. Despite their utility, however, these methods are limited because they do not account for many microbiome features, including the dispersion and correlation, which can result in substantial loss of power. In light of these limitations, other methods have been proposed. Of note, and more recent, include zero-inflated generalized Dirichlet multinomial (ZIGDM) (Tang and Chen, 2018), and kernel-penalized regression (KPR) (Randolph et al., 2018). The ZIGDM model provides a more flexible way to model zero inflation, and the correlation structure in taxon count data. However, the approach does not account for between-samples correlations, in which case all samples are regarded to have similar covariance structure. This can compromise power, since in a practical microbiome study, there is likely to be similarity both within and between samples due to microbial interactions and shared environment or other factors. Meanwhile, KPR addresses the taxa abundance information and uses (dis)similarity kernels to incorporate extrinsic information such as phylogeny in identifying taxon-specific association with a phenotype. However, this method just like others such as (Chen et al., 2012b; Lin et al., 2014; Xiao et al., 2018b), proceed by imposing a smoothness penalty on the regression coefficients, under the premise that response can be explained by only few predictors. Such assumption, although reduces dimensionality and thereby can increase the interpretability of the selected taxa, ignores the important scenario that although the effect of some individual taxa may be weak, their joint effect may be strong. Even crucially, these are primarily tailored to address the question of whether the overall microbiome community is associated with a given host state, and thus cannot be used to investigate the equally interesting question of which specific OTUs are involved in the association.

In this work, a Bayesian framework (herein named, **phylogenetic Dirichlet Process random effects Model**; phyloDPM) is proposed for identifying OTU-

specific association with a given host's phenotype. The proposed model uses a weighted combination of the phylogenetic- and radial basis function-defined similarity kernels to model the covariance structure of taxa effects, and a Dirichlet process (DP) prior to model latent heterogeneity across samples. As the phylogenetic tree itself can be noisy or uninformative with respect to the phenotype under study (Xiao et al., 2018b), using composite kernels leads to improved inference due to its use of contextual information, incorporating/modelling information that a single kernel is unable to. Meanwhile, by modelling the latent variable non-parametrically, the restrictive distributional assumption often made on such random effects is avoided and, instead, the distribution for the parameters is allowed to adapt to the data. In this way, the model remains robust under a wide range of sample features, including multi-modality, heavy tailedness, and skewness; all of which would otherwise result in false inference about the true magnitude of the estimated effects if the true data distribution differs from the assumed model.

4.2 Materials and methods

4.2.1 Model specification

Consider a microbiome study in which a total of p OTUs have been profiled from n individuals. Suppose also that the phenotype of interest has been recorded for each individual. The OTU data can be represented in an $n \times p$ matrix, with rows corresponding to the samples and columns corresponding to the OTUs. A characteristic feature of such data is that the total number of OTUs (which is usually in hundreds or thousands) is much larger than the sample size n (which is usually in tens or hundreds). Another important feature of this data is the correlation among the OTUs, imposed by the phylogenetic tree structure. The phylogenetic tree structure can be specified by a $p \times p$ matrix D of pairwise patristic distances between OTUs; this distance matrix being independent of the

observations in the $n \times p$ data matrix.

We are interested in modelling the association between the OTUs and the phenotype (clinical outcome) of the host. To cast this into a mathematical formulation, let $\mathbf{y} = (y_1, y_2, \dots, y_n)'$ be the binary response of the n individuals; X , a $n \times q$ matrix of covariates (e.g gender, age, etc); and Z , a $n \times p$ matrix of normalized taxa abundance associated with the observations. Let $\boldsymbol{\psi} = (\psi_1, \psi_2, \dots, \psi_n)'$ be a latent variable that accounts for subject-specific deviation from the underlying model. The association of the OTUs with the phenotype is then modeled by the logistic random effects model given by

$$y_i | X_i, Z_i, \psi_i \sim \text{Bernoulli}(\mu_i), \quad \text{where } \mu_i = \frac{\exp [X_i \boldsymbol{\alpha} + Z_i \boldsymbol{\beta} + \psi_i]}{1 + \exp [X_i \boldsymbol{\alpha} + Z_i \boldsymbol{\beta} + \psi_i]}, \quad (4.1)$$

$\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_q)'$ is a q -dimensional vector of covariate effects including an intercept, and $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)'$ is a p -dimensional vector of OTU effects. We seek two primary aims: (1) to account for multicollinearity among bacterial taxa at various phylogenetic depths, by using a weighted combination of co-informative kernels - the phylogenetic and radial basis function (RBF) kernels - to define the covariance structure for taxa effects, and (2) to account for latent structuring (similarity) across samples, by modelling the latent variable $\boldsymbol{\psi}$ as realizations from a Dirichlet process (DP). In the sequel, we description of the formulation the weighted kernel and the DP part.

4.2.2 Modelling correlation structure among microbial taxa

Phylogenetically-close microbial species tends to have similar biological function, and are likely to be associated with the clinical outcome simultaneously; and, accordingly, has been shown to improve regression and prediction (Wang et al., 2019c,c,b; Xiao et al., 2018b,a; Randolph et al., 2018; Chen et al., 2013; Lin et al., 2014; Tamames et al., 2016; Purdom, 2011). In addition, community level effects have been observed where a large number of functionally dependent species in a community are jointly associated with the clinical outcome (Xiao et al., 2018a;

Awany et al., 2019). This highlights the need to incorporate phylogenetic tree structure in the analysis. That being said, however, the phylogenetic tree is in some cases unavailable, misspecified or uninformative with respect to the joint effect of taxa on phenotypic expressions. In this case, a weighted kernel, formed by a linear combination of kernels is more desirable.

4.2.2.1 Phylogeny-based correlation among OTUs

To incorporate the phylogenetic tree information into the analysis, we base on the patristic distance matrix D to define the correlation structure among the OTUs. Using the trait evolution model (Martins and Hansen, 1997), the correlation of traits between OTU j and OTU k , C_{jk} , can be modelled using the following exponential function

$$C_{jk} = \exp\left(-2\delta D_{jk}^\lambda\right), \quad j, k = 1, 2, \dots, p,$$

where D_{jk} is the patristic distance between OTU j and OTU k , and $\delta \in [0, \infty)$ is an evolutionary rate parameter. Following Xiao and Chen (2017), we introduce a power transformation parameter λ on the patristic distances to modulate the decay in correlation coefficient as a function of the patristic distance (Figure 4.1).

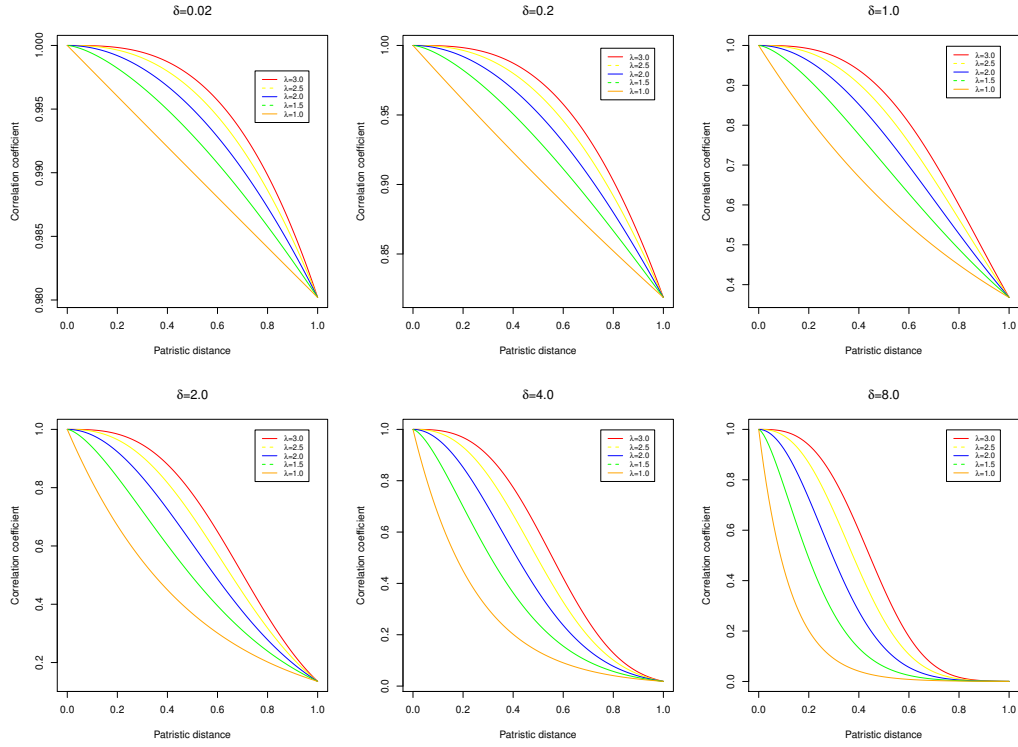


Figure 4.1: The relationship between correlation coefficient and patristic distance, plotted for various values of evolutionary rate parameter and power transformation parameter.

Considering δ , a $\delta = 0$ means that there is no evolution. In this case, $C_{jk} = 1 \forall j, k$ which implies that all traits are identical with maximum phylogenetic relatedness. Meanwhile, $\delta \rightarrow \infty$ means extremely rapid evolution. The result of this is that $C_{jk} \rightarrow 0 \forall j \neq k$, with the implication that the traits have no phylogenetic relationship. From this observation, δ can also be viewed as an OTU dependency measure, by which $\delta = 0$ means all the OTUs are dependent (intricately linked) and $\delta \rightarrow \infty$ all the OTUs are independent. As a corollary, the magnitude of δ also determines the OTU phylogenetic relatedness at the various taxonomic ranks. Let us denote the phylogeny-based correlation matrix by C_{phy} .

4.2.2.2 Radial basis function-based correlation among OTUs

The relationship matrix, C_{phy} , defined above captures signals from OTU effects, permitting a parametric analysis and interpretation with respect to underlying genealogical architecture of microbial taxa. It is also desirable to invoke a similarity measure that captures characteristics of the data, regardless of the underlying phylogenetic relationship. Following the theoretical foundation in [Läuter \(1988\)](#) and the development in [Gianola et al. \(2006\)](#), we define the radial basis function (RBF) kernel as similarity matrix. This non-parametric kernel allows interpretation of classical relatedness as spatial distances between taxa.

The Gaussian RBF kernel, which we make use of in our modelling, embeds taxa in the Euclidean space locussed by a squared Euclidean norm. For example, suppose that each of n individuals possesses measurements, $\mathbf{z}$, of the abundance for p OTUs. Without loss generality, we will regard the vector $\mathbf{z}$ as continuous. Then, the spatial distance between OTUs j and k is given by

$$K(\mathbf{z}_j, \mathbf{z}_k) = \exp\left(-\theta d_{jk}^2\right) = \exp\left(-\theta \|\mathbf{z}_j - \mathbf{z}_k\|^2\right),$$

where $\theta \in \mathbb{R}^+$ is a scaling parameter that determines the overall scale on which the microbial abundance varies across taxa. For very rapidly varying abundance, it is small, and for slowly varying abundance it is large. The smaller the Euclidean distance between two OTUs, the stronger the similarity them. Note, however, that applying exponentiation to the negative Euclidean distance reverses the direction of relationship between inter-OTU distance and RBF kernel. That is, a larger inter-OTU Euclidean distance results in a smaller value of the kernel and a smaller spatial OTU similarity. The kernel matrix serves as the covariance matrix for modelling OTU effects; denote this kernel matrix by C_{rbf} .

Since both C_{phy} and C_{rbf} are Gram matrices ([Greub, 2012](#)), they are positive definite. The composite covariance matrix, C_T , is set up as

$$C_T = wC_{phy} + (1 - w)C_{rbf},$$

where $w \in [0, 1]$ is a weighting parameter. In this exposition, the correlation is modelled as a linear combination of a phylogenetically- and RBF-based similarities. w , which is constrained to ensure positive semi-definiteness of the composite kernel, controls the relative contribution of the phylogenetic and radial basis function kernels to defining the correlation structure among microbial taxa. When $w = 1$, the implicit assumption is that the phylogenetic tree completely specifies correlation among the microbial taxa and the RBF kernel is uninformative. A $w = 0$, however, implies the RBF sufficiently describes the microbial correlation structure and the phylogenetic tree is entirely uninformative. When w is selected as 0.5, the radial basis function and the phylogenetic structure are allowed to contribute equally to define the correlation structure. The optimal choice for w is dependent on the data under consideration, and can, in practice, be obtained through optimization. We here, however, chose to let it be a fixed parameter whose value is specified by the investigator. By doing so, we allow the investigator to control the contributions from the biologically-motivated phylogenetic kernel and the mathematically-derived RBF kernel.

The joint covariance, C , is used to regularize taxa effects. Specifically, the coefficient vector β is set up hierarchically with C as:

$$\beta \sim \mathcal{N}(\mu, \sigma_\beta^2 I); \mu \sim \mathcal{N}(\gamma, C); \gamma \sim \mathcal{N}(\mathbf{0}, \tau_\gamma^2 I),$$

where I is the $p \times p$ identity matrix, σ_β^2 and τ_γ^2 are unknown variance components, and the vector μ is the mean of the Gaussian distribution for the main (OTU) effects. The hierarchical structure allows incorporation of prior correlation structure into the model.

4.2.3 Dirichlet process for the random effects ψ

Consider a stylized situation in which each of a series of individuals possesses measurements for some phenotype of interest, as well as the abundance of microbial OTUs, as obtained from 16S rRNA gene sequencing. In our modelling

framework for identifying OTUs associated with the phenotype of interest, we would like to model random effects, thereby providing us with a more suitable model capable of modelling not only heterogeneity (accounting for unknown sources of variability), most of whose biological relevance are either unknown or unmeasured, but also correlations in the measured observations.

An apt approach to this issue is to add another level of model uncertainty by introducing a random vector and imposing distributional assumptions on it. In this situation, one may be tempted to model the random vector using simple distributions, such as normal or log normal. However, doing so is restricting, and generally not very realistic for empirical datasets. For instance, what if the data is characterized by distinct groupings that would suggest a multimodal distribution? Unfortunately, even more troubling, is the fact that such random effects cannot be checked, as they are not residuals (Burr and Doss, 2005). Bayesian non-parametric models provides an elegant approach to this problem. One such model, which we introduce to our modelling framework, is the Dirichlet process (DP). The DP prior has several appealing features, including incorporating multimodality, expressing our uncertainty about the form of the underlying random effects, and robustness to errors in model specification.

The DP, by the formulation of Ferguson (1973), is a probability measure defined on the probability space $(\Omega, \mathcal{A}, G)$ of all measures G such that for any arbitrary and finite partition $\{B_1, B_2, \dots, B_k\}$ of Ω , the joint distribution of the vector $[G(B_1), G(B_2), \dots, G(B_k)]$ is a Dirichlet distribution with parameters $[mG_0(B_1), mG_0(B_2), \dots, mG_0(B_k)]$, $m(> 0)$, is a concentration (precision) parameter and G_0 is a base distribution defined on Ω . For any such measurable set B , we have $E[G(B)] = G_0(B)$.

The DP provides a flexible approach for modelling latent heterogeneity through construction of probability measures on the space of distribution functions. Assume each ψ_i is a realization from a general distribution that follows a DP, and

suppose further that the base distribution of the DP is normal. That is,

$$\psi_i \sim G; \quad G \sim \text{DP}(m, G_0), \quad G_0 = \text{N}(0, \tau^2).$$

In this formalism, we postulate the random effects to be generated by a mixture of “atoms” of underlying distributions of known parametric form; Gaussian in this case.

The two elements G_0 , the base distribution, and m , the concentration parameter, completely specify the DP. G_0 defines the probability distribution over the space of component parameters, and serves as a prior for the unknown distribution G . Meanwhile, m defines a probability distribution on the component weights, and expresses our uncertainty about the form of G . More intuitively, m controls the variability of the realizations around G_0 , with larger values of it reducing the variability in the realizations.

The predictive distribution of $\boldsymbol{\psi} = (\psi_1, \psi_2, \dots, \psi_n)$ is

$$\pi(\boldsymbol{\psi}|G_0, m) = \int_G \left[\prod_{i=1}^n \pi(\psi_i|G) \pi(G|G_0, m) \right] dG, \quad (4.2)$$

which can be decomposed as

$$\pi(\boldsymbol{\psi}|G_0, m) = \pi(\psi_1|G_0, m) \times \pi(\psi_2|G_0, m, \psi_1) \times \dots \times \pi(\psi_n|G_0, m, \psi_1, \dots, \psi_{i-1}).$$

Following [Blackwell et al. \(1973\)](#), we have $\psi_1|G_0, m \sim G_0$, and the successive conditional distributions are

$$\psi_i|G_0, m, \psi_1, \dots, \psi_{i-1} \sim \frac{m}{i-1+m} G_0(\psi_i) + \frac{1}{i-1+m} \sum_{j=1}^{i-1} \delta(\psi_j = \psi_i), \quad (4.3)$$

where $\delta(\cdot)$ is the usual Dirac delta function.

The construction shows that the predictive distribution of ψ_i is a mixture of

the empirical distributions of G_0 and $(\psi_1, \psi_2, \dots, \psi_{i-1})$, where the weight is a function of m and $(i - 1)$, the number of realizations being conditioned upon. From this formulation, it can be seen that there is a non-zero probability that there will be ties among the realizations. The implication of this is that the DP generates clusters among the observations. The clustering assigns different base distribution parameters across groups and the same parameter within groups. Thus, ψ_i values are independent and identically distributed (i.i.d) only if they are within the same group. This, which is in contrast to the usual parametric modelling strategy that presumes that the random effects are all i.i.d, represents a more general method.

4.3 Model likelihood

Given n independent measurements, the joint distribution of the observed data in Eqn.4.1 can be expressed in a general form as

$$y_1, \dots, y_n \sim f(y_1, \dots, y_n | \boldsymbol{\theta}, \psi_1, \dots, \psi_n) = \prod_{i=1}^n f(y_i | \boldsymbol{\theta}, \psi_i), \quad (4.4)$$

where $\boldsymbol{\theta}$ is a vector of all except the DP model parameters. The likelihood function, $L(\boldsymbol{\theta} | \mathbf{y})$, can then be obtained, by definition, by integrating over the random effects. Therefore

$$L(\boldsymbol{\theta} | \mathbf{y}) = \int f(y_1, \dots, y_n | \boldsymbol{\theta}, \psi_1, \psi_n) \times \pi(\psi_1, \dots, \psi_n) d\psi_1 \dots d\psi_n, \quad (4.5)$$

where $\pi(\psi_1, \dots, \psi_n) = \pi(\boldsymbol{\psi})$ is as previously defined in (4.2).

The clustering in the DP model leads to a succinct mathematical formalism in which the observations can be represented by a binary matrix. Following [Kyung et al. \(2010\)](#); [Li et al. \(2018\)](#), let A be a $n \times k$ binary matrix associated with a particular partition of the sample of size n into k groups, $k = 1, \dots, n$. A partition as such is a cluster or, in a more appropriate terminology, a “subcluster” C , since

the grouping is done non-parametrically and the resulting clusters may differ from that determined by some substantive criteria. The matrix A has the general form,

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1k} \\ a_{21} & a_{22} & \cdots & a_{2k} \\ \vdots & & \cdots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nk} \end{pmatrix}$$

Each row of A has all zeros except for a 1 in one position indicating the group the observation belongs. Meanwhile, the column sums of A represent the number of observations in each of k groups. This means that, using the matrix A , the n -dimensional vector ψ can be represented by a $k(\leq n)$ -dimensional vector η . Then $\psi_i = \eta_j$ so that $\psi = A\eta$, where $\eta \sim N_k(0, \tau^2 I_k)$.

The likelihood function, the joint density of the observations as a function of the parameters, becomes

$$\begin{aligned} L_k(\theta|\mathbf{y}) &= f(\mathbf{y}|\theta) = \int \prod_{i=1}^n \left[\frac{1}{1 + \exp \{ - (X'_i \alpha + Z'_i \beta + (A\eta)_i) \}} \right]^{y_i} \times \\ &\quad \left[\frac{1}{1 + \exp \{ X'_i \alpha + Z'_i \beta + (A\eta)_i \}} \right]^{1-y_i} dG_0(\eta) \\ &= \prod_{i=1}^n \left[\frac{1}{1 + \exp \{ - (X'_i \alpha + Z'_i \beta + (A\eta)_i) \}} \right]^{y_i} \times \\ &\quad \left[\frac{1}{1 + \exp \{ X'_i \alpha + Z'_i \beta + (A\eta)_i \}} \right]^{1-y_i} \times \prod_{j=1}^k \left(\frac{1}{2\pi\tau^2} \right)^{\frac{1}{2}} \exp \left(-\frac{1}{2\tau^2} \eta_j^2 \right). \quad (4.6) \end{aligned}$$

To complete the Bayesian modelling, we specify the following priors on the other model parameters:

$$\begin{aligned} \alpha | \sigma_\alpha^2 &\sim N_q(\mathbf{0}, \sigma_\alpha^2 I_q), \quad \tau^2 \sim \text{IG}(a_\tau, b_\tau) \\ \sigma_\alpha^2 &\sim \text{IG}(a_\alpha, b_\alpha), \quad \text{and} \quad \sigma_\beta^2 \sim \text{IG}(a_\beta, b_\beta). \end{aligned} \quad (4.7)$$

4.4 Bayesian inference of the parameters

In this section, we consider estimation of the parameters involved in our Dirichlet process logistic random effects model. The ensemble of parameters can be considered into three sets: (i) A , the matrix of sub-clusters (ii) m , the precision parameter of the Dirichlet process, and (iii) $\theta = (\alpha, \beta, \mu, \gamma, \tau_\gamma^2, \eta, \tau^2, \sigma_\alpha^2, \sigma_\beta^2)$, the other model parameters. Using the algorithmic idea in [Kyung et al. \(2009\)](#), these three sets of parameters can be iteratively estimated group-wise as follows: conditional on A and m , simulate θ ; then conditional on θ and A , simulate m ; finally, conditional on θ and m , simulate A .

4.4.1 Estimation of model parameters, θ .

Bayesian inference treats the observed data as fixed and the model parameters as random variables. Given a parameter θ and a set of observed data y , the Bayesian approach seeks to obtain the probability of the parameter θ given the set of data available y , mathematically expressed as $p(\theta|y)$. Given a prior information about the parameter θ , which is formation is expressed as a probability distribution $p(\theta)$, interest is in the posterior distribution of the parameter given then data, $p(\theta|y)$. This posterior distribution is obtained by combining the prior distribution with the likelihood function, using Baye's theorem. Applying this theorem, the likelihood in Eqn.4.6 can be combined with the priors in Eqn.4.7 to give the following joint posterior distribution.

$$\begin{aligned}
 f(\theta|A, \mathbf{y}) &\propto L(\theta|A, \mathbf{y}) \times \left(\frac{1}{\tau^2}\right)^{a_\tau+1} \exp\left(-\frac{b_\tau}{\tau^2}\right) \times \left(\frac{1}{\sigma_\alpha^2}\right)^{\frac{q}{2}} \exp\left(-\frac{|\alpha|^2}{2\sigma_\alpha^2}\right) \times \\
 &\left(\frac{1}{\sigma_\beta^2}\right)^{\frac{p}{2}} \exp\left(-\frac{(\beta - \mu)^2}{2\sigma_\beta^2}\right) \times \left(\frac{1}{\sigma_\alpha^2}\right)^{a_\alpha+1} \exp\left(-\frac{b_\alpha}{\sigma_\alpha^2}\right) \times \left(\frac{1}{\sigma_\beta^2}\right)^{a_\beta+1} \times \\
 &\exp\left(-\frac{b_\beta}{\sigma_\beta^2}\right) \times \exp\left(-\frac{(\mu - \gamma)C_T^{-1}(\mu - \gamma)}{2}\right) \times \exp\left(-\frac{|\gamma|^2}{2\tau_\gamma^2}\right). \quad (4.8)
 \end{aligned}$$

The complexity of the dependency structure of the model, imposed by the DP component, makes direct Markov Chain Monte Carlo (MCMC) sampling of the lower-level parameters using the Gibbs sampler impossible to implement, because the resulting conditional distributions assume awkward forms. In this case, one may be tempted to turn to the Metropolis–Hastings algorithms. However, these algorithms can be difficult to set up for such complex functional forms, and may require ‘tuning’ to achieve satisfactory performance. Moreover, even though the ‘black box’ random variate generation techniques such as the ratio-of-uniforms method (Wakefield et al., 1991) may seem an alternative, their use can be daunting since they also require tuning to achieve reliable and efficient sampling.

To circumvent these difficulties, we invoke the auxiliary variable (*slice* sampling) method (Damien et al., 1999). The idea of the method is to augment the parameter space with a set of convenient random (auxiliary) variables that keeps the marginal posterior distribution of interest unchanged but converts all of the full conditionals into a set of easily sampled ‘standard’ full conditional distributions. Additionally, the technique also improves sampling efficiency (Lu et al., 2018; Kyung et al., 2009). By this, the slice-Gibbs sampling can be used to sample the lower-level parameters, whereas the higher-level parameters can be directly sampled using the Gibbs sampler applied to their respective full-conditional distributions.

4.4.2 Full conditional distributions

We now proceed to derive the full-conditional distribution for each parameter in the model. For notational convenience, let $\boldsymbol{\theta}$ be the vector of all model parameters and $\boldsymbol{\theta}_{-p}$ be the vector of all these parameters except p . Let

$$p_i = P(y_i = 1 | \boldsymbol{\theta}, X, Z, \mathbf{y}) = \frac{\exp \{X'_i \boldsymbol{\alpha} + Z'_i \boldsymbol{\beta} + (A\boldsymbol{\eta})_i\}}{1 + \exp \{X'_i \boldsymbol{\alpha} + Z'_i \boldsymbol{\beta} + (A\boldsymbol{\eta})_i\}} \quad (4.9)$$

and let $q_i = 1 - p_i$. For each response variable y_i , we introduce two mutually independent latent variables λ_i and ϕ_i , where each is assumed to be uniformly

distributed over the interval $(0, 1)$. Then, according to Eqn.4.9, the variables λ_i and ϕ_i are constrained by the response variable y_i . When $y_i = 1$, we have $p_i = E_\lambda [I(0 < \lambda_i \leq p_i)]$, and when $y_i = 0$, we have $q_i = E_{\phi_i} [I(0 < \phi_i \leq q_i)]$. Therefore, the likelihood function in Eqn.4.6 can be rewritten using the expectation of the auxiliary variables. That is,

$$L(\boldsymbol{\theta}|A, \mathbf{y}) = \prod_{i=1}^n [I(y_i = 1)I(0 < \lambda_i \leq p_i) + I(y_i = 0)I(0 < \phi_i \leq q_i)] \times \prod_{j=1}^k \left(\frac{1}{2\pi\tau^2} \right)^{\frac{1}{2}} \exp \left(-\frac{1}{2\tau^2} \eta_j^2 \right). \quad (4.10)$$

Using (4.8) and (4.10), we obtain the following full conditional distributions:

(a) Full conditionals for λ and ϕ are both uniform.

For $i = 1, \dots, n$, we have

$$\lambda_i | \boldsymbol{\theta}_{-\lambda_i}, X, Z, \mathbf{y} \sim \text{Uniform}(0, p_i), \quad \text{if } y_i = 1 \quad (4.11)$$

$$\phi_i | \boldsymbol{\theta}_{-\phi_i}, X, Z, \mathbf{y} \sim \text{Uniform}(0, q_i), \quad \text{if } y_i = 0 \quad (4.12)$$

The full conditional distribution for $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $\boldsymbol{\eta}$ are, respectively, the following truncated distributions:

(b) Full conditional distribution for α_j ; $j \in \{1, \dots, q\}$:

Let $S_1 = \{(i, j) : y_i = 1, X_{ij} > 0\}$ and $S_2 = \{(i, j) : y_i = 0, X_{ij} > 0\}$, then we have

$$\alpha_j | \boldsymbol{\theta}_{-\alpha_j}, X, Z, \mathbf{y} \sim \begin{cases} N(0, \sigma_\alpha^2), & \alpha_j \in [\alpha_j^L, \alpha_j^U] \\ 0, & \text{otherwise} \end{cases} \quad (4.13)$$

where

$$\alpha_j^L = \max_{i \in S_1} \left\{ \frac{1}{X_{ij}} \left[\log \left(\frac{\lambda_i}{1 - \lambda_i} \right) - \sum_{k \neq j} X_{ik} \alpha_k - Z_i' \boldsymbol{\beta} - (A\boldsymbol{\eta})_i \right] \right\}, \text{ and}$$

$$\alpha_j^U = \min_{i \in S_2} \left\{ \frac{1}{X_{ij}} \left[\log \left(\frac{1 - \phi_i}{\phi_i} \right) - \sum_{k \neq j} X_{ik} \alpha_k - Z'_i \boldsymbol{\beta} - (A\boldsymbol{\eta})_i \right] \right\}.$$

(c) Full conditional distribution for β_j ; $j \in \{1, \dots, p\}$.

Because some elements of the matrix L can be negative, let $S_1 = \{(i, j) : y_i = 1, Z_{ij} > 0\}$, $S_2 = \{(i, j) : y_i = 0, Z_{ij} < 0\}$, $S_3 = \{(i, j) : y_i = 1, Z_{ij} < 0\}$, and $S_4 = \{(i, j) : y_i = 0, Z_{ij} > 0\}$. Then, we have

$$\beta_j | \boldsymbol{\theta}_{-\beta_j}, X, Z, \mathbf{y} \sim \begin{cases} N(\mu_j, \sigma_\beta^2), & \beta_j \in [\beta_j^L, \beta_j^U] \\ 0, & \text{otherwise} \end{cases} \quad (4.14)$$

where

$$\begin{aligned} \beta_j^L = \max & \left\{ \max_{i \in S_1} \left\{ \frac{1}{Z_{ij}} \left[\log \left(\frac{\lambda_i}{1 - \lambda_i} \right) - X'_i \boldsymbol{\alpha} - \sum_{k \neq j} Z_{ik} \beta_k - (A\boldsymbol{\eta})_i \right] \right\}, \right. \\ & \left. \max_{i \in S_2} \left\{ \frac{1}{Z_{ij}} \left[\log \left(\frac{1 - \phi_i}{\phi_i} \right) - X'_i \boldsymbol{\alpha} - \sum_{k \neq j} Z_{ik} \beta_k - (A\boldsymbol{\eta})_i \right] \right\} \right\} \end{aligned}$$

and

$$\begin{aligned} \beta_j^U = \min & \left\{ \min_{i \in S_3} \left\{ \frac{1}{Z_{ij}} \left[\log \left(\frac{\lambda_i}{1 - \lambda_i} \right) - X'_i \boldsymbol{\alpha} - \sum_{k \neq j} Z_{ik} \beta_k - (A\boldsymbol{\eta})_i \right] \right\}, \right. \\ & \left. \min_{i \in S_4} \left\{ \frac{1}{Z_{ij}} \left[\log \left(\frac{1 - \phi_i}{\phi_i} \right) - X'_i \boldsymbol{\alpha} - \sum_{k \neq j} Z_{ik} \beta_k - (A\boldsymbol{\eta})_i \right] \right\} \right\}. \end{aligned}$$

(d) Full conditional distribution for η_k ; $k \in \{1, \dots, K\}$. Let $S_1 = \{i : y_i = 1\}$ and $S_2 = \{i : y_i = 0\}$, then we have

$$\eta_k | \boldsymbol{\theta}_{-\eta_k}, X, Z, \mathbf{y} \sim \begin{cases} N(0, \tau^2), & \eta_k \in [\eta_k^L, \eta_k^U] \\ 0, & \text{otherwise} \end{cases} \quad (4.15)$$

where $\eta_k^L = \max_{S_1} \left\{ \log \left(\frac{\lambda_i}{1 - \lambda_i} \right) - X_i' \boldsymbol{\alpha} - Z_i' \boldsymbol{\beta} \right\}$, and
 $\eta_k^U = \min_{S_2} \left\{ \log \left(\frac{1 - \phi_i}{\phi_i} \right) - X_i' \boldsymbol{\alpha} - Z_i' \boldsymbol{\beta} \right\}.$

(e) The full conditional distribution for $\boldsymbol{\mu}$.

$$f(\boldsymbol{\mu} | \dots) \propto \exp \left\{ -\frac{1}{2} \left[\frac{1}{\sigma_\beta^2} (\boldsymbol{\beta} - \boldsymbol{\mu})^2 + (\boldsymbol{\mu} - \boldsymbol{\gamma}) C_T^{-1} (\boldsymbol{\mu} - \boldsymbol{\gamma}) \right] \right\} \quad (4.16)$$

The quadratic forms of $\boldsymbol{\mu}$ in Eqn.4.16 can be combined using the [Box and Tiao \(1973\)](#) identity for quadratic functions. The identity states that

$$\begin{aligned} (\mathbf{z} - \mathbf{m})' \mathbf{M} (\mathbf{z} - \mathbf{m}) + (\mathbf{z} - \mathbf{b})' \mathbf{B} (\mathbf{z} - \mathbf{b}) &= (\mathbf{z} - \mathbf{c})' (\mathbf{M} + \mathbf{B}) (\mathbf{z} - \mathbf{c}) + \\ &\quad (\mathbf{m} - \mathbf{b})' \mathbf{M} (\mathbf{M} + \mathbf{B})^{-1} \mathbf{B} (\mathbf{m} - \mathbf{b}), \end{aligned} \quad (4.17)$$

where $\mathbf{c} = (\mathbf{M} + \mathbf{B})^{-1} (\mathbf{Mm} + \mathbf{Bb})$.

