

Metagenomic analyses of the human gut microbiome reveal connections to the immune system

Tommi Vatanen



Metagenomic analyses of the human gut microbiome reveal connections to the immune system

Tommi Vatanen

A doctoral dissertation completed for the degree of Doctor of Science (Technology) to be defended, with the permission of the Aalto University School of Science, at a public examination held at the lecture hall AS1 of the school on March 24, 2017 at 12 noon.

Aalto University
School of Science
Department of Computer Science
Computational systems biology group

Supervising professor

Prof. Harri Lähdesmäki, Aalto University, Finland

Thesis advisors

Prof. Ramnik Xavier, Harvard Medical School, The Broad Institute of MIT and Harvard, Massachusetts General Hospital, Massachusetts Institute of Technology, USA

Prof. Curtis Huttenhower, Harvard School of Public Health, The Broad Institute of MIT and Harvard, USA

Preliminary examiners

Prof. Anders Andersson, KTH Royal Institute of Technology, Sweden

Dr. Jarkko Salojärvi, University of Helsinki, Finland

Opponent

Prof. David Berry, University of Vienna, Austria

Aalto University publication series

DOCTORAL DISSERTATIONS 35/2017

© Tommi Vatanen

ISBN 978-952-60-7314-9 (printed)

ISBN 978-952-60-7313-2 (pdf)

ISSN-L 1799-4934

ISSN 1799-4934 (printed)

ISSN 1799-4942 (pdf)

<http://urn.fi/URN:ISBN:978-952-60-7313-2>

Unigrafia Oy

Helsinki 2017

Finland



Author
Tommi Vatanen

Name of the doctoral dissertation
Metagenomic analyses of the human gut microbiome reveal connections to the immune system

Publisher School of Science

Unit Department of Computer Science

Series Aalto University publication series DOCTORAL DISSERTATIONS 35/2017

Field of research Computational systems biology

Manuscript submitted 12 October 2016 **Date of the defence** 24 March 2017

Permission to publish granted (date) 16 January 2017 **Language** English

☐ **Monograph** ☒ **Article dissertation** ☐ **Essay dissertation**

Abstract

Mounting evidence shows that the gut microbiome has an important role in human health. This thesis utilizes metagenomic sequencing data to examine the role of the gut microbiota in human health, specifically in juvenile autoimmune disorders such as type 1 diabetes (T1D). Numerous studies suggest that the intestinal microbes contribute to many immunological disorders and other conditions such as obesity. According to the hygiene hypothesis, lack of certain microbial exposures are central in development of autoimmunity in early childhood. However, a clear distinction between beneficial and harmful bacteria as well as mechanistic understanding of how the microbiome leads to aberrations in immune development are lacking.

We aimed to elucidate the mechanisms behind the hygiene hypothesis by studying the gut microbiome of infants at risk for autoimmune disorders in Northern Europe. We also investigated the gut microbiome of 1135 healthy Dutch adults for connections with various intrinsic and extrinsic factors, and conducted a fecal microbial transplantation study in active Crohn's disease (CD).

We used whole metagenome shotgun (WMS) and 16S rRNA gene sequencing, and computational analysis methods to taxonomically and functionally profile the gut microbial communities in three separate cohorts. By investigating the gut microbiome of 294 infants from Finland, Estonia and Russian Karelia, we characterized the developing infant gut microbiome and identified the immunogenicity of lipopolysaccharide (LPS) produced by *Bacteroides* species as a novel factor contributing to the higher incidence of T1D in Finland and Estonia compared to Russian Karelia. LPS derived from *Bacteroides dorei*, a species that has been previously linked to T1D pathogenesis, harbored tetra- and penta-acylated LPS structures which failed to induce immune stimulation in human cells and inhibited immune stimulation and endotoxin tolerance by *Escherichia coli*-derived LPS. We also found that recurrent antibiotic treatments lead to decrease in microbial diversity and increase in antibiotic resistance genes.

Using WMS sequencing data from healthy adults, we identified chromogranin A as a novel biomarker for gut health. By investigating the gut microbiome of 19 CD patients before and after colonoscopic fecal microbial transplantation (FMT), we concluded that the FMT was safe and resulted in a shift towards the donor microbiota.

This thesis contributes to novel functional and mechanistic understanding of the human gut microbiome in infancy and adulthood, health and disease. I provide new computational methodologies for analysing microbiome on strain level and connecting its functionalities with human health. The findings of this thesis pave way towards therapeutic interventions modifying the gut microbiome to prevent immune mediated and other disorders in humans.

Keywords gut microbiome, metagenomic sequencing, type 1 diabetes, hygiene hypothesis

ISBN (printed) 978-952-60-7314-9

ISBN (pdf) 978-952-60-7313-2

ISSN-L 1799-4934

ISSN (printed) 1799-4934

ISSN (pdf) 1799-4942

Location of publisher Helsinki

Location of printing Helsinki **Year** 2017

Pages 149

urn <http://urn.fi/URN:ISBN:978-952-60-7313-2>

Tekijä

Tommi Vatanen

Väitöskirjan nimi

Ihmisen suolistomikrobiston perimän analyysit paljastavat kytköksiä immuunijärjestelmään

Julkaisija Perustieteiden korkeakoulu**Yksikkö** Tietotekniikan laitos**Sarja** Aalto University publication series DOCTORAL DISSERTATIONS 35/2017**Tutkimusala** Laskennallinen systeemibiologia**Käsitteilyajon pvm** 12.10.2016**Väitöspäivä** 24.03.2017**Julkaisuluvan myöntämispäivä** 16.01.2017**Kieli** Englanti☐ **Monografia**☒ **Artikkeliväitöskirja**☐ **Esseeväitöskirja****Tiivistelmä**

Useat tutkimukset ovat osoittaneet ihmisen suolistomikrobiston terveysvaikutukset. Tässä väitöskirjassa tutkittiin suolistomikrobiston yhteyttä terveyteen käyttäen moderneja DNA-sekvensointiin perustuvia menetelmiä, keskittyen erityisesti varhaislapsuuden autoimmuunisairauksiin kuten tyypin 1 diabetekseen (T1D). Hygieniahypoteesin mukaan vähäinen altistus ympäristön mikrobeille varhaislapsuudessa voi johtaa häiriöihin immuunijärjestelmän kehityksessä. Aikaisempien tutkimusten perusteella suolistobakteerit ovatkin tärkeässä roolissa monissa immuunijärjestelmän häiriöissä, mutta useat tekijät kuten jako hyödyllisiin ja haitallisiin bakteereihin sekä mekanismit joilla nämä vaikutukset välittyvät ovat edelleen epäselviä.

Tarkastelimme hygieniahypoteesin takana olevia mekanismeja tutkimalla varhaislapsuuden suolistomikrobistoa T1D-altistuneilla lapsilla. Selvitimme myös suolistomikrobiston yhteyksiä luontaisiin ja ulkoisiin tekijöihin 1135:n terveen alankomaalaisen aikuisen aineistolla, sekä suoritimme ulosteensiirtotutkimuksen Crohnin tautia sairastavilla potilailla.

Analysoimme keräämämme ulostenäytteet käyttäen ns. metagenomin sekvensointia ja 16S rRNA geenin sekvensointia, sekä laskennallisia työkaluja. Näiden avulla karakterisoimme ulostenäytteet sekä niiden bakteeritaksonomioiden että bakteerien metabolisten ominaisuuksien osalta. Tutkimalla 294 lapselta Suomessa, Virossa sekä Venäjän Karjalassa kerättyjä ulostenäytteitä määritimme tyypillisiä tekijöitä varhaislapsuuden suolistoflooran kehityksestä ja tunnistimme *Bacteroides*-lajien tuottaman lipopolysakkaridin (LPS) erityispiirteet tekijäksi, joka altistaa suomalais- ja virolaislapsia diabetekselle. *Bacteroides dorei*, joka on aiemmin yhdistetty diabeteksen puhkeamiseen, tuotti immuunijärjestelmälle näkymätöntä LPS:a, joka ei kokeissa aktivoitunut valkosoluja ja pystyi myös hiljentämään kolibakteerin LPS:n vaikutuksia sekä endotoksiinitoleranssia. Näytimme myös mm. että varhaislapsuuden antibioottikuurit johtavat köyhtyneeseen suolistoflooraan ja lisäävät antibioottivastustuskykyä välittävien geenien määrää.

Lisäksi löysimme uuden suoliston terveydestä kertovan biomarkkerin, kromograniniini A, terveiden alankomaalaisten aineistosta. Lisäksi suoritimme ulosteensiirron 19 Chronin tautia sairastaneelle potilaalle ja totesimme että operaatio on paitsi turvallinen, se johtaa muuntuneeseen suolistomikrobistoon, joka muistuttaa luovuttajan mikrobistoa.

Tämä väitöskirja sisältää uutta toiminnallista ja mekanistista tietoa siitä kuinka suolistobakteerit vaikuttavat terveyteen lapsuudessa ja aikuisuudessa. Osoitimme että DNA-sekvensointi-aineistojen laskennallisilla analyyseillä voidaan saada yksityiskohtaista tietoa bakteerien vaikutuksista terveyteen. Tämän väitöskirja osoittaa tietä uusille mikrobistoa muokkaaville terapimuodoille, joilla voitaneen tulevaisuudessa ehkäistä autoimmuunisairauksien puhkeamista.

Avainsanat suolistomikrobisto, metagenomiikka, tyypin 1 diabetes, hygieniahypoteesi**ISBN (painettu)** 978-952-60-7314-9**ISBN (pdf)** 978-952-60-7313-2**ISSN-L** 1799-4934**ISSN (painettu)** 1799-4934**ISSN (pdf)** 1799-4942**Julkaisupaikka** Helsinki**Painopaikka** Helsinki**Vuosi** 2017**Sivumäärä** 149**urn** <http://urn.fi/URN:ISBN:978-952-60-7313-2>

Preface

This thesis contributes to the fields of microbial ecology and human microbiome which have reached public attention in the recent years with numerous fascinating breakthroughs and discoveries. The research for this thesis has been conducted in two research institutions: in the Computational Systems Biology group led by Prof. Harri Lähdesmäki at the Department of Computer Science at Aalto University, Finland and in the Laboratory of Prof. Ramnik Xavier at the Broad Institute of MIT and Harvard in Boston, USA.

I had an opportunity to be introduced to this field roughly four years ago, when Prof. Harri Lähdesmäki, Prof. Ramnik Xavier and Prof. Mikael Knip started a collaboration related to the DIABIMMUNE study. My supervisor Prof. Lähdesmäki supported my visit to Prof. Xavier's laboratory in the Broad Institute of MIT and Harvard. Since then I have been privileged to learn about the microbiome from scientists and mentors from numerous different disciplines including computer science, microbiology and immunology to mention just a few. I spend most of my time between September 2013 and December 2016 in the Broad Institute as a visiting scholar in Prof. Xavier's laboratory.

This work has been generously supported and funded by former Helsinki Doctoral Programme in Computer Science (Hecse), the Juvenile Diabetes Research Foundation (JDRF) and Academy of Finland's Centre of Excellence in Molecular Systems Immunology and Physiology Research (SyMMyS). I am grateful to these organizations for their generous financial support.

I would like to thank my supervisor and instructor Prof. Harri Lähdesmäki. His visionary ability to form collaborations across department lines, country borders and scientific disciplines, and his consistent support have been instrumental for enabling this thesis.

I would like to thank my numerous advisors, mentors and collaborators

in the Broad Institute. I thank my thesis advisors Prof. Ramnik Xavier and Prof. Curtis Huttenhower for introducing me into the field of microbiome. I thank Dr. Dirk Gevers and Dr. Alex Kostic for the warm welcome I received in the laboratory. I thank Dr. Hera Vlamakis for her consistent support and valuable mentoring. I also thank all current and former members of the laboratories of Prof. Xavier and Prof. Huttenhower I have worked with, particularly Dr. Moran Yassour, Dr. Raivo Kolde, Dr. Eric Franzosa and Dr. Timothy Tickle.

I would like to address warm thanks to all my collaborators; without you this thesis would not be possible. I thank Dr. Mikael Knip for his continuous support and feedback, and for his persistent drive to understand the causes of type 1 diabetes. I thank our collaborators at Novartis Institutes of Biomedical Research, Dr. Eva d’Hennezel and Dr. Thomas W. Cullen, for valuable contributions and fruitful scientific exchange. I thank Dr. Alexandra Zhernakova, Dr. Jingyuan Fu, Dr. Cisca Wijmenga, Dr. Byron Vaughn and Dr. Alan Moss for their valuable work and scientific collaboration.

I would also like to thank the pre-examiners of this thesis, Prof. Anders Andersson and Dr. Jarkko Salojärvi, for their insightful and detailed comments which undoubtedly improved this thesis.

Finally, I want to thank my family for their unconditional support for anything I have decided to pursue in my life. I thank many friends who have given important support and advice during the preparation of this thesis. I also want to thank my partner Rachel for her support and her continuing interest in what I do. Thank you for providing a perfect environment for writing this thesis in Auckland, New Zealand.

Auckland, New Zealand, February 17, 2017,

Tommi Vatanen

Contents

Preface	1
Contents	3
List of Publications	5
Author's Contribution	7
List of Abbreviations	9
1. Introduction	11
2. Methods for microbial community analysis	17
2.1 16S rRNA gene sequencing	17
2.2 Whole metagenome shotgun sequencing	19
2.2.1 Taxonomic profiling	20
2.2.2 Functional profiling	23
2.3 Downstream data analysis	26
2.3.1 Microbial diversity and stability	26
2.3.2 Ordination of microbial profiles	28
2.3.3 Statistical modeling	29
2.3.4 Classification	32
3. Materials and study designs	33
3.1 DIABIMMUNE study	33
3.2 LifeLines DEEP study	35
3.3 Fecal microbial transplantation in Crohn's disease	36
4. Results	37
4.1 Gut microbiome in early childhood	37

4.1.1	Microbial diversity is established during the first three years of life	37
4.1.2	Stability of the microbiome is decreased in infancy . .	39
4.1.3	The microbiome is modulated by diet	40
4.1.4	Antibiotic treatments perturb microbial taxa and genes	42
4.1.5	Population-level differences in microbiota composition	43
4.1.6	Variation in LPS structure impacts immune development	45
4.2	Adult gut microbiome	48
4.2.1	Factors associated with gut microbiome variation in Dutch population	49
4.2.2	FMT in active Crohn's disease	51
5.	Discussion	53
	References	61
	Publications	77

List of Publications

This thesis consists of an overview and of the following publications which are referred to in the text by their Roman numerals.

I T. Vatanen, A. D. Kostic, E. d’Hennezel, H. Siljander, E. A. Franzosa, M. Yassour, R. Kolde, H. Vlamakis, T. D. Arthur, A. M. Hämäläinen, A. Peet, V. Tillmann, R. Uibo, S. Mokurov, N. Dorshakova, J. Ilonen, S. M. Virtanen, S. J. Szabo, J. A. Porter, H. Lähdesmäki, C. Huttenhower, D. Gevers, T. W. Cullen, M. Knip, Diabimmune Study Group, R. J. Xavier. Variation in Microbiome LPS Immunogenicity Contributes to Autoimmunity in Humans. *Cell*, Volume 165, issue 4, pages 842-853, ISSN: 1097-4172, DOI: 10.1016/j.cell.2016.04.007, May 2016.

II A. D. Kostic, D. Gevers, H. Siljander, T. Vatanen, T. Hyötyläinen, A. M. Hämäläinen, A. Peet, V. Tillmann, P. Pöhö, I. Mattila, H. Lähdesmäki, E. A. Franzosa, O. Vaarala, M. de Goffau, H. Harmsen, J. Ilonen, S. M. Virtanen, C. B. Clish, M. Oresic, C. Huttenhower, M. Knip, Diabimmune Study Group, R. J. Xavier. The dynamics of the human infant gut microbiome in development and in progression toward type 1 diabetes. *Cell Host & Microbe*, Volume 17, issue 2, pages 260-273, ISSN: 1934-6069. DOI: 10.1016/j.chom.2015.01.001, Feb 2015.

III M. Yassour, T. Vatanen, H. Siljander, A. M. Hämäläinen, T. Härkönen, S. J. Ryhänen, E. A. Franzosa, H. Vlamakis, C. Huttenhower, D. Gevers, E. S. Lander, M. Knip, DIABIMMUNE Study Group, R. J. Xavier. Natural history of the infant gut microbiome and impact of antibiotic treatment on bacterial strain diversity and stability. *Science Transla-*

tional Medicine, Volume 8, issue 343, 343RA81, ISSN: 1946-6234, DOI: 10.1126/scitranslmed.aad0917, Jun 2016.

IV A. Zhernakova, A. Kurilshikov, M. J. Bonder, E. F. Tigchelaar, M. Schirmer, T. Vatanen, Z. Mujagic, A. V. Vila, G. Falony, S. Vieira-Silva, J. Wang, F. Imhann, E. Brandsma, S. A. Jankipersadsing, M. Joossens, M. C. Cenit, P. Deelen, M. A. Swertz, R. K. Weersma, E. J. M. Feskens, M. G. Netea, D. Gevers, D. Jonkers, L. Franke, Y. S. Aulchenko, C. Huttenhower, J. Raes, M. H. Hofker, R. J. Xavier, C. Wijmenga, J. Fu, LifeLines cohort study. Population-based metagenomics analysis reveals markers for gut microbiome composition and diversity. *Science*, Volume 352, issue 6285, pages 565-569, ISSN: 0036-8075, DOI: 10.1126/science.aad3369, Apr 2016.

V B. P. Vaughn, T. Vatanen, J. R. Allegretti, A. Bai, R. J. Xavier, J. Korzenik, D. Gevers, A. Ting, S. C. Robson, A. C. Moss. Increased Intestinal Microbial Diversity following Fecal Microbiota Transplant for Active Crohn's Disease. *Inflammatory Bowel Diseases*, Volume 22, issue 9, pages 2182-2190, ISSN: 1078-0998, DOI: 10.1097/MIB.0000000000000893, Sep 2016.

Author's Contribution

Publication I: “Variation in Microbiome LPS Immunogenicity Contributes to Autoimmunity in Humans”

Vatanen served as a project leader, contributed to the sequencing study design, sample management pipelines and method development, and conducted all 16S sequencing and WMS sequencing data analysis. Vatanen had a leading role in interpreting and consolidating the results, and in writing and revising the manuscript.

Publication II: “The dynamics of the human infant gut microbiome in development and in progression toward type 1 diabetes”

Vatanen participated in the method development and carried out analysis tasks for 16S sequencing and WMS sequencing data (diversity and stability analyses, associations with metabolomic data, metagenomic gene counts). Vatanen was involved in interpreting the consolidating the results and commented on the manuscript.

Publication III: “Natural history of the infant gut microbiome and impact of antibiotic treatment on bacterial strain diversity and stability”

Vatanen had a role in the method development for 16S and WMS sequencing data (strain-level analysis, stability analysis, AR gene quantifications). Vatanen commented on the manuscript.

Publication IV: “Population-based metagenomics analysis reveals markers for gut microbiome composition and diversity”

Vatanen contributed to analysis and method development of WMS sequencing data (MetaPhlAn and HUMAnN profiling, ordination analysis, pathway stability analysis). Vatanen commented on the manuscript.

Publication V: “Increased Intestinal Microbial Diversity following Fecal Microbiota Transplant for Active Crohn’s Disease”

Vatanen analyzed the WMS sequencing data and contributed on writing the manuscript. Interpretation of the data analysis results and writing the corresponding parts of the paper was conducted by Vatanen.

