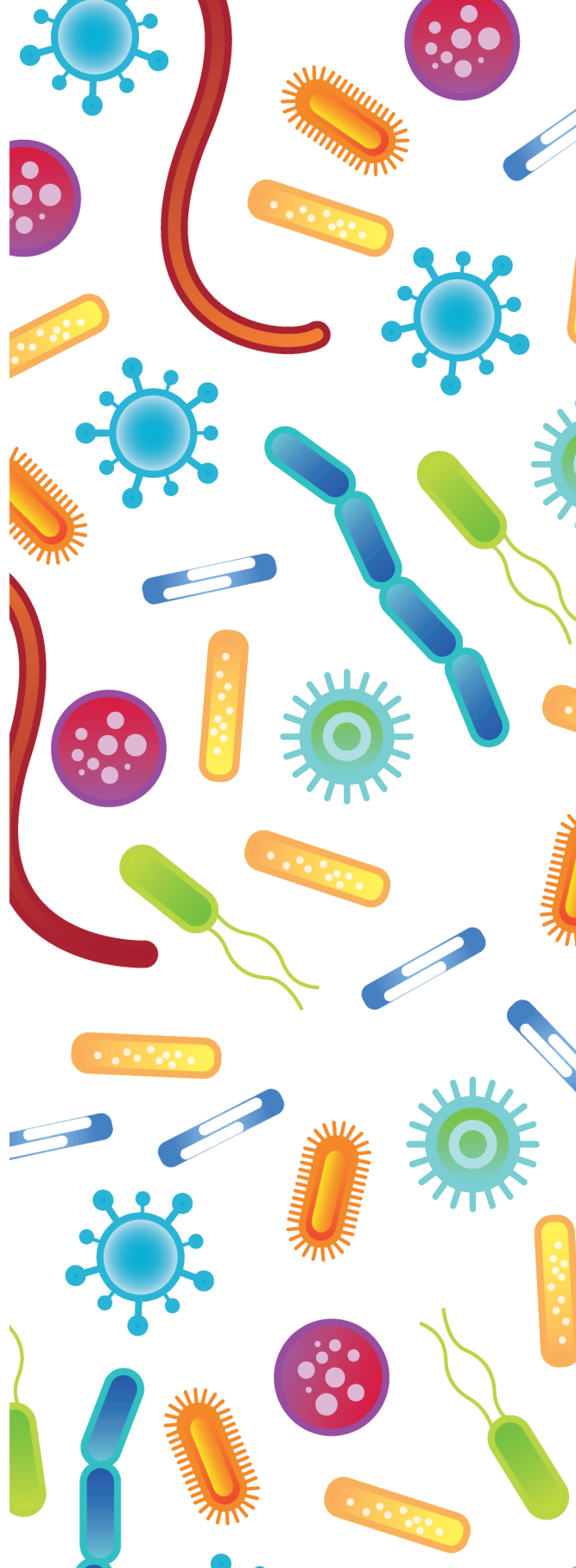


ON THE ROLE OF HOST MICROBIOTA ASSOCIATED WITH HEALTH AND DISEASE

A METAGENOMICS SURVEY
THROUGH MARKER-GENE
SEQUENCING AND
GENOMICS APPROACHES

THOMAS H.A. EDERVEEN



On the role of host microbiota associated with health and disease

a metagenomics survey through marker-gene sequencing and
genomics approaches

Thomas H.A. Ederveen

Colofon

The research presented in this thesis was conducted at the Bacterial Genomics group of the Center for Molecular and Biomolecular Informatics, and the Department of Dermatology, both of the Radboud university medical center (Radboudumc), Nijmegen, The Netherlands.

Printing of this thesis was financially supported by the Radboudumc.

ISBN

978-94-93118-09-6

Thesis design and cover

Koen Ederveen

Thesis layout

Thomas Ederveen

Print

Ipskamp Printing Nijmegen

© T. Ederveen, 2019

On the role of host microbiota associated with health and disease

a metagenomics survey through marker-gene sequencing and
genomics approaches

Proefschrift

ter verkrijging van de graad van doctor
aan de Radboud Universiteit Nijmegen
op gezag van de rector magnificus prof. dr. J.H.J.M. van Krieken,
volgens besluit van het college van decanen
in het openbaar te verdedigen op

donderdag 28 maart 2019
om 16.30 uur precies

door

Thomas Hendrikus Antonius Ederveen
geboren op 13 juni 1987
te Nijmegen

Promotor

prof. dr. J. Schalkwijk

Copromotoren

dr. S.A.F.T. van Hijum (NIZO food research, Ede)

dr. J. Boekhorst (NIZO food research, Ede)

dr. P.L.J.M. Zeeuwen

Manuscriptcommissie

prof. dr. H.F.L. Wertheim (voorzitter)

prof. dr. G. Muijzer (Universiteit van Amsterdam)

dr. B.E. Dutilh (Universiteit Utrecht)

"Looking for that one constant floating in a sea of variables "

- adapted from Mr. Robot (TV series 2015)

Table of Contents

INTRODUCTION

Glossary	8
Chapter 1. General introduction and scope	12
Thesis summarizing outline	37

PART I

Bioinformatics Tools and Analyses

Chapter 2. COMPANION: Reduce manual curation by combining gene predictions from multiple annotation engines: <i>a case study of start codon prediction.</i>	42
Chapter 3. TaxPhlAn: A generic workflow for Single Locus Sequence Typing (SLST) design and subspecies characterization of microbiota: <i>a case-study of beyond species-level profiling of Staphylococcus in atopic dermatitis.</i>	62

PART II

The Skin Microbiome

Chapter 4. An <i>in vitro</i> model for bacterial growth on human stratum corneum.	86
Chapter 5. Gram-positive anaerobe cocci (GPAC) are underrepresented in the microbiome of filaggrin-deficient human skin.	102

PART III

The Gut Microbiome

Chapter 6. Aberrant intestinal microbiota due to IL-1 receptor antagonist deficiency promotes IL-17- and TLR4-dependent arthritis.	128
Chapter 7. The gut microbiome in ADHD and its relation to neural reward anticipation.	150

PART IV

The Nasal Microbiome

Chapter 8. <i>Haemophilus</i> is overrepresented in the nasopharynx of infants hospitalized with RSV infection and associated with increased viral load and enhanced mucosal CXCL8 responses.	172
--	-----

CONCLUSIONS

Chapter 9. General conclusion and discussion	192
Future perspective	199
Final considerations	201

Chapter 10. List of references	202
Chapter 11. Nederlandse Samenvatting	222

ADDENDUM

About the author	230
Dankwoord	232
Research data stewardship and accessibility (FAIR)	243
List of publications	245
List of abbreviations	246
notes	

Glossary

Common terms universal to this thesis are explained in this section, in order to streamline terminology. Full list of abbreviations can be found in the back of this thesis.