Using this identity, we can combine the sums in Eqn.4.16 as

$$\begin{aligned} \frac{1}{\sigma_\beta^2} (\boldsymbol{\beta} - \boldsymbol{\mu})^2 + (\boldsymbol{\mu} - \boldsymbol{\gamma}) C_T^{-1} (\boldsymbol{\mu} - \boldsymbol{\gamma}) &= (\boldsymbol{\mu} - \hat{\boldsymbol{\mu}})' \left(\frac{1}{\sigma_\beta^2} I + C^{-1} \right) (\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}) + \\ &\quad (\boldsymbol{\beta} - \boldsymbol{\gamma})' \frac{1}{\sigma_\beta^2} \left(\frac{1}{\sigma_\beta^2} I + C^{-1} \right)^{-1} C^{-1} (\boldsymbol{\beta} - \boldsymbol{\gamma}), \end{aligned} \quad (4.18)$$

where $\hat{\boldsymbol{\mu}} = \left(\frac{1}{\sigma_\beta^2} I + C^{-1} \right)^{-1} \left(\frac{1}{\sigma_\beta^2} \boldsymbol{\beta} + C^{-1} \boldsymbol{\gamma} \right).$

Putting Eqn.4.18 in (4.16) and retaining only the parts that varies with $\boldsymbol{\mu}$, the full conditional distribution can be expressed as

$$f(\boldsymbol{\mu} | \dots) \propto \exp \left\{ -\frac{1}{2} (\boldsymbol{\mu} - \hat{\boldsymbol{\mu}})' \left(\frac{1}{\sigma_\beta^2} I + C^{-1} \right) (\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}) \right\},$$

which is immediately recognized as the kernel of a Gaussian distribution with mean vector $\hat{\boldsymbol{\mu}}$, as defined above, and variance-covariance matrix $\left(\frac{1}{\sigma_\beta^2}I + C^{-1}\right)$. Thus, component-wise, we can write

$$\mu_j | \dots \sim \mathcal{N} \left(\left\{ \left(\frac{1}{\sigma_\beta^2} + (C'_j C_j)^{-1} \right)^{-1} \left(\frac{1}{\sigma_\beta^2} \beta_j + (C'_j C_j)^{-1} \gamma_j \right) \right\}, \right. \\ \left. \left(\frac{1}{\sigma_\beta^2} + (C'_j C_j)^{-1} \right) \right)$$

(f) The full conditional distribution for γ .

Following similar exposition of part **(e)** above, we find that

$$\gamma_j | \dots \sim \mathcal{N} \left(\left\{ \left(\frac{1}{\tau_\gamma^2} + (C'_j C_j)^{-1} \right)^{-1} \left((C'_j C_j)^{-1} \mu_j \right) \right\}, \right. \\ \left. \left(\frac{1}{\tau_\gamma^2} + (C'_j C_j)^{-1} \right) \right)$$

(g) Finally, the full conditional distribution for τ^2 , σ_α^2 , and σ_β^2 are, respectively:

$$\tau^2 | \boldsymbol{\theta}_{-\tau^2}, X, Z, \mathbf{y} \sim \text{IG} \left(\frac{k}{2} + a_\tau, \frac{1}{2} |\boldsymbol{\eta}|^2 + b_\tau \right) \quad (4.19)$$

$$\sigma_\alpha^2 | \boldsymbol{\theta}_{-\sigma_\alpha^2}, X, Z, \mathbf{y} \sim \text{IG} \left(\frac{q}{2} + a_\alpha, \frac{1}{2} |\boldsymbol{\alpha}|^2 + b_\alpha \right) \quad (4.20)$$

$$\sigma_\beta^2 | \boldsymbol{\theta}_{-\sigma_\beta^2}, X, Z, \mathbf{y} \sim \text{IG} \left(\frac{p}{2} + a_\beta, \frac{1}{2} (\boldsymbol{\beta} - \boldsymbol{\mu})^2 + b_\beta \right), \quad (4.21)$$

where $\text{IG}(\cdot, \cdot)$ denotes the inverse Gamma distribution.

4.4.3 Generation of sub-clusters matrix A .

For the generation of the sub-clusters matrix A , we apply the method of [Kyung et al. \(2009\)](#), where a Metropolis-Hastings algorithm with a Dirichlet proposal distribution is used. The method is based on recognizing the fact that the marginal distribution of any k components taken from an n -dimensional multinomial/Dirichlet distribution is also multinomial/Dirichlet. Therefore, a sampling scheme can be employed with candidate draws taken from a multinomial/Dirichlet distribution.

Starting from the model parameters at iteration t :

$\theta^{(t)} = (\alpha^{(t)}, \beta^{(t)}, \mu^{(t)}, \gamma^{(t)}, \tau_\gamma^{2(t)}, \eta^{(t)}, \tau^{2(t)}, \sigma_\alpha^{2(t)}, \sigma_\beta^{2(t)})$ and $A^{(t)}$, the Dirichlet distribution parameters is first simulated from:

$$\begin{aligned} \mathbf{q}^{(t+1)} &= (q_1^{(t+1)}, q_2^{(t+1)}, q_n^{(t+1)}) \\ &\sim \text{Dirichlet}(s_1^{(t)} + 1, s_2^{(t)} + 1, \dots, s_k^{(t)} + 1, \overbrace{1, 1, 1, \dots, 1}^{n-k \text{ components}}), \end{aligned}$$

where $s_j \geq 1$ are the non-zero column sums of A , and $s_1 + s_2 + \dots + s_k = n$. Then based on the value of $\mathbf{q}^{(t+1)}$, we generate $A^{(t+1)}$ from a Multinomial distribution,

$$A^{(t+1)} \sim P(A^{(t)})f(y|\theta^{(t+1)}, A^{(t)}) \binom{n}{\mathbf{n}_1, \dots, \mathbf{n}_{k'}} \prod_{j=1}^{k'} [q_j^{(t+1)}]^{n'_j}.$$

The columns with zero column sums are then deleted, resulting in a $n \times k$ matrix. Deleting the zero sum columns is a marginalization of the multinomial distribution.

4.4.3.1 Estimation of precision parameter m .

The precision parameter m , unknown before hand, plays pivotal role in the distribution of ψ as it determines the number of unique clusters. Two general approaches have been pursued to estimate the precision parameter of Dirichlet

processes. The first employs empirical Bayes method in which m is obtained by the maximum likelihood estimate (MLE) method, and the second undertakes a Bayesian approach so that a prior distribution is placed on m and estimated using Markov Chain Monte Carlo (MCMC) method. Although appealing due to its insensitivity to the choice of prior distribution, the MLE approach cannot account for the errors in estimation and has been shown to be unstable in some situations (Dorazio, 2009). Accordingly, we undertake the Bayesian approach for our current framework. The main appeal of the Bayesian approach here is three fold. First, is the ability to account for uncertainty in estimation of the ‘true’ value. Second, is the flexibility that in the absence of prior knowledge about the value, which is usually the case, one can specify a prior distribution with large variance; this assigns a mass across all possible and realistic values. Third, is computational consideration; Bayesian prior of is readily amenable to the many established sampling schemes.

In the Bayesian specification, we elicit a Gamma (a, b) prior for m , where a and b are the hyper-parameters, and use Escobar and West (1995)’s computational sampling scheme to sample m from its conditional posterior distribution. By this, m is sampled from a mixture of two Gamma densities augmented with a Beta distributed random variable. Specifically, suppose clustering matrix A has been sampled. In doing so, a value for k has been sampled, and the specific configuration of the samples into k groups. Based on this, the sampling proceeds as follow. First, an auxiliary variable ν is sampled from a Beta distribution, i.e. $\nu|m, k, \dots \sim \text{Beta}(m+1, n)$. Then, based on this, m is sampled from the mixture:

$$m|\nu, k, \dots \sim \frac{c+k-1}{c+k-1+n(d-\log(\nu))} \times \text{Gamma}(c+k, d-\log(\nu)) + \frac{n(d-\log(\nu))}{c+k-1+n(d-\log(\nu))} \times \text{Gamma}(c+k-1, d-\log(\nu)).$$

In this way, m can be iteratively updated within the general Gibbs sampling framework, along with the other model parameters.

4.5 Application

4.5.1 Simulation studies

We conducted a comprehensive simulation study to demonstrate the power of the proposed model under various conditions of microbiome associations, and to justify the use of multiple kernels - the phylogenetic and RBF kernels - to model correlation structure among OTUs. These simulations were designed to answer at least two key questions: (1) what is the relative performance of the proposed model in identifying bacterial taxa associated with the phenotype, and how does this performance scale with increasing number of microbial taxa? (2) what gain in performance do we achieve by using multiple kernels to model correlation structures among the OTUs?

4.5.1.1 Simulation setting and data generation

Recent studies have performed simulations for microbiome count data, using negative binomial (NB), log-normal, multinomial or Dirichlet multinomial (DM) distributions (Chen et al., 2012a; Chen and Li, 2013; Sohn et al., 2015; Xu et al., 2015; Tang and Chen, 2018; Xiao et al., 2018b). However, due to the high degree of heterogeneity among microbiome samples, the DM and NB distributions are more suitable because of their ability to more sufficiently model overdispersion in counts. Accordingly, in this application, OTU counts were simulated using a DM distribution. In order to make the simulation more realistic, the parameters of the DM distribution (mean and dispersion) were estimated from real microbiome data of human upper respiratory microbiome in the study by Charlson et al. (2010). This dataset contains the counts of 856 OTUs from 60 samples. Also available from this dataset is the phylogenetic tree structure for these OTUs.

The $p \in \{200, 400, 800\}$ most abundant OTUs from this dataset were used for parameter estimation, and, accordingly, p OTUs were subsequently simulated for

$n = 500$ individuals. The simulated OTU counts data closely mimics real data (Figure 4.2), characterized by zero-inflation and overdispersion.

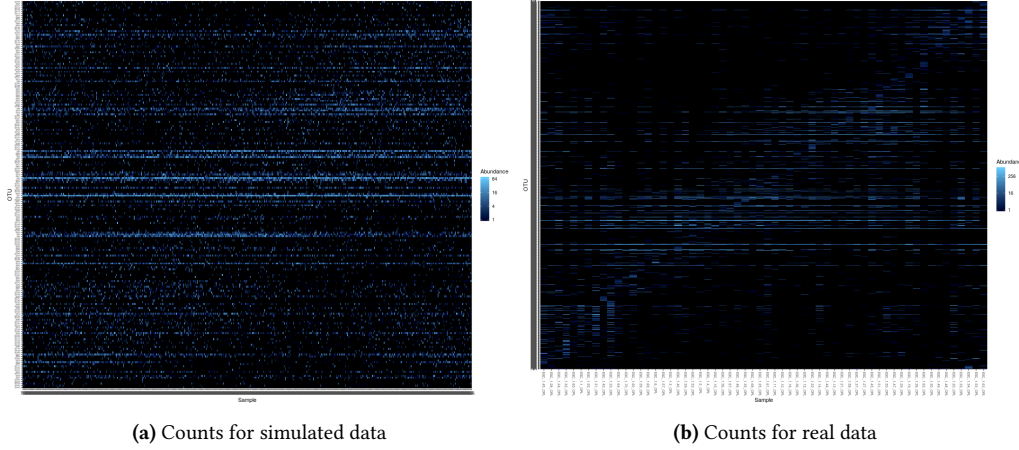


Figure 4.2: Heatmap showing the OTU abundance distribution for (a) simulated microbial community of 200 OTUs from 500 samples, and (b) real microbial community of 856 OTUs obtained from the human upper respiratory tract from 60 samples. The scale indicates abundance level of the OTU counts with dark representing zero counts (see legends).

The resulting OTU count data for each sample was normalized by the total read count for each sample, where the total read count for a sample was obtained by sampling from a negative binomial distribution with mean 5000 and dispersion 25, which reflects typical sequencing depths in empirical microbiome studies. The simulation steps are summarized below:

- (i) Using the real microbiome data, estimate the mean parameters, $\boldsymbol{\pi} = (\pi_1, \pi_2, \dots, \pi_p)$, and dispersion parameter, θ , via maximum likelihood.
- (ii) Simulate OTU counts using DM distribution with parameters $\boldsymbol{\pi}$ and θ .
- (iii) Generate the total read count for each sample by sampling from negative binomial distribution with mean 5000 and dispersion 25.
- (iv) Normalize the simulated OTU data for each sample by its simulated total read count. This normalization process is essential before any analysis can be done, to mitigate bias arising from variability in sequencing depth across samples (Li et al., 2015).

4.5.1.2 Simulating phenotype

Having simulated the OTU data for the individuals, we now generate the corresponding phenotype for the individuals. To simulate a phylogeny-informative scenario, we adapted the simulation strategy by [Xiao et al. \(2018b\)](#). The partitioning-around-medoids (PAM) algorithm was used to partition the p OTUs into clusters into 50 different sizes, based on the patristic distances between them. Then, one cluster with the highest abundance was chosen and ten most-abundant OTUs from this cluster were selected to be associated to the phenotype.

We need to simulate clustered samples. So, before we simulate the phenotype, we first generate a clustering effect that represent a more general feature of sample heterogeneity. We achieve this by simulating random effects, ψ , that follow a Dirichlet process. For each individual,

$$\psi_i \sim G, \quad \text{and } G \sim \text{DP}(m, G_0).$$

To simulate this random vector, we require the two elements of the DP: G_0 and m . We let $G_0 \sim N(0, 1)$. To set a value for m , we fix the number of subclusters, K , to 40 (8% of the sample size), and make use of the mathematical relationship between K and m , which is $K = \sum_{i=1}^n \frac{m}{m+i-1}$, to obtain m . This yields $m = 24.21$. With G_0 and m , we use the Polya-urn model in (4.3) to simulate ψ_i , ($i = 1, \dots, n$).

Finally, the phenotype, y_i , ($i = 1, \dots, n$) is simulated from the model

$$y_i = \text{Bin}(p_i); \quad p_i = \frac{1}{1 + \exp(-\{\sum_{j \in \mathcal{A}} Z_{ij} \beta_j + \psi_i\})},$$

where $\mathcal{A}$ is the set containing the associated OTUs, and β_j is the effect size of the j^{th} OTU.

4.5.1.3 Choice of hyperparameters and sensitivity analysis

In applying the model to the datasets, the hyper-parameters for the Bayesian priors were set as follows: In light of the simulation study by [Murugiah and Sweeting \(2012\)](#), we let the parameters c and d for the Gamma distribution of the precision parameter $m \sim \Gamma(c, d)$ be $c = d = \exp(-0.033 \times n)$ where n is the sample size. For the variance components, while the use of weakly informative inverse gamma hyper-parameters such as $\text{Gamma}(0.1, 0.1)$ or $\text{Gamma}(1, 0.001)$ is popular, our test simulations showed their use leads to very poor MCMC posterior samples; possibly given that such hyperparameters place the majority of the prior mass away from zero. Therefore, it is imperative to carry out a sensitivity analysis to assess the impact of hyper-parameter values on the performance of the model, and, subsequently, to specify appropriate hyper-parameter values. To conduct sensitivity analysis with respect to the hyper-parameter values, we performed several simulation runs, varying hyper-parameter values (a_τ, b_τ) and (a_β, b_β) from 1 to 20, and used Matthews correlation coefficient (MCC) ([Matthews, 1975](#)) to compare the results from the different simulations.

Specifically, we first conduct initial test runs with hyper-parameter values (a_τ, b_τ) and (a_β, b_β) from 1 to 20, monitoring the Bayesian posterior distribution of the parameters. Based on this, (a_τ, b_τ) is fixed to a value $(a_\tau^*, b_\tau^*) \in \{1, 20\}$ that provide stable posterior estimates, and several simulation runs are performed for various values of (a_β, b_β) from 1 to 20. Then, in each replicate, the discriminatory power of the model (in discerning case from control subjects) is evaluated by calculating the number of true positive (TP), true negative (TN), false positive (FP) and false negative (FN), from where MCC is calculated as

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}}.$$

MCC score takes values in the interval $[-1, 1]$, with 1 showing a complete agreement (each case subject is classified as a case and each control subject is classified as a control), and -1 a complete disagreement (a case subject is classified as

control and *vice versa*). MCC provides a good measure of global accuracy, since it accounts for both sensitivity and specificity. Because of this, it is widely used as a performance metric in Bioinformatics (Boughorbel et al., 2017). Following the above procedure, we take the ‘appropriate’ hyper-parameter pair (a_β, b_β) to be the pair that yields the highest value of the MCC out of all the tested hyper-parameter settings. Similarly, we determine the ‘appropriate’ hyper-parameter (a_τ, b_τ) by fixing (a_β, b_β) at the obtained optimal value, and varying (a_τ, b_τ) over the same interval (1 to 20) and calculating MCC for several replicates. Figure 4.3 displays the heat-map of average MCC values with respect to various choices of inverse Gamma hyper-parameters. As can be observed from the plot, for τ^2 , the MCC is undesirable for values of $a_\tau \leq 1$ and $b_\tau \leq 1$. Similarly, for σ_β^2 , MCC is sub-optimal for $a_\beta \leq 5$ and $b_\beta \leq 5$. Meanwhile, fixing one of hyper-parameter values a_τ (a_β for σ_β^2) or b_τ (b_β for σ_β^2) to a small value and increasing the other component still result in a poor posterior, possibly since the associated $IG(a, b)$ prior with mean $b/(a-2)$ and variance $b^2/(a-2)^2$ is strongly informative in this scenario. We complemented the sensitivity analysis by monitoring the changes in Bayesian posterior estimates of the parameters, on a grid of values of the hyper-parameters in the same interval (1 to 20). From these analyses, we chose $a_\tau = b_\tau = 20$ and $a_\beta = 20, b_\beta = 15$ as the well performing hyper-parameter configurations for the variance parameters σ_β^2 and τ^2 respectively.

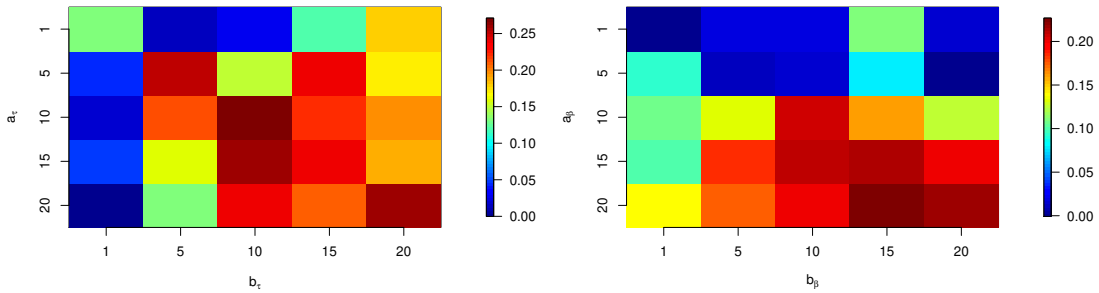


Figure 4.3: Heatmap showing Matthews correlation coefficient (MCC) values for various choices of (a, b) hyper-parameters for the inverse-gamma prior on the variance components τ^2 and σ_β^2 . Each MCC value is the average obtained for 10 replications.

The kernel coefficient of the RBF is fixed to 10^{-3} for all analyses. This parameter setting achieved optimal performance in test runs. Each simulation experiment was conducted for 100 replications.

4.5.1.4 Results & discussion: simulation experiment

We aimed to assess the performance of the proposed model on simulated microbiome OTU datasets for case-control subjects. To achieve this, we simulate OTUs across a number of individuals and set some to be ‘causal’. The final outcome is to apply the proposed model to the simulated data to estimate individual OTU effects, and compare with the preset known values. In this way, we can test the model’s ability to recover simulated values. To this end, we compute the summary statistics and a 95% credible interval and see if the true parameter value lies within the interval, which is indicative of good model fit (Brody-Moore, 2019).

In Bayesian methods employing MCMC sampling techniques, it is very important to assess convergence of the Markov chains, as non-convergence in the simulated MCMC samples means that the generated chain does not converge, in distribution, to the target posterior distribution of interest, and any inference from such posterior samples would be unreliable. Therefore, for each simulation scenario, we first conducted test runs to check the convergence property of the chain, and found that in each case the chain mixes properly within 5000 iterations. Convergence of the chains were assessed by visual inspection using trace plot, auto-correlation function plot and cumulative quantile plot (see plots of posterior distribution of the parameters in Figure 4.6), as well as formal diagnostic test using Geweke diagnostic (Geweke et al., 1991). All these showed no evidence of non-convergence. We observe from the plots that the chain moves quite quickly over the possible values of the parameters, moving directly to the centre of the target posterior distribution. Following this, the Markov chain was run for 50000 iterations from where the first 5000 was discarded as burn-in, and, to remove the autocorrelation, a lag of 10 was considered. The resulting effective samples were the ones finally used for posterior inference to produce summary statistics

(Figure 4.1). Because we know the true values of the simulated effects, we can now assess the model's performance.

Table 4.1: Results of estimation in the simulation study. Estimate of the parameter (Estim) and standard error of the estimate (SE) based on 100 replications.

Parameter	$\mu = 0.0$						$\mu = 0.2$					
	$p = 200$		$p = 400$		$p = 800$		$p = 200$		$p = 400$		$p = 800$	
	Estim	SE	Estim	SE	Estim	SE	Estim	SE	Estim	SE	Estim	SE
$\beta_1 = 0.0$	0.040	0.034	0.054	0.033	0.089	0.055	0.001	0.037	0.005	0.048	0.045	0.062
$\beta_2 = 0.2$	0.138	0.033	0.153	0.031	0.180	0.046	0.080	0.036	0.180	0.050	0.181	0.056
$\beta_3 = 0.5$	0.295	0.032	0.213	0.032	0.263	0.054	0.222	0.036	0.154	0.045	0.849	0.060
$\beta_4 = 0.8$	0.574	0.036	0.404	0.032	0.428	0.049	0.358	0.036	0.399	0.052	0.455	0.054
$\beta_5 = 0.0$	0.057	0.034	0.106	0.030	0.120	0.058	0.461	0.036	0.082	0.047	0.185	0.067
$\beta_6 = 1.0$	0.591	0.034	0.587	0.032	0.663	0.052	0.529	0.037	0.448	0.044	0.497	0.062
$\beta_7 = 0.5$	0.302	0.036	0.224	0.034	0.448	0.051	1.215	0.035	0.251	0.043	0.870	0.060
$\beta_8 = 0.2$	0.345	0.034	0.126	0.032	0.168	1.316	0.465	0.037	0.267	0.053	0.136	0.058
$\beta_9 = 0.0$	-0.011	0.031	0.097	0.030	0.239	0.053	0.145	0.037	0.173	0.050	0.171	0.059
$\beta_{10} = 0.3$	0.178	0.034	0.208	0.031	0.607	0.047	0.463	0.033	0.193	0.050	0.207	0.057

Parameter	$\mu = 0.8$						$\mu = 1.0$					
	$p = 200$		$p = 400$		$p = 800$		$p = 200$		$p = 400$		$p = 800$	
	Estim	SE	Estim	SE	Estim	SE	Estim	SE	Estim	SE	Estim	SE
$\beta_1 = 0.0$	-0.080	0.053	0.016	0.089	0.053	0.134	0.001	0.078	0.071	0.114	0.138	0.188
$\beta_2 = 0.2$	0.164	0.054	0.147	0.077	0.122	0.126	0.190	0.072	0.169	0.111	0.136	0.156
$\beta_3 = 0.5$	0.282	0.056	0.254	0.083	0.371	0.128	0.980	0.073	0.757	0.118	0.202	0.170
$\beta_3 = 0.8$	0.440	0.054	0.653	0.084	0.530	0.133	0.413	0.074	0.401	0.122	0.503	0.171
$\beta_5 = 0.0$	0.105	0.056	0.112	0.081	0.109	0.125	0.182	0.074	0.131	0.111	0.059	0.154
$\beta_6 = 1.0$	0.558	0.055	0.572	0.086	0.550	0.134	0.524	0.075	0.406	0.111	0.566	0.142
$\beta_7 = 0.5$	0.350	0.056	0.208	0.089	0.367	0.132	0.224	0.078	0.383	0.115	0.395	0.161
$\beta_8 = 0.2$	0.338	0.056	0.284	0.078	0.294	0.112	0.216	0.080	0.254	0.103	1.132	0.159
$\beta_9 = 0.0$	0.094	0.056	0.130	0.081	0.111	0.120	0.103	0.073	0.127	0.107	0.066	0.136
$\beta_{10} = 0.3$	0.231	0.056	0.506	0.084	0.427	0.108	0.218	0.070	0.160	0.117	0.605	0.142

Going through the analysis, we find that the model performs satisfactorily well under all the scenarios considered. Most noticeably, the posterior means are good point estimators of the parameter values, and the standard errors for virtually all coefficients are small, which is indicative of a good model fit. However, performance degrades in the higher dimensional setting (large number of predictors). We now turn to the impact of the tuning parameter μ and number of OTUs p , based on the standard errors (see plots of standard errors in Figure 4.4–4.5). If the SEs were the same under the different settings, then the plotted points in the

figure would all fall along the diagonal line. It can be seen that both μ and p had discernible influence on empirical performance, for the fixed sample size. For $\mu = 0$, the standard errors for the model when $p = 200$ and $p = 400$ are spread roughly along the diagonal, meaning that the standard errors are roughly similar. However, as p increases from 200 to 800, the standard errors become concentrated in the lower right of the diagonal, meaning the standard errors are higher for $p = 800$ than for $p = 200$. Similar trends are observed for $\mu = 0.2, 0.8$ and 1.0 ; standard errors are higher for higher values of p . This may be attributed to the fact that in high dimension (large p , as compared to n), and without regularization, there is substantially large variance that reduces the overall prediction accuracy of the model.

Meanwhile, for any fixed value of p , the standard errors increase with increasing value of μ (Figure 4.5). A possible explanation for this is that the phylogenetic tree is ‘noisy’ or uninformative - an issue that has been noted for phylogenetic tree usage in microbiome analysis (Xiao et al., 2018a) - and therefore as μ is increased, by which the phylogeny-defined similarity becomes more favoured than the radial basis function, the noise introduced in the data reduces model fit.

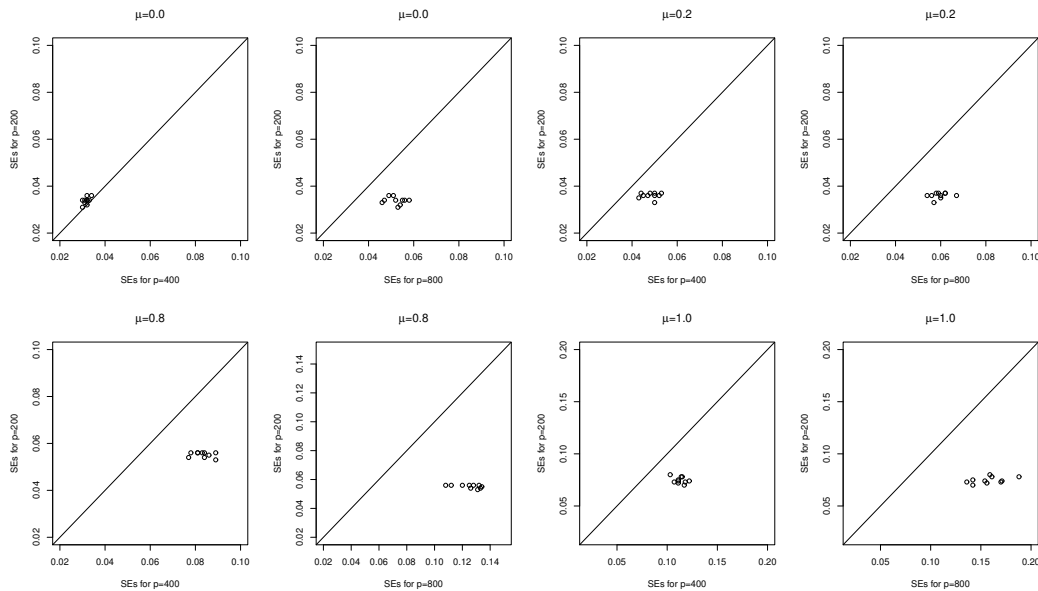


Figure 4.4: Comparison of standard errors (SEs) of the model for fixed μ and varying p .

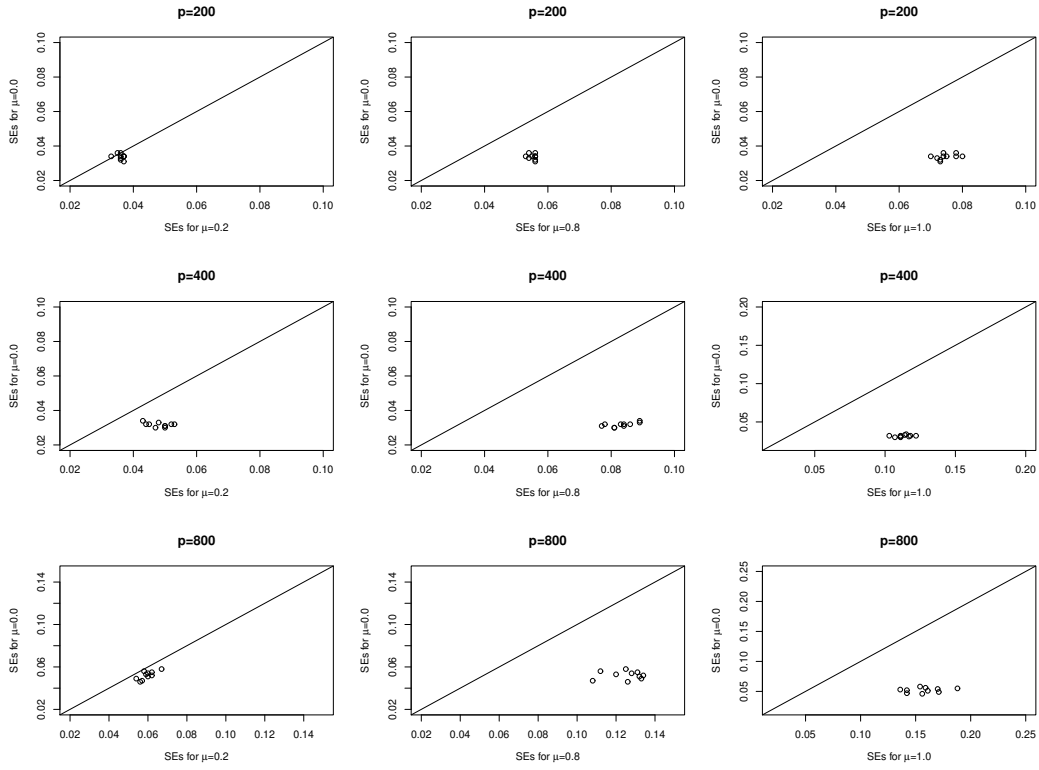


Figure 4.5: Comparison of standard errors (SEs) of the model for fixed p and varying μ .