List of Abbreviations

AR	Antibiotic resistance
AUC	Area under [ROC] curve
<i>B. dorei</i>	<i>Bacteroides dorei</i>
<i>C. difficile</i>	<i>Clostridium difficile</i>
CD	Crohn's disease
CgA	Chromogranin A
CRP	C-reactive protein
<i>D. invisus</i>	<i>Dialister invisus</i>
DNA	Deoxyribonucleic acid
<i>E. coli</i>	<i>Escherichia coli</i>
FMT	Fecal microbial transplantation
GO	Gene ontology
HDL	High density lipoprotein
HBI	Harvey Bradshaw Index
HLA	Human leukocyte antigen
HMO	Human milk oligosaccharide
IBD	Inflammatory bowel disease
IBS	Inflammatory bowel syndrome
KL	Kullback-Leibler [divergence]
LPS	Lipopolysaccharide

MDS	Multidimensional scaling
mOTU	Metagenomic operational taxonomic unit (a software tool)
NCBI	The National Center for Biotechnology Information (USA)
NF- κ B	Nuclear factor κ -light-chain-enhancer of activated B cells
OTU	Operational taxonomic unit
PBMC	Peripheral blood mononuclear cell
PCA	Principal component analysis
PCoA	Principal coordinate analysis
PCR	Polymerase chain reaction
RNA	Ribonucleic acid
ROC	Receiver operating characteristic
RPK	Reads per kilobase
RPKM	Reads per kilobase per million reads
RF	Random forest
rRNA	Ribosomal ribonucleic acid
sIBDQ	Short Inflammatory Bowel Disease Questionnaire
SNP	Single nucleotide polymorphism
t-SNE	t-Distributed stochastic neighborhood embedding
T1D	Type 1 diabetes mellitus
TLR4	Toll-like receptor 4
WMS	Whole metagenome shotgun

1. Introduction

The human gastrointestinal tract is colonized by a dense and diverse community of resident commensal¹ microorganisms known as the gut microbiota². This neglected endocrine organ (O’Hara and Shanahan, 2006; Clarke et al., 2014) contains tens of trillions³ of bacterial cells—a number that is in the same order of magnitude with the number of cells in an adult human body (Sender et al., 2016). The gut microbiota has co-evolved with the host species to conduct functions that the host cells are lacking: they help digest food (Conlon and Bird, 2015; Sonnenburg and Bäckhed, 2016), produce important vitamins (Martens et al., 2002; Conly et al., 1994) and suppress the growth of harmful bacteria (Bäumler and Sperandio, 2016) among other things. Indeed, the number of genes in the human gut microbiome is estimated to be at least two orders of magnitude larger than the number of genes in the human genome (Qin et al., 2010). Based on the mutualism between the host and its microbiome, it has been suggested that we are all holobionts, entities which consist of ourselves and our microbiomes (Bordenstein and Theis, 2015; Gordon et al., 2013). However, the dynamics of the gut microbial ecosystems as well as the complexity of their functions and host-microbiome interactions remain largely unexplored.

Recent developments in DNA sequencing techniques—so called high-throughput sequencing techniques (Goodwin et al., 2016)—have revolutionized the research of microbial communities. In contrast to traditional culture-based approaches, where bacteria of interest are grown in laboratory conditions, DNA sequencing-based approaches allow characterizing

¹Commensalism refers to non-harmful coexistence.

²The term *microbiota* refers to a collection of microorganisms, whereas the term *microbiome* is used to refer to a collection of microbial genomes or genetic material.

³One trillion equals 10^{12} .

whole microbial communities in an unbiased manner. For example, traditional culture methods can capture only a fraction, 10%–30%, of the gut microbiota (Suau et al., 1999; Tannock, 2001). Together with other molecular techniques, including RNA sequencing, metabolomics and proteomics, high-throughput DNA sequencing has, figuratively speaking, triggered a golden age on the field of microbiology. Researchers on many branches of molecular biology are participating in this joint effort to better understand these miniature ecosystems, their complex dynamics and interactions with their environment.

The human gut microbiome matures to closely resemble the adult composition during the first 2–3 years of life (Tamburini et al., 2016; Backhed et al., 2015; Yatsunenko et al., 2012; Jakobsson et al., 2014), even though there is evidence suggesting that the microbiota is still developing towards toward adult-like configuration at the age of five (Cheng et al., 2015). There is some evidence of microbial exposure *in utero* (Aagaard et al., 2014; Fox and Eichelberger, 2015), but the large-scale microbial colonization of the gut begins at delivery. In vaginal birth, the neonate is exposed to the vaginal microbiota of the mother (Dominguez-Bello et al., 2010b), whereas in caesarean section the first bacterial contacts consists of common skin and environmental microbes (Tamburini et al., 2016). Afterwards, the establishment of the gut microbial community is largely shaped by oligosaccharides and microbes in human milk which usually constitutes the cornerstone of neonate’s diet during the first months of life (McGuire and McGuire, 2015; Jost et al., 2015; Gomez-Gallego et al., 2016). The maturation of the gut microbiota is further influenced by the growing complexity of diet, potential antibiotic exposures, the host genetics and numerous other environmental factors. The assembly and transmission of maternal and early childhood microbial communities is an area of active research with many lessons to be learned (Charbonneau et al., 2016).

Mounting evidence connects aberrations in infant and childhood gut microbiota with not only immunological disorders including type 1 diabetes (T1D)(Rewers and Ludvigsson, 2016; Knip and Siljander, 2016), asthma (Arrieta et al., 2015; Abrahamsson et al., 2014), juvenile rheumatoid arthritis (Arvonen et al., 2016), allergic disease (Simonyte Sjodin et al., 2016) and pediatric IBD (Lewis et al., 2015) but also with conditions such as obesity (Dogra et al., 2015), eczema (Nylund et al., 2013; Abrahamsson et al., 2012) and even autism spectrum disorder (Kang et al.,

2013). According to the hygiene hypothesis, exposure to specific microorganisms early in life benefits the developing immune system and protects against immune-mediated diseases (Strachan, 1989). Indeed, the complex relationship between the microbiome and both innate (Thaiss et al., 2016) and adaptive (Honda and Littman, 2016) immune systems is an active research area. The hygiene hypothesis is supported by a substantial amount of correlative evidence, reviewed in (Bach, 2002; Bach and Chateaunoud, 2012), but the distinction between beneficial and harmful microorganisms as well as mechanisms underlying their effects are still poorly understood.

The gut microbial communities in healthy adults are stable over long periods of time (Faith et al., 2013; Coyte et al., 2015; Rajilić-Stojanović et al., 2013) but exhibit significant inter-personal (Human Microbiome Project, 2012; Franzosa et al., 2015) and inter-cultural (Brito et al., 2016) variation. Aberrations in the adult gut microbiota have been linked to various conditions, such as inflammatory bowel disease (IBD) (Frank et al., 2007), type 2 diabetes (Pedersen et al., 2016), obesity (Ley et al., 2006), and even depression (Naseribafrouei et al., 2014). Eventually, manipulations of the gut microbiota may prove to help treating at least some of these conditions but the current understanding of what constitutes a healthy gut microbiota and how it can be manipulated is still very limited (Gilbert et al., 2016). Among the first examples of microbiome based therapies is the fecal microbial transplantation (FMT) from healthy donors which has proven to be an effective treatment for restoring the healthy gut microbiota in recurring *Clostridium difficile* infections (Kelly et al., 2012; Rohlke and Stollman, 2012). Using FMT to treat other conditions with microbial aberrations is an active research area (Kelly et al., 2015) with ongoing studies on different subtypes of IBD.

The nature and amount of data generated by DNA sequencing techniques has granted computational scientists a central role in many microbiome sequencing studies. DNA sequencing results in usually millions of short, roughly 100 basepairs long, DNA fragments, which need to be processed computationally to gain any biological insights. Usually, sequencing reads are first quality controlled computationally and then either aligned against reference databases or assembled *de novo* to form longer segments called contigs that represent a consensus region of DNA. Computational analyses enable profiling microbial communities taxonomically as well as quantifying metabolic pathways and microbial genes.

These information together with sample summary statistics, such as microbial community diversity, can be then analyzed statistically and compared with physiological and other metadata that has been collected.

The aim of this thesis was to improve the understanding of infant and adult gut microbiome while pushing the boundaries of the modern molecular and computational microbiome profiling techniques. By leveraging the existing databases of microbial genomes and proteins, and whole metagenome shotgun sequencing data of several large cohort studies, we developed new methodology to assess microbial diversity and stability on the strain level, and improved methods to assess functional profiles of the microbiome in connection to contributing organisms. In DIABIMMUNE study (Publications I–III) we set out to explain mechanisms behind the hygiene hypothesis by studying the gut microbiome of infants in Finland, Estonia and Russian Karelia—areas with contrasting occurrences of autoimmune diseases—from birth until the age of three. In Publication I, we identified a mechanism of microbial origin—differences in lipopolysaccharide immunogenicity—which may contribute to the difference in autoimmunity between Finland and Estonia versus Russian Karelia. In Publication II, we focused on T1D in case-control setting and found that infants progressing to T1D before the age of three harbored gut microbiome with decreased diversity, indicative of dysbiosis, before the disease diagnosis. In Publication III, we studied the effects of recurrent antibiotic treatments and saw decreased microbial diversity as well as increase in antibiotic resistance genes as a consequence of antibiotics. Together these three Publications provide, to date, the largest longitudinal functional profile of the infant gut microbiome. In Publications IV and V the focus was on healthy and disrupted adult gut microbiome, respectively. In Publication IV, we systematically analyzed 207 intrinsic and exogenous factors for connections to the gut microbiome of 1179 Dutch adults. In Publication V, we conducted the first prospective FMT study for adults with active Crohn’s disease (CD).

This thesis contributes to the understanding of the human gut microbiome and its interplay with immune system and surrounding environment in infancy, adulthood, health and disease. The content is structured as follows. In Chapter 2, I introduce the methods and techniques used in this thesis, with an emphasis on the computational analysis. In chapter 3, I discuss the materials and motivate the study designs of the individual studies. This is followed by the most important results of this thesis in

Chapter 4. Finally, I summarize the findings and discuss both theoretical and practical implications of this thesis in Chapter 5.

2. Methods for microbial community analysis

In the era of next-generation sequencing (reviewed in Goodwin et al. (2016)), microbial community profiling using DNA sequencing techniques has become a common practice. In this task, where either specific informative amplicons or all genetic material is sequenced, very little can be done with the raw sequencing data itself. Instead, downstream bioinformatic analyses are needed to understand the connection between the sequencing data—short reads—and the microbial communities in question. On high level, these analyses usually aim at answering two questions, (i) “who is there?”, that is, taxonomic profiling, and (ii) “what are they doing?”, that is, functional profiling. The first question can be addressed using either amplicon sequencing, which is usually 16S rRNA gene sequencing, or whole metagenome shotgun (WMS) sequencing. To answer the second question in detail, that is to generate functional profiles for microbial communities, WMS sequencing data is needed.

In this Chapter, I describe the experimental and computational methods used in analyzing 16S rRNA gene and WMS sequencing data in this thesis. Since WMS sequencing data has a more important role in this thesis, WMS sequencing analysis techniques are given more detailed scrutiny. For all analysis tasks, the most important complementary and/or competing methods are also briefly described. I also cover computational analysis tasks including data visualization and statistical testing.

2.1 16S rRNA gene sequencing

The 16S ribosomal RNA (rRNA) gene is encoding a part of the RNA component of the ribosome, a cellular component responsible for the protein synthesis in all living organisms. As such, the gene is found in all living organisms, and its variable, non-conserved regions (so-called V-regions)

can be used to compare different organisms in terms of evolutionary distance. 16S rRNA gene sequencing (later 16S sequencing) is the most common experimental procedure for characterizing microbial communities taxonomically, that is, tackling the question “who is there?”. In 16S sequencing, a chosen V-region(s) of 16S rRNA gene is amplified using polymerase chain reaction (PCR) and region-specific primers. In our experiments, we sequenced the V4 region of the 16S rRNA gene using Illumina HiSeq 2500 instruments by following the protocol described by Caporaso et al. (2012). More detailed experimental procedures and information about sample handling are given in accompanying publications (Publications I-III). In this thesis, all 16S sequencing data processing was conducted using QIIME 1.8.0 (Caporaso et al., 2010; Kuczynski et al., 2012), an open-source bioinformatics pipeline for performing microbiome analysis from raw 16S sequencing data, which is described below.

16S sequence data needs to be analyzed computationally to obtain any information about the microbial communities in question. An important step of processing 16S sequencing data involves clustering all sequences in groups based on sequence similarity. These clusters represent operational taxonomic units (OTUs), the lowest-level phylotypes detected by 16S sequencing. This step can be performed either in unsupervised manner using tools such as CD-HIT (Fu et al., 2012) and mothur (Schloss et al., 2009) incorporated in QIIME, or using a reference database of OTU representatives, in which case this step does not involve clustering but classification, strictly speaking. In this thesis, the OTU picking step was conducted using GreenGenes reference database (McDonald et al., 2012) with 97 % sequence similarity OTUs. Usually 97 % sequence similarity is considered to approximate species-level phylotypes. However, due to poor species-level annotations in reference databases and relatively short length of the V4 region (~255 basepairs), we constrained our 16S data analysis on genus level and above (Soergel et al., 2012).

Recently, many improvements in 16S sequencing analysis pipeline have been proposed. Edgar (2013) proposed a UPARSE pipeline which includes quality filtering, dereplication, discarding singletons, and *de novo* OTU clustering. More recently, Edgar and Flyvbjerg (2015) proposed improvements in 16S sequence analysis by filtering reads with high expected error count, by assembling overlapping read pairs, and by exploiting sequence abundances for correcting sequencing errors. UPARSE algorithm perform

chimera¹ filtering and OTU clustering simultaneously, which improves the accuracy of OTU picking. In this pipeline, taxonomy is assigned after the OTU clustering step and depending on the reference database a number of OTUs may remain unannotated. Callahan et al. (2016) provide another software package called DADA2 for modeling and correcting amplicon sequencing errors in Illumina data.

2.2 Whole metagenome shotgun sequencing

Where 16S sequencing provides taxonomic information about the microbial communities, WMS sequencing aims at sequencing all genetic material in a given sample. Therefore, this procedure provides near-complete view of microbial genes and genetic potential in a given sample. However, it is also more costly compared to 16S sequencing.

WMS sequencing data can be analyzed using two different and complementary approaches: read-based (and assembly-free) and assembly-based methods. In read-based methods, raw sequencing reads are analyzed together with existing genomes and gene catalogs to better understand taxonomic composition and genetic functional potential of the given microbial community. Assembly-based methods start by genome assembly, where short reads are first assembled *de novo* to form longer contigs (Howe and Chain, 2015). These contigs can then be annotated for coding sequences, and their taxonomic origin and functional potential can be assessed in many different ways. In this thesis, all work was limited to read-based methods and assembly-based methods are not covered.

The taxonomic composition and the functional potential of the microbiome are in many ways orthogonal questions. Sometimes strain level variation can have significant functional implications—for example, addition of a single genetic cassette, a virulence factor, may implicate pathogenicity—and other times taxonomically distant species can occupy the same functional niche. WMS sequencing data enables characterizing the functional potential of the microbial community, therefore providing a complementary view to the taxonomic composition. Below, taxonomic and functional profiling approaches for WMS sequencing data are described in detail.

¹Chimera or chimeric reads are artefacts of the PCR process, which contain DNA sequences originating from two or more genomes.

2.2.1 Taxonomic profiling

While 16S sequencing provides limited resolution in taxonomic assignments, WMS sequencing data provides means for assigning taxonomies on species and strain level. Given enough sequencing depth², WMS sequencing data contains reads covering all genomes in a given microbial community, providing means for detecting single nucleotide polymorphism (SNP) level differences between bacterial strains. In practice, reliable SNP detection requires at least 10x coverage for a given genome, which translates roughly to 1 % relative abundance of the given species with sequencing depth of 5 million short 100 nucleotide reads (Luo et al., 2015).

In this thesis, MetaPhlAn2 (Truong et al., 2015) was used for taxonomic profiling of WMS sequencing data. Additionally, WMS sequencing data was analyzed using ConStrains (Luo et al., 2015) for strain tracking in Publications I and III. In This section, I will describe these methods together with their closest competitors.

MetaPhlAn for taxonomic profiling

A clade is a group of organisms representing a single branch in a phylogenetic tree. MetaPhlAn utilized clade-specific marker sequences to estimate abundances of clades, which can be as specific as species and strains, or as generic as phyla and domains. Clade-specific marker sequences are short DNA segments of coding sequences, which (i) are highly conserved within the clade, and (ii) are not found in any genomes outside the clade. These properties guarantee that observing the marker sequences indicates presence of the clade in question, given vast enough reference collection. In theory, the definition of marker sequences, especially property (i), is sensitive to the size of database of sequenced genomes; there can always be unknown, yet-to-be-sequenced strains which do not carry any given marker sequence defined for the clade in question. However, experiments conducted by Segata et al. (2012) showed that in practice this approach is effective even with an incomplete collection of reference genomes. In the latest MetaPhlAn2—the second generation version of MetaPhlAn—marker database, there are approximately one million unique clade-specific marker sequences identified from roughly 17,000 reference genomes (~13,500 bacterial and archaeal, ~3,500 viral, and

²Generally, sequencing depth (or coverage) refers to the number of reads covering any given nucleotide in targeted genome. In WMS, sequencing depth usually refers to total number of short reads generated.

~110 eukaryotic).

MetaPhlAn2 aligns metagenomic reads against pre-computed marker sequence database using Bowtie2 (Langmead and Salzberg, 2012), which provides more than 10 times faster run times compared to BLAST alignment in MetaPhlAn. After alignment, the total number of reads mapped to each clade is normalized by the length of the marker sequences in question, and relative abundances are calculated by dividing the normalized read counts by the sum of all read counts. Details of the mapping and relative abundance estimation steps are given in Truong et al. (2015).

Other methods conducting taxonomic profiling for WMS sequencing data include mOTU (metagenomic operational taxonomic unit) (Sunagawa et al., 2013), Kraken (Davis et al., 2013) and MEGAN (Huson et al., 2016). mOTU utilizes universal marker genes, which occur in single copy in the vast majority of known organisms, to detect and quantify both known species and also clades that still lack reference information. Additionally, mOTU can exploit covariance data across multiple samples to combine different marker genes (mOTUs) into mOTU linkage groups. Kraken is a k-mer based approach, utilizing exact alignment of k-mers and a novel classification algorithm to assign taxonomic labels to short DNA reads. MEGAN is a reference based tool for both taxonomic and functional profiling. It conducts taxonomic profiling by aligning WMS reads directly against NCBI reference genomes and assigning taxonomy for each read according to the lowest common ancestor of all clades with a significant alignment (Huson et al., 2016). Up to date, there are only a few independent studies comparing methods for taxonomic profiling approaches for WMS data. In one such study, Lindgreen et al. (2016) conclude that "Picking a single best tool is not straightforward" and provide different performance metrics to help researchers decide based on their own demands.

ConStrains for strain-level profiling

Occasionally, it may be possible to detect and classify microbial strains using the marker gene approach described above. To achieve more generic and sensitive strain tracking, one needs to look at the sequence data on single nucleotide resolution. To obtain this, ConStrains (Luo et al., 2015) exploits polymorphism patterns—that is, conducts SNP haplotyping—on a set of bacterial genes found universally on the genomes of the given species. More specifically, ConStrains operates on two steps: (i) identify

species with enough coverage to detect and quantify SNPs, and (ii) transform individual SNPs into species-specific SNP profiles representing individual strains.

In the first step, ConStrains uses MetaPhlAn (Segata et al., 2012) for species level compositional profiling. Species-specific sequencing coverage is determined by multiplying the total sequencing depth (number of nucleotides sequenced) with the relative abundance of a given species and dividing the product with the average genome length of the species. By default, species with $> 10\times$ coverage are selected for downstream processing. Given these species, ConStrains constructs a custom database of marker genes using PhyloPhlAn marker set (Segata et al., 2013) and aligns the raw sequencing reads against this marker gene database. Resulting alignments are processed using SAMtools (Li et al., 2009) for SNP detection. In this step, the reference gene sequences are no longer used to minimize reference dependency.

In the second step, individual SNPs are combined into SNP profiles representing strains. Each strain is represented by its unique SNP barcode, termed uniGcode, spanning hundreds of genes. To determine the number of strains per species and to derive strain specific uniGcodes, ConStrains relies on clustering based approaches termed SNP-flow and SNP-type algorithms. Briefly, these algorithms use per-species, per-sample nucleotide occurrence proportions to determining strains-specific uniGcodes. Finally, ConStrains will match SNP profiles between samples—for example, in the case of longitudinal sampling—to obtain strain tracking across samples. Detailed description of ConStrains algorithm is given in Luo et al. (2015) and accompanying Online Methods.

As WMS sequencing is becoming commonplace, there are many new tools aiming at strain level profiling, such as Sigma (Ahn et al., 2015), PathoScope (Hong et al., 2014) and StrainPhlAn (yet unpublished, based on Truong et al. (2015)). Sigma and PathoScope rely on available reference strain collections and are therefore limited to detect only known strains. Similar to ConStrains, StrainPhlAn detects strain based on SNP patterns on core genes. However, StrainPhlAn will only detect the dominant strain per sample, which limits possible downstream analysis tasks. For example, it is not possible to measure haplotype diversity (see Section 2.3.1) using such strain profiling approach. Li et al. (2016) have conducted SNP based strain tracking using mOTU based method operating with universal marker genes. There are also methods for strain tracking based on

amplicon sequencing data, such as 16S sequencing. Oligotyping (Eren et al., 2013) and minimum entropy decomposition (Eren et al., 2015) can successfully detect strain specific SNP patterns on 16S sequencing data. However, they are limited to the SNPs within the span of the amplicon in question and thus provide very limited view of the strains in any given microbial community.

Strain profiling can reveal strain level shifts, which are left unobserved on species-level taxonomic profiling. In Publication I, we measured within-species, within-subject stability on strain level using ConStrains strain profiles and Bray-Curtis dissimilarity (see Equation 2.5 below). This analysis led us to observe occasional “strain sweeps”, in which the dominant strain was replaced by a new dominant strain between the two samples.