-omics	This suffix refers to a particular field of study in biology involving large-scale characterization and quantification.
16S	The 16S ribosomal RNA (rRNA) molecule of bacteria and archaea (prokaryotes). Term usually refers to the gene encoding this molecule, a marker gene which is used for taxonomic placement of prokaryotes. see <i>marker gene</i> . see <i>metataxonomics</i> .
AD	Abbreviation for atopic dermatitis. AD (eczema) is a type of inflammatory skin disease (dermatitis) that is characterized by red, itchy, lesional skin, and often associated by skin colonization by the bacterium <i>Staphylococcus aureus</i> . see <i>FLG</i> .
AMP	Abbreviation for antimicrobial protein / peptide. AMP are expressed by the host innate immune system in order to combat pathogens and infections.
antibiotics	Antimicrobial drugs used to treat and prevent bacterial infections. see <i>biotics</i> .
biome	A community of organisms, including all of its biotic components and abiotic factors. see <i>microbiome</i> .
biotics	The broad definition of molecules, (viable) microorganisms, or components thereof, which through modulating microbiota have a beneficial effect on the host. see <i>antibiotics</i> . see <i>postbiotics</i> . see <i>prebiotics</i> . see <i>probiotics</i> .
community	see <i>biome</i> . see <i>microbiome</i> . see <i>niche</i> .
FLG	Abbreviation for filaggrin. FLG is a protein that surrounds the structural keratin filaments, and plays a vital role in the structure and hydration of the skin. Its breakdown products are an important source for NMF. Genetic mutations in the FLG gene are a strong predisposition for development of skin disease. see <i>AD</i> . see <i>IV</i> . see <i>NMF</i> .
genome	The collection of genes in an organism.
GI	Abbreviation for gastrointestinal (tract).
homologue	Genes with shared ancestry in the evolutionary history of life. Note, high gene sequence similarity does not necessarily mean that genes are homologous. see <i>orthologue</i> . see <i>parologue</i> .
<i>Incertae sedis</i>	Term used for a taxonomic group where its broader relationships are unknown or undefined. Meaning "of uncertain placement". see <i>taxonomy</i> .
ITS	Abbreviation of the Internal Transcribed Spacer gene of fungi (eukaryotes). Term usually refers to the marker gene which is used for taxonomic placement of fungi. see <i>marker gene</i> . see <i>metataxonomics</i> .
IV	Abbreviation for ichthyosis vulgaris. Skin disease that is characterized by a very dry, scaly skin. Etiology largely involves a strong association with mutations in the FLG gene, encoding a protein which is an important source for NMF. see <i>FLG</i> . see <i>NMF</i> .
marker gene	Genetic marker for screening and phylogenetic placement of organisms. Typically these are groups of orthologous genes which share a common phylogenetic origin. see <i>16S</i> . see <i>ITS</i> . see <i>metataxonomics</i> . see <i>MGS</i> . see <i>orthologue</i> . see <i>phylogenetics</i> . see <i>SLST</i> .

meta-	This prefix is used to indicate a concept which is an abstraction behind another concept. In this thesis, it usually indicates the combined microbiota potential for a certain concept, in contrast to only the host.
metabolome	The metabolite profile(s) in any given organism, cell, or collection of cells (i.e. tissue, multicellular organism, etc.) at one point in time.
metabolomics	The study of metabolome(s).
metagenome	The collection of genomes and genes from the members of a microbiome. Typically obtained through WGS. see <i>genome</i> .
metagenomics	The study of metagenome(s).
metaproteomics	The collection of proteomes of microbiota. see <i>proteome</i> .
metataxonomics	The high-throughput process of identification, classification and naming of (groups of) organisms, usually relating to simultaneously characterizing all microbiota of the same microbiome. Term usually reserved for 16S marker gene sequencing. see <i>marker gene</i> . see <i>taxonomy</i> .
metatranscriptome	The collection of transcriptomes of microbiota. see <i>transcriptome</i> .
metatranscriptomics	The study of metatranscriptome(s).
MGS	Thesis abbreviation for marker gene sequencing. This method is commonly used for taxonomic placement of organisms. see <i>marker gene</i> . see <i>metataxonomics</i> . see <i>sequencing</i> .
microbe	see <i>microorganism</i> .
microbiome	The entire habitat, comprising its microorganisms (microbiota, as outlined above, but including viruses, viroids, phages, etc.), their genomes (i.e., genes), their derived components (i.e. waste products, available nutrition, metabolites, etc.), and the surrounding environmental abiotic conditions (i.e. temperature, humidity, acidity / pH, etc.). see <i>biome</i> .
microbiomics	The study of microbiome(s).
microbiota	All microorganisms (microbes) present in a specific environment, location and time.
microorganism	Mostly unicellular organisms that commonly relate to bacteria and archaea (prokaryotes) or fungi (eukaryotes), but can also comprise lower eukaryotes such as protists (e.g. amoebae) or chlorophytes (e.g. algae) and many more.
MLST	Abbreviation for multi-locus sequence typing. This method is used for taxonomic placement of organisms based on multiple loci. see <i>marker gene</i> . see <i>metataxonomics</i> .
NGS	Abbreviation for next-generation sequencing. Techniques that rely on high-throughput, massively parallel and deep sequencing of DNA samples. Usually referring to Roche 454, Illumina, SOLiD, Ion Torrent (2nd generation) and the more recent Nanopore (3rd generation). see <i>MGS</i> . see <i>sequencing</i> . see <i>shotgun</i> . see <i>WGS</i> .
niche	The environment where organisms live in. see <i>biome</i> . see <i>community</i> . see <i>microbiome</i> .
NMF	Abbreviation for natural moisturising factors. NMF are small molecules that are key in keeping the skin hydrated. see <i>FLG</i> . see <i>IV</i> .
orthologue	Homologous genes in different organisms (genomes) that originated by vertical descent from a single gene of the last common ancestor. see <i>homologue</i> .
OTU	Abbreviation for Operational Taxonomic Unit, used to classify groups of closely related microorganism individuals. see <i>metataxonomics</i> . see <i>MGS</i> . see <i>phylogenetics</i> .