4.5.2 Real data studies

To illustrate the applicability of the proposed method in empirical microbiome data analysis, we apply the model to three previously published real microbiome datasets of a schizophrenia, HIV/AIDS, and atherosclerosis. For these analyses, μ was fixed to 0.9 (favouring the phylogeny-based kernel more than the RBF-based kernel), and the Bayesian hyper-parameter settings maintained as in the simulation study.

4.5.2.1 Real data study 1: application to schizophrenia data

In this section, we apply the method to a real microbiome data in the study by [Nguyen et al. \(2019b\)](#), comparing the gut microbiome composition between persons with chronic schizophrenia and non-psychiatric individuals. The study collected stool samples and used 16S rRNA sequencing of the V4 gene region to compare, in a cross-sectional study, the microbiome composition of 25 chronic schizophrenic and 25 demographically matched non-psychiatric subjects. The dataset is available for download from Qiita (<https://qiita.ucsd.edu>) with study ID 11710. After profiling, the sequence reads were processed using QIIME2 ([Caporaso et al., 2010](#)); raw sequencing results demultiplexed and unique microbial OTUs identified using Deblur algorithm ([Amir et al., 2017](#)), and the resulting output feature table rarefied to 7,905 sequences per sample. A total of 3,564 OTUs were obtained. We applied our model to this dataset, with the goal of identifying bacterial taxa that are associated with the disease. The following pre-processing steps were performed on the data prior to analysis.

- (i) Sample filtering to remove outlier samples. One sample with sequencing depth < 1000 was removed.
- (ii) OTU filtering to remove less informative and noisy OTUs. Following [Xiao et al. \(2018b\)](#), we imposed two filters: (a) OTUs with overall prevalence $< 10\%$ across all samples, and (b) OTUs with median non-zero counts < 10 .
- (iii) Normalization to correct and reduce bias introduced by variable library sizes. This is a critical step before any meaningful statistical analyses can be performed on microbiome sequencing data ([Li et al., 2015](#)). We applied the recently proposed geometric mean of pairwise ratios (GMPR) ([Chen et al., 2018b](#)) normalization. Our choice for this normalization was due to its robustness for zero-inflated sequencing data.
- (iv) Winsorization to replace outlier observations. In this, any observation exceeding the α quantile of scaled counts, c_{ij}/s_i , for a given taxon is replaced with the α quantile of scaled counts for that taxon; where, s_i is the GMPR

scale factor for sample i , and herein α is chosen to be 97%. This procedure has demonstrated effectiveness in removing false positives (Chen et al., 2018a; Xiao et al., 2018a).

- (v) OTU data transformation to reduce the influence of highly abundant taxa. Here, the commonly used square-root transformation was applied.

Following this pre-process, a total of 1521 OTUs across 49 sample remained. The resulting OTU ('species' level) data were agglomerated to the class level, and analysis performed at this taxonomic rank. The corresponding phylogenetic tree was accordingly pruned to contain only these microbial taxa. Subsequent to this, the model was applied to analyse the data, with the model hyper-parameters set as in the simulation study. The Gibbs-slice sampling algorithm was run for 100,000 iterations. We observe from the posterior plots (Figure 4.7) that the chain moves quite quickly over the range of possible values of the parameters, moving directly to the center of the target posterior distribution. Convergence diagnostic using trace plots, cumulative quantile plots and auto-correlation function plots shows no evidence of non-convergence for each of the parameter components after 5,000 iterations. 100,000 posterior samples was drawn, burn-in period of 10,000 was considered, and the the remaining effective samples thinned by a factor of 5 before performing posterior inference. As a validation, we also analyse the same data using the standard generalized linear model, which essentially is a Bayesian probit model with highly diffuse priors (Jeff, 2007). As the standard generalized linear model accounts for neither sample clustering nor correlation among the predictors, our aim was not to view it as an alternative competing model, as this would certainly be biased, but rather to demonstrate the trade-offs and performance of schemes that incorporate phylogenetic information and/or account for samples clusterings with those that do not.

Results & discussion: schizophrenia data. The results are summarized in Table 4.2, showing the posterior means (i.e., the model coefficients) for the phyloDPM and the GLM. From it, we see that the samples exhibit substantial clus-

tering, as indicated by the precision parameter, m , of the DP and the number of clusters, $K(= 22)$, which is less than half the sample size $n(= 49)$. Comparing results from the two models, we observe that the sign of the estimated effects is consistent across the two models. Meanwhile, the magnitude of effects are smaller with the phyloDPM for some parameters and larger for others. This is interesting because the non-parametric part of the model in the phyloDPM is expected to better model heterogeneous sources of variability that tend to amplify ‘true’ effect sizes when parametric approaches are used to model such latent sources of variability. Given this observation, models with Dirichlet Processes (or non-parametric, in general) ought to be regarded as a robust modelling approach; providing a more conservative estimate of the coefficients rather than just correcting for overestimation of predictor effects.

Table 4.2: Summary of parameter estimates for the schizophrenia data: posterior mean for the coefficient for species effect (Coef), standard error (SE) and 95% credible interval (CI) for estimated parameters for OTU effects.

Parameter	Class	phyloDPM				GLM			
		Coef	SE	95% CI		Coef	SE	95% CI	
Intercept	-	0.062	0.016	-0.777	0.890	3.773	4.844	-5.721	13.268
β_1	<i>Clostridia</i>	-0.361	0.013	-0.797	0.040	-0.326	0.230	-0.777	0.126
β_2	<i>Bacilli</i>	-0.052	0.008	-0.036	-0.262	-0.086	0.133	-0.347	0.174
β_3	<i>Clostridia</i>	-0.005	0.007	-0.165	0.147	-0.011	0.055	-0.119	0.098
β_4	<i>Clostridia</i>	-0.013	0.003	-0.055	0.017	-0.017	0.0133	-0.043	0.009
β_5	<i>Bacteroidia</i>	0.071	0.010	-0.063	0.239	0.026	0.046	-0.065	0.117
β_6	<i>Actinobacteria</i>	0.063	0.004	-0.107	0.234	0.042	0.078	-0.112	0.196
β_7	<i>Coriobacteriia</i>	0.433	0.022	0.074	1.014	0.313	0.166	-0.013	0.640
β_8	<i>Bacteroidia</i>	-0.078	0.008	-0.259	0.083	-0.064	0.061	-0.184	0.055
β_9	<i>Deltaproteobacteria</i>	0.028	0.010	-0.358	0.327	0.019	0.178	-0.330	0.369
β_{10}	<i>Erysipelotrichi</i>	-0.030	0.007	-0.222	0.176	-0.034	0.074	-0.179	0.110
β_{11}	<i>Gammaproteobacteria</i>	0.033	0.001	-0.021	0.092	0.029	0.024	-0.019	0.076
β_{12}	<i>Bacteroidia</i>	0.019	0.002	-0.004	0.046	0.010	0.011	-0.012	0.032
β_{13}	<i>Clostridia</i>	-0.032	0.003	-0.088	0.017	-0.028	0.021	-0.068	0.013

To put these results into context, it is observed that the bacterial class *Coriobacteriia* and *Bacilli* tends to have a higher probability of association with the disease condition, with statistically significant effects at the 5% significance level. *Coriobacteriia* and *Bacilli* are Gram-positive class of bacteria within the *Actinobacteria* and *Firmicutes* phyla respectively. Previous studies have found both classes of bacteria to be enriched in subjects with schizophrenia as compared to their non-schizophrenic counterparts (Wu et al., 2016b; Nguyen et al., 2019b; Dickerson

et al., 2017; Huang et al., 2018). The results herein obtained using both models are therefore consistent with literature. By contrast, although there is substantive association of *Bacterioidia*, *Deltaproteobacteria*, *Erysipelotrichi*, and *Gammaproteobacteria*, the effects are not statistically significant at the 5% significance level.

Also noted, are the negative values of the coefficient estimates, which indicate that the *logit* components decrease with increase in the magnitude of these coefficients. Accordingly, this suggests that disease risk reduces in presence of *Clostridia*, *Bacilli*, and *Erysipelotrichi*, consistent with previous observation (Nguyen et al., 2019b). However, with the exception of *Bacilli*, the substantive results are not statistically significant.

4.5.2.2 Real data study 2: application to HIV/AIDS data

We now apply the model to a gut microbiome dataset for HIV/AIDS. The dataset is from the study by Neff et al. (2018), of the enteric microbiota composition in Human Immunodeficiency Virus (HIV) infection. The sample, with subjects all of whom recruited from Metropolitan Denver - USA, contains the microbiome profile from 42 healthy HIV-seronegative individuals and from 48 HIV-infected individuals, obtained by 16S rRNA targeted sequencing. The detailed protocol for sample collection, isolate preparation and DNA sequencing can be found in (Neff et al., 2018). The dataset can be downloaded from Qiita (<https://qiita.ucsd.edu>) with study ID 11669. Composition of the microbiota was determined by deep sequencing the PCR products from the V4 region of the 16S rDNA gene, using the Illumina MiSeq platform. 5002 taxa, belonging to several different genera, were detected for the species level data.

We applied the same pre-processing described in steps (i) through (iv) in the simulation study above. Following this, 69 abundant OTUs across 76 samples were retained for the subsequent analysis. After this, the underlying phylogenetic tree was pruned to contain only these remaining OTUs. We then used the model to detect associations between taxa and inflammation in individuals with

HIV. The analysis of the inflammatory properties of the enteric microbiota of HIV-infected individuals is of paramount interest because of existing evidence that bacterial translocation mediates chronic immune activation which has been shown to influence disease progression and persistence (Liovat et al., 2012; Eric et al., 2016; Ericson et al., 2016).

Results & discussion: HIV/AIDS data. Table 4.3 presents the Bayesian estimates of the parameters in the model. The Bayesian hyper-parameter setting were maintained as before, and, as before, 100,000 MCMC samples were drawn from the Bayesian posterior and the first 10,000 samples discarded as burn-in. The parameter estimates, using the resultant effective samples, are shown in Table 4.3.

As Table 4.3 shows, at the 95% credible level, the OTUs (parametrised by $\beta_1 \cdots \beta_{11}$, and $\beta_{52} \cdots \beta_{69}$) of the class *Bacteroidia*, *Clostridia*, *Gammaproteobacteria*, and *Erysipelotrichi* have profound, statistically significant, effects on inflammation in HIV-infected individuals, whereas the effects of the rest of the bacterial species are not statistically significant. These results are consistent with existing literature. As analyses of the gut microbiome continue to show that the enteric microbiome is compositionally distinct between HIV-positive and HIV-negative individuals, unifying patterns are emerging; similar to the pathologies observed in other diseases such as diabetes, obesity, inflammatory bowel disease, and *Clostridium difficile* infection, microbial community diversity is decreased in HIV-infected individuals. The enteric gut microbiome community in HIV-infected individuals has been reported to be enriched for taxa such as *Erysipelotrichaceae* and *Enterobacteriaceae*, known to have pathogenic potential, and depleted for taxa such as *Lactobacillaceae*, *Bacteroidaceae*, *Lachnospiraceae* and *Ruminococcaceae* that are reported to have anti-inflammatory properties (Ling et al., 2016; Dubourg et al., 2016; Gootenberg et al., 2017). The herein associated bacterial class: *Erysipelotrichi*, *Clostridia*, *Bacteroidia*, and *Gammaproteobacteria* have previously been observed in HIV infection (Neff et al., 2018; Pinto-Cardoso et al., 2017; Rocafort et al., 2019). *Erysipelotrichia* has been consistently reported to be increased in abundance and associated with HIV infection in multiple studies (Lozupone et al., 2013, 2014;

Dinh et al., 2015; Ling et al., 2016; Vujkovic-Cvijin et al., 2013).

Similarly, multiple lines of evidence points to *Clostridia* being among key phylotypes prevalent in HIV-1-infected patients (San-Juan-Vergara et al., 2018). Indeed, although there has been contrasting results with some studies, for example (Gootenberg et al., 2017; Sun et al., 2016), reporting decreased abundance while others, for example (Dein Terra Mota Ribeiro et al., 2017; Nowak et al., 2015; Dubourg et al., 2016), finding increased abundance of several bacteria within the *Clostridia* class, it has increasingly become apparent that *Clostridia* is strongly associated with HIV infection. Interestingly, this same community of bacteria has also been positively correlated with gut impairment markers as well as proinflammatory states and disease progression in HIV-infected patients (Dubourg et al., 2016).

Although it remains unclear whether compositional changes of the gut microbiota alone or other factors such as increased intestinal permeability, alteration of the function of the gut epithelial barrier towards pathogens or the joint effect of all of them is responsible for bacterial translocation and peripheral immune activation in HIV infection, the results of our analysis in concert with those of previous studies would suggest that bacteria within the *Clostridiaceae* and *Erysipelotrichaceae* families play an active role in driving gut and systemic immune activation, associated with inflammatory signatures in HIV-infected individuals.

4.5.2.3 Real data study 3: application to atherosclerosis data

Atherosclerosis, which is characterized by accumulation of cholesterol and recruitment of macrophages into artery walls, is the major risk factor for cardiovascular disease. While the influence of host genetics and other factors such as nutrition have been acknowledged and appreciated, a number of clinical and animal studies during the past decade have provided a substantial amount of association-based evidence indicating that gut dysbiosis can also contribute to its development and progression. As a final illustration of the performance of the proposed model,

we further analyse atherosclerosis data with 4623 OTUs from 73 samples. The dataset available from Qiita (<https://qiita.ucsd.edu>) with study ID 349, is from the study by Koren et al. (2011). The detailed information regarding sample preparation and DNA sequencing can be found in the author's article. We applied the same data-pre-processing steps as described above. Following the pre-processing procedure, a total of 652 OTUs remained across 72 samples. The underlying phylogenetic tree was also subsequently pruned to these remaining OTUs. The Bayesian hyper-parameter setting and MCMC sampling runs was maintained as in the applications earlier described.

Results & discussion: atherosclerosis data. Looking at the results presented in Table 4.4, there is an apparent interaction between the gut microbiome composition and the phenotypic condition. Specifically, the top associations (largest magnitude of the effect sizes) are predominated by bacterial OTUs belonging to the class *Clostridia*, *Erysipelotrichi*, and *Bacilli* with statistically significant effects, with the exception of *Erysipelotrichi*. The substantive effects for all OTUs belonging to the class *Erysipelotrichi* are not statistically significant. Also, of note, the coefficient estimates for all the significantly associated OTU effects are positive, with the implication that the *logit* components increase with increase in magnitude of the coefficients. The substantive results thus suggests that the abundance of these statistically significant taxa increases the risk for development and/or progression of atherosclerosis.

A close look at the biological significance of the associations identified by the model affirms their relevance. Existing body of evidence suggests commensal gut microbiota can contribute to atherosclerotic either directly by orchestrating and/or facilitating progression of atherosclerotic plaques, or indirectly by impacting its risk factors (Anthony and Komaroff, 2018; Jonsson and Bäckhed, 2017). Indeed, unique microbiome signatures associated with phenotypes such as acute coronary syndrome, obesity and type 2 diabetes, which are known to be risk factors for atherosclerosis, have been observed in patients with symptomatic atherosclerosis (Liu et al., 2019b; Anthony and Komaroff, 2018; Emoto et al., 2017; Pan et al., 2017).

The strong associations here detected, have been affirmed in atherosclerosis and its risk factors. Thus, the associations identified by the model replicates those reported in previous literature, at least at the class level; lending support to the robustness of the estimation and the biological relevance of the associations.

Table 4.3: Summary of parameter estimates for the HIV/AIDS data: posterior mean for the coefficient for species effect (Coef), standard error (SE) and 95% credible interval (CI) for estimated taxa effects. Also shown are the corresponding classes for the species. The (–) on the species column indicates the species information is unavailable.

Parameter	Specie	Class	Coef	SE	95% CI (lo)	95% CI (up)	Parameter	Specie	Class	Coef	SE	95% CI (lo)	95% CI (up)
β_1	–	<i>Bacteroidia</i>	0.077	0.030	0.057	0.097	β_{36}	–	<i>Clostridia</i>	-0.001	0.033	-0.023	0.021
β_2	–	<i>Erysipelotrichi</i>	0.063	0.032	0.041	0.085	β_{37}	–	<i>Clostridia</i>	-0.002	0.033	-0.024	0.020
β_3	–	<i>Clostridia</i>	0.060	0.034	0.037	0.083	β_{38}	–	<i>Clostridia</i>	-0.007	0.033	-0.029	0.015
β_4	–	<i>Clostridia</i>	0.060	0.032	0.038	0.082	β_{39}	–	<i>Clostridia</i>	-0.008	0.031	-0.029	0.013
β_5	–	<i>Clostridia</i>	0.048	0.030	0.028	0.068	β_{40}	–	<i>Clostridia</i>	-0.008	0.033	-0.030	0.014
β_6	–	<i>Clostridia</i>	0.040	0.032	0.018	0.062	β_{41}	<i>biforme</i>	<i>Erysipelotrichi</i>	-0.008	0.031	-0.029	0.013
β_7	–	<i>Clostridia</i>	0.039	0.033	0.017	0.061	β_{42}	–	<i>Clostridia</i>	-0.008	0.033	-0.030	0.014
β_8	–	<i>Clostridia</i>	0.032	0.032	0.010	0.054	β_{43}	–	<i>Clostridia</i>	-0.011	0.032	-0.033	0.011
β_9	–	<i>Clostridia</i>	0.030	0.034	0.007	0.053	β_{44}	–	<i>Bacteroidia</i>	-0.011	0.034	-0.034	0.012
β_{10}	–	<i>Clostridia</i>	0.029	0.032	0.007	0.051	β_{45}	<i>ovatus</i>	<i>Bacteroidia</i>	-0.011	0.031	-0.032	0.010
β_{11}	–	<i>Clostridia</i>	0.023	0.032	0.001	0.045	β_{46}	<i>prausnitzii</i>	<i>Bacteroidia</i>	-0.012	0.033	-0.034	0.010
β_{12}	–	<i>Clostridia</i>	0.019	0.032	-0.003	0.041	β_{47}	–	<i>Clostridia</i>	-0.016	0.033	-0.038	0.006
β_{13}	–	<i>Clostridia</i>	0.018	0.032	-0.004	0.040	β_{48}	<i>prausnitzii</i>	<i>Clostridia</i>	-0.018	0.031	-0.039	0.003
β_{14}	–	<i>Clostridia</i>	0.017	0.030	-0.003	0.037	β_{49}	–	<i>Bacilli</i>	-0.020	0.031	-0.041	0.001
β_{15}	–	<i>Clostridia</i>	0.017	0.031	-0.004	0.038	β_{50}	–	<i>Erysipelotrichi</i>	-0.021	0.032	-0.043	0.001
β_{16}	–	<i>Clostridia</i>	0.014	0.031	-0.007	0.035	β_{51}	–	<i>Clostridia</i>	-0.021	0.034	-0.044	0.002
β_{17}	–	<i>Clostridia</i>	0.014	0.033	-0.008	0.036	β_{52}	–	<i>Bacteroidia</i>	-0.023	0.032	-0.045	-0.001
β_{18}	–	<i>Clostridia</i>	0.013	0.031	-0.008	0.034	β_{53}	–	<i>Clostridia</i>	-0.025	0.033	-0.047	-0.003
β_{19}	–	<i>Clostridia</i>	0.013	0.031	-0.008	0.034	β_{54}	–	<i>Clostridia</i>	-0.027	0.031	-0.048	-0.006
β_{20}	<i>copri</i>	<i>Bacteroidia</i>	0.010	0.033	-0.012	0.032	β_{55}	–	<i>Clostridia</i>	-0.027	0.033	-0.049	-0.005
β_{21}	–	<i>Clostridia</i>	0.010	0.032	-0.012	0.032	β_{56}	–	<i>Clostridia</i>	-0.031	0.032	-0.053	-0.009
β_{22}	–	<i>Clostridia</i>	0.009	0.032	-0.013	0.031	β_{57}	–	<i>Clostridia</i>	-0.031	0.030	-0.051	-0.011
β_{23}	<i>aerofaciens</i>	<i>Coriobacteriia</i>	0.008	0.033	-0.014	0.030	β_{58}	–	<i>Clostridia</i>	-0.031	0.032	-0.053	-0.009
β_{24}	<i>prausnitzii</i>	<i>Clostridia</i>	0.008	0.031	-0.013	0.029	β_{59}	–	<i>Bacteroidia</i>	-0.033	0.031	-0.054	-0.012
β_{25}	–	<i>Clostridia</i>	0.007	0.031	-0.014	0.028	β_{60}	<i>prausnitzii</i>	<i>Clostridia</i>	-0.033	0.031	-0.054	-0.012
β_{26}	–	<i>Bacteroidia</i>	0.007	0.033	-0.015	0.029	β_{61}	–	<i>Clostridia</i>	-0.040	0.032	-0.062	-0.018
β_{27}	<i>prausnitzii</i>	<i>Clostridia</i>	0.006	0.030	-0.014	0.026	β_{62}	–	<i>Clostridia</i>	-0.042	0.032	-0.064	-0.020
β_{28}	–	<i>Clostridia</i>	0.006	0.033	-0.016	0.028	β_{63}	–	<i>Gammaproteobacteria</i>	-0.043	0.033	-0.065	-0.021
β_{29}	–	<i>Clostridia</i>	0.006	0.032	-0.016	0.028	β_{64}	–	<i>Clostridia</i>	-0.043	0.031	-0.064	-0.022
β_{30}	–	<i>Clostridia</i>	0.006	0.032	-0.016	0.028	β_{65}	–	<i>Bacteroidia</i>	-0.045	0.032	-0.067	-0.023
β_{31}	–	<i>Clostridia</i>	0.004	0.031	-0.017	0.025	β_{66}	<i>prausnitzii</i>	<i>Clostridia</i>	-0.049	0.032	-0.071	-0.027
β_{32}	–	<i>Clostridia</i>	0.003	0.032	-0.019	0.025	β_{67}	–	<i>Clostridia</i>	-0.059	0.031	-0.080	-0.038
β_{33}	<i>prausnitzii</i>	<i>Clostridia</i>	0.002	0.033	-0.020	0.024	β_{68}	–	<i>Clostridia</i>	-0.065	0.035	-0.089	-0.041
β_{34}	–	<i>Clostridia</i>	0.001	0.032	-0.021	0.023	β_{69}	–	<i>Clostridia</i>	-0.099	0.032	-0.121	-0.077
β_{35}	–	<i>Clostridia</i>	0.000	0.032	-0.022	0.022							

Variable	Species	Class	Coeff	SE	95% CI (Lo)	95% CI (Up)	Variable	Species	Class	Coeff	SE	95% CI (Lo)	95% CI (Up)
1076608	–	Bacilli	0.017	0.034	-0.006	0.040	192523	–	Chloridia	0.011	0.032	-0.011	0.033
124093	–	Bacilli	-0.018	0.001	-0.019	0.017	437901	–	Bacilli	0.009	0.003	-0.002	0.012
1736666	–	Chloridia	-0.017	0.005	-0.041	0.007	437941	–	Actinobacteria	0.012	0.002	-0.001	0.015
17786	–	Chloridia	-0.014	0.003	-0.036	0.008	20958	–	Chloridia	-0.011	0.005	-0.035	0.019
183804	–	Chloridia	-0.016	0.002	-0.038	0.006	448139	–	Gammaglutathione	0.007	0.004	-0.006	0.017
187712	–	Chloridia	-0.015	0.005	-0.037	0.007	193987	–	Chloridia	-0.010	0.002	-0.032	0.020
187894	–	Chloridia	-0.012	0.003	-0.034	0.010	193987	–	Chloridia	0.009	0.004	-0.014	0.025
191355	–	Erythrobacter	0.014	0.002	-0.008	0.016	193601	–	Actinobacteria	0.005	0.003	-0.002	0.012
192210	–	Chloridia	0.014	0.006	-0.010	0.038	192996	–	Chloridia	-0.008	0.003	-0.020	0.024
192284	–	Chloridia	0.013	0.002	-0.009	0.035	193008	–	Chloridia	-0.008	0.002	-0.021	0.022
192746	–	Chloridia	-0.017	0.001	-0.038	0.004	4439821	–	Bacteroidia	-0.011	0.003	-0.033	0.018
193208	–	Chloridia	-0.013	0.001	-0.034	0.008	4357713	–	Chloridia	0.006	0.003	-0.016	0.021
193551	–	Chloridia	0.012	0.004	-0.011	0.035	1939635	–	Bacilli	0.006	0.003	-0.008	0.021
193667	–	Chloridia	0.015	0.005	-0.009	0.039	210895	–	Chloridia	0.009	0.003	-0.013	0.019
194035	–	Chloridia	-0.014	0.004	-0.037	0.009	157804	–	Chloridia	0.007	0.001	-0.014	0.024
195719	–	Chloridia	-0.017	0.001	-0.038	0.008	4407586	–	Bacilli	-0.006	0.003	-0.015	0.021
196004	–	Chloridia	-0.017	0.004	-0.040	0.006	192240	–	Chloridia	0.012	0.004	-0.011	0.018
196087	–	Chloridia	-0.015	0.004	-0.038	0.008	300222	–	Chloridia	0.009	0.002	-0.013	0.027
196241	–	Chloridia	-0.013	0.001	-0.036	0.008	193801	–	Chloridia	0.006	0.002	-0.010	0.024
196242	–	Chloridia	-0.013	0.001	-0.036	0.008	193801	–	Chloridia	0.006	0.002	-0.010	0.024
197361	–	Chloridia	-0.012	0.004	-0.035	0.011	193594	–	Chloridia	-0.006	0.002	-0.028	0.023
200910	–	Chloridia	-0.012	0.004	-0.035	0.011	193594	–	Chloridia	-0.006	0.002	-0.028	0.023
212785	–	Chloridia	-0.015	0.000	-0.035	0.003	505984	–	Chloridia	-0.008	0.003	-0.032	0.024
2221284	–	Chloridia	-0.014	0.004	-0.037	0.009	194077	–	Chloridia	-0.008	0.003	-0.032	0.024
230109	–	Chloridia	-0.016	0.005	-0.040	0.008	193805	–	Chloridia	0.010	0.003	-0.012	0.024
244248	–	Chloridia	-0.014	0.003	-0.036	0.008	193919	–	Bacilli	-0.008	0.004	-0.031	0.025
246783	–	Bacteroidia	0.015	0.002	-0.007	0.037	362526	–	Chloridia	-0.009	0.003	-0.033	0.025
257492	–	Bacilli	0.015	0.001	-0.006	0.036	50611	–	Bacilli	0.009	0.002	-0.013	0.025
274257	–	Chloridia	-0.014	0.005	-0.038	0.010	4438826	–	Chloridia	-0.009	0.003	-0.033	0.025
287592	–	Chloridia	-0.015	0.004	-0.038	0.008	439319	–	Bacteroidia	-0.009	0.003	-0.033	0.025
290875	–	Bacilli	-0.017	0.003	-0.039	0.005	293389	–	Chloridia	-0.011	0.004	-0.034	0.027
299439	–	Bacteroidia	0.016	0.003	-0.006	0.038	196410	–	Chloridia	-0.009	0.003	-0.031	0.027
312866	–	Chloridia	-0.017	0.003	-0.039	0.005	441938	–	Chloridia	-0.010	0.003	-0.032	0.027
329703	–	Chloridia	-0.017	0.002	-0.039	0.003	434749	–	Actinobacteria	0.011	0.002	-0.011	0.027
338992	–	Chloridia	0.016	0.004	-0.007	0.039	52391	–	Chloridia	0.010	0.004	-0.013	0.027
353292	–	Chloridia	0.013	0.003	-0.009	0.040	447605	–	Chloridia	-0.010	0.003	-0.032	0.027
365336	–	Chloridia	0.012	0.002	-0.010	0.014	192646	–	Chloridia	0.006	0.004	-0.017	0.027
380759	–	Bacilli	0.017	0.002	-0.007	0.049	193790	–	Chloridia	-0.011	0.004	-0.034	0.027
4323100	–	Bacilli	-0.015	0.002	-0.037	0.007	1106323	–	Chloridia	-0.009	0.002	-0.031	0.027
432717	–	Bacilli	0.015	0.004	-0.008	0.038	460774	–	Actinobacteria	-0.009	0.002	-0.031	0.027
4340587	–	Bacteroidia	-0.018	0.001	-0.039	0.003	448755	–	Bacteroidia	0.012	0.004	-0.011	0.027
436272	–	Bacilli	-0.012	0.003	-0.034	0.010	294606	–	Chloridia	-0.005	0.003	-0.027	0.027
4362857	–	Chloridia	0.016	0.002	-0.006	0.038	192217	–	Chloridia	-0.010	0.000	-0.030	0.027
4396516	–	Chloridia	-0.017	0.002	-0.039	0.005	226639	–	Actinobacteria	-0.010	0.001	-0.031	0.027
4397680	–	Bacteroidia	0.017	0.003	-0.005	0.039	294617	–	Bacteroidia	0.011	0.002	-0.011	0.027
4427617	–	Gammaglutathione	-0.018	0.001	-0.039	0.003	448333	–	Chloridia	-0.012	0.003	-0.034	0.027
4447219	–	Actinobacteria	-0.014	0.004	-0.037	0.009	422295	–	Chloridia	-0.008	0.002	-0.030	0.027
448731	–	Bacteroidia	-0.017	0.003	-0.039	0.005	194112	–	Chloridia	0.011	0.003	-0.011	0.027
448550	–	Bacteroidia	0.015	0.002	-0.007	0.037	870594	–	Bacilli	-0.009	0.005	-0.033	0.027
446139	–	Gammaglutathione	-0.014	0.002	-0.036	0.008	190701	–	Chloridia	-0.007	0.003	-0.029	0.027
4478641	–	Bacteroidia	-0.015	0.003	-0.037	0.007	223957	–	Bacteroidia	-0.008	0.004	-0.031	0.027
4483174	–	Bacteroidia	0.012	0.002	-0.010	0.034	979991	–	Gammaglutathione	0.006	0.003	-0.016	0.027
4484389	–	Bacteroidia	0.016	0.005	-0.008	0.040	342427	–	Chloridia	-0.010	0.001	-0.031	0.027
512239	–	Bacilli	0.013	0.003	-0.009	0.035	446970	–	Chloridia	0.006	0.001	-0.015	0.027
551481	–	Actinobacteria	-0.014	0.004	-0.037	0.009	446973	–	Bacteroidia	-0.008	0.003	-0.030	0.027
565082	–	Chloridia	0.014	0.002	-0.008	0.036	434479	–	Actinobacteria	-0.007	0.004	-0.030	0.027
585585	–	Erythrobacter	-0.013	0.003	-0.036	0.010	4478125	–	Chloridia	0.006	0.005	-0.014	0.027
932653	–	Bacilli	0.017	0.003	-0.005	0.039	192225	–	Chloridia	-0.010	0.001	-0.031	0.027
965048	–	Bacteroidia	-0.014	0.009	-0.034	0.006	192930	–	Chloridia	0.007	0.002	-0.015	0.027
970796	–	Gammaglutathione	-0.018	0.004	-0.039	0.007	210867	–	Chloridia	0.010	0.006	-0.014	0.027

4.6 Concluding remarks

Microbiome association studies, prodded by the advent and development of high-throughput sequencing technologies, offer unparalleled opportunities to elucidate the underlying mechanisms of complex human traits and diseases. In modelling the link between microbiome taxonomic features and the host's phenotypic status, considerations ought to be made of the underlying dataset. A statistical model that incorporates taxa similarity and models sample heterogeneity will fit the data better and provide more accurate measure of taxa effects. Here, we formulated a Bayesian statistical framework, phyloDPM, for analysis of 16S rRNA microbiome data. Unlike many existing methods that associates overall microbiome profile with the host's phenotype, phyloDPM allows estimation of OTU-specific effects. Besides accounting for OTU clustering at the various phylogenetic depths, the method, importantly, models multi-collinearity among taxa as well as heterogeneity across samples. Moreover, it is robust to the possibility that the phylogenetic tree may be mis-specified or uninformative with respect to OTU effects on the host's phenotypic state. Jointly modelling taxa similarities as well as latent heterogeneity facilitates identification of truly phenotype-associated OTUs and their effects. Furthermore, phyloDPM employs the slice-Gibbs sampling, known to greatly improve mixing of MCMC samples [Lu et al. \(2018\)](#), which results in statistically reliable estimates with relatively short runs of the chain.