2.2.2 Functional profiling

Functional interpretation of WMS sequencing data is key to connecting microbial communities with their host or surrounding environment. This task involves identifying different metabolic pathways or functions present or absent in a given microbial community and determining their relative abundances. In this thesis, this task is tackled using the second generation version of the HMP Unified Metabolic Analysis Network (HUMAN_N), that is HUMAN_N2, described below. Additionally, ShortBRED was used to identify and to quantify antibiotic resistance (AR) genes in Publication III.

HUMAN_N2

Both the original HUMAN_N (Abubucker et al., 2012), and the second generation version (HUMAN_N2) described here are assembly-free methods for constructing functional profiles for microbial communities based on WMS sequencing data. Given WMS sequencing data—that is DNA reads—this method provides information on the functional potential of the communities, while the same analysis based on RNA-sequencing data provides profiles of actual functional activities.

A pangenome is a collection of genes found in any sequenced isolate of a bacterial species (Huang et al., 2014). HUMAN_N2 start by aligning quality controlled and human genome filtered WMS sequencing reads against a collection of pangenomes. In HUMAN_N2, genes on the pangenomes have been pre-annotated using a comprehensive, non-redundant protein sequence database UniRef50 (Suzek et al., 2015) to their respective UniRef50

gene families. Ambiguity in the alignment step is minimized by selecting the pangenomes used using an upstream taxonomic profiling conducted using MetaPhlAN2; only pangenomes of species with $>0.1\%$ relative abundance are used. Alignment to annotated pangenomes provides taxonomically stratified functional information; each aligned sequencing read is functionally annotated through UniRef50 with taxonomic origin given by the pangenomes.

WMS sequencing reads with no alignment in the pangenomes alignment step are mapped separately to the entire UniRef50 database by translated search with DIAMOND (Buchfink et al., 2015). This step allows detecting contributions from organisms that are lowly abundant or otherwise undetected—for example, species with no sequenced genome—in the taxonomic profiling step. Hits from both alignment steps described above are weighted based on alignment quality and target sequence length, and combined to produce community totals for each gene family in reads per kilobase (RPK) units.

The above process results in quantifications of typically tens of thousands of microbial UniRef50 gene families. These families can be combined to more meaningful and interpretable functional categories using different ontologies, such as Gene Ontology (GO) (Gene Ontology Consortium, 2015) or MetaCyc (Caspi et al., 2012). In Publications I, IV and V, we mapped UniRef50 gene families to GO terms using the mapping between UniProt proteins—UniRef50 gene families is a subset of UniProt proteins—and GO terms (Dimmer et al., 2012). We further isolated a subset of categories in GO terms by following previous work in Huang et al. (2007) and Zhou et al. (2002). We concentrated in a subset of “informative” GO terms associated with $> k$ proteins for which all descendant terms were associated with $< k$ proteins. In Publication I, we used $k = 2000$, which equates to roughly 1 GO term per 5,000 UniRef50 protein families. In that case, this procedure yielded a comprehensive but manageable set of 247 non-redundant GO Biological Process terms for subsequent analysis. By the nature of their construction, informative GO terms tend to provide more resolution for well-conserved and well-studied processes, which are annotated to many proteins, and place less focus on highly specific processes associated with only a small number of proteins.

There are many other tools for functional characterization of WMS sequencing data, which operate by searching the reads from a sequenced microbial community against pre-annotated databases of protein sequences.

Such approaches include IMG/M (Markowitz et al., 2014), MG-RAST (Glass et al., 2010) and MEGAN (Mitra et al., 2011). Since these approaches use sequenced genomes to build their databases, they may be biased towards known species and under-perform when analyzing microbial communities with many novel organisms. In HUMAnN2, the translated search against UniRef50 database aims at tackling this weakness.

ShortBRED

In addition to quantifying gene families and broad functional pathways, one may be interested in tracking presence/absence of specific microbial proteins and genes encoding them. ShortBRED (Short, Better Representative Extract Dataset) is a tool for high-specificity targeted functional profiling for WMS samples (Kaminski et al., 2015). In this thesis, we used ShortBRED for detecting and quantifying AR genes in Publication III. ShortBRED operates with short, highly representative protein sequences, but in contrast to MetaPhlAn marker sequences, ShortBRED markers are represented as amino acid sequences. The tool enables one (i) to identify such marker sequences for any proteins of interest, and (ii) to quantify these proteins in microbial communities analyzed using WMS sequencing.

In the identification step given a set of proteins of interest, ShortBRED identifies peptide marker sequences for these proteins against any comprehensive protein database by several clustering and alignment steps using CD-HIT (Fu et al., 2012), MUSCLE (Edgar, 2004) and BLAST (Altschul et al., 1990). Proteins are first clustered using CH-HIT, and consensus sequences are created for clustered protein families using multiple sequence alignment by MUSCLE. These consensus sequences are then queried against each other and against comprehensive protein database to identify marker sequences using BLAST. In this thesis, UniRef90 was used as the comprehensive protein database (Suzek et al., 2015).

In the quantification step, ShortBRED maps WMS sequencing reads against the marker sequences by translated search using USEARCH (Edgar, 2010). Hit counts are normalized by adjusting for average read length, marker length and sequencing depth to produce protein quantity estimates in reads per kilobase of reference sequence per million sample reads (RPKM).

2.3 Downstream data analysis

After taxonomic and functional profiling steps described above, one usually needs to conduct many additional computational analysis steps in order to answer scientific questions of interest. These tasks may include assessment of community diversity, comparisons of samples, data ordination or visualization, statistical modeling, classification and clustering. In this section, I describe the most important downstream data analysis methods used in this thesis.

2.3.1 Microbial diversity and stability

The within-sample diversity, or α -diversity, has been shown to be an important biomarker for disruptions in microbial communities. The condition of decreased α -diversity is sometimes referred as dysbiosis, referring to unbalanced microbiota. Specifically, intestinal dysbiosis has been reported in many diseases, such as IBD (Matsuoka and Kanai, 2015), colorectal cancer (Ahn et al., 2013) and T1D (Publication II).

In microbial community profiling, the (dis)similarity between two samples or microbial communities is usually called β -diversity. As such, it can be used to measure stability of the microbiota over time, given longitudinal sampling. Below, I describe the most common measures of α - and β -diversity.

Measures of α -diversity

The measure of α -diversity is related to the count of different bacteria and their distribution in the community. Sometimes α -diversity is measured by the mere count of distinct microbial species in the community, often referred to as richness. In sequencing based community profiling, there may often be lowly abundant but undetected bacteria, which means that richness cannot be observed directly. Instead, one can estimate the number of unseen taxa by looking at the distribution of observed taxa. Chao1 index (S_{est}) is a nonparametric estimate of richness based on frequencies of singleton and doubleton taxonomic groups (Chao, 1984):

$$S_{est} = S_{obs} + \frac{f_1^2}{2f_2}, \quad (2.1)$$

where S_{obs} is the number of observed species, f_1 is the number of singleton species and f_2 is the number of doubleton species. In Publication II, we used chao1 estimate to measure α -diversity based on 16S sequencing data.

Instead of focusing on the mere number of microbial species, one can measure α -diversity by assessing the abundance distribution of microbes in the community. Intuitively, of two communities with the same number of species, the one where the species are present in comparable abundances is more diverse compared to the other where one microbe dominates the community. Following this reasoning, any information theoretic measure of entropy can be used as an estimate of α -diversity. One commonly used measure is Shannon's diversity index (Shannon, 1948)

$$H = - \sum_{i=1}^T p_i \ln p_i, \quad (2.2)$$

where p_i are relative abundance of microbial taxa and T is the total number of taxa. In Publications I, III, IV and V Shannon's diversity index was used for estimating microbial α -diversity.

Similar to sample-specific microbial community diversity, it is possible to evaluate species-specific strain diversity based on ConStrains strain profiles. In Publication I, we formulated haplotype diversity score

$$H_{haplotype} = 1 - \sum_{i=1}^S p_i^2, \quad (2.3)$$

where p_i denotes the within-species abundance of strain i and S is the total number of strains. Haplotype diversity measures species- and sample-specific strain diversity and it is bounded between $[0,1]$. This measure was motivated by the concept of heterozygosity in population genetics. There is a convenient probabilistic interpretation for haplotype diversity score: given two randomly sampled bacterial cells from species X in sample Y , the corresponding haplotype diversity score reflects the probability that the two cells are derived from different bacterial strains.

Measures of β -diversity

In theory, any (dis)similarity measure, such as Euclidean distance, can be used to measure β -diversity between two microbial communities. However, given the proportional nature of profiles we are dealing with, the field of ecology offers many better-suited measures, which take the compositionality of the data into account.

Jaccard index measures the proportion of shared taxa in two microbial communities A and B :

$$J(A, B) = \frac{|A \cup B|}{|A \cap B|}. \quad (2.4)$$

As such, Jaccard index only concerns the presence or the absence of microbial taxa rather than their relative abundances. When used for mea-

asuring microbial stability in longitudinal setting, Jaccard index measures the introduction of new microbes and the elimination of existing community members. Using Jaccard index for measuring stability, (Faith et al., 2013) found that on average roughly 60 % microbial strains remain stable in adult gut microbiota over the course of five years. We used Jaccard index for measuring the stability of infant gut microbiota in Publications I-III.

Another commonly used measure for β -diversity is Bray-Curtis dissimilarity

$$BC_{ij} = \frac{\sum_{t=1}^T |x_{ti} - x_{tj}|}{\sum_{t=1}^T |x_{ti} + x_{tj}|}, \quad (2.5)$$

where x_{ti} is a count or relative abundance of taxon t in sample i and T is the total number of taxa. Bray-Curtis dissimilarity can be used for both count data, such as OTU counts in 16S sequencing, and relative abundance data. Bray-Curtis dissimilarity was used in all β -diversity based analyses in all Publications except for the longitudinal stability analyses described above.

2.3.2 Ordination of microbial profiles

Data ordination or visualization, where data points or other objects of interest are presented as point on a, usually two-dimensional, surface, is an important element of data analysis. Such analysis can provide a compact high-level view of the data convenient for further hypothesis generation and data summarization. In microbial community analysis, any data ordination graph is usually referred to as principal coordinate analysis (PCoA) plot, which is not to be confused with another dimensionality reduction and data visualization method principal component analysis (PCA)(Pearson, 1901). Operating with data covariance matrix, PCA finds a linear mapping to a lower-dimensional space in such way that maximum amount of variation is preserved. When operating with microbial community data, it is often desirable to use other measures for similarity rather than covariance (see Section 2.3.1 above). Different variants of multidimensional scaling (MDS)(Borg and Groenen, 2005), such as non-metric MDS, and t-Distributed Stochastic Neighbor Embedding (t-SNE)(van der Maaten and Hinton, 2008) enable data ordination based on any (dis)similarity matrix and are popular for generating PCoA plots.

In this thesis, visualization of stool samples, that is microbial communities, was conducted using t-SNE (van der Maaten and Hinton, 2008),

which is a state-of-the-art method for visualizing structure in large data sets. Briefly, t-SNE operates by minimizing the Kullback-Leiber (KL) divergence between two distributions, p_{ij} and q_{ij} :

$$KL(P||Q) = \sum_{i,j} p_{i,j} \log \frac{p_{i,j}}{q_{i,j}}, \quad (2.6)$$

where p_{ij} is a matrix of probabilities computed given the original data points and q_{ij} is a matrix of probabilities given the distances of the data points in the output mapping, that is the resulting visualization of the data points. In the case of microbial community data, any β -diversity matrix naturally takes the form of $p_{i,j}$'s (Van der Maaten and Hinton, 2012). The task is to find q_{ij} such that it is as close to p_{ij} as possible, that is, the KL divergence between these distributions is minimized. This is achieved iteratively using gradient descent, and the process is further accelerated using Barnes-Hut algorithm in the latest implementation (van der Maaten, 2014). In this thesis, Bray-Curtis dissimilarity matrix (Equation 2.5) was used as $p_{i,j}$.

2.3.3 Statistical modeling

Biological data collected from human subjects, such as the microbiome data studied in this thesis, is usually complex and noisy. Subjects under study are genetically different and they are exposed to environment with plethora of factors potentially contributing to the biological measurements in question. There may also be technical factors producing biases in the measurements. This is to say that the inter-sample variation observed is always a sum of all these factors and never reflects the comparisons of scientific interest, such as case-control comparisons, alone.

Statistical models provide a convenient mathematical framework for accounting any confounding factors while testing the scientific hypotheses at the same time. In this thesis, a statistical tool called Multivariate Association with Linear Models (MaAsLin) (Morgan et al., 2012) was used for all statistical modeling. Below, I describe the components of MaAsLin one by one.

MaAsLin

MaAsLin is a linear modeling system adapted for microbial community data and it consists of multiple steps. The goal of the first three steps is to filter and transform the data such that it better fits the assumptions of the linear mixed effects model used for modeling the data in step four.

Final step conducts multiple testing correction.

First, MaAsLin conducts outlier removal using Grubbs' test (Grubbs, 1950). This is important since both microbial relative abundances and clinical metadata often contain outliers which violate the assumptions—such as assumption of normally distributed data—of linear models and may thus result in spurious results. Grubbs' test is based on the assumption of normality and is therefore well-suited for this task.

Second, MaAsLin applies variance stabilizing arcsine square root transformation

$$y_{i,transformed} = \arcsin \sqrt{y_i}, \quad (2.7)$$

which is also known as arcsine transformation (Sokal and Rohlf, 1995). This will pull out the ends of the distribution making the data more normally distributed. Arcsine transformation is preferred over other transformations, such as logarithmic and logistic transformations, since it can be directly applied to data with zeros and ones, whereas adding pseudo counts or other steps need to be applied before these other transformations are applicable.

Third, MaAsLin reduces the complexity of the linear model to be fitted next by conducting a feature selection step per each microbial taxon. This means that only a minimal set of predictors, which are identified to be associated with the given microbial taxon, is used in each linear model fitted. Features are tested for a tentative association with the microbial taxon one by one. For continuous predictors associatedness is measured by correlation and for binary and categorical predictors non-parametric test is used to test for any association between the predictor and microbial taxon in question. By including all predictors with tentative association in the model simultaneously, MaAsLin provides a data-driven approach for associating strongest predictors with microbial taxa while other, possibly correlated and confounding predictors are included in the model to explain away their contribution to the variation in the microbial taxon in question. To resolve possible confounding effects more explicitly, one has to directly measure correlations between the predictors.

Fourth, MaAsLin fits a linear mixed effect model one microbial taxon at a time. A linear model is a statistical model where the observed quantities y_i —in this case arcsine transformed microbial relative abundances—are represented in terms of J predictors $x_{i,j}$ with a linear relationship to the target variable

$$y_i = \beta_1 x_{i,1} + \dots + \beta_J x_{i,J} + \varepsilon_i, \quad i = 1, \dots, n. \quad (2.8)$$

Variables selected in the previous step are used as predictors. There may be two kinds of predictors or effects, fixed effects and random effects, hence the name mixed effects model.

A random effect(s) model (Skron dal and Rabe-Hesketh, 2004) assumes that the data being analyzed contains a hierarchical structure and it is thus directly related to Bayesian hierarchical models (Gelman et al., 2014). For example, in the case of longitudinal study of human subjects, repeated measurements from the same test subject are correlated and thus violate assumptions of many statistical models, which assume independent and identically distributed (i.i.d.) data. In a random effect model, these correlations are explicitly modeled using a hierarchical structure where the subjects themselves represent the lower level of the hierarchy, and the subjects are members of a population that is the higher level of the hierarchy. As such, a random effect term is well-suited for modeling within-subject correlations in longitudinal study designs, and in biostatistics, terms “random effect” and “subject-specific effect” are often used interchangeably (Diggle, 2002).

In biostatistics, the term “fixed effect” is used to refer to terms other than random effects in a linear mixed effect model. This is in contrast to, for example, econometrics, where a fixed effects model is another type of linear model for longitudinal panel data analysis, where subject specific effects are assumed to correlate with other predictors (Wooldridge, 2010). Gardiner et al. (2009) provide a concise description of fixed effects and random effects with formal definitions and illustrative examples.

A linear mixed effects model is a statistical model containing both fixed effects and random effects and as such it is useful in longitudinal studies where repeated measurements are made on the same test subjects. MaAsLin fits a linear mixed effects model for one microbial taxon at a time using a penalized quasi-likelihood approach (Breslow and Clayton, 1993) provided in R package *MASS*. In the resulting model, each fixed effect term is given a p-value which is computed using respective degrees of freedom and the ratio between the fixed effect estimate and its standard error, t-value, as described in Pinheiro and Bates (2006).

Finally, MaAsLin conducts false discovery rate correction for all p-values obtained from the linear models fitted in the previous step, using Benjamini-Hochberg method (Benjamini and Hochberg, 1995). This correction involves p-values for all fixed effect terms that have passed the feature selection step for each taxon.

2.3.4 Classification

Classification is a branch of supervised machine learning where labeled training data are used to learn a model which predicts the class of new data. In addition to prediction, classification models can be used to evaluate separability between groups of data among other things. In this thesis, supervised learning method called Random Forest was used for classifying samples in Publication I.

Random Forest

Random Forest (RF) is a supervised learning technique which can be used to conduct classification and regression among other things (Breiman, 2001). It is a robust technique—that is, it is difficult to do over-fitting with RFs—with loose assumptions about the input data. These reasons have probably contributed to its current popularity on many fields applying machine learning techniques, including microbiome research.

RFs operate by training an ensemble of decision trees (Utgoff, 1989) while applying bootstrap aggregating (bagging)(Breiman, 1996) for the training data of each decision tree separately. This means that the training set for each decision tree in a RF is generated by sampling from the original input data with replacement. RFs also apply feature bagging, which means that a random subset of features is used at each candidate split in the learning process.

When used for prediction, RFs classify new data using majority vote principle, which means that the output is the mode of the classes of the individual trees. In regression task, the output is usually the mean of the individual trees' predictions. RFs also enable evaluating the performance of the model without separate test data or cross-validation procedure. This is achieved by using out-of-bag predictions—that is, predictions where a subset of trees that did not have the sample in question in their training bootstrap sample are used to generate the output—to compute classification or regression errors.

3. Materials and study designs

In this Chapter, I describe the research hypotheses and study cohorts used in this thesis. Specifically, I lay out the motivation behind the cohorts and describe the experimental methods used for analyzing the stool samples collected in our studies.

3.1 DIABIMMUNE study

In the DIABIMMUNE study, 678 infants and their families from Finland, Estonia and Russian Karelia were recruited in order to study how differences in lifestyle between these regions affect the early development of infants' immune systems. All infants were followed-up from birth until three years of age. Enrollment was based on human leukocyte antigen (HLA) genotyping, which can be used to predict susceptibility to autoimmune diseases and allergies (Pociot and Lernmark, 2016; Larizza et al., 2012; Sollid and Thorsby, 1993). HLA genotyping was used in order to control for genetic differences between the populations under study, allowing us to attribute any observed phenotypic differences between the populations to environmental factors. In this thesis, these three populations were used as a “living laboratory” to study mechanisms behind the hygiene hypothesis.

There is a steep gradient in incidence of autoimmune diseases and allergies between Finland and Russian Karelia despite the geographic proximity and the genetic similarity between the populations. For instance, the incidence of type 1 diabetes (T1D) is 5–6 fold higher (Kondrashova et al., 2008a,b) and incidence of allergic diseases are 2-6 fold higher (Seiskari et al., 2007) in Finland relative to Russian Karelia across the border. In Estonia, the incidence of aforementioned disease has been increasing rapidly from rates similar to Russian Karelia to incidence similar to Fin-

land together with modernization of the lifestyle in recent decades (Teeaar et al., 2010; Voor et al., 2005).

All infants in DIABIMMUNE study were followed from birth till age of three by monthly stool sampling, together with collection of periodic blood samples at age of 3, 6, 12, 18, 24 and 36 months. Families filled in extensive questionnaires covering topics such as breastfeeding, diet, allergies, living conditions and use of drugs. Additionally, information about the mode of birth (vaginal birth vs. caesarean section) was recorded. This cohort provides, up to date, largest longitudinal collection of infant stool samples, enabling functional profiling of developing infant gut microbiome at unprecedented detail.

In Publication I, the underlying research hypothesis was that there are differences in early gut microbiome contributing to the differences in autoimmunity between the countries. We set out to compare the infant gut microbiome in these three populations—Finns, Estonians and Russian Karelians—which represent a microcosm of the global gradient in autoimmune disease incidence between western and developing nations to gain insights in mechanisms behind the hygiene hypothesis.

Studies in mice have demonstrated that early colonization with protective microbiota can convey decreased risk of autoimmune diabetes (Markle et al., 2013) and allergies (Stefka et al., 2014). In Publication II, we examined the relationship between T1D and microbiome in case-control setting by following gut microbiome of 11 infants with T1D associated autoantibodies (cases) and 22 infants with no observed autoantibodies (controls) until three years of age. Four infants developed T1D within the timeframe of the study.