paralogue	Homologous genes in the same organism (genome) that originated from the same ancestral gene by gene duplication events. see <i>homologue</i> .
phylogenetics	The study of evolutionary history and relationships among individuals or groups of organisms.
phylogeny	The result of the study of phylogenetics, such as represented in a phylogenetic tree. see <i>phylogenetics</i> .
postbiotics	Compounds such as metabolites of (cultures of) microorganisms which are considered beneficial for the host. Postbiotics no longer contain viable microorganisms. see <i>biotics</i> .
prebiotics	Food ingredients that induce the growth or activity of microorganisms which are considered beneficial for the host. see <i>biotics</i> .
probiotics	Microorganisms which are considered beneficial for the host. see <i>biotics</i> .
proteome	The protein profile(s) in any given microorganism, cell, or collection of cells (i.e. tissue, multicellular organism, etc.) at one point in time.
proteomics	The study of proteome(s).
Sanger	Classic chain-termination sequencing method which is primarily suitable for single, homogenous DNA fragments. Referred to as first-generation sequencing. see <i>NGS</i> . see <i>sequencing</i> .
sequencing	The process of determining the order of bases in a fragment of DNA such as in a gene, PCR amplicon or chromosomal fragment. see <i>NGS</i> . see <i>Sanger</i> .
shotgun	Popular method for whole-genome sequencing (WGS) of organisms. Term usually reserved for sequencing a metagenome. Typically, shotgun allows for sequencing the entire genetic content of a microbiome. see <i>sequencing</i> . see <i>WGS</i> .
SLST	Thesis abbreviation for single-locus sequence typing. This method is used for taxonomic placement of organisms based on a single locus. see <i>marker gene</i> . see <i>metataxonomics</i> .
taxonomy	Study of the identification, classification and naming of organisms.
transcriptome	The collection of DNA transcripts (mRNA) of one microorganism, cell, or collection of cells (i.e. tissue, multicellular organism, etc.) at one point in time. Term usually refers to the host in relation to its microbiome, such as a human individual, in which case the transcriptome can be derived from multiple types of tissues or organs.
transcriptomics	The study of transcriptome(s).
WGS	Common abbreviation for whole-genome sequencing. Term usually reserved for sequencing one particular organism. see <i>sequencing</i> . see <i>shotgun</i> .

CHAPTER 1

GENERAL INTRODUCTION AND SCOPE

General introduction and scope

1. Microorganisms are omnipresent.

Terminology of the unseen.

Consortia of microbiota are found in many niches of the earth, such as on various sites of animals and plants, in soil, in water or even in the atmosphere, but also in industrial fermentations and biofilms. Microbiota are defined as mostly unicellular organisms that commonly relate to bacteria and archaea (prokaryotes) or fungi (eukaryotes), but also comprises other lower eukaryotes such as protists (e.g. amoebae) or chlorophytes (e.g. algae) and many more. The microbiome definition was recently restated and proposed to be, as partly adopted from Marchesi *et al.*, 2015: *The entire habitat, comprising its microorganisms, their genomes, their derived components and the surrounding environmental abiotic conditions* [1]. See the [Glossary](#) for a more detailed description of 'microbiome' and other related and thesis-relevant vocabulary. The relevance of bacteria for humans becomes clear by only looking at the numbers, as bacteria colonize our body sites with a ratio of 1:1 bacterium to human cell (in contrast to the widely-cited 10:1 ratio, that was recently revised) [2]. Even if they would not evoke an effect on bodily processes – but which many of them do – their sheer numbers in members, genes, but also their variation in comparison to their host, make them interesting subjects of research.

Exploring microbiome interactions.

Microbiota are in closely regulated interaction with their environment; and vice versa: characteristics of the niche they thrive in greatly influence presence and absence of microbial entities. This is for example nicely illustrated for human skin, where different skin sites bring about niche-specific communities of microbes, but where mechanical disruption of the skin barrier results in temporary change in microbial composition up until recovery of the skin [3]. Microbial diversity is considerable, and the current challenge lies in determining which microbes and (corresponding) functionality is of importance to a given ecological niche. Furthermore, as there is increasing evidence of microbial involvement in health and disease, the need arises to fundamentally understand microbiome processes for application in healthcare, nutrition and personal care products (e.g. cosmetics, probiotics, hygiene instruction and education). This thesis is an in-depth exploration of these complex interactions between microbiota, their host (if applicable) and environment in different niches from a wide range of different sources: from mouse and dog animal models, to relevant gut, skin and airway studies in humans.

2. Microbiota and their societal relevance in biotechnology and nutrition.

Microbes in medicine and biotechnology.

In the medical and biotechnological sector, microbes are cultivated and adopted for industrial production of a wide range of compounds such as proteins, enzymes and metabolites. One example of such success in these sectors is the adaptation and application of *Escherichia coli* as small cell factories for production of bioactive compounds such as recombinant human insulin for treatment of diabetes [4], or

manufacturing of riboflavin (vitamin B₂) for supplementation in animal feed [5]. Alternatively, harmless microorganisms can be adopted for production of pathogen-specific compounds for medical use, thereby reducing risk for undesired harmful compounds such as toxins or virulence factors. An example is the case where a safe *Lactococcus lactis* bacterial strain is genetically modified to produce a recombinant proteolytically inactive, but immunoactive form of the virulent *Staphylococcus aureus* protein serine surface protease HtrA for application in vaccines [6].

Microbes in health and nutrition.