By simulation studies, we demonstrated that phyloDPM has satisfactory performance. We also showed that while performance diminished as the number of OTUs in the analysis increased, the loss in power was not substantially appreciable, as indicated by the corresponding standard errors. In the real microbiome data studies of schizophrenia, HIV, and atherosclerosis, we further confirmed the merits of the method in capturing association signals. Of note, we identify bacteria of *Actinobacteria* and *Firmicutes* phyla to be strongly associated with schizophrenia; *Clostridia* and *Erysipelotrichi* class with inflammation in HIV-infected individuals; and *Clostridia* and *Bacilli* class with atherosclerosis. Importantly, the standard

errors and the 95% credible intervals of the parameters are small for virtually all the coefficients, showing that the method maintains its satisfactory performance. Thus, we believe the model and method is useful for understanding the underlying link between microbial taxa from 16S rDNA sequencing and complex human diseases. Since the majority of microbiome studies use the 16S rDNA sequencing platform and most bioinformatics pipelines generate operational taxonomic units (OTUs), the proposed approach can be readily used.

Some limitations and assumptions of the phyloDPM method require mention. First, it makes a key assumption that OTU effects is additive and that higher order interaction effects is negligible, which permits modelling the joint effect of OTUs as a linear combination of individual OTU effects using only main effects. Although often sufficient in practice, we expect that the inclusion of a few higher order terms could prove useful by way of improving precision on parameter estimates. Second, is the relatively high computational cost. As sample size and the number of OTUs in the analysis increases, the number of CPU time also proportionately increases, as the algorithm makes component-wise conditional update on each parameter. In our analysis with 652 OTUs for 72 samples, for example, the sampling algorithm ~ 3.2 hrs to complete 100,000 iterations. One way to address this could be to apply a more stringent pre-processing procedure as a way of reducing dimensionality, so that many of the low-abundant OTUs, which can be regarded as being less biologically informative and noisy, are filtered out. This must, however, be done with care; otherwise rare but phenotype-informative taxa may be erroneously discarded. Alternatively, the simulation-based Approximate Bayesian Computation (ABC) methodology can be pursued as a way to circumvent computational cost. Third, besides computational cost, we point out that the auxiliary variables sampling on which the model estimation is based has high auto-correlation and slow convergence. Finally, the model is primarily developed for case-control phenotypes. Nevertheless, extension to quantitative phenotype can be made in a straight-forward way. The only modification would be replacing the logistic model structure with its linear counterpart. Gratifyingly, in this case, the conditional distributions of the

parameters assume standard distributional forms that can be easily sampled from, and hence Gibbs sampling, with a Metropolis step for the precision parameter of the Dirichlet process component, can be directly applied.

In conclusion, the proposed framework represents a robust way to model host's OTU-phenotype relationship, offering the opportunity to provide insight into the microbiome-basis of human diseases. We were able to provide a list of plausible candidate bacterial taxa that are intimately involved in the pathogenesis of schizophrenia, HIV, and atherosclerosis. However, further validation and identification of the causal bacterial taxa is necessary.

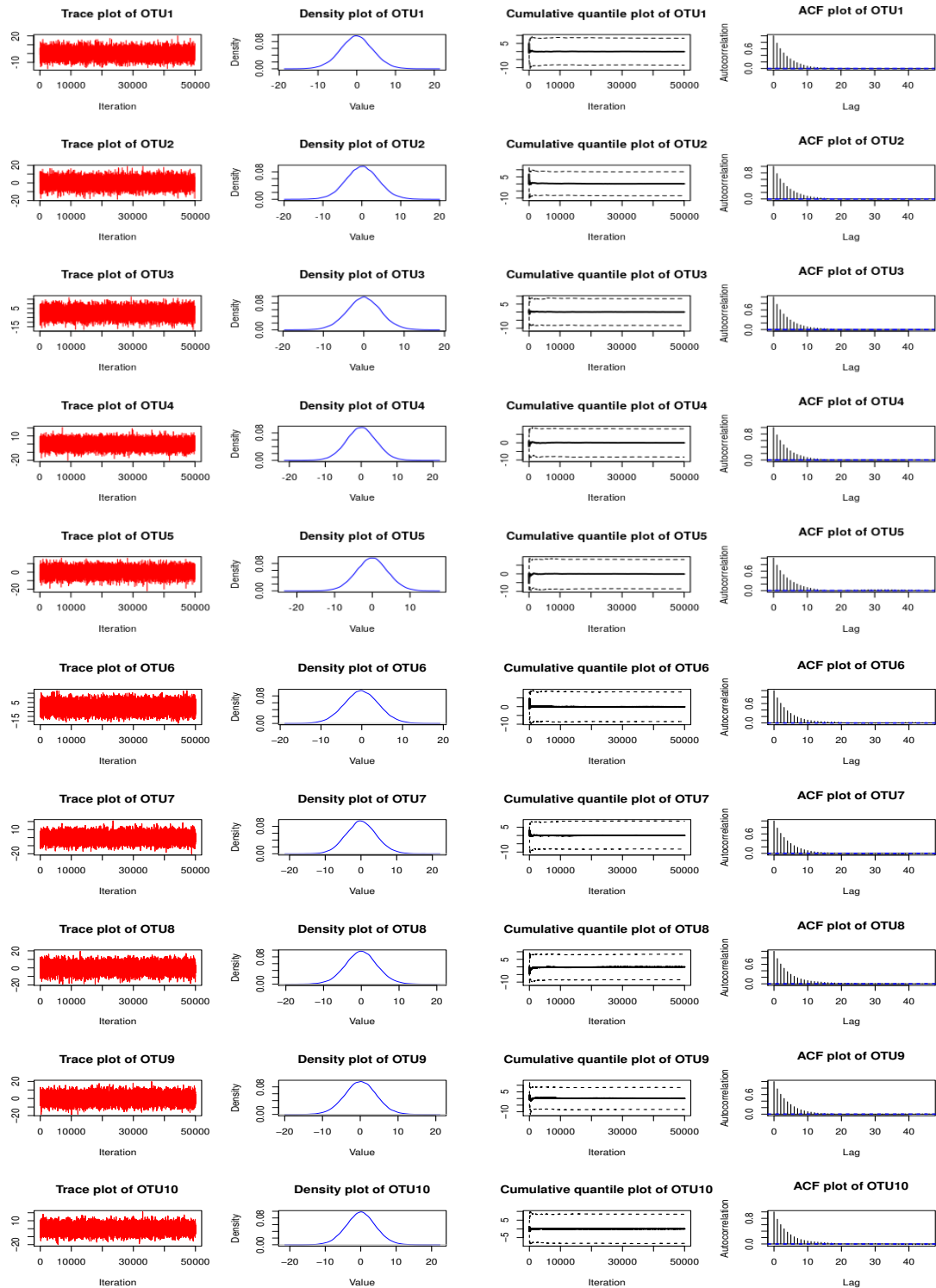


Figure 4.6: Plots to assess convergence of the posterior distribution for parameters of the model. Plots are from posterior draws of the model applied to the simulated OTU dataset: trace plots, estimated posterior densities, cumulative quantile plot and sample auto-correlation function plot for a representative sampling chain for the effects of the first 10 OTUs.

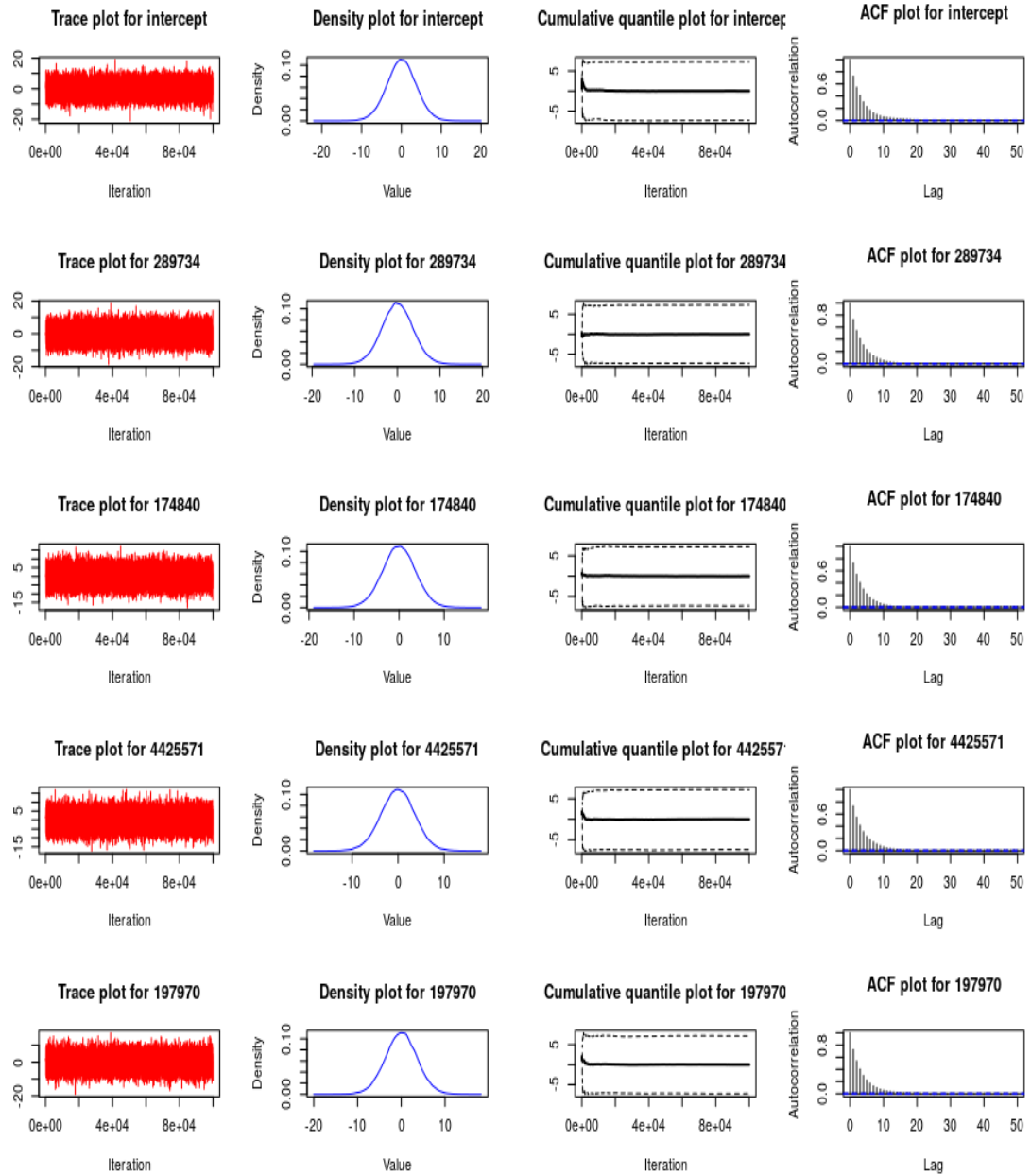


Figure 4.7: Plots to assess convergence of the posterior distribution for parameters of the model. Trace plots, estimated posterior densities, cumulative quantile plot and sample auto-correlation function plot for the first 4 species, using schizophrenia data.

Conclusions and Perspectives

"The more we know, the better we realize that our knowledge is a little island in the midst of an ocean of ignorance".

- Theodosius Dobzhansky, *The Great Ideas Today*

Genetic research over the past decade has provided fundamental insights into the biological basis of various human complex diseases and traits; identifying several loci and biological pathways through which they operate in trait variation and disease pathogenesis. These discoveries have fuelled widespread expectation that genome-wide screening of disease-predisposing genetic variations will revolutionize clinical and medical practice, from disease risk prediction & prevention to drug development and 'personalization' of medicine to maximize clinical efficacy and minimize toxicity. Indeed, while no appreciable clinical utility have yet been realized from GWAS findings so far, accumulating body of knowledge from GWAS regarding functionally-important variants, pathways and potential mechanisms is providing an optimistic outlook for this pursuit.

Equally, with these successful discoveries and the gratifying promise comes the sobering realization that the genetic basis of most human phenotypes is even more complex than had earlier been thought. A large majority of human diseases are multi-factorial, in that they result from the influence of multiple host DNA and non host DNA factors. While the complexity of the problem maybe disappointing, the biological reality has to be confronted if clinical benefits are to be reaped. Unlike environmental exposure whose role on host phenotypes has long been appreciated, it has only recently been established that the contribution of the

microbiome, a non host DNA factor, to its host phenotypic plasticity is not only considerable but is also as important, if not more important at least for some phenotypes, as the host's genetic variation. Accordingly, there is clearly enormous potential in leveraging microbiome information into host's genome-wide association studies for a better understanding of the molecular mechanisms, processes and pathways involved in disease pathogenesis and variability in phenotypes across individuals. The success of this new paradigm of association mapping will lead to functional characterization of the basis of complex human diseases and elucidate the underlying functional mechanisms, making precision medicine a forthcoming possibility.

Through the studies presented in this thesis we explored the role of the human microbiome and genotype on its host's phenotype, and the interaction of host host genetics with the microbiome. We further explored existing statistical frameworks for host and microbiome association mapping and proposed a statistical framework for studying association between microbial OTUs and the host's phenotype. To begin with, we presented, in Chapter 2, human genetic variation and its role on human complex traits/diseases as established through GWAS, and discussed the statistical methodologies for GWAS. We then presented the knowledge-so-far gained from GWAS of common human complex traits/diseases and discussed the implications of these findings for disease scoring statistics. In Chapter 3, we took a detailed study of the link between the human microbiome and human traits and diseases, and the interaction of human genetic polymorphisms with the microbiome. Following this, we dissected genome-wide studies for microbiome-related metabolic diseases. In Chapter 4, we presented the proposed statistical framework, based on the Dirichlet process random effects model, for studying the association between microbial OTUs and their host's phenotype.

This chapter concludes the thesis by first summarizing the major contributions from the work herein, and then discussing a perspective on how and why extending the human genome-wide association repertoire to include the inherent microbiome information may enhance our understanding of the spectrum and relative strengths of the causative factors for human phenotypic variation. It also

points out some limitations of the model presented in the thesis, and provides suggestions for potential directions for future work.

5.1 Summary of contributions

To recap, the contributions of this thesis can be summarized as follows:

1. A detailed study of the human microbiome and its interaction with human genetic polymorphisms to further investigations into the genetic and microbiome basis of complex human traits and diseases.
2. A comprehensive review of disease scoring models for genome-wide and microbiome-wide human trait/disease mapping, highlighting the strengths and limitations of contemporary disease scoring models, and proposing methodological approaches to address some of the limitations to improve association inference.
3. A dissection of the genetics of microbiome-related metabolic diseases and the leveraging of whole genome sequence (WGS) samples from world-wide ethnic groups to examine genetic diversity in variants linked to metabolic diseases, demonstrating the marked differences in gene-specific proportion of pathogenic SNPs and gene-specific SNPs minor allele frequencies across ethnic groups.
4. The development of a statistical framework for studying association between microbial OTUs and the host's phenotype. The approach leverages the phylogenetic tree structure and the radial basis function in modelling microbial OTU effects, and uses a non-parametrically specified random variable to model heterogeneity across samples. The resulting model (Dirichlet process random effects model) offering considerable promise in addressing multicollinearity, and unobserved heterogeneity and over-dispersion in microbiome data.

5.2 Perspectives

5.2.1 Insights into the missing heritability mystery

Finding the missing heritability of complex human phenotypes has become an important challenge for GWAS, given its importance in delineating the upper limit to which phenotypic variation can be attributed to genetic variation (van Rheenen et al., 2019). Notwithstanding the polemics about hypotheses on the missing heritability problem, it is clear that the assumptions of GWAS represents a limitation crucial for heritability inference. As pointed out in the preceding section, GWAS assumes homogeneity in ‘environmental’ (non-genetic) exposure of the study subjects; an assumption which can be violated leading to discrepancy between heritability estimates from different studies (Sandoval-Motta et al., 2017b; Bourrat et al., 2017). In classical twin studies, it is assumed that the environments experienced by the twins do not differ in any way that may be relevant for the development of the phenotype, and that host genotype and environment are mutually independent with respect to the phenotype (Henriksen et al., 2017). In practice, however, there is a non-shared environment that contributes to phenotypic difference. For instance, microbiome profile, that can be distinct between any two individuals, can have differential influence on host phenotype. The result of this is that phenotypic resemblance may occur between genetically-distinct individuals, and genetically-similar individuals may have markedly different phenotypic expressions. Consequently, this can result in overestimation of heritability from twin-based studies (López-Cortegano and Caballero, 2019; Sandoval-Motta et al., 2017b; Zuk et al., 2012).

The relevance of this observation to host GWAS is that we might have been making a ‘black hole’ pursuit; perhaps the upper limit of the narrow-sense heritability attributable to host genetics is not as high as had been posited by twin-based studies. Along this same line, it is perhaps also the case that, at least for some complex phenotypes, host genotype by itself is not as important in the trait as we

had thought; but instead, other components such as the human microbiome, which can also be vertically transmitted (at least partially), might be the major player. Given this, coupled with the fact that GWAS only considers the effect of genetic variation in human cells and do not take into account the effect of the microbiome on the phenotype, we hypothesize that incorporating the compositional and functional diversity of the microbiome into host genome-wide association studies will provide novel insights into missing heritability mystery. This perspective is discussed in our previous article, “*Heritability jointly explained by host genetics and microbiome: will improve traits prediction?*”.

In order to integrate host genetics and the microbiome in the same analytical model, a community view has to be taken in which host genetics and the microbiome are viewed as a single entity. To this end, we assume the phenotype (y) is shaped by host genetics (g), the microbiome (b), and the environment (e) so that the general model becomes

$$y = g + b + e , \quad (5.1)$$

where $g = Z\beta$, $b = W\alpha$, $g_i \sim N(0, \sigma_g^2)$, and $b_i \sim N(0, \sigma_b^2)$. Z is a $n \times p$ matrix of host genotypes, and W is a $n \times q$ matrix of microbiome data; n being the sample size, p the number of SNPs, and q the number of microbial taxa.

It is generally not unreasonable, for purposes of estimating heritability, to concentrate on the main effects part of the model, partly for simplicity of exposition as it is very difficult to identify interactions terms with a reasonable accuracy in such high dimensional setting (Mäki-Tanila and Hill, 2014; Huang and Mackay, 2016; Rothschild et al., 2018) and partly because the first order of Taylor expansion of the model function can accurately approximate interaction effects, which are essentially encoded in lower order terms (Rothschild et al., 2018). However, if the true covariance between host genetics and the microbiome is appreciable, omitting it from the model can lead to over-estimation of the underlying heritability.

The phenotypic variance-covariance matrix is expressed as

$$\begin{aligned}\text{Cov}(y) &= \frac{ZZ'\sigma_g^2}{m} + \frac{WW'\sigma_b^2}{l} + C_r\sigma_r^2 + I\sigma_e^2 \\ &= A\sigma_g^2 + B\sigma_b^2 + C_r\sigma_r^2 + I\sigma_e^2,\end{aligned}$$

where A is host genetic relationship matrix defined on the causal variants, B is the host microbial taxa similarity matrix defined over the associated taxa, σ_g^2 is the total total additive genetic effect and σ_b^2 is the total total additive microbial effect, $C_r\sigma_r^2 = 2\text{Cov}(A\sigma_g^2, B\sigma_b^2)$, m and l is the total number of causal host genetic variants and microbial taxa, respectively.

Adapting the classical GWAS methodology, A can be estimated using the all SNPs genotyped in the study, as always done. The microbiome similarity matrix can, however, be defined in two ways. First, the phylogenetic distance measures, which accounts for the phylogenetic relationship among microbial taxa, could be used to define sample similarity matrix B . This approach, that has a solid foundation in the field of microbial ecology, would allow B to incorporate the degree of divergence between sequences, thereby estimating similarity among individuals based on phylogenetic relatedness of microbial communities in their bodies. This idea is further supported by the observation that host humans have co-evolved with their microbiome (Ley et al., 2008; Ochman et al., 2010; Sandoval-Motta et al., 2017a) and microbiome-related phenotypes can be transmitted between phylogenetically close humans (Sandoval-Motta et al., 2017a). Second, for each microbial specie, the abundance data can be discretized into ‘categories’ such as 0,1,2 corresponding to low, medium, high abundance respectively, based on some biologically-plausible scale. The frequency of each category can be calculated based on population values. This is, however, not straightforward as it requires knowledge of community-composition of each microbial specie. In the absence of this information, the complete set of individuals in the study samples may be used as proxy for the population. Once this information has been obtained, the sample similarity between the individuals can be obtained, as described in the sequel.

Consider model (5.1) and let R be an incidence matrix that maps different categories of microbial taxa to each subject. In the above case, the elements of B are 0, 1, or 2. Following the usual definition of Euclidean distance similarity, if we $K = RR'$ then the diagonals of K give the subject's relationship to itself while the off-diagonal elements gives the number of elements shared by the subjects. We are interested in investigating over-representation of microbial taxa. Accordingly, as in GWAS, it is possible to define a 'reference category' and determine the corresponding frequency. For our example, if we suppose category "2" is the reference category, then define F to be a matrix containing the frequencies of each category. The j^{th} column of F is $1f_j$, where f_j is the expected value of frequency of the reference category in the j^{th} taxon. With this, the matrix $(R - F)$ becomes the mean-centred form of R , which can essentially be interpreted as setting the mean value of taxa effects to zero. The microbial relationship matrix B is then calculated using

$$B = \frac{(R - F)(R - F)'}{\sum_{j=1}^m \text{Var}(f_j)}.$$

The normalization by $\sum_{j=1}^m \text{Var}(f_j)$ scales B in a way similar to the usual kinship matrix in GWAS.

Finally, the phenotypic variance explained by additive variation at all common SNPs, which we denote by h_{geno}^2 , is calculated from

$$h_{\text{geno}}^2 = \sigma_g^2 / (\sigma_g^2 + \sigma_b^2 + \sigma_r^2 + \sigma_e^2).$$

h_{geno}^2 is the narrow-sense host's genetic heritability of the trait, after accounting for the additive effects of host genetics and the microbiome. We call this 'geno' heritability.

Alternatively, with the emerging 'holobiont' view (Sandoval-Motta et al., 2017a), one can define 'genobiome' heritability as the proportion of phenotypic variance

due to both host genetics and microbiome variances. That is,

$$h_{\text{genobiome}}^2 = \frac{\sigma_g^2 + \sigma_b^2}{\sigma_g^2 + \sigma_b^2 + \sigma_r^2 + \sigma_e^2}. \quad (5.2)$$

The genobiome-heritability, so defined, is the narrow-sense host's genetic and microbiome heritability of the trait, and represents the heritability of the trait jointly explained by the host's genetics and microbiome. The amenability of the microbiome to a plethora of host environmental factors means, however, that the $h_{\text{genobiome}}^2$ is so vulnerable to inflation due to confounding.

5.2.2 Statistical power in genetic association mapping

The statistical power in detecting truly associated genetic variant remains a great issue of concern to geneticists. As this detection power is a function of both the effect size and minor allele frequency of the associated variant, limitations in discovery as well as replication of variants resulting from the generally low effect sizes of variants and the limited allelic polymorphisms across population groups is a notable challenge.

Incorporating information gleaned from inter-individual variability in the microbiome into host GWAS could, by reducing residual variance of the phenotype, improve statistical power. One way towards this end would be to undertake a two-step association procedure, where in the first a MWAS is performed, as described in Chapter 4 in this thesis, to identify microbial OTUs associated (putative causal) with the phenotype of interest, and in the second step a host GWAS is performed, incorporating information (call it 'OTU scores') from the associated OTUs. For this, the individual-level OTU scores, S_i , could be the weighted effect sizes from the putative causal, computed as:

$$S_i = \sum_{j \in \mathcal{A}} \hat{\beta}_j d_{ij},$$

where $\hat{\beta}_j > 0$ is the discovery MWAS log-odds ratio for the putative causal OTU, d_{ij} is individual i 's abundance discretized to some range, say 0, 1, 2 corresponding, respectively, to low, medium and high abundance, and $\mathcal{A}$ is the set of the all putative causal OTUs.

To illustrate this perspective empirically, we simulated a scenario where the true disease status (case-control) is due to the main effect of the host's SNPs and some OTUs, and assessed the p values in the model with and without OTU scores incorporated. In summary, genotype, microbiome, and phenotype of 1000 individuals were simulated, in which some host SNPs and OTUs were set to be 'causal' (see Appendix A.1 and A.2 for simulation details). Two models were then defined: a joint model incorporating host's genotype and microbiome, and a marginal model with host's genotype only. Finally, a association analysis was performed with each model. Figure 5.1 shows the quantile-quantile (QQ) plots for the results with the inflation factor (GC), calculated as the median of the observed joint tests divided by the median χ^2 random variable, to quantify the observation.

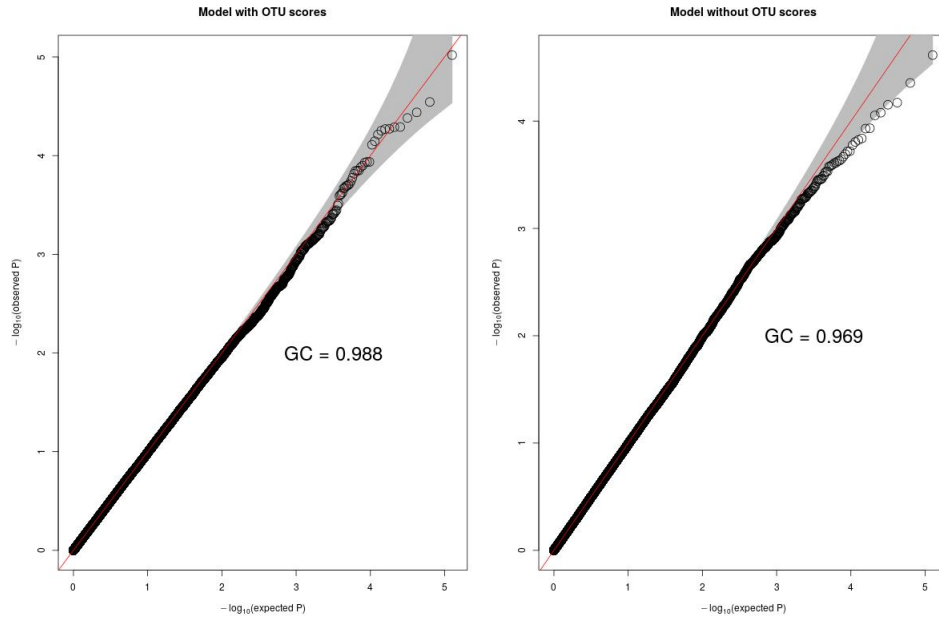


Figure 5.1: Quantile-Quantile plot of observed p values versus the expected p values (on a $-\log_{10}$ scale) for the genome-wide association analysis of the joint (host's genotype and microbiome OTU scores) model and marginal (host's genotype only) model.

We observe, from the plot, a genome-wide inflation factor of 0.988 for the joint model incorporating the host's genotype and microbiome OTU scores, and an inflation factor of 0.969 for the marginal model with host's genotype only. The results reveal some difference in tests from the two models. Specifically, the test from the marginal model show some inflation under the null hypothesis. Indeed, the individual SNP p values from the joint and marginal models (Table 5.1) show appreciable discrepancies.

Table 5.1: p values of SNPs from genome-wide association analysis with the joint (host's genotype and microbiome OTU scores) model and marginal (host's genotype only) model. p - the joint model; p^* - the marginal model. Results displayed for the first 20 SNPs.

SNP	p	p^*	SNP	p	p^*
rs12255619	0.105	0.033	rs4495823	0.635	0.371
rs11252546	0.031	0.106	rs11253478	0.896	0.582
rs7909677	0.134	0.045	rs9419557	0.223	0.563
rs10904494	0.031	0.106	rs9286070	0.223	0.563
rs11591988	0.950	0.584	rs9419560	0.223	0.563
rs4508132	0.705	0.406	rs9419561	0.214	0.538
rs10904561	0.751	0.704	rs11253562	0.216	0.311
rs7917054	0.658	0.962	rs11594819	0.356	0.374
rs7906287	0.638	0.362	rs7909028	0.812	0.536
rs12775579	0.013	0.062	rs7476951	0.322	0.173

The bias (from the marginal model) results from failure to account for phenotype-modifying predictors - the OTUs. Such bias can lead to an increase in false discovery rate under the null (no true association between a genetic variant and the phenotype). From this brief foray, it is plausible that incorporating microbiome information may increase the statistical power for association mapping.

5.2.3 Disease risk prediction

Disease risk prediction using individual-level genetic variation has long been heralded as a potentially valuable clinical application of genomic data. Despite the promise, however, as discussed earlier in this work, risk prediction using host's genetic data is of yet very limited power. While this observation may not be unexpected given that complex diseases have non-host's genetic components and that even though genetic risk attributes might correlate with a disease they may not be informative of the other risk factors associated with the disease, it underscores the need to integrate other phenotypic predictors in disease risk prediction models. Given the association of the human microbiome with its host's health and disease states, and as the microbiome can be expected to harbour unique signatures on its host's genetic variation and phenotypic plasticity owing to their co-evolution, we contend that microbiome-informed risk models will provide mechanistic insights into, and improve, disease risk prediction.

To what extent would leveraging the microbiome in host GWAS improve trait prediction? This is likely to depend on the trait under investigation. Some complex traits such as height are likely to be more strongly determined by the host's genetics than by the microbiome, and in which case the joint heritability model would yield very little to no improvement when compared to classical trait prediction based solely on host genotype; whereas other traits such as metabolic diseases (e.g obesity, and type 2 diabetes) are likely to have strong contribution from both host genetics and microbiome. For such traits, the jointly model would substantially improve trait prediction. Most importantly, the joint model would not vitiate but rather maintain robustness for trait prediction, irrespective of the underlying relative contribution of the predisposing factors.

5.2.4 Extension to the model

There are a number of possible extensions to the model presented in Chapter 4. First, the framework is primarily developed for case-control phenotypes. Extension to quantitative the linear model setting would thus enable it to handle quantitative phenotypes. This can be achieved in a straight-forward way. The only modification would be replacing the logistic model structure with its linear counterpart. Gratifyingly, in this case, the conditional distributions of the model parameters would assume standard distributional forms that can be easily sampled from and hence Gibbs sampling with a Metropolis step for the precision parameter of the Dirichlet process component, can be directly applied.

Second, in light of the empirical observation that functionally related OTUs can form ‘clustered signals’ at varying phylogenetic depths ([Xiao et al., 2018b](#)), it would be interesting to extend the model, allowing it to perform analysis at specific phylogenetic depth (taxonomic rank). This would be especially interesting as the taxonomic resolution of the 16S rRNA gene sequencing differs at each taxonomic rank; degrading at species and strain levels.