Antibiotic treatments have both short- and long-term effect in gut microbial communities (Dethlefsen and Relman, 2011; Jakobsson et al., 2010). In Publication III, we leveraged differences in early antibiotic prescription rates in DIABIMMUNE study by composing a cohort of infants with no antibiotic treatments or more than nine antibiotic treatments during the first three years of life. This cohort of 39 infants enabled us to examine effects of repeated antibiotic treatments in early life in comparison to infants with no such perturbations but otherwise similar environmental exposures.

In addition to the cohort specific questions described above, all three cohorts in Publications I–III were used to study various common questions on the developing infant gut microbiome, such as its diversity and

	# of infants	# of samples with 16S sequencing	# of samples with WMS sequencing
Publication I	222	1584	785
Publication II	33	777	124
Publication III	39	1069	240
Total	294	3430	1149

Table 3.1. Number of infants and stools samples in DIABIMMUNE study

stability. Results regarding these generic questions are combined and presented in coherent sections in Chapter 4.

The development of the gut microbial communities of infants in DIABIMMUNE study was characterized using taxonomic and functional profiling approaches, which are described in more detail below (experimental part) and in Chapter 2 (computational part). We analyzed a total of 3430 stool samples from 294 infants using 16S sequencing. Since WMS sequencing is more expensive compared to 16S sequencing, in DIABIMMUNE cohorts it was applied to a subset of samples analyzed using 16S sequencing. In Publication II and Publication III, samples for WMS sequencing were selected “manually”, by targeting samples of specific interest. In Publication I, we applied unsupervised selection tool called microPITA (Tickle et al., 2013) to select 785 samples, which represent the variation in 16S sequencing data as well as possible. The total of 1149 stool samples were analyzed using WMS sequencing. The sample counts are broken down by Publications I-III in Table 3.1.

3.2 LifeLines DEEP study

LifeLines is a prospective population cohort of 165000 Dutch adults who will be followed for 30 years. LifeLines DEEP is a sub-cohort of LifeLines where detailed ‘omics profiling data has been collected for approximately 1500 individuals (Tigchelaar et al., 2015). Goals of the LifeLines DEEP study include elaborating the concept of “normal” adult gut microbiome as well as investigating associations between the gut microbiome and other intrinsic and extrinsic host factors. For this purpose, stool samples were collected from 1179 LifeLines DEEP participants.

In Publications IV, we reported WMS sequencing analysis results of

1135 stool samples together with 207 exogenous and intrinsic host factors. These factors included 39 self-reported diseases, 44 drug categories, smoking status, 78 dietary factors and 41 intrinsic factors of various physiological and biomedical measures. These analyses provide important steps towards better understanding of the healthy human gut microbiome.

3.3 Fecal microbial transplantation in Crohn's disease

Fecal microbial transplantation (FMT) is an efficient technique for treating recurrent *C. difficile* infections (Kelly et al., 2012; Rohlke and Stollman, 2012). It has been proposed that it could be used for treating other conditions where gut dysbiosis has been implicated (Kelly et al., 2015). In Publication V, our goal was to study the efficacy of FMT in treating Crohn's disease (CD)—a subtype of IBD—in a prospective setting. We conducted an uncontrolled, prospective open-label study of FMT from healthy donor to patients with active CD with 19 subjects and three donors. A single FMT per patient was performed via colonoscopy.

We followed the microbiome of recipients by collecting three stool samples per subject: one pre-treatment sample preceding the FMT and two post-treatment samples four and eight weeks after the FMT. We also measured several clinical parameters including Harvey Bradshaw Index (HBI), short Inflammatory Bowel Disease Questionnaire (sIBDQ), Crohn's Disease Endoscopic Index of Severity (CDEIS), and C-reactive protein (CRP) levels in blood during 12 weeks period following FMT. Stool samples were subject to WMS sequencing and the data was analyzed for microbial signatures explaining treatment response. This is a first study reporting the impact of colonoscopic FMT on microbial and immunological parameters in patients with active CD.

4. Results

In this chapter, I describe the main findings in the publications in this thesis. First, I describe our results in characterizing the developing infant gut microbiome in the DIABIMMUNE study (Publications I–III). Next, I highlight some findings from the LifeLines DEEP study, where we characterized gut microbiomes of 1135 Dutch adults in Publication IV. The chapter is concluded with results from the FMT study in Publication V.

4.1 Gut microbiome in early childhood

The gut microbial communities in infancy are dynamic as they develop toward adult composition. In DIABIMMUNE, we have studied systematic differences between populations (Publication I), local differences between diseased and healthy infants (Publication II), and perturbations such as antibiotic treatments (Publication III) and their connections to the development of the immune system. These results described below, give many new insights in the developing infant gut microbiome.

4.1.1 Microbial diversity is established during the first three years of life

The infant gut microbial community is dynamic during the first three years of life. Accordingly, age of the subjects was a major source of variation in all three cohorts (Publications I–III). This could be seen in PCoA plots (see Figure 4.1 for PCoA plots in Publication I) and was also reflected by α -diversity, measured by Shannon diversity index (Equation 2.2); In all three cohorts, we observed consistent increase in α -diversity with respect to age. This trend was later controlled using linear or non-linear modeling when conducting other statistical comparison, such as case-control comparisons or comparisons between countries.

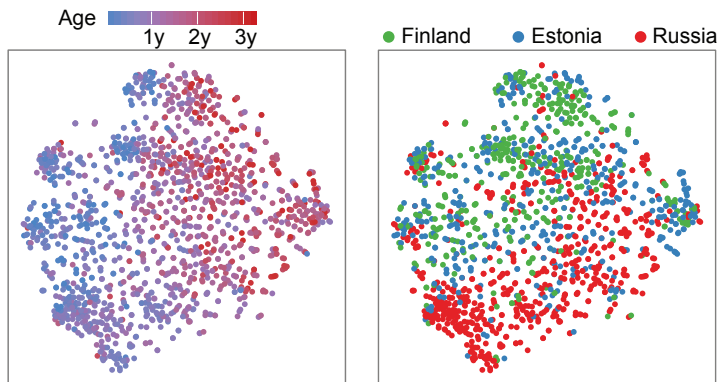


Figure 4.1. PCoA ordination of DIABIMMUNE 16S samples in Publication I, colored by age (left) and country (right). Each point represents an individual stool sample.

Healthy and stable microbial communities in the gut are typically characterized by high diversity (Lloyd-Price et al., 2016), whereas decreased diversity is usually implicated in disease or microbial dysbiosis (Manichanh et al., 2012). In Publication II, we found that decreased diversity is also implicated in T1D pathogenesis. Infants who progressed to T1D in our case-control cohort were described by decreased microbial diversity before clinical diagnosis but after autoantibody seropositivity. This finding suggests possible microbial elements in the onset of the disease but is not sufficient evidence to prove a causal relationship between microbiome and T1D.

In Publication I, where Russians were compared to Finnish and Estonian infants due to their lower susceptibility to autoimmunity, we surprisingly found that Russians had lower microbial diversity during the first year of life. This result could be explained by differences in human milk oligosaccharide (HMO) utilizing bacteria between the countries; in Russian gut communities, a few *Bifidobacterium* species were responsible for HMO utilization, whereas in Finnish and Estonian guts many *Bacteroides* species were responsible for HMO utilization in addition to *Bifidobacterium*. Regardless of lower diversity in early life, Russians developed gut communities with on average higher microbial diversity after the first year of life but before the end of the follow-up period. This suggests that higher microbial diversity of gut communities in early infancy may not necessarily be beneficial in similar manner that is reported in the literature covering adult gut microbiome. This observation is further connected to innate immunity in Section 4.1.6.

Antibiotic treatments have been shown to cause perturbations in both

the human (Dethlefsen and Relman, 2011; Korpela et al., 2016) and murine (Nobel et al., 2015) gut microbiome. In Publication III, we reported decreased microbial diversity (Equation 2.2) in infants with repeated antibiotic treatments. Birth by caesarean section results in a distinct microbial profile characterized by absence of *Bacteroides* species and decreased microbial diversity (Dominguez-Bello et al., 2010a; Jakobsson et al., 2014). Analyzing data of 39 infants in Publication III, we found that infants born by caesarean section together with infants with no early *Bacteroides* species (phenomenon which is further described in Section 4.1.3) had decreased microbial diversity.

We profiled the WMS data in Publication I and Publication III on strain level using ConStrains (see Section 2.2.1). Briefly, ConStrains infers microbial strains by conducting haplotyping of SNPs on species-specific marker genes and enables following subject-specific strains in time. We further quantified within-species diversity using haplotype diversity (Equation 2.3), a measure of within-species strain diversity. In Publication I, we observed an increasing trend with respect to age similar to diversity on community level (α -diversity). We also found that in 60 % of the strain profiles there was one dominant strain per microbial species accounting for more than 90 % of within-species abundance. However, some genera, such as *Veillonella* and *Faecalibacterium*, had bimodal haplotype diversity distributions indicating more complex strain behavior. In Publication III, we observed significantly lower haplotype diversity scores in infants with repeated antibiotic treatments. Consequently, these infants were more often colonized with species with a single dominant strain.

4.1.2 Stability of the microbiome is decreased in infancy

Faith et al. (2013) found that microbial stability follows power-law function in adults; within-subject similarity of microbial communities is decreased rapidly in time but there is certain number of taxa, which are persistent over long periods of time. In Publication II, we found that this holds also for the infant gut microbiome. Our model suggested that roughly 10 % of bacterial strains obtained soon after birth were maintained until three years of age. This means that despite the dynamic nature of the microbial community in infancy there is a “core” microbiota that is persistent over this period. In adulthood, this “core” part accounts for more than 50 % of the microbial species inhabiting the gut (Faith et al., 2013).

In addition to measuring stability on community level, we investigated stability within taxonomic groups in Publication I. First, we applied power-law model on genus level for genera represented with more than 10 OTUs in our data. *Bifidobacterium* were more stable in Russian infants whereas all other genera compared were more stable in Finns and Estonians. This can be also viewed such that Russians have more plastic microbiota during this early dynamic period while beneficial *Bifidobacterium* are able to colonize in a stable manner regardless.

Strain level variation may play important roles in determining functional potentials and even pathogenicity of microbes (Scholz et al., 2016). We investigated within-species stability using strain profiles generated by ConStrains and observed two kinds of behaviors: microbial strains tended to remain stable with a single dominant strain per species, and occasionally experienced a “strain sweep” where an old dominant strain was replaced by a new one. These events may be implicated in weaning, antibiotics exposures and other life events.

In Publication III we found that exposure to antibiotics provided a perturbation in the gut microbiome, which was seen as a short-term decrease in stability. Infants with multiple antibiotic treatments during the first three years of life had overall less stable microbial communities when compared to infants with no antibiotics treatments. This observation underlines the importance of ongoing efforts to reduce over-prescription of antibiotics.

4.1.3 The microbiome is modulated by diet

Dietary intake influences the composition and activity of the gut microbiota in adulthood in both human (Conlon and Bird, 2015; David et al., 2014) and mice (Sonnenburg et al., 2016). In early childhood, diet and especially breastfeeding play an important role in the development of gut microbial communities towards an adult composition (Backhed et al., 2015). In DIABIMMUNE study, we have collected times of first introduction to various dietary elements to study and correct for changes in microbiota associated with accrual of dietary diversity.

Breastmilk is the sole nutrition source available to most infants during the first months of life. Breastmilk contains human milk oligosaccharides (HMOs) (Sela and Mills, 2010) and assortment of bacteria (Hunt et al., 2011) which provide important pre- and probiotic functions, respectively. In Publication I and Publication II, we found that breastmilk was highly

favorable to fecal *Bifidobacterium* and *Lactobacillus* species, an effect well known in the literature (Backhed et al., 2015). Genera *Blautia* and *Oscillospira*, and family Lachnospiraceae were among the most decreased taxonomic groups until weaning. However, in Publication III we noted that some Finnish infants had low abundance of *Bifidobacterium* species even during the breastfeeding period.

In addition to *Bifidobacterium*, *Bacteroides* species are also capable of utilizing HMOs (Marcobal et al., 2011). In Publication I, we found evidence for distinct microbiota-dependent HMO metabolism in infants from Finland and Estonia in comparison to Russian Karelian infants. By analyzing the abundance and origin of genes in the *bona fide* HMO gene cluster (Sela et al., 2008) we showed that while most of these genes were carried by *Bifidobacterium* in Russians, *Bacteroides* species were responsible for majority of HMO metabolism in Finnish and Estonian infants. The relative decrease of *Bifidobacterium* in Finns and Estonians was not explained by differences in breastfeeding; Russian infants were breastfed for a shorter period on average (mean \pm SD breastfeeding days: Finland 268 ± 149 , Estonia 307 ± 217 , Russia 199 ± 165). However, Finnish infants had higher abundance of *Bifidobacterium breve*—species incapable of metabolizing intact HMOs (Locascio et al., 2009)—compared to Russians.

In Publication III, we observed that both *Bifidobacterium* and *Bacteroides* dominated HMO metabolism can be found in Finnish infants as well. It is known that infants born by caesarean section lack *Bacteroides* species in their early gut microbiota (Jakobsson et al., 2014; Backhed et al., 2015). However, we found such signature in 7 of 35 (20%) infants born vaginally. In infants with no early *Bacteroides* species (all four infants born by caesarean section and seven vaginally born infants) *Bifidobacterium* were the main contributors to HMO metabolism. Despite our extensive search for clinical metadata which would explain this lack of *Bacteroides* species in vaginally born infants, we could not find any explanation for this newly described phenomenon.

Even though the influence of diet to the microbiome is recognized, there is only scattered information on the implications of different dietary elements on the microbiome. In Publication I, we utilized our dietary data to look for consistent changes in the microbiome co-occurring with first introduction to different dietary elements. Among other things, we observed an increase in bacterial family Lachnospiraceae to be significantly increased after introduction of vegetables and oat, and genus *Lachnospira*

to be increased after introduction of soy. De Filippis et al. (2015) have previously reported increased levels of *Lachnospira* in subjects following vegetarian diet. We also saw a consistent increase in *Ruminococcus* and Ruminococcaceae after introduction of egg in Publication I and Publication II, respectively. These bacteria can likely utilize proteins in eggs. More systematic study designs are required to assess dietary implications in early gut microbiome in more detail and with better confidence.

4.1.4 Antibiotic treatments perturb microbial taxa and genes

In addition to looking at global shifts in gut microbial communities caused by antibiotic treatments (see Section 4.1.1 and Section 4.1.2), we examined implications of antibiotics on taxonomic and functional levels. In Publication I, we observed that while genera *Clostridium* and *Haemophilus* and class Gammaproteobacteria tended to be most significantly decreased in association with antibiotics, class Deltaproteobacteria and genus *Bilophila* had an opposite trend being increased together with antibiotics. Atarashi et al. (2013) identified strains in so called *Clostridium* clusters IV and XIVa to have health-promoting effects through induction of T regulatory cells. In Publication III, we observed that children in antibiotic treatment group had lower levels of these bacterial species—a difference driven by difference in abundance of *Eubacterium rectale*—compared to children with no antibiotic treatments.

Bacterial resistance to antibiotic treatments is conveyed by antibiotic resistance (AR) genes which can be harbored chromosomally and transferred horizontally in mobile elements (Waters and Salyers, 2013). In Publication III, we leveraged the WMS sequencing data to quantify genes that are known to confer resistance to specific types on antibiotics (McArthur et al., 2013). We found that chromosomal AR genes peaked after antibiotic treatments and in many cases we were able to identify the bacterial species likely carrying the AR gene by corresponding peaks in their longitudinal relative abundance profiles. In contrast to this peaking behavior, we observed different patterns for some episomally encoded AR genes, that is, genes that are encoded on plasmids or other mobile elements. More specifically, their presence continued for much longer time after the antibiotic treatment, which may be explained by the fact that they can be distributed across wide variety of bacteria, whereas chromosomal AR genes are constrained to their harboring species. Finally, we detected AR genes in some children (11 of 39) prior to any antibiotic treatments,

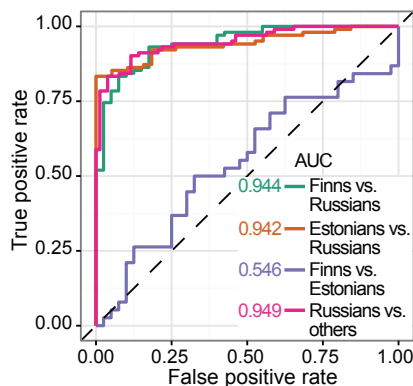


Figure 4.2. Receiver operating characteristic (ROC) curves for pairwise Random Forest classifiers predicting country based on 16S genus data using samples collected between 170 and 260 days of age in Publication I.

a phenomenon described earlier in the microbiome of infants (Backhed et al., 2015) and rural Amerindians with no access to modern antibiotics (Clemente et al., 2015).

4.1.5 Population-level differences in microbiota composition

There is a regional, and even a familial component in human gut microbiome implicating that people living in same area or household have more similar microbiome with each other compared to their peers living outside of their community (Lax et al., 2014; Yatsunenko et al., 2012). In Publication I, we found similar trend already in early infancy; Russian infants had distinct gut microbiota compared to their Finnish and Estonian peers with striking consistency. This difference could be seen on PCoA ordination (Figure 4.1) and was most clear in the earliest sampling points. We trained a Random Forest classifier to separate samples by country in early time window and obtained area under curve (AUC) equal to 0.949 (Figure 4.2) when comparing Russians to Finns and Estonians, implicating close-to-perfect separability between the groups.

As briefly described above, differences in *Bacteroides* and *Bifidobacterium* species were among the most significant taxonomic differences between the countries. *Bifidobacterium* species are usually increased during breastfeeding due to their specialization in HMO metabolism, whereas in addition to breastmilk HMOs, *Bacteroides* species can utilize a wide variety of dietary plant polysaccharides (Sela and Mills, 2010; Xu et al., 2003).

Previous studies suggest that early colonization of *Bacteroides* play a role in different immune mediated diseases, such as T1D (Davis-Richardson

et al., 2014) and celiac disease (Sanchez et al., 2011). In Publication I we found that *Bacteroides dorei*, a species previously connected T1D pathogenesis (Davis-Richardson et al., 2014), was the *Bacteroides* species with the biggest fold change between Russian and Finnish infants. Fecal *B. dorei* relative abundance correlated with serum insulin autoantibody levels both within Finland and cohort wide. This result was consistent with the literature and provided further evidence for the role of *B. dorei* in autoimmunity and T1D pathogenesis.

To better understand potential implications of the taxonomic differences between the countries, we also compared the countries by looking at the functional content of the microbiome. We found roughly 100 biological process GO terms with differences in abundance between Finland and Russia. Glycolytic functions, which are related to HMO metabolism, were more active in Russian infants. This is likely caused by higher abundance of *Bifidobacterium* species—highly specialized HMO metabolizing bacteria—in Russian intestines. We saw many GO terms, such as virulence and siderophore-related functionalities, increased in Finnish infants compared to Russians. While most of these findings have not been followed up in this thesis, the list may include many important, yet unstudied microbial pathways, which have implications in infant health and wellbeing, deserving studies of their own.

Most importantly, we observed that two lipopolysaccharide (LPS) related GO terms—LPS biosynthetic process and Lipid A biosynthetic process—showed notably increased abundance in Finnish infants compared to Russians. The difference in LPS biosynthesis was most substantial during early months of life and dissipated in time (Figure 4.3). LPS is a surface protein of gram-negative bacteria and it is known to elicit strong immune response in mammalian cells (Cullen et al., 2015). The known immunostimulatory properties of LPS and early differences in LPS production between Finland and Russia made us hypothesize that this signal might have implications on the development of the immune system.

Lipid A is a subunit of LPS, responsible for its immunostimulatory properties. Therefore, the lipid A biosynthesis pathway is more specific pathway and is encoded by a smaller set of genes compared to the LPS biosynthesis pathway. Using lipid A biosynthesis as a proxy to LPS biosynthesis, allowed us to avoid contributions from many genes that are important to multiple microbial pathways, and thus resulted in a less noisy signal. For example, when we deconvoluted LPS biosynthesis signal by contributing

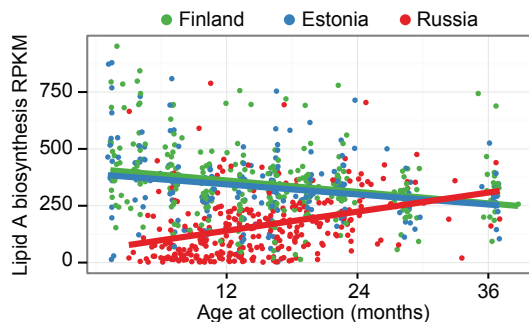


Figure 4.3. Lipid A biosynthesis pathway normalized read counts (RPKM) per sample and a linear fit per country in Publication I.

species, we saw contributions from many highly abundant gram-positive bacteria, which are known not to produce LPS. This was not the case when the deconvolution was conducted to the lipid A biosynthesis signal.