In the health and nutrition sector, microbes are typically applied in fermentation processes that in general can be distinguished in ethanol or lactic acid fermentation by yeast or bacteria, respectively. The well-known baker's or brewer's yeast *Saccharomyces cerevisiae* is by far the most popular microbe for production of bread and alcoholic beverages. Other examples of (wild) yeast species commonly used for food fermentations are *Candida* (*C. zemplinina*) and *Hanseniaspora* (*H. uvarum*) which are involved in wine making processes, albeit usually dominant in an earlier stage before succeeded by the more alcohol tolerant *S. cerevisiae* [7]. In the Japanese kitchen the fungus *Aspergillus* (*A. oryzae*) is traditionally used for the production of a number of popular and fermented foods such as rice wine, soy sauce and miso [8]. On the other side of the fermentation spectrum there is a wide range of lactic acid bacteria (LAB) such as *Lactobacillus* (primary examples are *L. casei*, *L. helveticus*, *L. plantarum* and *L. reuteri*) and *Lactococcus* (*L. lactis*), but also *Leuconostoc* (*L. mesenteroides*) [9] and *Streptococcus* (*S. thermophilus*) [10] which are used for pickling of foods such as kimchi, sauerkraut, chow-chow and many others. Alternatively, these LAB are used for application as starter cultures in dairy products such as cheese, yoghurt, kefir and many others. Apart from the apparent benefit of these microbes, they can in some cases also contribute to food spoilage, for instance, *L. brevis* is found in food products such as sauerkraut, and it is one of the most common species involved in beer spoilage [11, 12]. Hence, great care must be taken by food providers in order to discriminate between desirable and undesirable microbes, for instance through careful genetic monitoring of fermentation processes in order to predict and prevent spoilage [13]. In pursue of this importance, in this thesis I present and adopt a multitude of techniques and bioinformatics tools and workflows that allow for discriminating between a plethora of microbes-of-interest.

3. Studying microbiota composition and dynamics – how and why?

Microscopy and culture-based analysis.

Being able to define the exact composition and abundance of microbial communities is of great relevance when studying the role of microbiota in a microbiome, and consequently, its impact on the environment such as that of the human host [14]. Studying microbiota dynamics in a defined habitat of interest, for example during health and disease in man, requires a comprehensive set of microscopy-, culture- or DNA amplification-based tools (Table 1). However, in practice, microscopic examination of cell morphology for classification purposes does not allow for high resolution identification of microbiota. More importantly, the vast majority of (human host-associated) microbiota are yet

uncultivable in the laboratory, thereby hindering differentiation of isolated colonies by chemical- and cell-biological approaches such as cell staining (e.g. Gram or Ziehl-Neelsen [15]) or compound conversion enzyme assays (e.g. metabolism traits such as starch hydrolysis or H₂S production [16, 17]), respectively.

Table 1. Overview of current sequencing-based and alternative methods for microbiota identification and classification, including their (dis)advantages.

Legend as follows: *cost*: estimation of method (commercial) prices; *return time*: rough indication of time until results are available; *data analysis*: complexity of data analysis; *phenotyping*: methods that allow for accurate phenotyping; *genomic potential*: amount of information with regard to a microbe its genomic (functional) potential; *composition*: how well a method is able to characterize the (full) microbial taxonomic composition of a sample; *resolution*: accuracy of taxonomic resolution by which a microbe can be identified (from phylum to strain). The symbols represent pie charts, open versus closed pie chart meaning, for *cost*: low vs. high; for *return time*: fast vs. slow; for *data analysis*: simple vs. complex; for *phenotyping*: inapplicable vs. applicable; for *genomic potential*: low vs. high; for *composition*: unsuitable vs. suitable; for *resolution*: low vs. high.

		cost	return time	data analysis	pheno-typing	genomic potential	composition	resolution
altern.	Culturing							
	RFLP / GTG5 / AFLP							
	PCR							
sequencing	PCR combined with Sanger							
	Genome sequencing							
	16S marker gene seq.							
	SLST marker seq.							
	Shotgun metagenomics							

Analysis based on DNA amplification and sequencing.

With regard to DNA-based microbiota analysis tools, depending on their technique and application, PCR- and sequencing-by-synthesis-based methods allow for identifying microbes in high-throughput screening assays, with high sensitivity and specificity, great in-depth resolution and at fairly low costs. Historically, (microbial) DNA sequences of maximally 150 kbp such as small PCR amplicons, full genes or large chromosomal chunks were sequenced by inserting them into a known bacterial artificial chromosome (BAC) plasmid, in a process called clone-by-clone sequencing. These plasmids with DNA insert were introduced into competent bacteria, amplified by these organisms to produce many identical copies, and split into even smaller, overlapping fragments of 500 bp with a more manageable size for sequencing. Subsequently, these smaller fragments were cloned into vectors and sequenced by techniques such as Sanger, a first-generation chain-termination sequencing method. Nowadays, next-generation sequencing (NGS) advances allow for high-throughput, massively parallel and deep sequencing of DNA samples, thereby dismissing the need for vector-based cloning of sequences. Most notable examples are the second-generation Roche 454, Illumina, SOLiD, Ion Torrent, and the more recent third-generation Nanopore. NGS gave an enormous boost to the field of genomics, microbiomics and bioinformatics, amongst others; mainly due to its substantially reduced sequencing costs and ultra-high-throughput application. Arguably, NGS is one of the technological innovations that strongly contributed to bioinformatics becoming the full-grown scientific community it is today.

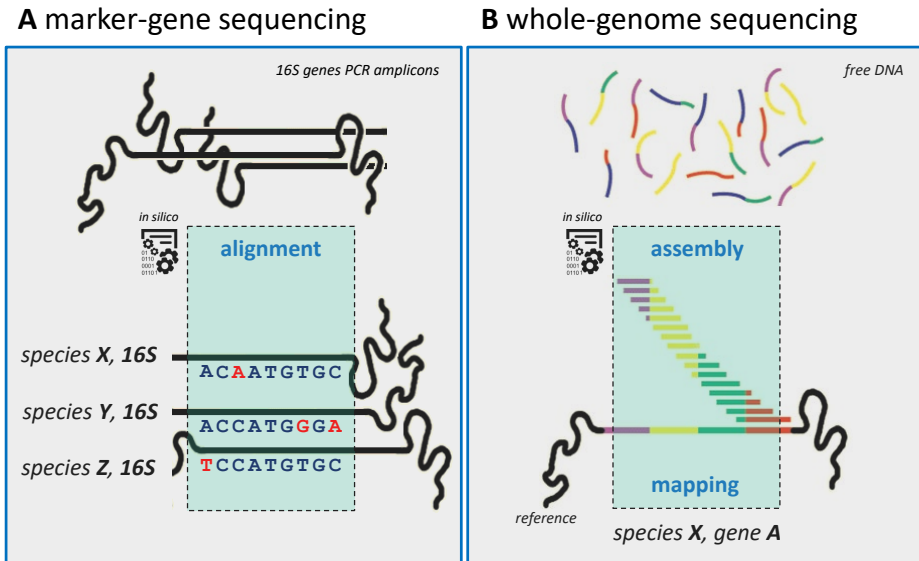


Figure 1. Principal microbiota sequencing approaches.