Finally, extension by integrating the model into the classic host GWAS frameworks for a joint analysis of traits/disease using both host’s genotype and microbiome profile could enhance our understanding of phenotypic variation and the basis of human complex traits/diseases. One approach towards this end is to undertake a step-wise approach where, in the first step, bacterial taxa informative of the disease/trait under investigation are first determined, as discussed in Chapter 4, and then attributes derived from these putative causal taxa are used as predictors alongside host genotype in classical GWAS frameworks. The mechanic of this integration ought, however, to be carefully investigated to avoid collider bias.

Exploring these new dimensions will offer new perspectives on complex disease variation and pathogenesis, and catalyse the translation of our understanding into clinical and medical practice.

References

- Abraham, G. and Inouye, M. (2015). Genomic risk prediction of complex human disease and its clinical application. *Current opinion in genetics & development*, 33:10–16.
- Abraham, G., Kowalczyk, A., Zobel, J., and Inouye, M. (2013). Performance and robustness of penalized and unpenalized methods for genetic prediction of complex human disease. *Genetic epidemiology*, 37(2):184–195.
- Adams, C. and Gutierrez, B. (2018). The microbiome has multiple influences on human health. *Research and Reviews: Journal of Microbiology and Biotechnology*, 7.
- Ahmad, A., Dwivedi, G., O’Gara, F., Caparros-Martin, J., and Ward, N. (2019). The gut microbiome and atherosclerosis: current knowledge and clinical potential. *American Journal of Physiology-Heart and Circulatory Physiology*.
- Aitchison, J. (1982). The statistical analysis of compositional data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 44(2):139–160.
- Aitchison, J. (2003). The statistical analysis of compositional data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 44(2):139–177.
- Amir, A., McDonald, D., Navas-Molina, J. A., Kopylova, E., Morton, J. T., Xu, Z. Z., Kightley, E. P., Thompson, L. R., Hyde, E. R., Gonzalez, A., et al. (2017). Deblur rapidly resolves single-nucleotide community sequence patterns. *MSystems*, 2(2):e00191–16.
- Amos, W., Driscoll, E., and Hoffman, J. (2010). Candidate genes versus genome-wide associations: which are better for detecting genetic susceptibility to infectious disease? *Proceedings of the Royal Society B: Biological Sciences*, 278(1709):1183–1188.

- Andreoli, T. E., Fitz, J. G., Benjamin, I., Griggs, R. C., and Wing, E. J. (2010). *Andreoli and Carpenter's Cecil Essentials of Medicine E-Book: With STUDENT CONSULT Online Access*. Elsevier Health Sciences.
- Anthony, L. and Komaroff, M. (2018). The microbiome and risk for atherosclerosis.
- Antonarakis, S. E. and Beckmann, J. S. (2006). Mendelian disorders deserve more attention. *Nature Reviews Genetics*, 7(4):277–282.
- Awany, D., Allali, I., Dalvie, S., Hemmings, S., Mwaikono, K. S., Thomford, N. E., Gomez, A., Mulder, N., and Chimusa, E. R. (2019). Host and microbiome genome-wide association studies: current state and challenges. *Frontiers in genetics*, 9.
- Balding, D. J., Bishop, M., and Cannings, C. (2008). *Handbook of statistical genetics*. John Wiley & Sons.
- Barreiro, L. B., Laval, G., Quach, H., Patin, E., and Quintana-Murci, L. (2008). Natural selection has driven population differentiation in modern humans. *Nature genetics*, 40(3):340.
- Bergot, A.-S., Giri, R., and Thomas, R. (2020). The microbiome and rheumatoid arthritis. *Best Practice & Research Clinical Rheumatology*, page 101497.
- Blackwell, D., MacQueen, J. B., et al. (1973). Ferguson distributions via pólya urn schemes. *The annals of statistics*, 1(2):353–355.
- Blekhman, R., Goodrich, J. K., Huang, K., Sun, Q., Bukowski, R., Bell, J. T., Spector, T. D., Keinan, A., Ley, R. E., Gevers, D., et al. (2015). Host genetic variation impacts microbiome composition across human body sites. *Genome biology*, 16(1):191.
- Bogdan, R., Baranger, D. A., and Agrawal, A. (2018). Polygenic risk scores in clinical psychology: bridging genomic risk to individual differences. *Annual review of clinical psychology*, 14:119–157.

- Bomba, L., Walter, K., and Soranzo, N. (2017). The impact of rare and low-frequency genetic variants in common disease. *Genome biology*, 18(1):1–17.
- Bonder, M. J., Kurilshikov, A., Tigchelaar, E. F., Mujagic, Z., Imhann, F., Vila, A. V., Deelen, P., Vatanen, T., Schirmer, M., Smekens, S. P., et al. (2016). The effect of host genetics on the gut microbiome. *Nature genetics*, 48(11):1407–1412.
- Boughorbel, S., Jarray, F., and El-Anbari, M. (2017). Optimal classifier for imbalanced data using matthews correlation coefficient metric. *PloS one*, 12(6).
- Bourrat, P., Lu, Q., and Jablonka, E. (2017). Why the missing heritability might not be in the dna. *BioEssays*, 39(7):1700067.
- Box, G. E. and Tiao, G. C. (1973). *Bayesian inference in statistical analysis*, volume 40. John Wiley & Sons.
- Bray, J. R. and Curtis, J. T. (1957). An ordination of the upland forest communities of southern wisconsin. *Ecological monographs*, 27(4):325–349.
- Breen, G., Li, Q., Roth, B. L., O'Donnell, P., Didriksen, M., Dolmetsch, R., O'Reilly, P. F., Gaspar, H. A., Manji, H., Huebel, C., et al. (2016). Translating genome-wide association findings into new therapeutics for psychiatry. *Nature neuroscience*, 19(11):1392–1396.
- Brody-Moore, P. (2019). Bayesian hierarchical meta-analysis of asymptomatic ebola seroprevalence.
- Brooks, A. W., Priya, S., Blekhman, R., and Bordenstein, S. R. (2018). Gut microbiota diversity across ethnicities in the united states. *PLoS biology*, 16(12).
- Bulik-Sullivan, B. K., Loh, P.-R., Finucane, H. K., Ripke, S., Yang, J., Patterson, N., Daly, M. J., Price, A. L., and Neale, B. M. (2015). Ld score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature genetics*, 47(3):291–295.
- Burns, M. B. and Blekhman, R. (2019). Integrating tumor genomics into studies of the microbiome in colorectal cancer. *Gut microbes*, 10(4):547–552.

- Burr, D. and Doss, H. (2005). A bayesian semiparametric model for random-effects meta-analysis. *Journal of the American Statistical Association*, 100(469):242–251.
- Bush, W. S. and Moore, J. H. (2012). Genome-wide association studies. *PLoS computational biology*, 8(12).
- Callahan, B. J., McMurdie, P. J., and Holmes, S. P. (2017). Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *The ISME journal*, 11(12):2639.
- Cani, P. D. (2019). 16 gut microbiome and obesity. *Handbook of Obesity, Two-Volume Set*, page 183.
- Cao, C. and Moulton, J. (2014). Gwas and drug targets. *BMC genomics*, 15(4):S5.
- Cao, S., Qin, H., Deng, H.-W., and Wang, Y.-P. (2014). A unified sparse representation for sequence variant identification for complex traits. *Genetic epidemiology*, 38(8):671–679.
- Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., Fierer, N., Pena, A. G., Goodrich, J. K., Gordon, J. I., et al. (2010). Qiime allows analysis of high-throughput community sequencing data. *Nature methods*, 7(5):335.
- Cardon, L. R. and Bell, J. I. (2001). Association study designs for complex diseases. *Nature Reviews Genetics*, 2(2):91–99.
- Chakravorty, S., Helb, D., Burday, M., Connell, N., and Alland, D. (2007). A detailed analysis of 16s ribosomal rna gene segments for the diagnosis of pathogenic bacteria. *Journal of microbiological methods*, 69(2):330–339.
- Charlson, E. S., Chen, J., Custers-Allen, R., Bittinger, K., Li, H., Sinha, R., Hwang, J., Bushman, F. D., and Collman, R. G. (2010). Disordered microbial communities in the upper respiratory tract of cigarette smokers. *PloS one*, 5(12):e15216.

- Chen, G.-B., Lee, S. H., Montgomery, G. W., Wray, N. R., Visscher, P. M., Gearry, R. B., Lawrance, I. C., Andrews, J. M., Bampton, P., Mahy, G., et al. (2017). Performance of risk prediction for inflammatory bowel disease based on genotyping platform and genomic risk score method. *BMC medical genetics*, 18(1):94.
- Chen, H., Li, J., Wang, Y., Ng, P. K.-S., Tsang, Y. H., Shaw, K. R., Mills, G. B., and Liang, H. (2020). Comprehensive assessment of computational algorithms in predicting cancer driver mutations. *Genome biology*, 21(1):1–17.
- Chen, J., Bittinger, K., Charlson, E. S., Hoffmann, C., Lewis, J., Wu, G. D., Collman, R. G., Bushman, F. D., and Li, H. (2012a). Associating microbiome composition with environmental covariates using generalized unifracs distances. *Bioinformatics*, 28(16):2106–2113.
- Chen, J., Bushman, F. D., Lewis, J. D., Wu, G. D., and Li, H. (2012b). Structure-constrained sparse canonical correlation analysis with an application to microbiome data analysis. *Biostatistics*, 14(2):244–258.
- Chen, J., Bushman, F. D., Lewis, J. D., Wu, G. D., and Li, H. (2013). Structure-constrained sparse canonical correlation analysis with an application to microbiome data analysis. *Biostatistics*, 14(2):244–258.
- Chen, J., King, E., Deek, R., Wei, Z., Yu, Y., Grill, D., and Ballman, K. (2018a). An omnibus test for differential distribution analysis of microbiome sequencing data. *Bioinformatics*, 34(4):643–651.
- Chen, J. and Li, H. (2013). Variable selection for sparse dirichlet-multinomial regression with an application to microbiome data analysis. *The annals of applied statistics*, 7(1).
- Chen, L. and Chen, J. (2017). Gmpr: A novel normalization method for microbiome sequencing data. *bioRxiv*, page 112565.
- Chen, L., Reeve, J., Zhang, L., Huang, S., Wang, X., and Chen, J. (2018b). Gmpr: A robust normalization method for zero-inflated count data with application to microbiome sequencing data. *PeerJ*, 6:e4600.

- Clarridge, J. E. (2004). Impact of 16s rRNA gene sequence analysis for identification of bacteria on clinical microbiology and infectious diseases. *Clinical microbiology reviews*, 17(4):840–862.
- Clemente, J. C., Ursell, L. K., Parfrey, L. W., and Knight, R. (2012). The impact of the gut microbiota on human health: an integrative view. *Cell*, 148(6):1258–1270.
- Consortium, . G. P. et al. (2015). A global reference for human genetic variation. *Nature*, 526(7571):68.
- Consortium, I. H. et al. (2003). The international hapmap project. *Nature*, 426(6968):789.
- Cordell, H. J. and Clayton, D. G. (2005). Genetic association studies. *The Lancet*, 366(9491):1121–1131.
- Council, N. R. et al. (2007). *The new science of metagenomics: revealing the secrets of our microbial planet*. National Academies Press.
- Damlen, P., Wakefield, J., and Walker, S. (1999). Gibbs sampling for bayesian non-conjugate and hierarchical models by using auxiliary variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(2):331–344.
- Davenport, E. R., Cusanovich, D. A., Michelini, K., Barreiro, L. B., Ober, C., and Gilad, Y. (2015). Genome-wide association studies of the human gut microbiota. *PloS one*, 10(11).
- De Los Campos, G., Gianola, D., and Allison, D. B. (2010). Predicting genetic predisposition in humans: the promise of whole-genome markers. *Nature Reviews Genetics*, 11(12):880.
- Deering, K. E., Devine, A., O’Sullivan, T. A., Lo, J., Boyce, M. C., and Christophersen, C. T. (2020). Characterizing the composition of the pediatric gut microbiome: A systematic review. *Nutrients*, 12(1):16.
- Dein Terra Mota Ribeiro, A. B., Heimesaat, M. M., and Bereswill, S. (2017). Changes of the intestinal microbiome—host homeostasis in hiv-infected individuals—a

- focus on the bacterial gut microbiome. *European Journal of Microbiology and Immunology*, 7(3):158–167.
- Deschasaux, M., Bouter, K. E., Prodan, A., Levin, E., Groen, A. K., Herrema, H., Tremaroli, V., Bakker, G. J., Attaye, I., Pinto-Sietsma, S.-J., et al. (2018). Depicting the composition of gut microbiota in a population with varied ethnic origins but shared geography. *Nature medicine*, 24(10):1526–1531.
- DeWan, A., Liu, M., Hartman, S., Zhang, S. S.-M., Liu, D. T., Zhao, C., Tam, P. O., Chan, W. M., Lam, D. S., Snyder, M., et al. (2006). Htra1 promoter polymorphism in wet age-related macular degeneration. *Science*, 314(5801):989–992.
- Di Noia, A., Martino, A., Montanari, P., and Rizzi, A. (2020). Supervised machine learning techniques and genetic optimization for occupational diseases risk prediction. *Soft Computing*, 24(6):4393–4406.
- Dickerson, F., Severance, E., and Yolken, R. (2017). The microbiome, immunity, and schizophrenia and bipolar disorder. *Brain, behavior, and immunity*, 62:46–52.
- Ding, R.-x., Goh, W.-R., Wu, R.-n., Yue, X.-q., Luo, X., Wei, T. K. W., Wu, J.-r., and Lee, Y.-K. (2019). Revisit gut microbiota and its impact on human health and disease. *journal of food and drug analysis*.
- Dinh, D. M., Volpe, G. E., Duffalo, C., Bhalchandra, S., Tai, A. K., Kane, A. V., Wanke, C. A., and Ward, H. D. (2015). Intestinal microbiota, microbial translocation, and systemic inflammation in chronic hiv infection. *The Journal of infectious diseases*, 211(1):19–27.
- Dorazio, R. M. (2009). On selecting a prior for the precision parameter of dirichlet process mixture models. *Journal of Statistical Planning and Inference*, 139(9):3384–3390.
- Dubourg, G., Lagier, J.-C., Hüe, S., Surenaud, M., Bachar, D., Robert, C., Michelle, C., Ravaux, I., Mokhtari, S., Million, M., et al. (2016). Gut microbiota associated with hiv infection is significantly enriched in bacteria tolerant to oxygen. *BMJ open gastroenterology*, 3(1):e000080.

- Duncan, L. E., Ostacher, M., and Ballon, J. (2019). How genome-wide association studies (gwas) made traditional candidate gene studies obsolete. *Neuropsychopharmacology*, 44(9):1518–1523.
- Dunlap, D., Yang, L., Ries, J., Qin, S., Li, K., Fitch, A., Huang, L., McVerry, B., Methe, B., and Morris, A. (2019). Hiv copd: Exploring the role of host recognition and response to the lung microbiome in pulmonary disease. In *D92. THE MICROBIOME AND LUNG DISEASE*, pages A7025–A7025. American Thoracic Society.
- Eichler, E. E. (2019). Genetic variation, comparative genomics, and the diagnosis of disease. *New England Journal of Medicine*, 381(1):64–74.
- Eichler, E. E., Flint, J., Gibson, G., Kong, A., Leal, S. M., Moore, J. H., and Nadeau, J. H. (2010). Missing heritability and strategies for finding the underlying causes of complex disease. *Nature Reviews Genetics*, 11(6):446.
- Elinav, E., Garrett, W. S., Trinchieri, G., and Wargo, J. (2019). The cancer microbiome. *Nature Reviews Cancer*, 19(7):371–376.
- Emoto, T., Yamashita, T., Kobayashi, T., Sasaki, N., Hirota, Y., Hayashi, T., So, A., Kasahara, K., Yodoi, K., Matsumoto, T., et al. (2017). Characterization of gut microbiota profiles in coronary artery disease patients using data mining analysis of terminal restriction fragment length polymorphism: gut microbiota could be a diagnostic marker of coronary artery disease. *Heart and vessels*, 32(1):39–46.
- Eren, A. M., Maignien, L., Sul, W. J., Murphy, L. G., Grim, S. L., Morrison, H. G., and Sogin, M. L. (2013). Oligotyping: differentiating between closely related microbial taxa using 16s rRNA gene data. *Methods in Ecology and Evolution*, 4(12):1111–1119.
- Eren, A. M., Morrison, H. G., Lescault, P. J., Reveillaud, J., Vineis, J. H., and Sogin, M. L. (2015). Minimum entropy decomposition: unsupervised oligotyping for

- sensitive partitioning of high-throughput marker gene sequences. *The ISME journal*, 9(4):968–979.
- Eric, N., Janet, L., and Steven, K. G. (2016). Inflammation, immune activation, and cardiovascular disease in hiv. *AIDS (London, England)*, 30(10):1495.
- Ericksen, A. J., Lauck, M., Mohns, M. S., DiNapoli, S. R., Mutschler, J. P., Greene, J. M., Weinfurter, J. T., Lehrer-Brey, G., Prall, T. M., Gieger, S. M., et al. (2016). Microbial translocation and inflammation occur in hyperacute immunodeficiency virus infection and compromise host control of virus replication. *PLoS pathogens*, 12(12).
- Erny, D., de Angelis, A. L. H., Jaitin, D., Wieghofer, P., Staszewski, O., David, E., Keren-Shaul, H., Mahlakoiv, T., Jakobshagen, K., Buch, T., et al. (2015). Host microbiota constantly control maturation and function of microglia in the cns. *Nature neuroscience*, 18(7):965.
- Escobar, M. D. and West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the american statistical association*, 90(430):577–588.
- Falconer, D. S. (1965). The inheritance of liability to certain diseases, estimated from the incidence among relatives. *Annals of human genetics*, 29(1):51–76.
- Falconer, D. S. et al. (1960). Introduction to quantitative genetics. *Introduction to quantitative genetics*.
- Fallin, M. D., Duggal, P., and Beaty, T. H. (2016). Genetic epidemiology and public health: the evolution from theory to technology. *American journal of epidemiology*, 183(5):387–393.
- Falony, G., Vandeputte, D., Caenepeel, C., Vieira-Silva, S., Daryoush, T., Vermeire, S., and Raes, J. (2019). The human microbiome in health and disease: hype or hope. *Acta Clinica Belgica*, 74(2):53–64.
- Ferguson, T. S. (1973). A bayesian analysis of some nonparametric problems. *The annals of statistics*, pages 209–230.

- Ferretti, P., Pasoli, E., Tett, A., Asnicar, F., Gorfer, V., Fedi, S., Armanini, F., Truong, D. T., Manara, S., Zolfo, M., et al. (2018). Mother-to-infant microbial transmission from different body sites shapes the developing infant gut microbiome. *Cell host & microbe*, 24(1):133–145.
- Fisher, R. A. (1919). Xv.—the correlation between relatives on the supposition of mendelian inheritance. *Earth and Environmental Science Transactions of the Royal Society of Edinburgh*, 52(2):399–433.
- Floyd, J. S. and Psaty, B. M. (2016). The application of genomics in diabetes: barriers to discovery and implementation. *Diabetes care*, 39(11):1858–1869.
- Fournier, T., Abou Saada, O., Hou, J., Peter, J., Caudal, E., and Schacherer, J. (2019). Extensive impact of low-frequency variants on the phenotypic landscape at population-scale. *Elife*, 8:e49258.
- Frazer, K. A., Murray, S. S., Schork, N. J., and Topol, E. J. (2009). Human genetic variation and its contribution to complex traits. *Nature Reviews Genetics*, 10(4):241.
- Garrett, W. S. (2019). The gut microbiota and colon cancer. *Science*, 364(6446):1133–1135.
- Génin, E. (2019). Missing heritability of complex diseases: case solved? *Human genetics*, pages 1–11.
- Geschwind, D. H. (2018). Evolving views of human genetic variation and its relationship to neurologic and psychiatric disease. In *Handbook of clinical neurology*, volume 147, pages 37–42. Elsevier.
- Geweke, J. et al. (1991). *Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments*, volume 196. Federal Reserve Bank of Minneapolis, Research Department Minneapolis, MN.
- Gianola, D., Fernando, R. L., and Stella, A. (2006). Genomic-assisted prediction of genetic value with semiparametric procedures. *Genetics*, 173(3):1761–1776.

- Gilbert, J. A., Quinn, R. A., Debelius, J., Xu, Z. Z., Morton, J., Garg, N., Jansson, J. K., Dorrestein, P. C., and Knight, R. (2016). Microbiome-wide association studies link dynamic microbial consortia to disease. *Nature*, 535(7610):94.
- Gill, S. R., Pop, M., DeBoy, R. T., Eckburg, P. B., Turnbaugh, P. J., Samuel, B. S., Gordon, J. I., Relman, D. A., Fraser-Liggett, C. M., and Nelson, K. E. (2006). Metagenomic analysis of the human distal gut microbiome. *science*, 312(5778):1355–1359.
- Gola, D., Erdmann, J., Müller-Myhsok, B., Schunkert, H., and König, I. R. (2020). Polygenic risk scores outperform machine learning methods in predicting coronary artery disease status. *Genetic Epidemiology*.
- Goodrich, J. K., Davenport, E. R., Beaumont, M., Jackson, M. A., Knight, R., Ober, C., Spector, T. D., Bell, J. T., Clark, A. G., and Ley, R. E. (2016). Genetic determinants of the gut microbiome in uk twins. *Cell host & microbe*, 19(5):731–743.
- Gootenberg, D. B., Paer, J. M., Luevano, J.-M., and Kwon, D. S. (2017). Hiv-associated changes in the enteric microbial community: potential role in loss of homeostasis and development of systemic inflammation. *Current opinion in infectious diseases*, 30(1):31.
- Gorlov, I. P., Gorlova, O. Y., Frazier, M. L., Spitz, M. R., and Amos, C. I. (2011). Evolutionary evidence of the effect of rare variants on disease etiology. *Clinical genetics*, 79(3):199–206.
- Grady, B. J., Torstenson, E. S., and Ritchie, M. D. (2011). The effects of linkage disequilibrium in large scale snp datasets for mdr. *BioData mining*, 4(1):11.
- Greub, W. H. (2012). *Linear algebra*, volume 23. Springer Science & Business Media.
- Gundogdu, A. and Nalbantoglu, U. (2017). Human genome-microbiome interaction: metagenomics frontiers for the aetiopathology of autoimmune diseases. *Microbial genomics*, 3(4).

- Halfvarson, J., Brislawn, C. J., Lamendella, R., Vázquez-Baeza, Y., Walters, W. A., Bramer, L. M., D'Amato, M., Bonfiglio, F., McDonald, D., Gonzalez, A., et al. (2017). Dynamics of the human gut microbiome in inflammatory bowel disease. *Nature microbiology*, 2(5):17004.
- Hawkins, G., Clark, L., Sullivan, S., Keku, T., Brewster, W., and Bae-Jump, V. (2019). Impact of ethnicity and obesity on the uterine microbiome in women and mice with endometrial cancer. *Gynecologic Oncology*, 154:77.
- Hayes, B. (2013). Overview of statistical methods for genome-wide association studies (gwas). In *Genome-wide association studies and genomic prediction*, pages 149–169. Springer.
- Henriksen, M. G., Nordgaard, J., and Jansson, L. B. (2017). Genetics of schizophrenia: overview of methods, findings and limitations. *Frontiers in human neuroscience*, 11:322.
- Hindorff, L. A., Gillanders, E. M., and Manolio, T. A. (2011). Genetic architecture of cancer and other complex diseases: lessons learned and future directions. *Carcinogenesis*, 32(7):945–954.
- Hirschhorn, J. N. et al. (2009). Genomewide association studies—illuminating biologic pathways. *New England Journal of Medicine*, 360(17):1699.
- Hong, M., Ye, B. D., Yang, S.-K., Jung, S., Lee, H.-S., Kim, B. M., Lee, S. B., Hong, J., Baek, J., Park, S. H., et al. (2018). Immunochip meta-analysis of inflammatory bowel disease identifies three novel loci and four novel associations in previously reported loci. *Journal of Crohn's and Colitis*, 12(6):730–741.
- Hu, T., Gallins, P., and Zhou, Y.-H. (2018). A zero-inflated beta-binomial model for microbiome data analysis. *Stat*, 7(1):e185.
- Hua, X., Song, L., Yu, G., Goedert, J. J., Abnet, C. C., Landi, M. T., and Shi, J. (2015). Microbiomegwas: a tool for identifying host genetic variants associated with microbiome composition. *BioRxiv*, page 031187.

- Huang, W. and Mackay, T. F. (2016). The genetic architecture of quantitative traits cannot be inferred from variance component analysis. *PLoS genetics*, 12(11):e1006421.
- Huang, Y., Shi, X., Li, Z., Shen, Y., Shi, X., Wang, L., Li, G., Yuan, Y., Wang, J., Zhang, Y., et al. (2018). Possible association of firmicutes in the gut microbiota of patients with major depressive disorder. *Neuropsychiatric disease and treatment*, 14:3329.
- Janda, J. M. and Abbott, S. L. (2007). 16s rRNA gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils, and pitfalls. *Journal of clinical microbiology*, 45(9):2761–2764.
- Jeff, G. (2007). Bayesian methods for the social and behavioral sciences.
- Jeyakumar, T., Beauchemin, N., and Gros, P. (2019). Impact of the microbiome on the human genome. *Trends in parasitology*.
- Johnson, G. C., Esposito, L., Barratt, B. J., Smith, A. N., Heward, J., Di Genova, G., Ueda, H., Cordell, H. J., Eaves, I. A., Dudbridge, F., et al. (2001). Haplotype tagging for the identification of common disease genes. *Nature genetics*, 29(2):233–237.
- Jonsson, A. L. and Bäckhed, F. (2017). Role of gut microbiota in atherosclerosis. *Nature Reviews Cardiology*, 14(2):79.
- Jovel, J., Patterson, J., Wang, W., Hotte, N., O’Keefe, S., Mitchel, T., Perry, T., Kao, D., Mason, A. L., Madsen, K. L., et al. (2016). Characterization of the gut microbiome using 16s or shotgun metagenomics. *Frontiers in microbiology*, 7:459.
- Kalyuzhnaya, M. G., Lapidus, A., Ivanova, N., Copeland, A. C., McHardy, A. C., Szeto, E., Salamov, A., Grigoriev, I. V., Suci, D., Levine, S. R., et al. (2008). High-resolution metagenomics targets specific functional types in complex microbial communities. *Nature biotechnology*, 26(9):1029–1034.

- Karki, R., Pandya, D., Elston, R. C., and Ferlini, C. (2015). Defining “mutation” and “polymorphism” in the era of personal genomics. *BMC medical genomics*, 8(1):37.
- Khachatryan, Z. A., Ktsoyan, Z. A., Manukyan, G. P., Kelly, D., Ghazaryan, K. A., and Aminov, R. I. (2008). Predominant role of host genetics in controlling the composition of gut microbiota. *PloS one*, 3(8).
- Khan, I., Ullah, N., Zha, L., Bai, Y., Khan, A., Zhao, T., Che, T., and Zhang, C. (2019). Alteration of gut microbiota in inflammatory bowel disease (ibd): Cause or consequence? ibd treatment targeting the gut microbiome. *Pathogens*, 8(3):126.
- Khera, A. V., Chaffin, M., Aragam, K. G., Haas, M. E., Roselli, C., Choi, S. H., Natarajan, P., Lander, E. S., Lubitz, S. A., Ellinor, P. T., et al. (2018). Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nature genetics*, 50(9):1219–1224.
- Kho, Z. Y. and Lal, S. K. (2018). The human gut microbiome—a potential controller of wellness and disease. *Frontiers in microbiology*, 9.
- Klein, J. P. and Moeschberger, M. L. (1997). Statistics for biology and health. *Stat. Biol. Health, New York*, 27238.
- Klein, R. J., Zeiss, C., Chew, E. Y., Tsai, J.-Y., Sackler, R. S., Haynes, C., Henning, A. K., SanGiovanni, J. P., Mane, S. M., Mayne, S. T., et al. (2005). Complement factor h polymorphism in age-related macular degeneration. *Science*, 308(5720):385–389.
- Koboziev, I., Webb, C. R., Furr, K. L., and Grisham, M. B. (2014). Role of the enteric microbiota in intestinal homeostasis and inflammation. *Free Radical Biology and Medicine*, 68:122–133.
- Koenig, J. E., Spor, A., Scalfone, N., Fricker, A. D., Stombaugh, J., Knight, R., Angenent, L. T., and Ley, R. E. (2011). Succession of microbial consortia in the developing infant gut microbiome. *Proceedings of the National Academy of Sciences*, 108(Supplement 1):4578–4585.

- Koh, H., Li, Y., Zhan, X., Chen, J., and Zhao, N. (2019). A distance-based kernel association test based on the generalized linear mixed model for correlated microbiome studies. *Frontiers in genetics*, 10:458.
- Konstantinidis, K. T. and Tiedje, J. M. (2005). Genomic insights that advance the species definition for prokaryotes. *Proceedings of the National Academy of Sciences*, 102(7):2567–2572.
- Koren, O., Spor, A., Felin, J., Fåk, F., Stombaugh, J., Tremaroli, V., Behre, C. J., Knight, R., Fagerberg, B., Ley, R. E., et al. (2011). Human oral, gut, and plaque microbiota in patients with atherosclerosis. *Proceedings of the National Academy of Sciences*, 108(Supplement 1):4592–4598.
- Kurtz, Z. D., Müller, C. L., Miraldi, E. R., Littman, D. R., Blaser, M. J., and Bonneau, R. A. (2015). Sparse and compositionally robust inference of microbial ecological networks. *PLoS computational biology*, 11(5).
- Kuss, S. K., Best, G. T., Etheredge, C. A., Pruijssers, A. J., Frierson, J. M., Hooper, L. V., Dermody, T. S., and Pfeiffer, J. K. (2011). Intestinal microbiota promote enteric virus replication and systemic pathogenesis. *Science*, 334(6053):249–252.
- Kyung, M., Gill, J., and Casella, G. (2009). Sampling schemes for generalized linear dirichlet random effects models. Technical report, Technical Report, University of Florida, Department of Statistics. Available
- Kyung, M., Gill, J., Casella, G., et al. (2010). Estimation in dirichlet random effects models. *The Annals of Statistics*, 38(2):979–1009.
- Lange, K. (2003). *Mathematical and statistical methods for genetic analysis*. Springer Science & Business Media.
- Läuter, H. (1988). Silverman, bw: Density estimation for statistics and data analysis. chapman & hall, london–new york 1986, 175 pp., £ 12.—. *Biometrical Journal*, 30(7):876–877.

- Lee, S. H., Wray, N. R., Goddard, M. E., and Visscher, P. M. (2011). Estimating missing heritability for disease from genome-wide association studies. *The American Journal of Human Genetics*, 88(3):294–305.
- Lettre, G. (2014). Rare and low-frequency variants in human common diseases and other complex traits. *Journal of medical genetics*, 51(11):705–714.
- Levin, M. G. and Rader, D. J. (2020). Polygenic risk scores and coronary artery disease: Ready for prime time?
- Ley, R. E., Hamady, M., Lozupone, C., Turnbaugh, P. J., Ramey, R. R., Bircher, J. S., Schlegel, M. L., Tucker, T. A., Schrenzel, M. D., Knight, R., et al. (2008). Evolution of mammals and their gut microbes. *Science*, 320(5883):1647–1651.
- Li, C., Casella, G., and Ghosh, M. (2018). Estimation of regression vectors in linear mixed models with dirichlet process random effects. *Communications in Statistics-Theory and Methods*, 47(16):3935–3954.
- Li, M., Budding, A. E., van der Lugt-Degen, M., Du-Thumm, L., Vandeven, M., and Fan, A. (2019). The influence of age, gender and race/ethnicity on the composition of the human axillary microbiome. *International journal of cosmetic science*, 41(4):371–377.
- Li, P., Piao, Y., Shon, H. S., and Ryu, K. H. (2015). Comparing the normalization methods for the differential analysis of illumina high-throughput rna-seq data. *BMC bioinformatics*, 16(1):347.
- Li, Y., Zhao, H., Sun, J., and Kim, K. (2014). Nonparametric tests for panel count data with unequal observation processes. *Computational Statistics & Data Analysis*, 73:103–111.
- Lin, W., Shi, P., Feng, R., and Li, H. (2014). Variable selection in regression with compositional covariates. *Biometrika*, 101(4):785–797.
- Lin, Y.-L. and Gokcumen, O. (2019). Fine-scale characterization of genomic structural variation in the human genome reveals adaptive and biomedically relevant hotspots. *Genome biology and evolution*, 11(4):1136–1151.