By deconvoluting the lipid A biosynthesis signal, we were able to investigate computationally what species were contributing to LPS biosynthesis in the infants under study. We found that *Escherichia coli* was a major LPS producer in all countries. More importantly, many *Bacteroides* species contributed to lipid A and LPS biosynthesis in Finland and Estonia from early on. These bacteria were partly the same species that accounted for the differences in HMO metabolism between the countries, and the contrast in their abundance was most substantial during the first year of life, the time frame most important for the developing immune system. This prompted us to investigate properties of LPS and lipid A produced by these species in more detail in experiments which are described next.

4.1.6 Variation in LPS structure impacts immune development

Differences in LPS structure are implicated in its immunogenicity, that is, ability to invoke innate immune responses (Whitfield and Trent, 2014). Immune cells recognize LPS through the Toll-like receptor 4 (TLR4) complex, which is activated by binding of lipid A subunit described above (Kim et al., 2007). The number of acyl chains in lipid A is an important factor in determining the magnitude of the immune response (Hajjar et al., 2002; Needham et al., 2013). Lipid A from *E. coli* has six acyl chains (hexa-acylated) and provokes robust immune response (Needham et al., 2013) whereas lipid A structures with four or five acyl chains (tetra- and penta-acylated) have been previously shown to elicit reduced TLR4 activation

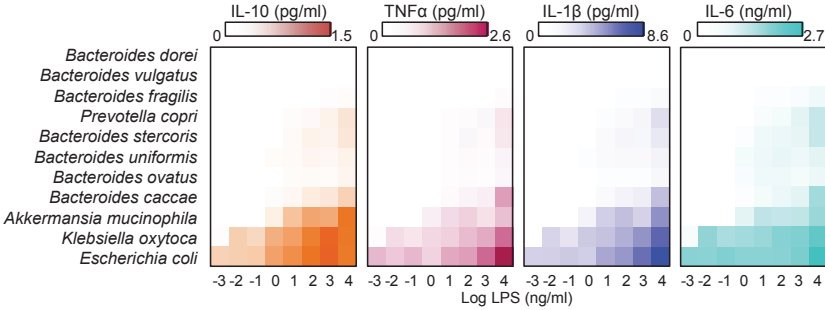


Figure 4.4. Mean cytokine production in PBMCs stimulated with the indicated doses of LPS purified from different bacterial species as assessed by cytokine bead array in Publication I.

(Herath et al., 2011). Taken together, structural changes in lipid A are likely to influence multiple facets of microbiota-host interactions, which prompted us to investigate LPS from the microbes in the guts of the infants in our cohort experimentally.

To study structure and function of LPS produced by early colonizing bacteria in infant in the cohort, we set out to purify LPS from bacteria that contributed to the LPS load in our samples. We were able to purify LPS from 11 type strains listed in Table 4.1. We first used these LPS to stimulate primary human peripheral blood mononuclear cells (PBMCs) and measured different necrosis factor κ B (NF- κ B)-dependent cytokines—interleukin-10 (IL-10), tumor necrosis factor α (TNF α), IL-1 β and IL6—to assess inflammatory response evoked by LPS. As expected, *E. coli* LPS produced a robust response even on small doses (Figure 4.4). Strikingly, LPS derived from *B. dorei* failed to elicit any response regardless of the dose. Response evoked by LPS from all other analyzed *Bacteroides* species and *Prevotella copri* was also greatly impaired compared to response shown by *E. coli* LPS. We confirmed these findings in human monocyte-derived dendritic cells and in TLR4-NF- α B reporter cells which both produced concordant results. Since *E. coli* and *B. dorei* LPS were responsible for the highest and lowest cytokine responses, respectively, we next used these two LPS subtypes to study certain properties of these molecules in more detail.

To probe the structural basis for the contrasting response between *E. coli* and *B. dorei* LPS, we used matrix-assisted laser desorption/ionization-time of flight mass spectrometry (MS) to examine the structure of the lipid A domain of these LPS subtypes. *E. coli* derived lipid A produced a pre-dominant peak at a mass-to-charge ratio (m/z) of 1798.3, consistent with

Name	Description	Reference
<i>Bacteroides dorei</i>	strain: 175	DSM 17855
<i>Bacteroides ovatus</i>	strain: NCTC11153	ATCC 8483
<i>Bacteroides vulgatus</i>	strain: NCTC11154	ATCC 8482
<i>Bacteroides stercoris</i>	strain: VPIB5-21	ATCC 43183
<i>Bacteroides fragilis</i>	strain: VPI2553	ATCC 25285
<i>Bacteroides uniformis</i>	strain: VPI0061	ATCC 8492
<i>Bacteroides caccae</i>	strain: VPI3452A	ATCC 43185
<i>Prevotella copri</i>	strain: 18205	DSM 18205
<i>Akkermansia mucinophila</i>	strain: CIP107961	ATCC BAA-835
<i>Klebsiella oxytoca</i>	strain: OCC-SAL-18A	ATCC 68831
<i>Escherichia coli</i>	strain: ECOR2	Ochman & Selander, J. Bacteriol (1984)

Table 4.1. Bacterial strains used in LPS Purification and following experiments with immune cells in Publication I.

predicted exact mass 1797.2 m/z of the published $[M-H]^-$ ion structure of *E. coli* lipid A containing two phosphate groups and six acyl chains (Needham et al., 2013). Lipid A extracted from *B. dorei* LPS produced two predominant peaks at m/z 1,690.9 and 1,436.2, corresponding to the $[M-H]^-$ ion structures with one phosphate group and four and five acyl chains, respectively (predicted exact mass: 1689.2 and 1435.0 m/z). These observations provided structural basis for the differences in immunogenicity of these LPS molecules we observed in our immune cell experiments.

Next, we set out to study possible interactions occurring when LPS from different bacteria co-occur in the infant gut. Using human primary immune cells and a base dose of *E. coli* LPS, we co-treated these cells with increasing dose of *B. dorei* LPS. We measured production of inflammatory cytokines with respect to the baseline *E. coli* stimulation and found that *B. dorei* LPS inhibited cytokine production elicited by *E. coli* LPS in primary human PBMCs and in monocyte-derived dendritic cells. We noticed that maximal inhibition was reached when *B. dorei* and *E. coli* LPS were used at a ratio of 10:1 which corresponded to the computational prediction of the ratio of these LPS subtypes typical for IAA seropositive infants in our cohort in Publication I. This result confirmed that immunologically silent *B. dorei* LPS has implications in immune activation, even when acting together with potent *E. coli* LPS, by inhibiting immune activation by

E. coli LPS.

Stimulating immune cells with LPS induces a temporary refractory state where the cells show decreased activation upon repeated immune challenge. This phenomenon is known as endotoxin tolerance (Watson and Kim, 1963). We hypothesized that the distinct properties of *B. dorei* LPS may have implications in induction of endotoxin tolerance and tested this using primary human monocytes. As expected, initial exposure to *E. coli* LPS inhibited TNF α production in restimulation at all conditioning doses used. When we used *B. dorei* LPS in the initial exposure, we observed substantially reduced potency to induce endotoxin tolerance. Consistently with our previous experiments, mixing *B. dorei* LPS with *E. coli* LPS during the endotoxin tolerance induction phase prevented the establishment of endotoxin tolerance by *E. coli* LPS in a dose-dependent manner. Even though the mechanisms of endotoxin tolerance are still poorly understood, these experiments connect our findings to immune development, as endotoxin tolerance is thought to underlie the immune protective effect conferred by microbial exposure suggested by the hygiene hypothesis (Biswas and Lopez-Collazo, 2009).

Finally, we wanted to demonstrate the relevance of the differences in LPS structure and function to T1D pathogenesis *in vivo* in non-obese diabetic (NOD) mouse model of T1D. LPS can be used to protect NOD mice from T1D by intraperitoneal (i.p.) injection (Aumeunier et al., 2010) and oral gavage (Sai and Rivereau, 1996). We found that while *E. coli* LPS given through i.p. injection protected NOD mice from T1D, *B. dorei* LPS did not have this protective effect. This shows that immunostimulatory LPS can play a role in protection from immune-mediated diseases, whereas immunologically silent *B. dorei* LPS has less potency to provide such protective effects.

4.2 Adult gut microbiome

In comparison to the infant gut microbiome, the adult gut microbiome is studied more extensively in both healthy adults and in different diseases. In Publication IV, we studied the gut microbiome of 1135 Dutch adults, representing the general population, using WMS sequencing of stool samples. This study provides a step towards a better understanding of complex environment-diet-microbe-host interactions. In Publication V, we reported the first FMT study in active Crohn's disease (CD), where we

followed the recipients by stool sampling for eight weeks. In both studies, WMS data was analyzed using MetaPhlAn2 for taxonomic profiling and HUMAnN2 for functional profiling similar to above.

4.2.1 Factors associated with gut microbiome variation in Dutch population

In Publication IV, we conducted extensive analysis of microbiomes from 1135 Dutch adults together with extensive metadata, including 39 self-reported diseases, 44 drug categories, smoking status, 78 dietary factors and 41 intrinsic factors of various physiological and biomedical measures. Similar to other studies, we found high inter-individual variation in our data, which was clearly visible on phylum level (Figure 4.5, top panel). Regardless, 23 nonredundant molecular function GO terms representing high level microbial functions showed remarkably stable profiles across the cohort (Figure 4.5, bottom panel), similar to previous reports (Human Microbiome Project, 2012; Lozupone et al., 2012). Out of all collected 207 metadata factors, 126 were found to significantly explain variation in the microbiome data. Together these factors explained 18.7 % of the variation in the data. Metadata also explained 13.7 % of the variation in microbial α -diversity, measured by Shannon's diversity index (Equation 2.2).

We conducted extensive association analysis between collected metadata and microbial species and pathways. Exhaustive results of these analyses are given in Publication IV and accompanying online materials. Here, I highlight some findings while connecting them to relevant literature.

Diet is known to be a major factor driving the taxonomic and functional composition of gut microbial communities (Sonnenburg et al., 2016; Turnbaugh et al., 2009). In our analyses, we observed positive correlation between fruit intake and *Alistipes shahii*, a species which have been previously associated with lower blood triglyceride levels (Fu et al., 2015). This finding suggests a link between the fruit intake and the blood triglyceride levels. We also founds that buttermilk (sour milk with a low fat content) consumption correlated with microbial diversity, whereas regular high-fat milk had inverse correlation with diversity. This suggests pre- and/or probiotic effects in buttermilk. In contrast to previous reports (Wu et al., 2011), we did not see a connection between carbohydrate consumption and genus *Prevotella*. Instead, we observed increased levels of bacteria from genera *Lactobacillus*, *Streptococcus* and *Roseburia* in low-carbohydrate

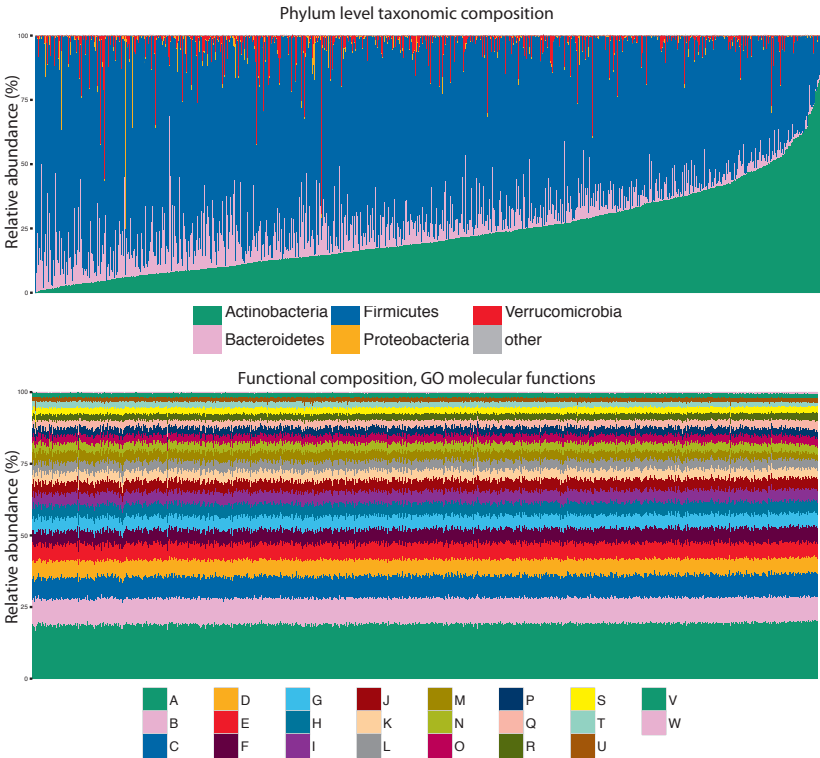


Figure 4.5. High-level taxonomic groups (top) and functional categories (bottom) in adult gut microbiome in Publication IV. Each column ($n = 1135$) represents a stool sample. Samples are sorted according to Actinobacteria relative abundance. Functional categories in the bottom panel are A: ATP binding; B: oxidoreductase activity; C: RNA binding; D: sequence-specific DNA binding transcription factor activity; E: active transmembrane transporter activity; F: ligase activity; G: lyase activity; H: oenzyme binding; I: isomerase activity; J: ATPase activity, coupled; K: methyltransferase activity; L: nucleotidyltransferase activity; M: sequence-specific DNA binding; N: zinc ion binding; O: nuclease activity; P: inorganic cation transmembrane transporter activity; Q: hydrolase activity, hydrolyzing O-glycosyl compounds; R: transferase activity, transferring glycosyl groups; S: transferase activity, transferring acyl groups other than amino-acyl groups; T: phosphorelay sensor kinase activity; U: protein binding; V: endopeptidase activity; W: protein serine/threonine kinase activity

diet.

Antibiotic use in the Netherlands is the lowest in Europe. Regardless, we saw significant shift in subjects who used antibiotics before sampling, with largest decrease in *Bifidobacterium adolescentis* and *Bifidobacterium longum* species. Other drugs, such as proton-pump inhibitors (PPIs), metformin, statins and laxatives were found to have an impact on the gut microbiota composition, as well.

Chromogranin A (CgA) is a member of the granine peptides and in LifeLines DEEP it was measured as a marker for neuro-endocrine system ac-

tivation. CgA is secreted by immune cells in many gastrointestinal tract disorders including IBD (Sciola et al., 2009) and inflammatory bowel syndrome (IBS) (Ohman et al., 2012). We found that fecal CgA levels explained 3 % of the variation in microbiota composition (adonis R² = 0.03, adjusted P = 0.0006). There was inverse correlation between CgA levels and microbial α -diversity, functional richness and high-density lipoprotein (HDL) concentration. High fecal CgA levels were associated with high fecal calprotectin levels, with high concentrations of triglycerides in blood, with high stool frequency, with soft stool type, and with self-reported IBS. After correcting for potential confounding effects by the factors above, we found 61 microbial species correlating with fecal CgA levels. Many species from phylum Bacteroidetes (24 out of all 36 species) showed inverse correlation with CgA.

4.2.2 FMT in active Crohn's disease

In Publication V, we conducted an uncontrolled, open-label FMT study for patients with active CD. Nineteen patients were treated with FMT from three donors and followed up over 26 weeks. Eleven patients showed a clinical response, measured by decrease greater than three in Harvey Bradshaw Index (HBI), a questionnaire-based index of Crohn's disease activity. We collected three stool samples per patient—one pre-treatment sample and two post-treatment samples four and eight weeks after the FMT treatment—and analyzed microbial composition of these samples using WMS sequencing. A stool sample from all three donors were analyzed in a similar manner.

We observed a shift in microbiota following the FMT, where responders showed significantly larger change towards the donor profiles, measured by Bray-Curtis dissimilarity (Equation 2.5). We also saw an increase in microbial α -diversity, which was larger in responders, following the FMT. Regardless of these community level differences between responders and non-responders, we did not observe any strong and consistent species level differences between these two groups. Overall, we found that questionnaire-based indices of disease severity—HBI and sIBDQ—improved after FMT, but there was no significant change in CDEIS or CRP scores 12 weeks after the FMT. Based on our study, FMT was safe and well tolerated in the setting of active Crohn's disease although the clinical response was variable.

5. Discussion

A growing body of evidence suggests that the gut microbiome is an important factor in many human conditions, such as autoimmunity and allergies. In this thesis, we used modern molecular methods to characterize the gut microbiome in two cohorts, DIABIMMUNE (Publications I–III) and LifeLines DEEP (Publication IV), and in a prospective FMT study in active CD (Publication V). We identified the lack of immunogenicity in *Bacteroides*-derived LPS as a novel molecular mechanism which may contribute to the prevalence of autoimmunity and T1D in Finland and Estonia. We also elucidated microbial changes preceding T1D diagnosis, and in connection to recurrent antibiotic treatments in infancy. With regards to the adult microbiome, we found signals in the gut microbiome of healthy Dutch adults corresponding to various intrinsic and extrinsic factors, most significantly chromogranin A. We also showed that FMT from healthy donors in active CD is safe and well-tolerated. Taken together this thesis expands the current understanding of the human gut microbiome in infancy and adulthood, and contributes novel mechanistic understanding in microbiome-host interactions in autoimmunity.

In Publication I, we found differences in the abundance of HMO degrading *Bacteroides* and *Bifidobacterium* species in the infant gut microbiome between Finnish and Estonian versus Russian Karelian infants (Figure 5.1). We also identified LPS biosynthesis as one of the largest functional differences between the populations, suggesting that the microbiome of Finnish and Estonian infants produced more LPS. By following up this pathway in mechanistic and structural experiments, we found out that LPS produced by different bacteria in the infant microbiota could either stimulate or actively inhibit TLR4, NF- κ B activation, and endotoxin tolerance. This difference is most likely driven by the difference of the number acyl chains in the lipid A component of LPS. Importantly, we showed that

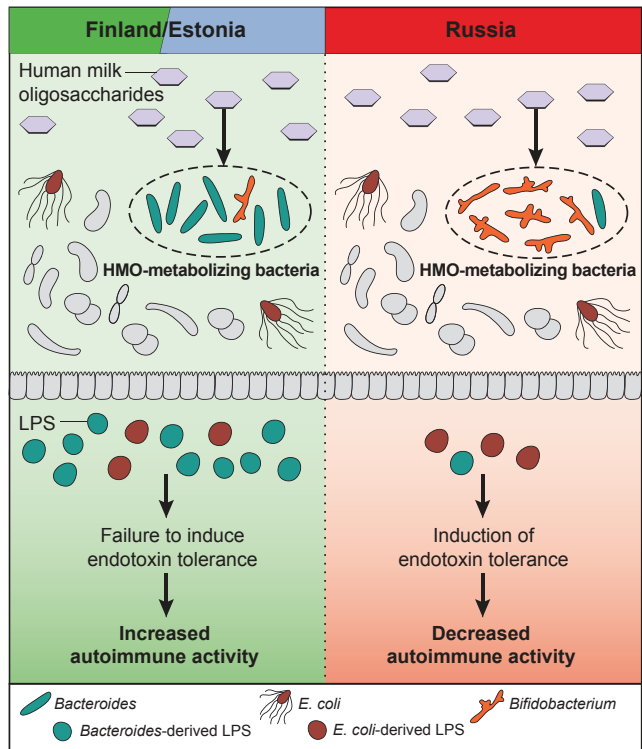


Figure 5.1. Graphical summary of Publication I: Human milk oligosaccharides can be metabolized by different prevalent microbes in Russia (primarily *Bifidobacterium* species) versus Finland and Estonia (primarily *Bacteroides* species). Potentially as a result of these population differences, *Bacteroides*-derived lipopolysaccharide (LPS) constitutes the major portion of LPS produced by microbes in Finnish and Estonian infants, whereas LPS in Russian infants is mostly derived from *E. coli*. *Bacteroides*-derived LPS is of an immunoinhibitory subtype, thus leading to differential immune education by means of endotoxin tolerance or other routes.

injections of non-immunogenic subtype of LPS from *B. dorei* did not protect NOD mice from T1D, whereas immunogenic LPS from *E. coli* both decreased the incidence of T1D and was able to elicit endotoxin tolerance *in vivo* in NOD mice. These results suggest that the immune activation triggered by the gut microbiome derived LPS is determined by the nature and composition of different subtypes of LPS rather than the amount of LPS alone.

In Publication II, we found a marked drop in α -diversity in infants who progressed to T1D in comparison to infants who seroconverted but did not progress to T1D and controls with no disease or T1D specific AABs. This shift occurred prior to the onset of the disease but after seroconversion and was accompanied by spikes in inflammation-favoring organisms, such as *Ruminococcus gnavus* and *Streptococcus infantarius*. In Publi-

cation III, we compared infants undergoing recurrent antibiotics treatments (9 or more treatments per infant) with infants with no antibiotic treatments during the first three years of life and saw a reduction in microbial diversity and increase in AR genes in connection to antibiotic treatments. In this study, we also noticed that 20 % of the children born vaginally lacked *Bacteroides* species in the first 6 to 18 months of life. This low-*Bacteroides* signature is known to be typical for infant born by caesarean section (Dominguez-Bello et al., 2010b; Jakobsson et al., 2014) and there are also reports arguing that this is a typical profile in infancy (Arrieta et al., 2014). Despite our extensive search for a clinical variable that would explain this behavior, we were not able to find any explaining variable(s) for the low-*Bacteroides* group.