We roughly distinguish two main applications for next-generation sequencing of microbial communities. These are marker-gene sequencing (MGS) for metataxonomics, and whole-genome sequencing (WGS) for metagenomics, and are not to be confused with sequencing platforms or techniques. In MGS a single gene is sequenced, typically a marker gene such as 16S, ITS, or a specific SLST target. In this MGS example (A), 16S is selected as marker gene, which is extracted from a mixed microbial population by PCR (not shown), and sequenced by NGS techniques. After MGS sequencing, reads, typically much longer than for WGS (± 500 bp), are aligned, and based on informative positional differences in the 16S gene known reference microbiota can be assigned, or novel taxonomies can be inferred. With WGS one can extract genomic potential and function information, in contrast to MGS, with which one can only extract taxonomic information. In this WGS example (B), typically small sequences (100–150 bp), deriving from the full genomic content of a mixed microbial population, are assembled into genes of all microbiota present. Or even into full genomes, depending on microbial complexity and sequencing depth.

Bacterial consortia are currently mostly analysed either by marker gene sequencing (MGS) metataxonomics, or by whole-genome sequencing (WGS) metagenomics (Figure 1). Metagenomics is a major topic in this thesis, and is defined as the study of the collection of genomes and genes from the members of a microbiome [1]. Metagenomics is a term reserved primarily for WGS and will in this thesis be referred to as shotgun, in short. It is not to be confused with MGS initiatives such as bacterial 16S or its fungal counterpart ITS, which are better described as metataxonomics: the high-throughput identification, classification and naming of microbiota. The metagenome of the gut greatly outmatches the coding capacity of the human genome with over 3.3 million versus 20,000 genes [18]. This is also the case for the RNA derivative of the metagenome: the metatranscriptome, counterpart of the human transcriptome. In later chapters I will discuss in-depth the applications and limitations of NGS.

DNA fingerprinting techniques and chemotaxonomy for microbial typing.

Another well-established classification method for microbial strain-, phylo- or genotyping is multi-locus strain typing (MLST), which is based on sequence variety in a number of marker core genes, typically on the species- or even strain level. MLST is conventionally applied PCR-based with subsequent Sanger sequencing or on single strain isolate full genome assemblies. Alternatively, a set of single nucleotide polymorphism (SNP) specific primers can be used. Recently, its concept has been adopted for high-resolution bacterial classification in next-generation sequencing efforts by initiatives such as PathoScope [19] and PhyloPhlAn [20] that provide sets of representative marker genes and algorithms for phylogenetic inference of sequencing reads. Likewise, metagenomics analysis methods such as AMPHORA, ConStrains, MetaPhlAn and StrainPhlAn effectively apply in high-throughput this concept of single nucleotide variant patterns in marker genes (these methods will in more detail be discussed in section 1.4 of this introduction).

Alternatively, chemotaxonomy methods based on strain-specific cell components (proteins, sugars and lipids, mainly) are available by various forms of mass spectrometry such as MALDI, electrospray ionization or gas chromatography, typically combined with a time-of-flight detector. Mass spectrometry-based typing methods are highly strain specific, but mass-to-charge spectrometry peak profile data analysis is challenging (especially high-throughput) and in its current form is mainly feasible for single-isolates and hardly on complex mixed cultures, which is a severe limitation for application in microbiomics field.

Furthermore, DNA fingerprinting techniques remain popular for typing microbiota. These commonly applied alternatives to sequencing-based methods are cost- and labor efficient. Mainly, pulsed field gel electrophoresis (PFGE) [21], multiple-locus variable-number tandem repeat analysis (VNTR), restriction fragment length polymorphism (RFLP), denaturing gradient gel electrophoresis (DGGE), restriction enzyme-directed ribotyping [22], and PCR fingerprinting by microsatellite-specific oligonucleotides such as (GTG)₅ [23]. Finally, taxonomy- and strain-specific (q)PCR profiling, based on shared core genome and unique pan-genome targets, respectively, remains one of the most popular bacterial typing methods. With (q)PCR, microbial identification can be performed typically strain-specific, or on different taxonomic levels such as species, genus, family, etc., or on a particular clade-of-interest. Alternatively, one can use fluorescently labelled oligonucleotide probes (FISH) [24] for microbial identification instead of detection by (q)PCR, by hybridizing complementary genome-specific target sequences.

Single locus sequence typing for bacterial strain tracking.

Recently, single locus sequence typing (SLST) has been described for determining down-to strain level identification of *Propionibacterium acnes* [25]. It is a technique analogous to 16S rRNA marker gene sequencing (MGS) that allows cost-effective profiling of specific bacteria up-to and beyond the species level. The difference with 16S is that with SLST a clade-specific, single-copy gene target is chosen that allows for discrimination between members (typically species, but also sub-species) of the microbial clade-of-

interest. Representative SLST studies are limited in number, most notable examples are by van Bokhorst-van de Veen and co-workers who applied the concept in order to study gastrointestinal (GI) track survival of a mixture of ten *Lactobacillus plantarum* strains *in vivo* in human volunteers [26]. Furthermore, Fernández-Ramírez and colleagues applied SLST in order to monitor interaction and dynamics of a mixture of 12 *L. Plantarum* strains in *in vitro* biofilm formation competition experiments (manuscript in preparation, [27]).

Global microbiome consortia and initiatives.

Initiatives for studying microbiomes are great in number, most notably the Human Microbiome Project (HMP) [28, 29] by U.S. National Institute of Health (NIH), focussing on 47 different human host-associated microbial niches. Furthermore, gut-associated microbiota initiatives are numerous with amongst others the Human Gut Microbiome Initiative (HGMI) [30], American Gut (a crowdfunded project) [31], and the European project METAGENOMICS of the Human Intestinal Tract (MetaHIT consortium) [18, 32]. Other initiatives focus more on environmental microbiota, such as the Earth Microbiome Project (EMP) [33], International Census of Marine Microbes (ICoMM) [34] and TerraGenome [35].