- Linck, E. and Battey, C. (2019). Minor allele frequency thresholds strongly affect population structure inference with genomic data sets. *Molecular Ecology Resources*, 19(3):639–647.
- Ling, Z., Jin, C., Xie, T., Cheng, Y., Li, L., and Wu, N. (2016). Alterations in the fecal microbiota of patients with hiv-1 infection: an observational study in a chinese population. *Scientific reports*, 6(1):1–12.
- Liovat, A.-S., Rey-Cuillé, M.-A., Lecuroux, C., Jacquelin, B., Girault, I., Petitjean, G., Zitoun, Y., Venet, A., Barré-Sinoussi, F., Lebon, P., et al. (2012). Acute plasma biomarkers of t cell activation set-point levels and of disease progression in hiv-1 infection. *PloS one*, 7(10).
- Liu, D., Chen, Z., Zhang, Z., Sun, H., Ma, P., Zhu, K., Liu, G., Wang, Q., and Pan, Y. (2019a). Detection of genome-wide structural variations in the shanghai holstein cattle population using next-generation sequencing. *Asian-Australasian journal of animal sciences*, 32(3):320.
- Liu, H., Chen, X., Hu, X., Niu, H., Tian, R., Wang, H., Pang, H., Jiang, L., Qiu, B., Chen, X., et al. (2019b). Alterations in the gut microbiome and metabolism with coronary artery disease severity. *Microbiome*, 7(1):68.
- Lloyd-Jones, L. R., Zeng, J., Sidorenko, J., Yengo, L., Moser, G., Kemper, K. E., Wang, H., Zheng, Z., Magi, R., Esko, T., et al. (2019). Improved polygenic prediction by bayesian multiple regression on summary statistics. *Nature communications*, 10(1):1–11.
- Loeys, T., Moerkerke, B., De Smet, O., and Buysse, A. (2012). The analysis of zero-inflated count data: Beyond zero-inflated poisson regression. *British Journal of Mathematical and Statistical Psychology*, 65(1):163–180.
- López-Cortegano, E. and Caballero, A. (2019). Gwehs: A genome-wide effect sizes and heritability screener. *Genes*, 10(8):558.

- Lozupone, C. and Knight, R. (2005). Unifrac: a new phylogenetic method for comparing microbial communities. *Applied and environmental microbiology*, 71(12):8228–8235.
- Lozupone, C. A., Li, M., Campbell, T. B., Flores, S. C., Linderman, D., Gebert, M. J., Knight, R., Fontenot, A. P., and Palmer, B. E. (2013). Alterations in the gut microbiota associated with hiv-1 infection. *Cell host & microbe*, 14(3):329–339.
- Lozupone, C. A., Rhodes, M. E., Neff, C. P., Fontenot, A. P., Campbell, T. B., and Palmer, B. E. (2014). Hiv-induced alteration in gut microbiota: driving factors, consequences, and effects of antiretroviral therapy. *Gut microbes*, 5(4):562–570.
- Lu, J., Zhang, J., and Tao, J. (2018). Slice–gibbs sampling algorithm for estimating the parameters of a multilevel item response model. *Journal of Mathematical Psychology*, 82:12–25.
- Luca, F., Kupfer, S. S., Knights, D., Khoruts, A., and Blekhman, R. (2018). Functional genomics of host–microbiome interactions in humans. *Trends in Genetics*, 34(1):30–40.
- Lush, J. L. (1947). Family merit and individual merit as bases for selection. part i. *The American Naturalist*, 81(799):241–261.
- Lv, L.-X., Jiang, H.-Y., Yan, R., and Li, L. (2019). Interactions between gut microbiota and hosts and their role in infectious diseases. *Infectious Microbes & Diseases*, 1(1):3–9.
- Lynch, J., Tang, K., Priya, S., Sands, J., Sands, M., Tang, E., Mukherjee, S., Knights, D., and Blekhman, R. (2017). Hominid: A framework for identifying associations between host genetic variation and microbiome composition. *GigaScience*, 6(12):gix107.
- Lynch, M., Walsh, B., et al. (1998). *Genetics and analysis of quantitative traits*, volume 1. Sinauer Sunderland, MA.

- Mahajan, A., Go, M. J., Zhang, W., Below, J. E., Gaulton, K. J., Ferreira, T., Horikoshi, M., Johnson, A. D., Ng, M. C., Prokopenko, I., et al. (2014). Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. *Nature genetics*, 46(3):234–244.
- Mahmud, Z., Malik, S. U. F., Ahmed, J., and Azad, A. K. (2016). Computational analysis of damaging single-nucleotide polymorphisms and their structural and functional impact on the insulin receptor. *BioMed research international*, 2016.
- Mäki-Tanila, A. and Hill, W. G. (2014). Influence of gene interaction on complex trait variation with multilocus models. *Genetics*, 198(1):355–367.
- Mandal, S., Van Treuren, W., White, R. A., Eggesbø, M., Knight, R., and Peddada, S. D. (2015). Analysis of composition of microbiomes: a novel method for studying microbial composition. *Microbial ecology in health and disease*, 26(1):27663.
- Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., McCarthy, M. I., Ramos, E. M., Cardon, L. R., Chakravarti, A., et al. (2009). Finding the missing heritability of complex diseases. *Nature*, 461(7265):747.
- Martins, E. P. and Hansen, T. F. (1997). Phylogenies and the comparative method: a general approach to incorporating phylogenetic information into the analysis of interspecific data. *The American Naturalist*, 149(4):646–667.
- Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure*, 405(2):442–451.
- Mavaddat, N., Michailidou, K., Dennis, J., Lush, M., Fachal, L., Lee, A., Tyrer, J. P., Chen, T.-H., Wang, Q., Bolla, M. K., et al. (2019). Polygenic risk scores for prediction of breast cancer and breast cancer subtypes. *The American Journal of Human Genetics*, 104(1):21–34.

- Mitchell, K. J. (2012). What is complex about complex disorders? *Genome biology*, 13(1):1–11.
- Mittag, F., Büchel, F., Saad, M., Jahn, A., Schulte, C., Bochdanovits, Z., Simón-Sánchez, J., Nalls, M. A., Keller, M., Hernandez, D. G., et al. (2012). Use of support vector machines for disease risk prediction in genome-wide association studies: Concerns and opportunities. *Human mutation*, 33(12):1708–1718.
- Moore, J. H. and Williams, S. M. (2009). Epistasis and its implications for personal genetics. *The American Journal of Human Genetics*, 85(3):309–320.
- Motsinger, A. A., Haas, D. W., Todd, H., and Ritchie, M. D. (2007). Human genomic association studies: a primer for the infectious diseases specialist. *The Journal of infectious diseases*, 195(12):1737–1744.
- Murugiah, S. and Sweeting, T. (2012). Selecting the precision parameter prior in dirichlet process mixture models. *Journal of Statistical Planning and Inference*, 142(7):1947–1959.
- Neff, C. P., Krueger, O., Xiong, K., Arif, S., Nusbacher, N., Schneider, J. M., Cunningham, A. W., Armstrong, A., Li, S., McCarter, M. D., et al. (2018). Fecal microbiota composition drives immune activation in hiv-infected individuals. *EBioMedicine*, 30:192–202.
- Nguyen, N.-P., Warnow, T., Pop, M., and White, B. (2016). A perspective on 16s rrna operational taxonomic unit clustering using sequence similarity. *NPJ biofilms and microbiomes*, 2(1):1–8.
- Nguyen, P. A., Born, D. A., Deaton, A. M., Nioi, P., and Ward, L. D. (2019a). Phenotypes associated with genes encoding drug targets are predictive of clinical trial side effects. *Nature communications*, 10(1):1–11.
- Nguyen, T. T., Kosciolk, T., Maldonado, Y., Daly, R. E., Martin, A. S., McDonald, D., Knight, R., and Jeste, D. V. (2019b). Differences in gut microbiome composition between persons with chronic schizophrenia and healthy comparison subjects. *Schizophrenia research*, 204:23–29.

- Niccolai, E., Boem, F., Russo, E., and Amedei, A. (2019). The gut–brain axis in the neuropsychological disease model of obesity: A classical movie revised by the emerging director “microbiome”. *Nutrients*, 11(1):156.
- Nolte, I. M., van der Most, P. J., Alizadeh, B. Z., de Bakker, P. I., Boezen, H. M., Bruinenberg, M., Franke, L., van der Harst, P., Navis, G., Postma, D. S., et al. (2017). Missing heritability: is the gap closing? an analysis of 32 complex traits in the lifelines cohort study. *European Journal of Human Genetics*, 25(7):877.
- Norouzi, S., Adulcikas, J., Sohal, S. S., and Myers, S. (2017). Zinc transporters and insulin resistance: therapeutic implications for type 2 diabetes and metabolic disease. *Journal of biomedical science*, 24(1):87.
- Nowak, P., Troseid, M., Avershina, E., Barqasho, B., Neogi, U., Holm, K., Hov, J. R., Noyan, K., Vesterbacka, J., Svärd, J., et al. (2015). Gut microbiota diversity predicts immune status in hiv-1 infection. *Aids*, 29(18):2409–2418.
- Ochman, H., Worobey, M., Kuo, C.-H., Ndjango, J.-B. N., Peeters, M., Hahn, B. H., and Hugenholtz, P. (2010). Evolutionary relationships of wild hominids recapitulated by gut microbial communities. *PLoS biology*, 8(11):e1000546.
- O’Connor, L. J. (2019). *Genetic Architecture of Common Diseases and Complex Traits*. PhD thesis.
- Ott, J. (1999). *Analysis of human genetic linkage*. JHU Press.
- Overington, J. P., Al-Lazikani, B., and Hopkins, A. L. (2006). How many drug targets are there? *Nature reviews Drug discovery*, 5(12):993–996.
- Pallares, L. F. (2019). Genetic variation: Searching for solutions to the missing heritability problem. *eLife*, 8:e53018.
- Pan, L.-L., Qin, M., Liu, X.-H., and Zhu, Y.-Z. (2017). The role of hydrogen sulfide on cardiovascular homeostasis: an overview with update on immunomodulation. *Frontiers in pharmacology*, 8:686.

- Paulson, J. N., Stine, O. C., Bravo, H. C., and Pop, M. (2013). Differential abundance analysis for microbial marker-gene surveys. *Nature methods*, 10(12):1200.
- Pearson, E. (2014). Zinc transport and diabetes risk. *Nature genetics*, 46(4):323–324.
- Pedersen, C.-E. T., Lohmueller, K. E., Grarup, N., Bjerregaard, P., Hansen, T., Siegmund, H. R., Moltke, I., and Albrechtsen, A. (2017). The effect of an extreme and prolonged population bottleneck on patterns of deleterious variation: Insights from the greenlandic inuit. *Genetics*, 205(2):787–801.
- Peng, J., Xiao, X., Hu, M., and Zhang, X. (2018). Interaction between gut microbiome and cardiovascular disease. *Life sciences*, 214:153–157.
- Peng, X., Li, G., and Liu, Z. (2016). Zero-inflated beta regression for differential abundance analysis with metagenomics data. *Journal of Computational Biology*, 23(2):102–110.
- Pinto-Cardoso, S., Lozupone, C., Briceño, O., Alva-Hernández, S., Téllez, N., Adriana, A., Murakami-Ogasawara, A., and Reyes-Terán, G. (2017). Fecal bacterial communities in treated hiv infected individuals on two antiretroviral regimens. *Scientific reports*, 7:43741.
- Platt, A., Pivrotto, A., Knoblauch, J., and Hey, J. (2019). An estimator of first coalescent time reveals selection on young variants and large heterogeneity in rare allele ages among human populations. *PLoS genetics*, 15(8):e1008340.
- Platt, A., Vilhjálmsson, B. J., and Nordborg, M. (2010). Conditions under which genome-wide association studies will be positively misleading. *Genetics*, 186(3):1045–1052.
- Plottel, C. S. and Blaser, M. J. (2011). Microbiome and malignancy. *Cell host & microbe*, 10(4):324–335.
- Price, A. L., Zaitlen, N. A., Reich, D., and Patterson, N. (2010). New approaches to population stratification in genome-wide association studies. *Nature Reviews Genetics*, 11(7):459–463.

- Pritchard, J. K. and Cox, N. J. (2002). The allelic architecture of human disease genes: common disease–common variant... or not? *Human molecular genetics*, 11(20):2417–2423.
- Purdom, E. (2011). Analysis of a data matrix and a graph: Metagenomic data and the phylogenetic tree. *The Annals of Applied Statistics*, pages 2326–2358.
- Randolph, T. W., Zhao, S., Copeland, W., Hullar, M., Shojaie, A., et al. (2018). Kernel-penalized regression for analysis of microbiome data. *The Annals of Applied Statistics*, 12(1):540–566.
- Ranjan, R., Rani, A., Metwally, A., McGee, H. S., and Perkins, D. L. (2016). Analysis of the microbiome: Advantages of whole genome shotgun versus 16s amplicon sequencing. *Biochemical and biophysical research communications*, 469(4):967–977.
- Rausch, P., Rehman, A., Künzel, S., Häsler, R., Ott, S. J., Schreiber, S., Rosenstiel, P., Franke, A., and Baines, J. F. (2011). Colonic mucosa-associated microbiota is influenced by an interaction of crohn disease and fut2 (secretor) genotype. *Proceedings of the National Academy of Sciences*, 108(47):19030–19035.
- Raynes, Y., Halstead, A. L., and Sniegowski, P. D. (2014). The effect of population bottlenecks on mutation rate evolution in asexual populations. *Journal of evolutionary biology*, 27(1):161–169.
- Regalado, J., Lundberg, D. S., Deusch, O., Kersten, S., Karasov, T. L., Poersch, K., Shirsekar, G., and Weigel, D. (2019). Combining whole genome shotgun sequencing and rDNA amplicon analyses to improve detection of microbe-microbe interaction networks in plant leaves. *BioRxiv*, page 823492.
- Risch, N. and Merikangas, K. (1996). The future of genetic studies of complex human diseases. *Science*, 273(5281):1516–1517.
- Robert, F. and Pelletier, J. (2018). Exploring the impact of single-nucleotide polymorphisms on translation. *Frontiers in genetics*, 9:507.

- Robinson, J. G., Farnier, M., Krempf, M., Bergeron, J., Luc, G., Aversa, M., Stroes, E. S., Langslet, G., Raal, F. J., El Shahawy, M., et al. (2015). Efficacy and safety of alirocumab in reducing lipids and cardiovascular events. *New England Journal of Medicine*, 372(16):1489–1499.
- Robinson, M. R., Wray, N. R., and Visscher, P. M. (2014). Explaining additional genetic variation in complex traits. *Trends in Genetics*, 30(4):124–132.
- Rocafort, M., Noguera-Julian, M., Rivera, J., Pastor, L., Guillén, Y., Langhorst, J., Parera, M., Mandomando, I., Carrillo, J., Urrea, V., et al. (2019). Evolution of the gut microbiome following acute hiv-1 infection. *Microbiome*, 7(1):73.
- Rodríguez, J. M., Murphy, K., Stanton, C., Ross, R. P., Kober, O. I., Juge, N., Avershina, E., Rudi, K., Narbad, A., Jenmalm, M. C., et al. (2015). The composition of the gut microbiota throughout life, with an emphasis on early life. *Microbial ecology in health and disease*, 26(1):26050.
- Rothschild, D., Weissbrod, O., Barkan, E., Kurilshikov, A., Korem, T., Zeevi, D., Costea, P. I., Godneva, A., Kalka, I. N., Bar, N., et al. (2018). Environment dominates over host genetics in shaping human gut microbiota. *Nature*, 555(7695):210–215.
- Safari, Z. and Gérard, P. (2019). The links between the gut microbiome and non-alcoholic fatty liver disease (nafld). *Cellular and Molecular Life Sciences*, 76(8):1541–1558.
- San-Juan-Vergara, H., Zurek, E., Ajami, N. J., Mogollon, C., Peña, M., Portnoy, I., Vélez, J. I., Cadena-Cruz, C., Diaz-Olmos, Y., Hurtado-Gómez, L., et al. (2018). A lachnospiraceae-dominated bacterial signature in the fecal microbiota of hiv-infected individuals from colombia, south america. *Scientific reports*, 8(1):1–13.
- Sandoval-Motta, S., Aldana, M., and Frank, A. (2017a). Evolving ecosystems: Inheritance and selection in the light of the microbiome. *Archives of medical research*, 48(8):780–789.

- Sandoval-Motta, S., Aldana, M., Martínez-Romero, E., and Frank, A. (2017b). The human microbiome and the missing heritability problem. *Frontiers in genetics*, 8:80.
- Sanseau, P., Agarwal, P., Barnes, M. R., Pastinen, T., Richards, J. B., Cardon, L. R., and Mooser, V. (2012). Use of genome-wide association studies for drug repositioning. *Nature biotechnology*, 30(4):317.
- Sanz, Y., Olivares, M., Moya-Pérez, Á., and Agostoni, C. (2015). Understanding the role of gut microbiome in metabolic disease risk. *Pediatric research*, 77(1-2):236–244.
- Saus, E., Iraola-Guzmán, S., Willis, J. R., Brunet-Vega, A., and Gabaldón, T. (2019). Microbiome and colorectal cancer: Roles in carcinogenesis and clinical potential. *Molecular aspects of medicine*.
- Scherer, S. W. and Visscher, P. M. (2016). *Genome-wide association studies: From polymorphism to personalized medicine*. Cambridge University Press.
- Schmidt, T. S., Raes, J., and Bork, P. (2018). The human gut microbiome: from association to modulation. *Cell*, 172(6):1198–1215.
- Schuijt, T. J., van der Poll, T., de Vos, W. M., and Wiersinga, W. J. (2013). The intestinal microbiota and host immune interactions in the critically ill. *Trends in microbiology*, 21(5):221–229.
- Scotti, E., Boué, S., Sasso, G. L., Zanetti, F., Belcastro, V., Poussin, C., Sierro, N., Battey, J., Gimalac, A., Ivanov, N. V., et al. (2017). Exploring the microbiome in health and disease: Implications for toxicology. *Toxicology Research and application*, 1:2397847317741884.
- Shah, N., Tang, H., Doak, T. G., and Ye, Y. (2011). Comparing bacterial communities inferred from 16s rrna gene sequencing and shotgun metagenomics. In *Biocomputing 2011*, pages 165–176. World Scientific.

- Shah, S., Bonder, M. J., Marioni, R. E., Zhu, Z., McRae, A. F., Zhernakova, A., Harris, S. E., Liewald, D., Henders, A. K., Mendelson, M. M., et al. (2015). Improving phenotypic prediction by combining genetic and epigenetic associations. *The American Journal of Human Genetics*, 97(1):75–85.
- Shi, P., Zhang, A., Li, H., et al. (2016). Regression analysis for microbiome compositional data. *The Annals of Applied Statistics*, 10(2):1019–1040.
- Singer-Englar, T., Barlow, G., and Mathur, R. (2019). Obesity, diabetes, and the gut microbiome: an updated review. *Expert review of gastroenterology & hepatology*, 13(1):3–15.
- Sirugo, G., Williams, S. M., and Tishkoff, S. A. (2019). The missing diversity in human genetic studies. *Cell*, 177(1):26–31.
- Slatkin, M. (2009). Epigenetic inheritance and the missing heritability problem. *Genetics*, 182(3):845–850.
- Sohn, M. B., Du, R., and An, L. (2015). A robust approach for identifying differentially abundant features in metagenomic samples. *Bioinformatics*, 31(14):2269–2275.
- Song, S., Jiang, W., Hou, L., and Zhao, H. (2020). Leveraging effect size distributions to improve polygenic risk scores derived from summary statistics of genome-wide association studies. *PLOS Computational Biology*, 16(2):e1007565.
- Speed, D., Hemani, G., Johnson, M., and Balding, D. (2012). Improved heritability estimation from genome-wide snp data. *Am. J. Hum*, 116:1011–1021.
- Speed, D., Holmes, J., and Balding, D. J. (2020). Evaluating and improving heritability models using summary statistics. *Nature genetics*, 52(4):458–462.
- Spitz, M. R., Amos, C. I., Dong, Q., Lin, J., and Wu, X. (2008). The chrna5-a3 region on chromosome 15q24-25.1 is a risk factor both for nicotine dependence and for lung cancer. *JNCI: Journal of the National Cancer Institute*, 100(21):1552–1556.

- Stewart, J. A., Chadwick, V. S., and Murray, A. (2005). Investigations into the influence of host genetics on the predominant eubacteria in the faecal microflora of children. *Journal of medical microbiology*, 54(12):1239–1242.
- Sud, A., Kinnersley, B., and Houlston, R. S. (2017). Genome-wide association studies of cancer: current insights and future perspectives. *Nature Reviews Cancer*, 17(11):692–704.
- Sukhumsirichart, W. (2018). Polymorphisms. In *Genetic Diversity and Disease Susceptibility*. IntechOpen.
- Sul, J. H., Martin, L. S., and Eskin, E. (2018). Population structure in genetic studies: Confounding factors and mixed models. *PLoS genetics*, 14(12).
- Sun, Y., Ma, Y., Lin, P., Tang, Y.-W., Yang, L., Shen, Y., Zhang, R., Liu, L., Cheng, J., Shao, J., et al. (2016). Fecal bacterial microbiome diversity in chronic hiv-infected patients in china. *Emerging microbes & infections*, 5(1):1–7.
- Tamames, J., Sánchez, P. D., Nikel, P. I., and Pedrós-Alió, C. (2016). Quantifying the relative importance of phylogeny and environmental preferences as drivers of gene content in prokaryotic microorganisms. *Frontiers in microbiology*, 7:433.
- Tang, Z.-Z. and Chen, G. (2018). Zero-inflated generalized dirichlet multinomial regression model for microbiome compositional data analysis. *Biostatistics*.
- Taylor, V. H. (2019). The microbiome and mental health: Hope or hype? *Journal of psychiatry & neuroscience: JPN*, 44(4):219.
- Thursby, E. and Juge, N. (2017). Introduction to the human gut microbiota. *Biochemical Journal*, 474(11):1823–1836.
- Tikhonov, M., Leach, R. W., and Wingreen, N. S. (2015). Interpreting 16s metagenomic data without clustering to achieve sub-otu resolution. *The ISME journal*, 9(1):68–80.
- Tseng, C.-H. and Wu, C.-Y. (2019). The gut microbiome in obesity. *Journal of the Formosan Medical Association*, 118:S3–S9.

- Turley, P., Walters, R. K., Maghzian, O., Okbay, A., Lee, J. J., Fontana, M. A., Nguyen-Viet, T. A., Wedow, R., Zacher, M., Furlotte, N. A., et al. (2018). Multi-trait analysis of genome-wide association summary statistics using mtag. *Nature genetics*, 50(2):229–237.
- Turnbaugh, P. J., Hamady, M., Yatsunenko, T., Cantarel, B. L., Duncan, A., Ley, R. E., Sogin, M. L., Jones, W. J., Roe, B. A., Affourtit, J. P., et al. (2009). A core gut microbiome in obese and lean twins. *nature*, 457(7228):480–484.
- Turnbaugh, P. J., Ley, R. E., Hamady, M., Fraser-Liggett, C. M., Knight, R., and Gordon, J. I. (2007). The human microbiome project. *Nature*, 449(7164):804–810.
- van Rheenen, W., Peyrot, W. J., Schork, A. J., Lee, S. H., and Wray, N. R. (2019). Genetic correlations of polygenic disease traits: from theory to practice. *Nature Reviews Genetics*, 20(10):567–581.
- Vassy, J. L., Hivert, M.-F., Porneala, B., Dauriz, M., Florez, J. C., Dupuis, J., Siscovick, D. S., Fornage, M., Rasmussen-Torvik, L. J., Bouchard, C., et al. (2014). Polygenic type 2 diabetes prediction at the limit of common variant detection. *Diabetes*, 63(6):2172–2182.
- Vilhjálmsdóttir, B. J., Yang, J., Finucane, H. K., Gusev, A., Lindström, S., Ripke, S., Genovese, G., Loh, P.-R., Bhatia, G., Do, R., et al. (2015). Modeling linkage disequilibrium increases accuracy of polygenic risk scores. *The american journal of human genetics*, 97(4):576–592.
- Visser, P. M., Medland, S. E., Ferreira, M. A., Morley, K. I., Zhu, G., Cornes, B. K., Montgomery, G. W., and Martin, N. G. (2006). Assumption-free estimation of heritability from genome-wide identity-by-descent sharing between full siblings. *PLoS genetics*, 2(3):e41.
- Visser, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A., and Yang, J. (2017). 10 years of gwas discovery: biology, function, and translation. *The American Journal of Human Genetics*, 101(1):5–22.

- Vizmanos, J. L., Cross, N. C., and Novo, F. J. (2020). The nature of traits, genes and variation. In *Principles of Nutrigenetics and Nutrigenomics*, pages 3–9. Elsevier.
- Vujkovic-Cvijin, I., Dunham, R. M., Iwai, S., Maher, M. C., Albright, R. G., Broadhurst, M. J., Hernandez, R. D., Lederman, M. M., Huang, Y., Somsouk, M., et al. (2013). Dysbiosis of the gut microbiota is associated with hiv disease progression and tryptophan catabolism. *Science translational medicine*, 5(193):193ra91–193ra91.
- Wade, K. H. and Hall, L. J. (2019). Improving causality in microbiome research: can human genetic epidemiology help? *Wellcome Open Research*, 4(199):199.
- Wadsworth, W. D., Argiento, R., Guindani, M., Galloway-Pena, J., Shelburne, S. A., and Vannucci, M. (2017). An integrative bayesian dirichlet-multinomial regression model for the analysis of taxonomic abundances in microbiome data. *BMC bioinformatics*, 18(1):94.
- Wakefield, J., Gelfand, A., and Smith, A. (1991). Efficient generation of random variates via the ratio-of-uniforms method. *Statistics and Computing*, 1(2):129–133.
- Wang, B., Yao, M., Lv, L., Ling, Z., and Li, L. (2017). The human microbiota in health and disease. *Engineering*, 3(1):71–82.
- Wang, J., Thingholm, L. B., Skiecevičienė, J., Rausch, P., Kummen, M., Hov, J. R., Degenhardt, F., Heinsen, F.-A., Rühlemann, M. C., Szymczak, S., et al. (2016). Genome-wide association analysis identifies variation in vitamin d receptor and other host factors influencing the gut microbiota. *Nature genetics*, 48(11):1396.
- Wang, S., Ryan, C. A., Boyaval, P., Dempsey, E. M., Ross, R. P., and Stanton, C. (2019a). Maternal vertical transmission affecting early-life microbiota development. *Trends in microbiology*.
- Wang, Y., Randolph, T., Shojaie, A., and Ma, J. (2019b). The gmd-biplot and its application to microbiome data. *BioRxiv*, page 814269.

- Wang, Y., Shojaie, A., Randolph, T., and Ma, J. (2019c). Generalized matrix decomposition regression: Estimation and inference for two-way structured data.
- Wei, Y., Li, Y., Yan, L., Sun, C., Miao, Q., Wang, Q., Xiao, X., Lian, M., Li, B., Chen, Y., et al. (2020). Alterations of gut microbiome in autoimmune hepatitis. *Gut*, 69(3):569–577.
- Wei, Z., Wang, W., Bradfield, J., Li, J., Cardinale, C., Frackelton, E., Kim, C., Mentch, F., Van Steen, K., Visscher, P. M., et al. (2013). Large sample size, wide variant spectrum, and advanced machine-learning technique boost risk prediction for inflammatory bowel disease. *The American Journal of Human Genetics*, 92(6):1008–1012.
- Weinberg, W. (1909). Über vererbungsgesetze beim menschen ii. *Z. indukt. Abstamm. Vererbungs*, 2:276–330.
- Weinberger, D. R. (2002). Biological phenotypes and genetic research on schizophrenia. *World Psychiatry*, 1(1):2.
- Weiss, S., Xu, Z. Z., Peddada, S., Amir, A., Bittinger, K., Gonzalez, A., Lozupone, C., Zaneveld, J. R., Vázquez-Baeza, Y., Birmingham, A., et al. (2017). Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome*, 5(1):27.
- Whitlock, M. C. and Bürger, R. (2004). Fixation of new mutations in small populations.
- Winter, G., Hart, R. A., Charlesworth, R. P., and Sharpley, C. F. (2018). Gut microbiome and depression: what we know and what we need to know. *Reviews in the Neurosciences*.
- Witte, J. S. (2010). Genome-wide association studies and beyond. *Annual review of public health*, 31:9–20.

- Wray, N. R. (2005). Allele frequencies and the r^2 measure of linkage disequilibrium: impact on design and interpretation of association studies. *Twin Research and Human Genetics*, 8(2):87–94.
- Wright, A. F. (2001). Genetic variation: polymorphisms and mutations. *e LS*.
- Wu, C., Chen, J., Kim, J., and Pan, W. (2016a). An adaptive association test for microbiome data. *Genome medicine*, 8(1):56.
- Wu, J., Peters, B. A., Dominianni, C., Zhang, Y., Pei, Z., Yang, L., Ma, Y., Purdue, M. P., Jacobs, E. J., Gapstur, S. M., et al. (2016b). Cigarette smoking and the oral microbiome in a large study of american adults. *The ISME journal*, 10(10):2435.
- Xiao, J. and Chen, J. (2017). Phylogeny-based kernels with application to microbiome association studies. In *New Advances in Statistics and Data Science*, pages 217–237. Springer.
- Xiao, J., Chen, L., Johnson, S., Zhang, X., and Chen, J. C. (2018a). Predictive modeling of microbiome data using a phylogeny-regularized generalized linear mixed model. *Frontiers in microbiology*, 9:1391.
- Xiao, J., Chen, L., Yu, Y., Zhang, X., and Chen, J. (2018b). A phylogeny-regularized sparse regression model for predictive modeling of microbial community data. *Frontiers in microbiology*, 9.
- Xu, L., Paterson, A. D., Turpin, W., and Xu, W. (2015). Assessment and selection of competing models for zero-inflated microbiome data. *PloS one*, 10(7):e0129606.
- Xu, T., Demmer, R. T., and Li, G. (2019). Zero-inflated poisson factor model with application to microbiome absolute abundance data. *arXiv preprint arXiv:1910.11985*.
- Yang, J., Bakshi, A., Zhu, Z., Hemani, G., Vinkhuyzen, A. A., Lee, S. H., Robinson, M. R., Perry, J. R., Nolte, I. M., van Vliet-Ostaptchouk, J. V., et al. (2015). Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. *Nature genetics*, 47(10):1114.

- Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., Madden, P. A., Heath, A. C., Martin, N. G., Montgomery, G. W., et al. (2010). Common snps explain a large proportion of the heritability for human height. *Nature genetics*, 42(7):565.
- Yang, J., Lee, S. H., Goddard, M. E., and Visscher, P. M. (2011). Gcta: a tool for genome-wide complex trait analysis. *The American Journal of Human Genetics*, 88(1):76–82.
- Yang, J., Zeng, J., Goddard, M. E., Wray, N. R., and Visscher, P. M. (2017). Concepts, estimation and interpretation of snp-based heritability. *Nature genetics*, 49(9):1304.
- Yin, W., Gao, C., Xu, Y., Li, B., Ruderfer, D. M., and Chen, Y. (2018). Learning opportunities for drug repositioning via gwas and phewas findings. *AMIA Summits on Translational Science Proceedings*, 2018:237.
- Young, A. I. (2019). Solving the missing heritability problem. *PLoS genetics*, 15(6):e1008222.
- Young, V. B. (2017). The role of the microbiome in human health and disease: an introduction for clinicians. *Bmj*, 356:j831.
- Zaitlen, N., Kraft, P., Patterson, N., Pasaniuc, B., Bhatia, G., Pollack, S., and Price, A. L. (2013). Using extended genealogy to estimate components of heritability for 23 quantitative and dichotomous traits. *PLoS genetics*, 9(5):e1003520.
- Zeng, X., Zhu, S., Hou, Y., Zhang, P., Li, L., Li, J., Huang, L. F., Lewis, S. J., Nussinov, R., and Cheng, F. (2020). Network-based prediction of drug–target interactions using an arbitrary-order proximity embedded deep forest. *Bioinformatics*.
- Zhang, G. (2015). Genetic architecture of complex human traits: What have we learned from genome-wide association studies? *Current Genetic Medicine Reports*, 3(4):143–150.

- Zhang, X., Mallick, H., Tang, Z., Zhang, L., Cui, X., Benson, A. K., and Yi, N. (2017). Negative binomial mixed models for analyzing microbiome count data. *BMC bioinformatics*, 18(1):4.
- Zhao, N., Chen, J., Carroll, I. M., Ringel-Kulka, T., Epstein, M. P., Zhou, H., Zhou, J. J., Ringel, Y., Li, H., and Wu, M. C. (2015). Testing in microbiome-profiling studies with mirkat, the microbiome regression-based kernel association test. *The American Journal of Human Genetics*, 96(5):797–807.
- Zhu, Z., Bakshi, A., Vinkhuyzen, A. A., Hemani, G., Lee, S. H., Nolte, I. M., van Vliet-Ostaptchouk, J. V., Snieder, H., Esko, T., Milani, L., et al. (2015). Dominance genetic variation contributes little to the missing heritability for human complex traits. *The American Journal of Human Genetics*, 96(3):377–385.
- Zitvogel, L., Ma, Y., Raoult, D., Kroemer, G., and Gajewski, T. F. (2018). The microbiome in cancer immunotherapy: Diagnostic tools and therapeutic strategies. *Science*, 359(6382):1366–1370.
- Zoetendal, E. G., Akkermans, A. D., Akkermans-van Vliet, W. M., de Visser, J. A. G., and de Vos, W. M. (2001). The host genotype affects the bacterial community in the human gastrointestinal tract. *Microbial ecology in health and disease*, 13(3):129–134.
- Zuk, O., Hechter, E., Sunyaev, S. R., and Lander, E. S. (2012). The mystery of missing heritability: Genetic interactions create phantom heritability. *Proceedings of the National Academy of Sciences*, 109(4):1193–1198.

Appendices

A.1 Simulating genotypes, OTU scores, and phenotype

Simulating genotypes. Using real genotype data of the European population obtained from the 1000 Genomes Phase 3 reference data, together with its LD pattern, we utilize HapGen2 software to simulate genotypes of 1000 individuals, based on chromosome 10 (arbitrarily chosen) of the reference haplotype. The chromosome contains 73,832 SNPs, making it adequate for model testing.

Simulating OTUs. We simulate 200 OTUs for the 1000 individuals. We simulate OTU counts using the DM distribution with the parameter (mean and dispersion) estimated from real microbiome data. Then, for each OTU sample, the simulated counts are normalized by dividing by the total read count; where the total read count for each sample is obtained by sampling from a negative binomial distribution with mean 5000 and dispersion 25, which reflects typical sequencing depths in empirical studies. From these OTUs, we randomly choose 20 to be putatively causal, setting their effect sizes as $\beta \in \{0.1, 2\}$ and 0 for the non-causal ones. Then, the OTU scores, S_i , are computed using $S_i = \sum_{j \in \mathcal{A}} \beta_j d_{ij}$, d_{ij} is the abundance of the putative causal OTUs discretized to the range 0, 1, 2.

Simulating phenotypes. With the simulated genotypes and OTU scores, the case-control statuses of the individuals is generated from the following model:

$$\begin{aligned} \lambda_1^{(g)}, \lambda_2^{(g)}, \dots, \lambda_{73832}^{(g)} &\in \{0, 1\} ; & \sum_{l=1}^{73832} \lambda_l^{(g)} &= 2 \\ \lambda_1^{(m)}, \lambda_2^{(m)}, \dots, \lambda_{200}^{(m)} &\in \{0, 1\} ; & \sum_{l=1}^{200} \lambda_l^{(m)} &= 20 , \\ \text{logit} [P(y_i = 1 | \mathbf{x}_i, \mathbf{z}_i)] &= \mu + \sum_{j=1}^{73832} \lambda_j^{(g)} \beta_j x_{ij} + \sum_{k=1}^{200} \lambda_k^{(m)} S_{ki} , \end{aligned}$$

where x_{ij} is the j^{th} SNP on the i^{th} individual, $(\cdot)^{(g)}$ and $(\cdot)^{(m)}$ denote genotype

and microbiome parameters respectively, and μ is set to 0.5.

A.2 Scripts to simulate genotypes, OTU scores, and phenotypes and then perform GWAS analysis

script simOTU.R: simulate OTUs

```

1 # -----
2 # Simulating Microbiome Abundance Data
3 # -----
4 library(dirmult) # for simulating from DM distribution.
5 library(GUniFrac) # contains throat data (real data) to be used to obtain
   realistic sim params.
6 library(picante) # to prune phylogenetic tree.
7
8 # Real data for simulation parameters
9 data(throat.otu.tab)
10 data(throat.tree)
11
12 n <- 1000 # number of samples to simulate
13 m <- 200 # number of OTUs to simulate
14
15 # Only 200 most abundant taxa were used for parameter estimation.
16 otu.pruned <- throat.otu.tab[order(colSums(throat.otu.tab))[1:m]]
17 otu.pruned <- otu.pruned[rowSums(otu.pruned) !=0,]
18 DM.params <- dirmult(data=otu.pruned,epsilon=10^(-4), trace=TRUE) #obtain 'pi'
   and 'theta'.
19
20 # simulate OTU counts
21 otu.counts <- simPop(J=n, # number of samples
   n=m, # number of OTUs
22   pi=DM.params$pi,
23   theta=DM.params$theta)$data
24
25 # Read counts sampled from negative binom with mean 5000 and dispersion 25:
26 #rnbinom(n=1,mu=5000,size=25); and normalize the counts.
27 otu.abun <- as.data.frame(apply(otu.counts,2,
28   function(x){return(x/rnbinom(n=1,mu=5000,size=25)
29     )}) ));
30 colnames(otu.abun)=paste0("OTU",1:ncol(otu.counts))
31

```

```

# -----
33 write.table(as.data.frame(otu.abun), file="simulatedOTUs.txt", row.names=FALSE)
# -----

```

script simPhenotype.py: simulate OTUs scores and phenotype

```

=====
2 # Processing genotype data
##=====
4 import pandas as pd
import numpy as np
6
genotype=pd.read_table("genotypes_chr10_hapgen.controls.gen", sep=' ', header=
    None)
8
#Select only genotype data [minor allele, assuming dominant model] for the
    purpose
10 #of simulating the phenotype.

12 # The loaded genotype data is SNP x sample info. Extract only genotype info[index
    5 onwards
    # Then, transpose the data to have SNPs in columns (for convenience)
14 genotype_val=genotype.iloc[:,5::3]
    genotype_val=genotype_val.T
16
n=genotype_val.shape[0] # number of samples
18 m=genotype_val.shape[1] # number of SNPs

20 #Set effects for 2 SNPs that will be assumed to be 'associated'
    causal_snp_effects=np.array([1.2,2.0])
22
#location of the causal SNPs
24 loc_causal_snp=[6,83]
26
##=====
28 # Processing OTU (microbiome) data
##=====
30 otu_data=pd.read_table('simulatedOTUs.txt',header=0,sep=' ')

32 # Discretize OTU abundance.
    # to 3 bins: 'low', 'medium', 'high'
34 func=lambda x: pd.cut(x,bins=3,labels=False)
    otu_data_discreized=otu_data.apply(func,axis=1) # along columns (OTUs)
36
# Compute individual-level scores for the 'causal' OTUs.
38 #Set effects for 20 OTUs that are assumed to be 'associated' with the phenotype.
    causal_otu_effects=np.array

```

```

    ([0.1,0.4,0.5,0.9,0.9,0.9,1.0,1.0,1.0,1.2,1.3,1.4,1.4,1.6,1.7,1.7,1.8, 1.8,
    1.8,2.0]))
40
#location of the causal SNPs
42 loc_causal_otu=np.array
    ([12,21,25,45,49,55,58,62,73,105,118,136,150,159,161,165,167,171,186,187]))

44 # Compute OTU scores for individuals, for the associated OTUs.
otu_score=np.dot(otu_data_discreized.iloc[:,loc_causal_otu],causal_otu_effects)
46
# Write OTU scores to file.
48 otu_score1=pd.Series(otu_score)
otu_score1.to_csv('otu_score.txt', index=False,sep='\t',header=True)
50
##=====
52 # Simulating phenotype.
##=====
54 #z=intercept + X*beta + otu_score
z=-0.5+ np.dot(genotype_val.iloc[:,loc_causal_snp],causal_snp_effects) +
    otu_score
56 pr=np.exp(z)/(1+np.exp(z))
y=np.random.binomial(np.repeat(1,n),pr)
58 #print(y)
y=pd.Series(y)
60 y.to_csv('phenotype.txt', index=False,sep='\t',header=True)
# -----
62
64 # END OF SCRIPT

```

script Rplot.R: plots QQ plot from GWAS p values

```

library(GWASTools)
2
## The GEMMA tool create a 'output' directory and writes results into it.
4 # Read the data with association pvalues
pvals1=read.table('./output/results1_chr10_cov_pvalues.txt')[,2]
6 pvals2=read.table('./output/results2_chr10_nocov_pvalues.txt')[,2]

8 # Calculate genomic inflation factor
chisq1=qchisq(1-pvals1,1)
10 GC1=median(chisq1)/qchisq(0.5,1)

12 chisq2=qchisq(1-pvals2,1)
GC2=median(chisq2)/qchisq(0.5,1)
14 # -----
16 # QQ plots

```

```

jpeg("./output/qqPlots.jpg", width=1200,height=800) # or pdf("qqPlots.pdf")
18 par(mfrow=c(1,2))
qqPlot(pvals1, truncate=FALSE,main="Model with OTU scores",cex=2)
20 text(5,3, paste('GC =',round(GC1,3)),cex=2)
qqPlot(pvals2, truncate=FALSE,main="Model without OTU scores",cex=2)
22 text(5,3, paste('GC =',round(GC2,3)),cex=2)
dev.off()
24 # -----

26 # Table for p values of SNPs at the causal SNPs
snps1=read.table('./output/results1_chr10_cov_pvalues.txt')
28 snps2=read.table('./output/results2_chr10_nocov_pvalues.txt')
#tab=data.frame(cbind(snps1[c(6,31,83,56,101),],snps2[c(6,31,83,56,101),2])) #
30 tab=as.data.frame(cbind(snps1[1:20,],snps2[1:20,2])) # p.vals for first 20 SNPs
write.table(tab, './output/pvalsCausalSNPs.txt')

```

analysis.sh: overall script that calls the simOTU.R plus simPhenotype.py plus Rplot.R and performs GWAS analysis

```

#!/bin/bash
2

4 work_dir=/home/denaw/Documents/Others/SIMULATION/simGenotype
ref_data=/home/denaw/Documents/Others/SIMULATION/RefData
6

8 #####
# SIMULATE GENOTYPES; using 'hapgen' tool
10 #####

12 for chr in 10 # 'seq 10 22' # we'll use only chromosome 10 for this analysis.
do
14     #Can't get hapgen to work without the -dl flag (should be possible since
    #v2.0.2) so create dummy variable, that can be passed on
16
    dummyDL='sed -n '2'p $ref_data/hapmap3.r2.b36.chr${chr}.legend | cut -d
        ' ' -f 2'
18     #dummyDL='sed -n '2'p $ref_data/CEU.0908.chr${chr}.legend | cut -d ' ' -
        f 2'
    ./hapgen2 -m $ref_data/genetic_map_chr${chr}_combined_b36.txt \
20     -l $ref_data/hapmap3.r2.b36.chr${chr}.legend \
    -h $ref_data/CEU.chr${chr}.hap \
22     -o genotypes_chr${chr}_hapgen \
    -n 1000 0 \
24     -dl $dummyDL 0 0 0
    #-no_haps_output
26
done

```

```

28 ##-----
30
32 ##=====
33 # FILTER (QC) THE THE SIMULATED GENOTYPES; use 'PLINK' and/or 'gtool'
34 ##=====
35 # First, the simulated data is in .gen and .sample format; convert to plink
36 # format, then
37 # prune the genotypes.
38 for chr in 10
39 do
40     # issues with generating .bim and .ped files using this plink; so we'll
41     # se gtool to achieve this.
42     #./plink --data genotypes_chr${chr}_hapgen.controls \
43     #--oxford-single-chr $chr \
44     #--make-bed \
45     #--out genotypes_chr${chr}_hapgen.controls
46     ./gtool -G --g genotypes_chr${chr}_hapgen.controls.gen \
47     --s genotypes_chr${chr}_hapgen.controls.sample \
48     --ped genotypes_chr${chr}_hapgen.controls.ped \
49     --map genotypes_chr${chr}_hapgen.controls.map --chr ${chr}
50
51     #the gtool above does not generate the needed .fam file; so, we'll use
52     #plink for it.
53     ./plink --file genotypes_chr${chr}_hapgen.controls --make-bed --out
54     genotypes_chr${chr}_hapgen.controls
55
56     #prune variants to produce markers that are in approximate LD with each
57     #other
58     #parameters: window size in variant count or kilobase, variant count to
59     #shift window at the end of each step, and
60     #pairwise r2 threshold.
61     ./plink --bfile genotypes_chr${chr}_hapgen.controls \
62     --indep-pairwise 50kb 5 1 \
63     --out genotypes_chr${chr}_hapgen.controls
64
65     # Do not set ambiguously-sexed individuals missing.
66     ./plink --bfile genotypes_chr${chr}_hapgen.controls --allow-no-sex --make
67     -bed --out genotypes_chr${chr}_hapgen.controls
68
69 done
70 ##-----
71
72 ##=====

```

```

# SIMULATE OTUs (Microbiome); using the R script 'simOTU.R' tool
70 ##=====
Rscript simOTU.R
72 ##-----

74
##=====
76 # COMPUTE OTU SCORES, AND PHENOTYPE; using the python script 'simPhenotype.py'
##=====
78 # Compute OTU scores for the individuals; set 'causal' OTUs and SNPs; and
# simulate case-control statuses of the individuals.
80 python3 simPhenotype.py
##-----

82
84
##=====
86 # PERFORM GWAS ANALYSIS; using 'GEMMA' tool
##=====
88 # GEMMA requires that phenotypes, kinship and non-genetic covariates be in csv
and GEMMA-specific format
# Genotype file need be in BIMBAM format (obtainable by Bash script below); s (
below) is number of samples.
90
for chr in 10
92 do
cat $work_dir/genotypes_chr${chr}_hapgen.controls.gen | awk -v s=1000 \
94 '{ printf $2 ", " $4 ", " $5; for(i=1; i<=(s+1); i++) printf ", " $(i*3+3)
*2+$(i*3+4); printf "\n"}' \
> $work_dir/genotypes_chr${chr}_hapgen.controls.gemma

96
# Estimate relatedness matrix from genotype
98 ./gemma-0.98.1-linux-static -g $work_dir/genotypes_chr${chr}_hapgen.
controls.gemma -p $work_dir/phenotype.txt -gk 2 -o kinship

100
# And finally, perform association using GEMMA (applying LMM)
#./gemma-0.98.1-linux-static -g $work_dir/genotypes_chr${chr}_hapgen.
controls.gemma -p $work_dir/phenotype.txt -c $work_dir/otu_score.txt
-lm 2 -o results1_chr${chr}_cov
102 ./gemma-0.98.1-linux-static -g $work_dir/genotypes_chr${chr}_hapgen.
controls.gemma -p $work_dir/phenotype.txt -k $work_dir/output/kinship
.sXX.txt -c $work_dir/otu_score.txt -lmm 2 -o results1_chr${chr}_cov
./gemma-0.98.1-linux-static -g $work_dir/genotypes_chr${chr}_hapgen.
controls.gemma -p $work_dir/phenotype.txt -k $work_dir/output/kinship
.sXX.txt -lmm 2 -o results2_chr${chr}_nocov
104 #./gemma-0.98.1-linux-static -g $work_dir/genotypes_chr${chr}_hapgen.
controls.gemma -p $work_dir/phenotype.txt -lm 2 -o results2_chr${chr}
_nocov

```

```

done
106 ##-----
108
108 ##=====
110 # PROCESS OUTPUT FILES TO STUDY PERFORMANCE
110 ##=====
112 # We extract the columns of interest for power analysis (simply based on p-values
    ). The last
    # column contains the p-values of the analyses, the second column contains the
    genetic variant identifier.
114 # We extract these columns and write to a file.

116 for chr in 10
do
118     columns_LM_cov='head -1 $work_dir/output/results1_chr${chr}_cov.assoc.txt
        |wc -w'

120     tail -n +2 $work_dir/output/results1_chr${chr}_cov.assoc.txt | cut \
        -f 2,$columns_LM_cov >> $work_dir/output/results1_chr${chr}_cov_pvalues.
        txt

122     columns_LM_nocov='head -1 $work_dir/output/results2_chr${chr}_nocov.assoc
        .txt|wc -w'

124     tail -n +2 $work_dir/output/results2_chr${chr}_nocov.assoc.txt | cut \
126     -f 2,$columns_LM_nocov >> $work_dir/output/results2_chr${chr}
        _nocov_pvalues.txt

done
128 ##-----
130
132 ##=====
132 # PLOT RESULTS; using the R script 'Rplots.R'
134 ##=====
134 # Use R (the 'GWASTools') to plot [QQ plot] the results

136 Rscript Rplot.R
138 ##-----

```

B. R script to implement the phyloDPM model

```

R script for the phyloDPM model

# *****#
2 #
# SCRIPT: phyloDPM.R
4 # AUTHOR: Denis Awany; Emile R. Chimusa
# DATE: 30/03/2020
6 # REV: 1.0
# VERSION: v1.0
8 # PLATFORM: Not Platform Dependent
#
10 # PURPOSE: This script is used to analyse microbiome data.
#           It is based on the logistic Dirichlet process random effects model.
12 #           Produces Bayesian posterior estimates for OTU effects.

14
# =====
16 # BEGINNING OF MAIN
# =====
18 phyloDPM.main <- function(Y, #case-control phenotype [vector]
                           Z, # microbiome abundance data [matrix or dataframe
                              object]
                           mTree, #phylogenetic tree[matrix or phylo class object]
                           Nsim, #number of MCMC iterations [integer]
                           thin, # thinning value for posterior samples [integer]
                           burnin, #number of burn-in samples [integer]
                           post.plots, # whether or not to display plots of
24                                posterior samples
                           mu, # tuning parameter for phylogenetic and RBF kernels
                           .
26                                alpha.par, #first hyper-parameter for Gamma prior of
                                precision parameter
                                sig2.bet.par, #variance parameter on beta
                                tau2.par, # variance parameter on the base parameter of
28                                the DP.
                                rbf.par, # kernel coefficient for the RBF kernel.
                                saveSim, # save raw simulated samples to a file in the
30                                current directory.
                                evol.rate, # evolutionary rate parameter for the
                                phylogenetic kernel.
                                otu.corr, #logical; whether or not to consider
32                                correlation among OTUs.
                                ...){
34
start.time = proc.time()[3] # time the duration of the code execution.

```

```

36 | welcome=function()
38 | {
    |   message("\n");
40 |   message("*****");
    |   message("                phyloDPM v1.0                ");
42 |   message("                ");
    |   message("                Awany Denis; Emile R. Chimusa                ");
44 |   message("                Human Genetics, Dept. of Pathology                ");
    |   message("                University of Cape Town                ");
46 |   message("*****");
    |   message("\n");
48 | }

50 | welcome()

52 |
    | # -----
54 | #               Check user inputs
    | # -----
56 |
    | if(missing(Y) | missing (Z)) {
58 |   stop("Response, and OTU data are required!")}
    | if(!is.vector(Y)){
60 |   stop('Response must be a vector.')}
    | if(class(Z)!='data.frame' & class(Z)!='matrix'){
62 |   stop('OTU data must be given as a dataframe or matrix.')}
    | if(length(Y) != nrow(Z)) {
64 |   stop("Response vector must have the same length as the microbiome sample size
    |       !")}
    | if(!is.finite(mu) | !is.numeric(mu) | mu<0 | mu>1){stop("invalid entry: mu must
    |       be a numeric between 0 and 1.")}
66 | if(is.null(mTree)){
    |   message("Phylogenetic tree data is missing; taxa similarity will be based on
    |       the radial basis function.")
68 |   if(mu >0){message("Warning: mu ignored, as mTree is missing.")}
    | }else{
70 |   if(class(mTree)!="phylo" & class(mTree)!="matrix"){
    |     stop("mTree must be a distance matrix or an object of the class 'phylo' ")
72 |   if(mu==1){message("Warning: radial basis function information will not be
    |       used, as 'mu' is set to 1.")}
    |   if(mu==0){message("Warning: phylogenetic tree information will not be used,
    |       as 'mu' is set to 0.")}
74 |   }
    | if (thin >= Nsim ) {stop("Error: thin cannot be equal to or greater than the
    |       number of posterior draws (Nsim).")}
76 | if (thin < 1) {stop("Error: thin should be greater than or equal to 1.")}
    | if(is.null(burnin)){

```

```

78     burnin = round(0.1 * Nsim) # if burn-in not supplied, set it to 10% of number
      of posterior draws (Nsim).
    }else{
80     if(burnin < 0){stop("Error! burnin should be a positive number.") }
      if(burnin >= Nsim){stop("Error! burnin should be less than the number of
        posterior draws (Nsim).") }
82   }
    if(!is.finite(evol.rate)| !is.numeric(evol.rate) | evol.rate < 0){stop("Error: '
      evol.rate' parameter must be a numeric, positive value.")]
84
    message("\nInput data succesfully verified ... \n")
86
88   # Load required packages
    if(post.plots==TRUE) {
90     message("The 'coda' package is required for some posterior plots. It will be
      installed and/or loaded.")
      if (!require("coda")) { # Needed for Bayesian posterior plots
92     install.packages("coda")
      library(coda)
94     }
    }
96
    # =====
98     #             DEFINE FUNCTIONS HERE
    # =====
100
    # thinning Bayesian posterior samples. From boom 20/07/2019.
102 thin.samples <- function (x, thin) {
    ## Thins the data in x, so that every thin'th observation is
104     ## returned. This is useful for obtaining aprximately uncorrelated
    ## posterior samples from Bayesian posterior.
106     ##
    ## Args:
108     ##   x: A numeric vector, matrix, or array to be thinned. If a
    ##       matrix or array then the first dimension corresponds to the
110     ##       observation number, and will be used for thinning.
    ##   thin: The frequency with which to thin. E.g. if thin == 10
112     ##       then every 10th observation will be returned.
    ##
114     ## Returns:
    ##   The thinned subset of x.
116     stopifnot(is.numeric(thin) && length(thin) == 1)
    if (thin <= 1)
118       return(x)
    if (is.array(x)) {
120       nr <- dim(x)[1]
    } else if (is.numeric(x)) {

```

```

122     nr <- length(x)
    } else stop("x must be a numeric type in thin()")
124
    top <- floor(nr/thin)
126     indx <- (1:top) * thin

128     if (is.matrix(x)) {
        return(x[indx, ])
130     } else if (is.numeric(x)) {
        return(x[indx])
132     } else if (is.array(x)) {
        stop("Error! you cannot drop all samples.")
134     }
    }
136
    # rdirichlet from LearnBayes (24/03/2019)
138     rdirichlet <- function(n, alpha)
    ## sampling from Dirichlet process using the Polya Urn scheme.
140     ##
    ## Args:
142     ##   n: number of samples; likened to number of balls in the Urn
    ##   alpha: concentration (precision) parameter.
144     ##
    ## Returns:
146     ##   samples distributed as a DP.
    {
148         k = length(alpha)
        z = array(0, dim = c(n, k))
150         s = array(0, dim = c(n, 1))
        for (i in 1:k) {
152             z[, i] = rgamma(n, shape=alpha[i])
            s = s + z[, i]
154         }
        for (i in 1:k) {
156             z[, i] = z[, i]/s
        }
158         return(z)
    }
160
    #Posterior predictive p-value, adapted from bayespval func in 'contig' package
    by Antony O.
162     # 21/07/2019.
    bayespval <- function(Y, Z, Beta, statistic="X2"){
164         MU <- exp(as.matrix(Z )%*% as.vector(Beta))
        Pred <- matrix(rbinom(n=prod(dim(MU)), size=1, prob=MU/(1+MU)),ncol=dim(MU)
            [2])
166         Obs <- Y

```

```

168   Tpred<-apply(((Pred-MU)^2)/MU,1,sum)
169   Tobs<-apply(((Obs-MU)^2)/MU,1,sum)
170
171   pval<-mean(as.numeric(Tpred>Tobs))
172   pval
173 }
174
175 HPDinterval.Bayesian <- function(obj, prob = 0.95, ...) {
176   obj <- as.matrix(obj)
177   vals <- apply(obj, 2, sort)
178   if (!is.matrix(vals)) stop("obj must have nsamp > 1")
179   nsamp <- nrow(vals)
180   npar <- ncol(vals)
181   gap <- max(1, min(nsamp - 1, round(nsamp * prob)))
182   init <- 1:(nsamp - gap)
183   inds <- apply(vals[init + gap, ,drop=FALSE] - vals[init, ,drop=FALSE],
184                 2, which.min)
185   ans <- cbind(vals[cbind(inds, 1:npar)],
186               vals[cbind(inds + gap, 1:npar)])
187   dimnames(ans) <- list(colnames(obj), c("lower", "upper"))
188   attr(ans, "Probability") <- gap/nsamp
189   ans
190 }
191
192 # cophenetic.phylo from 'ape' (06/04/2019)
193 dist.nodes <- function(x)
194 {
195   x <- reorder(x) # required for the C code
196   n <- Ntip(x)
197   m <- x$Nnode
198   nm <- n + m
199   d <- .C("dist_nodes", as.integer(n), as.integer(m), # "dist_nodes", from a
200           C file named 'dist_nodes.c'.
201           as.integer(x$edge[, 1] - 1L), as.integer(x$edge[, 2] - 1L),
202           as.double(x$edge.length), as.integer(Nedge(x)),
203           double(nm * nm), NAOK = TRUE)[[7]]
204   dim(d) <- c(nm, nm)
205   dimnames(d) <- list(1:nm, 1:nm)
206   d
207 }
208
209 cophenetic.phylo <- function(x)
210 {
211   n <- length(x$tip.label)
212   ans <- dist.nodes(x)[1:n, 1:n]
213   dimnames(ans)[1:2] <- list(x$tip.label)
214   ans
215 }

```

```

216 # rbfkernel from rdetools (02/04/2019)
rbfkernel <- function(X, sigma = 1, Y = NULL)
218 {
    # test if X is a matrix
220     if(!is.matrix(X))
    {
222         print("X must be a matrix containing samples in its rows")
        return()
224     }
    # test if sigma is a number and > 0
226     if(length(sigma) != 1 || sigma <= 0)
    {
228         print("sigma must be a number > 0 specifying the rbf-kernel width")
        return()
230     }
    if(!is.null(Y))
232     {
        # test if Y is a matrix
234         if(!is.matrix(Y))
        {
236             print("Y must be a matrix containing samples in its rows or NULL if
                it should not be used")
            return()
238         }
        # test if vectors in X and Y have same dimension
240         if(ncol(X) != ncol(Y))
        {
242             print("The samples in the rows of X and Y must be of same dimension
                ")
            return()
244         }
        }
246
    n <- nrow(X) # number of samples in X
248
    if(is.null(Y))
    {
250        # calculate distance matrix
        XtX <- tcrossprod(X)
252        XX <- matrix(1, n) %*% diag(XtX)
        D <- XX - 2*XtX + t(XX) # distance matrix
254    }
    else
256    {
258        m <- nrow(Y) # number of samples in Y
        # calculate distance matrix (between vectors of X and Y)
260        XX <- matrix(apply(X ^ 2, 1, sum), n, m)

```

```

262     YY <- matrix(apply(Y ^ 2, 1, sum), n, m, byrow = TRUE)
263     XY <- tcrossprod(X, Y)
264     D <- XX - 2*XY + YY
265   }
266
267   # calculate rbf-kernel matrix
268   K <- exp(-D/(2*sigma^2))
269
270   return(K)
271 }
272
273 # rdiscrete from e1071 (24/03/2019)
274 rdiscrete <- function (n, probs, values = 1:length(probs), ...)
275 {
276   sample(values, size = n, replace = TRUE, prob = probs)
277 }
278
279 # rinvgamma from MCMCpack (24/03/2019)
280 rinvgamma <- function (n, shape, scale = 1)
281 {
282   return(1/rgamma(n = n, shape = shape, rate = scale))
283 }
284
285 # rmvnorm from mvtnorm (24/03/2019)
286 rmvnorm <- function (n, mean = rep(0, nrow(sigma)), sigma = diag(length(mean)),
287   method = c("eigen", "svd", "chol"))
288 {
289   if (!isSymmetric(sigma, tol = sqrt(.Machine$double.eps))) {
290     stop("sigma must be a symmetric matrix")
291   }
292   if (length(mean) != nrow(sigma)) {
293     stop("mean and sigma have non-conforming size")
294   }
295   sigma1 <- sigma
296   dimnames(sigma1) <- NULL
297   if (!isTRUE(all.equal(sigma1, t(sigma1)))) {
298     warning("sigma is numerically not symmetric")
299   }
300   method <- match.arg(method)
301   if (method == "eigen") {
302     ev <- eigen(sigma, symmetric = TRUE)
303     if (!all(ev$values >= -sqrt(.Machine$double.eps) * abs(ev$values[1])))
304     {
305       warning("sigma is numerically not positive definite")
306     }
307     retval <- ev$vectors %*% diag(sqrt(ev$values), length(ev$values)) %*%
308       t(ev$vectors)

```

```

308   }
    else if (method == "svd") {
310       sigsvd <- svd(sigma)
        if (!all(sigsvd$d >= -sqrt(.Machine$double.eps) * abs(sigsvd$d[1]))) {
312           warning("sigma is numerically not positive definite")
        }
314       retval <- t(sigsvd$v %*% (t(sigsvd$u) * sqrt(sigsvd$d)))
    }
316   else if (method == "chol") {
        retval <- chol(sigma, pivot = TRUE)
318       o <- order(attr(retval, "pivot"))
        retval <- retval[, o]
320   }
    retval <- matrix(rnorm(n * ncol(sigma)), nrow = n) %*% retval
322   retval <- sweep(retval, 2, mean, "+")
    colnames(retval) <- names(mean)
324   retval
  }
326
328 # rtnorm from msm (06/04/2019); truncated normal sampler
rtnorm <- function (n, mean = 0, sd = 1, lower = -Inf, upper = Inf)
330 {
    if (length(n) > 1)
332       n <- length(n)
    mean <- rep(mean, length = n)
334   sd <- rep(sd, length = n)
    lower <- rep(lower, length = n)
336   upper <- rep(upper, length = n)
    lower <- (lower - mean)/sd
338   upper <- (upper - mean)/sd
    ind <- seq(length = n)
340   ret <- numeric(n)
    alg <- ifelse(lower > upper, -1, ifelse(((lower < 0 & upper == Inf) | (
        lower ==
342         -Inf & upper > 0) | (is.finite(lower) & is.finite(upper) & (
            lower < 0) &
            (upper > 0) & (upper - lower > sqrt(2 * pi)))), 0, ifelse((
                lower >= 0 &
344         (upper > lower + 2 * sqrt(exp(1))/(lower + sqrt(lower^2 + 4))
            *
            exp((lower * 2 - lower * sqrt(lower^2 + 4))/4))), 1, ifelse(
                upper <= 0 &
346         (-lower > -upper + 2 * sqrt(exp(1))/(-upper + sqrt(upper^2 +
            4)) *
            exp((upper * 2 - -upper * sqrt(upper^2 + 4))/4))), 2, 3)))
348   ind.nan <- ind[alg == -1]
    ind.no <- ind[alg == 0]

```

```

350   ind.expl <- ind[alg == 1]
351   ind.expu <- ind[alg == 2]
352   ind.u <- ind[alg == 3]
353   ret[ind.nan] <- NaN
354   while (length(ind.no) > 0) {
355     y <- rnorm(length(ind.no))
356     done <- which(y >= lower[ind.no] & y <= upper[ind.no])
357     ret[ind.no[done]] <- y[done]
358     ind.no <- setdiff(ind.no, ind.no[done])
359   }
360   stopifnot(length(ind.no) == 0)
361   while (length(ind.expl) > 0) {
362     a <- (lower[ind.expl] + sqrt(lower[ind.expl]^2 + 4))/2
363     z <- rexp(length(ind.expl), a) + lower[ind.expl]
364     u <- runif(length(ind.expl))
365     done <- which((u <= exp(-(z - a)^2/2)) & (z <= upper[ind.expl]))
366     ret[ind.expl[done]] <- z[done]
367     ind.expl <- setdiff(ind.expl, ind.expl[done])
368   }
369   stopifnot(length(ind.expl) == 0)
370   while (length(ind.expu) > 0) {
371     a <- (-upper[ind.expu] + sqrt(upper[ind.expu]^2 + 4))/2
372     z <- rexp(length(ind.expu), a) - upper[ind.expu]
373     u <- runif(length(ind.expu))
374     done <- which((u <= exp(-(z - a)^2/2)) & (z <= -lower[ind.expu]))
375     ret[ind.expu[done]] <- -z[done]
376     ind.expu <- setdiff(ind.expu, ind.expu[done])
377   }
378   stopifnot(length(ind.expu) == 0)
379   while (length(ind.u) > 0) {
380     z <- runif(length(ind.u), lower[ind.u], upper[ind.u])
381     rho <- ifelse(lower[ind.u] > 0, exp((lower[ind.u]^2 -
382       z^2)/2), ifelse(upper[ind.u] < 0, exp((upper[ind.u]^2 -
383       z^2)/2), exp(-z^2/2)))
384     u <- runif(length(ind.u))
385     done <- which(u <= rho)
386     ret[ind.u[done]] <- z[done]
387     ind.u <- setdiff(ind.u, ind.u[done])
388   }
389   stopifnot(length(ind.u) == 0)
390   ret * sd + mean
391 }
392
393
394 # 'is.empty' from sjmisc: to check whether string, list or vector is empty
395   (09/04/2019)
396   is.empty <- function(x, first.only = FALSE) {
397     # do we have a valid vector?