Understanding the microbial mechanisms that influence the development of autoimmunity and allergic sensitization is necessary for the success of efforts to manipulate bacterial communities for prophylactic and therapeutic purposes. Though our studies concerning the DIABIMMUNE cohort did not address therapy, our results entice speculation that early colonization and engraftment by certain *Bifidobacterium* species may lower the load of *Bacteroides* and other tetra-acylated-LPS-producing species. This hypothesis is supported by a recent study (Uusitalo et al., 2016) suggesting that early probiotic administration, consisting of mostly *Bifidobacterium* and *Lactobacillus* species, may reduce the risk of islet autoimmunity in children with genetic risk to T1D. The exact environmental factors that would initially favor a colonization by the beneficial *Bifidobacterium* species in Russia remain to be assessed, but may reflect biodiversity differences at the macroscopic and microbial levels in these regions (von Hertzen et al., 2015). Indeed, one hypothesis suggests that antibiotic administration leads to loss of commensal microbes, especially taxa with low relative abundance (Blaser, 2016). This effect is cumulative across generations due to the fact that we inherit our microbiome largely from our mother. Indeed, in the DIABIMMUNE cohort, Finnish families reported roughly seven times more antibiotic courses per children during the first three years of life compared to Russian families (average 3.42 versus 0.46 courses per child during the first three years in Finland versus Russia).

Interestingly, in Publication II we identified *Dialister invisus* as being absent in infants who progressed to T1D, suggesting that it might confer disease protection. In the cohort of Publication I, *D. invisus* was highly

abundant in seven Russian subjects after the first year, but the corresponding difference between the countries was not statistically significant. Notably, *D. invisus* is a member of the ill-defined Negativicutes class, whose cell wall composition is atypical and poorly characterized. Indeed, we were unsuccessful in purifying LPS from our *D. invisus* strain and could not determine whether this strain produces LPS at all. Therefore, additional investigations are necessary to determine the exact mechanism of action underlying potential disease protection of *D. invisus*.

Altogether, Publications I–III expanded the current understanding of the developing infant gut microbiome in many ways. They extended the view that the adult gut microbiome has high inter-individual and inter-regional variation. This was supported by the marked differences between Russians and others in Publication I, and differences in early *Bacteroides* abundance in Publication III. The microbiome data generated in DIABIMMUNE are publicly available through DIABIMMUNE microbiome website¹. These data—3430 16S sequencing and 1149 WMS sequencing samples from 294 infants—will hopefully enable testing many other existing hypothesis on the infant gut microbiome and will hopefully facilitate many more interesting discoveries.

Publication I–III earned a lot of media attention² and this has spurred public discussion about the hygiene hypothesis and its implications in parenting. While this thesis does not fully explain the hygiene hypothesis or microbial components of T1D pathogenesis, it has provided an opportunity to remind parents, for example, about the side effects of unnecessary antibiotic courses. Many reporters have also educated the public about the benefits of a healthy and diverse gut microbiome; the idea that not all microbes are bad. This is an important counter force to the increasing use of antimicrobial products which may be harmful to our commensal microbiome and may also facilitate resistant strains.

Our analysis in Publication I followed a generalizable discovery and validation process (Figure 5.2). We computationally quantified metabolic pathways in the microbial communities and identified pathways with differential abundance between the phenotypically distinct populations in our study. We followed by assigning these functions to specific microbes and ultimately identified structural differences within the products of these pathways (e.g. LPS) that induced distinct immune responses *in*

¹<https://pubs.broadinstitute.org/diabimmune/>

²See <https://cell.altmetric.com/details/7059017>

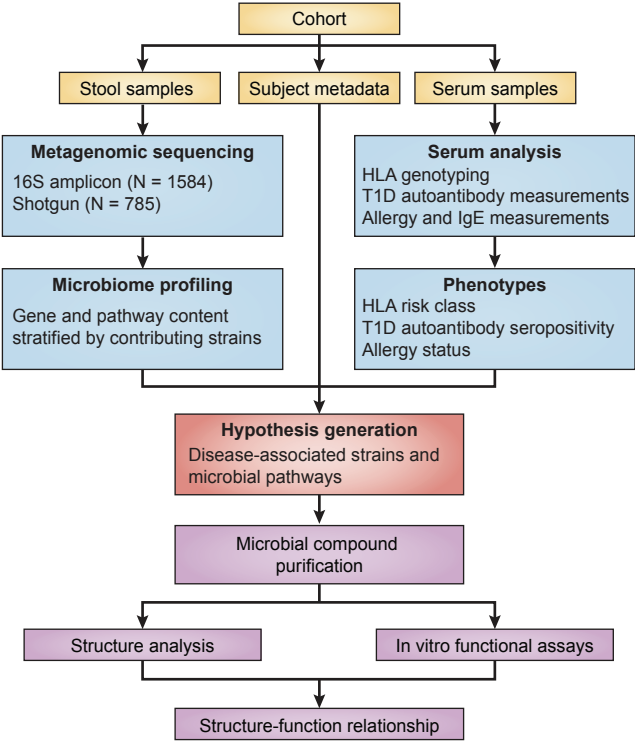


Figure 5.2. Generalizeable analysis workflow highlighting important steps in metagenomic data analysis and mechanistic experiments in Publication I.

vitro. Similar process can be used to identify and characterize microbial products with potential interactions with the host immune system in any study setting. In the DIABIMMUNE study population, there were many other functional differences between the countries, such as glycolysis and iron uptake, which may deserve further mechanistic investigations of their own.

WMS sequencing study in LifeLines DEEP revealed a lengthy list of associations between the gut microbiome and various intrinsic, environmental, dietary and medication parameters, and disease phenotypes. Even though 90 % of these associations could be replicated using 16S sequencing data generated in Fu et al. (2015), most of these associations need to be validated and followed up in new studies to understand their significance and possible causality in more detail. Most importantly, this study suggested fecal CgA as a novel biomarker for gut health. Further studies will show whether CgA can be used as a proxy for gut inflammation or microbial dysbiosis in a similar manner as, for example, fecal calprotectin (Konikoff and Denson, 2006).

In Publication V, we reported the impact of colonoscopic FMT on mi-

crobial and immunological parameters in patients with active CD. This study suggested that FMT for CD is safe and may provide symptomatic improvement in some patients potentially by increasing the overall diversity of the colonic microbiome. However, randomized controlled trials of FMT are needed to assess the clinical efficacy of FMT in CD and to identify components of the donor microbiota that can suppress potentially pathogenic bacterial families in the patients. This study complements previous, partly contradictory studies of FMT in IBD (Moayyedi et al., 2015; Rossen et al., 2015), and it is yet to be shown whether FMT can be efficiently and reproducibly used to cure different subtypes of IBD.

This thesis introduced novel methods for analyzing WMS sequencing data for microbial pathways as well as strain level taxonomic identification of microbes. HUMAnN2 analyses combined with downstream quantification of GO terms provides a convenient way of studying microbial communities at the functional level. These quantification can be improved in the future by identifying high-specificity genes per metabolic pathway, rather than using all contributing genes to quantify a pathway. In Publications I and III, we analyzed the WMS sequencing data using ConStrains, and developed novel downstream analyses for these data. We adapted the concept of heterozygosity from population genetics to measure within-species diversity on strain level and investigated phylogenetic relationships between different strains. There are many unexplored avenues on strain level analysis and it remains to be seen, for example, how to best interpret the implications of SNP level differences between strains.

Current databases utilized and populated by the microbiome researchers contain tens of thousands of microbial genomes (e.g., NCBI RefSeq) and tens of millions of protein sequences (e.g., Universal Protein Resource, UniProt). Regardless, this reference material covers only about 50 % of the human gut metagenome at most, usually significantly less (Joice et al., 2014). More is known about the gut microbiome of people with western lifestyle, whereas the gut microbiomes of isolated, indigenous communities are typically remarkably different from westerners and poorly covered by existing reference material (Schnorr et al., 2014; Obregon-Tito et al., 2015). This bias has been previously noted in psychology and dubbed “WEIRD”, according to the traits that are often overrepresented in scientific studies: Western, educated, industrial, rich and democratic (Henrich et al., 2010). Portion of the gut metagenome that lays beyond

our current understanding, the metagenomic “dark matter”, offers plenty of challenges for both bioinformaticians and microbiologists in years to come. Below I provide an outlook on how these challenges may be approached.

A portion of metagenomic reads that cannot be assigned to any known genomes can still be mapped to known proteins using translated search. While large proportion of catalogued proteins lack functional annotation, this offers a data-driven way to prioritize proteins with both high prevalence and high abundance in known metagenomes for functional characterization. Recent advances in genome engineering techniques, such as CRISPR (clustered regularly interspaced short palindromic repeats)/Cas9 (Peters et al., 2016) and chemical synthesis of entire bacterial genomes (Hutchison et al., 2016), offer multiple ways to study the function of any given gene by systematic phenotypic analyses.

Assembly-based methods offer many venues to extend our understanding of the dark matter. For example, it may be possible to assign a subset of tentative proteins to known clades in situations where they occur on longer contigs with confident taxonomic assignments. Offering benefits similar to assembly, a few competing sequencing techniques offer similar benefits by generating significantly longer reads (>10,000 base pairs) compared to Illumina platform. Indeed, Pacific Biosciences (PacBio) is already the choice of many microbiologists when the aim is to sequence a complete genome of a single organism. Using PacBio sequencing, one can obtain a high quality genome with rudimentary computational processing. Nanopore sequencing, where DNA is sequenced as it flows through small pore with a voltage applied across it, is developed by Oxford Nanopore Technologies. Their sequencing machine called MinION is truly portable in size (think of an usb-drive), and may revolutionize DNA sequencing, given that they are able to solve some final technical hurdles (Deamer et al., 2016).

In addition to sequencing-based studies and computational analysis, performing experiments on gut strains in the laboratory is essential to complement the results obtained using computational methods. Majority of the gut microbes are obligate anaerobes, which are challenging to isolate and cultivate in the laboratory. However, the development of specific complex media and adequate growth techniques (Browne et al., 2016) has enabled capturing a great majority of the gut microbial population in anaerobic chambers. In addition to monocultures, mimicking the gut mi-

crobial complexity in a bioreactor has shown to be key to the study of community composition and interdependencies in this elaborate ecosystem (McDonald et al., 2013). Investigators using these techniques in concert with data from large cohort studies are likely to spearhead the gut microbiome research with new discoveries in the coming years.

References

- Aagaard, K., Ma, J., Antony, K. M., Ganu, R., Petrosino, J., and Versalovic, J. The placenta harbors a unique microbiome. *Science Translational Medicine*, 6(237):237ra65–237ra65, 2014.
- Abrahamsson, T. R., Jakobsson, H. E., Andersson, A. F., Björkstén, B., Engstrand, L., and Jenmalm, M. C. Low diversity of the gut microbiota in infants with atopic eczema. *Journal of Allergy and Clinical Immunology*, 129(2):434–440, 2012.
- Abrahamsson, T., Jakobsson, H., Andersson, A. F., Björkstén, B., Engstrand, L., and Jenmalm, M. Low gut microbiota diversity in early infancy precedes asthma at school age. *Clinical & Experimental Allergy*, 44(6):842–850, 2014.
- Abubucker, S., Segata, N., Goll, J., Schubert, A. M., Izard, J., Cantarel, B. L., Rodriguez-Mueller, B., Zucker, J., Thiagarajan, M., Henrissat, B., White, O., Kelley, S. T., Methe, B., Schloss, P. D., Gevers, D., Mitreva, M., and Huttenhower, C. Metabolic reconstruction for metagenomic data and its application to the human microbiome. *PLOS Computational Biology*, 8(6):e1002358, 2012.
- Ahn, J., Sinha, R., Pei, Z., Dominianni, C., Wu, J., Shi, J., Goedert, J. J., Hayes, R. B., and Yang, L. Human gut microbiome and risk for colorectal cancer. *Journal of the National Cancer Institute*, 105(24):1907–1911, 2013.
- Ahn, T. H., Chai, J., and Pan, C. Sigma: strain-level inference of genomes from metagenomic analysis for biosurveillance. *Bioinformatics*, 31(2):170–177, 2015.
- Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410, 1990.
- Arrieta, M.-C., Stiemsma, L. T., Amenogbe, N., Brown, E. M., and Finlay, B. The intestinal microbiome in early life: health and disease. *Frontiers in Immunology*, 5:427, 2014.
- Arrieta, M.-C., Stiemsma, L. T., Dimitriu, P. A., Thorson, L., Russell, S., Yurist-Doutsch, S., Kuzeljevic, B., Gold, M. J., Britton, H. M., Lefebvre, D. L., Subbarao, P., Mandhane, P., Becker, A., McNagny, K. M., Sears, M. R., Kollmann, T., Mohn, W. W., Turvey, S. E., and Brett Finlay, B. Early infancy microbial and metabolic alterations affect risk of childhood asthma. *Science Translational Medicine*, 7(307):307ra152–307ra152, 2015.

- Arvonen, M., Berntson, L., Pokka, T., Karttunen, T. J., Vähäsalo, P., and Stoll, M. L. Gut microbiota-host interactions and juvenile idiopathic arthritis. *Pediatric Rheumatology*, 14(1):1–9, 2016.
- Atarashi, K., Tanoue, T., Oshima, K., Suda, W., Nagano, Y., Nishikawa, H., Fukuda, S., Saito, T., Narushima, S., Hase, K., Kim, S., Fritz, J. V., Wilmes, P., Ueha, S., Matsushima, K., Ohno, H., Olle, B., Sakaguchi, S., Taniguchi, T., Morita, H., Hattori, M., and Honda, K. Treg induction by a rationally selected mixture of Clostridia strains from the human microbiota. *Nature*, 500(7461):232–236, 2013.
- Aumeunier, A., Grela, F., Ramadan, A., Pham Van, L., Bardel, E., Gomez Alcala, A., Jeannin, P., Akira, S., Bach, J. F., and Thieblemont, N. Systemic toll-like receptor stimulation suppresses experimental allergic asthma and autoimmune diabetes in nod mice. *PLOS ONE*, 5(7):e11484, 2010.
- Bach, J. F. The effect of infections on susceptibility to autoimmune and allergic diseases. *New England Journal of Medicine*, 347(12):911–920, 2002.
- Bach, J. F. and Chatenoud, L. The hygiene hypothesis: an explanation for the increased frequency of insulin-dependent diabetes. *Cold Spring Harbor Perspectives in Medicine*, 2(2):a007799, 2012.
- Backhed, F., Roswall, J., Peng, Y., Feng, Q., Jia, H., Kovatcheva-Datchary, P., Li, Y., Xia, Y., Xie, H., Zhong, H., Khan, M. T., Zhang, J., Li, J., Xiao, L., Al-Aama, J., Zhang, D., Lee, Y. S., Kotowska, D., Colding, C., Tremaroli, V., Yin, Y., Bergman, S., Xu, X., Madsen, L., Kristiansen, K., Dahlgren, J., and Wang, J. Dynamics and stabilization of the human gut microbiome during the first year of life. *Cell Host & Microbe*, 17(5):690–703, 2015.
- Bäumler, A. J. and Sperandio, V. Interactions between the microbiota and pathogenic bacteria in the gut. *Nature*, 535(7610):85–93, 2016.
- Benjamini, Y. and Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*, 57(1):289–300, 1995.
- Biswas, S. K. and Lopez-Collazo, E. Endotoxin tolerance: new mechanisms, molecules and clinical significance. *Trends in Immunology*, 30(10):475–487, 2009.
- Blaser, M. J. Antibiotic use and its consequences for the normal microbiome. *Science*, 352(6285):544–545, 2016.
- Bordenstein, S. R. and Theis, K. R. Host biology in light of the microbiome: Ten principles of holobionts and hologenomes. *PLOS Biology*, 13(8):1–23, 08 2015.
- Borg, I. and Groenen, P. J. *Modern multidimensional scaling: Theory and applications*. Springer Science & Business Media, 2005.
- Breiman, L. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.
- Breiman, L. Random forests. *Machine learning*, 45(1):5–32, 2001.
- Breslow, N. E. and Clayton, D. G. Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88(421):9–25, 1993.

- Brito, I., Yilmaz, S., Huang, K., Xu, L., Jupiter, S., Jenkins, A., Naisilisili, W., Tamminen, M., Smillie, C., Wortman, J., Birren, B., Xavier, R., Blainey, P., Singh, A., Gevers, D., and Alm, E. Mobile genes in the human microbiome are structured from global to individual scales. *Nature*, 535(7612):435–439, 2016.
- Browne, H. P., Forster, S. C., Anonye, B. O., Kumar, N., Neville, B. A., Stares, M. D., Goulding, D., and Lawley, T. D. Culturing of ‘unculturable’ human microbiota reveals novel taxa and extensive sporulation. *Nature*, 533(7604):543–546, 2016.
- Buchfink, B., Xie, C., and Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nature Methods*, 12(1):59–60, 2015.
- Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., and Holmes, S. P. DADA2: High-resolution sample inference from illumina amplicon data. *Nature Methods*, 13(7):581–583, 2016.
- 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., Huttley, G. A., Kelley, S. T., Knights, D., Koenig, J. E., Ley, R. E., Lozupone, C. A., McDonald, D., Muegge, B. D., Pirrung, M., Reeder, J., Sevinsky, J. R., Turnbaugh, P. J., Walters, W. A., Widmann, J., Yatsunenko, T., Zaneveld, J., and Knight, R. QIIME allows analysis of high-throughput community sequencing data. *Nature Methods*, 7(5):335–336, 2010.
- Caporaso, J. G., Lauber, C. L., Walters, W. A., Berg-Lyons, D., Huntley, J., Fierer, N., Owens, S. M., Betley, J., Fraser, L., Bauer, M., Cormley, N., Gilbert, J. A., Smith, G., and Knight, R. Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *ISME Journal*, 6(8):1621–1624, 2012.
- Caspi, R., Altman, T., Dreher, K., Fulcher, C. A., Subhraveti, P., Keseler, I. M., Kothari, A., Krummenacker, M., Latendresse, M., Mueller, L. A., Ong, Q., Paley, S., Pujar, A., Shearer, A. G., Travers, M., Weerasinghe, D., Zhang, P., and Karp, P. D. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Research*, 40(Database issue):D742–D753, 2012.
- Chao, A. Nonparametric estimation of the number of classes in a population. *Scandinavian Journal of Statistics*, 11(4):265–270, 1984.
- Charbonneau, M. R., Blanton, L. V., DiGiulio, D. B., Relman, D. A., Lebrilla, C. B., Mills, D. A., and Gordon, J. I. A microbial perspective of human developmental biology. *Nature*, 535(7610):48–55, 2016.
- Cheng, J., Ringel-Kulka, T., Heikamp-de Jong, I., Ringel, Y., Carroll, I., de Vos, W. M., Salojärvi, J., and Satokari, R. Discordant temporal development of bacterial phyla and the emergence of core in the fecal microbiota of young children. *ISME Journal*, 10(4):1002–1014, 2015.
- Clarke, G., Stilling, R. M., Kennedy, P. J., Stanton, C., Cryan, J. F., and Dinan, T. G. Minireview: Gut microbiota: the neglected endocrine organ. *Molecular endocrinology*, 28(8):1221–1238, 2014.