4. Next-generation sequencing of microbiota – applications and their limitations.

Marker gene sequencing.

16S rRNA marker gene sequencing (MGS) focuses on the 16S rRNA genes present in all prokaryotes and archaea (16S, in short). It is relatively cheap and data analysis is straight-forward. In theory, the 16S rRNA gene can be targeted by universal PCR primers, and the technique does therefore not require bacterial reference genomes for analysis. However, for classifying 16S sequencing reads – that is, assigning reads to their corresponding bacterium-of-origin – prior knowledge in the form of 16S rRNA gene databases with corresponding taxonomy information is required. Most notable, the Ribosomal Database Project (RDP) [36], Greengenes [37] and SILVA [38] are well-established examples of such databases. 16S allows for confidently profiling bacteria down-to the genus level (Table 1). The process of analysing 16S sequencing reads, in short, often involves clustering of sequencing reads in operational taxonomic unit (OTU), where after they are classified (Figure 1A). This so called OTU-picking can be performed in different modes: closed, using reference 16S reads only; *de novo*, allowing finding novel biodiversity without using a reference database; or open, a combination of both. In practice, for 16S, there is a delicate trade-off between taxonomic sensitivity and resolution in terms of microbiota classification. In other words, in order to identify as many microbial entities as possible one loses microbial resolution in terms of taxonomic discriminative power (classification); and vice versa: when focussing with higher resolution on a particular microbiota niche, one is likely to lose information on all microbes outside the niche-of-interest. We call this the *sensitivity-to-classification* problem. Hence, choosing the right 16S primers is crucial (Figure 2A). The 16S rRNA gene has multiple alternating conserved and variable (V1 to V8) regions, with a total length of roughly 1.5 kbp (Figure 2B) [39]. For application in sequencing, the longer

the 16S sequencing read the more confident microbial classification can be performed. Therefore, depending on the sequencing platform, one best selects those V regions that maximize potential of its sequencing platform with regard to read length. Furthermore, every 16S primer (combination) has its pros and cons when it comes to sensitivity and specificity to different bacterial families (genera) [40]. Primer choice also depends on the target niche-of-interest, such as for gut, skin or oral samples, which therefore is motivated by bacteria typically present on those sites [41].

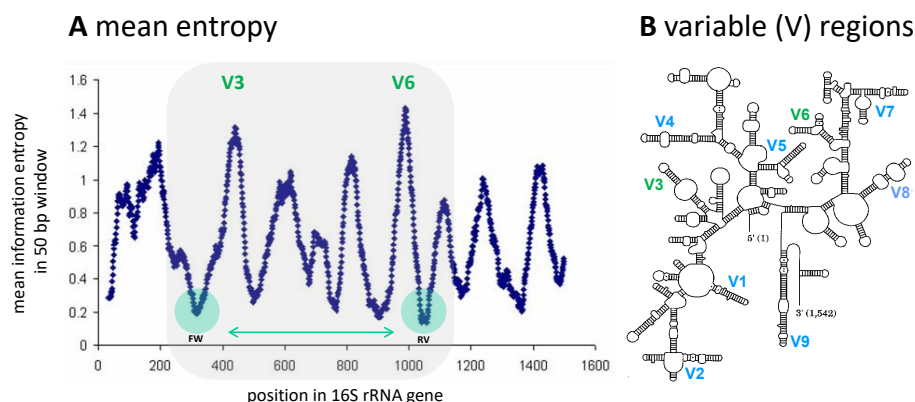


Figure 2. 16S rRNA marker gene characteristics.

The 16S rRNA gene in bacteria is widely used for metataxonomics. Between different clades of phylogenetically-related bacteria, this gene varies strongly in terms of conservation and variation, as shown in the left panel. (A) Visualization of the mean information entropy for each position of the 16S gene (± 1.5 kbp in length), based on all known 16S genes present in the Ribosomal Database Project (RDP). One can see peaks (strong variation) and valleys (strong conservation) in different regions of the 16S gene sequence, regions which can be used for taxonomic discrimination and primer design, respectively. In this example, we observe 9 peaks (variable regions), of which V3 and V6 show the largest peaks and deepest valleys. Therefore, the gene region from V3 up-to V6 is very suitable for primer design and marker gene sequencing (this thesis). (B) The 9 different variable regions, structurally visualized over the full length of the 16S gene. The locations of V3 and V6 are indicated in green.

Many different methods and bioinformatics tools are available to analyse MGS data, most notably operational taxonomic unit-based (OTU) methods like QIIME (v2) (Quantitative Insights Into Microbial Ecology [42-44]), based on the UCLUST OTU clustering algorithm [45], and Mothur [46], based on average linkage OTU clustering. Likewise, analogously to the bacterial 16S rRNA gene, the fungal rRNA gene has two ITS (internal transcribed spacer) variable regions (ITS₁ to ITS₂) that can be adopted for universal fungal MGS [47]. The same MGS data analysis tools used for 16S can be applied to ITS datasets. Fungal-application, however, requires ITS-specific databases such as UNITE, containing fungal ITS sequences and annotation [48]; these data can relatively straight-forward be incorporated in conventional 16S data analysis pipelines.

Metagenomics sequencing.

Shotgun metagenomics, on the other hand, does not suffer from the aforementioned *sensitivity-to-specificity* problem as it sequences all (free) DNA present in a sample (Figure 1B). Only certain types of samples, such as skin swabs, necessitate DNA

amplification during library preparation because of typical low DNA yield [3]; however, in these cases all DNA is (in theory) equally amplified with random oligonucleotide primers. The major advantage of shotgun over MGS is that it provides insight into gene and metabolic function potential of a sample based on its metagenome. This allows for pathway analysis and even mining of for instance virulence factors, antibiotic resistance, (pathogen) lineage-specific markers, or novel enzymes for catalysis of reactions in pharmaceutical-, food-, or industrial processes which may currently be expensive or difficult [49, 50]. Alternatively, complete bacterial genomes can be extracted from shotgun samples if the metagenomics data is assembled, provided the genomes are present in high enough numbers to achieve a minimum feasible reads-to-genome coverage to allow for confidential assembly: ideally 20x, to assemble genomes of species with 1% relative abundance [51]. Or one can use a combination of assembly and assembly-free methods by first assembling genomes, and thereafter mapping reads to these new, study-specific reference genomes. Furthermore, apart from bacteria and fungi, shotgun enables retrieval of DNA sequences from viruses, bacteriophages, (lower) eukaryotes, and even host cells. The last can be problematic in terms of contamination, which for example is a notorious problem in cutaneous microbiome studies as skin has a low microbe-to-host ratio. This could result in few bacteria and a large number of dead host cells that are easily taken along upon skin sampling [3]. Hence, the 'correct' way of microbial DNA sampling is of importance to yield high-quality sequencing data with minimal host DNA contamination; in this example, to minimize mechanical disruption of the skin such as by swabbing, in contrast to the conventional scraping alternative.