```

```

398   if (!is.null(x)) {
# if it's a character, check if we have only one element in that vector
400   if (is.character(x)) {
# characters may also be of length 0
402     if (length(x) == 0) return(TRUE)
# else, check all elements of x
404     zero_len <- nchar(x) == 0
# return result for multiple elements of character vector
406     if (first.only) {
zero_len <- .is_true(zero_len[1])
408       if (length(x) > 0) x <- x[!is.na(x)][1]
} else {
410       return(unname(zero_len))
}
# we have a non-character vector here. check for length
412 } else if (is.list(x)) {
x <- purrr::compact(x)
414 zero_len <- length(x) == 0
} else {
416 zero_len <- length(x) == 0
}
418 }

420 any(is.null(x) || zero_len || all(is.na(x)))

422 }

424 # 'is.empty' from rapporttools: to check whether, vector value is empty
(12/04/2019)
is_empty <- function(x, trim = TRUE, ...) {
426   if (length(x) <= 1) {
if (is.null(x))
428   return (TRUE)
if (length(x) == 0)
430   return (TRUE)
if (is.na(x) || is.nan(x))
432   return (TRUE)
if (is.character(x) && nchar(ifelse(trim, trim.space(x), x)) == 0)
434   return (TRUE)
if (is.logical(x) && !isTRUE(x))
436   return (TRUE)
if (is.numeric(x) && x == 0)
438   return (TRUE)
return (FALSE)
440   } else
sapply(x, is_empty, trim = trim, ...)
442 }

```

```

444 # con_par; concentration parameter of DP (29/03/2019)
con_par <- function(initial_m_value, n, K, prior_m_params)
446 {
  x <- rbeta(1, initial_m_value + 1, n)
448 pi1 <- prior_m_params[1] + K - 1
  pi2 <- n * (prior_m_params[2] - log(x))
450 pi1 <- pi1/(pi1 + pi2)
  if (runif(1) < pi1) {
452   g1 <- rgamma(1, prior_m_params[1] + K, prior_m_params[2] - log(x))
   new_m_value <- g1
454 } else {
   g2 <- rgamma(1, prior_m_params[1] + K - 1, prior_m_params[2] - log(x))
456   new_m_value <- g2
  }
458 #new_concentration parameter
  return(new_m_value)
460 }

462 hpd_interval <- function(obj, prob = 0.95, ...) {
  obj <- as.matrix(obj)
464 vals <- apply(obj, 2, sort)
  if (!is.matrix(vals)) stop("obj must have nsamp > 1")
466 nsamp <- nrow(vals)
  npar <- ncol(vals)
468 gap <- max(1, min(nsamp - 1, round(nsamp * prob)))
  init <- 1:(nsamp - gap)
470 inds <- apply(vals[init + gap, ,drop=FALSE] - vals[init, ,drop=FALSE],
    2, which.min)
472 ans <- cbind(vals[cbind(inds, 1:npar)],
    vals[cbind(inds + gap, 1:npar)])
474 dimnames(ans) <- list(colnames(obj), c('lower', 'upper'))
  attr(ans, "Probability") <- gap/nsamp
476 round(ans,3)
  }
478
480
482 # -----
# Check user input and get basic info from data
# -----
484
# -----
486 # Get basic info from data/ convert to convenient forms for computations.
# -----
488 Y = as.vector(Y)
Z = otu.mat = as.matrix(Z)
490 n = num.sam = length(Y)
num.otu = ncol(Z)

```

```

492 # -----
494 # Initialize the model parameters
496 # -----
496 precis_param <- alpha.par #Gamma (a,b) parameters for the precision parameter
    of the DP.
    if(is.null(precis_param)) {precis_param = c(exp(-0.033*num.sam), exp(-0.033*num
        .sam))} # these are the parameters of Gamma(a,b) prior on m.
498 a.tau      <- tau2.par[1] # gamma distr parameter for tau - the parameter
    for base measure of DP.
    b.tau      <- tau2.par[2]
500 a.beta      <- sig2.bet.par[1] # inv-gamma parameter for sig.beta2 - the
    parameter for variance component of Beta.
    b.beta      <- sig2.bet.par[2] #
502
    tau.sq      <- rinvgamma(1,shape=a.tau,scale=b.tau)
504 if(!is.finite(tau.sq)) {tau.sq <- 1.0e+50} # If Infinity, use 1.0e+50 to
    represent the large enough infinite value!!
    tau      <- sqrt(tau.sq)
506
    sig.sq.beta <- rinvgamma(1,shape=a.beta,scale=b.beta)
508 if(!is.finite(sig.sq.beta)) {sig.sq.beta <- 1.0e+50} # If Infinity, use 1.0e+50
    to represent the large enough infinite value!!
    sig.beta   <- sqrt(sig.sq.beta)
510
    # Also initialize the DP parameters.
512 # Algorithm to simulate DP parameters adapted from Kyung et al (2010).
    q      <- c(rdirichlet(1, rep(1, num.sam))) # PROBABILITY OF
        ASSIGNMENT
514 A.n      <- sample(1:num.sam,num.sam,replace=TRUE,prob=q) # CREATES
        THE ASSIGNMENTS.
    A.K      <- table(A.n) # CREATES THE LIST OF
        OCCUPIED
516 A.K.labels <- as.numeric(names(A.K)) # LOCATIONS OF OCCUPIED
    K      <- length(A.K) # NUMBER OF OCCUPIED
        TABLES
518 B      <- rep(0,length=num.sam) # LENGTH num.sam
        FOR ALL CASES
    B[A.K.labels] <- 1 # 1 WHERE OCCUPIED >0
520 B      <- B+rnorm(num.sam,0,tau) # ASSIGN psi
        VALUES TO OCCUPIED
    nk      <- rep(0,num.sam); for (i in 1:num.sam) nk[A.n[i]] <- nk[A.n[i]]
        + 1 # COUNTS OCCUPANTS AT TABLES
522 psi     <- B[A.n] # MAP TABLE psi'S TO
        CASES
    eta     <- as.vector(rmvnorm(1, mean=rep(0,K), diag(rep(tau.sq), K)))
524

```

```

526  ## -----
    # Compute OTU similarity measures
528
    # Create 'RBF kernel
530  #X must be a matrix containing samples in its rows
    rbf.kernel = as.matrix(rbfkernel(X=t(Z[, ]), sigma=rbf.par)) #transpose to apply
        across taxa
532
    # Create Phylogenetic tree kernel
534  if(is.null(mTree)){
        JointKernel = rbf.kernel
536  }else{
        if(class(mTree)=='matrix'){
538            # if mTree supplied as a similarity matrix, convert to dissimilarity matrix
            # as exponentiation converts back to similarity matrix.
540            if(mTree[1,1]==1){ # testing if a similarity matrix.
                mTree.dissim.matrix <- (1-mTree)
542            }else if(mTree[1,1]==0){ # testing if a dissimilarity matrix
                mTree.dissim.matrix <- mTree # if mTree given as dissimilarity matrix
544            }else if (class(mTree)=='matrix' & mTree[1,1]!=0 & mTree[1,1]!=1 ) {
                stop("Error: the mTree matrix object supplied is not a similarity measure
                    . Please check!")
546            }
        }
548        if(class(mTree)=='phylo'){
            mTree.dissim.matrix <- as.matrix(cophenetic.phylo(mTree))
550        }

552        phylo.kernel <- exp(-2.0*evol.rate*mTree.dissim.matrix^2.0) #
        JointKernel <- mu*phylo.kernel + (1-mu)*rbf.kernel #Composite kernel
554    }

556    joint.cov <- JointKernel
    if(otu.corr==FALSE) joint.cov=diag(num.otu)
558
    ## Add inrtercept to the data
560    Z      <- cbind(1,Z) #

562    ## With intercept, add data of zero correlations for it.
    int.mat <- matrix(0, nrow=ncol(Z),ncol=ncol(Z)) ; int.mat[1,1] <- 1
564    int.mat[-1,-1] <- joint.cov
    joint.cov <- int.mat
566
    Beta1 <- rmvnorm (1, mean = rep(0, ncol(Z)), sigma = diag(ncol(Z)))
568

570  ## -----
    # Declare storage matrix for the sampled parameters.

```

```

572 # -----
Beta      <- rbind(Beta1, matrix(NA, nrow=Nsim, ncol=ncol(Z)))
574 tau.sq   <- c(tau.sq, rep(NA,Nsim))
sig.sq.beta <- c(sig.sq.beta, rep(NA,Nsim))
576
#name the matrices and vectors [variables]:
578 colnames(Beta) <- c("intercept",colnames(otu.mat))
names(tau.sq)   <- "tau.sq"
580 names(sig.sq.beta) <- "sig.sq.beta"

582 # -----
# calculate the likelihood.
584 pr <- 1 - 1/(1 + exp(Z%*%t(Beta1) + psi)) #
like.K <- 0 #
586 for (i in 1:n){
  like.K <- like.K + (dbinom(Y[i], size=1, prob=pr[i], log=TRUE)) + (dnorm(psi[
    i], mean=0, sd=tau,log=TRUE))
588 like.K <- exp(like.K + sum(lgamma(A.K))) #
}
590
# For the prior of m, we use Gamma(a,b)
592 m = initial.m = n/log(n) # precision parameter of the DP part. Initialize by n/
  log(n)
m = c(m, rep(NA,Nsim))
594 names(m) = "m"

596 Mu   = rnorm(ncol(Z),0,1)
Gama = rnorm(ncol(Z),0,1)
598 tau2.gam = 1

600
cat('\nPerforming MCMC sampling ... \n\n')
602
# -----
604 # Perform MCMC (Gibbs-slice) sampling for Nsim iterations.
# -----
606 for (iter in 2:(Nsim+1)){ # simulate samples (particles).

  ## print iteration to screen, if print.iter==TRUE
  if ( iter%%8==0 | iter%%20==0 ){
610    cat('Completed', format(round((iter/(Nsim+1))*100,1),nsmall=1),"%.", "
      Wait...", "\r")
  }
612

# -----
614 # Performing M-H algorithm to generate 'A' matrix.
# -----
616 p.A.old = ((m[iter-1])^K)

```

```

f.y.old = like.K
618 mult.old = dmultinom(x=A.K, prob=q[A.K.labels])

620 # Create new 'A' matrix; 'can' stands for candidate
pq      <- nk +1
622 new.q      <- rdirichlet(1, pq)
A.n.can  <- sample(1:num.sam,num.sam,replace=TRUE,new.q)
624 A.K.can    <- table(A.n.can)
A.K.labels.can <- as.numeric(names(A.K.can))
626 K.can     <- length(A.K.can)
B        <- rep(0,length=num.sam)
628 B[A.K.labels.can] <- 1
B        <- B*rnorm(num.sam,0, tau)
630 nk.can    <- rep(0,num.sam); for (i in 1:num.sam) nk.can[A.n.can[i]]
      <- nk.can[A.n.can[i]] + 1
psi.can    <- B[A.n.can]

632 pr <- 1 - 1/(1 + exp(Z%*%Beta[iter-1,] + psi))
634 like.K.can <- 0
for(i in 1:n){
636   like.K.can <- like.K.can + (dbinom(Y[i], size=1, prob=pr[i], log=TRUE)) + (
     dnorm(psi[i], mean=0, sd=tau, log=TRUE)) #
}
638 like.K.can <- exp(like.K.can + sum(lgamma(A.K.can))) #

640 p.A.can <- (m[iter-1]^K.can)
f.y.can <- like.K.can
642 mult.can <- dmultinom(x=A.K.can, prob=new.q[A.K.labels.can])

644 # Define the M-H acceptance probability to ACCEPT or REJECT simulated
     candidates.
p.ratio <- p.A.can/p.A.old; f.ratio <- f.y.can/f.y.old; mult.ratio <- mult.
     can/mult.old
646 if (!is.finite(p.ratio)) {p.ratio      <- 1} # If value NOT finite, set it
     to 1.
if (!is.finite(f.ratio)) {f.ratio      <- 1}
648 if (!is.finite(mult.ratio)) {mult.ratio <- 1}
rho <- p.ratio * f.ratio * mult.ratio # Acceptance probability/ratio.
650 if (rho>runif(1)) {
     A.n      <- A.n.can
652     A.K      <- A.K.can
     A.K.labels <- A.K.labels.can
654     K        <- K.can
     nk        <- nk.can
656     psi      <- psi.can
     like.K    <- like.K.can
658 }

```

```

660  ## =====
661  # Simulate uniform random variates; delta and lamb.
662  func      <- Z%%Beta[iter-1,] + psi
663  prob.fun  <- exp(func)/(1+exp(func)) #
664  delta     <- runif(num.sam, rep(0,num.sam), prob.fun)

665  #And for lamb, we have:
666  lamb <- runif(num.sam, rep(0,num.sam), 1-prob.fun)

667  # Simulate Beta
668  for (j in 1:dim(Z)[2]){
669      Gama[j] <- rnorm(n=1, mean=(1/(as.numeric(solve(t(joint.cov[,j]))%%joint.
670          cov[,j])) + 1/tau2.gam)) *
671          (as.numeric(solve(t(joint.cov[,j]))%%joint.cov[,j]))*Mu[j]) ,
672          sd=as.numeric(solve(t(joint.cov[,j]))%%joint.cov[,j])) + 1/tau2.
673          gam)
674
675      Mu[j] <- rnorm(n=1, mean=1/(1/sig.beta^2 + as.numeric(solve(t(joint.cov[,
676          j]))%%joint.cov[,j]))) *
677          ((1/sig.beta^2)*Beta[iter-1,j] + as.numeric(solve(t(joint.cov[,j
678          ]))%%joint.cov[,j]))%%Gama[j]),
679          sd=1/sig.beta^2 + as.numeric(solve(t(joint.cov[,j]))%%joint.cov[,
680          j])) )
681
682      # Define the truncation limits (intervals) to simulate 'Beta'.
683      S1 <- (1/Z[Y==1,j]) * (log(delta[Y==1]/(1-delta[Y==1])) - Z[Y==1,-j] %%%
684          Beta[iter-1,-j] - psi[Y==1])
685      S2 <- (1/Z[Y==0,j]) * (log((1-lamb[Y==0])/lamb[Y==0]) - Z[Y==0,-j] %%%
686          Beta[iter-1,-j] - psi[Y==0])
687
688      #left truncation limit
689      ind1 = Z[Y==1,j] > 0
690      A = ind1*S1
691      A[A==0] = NaN #
692      left1 = suppressWarnings(max(A, na.rm=TRUE))
693
694      ind2 = Z[Y==0,j] < 0
695      B = ind2*S2
696      B[B==0] = NaN
697      left2 = suppressWarnings(max(B, na.rm=TRUE))
698      beta.left = max(left1, left2)
699
700      #right truncation limit
701      ind3 = Z[Y==1,j] < 0
702      C = ind3*S1
703      C[C==0] = NaN
704      right1 = suppressWarnings(min(C, na.rm=TRUE))

```

```

702     indx4 = Z[Y==0,j] >0
       D = indx4*S2
       D[D==0] = NaN
704     right2 = suppressWarnings(min(D, na.rm=TRUE))
       beta.right = min(right1, right2)
706
       # simulating from truncated normal distribution.
708     #Either:
       unif.random1 <- runif(1,0,1)
710     a1 <- pnorm(beta.left, Mu[j], sig.beta) #
       a2 <- pnorm(beta.right, Mu[j], sig.beta)
712     pa <- a1 + unif.random1*(a2-a1)
       if(pa > 0.99999) {pa=0.99999} # Avoid numerical issues, since qnorm(1,
           mean,sd)==Inf
714     current.beta.j <- qnorm(pa, Mu[j], sig.beta)
       #Or:
716     #current.beta.j <- rtnorm(n=1, mean = Mu[j], sd = sig.beta, lower = beta.
           left, upper = beta.right)

718     Beta[iter,j] <- current.beta.j
720 }

722 ## =====
       # Simulate 'eta'.
724 A.m1 = matrix(0, nrow=num.sam, ncol=num.sam)
       for (i in 1:n){ A.m1[i,A.n[i]] <- 1 }
726 A.K.m = A.m1[,which(apply(A.m1,2,sum)>0)]

728 for (k in 1:K){
       # Lower limit for eta
730     eta.lim1 = log(delta[Y==1]/(1-delta[Y==1])) - Z[Y==1,] %*% Beta[iter,] - A
           .K.m[Y==1,-k] %*% eta[1:K][-k] #
       eta.lim2 = log((1-lamb[Y==0])/lamb[Y==0]) - Z[Y==0,] %*% Beta[iter,] - A.K
           .m[Y==0,-k] %*% eta[1:K][-k] #
732
       eta.left = max(eta.lim1) #
734     if(!is.finite(eta.left)) eta.left=-Inf
       eta.right = min(eta.lim2) #
736     if(!is.finite(eta.right)) eta.right=Inf

738     #simulating from truncated normal distribution.
       #Either:
740     unif.random2 <- runif(1,0,1)
       b1 <- pnorm(eta.left, 0, tau)
742     b2 <- pnorm(eta.right, 0, tau)
       pb <- b1 + unif.random2*(b2-b1)

```

```

744     if(pb > 0.99999) {pb=0.99999} #Avoid numerical issues, since qnorm(1,mean,
        sd)==Inf
        current.eta.k <- qnorm(pb, 0, tau)
746     #Or:
        #current.eta.k <- rtnorm(n=1, mean=0, sd=tau, lower=eta.left, upper=eta.
            right)
748     eta[k] <- current.eta.k
    }

750     ## =====
752     # Use simulated eta to generate psi.
    psi <- A.K.m %*% eta[1:K]
754     ## =====
756     # Perform conditional update on tau2, sigma2.beta.
    # simulate tau2
758     tau2 <- rinvgamma(1,shape=(0.5*K)+a.tau, scale=(0.5*sum(eta^2))+b.tau )

760     if (is.finite(tau2)){
        tau <- sqrt(tau2)
762     }else{ #since tau2 can go to infinity.
        tau2 <- 1e50 # if value goes infinite, 1e50 can represent infinity??
764         tau <- sqrt(tau2)
    } #
766     tau.sq[iter] <- tau2 # save tau2

768     ## =====
770     # simulate sig.beta.sq
    sig2.beta <- rinvgamma(1, shape=(0.5*ncol(Z)+a.beta), scale=(0.5*sum((Beta[
        iter,]-Mu)^2)+b.beta))
772     if(is.finite(sig2.beta)){
        sig.beta <- sqrt(sig2.beta)
774     }else{ #since sig2.beta can go to infinity.
        sig2.beta = 1e50 # If infinite, 1e50 can represent infinity??
776         sig.beta <- sqrt(sig2.beta) #
    } #
778     sig.sq.beta[iter] <- sig2.beta

780     ## =====
782     # Update concentration parameter, 'm':
    # inputs: [previous (m), sample size (n), cluster size (K), Gamma params (a,b
        ) on m].
    m.cur <- con_par(m[iter-1], n, K, precis_param) #
784     m[iter]<- m.cur

786     if(iter==(Nsim+1)){print('Sampling completed.', quote = FALSE)}

```

```

788 } # End of Nsim.

790
792 # Write (if save is TRUE) the raw simulation output to a file.
792 if (saveSim){
794   cat("\nWriting raw simulation data to a file in the present directory ...\n
      ")
794   out.df <- cbind(Beta, sig.sq.beta, m, tau.sq)
794   column.names <- c(paste0("OTU", 1:num.otu), "sigma2.beta", "m", "tau.sq")
796   #In writing the output to the file, exclude first row (initialization
      values for the simulations)
796   write.table(out.df[-1,], file="simParams.txt", quote=FALSE, row.names =
      FALSE, col.names = column.names)
798 }

800 cat('\nPerforming posterior inference ...\n')

802 # Estimate of Parameters
802 # Discard burn-in samples and do thinning (or not, when thin =1) to posterior
      samples before conducting
804 # inference on the parameters.
804 # Burn-in
806 Beta.Burn      <- Beta[(burnin+1):Nsim, ]
806 tau2.Burn      <- tau.sq[(burnin+1):Nsim]
808 sig2.beta.Burn <- sig.sq.beta[(burnin+1):Nsim]
808 m.Burn         <- m[(burnin+1):Nsim]
810
810 # thinning
812 Beta.BurnThin  <- thin.samples(as.matrix(Beta.Burn), thin)
812 tau2.BurnThin  <- thin.samples(as.vector(tau2.Burn), thin)
814 sig2.beta.BurnThin<- thin.samples(as.vector(sig2.beta.Burn), thin)
814 m.BurnThin     <- thin.samples(as.vector(m.Burn), thin)
816
818 cat(paste0("\nNumber of effective posterior samples ", nrow(Beta.BurnThin), "\n
      "))

820
822 # =====
822 #               Posterior summary
822 # =====
824 posterior.summary <- function(object, alpha=0.5, ...){
826   #
826   # S3 method to compute and print posterior summaries for a matrix of draws
826   # Modified from package bayeslm, author Hahn et al.(2017).
828   #
828   numEff=
830   function(x,m=as.integer(min(length(x), (100/sqrt(5000))*sqrt(length(x)))))

```

```

832 {
# purpose:
# compute N-W std error and relative numerical efficiency
834 #
# Arguments:
836 # x is vector of draws
# m is number of lags to truncate acf
838 # def is such that m=100 if length(x)= 5000 and grows with sqrt(length)
#
840 # Output:
# list with numerical std error and variance multiple (f)
842 #
wgt=as.vector(seq(m,1,-1))/(m+1)
844 z=acf(x,lag.max=m,plot=FALSE)
f=1+2*wgt %*% as.vector(z$acf[-1])
846 stderr=sqrt(var(x)*f/length(x))
list(stderr=stderr,f=f,m=m)
848 }
X = object
850 if(mode(X) == "list") stop("list entered \n Please extract sampling matrix from
the list \n")
if(mode(X) != "numeric") stop("Requires numeric argument \n")
852 if(is.null(attributes(X)$dim)) X=as.matrix(X)
nx=ncol(X)
854
if(is.null(colnames(X))){
856 names=as.character(1:nx)
} else {
858 names = colnames(X)
}
860
post.mean.beta=round(apply(X,2,mean), digits=3)
862
num_se=double(nx); rel_eff=double(nx); eff_s_size=double(nx)
864 for(i in 1:nx) # posterior numerical standard errors.
{out=numEff(X[,i])
866 if(is.nan(out$stderr))
{num_se[i]=-9999; rel_eff[i]=-9999; eff_s_size[i]=-9999}
868 else
{num_se[i]=out$stderr; rel_eff[i]=out$f; eff_s_size[i]=nrow(X)/ceiling(
out$f)}
870 }
num_se = round(num_se,3); rel_eff= format(round(rel_eff,3), nsmall=3)
872
sds.beta= round(apply(X,2,sd),3) # posterior stand devs.
874 lo = post.mean.beta - qnorm(1-alpha/2) * num_se
up = post.mean.beta + qnorm(1-alpha/2) * num_se
876 CI = format(round(cbind(lo, up),3), nsmall=3)

```

```

878 mat = cbind(format(post.mean.beta,nsml=3), format(num_se,nsml=3), CI)

880 #names of variables
      rownames(mat)=c("(intercept)",colnames(otu.mat),"m")
882 colnames(mat)[1]="Coef" #Estimated coefficients
      c#colnames(mat)[2]="std.dev" #posterior stan devs
884 colnames(mat)[2]="SE" # numerical standard errors
      #colnames(mat)[4]="rel.eff" # relative efficiency
886 colnames(mat)[3]="95% CI(lo)" # Credible Interval (lower)
      colnames(mat)[4]="95% CI(up)" # (upper)
888
      # format the output results
890 mat=format(mat, justify = "centre", digits=3, scientific=FALSE)

892 cat("\nEstimates of the parameters:\n")
      print((mat),digits=3, quote=FALSE)
894
      #return(invisible(1))
896 return(mat)
      # invisible(t(mat))
898 } # End of 'posterior.summary' function

900
      # =====
902 # Posterior Plots (traceplot, ACF plot) on the full simulated data
      # =====
904 graphs.function <- function(X, tvalues,...){

906   nx=ncol(X)

908   if(nx==1) par(mfrow=c(1,1))
      if(nx==2) par(mfrow=c(2,1))
910   if(nx==3) par(mfrow=c(3,1))
      if(nx==4) par(mfrow=c(2,2))
912   if(nx>=5) par(mfrow=c(5,4))

914   if(post.plots){
      if(nx==1) par(mfrow=c(1,4))
916   if(nx==2) par(mfrow=c(2,4))
      if(nx>=3) par(mfrow=c(5,4))
918   for(index in 1:nx){
      plot(as.vector(X[,index]),xlab="Iteration",ylab="",main=paste0("Trace
      plot for ",colnames(X)[index]),type="l",col="red") #traceplot
920   plot(density(X[,index]), xlab="Value",ylab="Density",main=paste0("Density
      plot for ",colnames(X)[index]), type="l",col="blue") #density plot
      cumuplot(mcmc(X[,index]),probs=c(0.025,0.5,0.975),ylab="",xlab="Iteration
      ",main=paste0("Cumulative quantile plot for ",colnames(X)[index]),

```

```

922         lty=c(2,1),lwd=c(1,2),type="l",ask=FALSE,auto.layout=FALSE, col=1) #
           cumulative quantile plot
           if(!missing(tvalues)) abline(h=tvalues[index],lwd=2,col="blue")
924         if(var(X[,index])>1.0e-20) {acf(as.vector(X[,index]),xlab="Lag",ylab="
           Autocorrelation",main=paste0("ACF plot for ",colnames(X)[index]))} #
           ACF plot
           else {plot.default(X[,index],xlab="",ylab="",type="n",main="No ACF
           Produced")}}
926     }
    }
928     invisible()
    } # end of graphs.function
930
932 # =====
    #             Posterior Inference
934 # =====
    # # Do posterior inference on 'effective' data (after thinning and burn-in).
936 post.est <- posterior.summary(cbind(Beta.BurnThin,m.BurnThin))

938 # # Make posterior plots using ALL the simulated data (no thin and no burn-in)
    if(post.plots){
940     graphs.function(cbind(Beta[,1:5])) #posterior plots for OTU effects..
    }
942
944 # Stop the clock
    end.time <- proc.time()[3] # '3' in seconds.
946 time.taken <- end.time - start.time # calculate difference
    #In realistic use (with large Nsim), 'time.taken' will be several minutes to
    hours.
948 if(time.taken < 60){
    cat("\nTotal time elapsed:", round(time.taken,digits=0), "seconds.\n")
950 }else if(time.taken >=60 & time.taken < 3600){
    cat("\nTotal time elapsed:", round(time.taken/60, digits=0),
952     ifelse(round(time.taken/60, digits=0)==1,"minute\n", "minutes.\n"))
    }else{ # for time in hours (>3600 seconds)
954     cat("\nTotal time elapsed:", (time.taken %/%3600),
    ifelse((time.taken %/%3600) < 2,"hour","hours"),
956     round((time.taken %/% 3600)/60,digits=0),
    ifelse(round((time.taken %/% 3600)/60,digits=0) < 2,"minute.\n","minutes.\n
    "))
958 }

960 cat("\nDone! Thank you.\n")
    cat("-----\n")
962

```

```

964 #return posterior means of parameters
return(list(out.all = data.frame(post.est), beta=as.numeric(post.est[1:(ncol(Beta)
)+1), 'Coef'])))
966
} # End of 'phyloDPM.main' function.
968
#=====
970 # We create a generic user-input function called "PhyloDPM" (18/07/2019)
#=====
972 phyloDPM <- function(pheno, otu.data, mTree, Nsim, thin=1, burnin=NULL, post.
plots=FALSE, mu=0,
alpha.par=NULL, tau2.par=c(20,20), sig2.bet.par=c(20,15), rbf
.par=0.001, saveSim=FALSE, evol.rate=10^3, otu.corr=TRUE
,...) {
974
est <- phyloDPM.main(Y = pheno, Z = otu.data, mTree = mTree, tau2.par=tau2.par,
sig2.bet.par=sig2.bet.par,
976 alpha.par=alpha.par, Nsim=Nsim, thin=thin, burnin=burnin,
post.plots=post.plots, mu=mu, rbf.par=rbf.par, saveSim=
saveSim, evol.rate=evol.rate, otu.corr=otu.corr,...)
978 #class(est) <- "phyloDPM"
est
980 }
982 # =====
#
#           END OF phyloDPM.R SCRIPT
984 # =====

```