- Clemente, J. C., Pehrsson, E. C., Blaser, M. J., Sandhu, K., Gao, Z., Wang, B., Magris, M., Hidalgo, G., Contreras, M., Noya-Alarcon, O., Lander, O., McDonald, J., Cox, M., Walter, J., Oh, P. L., Ruiz, J. F., Rodriguez, S., Shen, N., Song, S. J., Metcalf, J., Knight, R., Dantas, G., and Dominguez-Bello, M. G. The microbiome of uncontacted amerindians. *Science Advances*, 1(3), 2015.
- Conlon, M. A. and Bird, A. R. The impact of diet and lifestyle on gut microbiota and human health. *Nutrients*, 7(1):17–44, 2015.
- Conly, J., Stein, K., Worobetz, L., and Rutledge-Harding, S. The contribution of vitamin K 2 (Menaquinones) produced by the intestinal microflora to human nutritional requirements for vitamin K. *American Journal of Gastroenterology*, 89(6), 1994.
- Coyte, K. Z., Schluter, J., and Foster, K. R. The ecology of the microbiome: Networks, competition, and stability. *Science*, 350(6261):663–666, 2015.
- Cullen, T. W., Schofield, W. B., Barry, N. A., Putnam, E. E., Rundell, E. A., Trent, M. S., Degnan, P. H., Booth, C. J., Yu, H., and Goodman, A. L. Gut microbiota. antimicrobial peptide resistance mediates resilience of prominent gut commensals during inflammation. *Science*, 347(6218):170–175, 2015.
- David, L. A., Maurice, C. F., Carmody, R. N., Gootenberg, D. B., Button, J. E., Wolfe, B. E., Ling, A. V., Devlin, A. S., Varma, Y., Fischbach, M. A., Biddinger, S. B., Dutton, R. J., and Turnbaugh, P. J. Diet rapidly and reproducibly alters the human gut microbiome. *Nature*, 505(7484):559–563, 2014.
- Davis, M. P., van Dongen, S., Abreu-Goodger, C., Bartonicek, N., and Enright, A. J. Kraken: a set of tools for quality control and analysis of high-throughput sequence data. *Methods*, 63(1):41–49, 2013.
- Davis-Richardson, A. G., Ardisson, A. N., Dias, R., Simell, V., Leonard, M. T., Kempainen, K. M., Drew, J. C., Schatz, D., Atkinson, M. A., Kolaczowski, B., Ilonen, J., Knip, M., Toppaari, J., Nurminen, N., Hyoty, H., Veijola, R., Simell, T., Mykkanen, J., Simell, O., and Triplett, E. W. *Bacteroides dorei* dominates gut microbiome prior to autoimmunity in finnish children at high risk for type 1 diabetes. *Frontiers in Microbiology*, 5:678, 2014.
- De Filippis, F., Pellegrini, N., Vannini, L., Jeffery, I. B., La Stora, A., Laghi, L., Serrazanetti, D. I., Di Cagno, R., Ferrocino, I., Lazzi, C., Turrone, S., Colicin, L., Brigidi, P., Neviani, E., Gobetti, M., O'Toole, P. W., and Ercolini, D. High-level adherence to a mediterranean diet beneficially impacts the gut microbiota and associated metabolome. *Gut*, 2015.
- Deamer, D., Akeson, M., and Branton, D. Three decades of nanopore sequencing. *Nature Biotechnology*, 34(5):518–524, 2016.
- Dethlefsen, L. and Relman, D. A. Incomplete recovery and individualized responses of the human distal gut microbiota to repeated antibiotic perturbation. *Proceedings of the National Academy of Sciences*, 108 Suppl 1:4554–4561, 2011.
- Diggle, P. *Analysis of longitudinal data*. Oxford University Press, 2002.
- Dimmer, E. C., Huntley, R. P., Alam-Faruque, Y., Sawford, T., O'Donovan, C., Martin, M. J., Bely, B., Browne, P., Mun Chan, W., Eberhardt, R., Gardner,

- M., Laiho, K., Legge, D., Magrane, M., Pichler, K., Poggioli, D., Sehra, H., Auchincloss, A., Axelsen, K., Blatter, M. C., Boutet, E., Braconi-Quintaje, S., Breuza, L., Bridge, A., Coudert, E., Estreicher, A., Famiglietti, L., Ferro-Rojas, S., Feuermann, M., Gos, A., Gruaz-Gumowski, N., Hinz, U., Hulo, C., James, J., Jimenez, S., Jungo, F., Keller, G., Lemercier, P., Lieberherr, D., Masson, P., Moinat, M., Pedruzzi, I., Poux, S., Rivoire, C., Roechert, B., Schneider, M., Stutz, A., Sundaram, S., Tognolli, M., Bougueleret, L., Argoud-Puy, G., Cusin, I., Duek-Roggli, P., Xenarios, I., and Apweiler, R. The UniProt-GO annotation database in 2011. *Nucleic Acids Research*, 40(Database issue):D565–D570, 2012.
- Dogra, S., Sakwinska, O., Soh, S.-E., Ngom-Bru, C., Brück, W. M., Berger, B., Brüssow, H., Lee, Y. S., Yap, F., Chong, Y.-S., Godfrey, K. M., and Holbrook, J. D. Dynamics of infant gut microbiota are influenced by delivery mode and gestational duration and are associated with subsequent adiposity. *mBio*, 6(1), 2015.
- Dominguez-Bello, M. G., Costello, E. K., Contreras, M., Magris, M., Hidalgo, G., Fierer, N., and Knight, R. Delivery mode shapes the acquisition and structure of the initial microbiota across multiple body habitats in newborns. *Proceedings of the National Academy of Sciences*, 107(26):11971–11975, 2010a.
- Dominguez-Bello, M. G., Costello, E. K., Contreras, M., Magris, M., Hidalgo, G., Fierer, N., and Knight, R. Delivery mode shapes the acquisition and structure of the initial microbiota across multiple body habitats in newborns. *Proceedings of the National Academy of Sciences*, 107(26):11971–11975, 2010b.
- Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research*, 32(5):1792–1797, 2004.
- Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, 26(19):2460–2461, 2010.
- Edgar, R. C. UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nature Methods*, 10(10):996–998, 2013.
- Edgar, R. C. and Flyvbjerg, H. Error filtering, pair assembly and error correction for next-generation sequencing reads. *Bioinformatics*, 31(21):3476–3482, 2015.
- Eren, A. M., Maignien, L., Sul, W. J., Murphy, L. G., Grim, S. L., Morrison, H. G., and Sogin, M. L. Oligotyping: Differentiating between closely related microbial taxa using 16S rRNA gene data. *Methods in Ecology and Evolution*, 4(12), 2013.
- Eren, A. M., Morrison, H. G., Lescault, P. J., Reveillaud, J., Vineis, J. H., and Sogin, M. L. Minimum entropy decomposition: unsupervised oligotyping for sensitive partitioning of high-throughput marker gene sequences. *ISME Journal*, 9(4):968–979, 2015.
- Faith, J. J., Guruge, J. L., Charbonneau, M., Subramanian, S., Seedorf, H., Goodman, A. L., Clemente, J. C., Knight, R., Heath, A. C., Leibel, R. L., Rosenbaum, M., and Gordon, J. I. The long-term stability of the human gut microbiota. *Science*, 341(6141):1237439, 2013.

- Fox, C. and Eichelberger, K. Maternal microbiome and pregnancy outcomes. *Fertility and sterility*, 104(6):1358–1363, 2015.
- Frank, D. N., Amand, A. L. S., Feldman, R. A., Boedeker, E. C., Harpaz, N., and Pace, N. R. Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases. *Proceedings of the National Academy of Sciences*, 104(34):13780–13785, 2007.
- Franzosa, E. A., Huang, K., Meadow, J. F., Gevers, D., Lemon, K. P., Bohannan, B. J., and Huttenhower, C. Identifying personal microbiomes using metagenomic codes. *Proceedings of the National Academy of Sciences*, 112(22):E2930–E2938, 2015.
- Fu, J., Bonder, M. J., Cenit, M. C., Tigchelaar, E. F., Maatman, A., Dekens, J. A., Brandsma, E., Marczyńska, J., Imhann, F., Weersma, R. K., Franke, L., Poon, T. W., Xavier, R. J., Gevers, D., Hofker, M. H., Wijmenga, C., and Zhernakova, A. The gut microbiome contributes to a substantial proportion of the variation in blood lipids. *Circulation Research*, 117(9):817–824, 2015.
- Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, 28(23):3150–3152, 2012.
- Gardiner, J. C., Luo, Z., and Roman, L. A. Fixed effects, random effects and GEE: What are the differences? *Statistics in Medicine*, 28(2):221–239, 2009.
- Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. *Bayesian data analysis*, volume 2. Chapman & Hall/CRC Boca Raton, FL, USA, 2014.
- Gene Ontology Consortium. Gene Ontology Consortium: going forward. *Nucleic Acids Research*, 43(Database issue):D1049–D1056, 2015.
- Gilbert, J. A., Quinn, R. A., Debelius, J., Xu, Z. Z., Morton, J., Garg, N., Jansson, J. K., Dorrestein, P. C., and Knight, R. Microbiome-wide association studies link dynamic microbial consortia to disease. *Nature*, 535(7610):94–103, 2016.
- Glass, E. M., Wilkening, J., Wilke, A., Antonopoulos, D., and Meyer, F. Using the metagenomics RAST server (MG-RAST) for analyzing shotgun metagenomes. *Cold Spring Harbor Protocols*, 2010(1):pdb.prot5368, 2010.
- Gomez-Gallego, C., Garcia-Mantrana, I., Salminen, S., and Collado, M. C. The human milk microbiome and factors influencing its composition and activity. In *Seminars in Fetal and Neonatal Medicine*. Elsevier, 2016.
- Goodwin, S., McPherson, J. D., and McCombie, W. R. Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, 17(6):333–351, 2016.
- Gordon, J., Knowlton, N., Relman, D. A., Rohwer, F., and Youle, M. Superorganisms and holobionts. *Microbe*, 8(4):152–153, 2013.
- Grubbs, F. E. Sample criteria for testing outlying observations. *The Annals of Mathematical Statistics*, pages 27–58, 1950.
- Hajjar, A. M., Ernst, R. K., Tsai, J. H., Wilson, C. B., and Miller, S. I. Human Toll-like receptor 4 recognizes host-specific LPS modifications. *Nature Immunology*, 3(4):354–359, 2002.

- Henrich, J., Heine, S. J., and Norenzayan, A. Most people are not WEIRD. *Nature*, 466(7302):29–29, 2010.
- Herath, T. D., Wang, Y., Seneviratne, C. J., Lu, Q., Darveau, R. P., Wang, C. Y., and Jin, L. Porphyromonas gingivalis lipopolysaccharide lipid a heterogeneity differentially modulates the expression of IL-6 and IL-8 in human gingival fibroblasts. *Journal of Clinical Periodontology*, 38(8):694–701, 2011.
- Honda, K. and Littman, D. R. The microbiota in adaptive immune homeostasis and disease. *Nature*, 535(7610):75–84, 2016.
- Hong, C., Manimaran, S., Shen, Y., Perez-Rogers, J. F., Byrd, A. L., Castro-Nallar, E., Crandall, K. A., and Johnson, W. E. PathoScope 2.0: a complete computational framework for strain identification in environmental or clinical sequencing samples. *Microbiome*, 2:33, 2014.
- Howe, A. and Chain, P. S. Challenges and opportunities in understanding microbial communities with metagenome assembly (accompanied by IPython Notebook tutorial). *Frontiers in Microbiology*, 6:678, 2015.
- Huang, K., Brady, A., Mahurkar, A., White, O., Gevers, D., Huttenhower, C., and Segata, N. MetaRef: a pan-genomic database for comparative and community microbial genomics. *Nucleic Acids Research*, 42(Database issue):D617–D624, 2014.
- Huang, Y., Li, H., Hu, H., Yan, X., Waterman, M. S., Huang, H., and Zhou, X. J. Systematic discovery of functional modules and context-specific functional annotation of human genome. *Bioinformatics*, 23(13):i222–i229, 2007.
- Human Microbiome Project, C. Structure, function and diversity of the healthy human microbiome. *Nature*, 486(7402):207–214, 2012.
- Hunt, K. M., Foster, J. A., Forney, L. J., Schutte, U. M., Beck, D. L., Abdo, Z., Fox, L. K., Williams, J. E., McGuire, M. K., and McGuire, M. A. Characterization of the diversity and temporal stability of bacterial communities in human milk. *PLOS ONE*, 6(6):e21313, 2011.
- Huson, D. H., Beier, S., Flade, I., Górski, A., El-Hadidi, M., Mitra, S., Ruscheweyh, H.-J., and Tappu, R. Megan community edition - interactive exploration and analysis of large-scale microbiome sequencing data. *PLOS Computational Biology*, 12(6):1–12, 06 2016.
- Hutchison, C. A., Chuang, R.-Y., Noskov, V. N., Assad-Garcia, N., Deerinck, T. J., Ellisman, M. H., Gill, J., Kannan, K., Karas, B. J., Ma, L., et al. Design and synthesis of a minimal bacterial genome. *Science*, 351(6280):aad6253, 2016.
- Jakobsson, H. E., Jernberg, C., Andersson, A. F., Sjölund-Karlsson, M., Jansson, J. K., and Engstrand, L. Short-term antibiotic treatment has differing long-term impacts on the human throat and gut microbiome. *PLOS ONE*, 5(3): 1–12, 03 2010.
- Jakobsson, H. E., Abrahamsson, T. R., Jenmalm, M. C., Harris, K., Quince, C., Jernberg, C., Björkstén, B., Engstrand, L., and Andersson, A. F. Decreased gut microbiota diversity, delayed Bacteroidetes colonisation and reduced Th1 responses in infants delivered by caesarean section. *Gut*, 63(4):559–566, 2014.

- Joice, R., Yasuda, K., Shafquat, A., Morgan, X., and Huttenhower, C. Determining microbial products and identifying molecular targets in the human microbiome. *Cell Metabolism*, 20(5):731–741, 2014.
- Jost, T., Lacroix, C., Braegger, C., and Chassard, C. Impact of human milk bacteria and oligosaccharides on neonatal gut microbiota establishment and gut health. *Nutrition reviews*, 73(7):426–437, 2015.
- Kaminski, J., Gibson, M. K., Franzosa, E. A., Segata, N., Dantas, G., and Huttenhower, C. High-specificity targeted functional profiling in microbial communities with ShortBRED. *PLOS Computational Biology*, 11(12):1–22, 12 2015.
- Kang, D.-W., Park, J. G., Ilhan, Z. E., Wallstrom, G., LaBaer, J., Adams, J. B., and Krajmalnik-Brown, R. Reduced incidence of prevotella and other fermenters in intestinal microflora of autistic children. *PLOS ONE*, 8(7):e68322, 2013.
- Kelly, C. R., de Leon, L., and Jasutkar, N. Fecal microbiota transplantation for relapsing *Clostridium difficile* infection in 26 patients: methodology and results. *Journal of Clinical Gastroenterology*, 46(2):145–149, 2012.
- Kelly, C. R., Kahn, S., Kashyap, P., Laine, L., Rubin, D., Atreja, A., Moore, T., and Wu, G. Update on fecal microbiota transplantation 2015: Indications, methodologies, mechanisms, and outlook. *Gastroenterology*, 149(1):223–237, 2015.
- Kim, H. M., Park, B. S., Kim, J. I., Kim, S. E., Lee, J., Oh, S. C., Enkhbayar, P., Matsushima, N., Lee, H., Yoo, O. J., and Lee, J. O. Crystal structure of the TLR4-MD-2 complex with bound endotoxin antagonist eritoran. *Cell*, 130(5):906–917, 2007.
- Knip, M. and Siljander, H. The role of the intestinal microbiota in type 1 diabetes mellitus. *Nature Reviews Endocrinology*, 12(3):154–167, 2016.
- Kondrashova, A., Mustalahti, K., Kaukinen, K., Viskari, H., Volodicheva, V., Haapala, A. M., Ilonen, J., Knip, M., Maki, M., Hyoty, H., and Epivir Study, G. Lower economic status and inferior hygienic environment may protect against celiac disease. *Annals of Medicine*, 40(3):223–231, 2008a.
- Kondrashova, A., Viskari, H., Haapala, A. M., Seiskari, T., Kulmala, P., Ilonen, J., Knip, M., and Hyoty, H. Serological evidence of thyroid autoimmunity among schoolchildren in two different socioeconomic environments. *Journal of Clinical Endocrinology & Metabolism*, 93(3):729–734, 2008b.
- Konikoff, M. R. and Denson, L. A. Role of fecal calprotectin as a biomarker of intestinal inflammation in inflammatory bowel disease. *Inflammatory bowel diseases*, 12(6):524–534, 2006.
- Korpela, K., Salonen, A., Virta, L. J., Kekkonen, R. A., Forslund, K., Bork, P., and de Vos, W. M. Intestinal microbiome is related to lifetime antibiotic use in finnish pre-school children. *Nature Communications*, 7:10410, 2016.
- Kuczynski, J., Stombaugh, J., Walters, W. A., Gonzalez, A., Caporaso, J. G., and Knight, R. Using QIIME to analyze 16S rRNA gene sequences from microbial communities. *Current Protocols in Microbiology*, Chapter 1:Unit 1E 5, 2012.
- Langmead, B. and Salzberg, S. L. Fast gapped-read alignment with bowtie 2. *Nature Methods*, 9(4):357–359, 2012.

- Larizza, D., Calcaterra, V., Klersy, C., Badulli, C., Caramagna, C., Ricci, A., Brambilla, P., Salvaneschi, L., and Martinetti, M. Common immunogenetic profile in children with multiple autoimmune diseases: the signature of HLA-DQ pleiotropic genes. *Autoimmunity*, 45(6):470–475, 2012.
- Lax, S., Smith, D. P., Hampton-Marcell, J., Owens, S. M., Handley, K. M., Scott, N. M., Gibbons, S. M., Larsen, P., Shogan, B. D., Weiss, S., Metcalf, J. L., Ursell, L. K., Vazquez-Baeza, Y., Van Treuren, W., Hasan, N. A., Gibson, M. K., Colwell, R., Dantas, G., Knight, R., and Gilbert, J. A. Longitudinal analysis of microbial interaction between humans and the indoor environment. *Science*, 345(6200):1048–1052, 2014.
- Lewis, J. D., Chen, E. Z., Baldassano, R. N., Otley, A. R., Griffiths, A. M., Lee, D., Bittinger, K., Bailey, A., Friedman, E. S., Hoffmann, C., Albenberg, L., Sinha, R., Compher, C., Gilroy, E., Nessel, L., Grant, A., Chehoud, C., Li, H., Wu, G. D., and Bushman, F. D. Inflammation, antibiotics, and diet as environmental stressors of the gut microbiome in pediatric Crohn’s disease. *Cell Host & Microbe*, 18(4):489–500, 2015.
- Ley, R. E., Turnbaugh, P. J., Klein, S., and Gordon, J. I. Microbial ecology: human gut microbes associated with obesity. *Nature*, 444(7122):1022–1023, 2006.
- Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and Genome Project Data Processing, S. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, 2009.
- Li, S. S., Zhu, A., Benes, V., Costea, P. I., Hercog, R., Hildebrand, F., Huerta-Cepas, J., Nieuwdorp, M., Salojärvi, J., Voigt, A. Y., Zeller, G., Sunagawa, S., de Vos, W. M., and Bork, P. Durable coexistence of donor and recipient strains after fecal microbiota transplantation. *Science*, 352(6285):586–589, 2016.
- Lindgreen, S., Adair, K. L., and Gardner, P. P. An evaluation of the accuracy and speed of metagenome analysis tools. *Scientific reports*, 6, 2016.
- Lloyd-Price, J., Abu-Ali, G., and Huttenhower, C. The healthy human microbiome. *Genome Medicine*, 8(1):51, 2016.
- Locascio, R. G., Ninonuevo, M. R., Kronewitter, S. R., Freeman, S. L., German, J. B., Lebrilla, C. B., and Mills, D. A. A versatile and scalable strategy for glycoprofiling bifidobacterial consumption of human milk oligosaccharides. *Microbial Biotechnology*, 2(3):333–342, 2009.
- Lozupone, C. A., Stombaugh, J. I., Gordon, J. I., Jansson, J. K., and Knight, R. Diversity, stability and resilience of the human gut microbiota. *Nature*, 489(7415):220–230, 2012.
- Luo, C., Knight, R., Siljander, H., Knip, M., Xavier, R. J., and Gevers, D. Con-Strains identifies microbial strains in metagenomic datasets. *Nature Biotechnology*, 33(10):1045–1052, 2015.
- Manichanh, C., Borruel, N., Casellas, F., and Guarner, F. The gut microbiota in IBD. *Nature Reviews Gastroenterology and Hepatology*, 9(10):599–608, 2012.
- Marcobal, A., Barboza, M., Sonnenburg, E. D., Pudlo, N., Martens, E. C., Desai, P., Lebrilla, C. B., Weimer, B. C., Mills, D. A., German, J. B., and Sonnenburg,

- J. L. Bacteroides in the infant gut consume milk oligosaccharides via mucus-utilization pathways. *Cell Host & Microbe*, 10(5):507–514, 2011.
- Markle, J. G., Frank, D. N., Mortin-Toth, S., Robertson, C. E., Feazel, L. M., Rolle-Kampczyk, U., von Bergen, M., McCoy, K. D., Macpherson, A. J., and Danska, J. S. Sex differences in the gut microbiome drive hormone-dependent regulation of autoimmunity. *Science*, 339(6123):1084–1088, 2013.
- Markowitz, V. M., Chen, I. M., Chu, K., Szeto, E., Palaniappan, K., Pillay, M., Ratner, A., Huang, J., Pagani, I., Tringe, S., Huntemann, M., Billis, K., Varghese, N., Tennessen, K., Mavromatis, K., Pati, A., Ivanova, N. N., and Kyrpides, N. C. IMG/M 4 version of the integrated metagenome comparative analysis system. *Nucleic Acids Research*, 42(Database issue):D568–D573, 2014.
- Martens, J.-H., Barg, H., Warren, M., and Jahn, D. Microbial production of vitamin B12. *Applied Microbiology and Biotechnology*, 58(3):275–285, 2002.
- Matsuoka, K. and Kanai, T. The gut microbiota and inflammatory bowel disease. *Seminars in Immunopathology*, 37(1):47–55, 2015.
- McArthur, A. G., Wagelchner, N., Nizam, F., Yan, A., Azad, M. A., Baylay, A. J., Bhullar, K., Canova, M. J., De Pascale, G., Ejim, L., Kalan, L., King, A. M., Koteva, K., Morar, M., Mulvey, M. R., O'Brien, J. S., Pawlowski, A. C., Piddock, L. J., Spanogiannopoulos, P., Sutherland, A. D., Tang, I., Taylor, P. L., Thaker, M., Wang, W., Yan, M., Yu, T., and Wright, G. D. The comprehensive antibiotic resistance database. *Antimicrobial Agents and Chemotherapy*, 57(7):3348–3357, 2013.
- McDonald, D., Price, M. N., Goodrich, J., Nawrocki, E. P., DeSantis, T. Z., Probst, A., Andersen, G. L., Knight, R., and Hugenholtz, P. An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *ISME Journal*, 6(3):610–618, 2012.
- McDonald, J. A., Schroeter, K., Fuentes, S., Heikamp-deJong, I., Khursigara, C. M., de Vos, W. M., and Allen-Vercoe, E. Evaluation of microbial community reproducibility, stability and composition in a human distal gut chemostat model. *Journal of Microbiological Methods*, 95(2):167–174, 2013.
- McGuire, M. K. and McGuire, M. A. Human milk: mother nature's prototypical probiotic food? *Advances in Nutrition*, 6(1):112–123, 2015.
- Mitra, S., Rupek, P., Richter, D. C., Urich, T., Gilbert, J. A., Meyer, F., Wilke, A., and Huson, D. H. Functional analysis of metagenomes and metatranscriptomes using SEED and KEGG. *BMC Bioinformatics*, 12 Suppl 1:S21, 2011.
- Moayyedi, P., Surette, M. G., Kim, P. T., Libertucci, J., Wolfe, M., Onischi, C., Armstrong, D., Marshall, J. K., Kassam, Z., Reinisch, W., and Lee, C. H. Fecal microbiota transplantation induces remission in patients with active Ulcerative Colitis in a randomized controlled trial. *Gastroenterology*, 149(1):102–109.e6, 2015.
- Morgan, X. C., Tickle, T. L., Sokol, H., Gevers, D., Devaney, K. L., Ward, D. V., Reyes, J. A., Shah, S. A., LeLeiko, N., Snapper, S. B., Bousvaros, A., Korzenik, J., Sands, B. E., Xavier, R. J., and Huttenhower, C. Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. *Genome Biology*, 13(9):R79, 2012.