Unfortunately, shotgun is relatively expensive, very resource exhaustive both for personnel and computation (Table 1), and should ideally be applied only when there is a specific hypothesis to be tested: e.g. *is particular (new) functionality such as antibiotic resistance or virulence factors present and is it linked to certain bacteria?* Typically, shotgun metagenomics dataset resolution is down-to the class / family level, but recent advances in methodology allow for classification to the level of genus / species with tools such as MetaPhlAn (v2) [52], MEGAN (v6) [53-55], AMPHORA (v2) [56, 57] and mOTUs [58]; or even down-to sub-species level with tools such as ConStrains [59], PanPhlAn [60] and StrainPhlAn [61]. However, these methods are confident only for (sub-)species with an overall relative abundance of one percent or above. In other words, with shotgun it is in principle possible to determine the presence of entities close to strain-level, but currently this is time consuming and not trivial to deduce because of the complex methods and typically large datasets involved.

For pathway investigation and analysis of potential functions of microbiota, various bioinformatics tools exist and can be adopted, most notably, MG-RAST [62], and HUMAnN (v2) (includes ChocoPhlAn) [63]; although some of the aforementioned tools for microbial classification also offer to some extent a function analysis. Noteworthy, methods such as PICRUSt [64], PanFP [65], and Tax4Fun [66] are cost-effective alternatives to shotgun in order to yield information about microbial function from 16S sequencing data. It deduces the presence of bacteria based on 16S reads, and infers from this information the total metabolic potential of a microbiome with a taxon-to-function reference database. Albeit conceivably inferior to shotgun metagenomics, it

does provide a coarse-grained overview of the metabolic capacity of a microbiome. In opposite fashion, one is able to 'mine' 16S rRNA or other marker genes from shotgun data, which could be interesting for a quick screen, or when there is specific interest in one particular phylogenetic clade or (associated) function.

Microbial strains are the most specific source of metabolic / regulatory potential [67], and follow-up experiments are done straight-forward with (combinations of) microbial strains. Hence, pinpointing the right candidate strains from studies with access to shotgun data is crucial. Unfortunately, application of shotgun sequencing is still very expensive, which limits its application in practice. Consequently, we observe that large shotgun data sets are only rarely being published, and the number of study initiatives that can afford an attainable number of shotgun samples are still very limited.

5. Microbiome research in health and disease.

The microbiota associated with a host niche are typically niche-specific. Example studies where imbalance or perturbation in microbial composition or diversity at these sites is indicative for local or systemic disease, are growing in number. Unfortunately, the majority of such examples / cases with putative correlations are associative and not yet causally proven. However, guided by Koch's postulates, they do provide valuable leads for hypothesis building and follow-up research. Furthermore, although correlations are not necessarily causal in nature, they can increase understanding of biology if they are replicated in different studies. Three different microbial niches in humans are focus of study in this thesis, and will briefly be introduced here: skin, gut and nasal cavity. Other important microbial niches that are subject of in-depth research by the scientific community are the oral cavity [68], female genital tract [69] and the airways [70]. The (species) diversity of a microbiome is defined as the combination of species richness, i.e. number of different species represented in a community, and evenness, i.e. how closely related species in a community are. Furthermore, the (species) diversity can be described as microbiota differences within an individual or sample (inter-individual: alpha diversity) or between individuals or samples (intra-individual: beta diversity) [71].

The intestinal microbiome.

The most studied microbiome of the human host is unquestionably the gut, consisting mainly of the small intestine and large intestine (colon). The small intestine usually implies the ileum, but also consists of the more proximal duodenum and jejunum. The gut furthermore comprises the stomach, cecum and rectum. It is estimated that more than a thousand species-level bacterial phylotypes can be found in the GI tract of the human population, albeit that diversity at an individual level is much lower with approximately 160 different species [18]. Gut microbiota play an important role not only in nutritional and metabolic processes, but also in physiological and immunological processes of the human body [72].

The healthy stomach (gastric microbiome) is richer in microbiota than one would expect given its acidic environment and fast flow-through of content, with an estimate of 10^3 - 10^4 bacterial cells / mL stomach content. The gastric microbiome is primarily

characterized by species of *Prevotella*, *Streptococcus*, *Veillonella* and *Rothia*, but more can be identified. Furthermore, composition at the genus level is rather dynamic, likely because it is easily affected by many factors such as diet and medication use [73]. The small intestine (SI), directly after the stomach, is the primary site where nutrients are extracted from the diet. The microbial concentration in the ileum is higher than in the more proximal parts of the SI resulting from a drop in acidity and reabsorption of bile acids [72]. The total density of bacteria in the SI (10^8 cells / mL) is much lower than in the colon (10^{11} cells / mL) [2]; in fact, the vast majority of human-associated microbiota reside in the colon (estimated 10^{13} - 10^{14} cells in total).

Interestingly, the SI is also the more immunological active site of the intestine. Especially in the ileum and distal parts of the jejunum consisting of gut-associated lymphoid tissue (GALT) that contain many Peyer's patches organized into follicles [74]. These lymphoid aggregates, counting roughly 100 to 200 in humans, bring about mixed cell population of mononuclear origin such as lymphocytes, dendritic cells and macrophages, and is an important location for immunoglobulin IgA-producing B cell maturation [74, 75]. IgA secreted from these Peyer's patches has a major contribution to regulation of intestinal homeostasis and protecting the gut barrier integrity [76]. Furthermore, it is thought that Peyer's patches have a pivotal role in immune tolerance or response by interaction with intestinal (commensal or pathogenic) bacteria and other host cells such as (follicle-associated) epithelium [74]. This crosstalk is amongst others regulated through innate pattern recognition receptors such as Toll-like receptors, NOD-like receptors and C-type lectin receptors which are able to sense a wide range of microbial antigens like lipopolysaccharide, peptidoglycan and mannose, respectively [77].