- Naseribafrouei, A., Hestad, K., Avershina, E., Sekelja, M., Linløkken, A., Wilson, R., and Rudi, K. Correlation between the human fecal microbiota and depression. *Neurogastroenterology & Motility*, 26(8):1155–1162, 2014.
- Needham, B. D., Carroll, S. M., Giles, D. K., Georgiou, G., Whiteley, M., and Trent, M. S. Modulating the innate immune response by combinatorial engineering of endotoxin. *Proceedings of the National Academy of Sciences*, 110(4): 1464–1469, 2013.
- Nobel, Y. R., Cox, L. M., Kirigin, F. F., Bokulich, N. A., Yamanishi, S., Teitler, I., Chung, J., Sohn, J., Barber, C. M., Goldfarb, D. S., Raju, K., Abubucker, S., Zhou, Y., Ruiz, V. E., Li, H., Mitreva, M., Alekseyenko, A. V., Weinstock, G. M., Sodergren, E., and Blaser, M. J. Metabolic and metagenomic outcomes from early-life pulsed antibiotic treatment. *Nature Communications*, 6:7486, 2015.
- Nylund, L., Satokari, R., Nikkilä, J., Rajilić-Stojanović, M., Kalliomäki, M., Isolauri, E., Salminen, S., and de Vos, W. M. Microarray analysis reveals marked intestinal microbiota aberrancy in infants having eczema compared to healthy children in at-risk for atopic disease. *BMC Microbiology*, 13(1):1–11, 2013.
- Obregon-Tito, A. J., Tito, R. Y., Metcalf, J., Sankaranarayanan, K., Clemente, J. C., Ursell, L. K., Xu, Z. Z., Van Treuren, W., Knight, R., Gaffney, P. M., et al. Subsistence strategies in traditional societies distinguish gut microbiomes. *Nature Communications*, 6, 2015.
- O'Hara, A. M. and Shanahan, F. The gut flora as a forgotten organ. *EMBO reports*, 7(7):688–693, 2006.
- Ohman, L., Stridsberg, M., Isaksson, S., Jerlstad, P., and Simren, M. Altered levels of fecal chromogranins and secretogranins in IBS: relevance for pathophysiology and symptoms? *The American Journal of Gastroenterology*, 107(3): 440–447, 2012.
- Pearson, K. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901.
- Pedersen, H. K., Gudmundsdottir, V., Nielsen, H. B., Hyötyläinen, T., Nielsen, T., Jensen, B. A. H., Forslund, K., Hildebrand, F., Prifti, E., Falony, G., Le Chatelier, E., Levenez, F., Dore, J., Mattila, I., Plichta, D. R., Pöhö, P., Hellgren, L. I., Arumugam, M., Sunagawa, S., Vieira-Silva, S., Jorgensen, T., Holm, J. B., Trost, K., Consortium, M., Kristiansen, K., Brix, S., Raes, J., Wang, J., Hansen, T., Bork, P., Brunak, S., Oresic, M., Ehrlich, S. D., and Pedersen, O. Human gut microbes impact host serum metabolome and insulin sensitivity. *Nature*, advance online publication, Jul 2016.
- Peters, J. M., Colavin, A., Shi, H., Czarny, T. L., Larson, M. H., Wong, S., Hawkins, J. S., Lu, C. H., Koo, B.-M., Marta, E., et al. A comprehensive, crispr-based functional analysis of essential genes in bacteria. *Cell*, 165(6): 1493–1506, 2016.
- Pinheiro, J. and Bates, D. *Mixed-effects models in S and S-PLUS*. Springer Science & Business Media, 2006.
- Pociot, F. and Lernmark, A. Genetic risk factors for type 1 diabetes. *Lancet*, 387 (10035):2331–2339, 2016.

- Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K. S., Manichanh, C., Nielsen, T., Pons, N., Levenez, F., Yamada, T., Mende, D. R., Li, J., Xu, J., Li, S., Li, D., Cao, J., Wang, B., Liang, H., Zheng, H., Xie, Y., Tap, J., Lepage, P., Bertalan, M., Batto, J.-M., Hansen, T., Le Paslier, D., Linneberg, A., Nielsen, H. B., Pelletier, E., Renault, P., Sicheritz-Ponten, T., Turner, K., Zhu, H., Yu, C., Li, S., Jian, M., Zhou, Y., Li, Y., Zhang, X., Li, S., Qin, N., Yang, H., Wang, J., Brunak, S., Dore, J., Guarner, F., Kristiansen, K., Pedersen, O., Parkhill, J., Weissenbach, J., Bork, P., Ehrlich, S. D., and Wang, J. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*, 464(7285): 59–65, 2010.
- Rajilić-Stojanović, M., Heilig, H. G., Tims, S., Zoetendal, E. G., and Vos, W. M. Long-term monitoring of the human intestinal microbiota composition. *Environmental microbiology*, 15(4):1146–1159, 2013.
- Rewers, M. and Ludvigsson, J. Environmental risk factors for type 1 diabetes. *Lancet*, 387(10035):2340–2348, 2016.
- Rohlke, F. and Stollman, N. Fecal microbiota transplantation in relapsing *Clostridium difficile* infection. *Therapeutic Advances in Gastroenterology*, 5(6):403–420, 2012.
- Rossen, N. G., Fuentes, S., van der Spek, M. J., Tijssen, J. G., Hartman, J. H., Duflo, A., Löwenberg, M., van den Brink, G. R., Mathus-Vliegen, E. M., de Vos, W. M., Zoetendal, E. G., D’Haens, G. R., and Ponsioen, C. Y. Findings from a randomized controlled trial of fecal transplantation for patients with Ulcerative Colitis. *Gastroenterology*, 149(1):110–118.e4, 2015.
- Sai, P. and Rivereau, A. S. Prevention of diabetes in the nonobese diabetic mouse by oral immunological treatments. comparative efficiency of human insulin and two bacterial antigens, lipopolysaccharide from *Escherichia coli* and glycoprotein extract from *Klebsiella pneumoniae*. *Diabetes & Metabolism*, 22(5): 341–348, 1996.
- Sanchez, E., De Palma, G., Capilla, A., Nova, E., Pozo, T., Castillejo, G., Varea, V., Marcos, A., Garrote, J. A., Polanco, I., Lopez, A., Ribes-Koninckx, C., Garcia-Novo, M. D., Calvo, C., Ortigosa, L., Palau, F., and Sanz, Y. Influence of environmental and genetic factors linked to celiac disease risk on infant gut colonization by *Bacteroides* species. *Applied and Environmental Microbiology*, 77(15):5316–5323, 2011.
- Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., Lesniewski, R. A., Oakley, B. B., Parks, D. H., Robinson, C. J., et al. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and environmental microbiology*, 75(23):7537–7541, 2009.
- Schnorr, S. L., Candela, M., Rampelli, S., Centanni, M., Consolandi, C., Basaglia, G., Turrone, S., Biagi, E., Peano, C., Severgnini, M., et al. Gut microbiome of the Hadza hunter-gatherers. *Nature Communications*, 5, 2014.
- Scholz, M., Ward, D. V., Pasolli, E., Tolio, T., Zolfo, M., Asnicar, F., Truong, D. T., Tett, A., Morrow, A. L., and Segata, N. Strain-level microbial epidemiology and population genomics from shotgun metagenomics. *Nature Methods*, 13(5): 435–438, 2016.

- Sciola, V., Massironi, S., Conte, D., Caprioli, F., Ferrero, S., Ciafardini, C., Peracchi, M., Bardella, M. T., and Piodi, L. Plasma chromogranin A in patients with inflammatory bowel disease. *Inflammatory Bowel Diseases*, 15(6):867–871, 2009.
- Segata, N., Waldron, L., Ballarini, A., Narasimhan, V., Jousson, O., and Huttenhower, C. Metagenomic microbial community profiling using unique clade-specific marker genes. *Nature Methods*, 9(8):811–814, 2012.
- Segata, N., Bornigen, D., Morgan, X. C., and Huttenhower, C. PhyloPhlAn is a new method for improved phylogenetic and taxonomic placement of microbes. *Nature Communications*, 4:2304, 2013.
- Seiskari, T., Kondrashova, A., Viskari, H., Kaila, M., Haapala, A. M., Aittoniemi, J., Virta, M., Hurme, M., Uibo, R., Knip, M., Hyoty, H., and group, E. s. Allergic sensitization and microbial load—a comparison between Finland and Russian Karelia. *Clinical & Experimental Immunology*, 148(1):47–52, 2007.
- Sela, D. A. and Mills, D. A. Nursing our microbiota: molecular linkages between bifidobacteria and milk oligosaccharides. *Trends in Microbiology*, 18(7):298–307, 2010.
- Sela, D. A., Chapman, J., Adeuya, A., Kim, J. H., Chen, F., Whitehead, T. R., Lapidus, A., Rokhsar, D. S., Lebrilla, C. B., German, J. B., Price, N. P., Richardson, P. M., and Mills, D. A. The genome sequence of *Bifidobacterium longum* subsp. *infantis* reveals adaptations for milk utilization within the infant microbiome. *Proceedings of the National Academy of Sciences*, 105(48):18964–18969, 2008.
- Sender, R., Fuchs, S., and Milo, R. Are we really vastly outnumbered? revisiting the ratio of bacterial to host cells in humans. *Cell*, 164(3):337–340, 2016.
- Shannon, C. E. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423, 1948.
- Simonyte Sjodin, K., Vidman, L., Ryden, P., and West, C. E. Emerging evidence of the role of gut microbiota in the development of allergic diseases. *Curr Opin Allergy Clin Immunol*, 2016.
- Skrondal, A. and Rabe-Hesketh, S. *Generalized latent variable modeling: Multi-level, longitudinal, and structural equation models*. Crc Press, 2004.
- Soergel, D. A., Dey, N., Knight, R., and Brenner, S. E. Selection of primers for optimal taxonomic classification of environmental 16S rRNA gene sequences. *ISME Journal*, 6(7):1440–1444, 2012.
- Sokal, R. and Rohlf, F. *Biometry*. W.H. Freeman and Co., New York, 1995.
- Sollid, L. M. and Thorsby, E. HLA susceptibility genes in celiac disease: genetic mapping and role in pathogenesis. *Gastroenterology*, 105(3):910–922, 1993.
- Sonnenburg, E. D., Smits, S. A., Tikhonov, M., Higinbottom, S. K., Wingreen, N. S., and Sonnenburg, J. L. Diet-induced extinctions in the gut microbiota compound over generations. *Nature*, 529(7585):212–215, 2016.
- Sonnenburg, J. L. and Bäckhed, F. Diet-microbiota interactions as moderators of human metabolism. *Nature*, 535(7610):56–64, 2016.

- Stefka, A. T., Feehley, T., Tripathi, P., Qiu, J., McCoy, K., Mazmanian, S. K., Tjota, M. Y., Seo, G. Y., Cao, S., Theriault, B. R., Antonopoulos, D. A., Zhou, L., Chang, E. B., Fu, Y. X., and Nagler, C. R. Commensal bacteria protect against food allergen sensitization. *Proceedings of the National Academy of Sciences*, 111(36):13145–13150, 2014.
- Strachan, D. P. Hay fever, hygiene, and household size. *British Medical Journal*, 299(6710):1259, 1989.
- Suau, A., Bonnet, R., Sutren, M., Godon, J.-J., Gibson, G. R., Collins, M. D., and Doré, J. Direct analysis of genes encoding 16s rRNA from complex communities reveals many novel molecular species within the human gut. *Applied and environmental microbiology*, 65(11):4799–4807, 1999.
- Sunagawa, S., Mende, D. R., Zeller, G., Izquierdo-Carrasco, F., Berger, S. A., Kultima, J. R., Coelho, L. P., Arumugam, M., Tap, J., Nielsen, H. B., Rasmussen, S., Brunak, S., Pedersen, O., Guarner, F., de Vos, W. M., Wang, J., Li, J., Dore, J., Ehrlich, S. D., Stamatakis, A., and Bork, P. Metagenomic species profiling using universal phylogenetic marker genes. *Nature Methods*, 10(12):1196–1199, 2013.
- Suzek, B. E., Wang, Y., Huang, H., McGarvey, P. B., Wu, C. H., and UniProt, C. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics*, 31(6):926–932, 2015.
- Tamburini, S., Shen, N., Wu, H. C., and Clemente, J. C. The microbiome in early life: implications for health outcomes. *Nature Medicine*, 22(7):713–722, 2016.
- Tannock, G. W. Molecular assessment of intestinal microflora. *The American journal of clinical nutrition*, 73(2):410s–414s, 2001.
- Teeaar, T., Liivak, N., Heilman, K., Kool, P., Sor, R., Paal, M., Einberg, U., and Tillmann, V. Increasing incidence of childhood-onset type 1 diabetes mellitus among Estonian children in 1999-2006. time trend analysis 1983-2006. *Pediatric Diabetes*, 11(2):107–110, 2010.
- Thaiss, C. A., Zmora, N., Levy, M., and Elinav, E. The microbiome and innate immunity. *Nature*, 535(7610):65–74, 2016.
- Tickle, T. L., Segata, N., Waldron, L., Weingart, U., and Huttenhower, C. Two-stage microbial community experimental design. *ISME Journal*, 7(12):2330–2339, 2013.
- Tigchelaar, E. F., Zhernakova, A., Dekens, J. A., Hermes, G., Baranska, A., Mujagic, Z., Swertz, M. A., Munoz, A. M., Deelen, P., Cenit, M. C., Franke, L., Scholtens, S., Stolk, R. P., Wijmenga, C., and Feskens, E. J. Cohort profile: LifeLines DEEP, a prospective, general population cohort study in the northern Netherlands: study design and baseline characteristics. *BMJ Open*, 5(8):e006772, 2015.
- Truong, D. T., Franzosa, E. A., Tickle, T. L., Scholz, M., Weingart, G., Pasolli, E., Tett, A., Huttenhower, C., and Segata, N. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nature Methods*, 12(10):902–903, 2015.
- Turnbaugh, P. J., Ridaura, V. K., Faith, J. J., Rey, F. E., Knight, R., and Gordon, J. I. The effect of diet on the human gut microbiome: a metagenomic analysis

- in humanized gnotobiotic mice. *Science Translational Medicine*, 1(6):6ra14, 2009.
- Utgoff, P. E. Incremental induction of decision trees. *Machine learning*, 4(2): 161–186, 1989.
- Uusitalo, U., Liu, X., Yang, J., Aronsson, C. A., Hummel, S., Butterworth, M., Lernmark, Å., Rewers, M., Hagopian, W., She, J.-X., et al. Association of early exposure of probiotics and islet autoimmunity in the TEDDY study. *JAMA Pediatrics*, 170(1):20–28, 2016.
- van der Maaten, L. Accelerating t-SNE using tree-based algorithms. *Journal of Machine Learning Research*, 15:3221–3245, 2014.
- Van der Maaten, L. and Hinton, G. Visualizing non-metric similarities in multiple maps. *Machine learning*, 87(1):33–55, 2012.
- van der Maaten, L. and Hinton, G. Visualizing high-dimensional data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, Nov 2008.
- von Hertzen, L., Beutler, B., Bienenstock, J., Blaser, M., Cani, P. D., Eriksson, J., Farkkila, M., Haahtela, T., Hanski, I., Jenmalm, M. C., Kere, J., Knip, M., Kontula, K., Koskenvuo, M., Ling, C., Mandrup-Poulsen, T., von Mutius, E., Makela, M. J., Paunio, T., Pershagen, G., Renz, H., Rook, G., Saarela, M., Vaarala, O., Veldhoen, M., and de Vos, W. M. Helsinki alert of biodiversity and health. *Annals of Medicine*, 47(3):218–225, 2015.
- Voor, T., Julge, K., Bottcher, M. F., Jenmalm, M. C., Duchon, K., and Bjorksten, B. Atopic sensitization and atopic dermatitis in Estonian and Swedish infants. *Clinical & Experimental Allergy*, 35(2):153–159, 2005.
- Waters, J. L. and Salyers, A. A. Regulation of CTnDOT conjugative transfer is a complex and highly coordinated series of events. *mBio*, 4(6):e00569–13, 2013.
- Watson, D. W. and Kim, Y. B. Modification of host responses to bacterial endotoxins. I. specificity of pyrogenic tolerance and the role of hypersensitivity in pyrogenicity, lethality, and skin reactivity. *Journal of Experimental Medicine*, 118:425–446, 1963.
- Whitfield, C. and Trent, M. S. Biosynthesis and export of bacterial lipopolysaccharides. *Annual Review of Biochemistry*, 83:99–128, 2014.
- Wooldridge, J. M. *Econometric analysis of cross section and panel data*. MIT press, 2010.
- Wu, G. D., Chen, J., Hoffmann, C., Bittinger, K., Chen, Y. Y., Keilbaugh, S. A., Bewtra, M., Knights, D., Walters, W. A., Knight, R., Sinha, R., Gilroy, E., Gupta, K., Baldassano, R., Nessel, L., Li, H., Bushman, F. D., and Lewis, J. D. Linking long-term dietary patterns with gut microbial enterotypes. *Science*, 334(6052):105–108, 2011.
- Xu, J., Bjursell, M. K., Himrod, J., Deng, S., Carmichael, L. K., Chiang, H. C., Hooper, L. V., and Gordon, J. I. A genomic view of the human-Bacteroides thetaiotaomicron symbiosis. *Science*, 299(5615):2074–2076, 2003.

- Yatsunenko, T., Rey, F. E., Manary, M. J., Trehan, I., Dominguez-Bello, M. G., Contreras, M., Magris, M., Hidalgo, G., Baldassano, R. N., Anokhin, A. P., Heath, A. C., Warner, B., Reeder, J., Kuczynski, J., Caporaso, J. G., Lozupone, C. A., Lauber, C., Clemente, J. C., Knights, D., Knight, R., and Gordon, J. I. Human gut microbiome viewed across age and geography. *Nature*, 486(7402): 222–227, 2012.
- Zhou, X., Kao, M. C., and Wong, W. H. Transitive functional annotation by shortest-path analysis of gene expression data. *Proceedings of the National Academy of Sciences*, 99(20):12783–12788, 2002.



ISBN 978-952-60-7314-9 (printed)
ISBN 978-952-60-7313-2 (pdf)
ISSN-L 1799-4934
ISSN 1799-4934 (printed)
ISSN 1799-4942 (pdf)

Aalto University
School of Science
Department of Computer Science
www.aalto.fi

**BUSINESS +
ECONOMY**

**ART +
DESIGN +
ARCHITECTURE**

**SCIENCE +
TECHNOLOGY**

CROSSOVER

**DOCTORAL
DISSERTATIONS**