Besides Peyer's patches and epithelial cells, other important cell types involved in cell-host-microbe interactions of the mucosal interface are Paneth cells, found in the crypts of the SI. These cells produce antimicrobial peptides (AMP) upon contact with microorganisms, such as HIP/PAP (better known as RegIII, its counterpart in mice), which binds peptidoglycan of Gram-positive bacteria, thereby permeabilizing the bacterial membrane and killing them. Or ANG4, a bactericidal RNase that is exclusively secreted by Paneth cells [78]. Epithelial cells are also capable of producing some types of AMP such as human β -defensin-3 (hBD-3) with a broad-spectrum of antimicrobial activity against both Gram-negative bacteria (e.g. strains of *Escherichia coli*, *Pseudomonas aeruginosa*) and Gram-positive bacteria (e.g. *Streptococcus pyogenes*, *Enterococcus faecium*), and even fungi like *Candida albicans* [79]. Likewise, defensins of the alpha type (α -defensin) are produced by Paneth cells and have comparable specificity and mode-of-action to β -defensins secreted by epithelial cells [78]. Interestingly, it is thought that some AMP induction by innate immune cells can act as a possible mechanism to alert epithelial cells to the loss of mucosal integrity [77].

One more final important component of the gut mucosal immune system is the lamina propria (LP): a thin layer of connective tissue located under the epithelial cells and present throughout the GI (and respiratory) tract that gives rise to specialised LP lymphocytes [75]. The CD4+ T helper cells Th17 (producing the inflammatory IL-17 interleukin, and IL-22) and Treg (regulatory T cell producing the anti-inflammatory IL-

10) are most abundant in gut-associated tissue, particularly the small intestinal LP [80]. The delicate balance between pro-inflammatory Th17 and immunosuppressive Treg has been demonstrated to be vital in autoimmune disorders such as rheumatoid arthritis (this thesis, [Chapter 6](#)), inflammatory bowel diseases and psoriasis [81]. It is therefore fascinating that intestinal microbiota drive Th1, Th17 and Treg development in the gut [82]. Notable examples are segmented filamentous bacteria (SFB) which are able to induce strong Th17 responses in the gut [80], or reversely, the bacterium *Bacteroides fragilis* that expresses the capsule molecule polysaccharide A (PSA) and which induces Treg responses [83].

The cecum lumen, connects the ileum and colon, and harbours large numbers of aerobic bacteria in contrast to other parts of the gut [84], who are thought to aid in the breakdown of dietary fibres from plant material such as cellulose [85]. The cecal microbiome has not attracted much attention so far, as it is very small in humans in comparison to other animals, primarily in comparison to herbivores. Whereas the stomach and first parts of the SI are aerobic, oxygen levels decrease gradually towards the distal end of the gut. The colon is generally anaerobic, and the pH is near neutral. Consequently, different bacterial species reside in the SI compared to the colon. The Bacteroidetes and Firmicutes phyla account for 90% of the bacteria in the colon, with Gram-negative *Bacteroides* sp. and Gram-positive Clostridia and Baccili as their most notable representatives [77]. Furthermore, the colon contains low numbers of *Bifidobacterium* (phylum Actinobacteria) and Enterobacteriaceae (phylum Proteobacteria) with primarily *Escherichia* sp. and its close relative *Shigella* sp. The SI, on the other hand, generally contains more members belonging to Lactobacillales or Proteobacteria [77].

It is important to realise that in humans, the SI and colon are not easy to reach for study purposes without invasive intervention. For this reason, in practice, many studies use the fecal microbiota as a proxy for intestinal microbiota. It should be noted that feces does not necessarily correspond to ileal or colonic microbial profiles so care should be taken in interpreting and translating the relevance of such results [86]. For experimental animal models this is less of an issue as animals may be dissected during or after the study, and consequently more is known about rodent microbiota residing in the direct intestine compartments in comparison to human. Exceptions in humans are patients with for example ileo- or colostoma, but these do not necessary represent healthy individuals. Alternatives for sampling the human immune-relevant ileum are underway, for example the IntelliCap system [87] that is currently being adapted for ileal or colonic sampling rather than drug delivery.

The biological concept of colonization resistance describes that transient microbes have little means of colonizing the host as long as the (commensal) microbiome is present, occupying the host mucosal membrane or skin and effectively inhibiting colonization, overgrowth and infection by (pathogenic) intruders. This can be done directly, for instance by releasing bacteriocins, by metabolic exclusion, or by more effectively competing for nutrients. Alternatively, this can be done accomplished in an indirect manner, for instance by alarming the host immune system to produce AMP

[88]. Metabolic exclusion is a strategy of the host commensals to create conditions within the niche that can reduce virulence gene expression or even inhibit the growth of invading pathogens. Whenever a niche-opening is created as an effect of disturbance or perturbations of the local commensals by for example antibiotics, diet, tissue damage or loss of immune function, then other microbes at that time present in the environment have a chance to nestle themselves into the community [88]. Reversely, pathogens are able to exploit antibiotics or virulence factors to trigger inflammation in order to disrupt host microbial homeostasis to create niche-openings themselves. If these newly introduced transient microbes are pathogenic in nature this can cause health problems for the host. Notorious examples are enteric infections with pathogenic strains of genera such as *Campylobacter*, *Clostridium*, *Escherichia*, *Salmonella* and *Shigella*. These bacteria are able to breach intestinal barrier function, leading to leakage of luminal content into the LP, with consequential inflammation, diarrhea, and loss of absorption of nutrients [89]. Conventional treatment of enteric infection primarily exists of broad-spectrum antibiotics, but population-level antibiotic resistance of gut microbiota is an enormous threat to human health [90], demanding treatment alternatives on the long run. The fecal transplantation of gut microbiota from healthy volunteers to patients is a relatively novel and highly effective alternative to fight enteric pathogens such as *Clostridium difficile*, for which an impressive 89% effectiveness in preventing recurrent infection has been shown after one treatment only [91]. More practical applications for fecal transplantation are currently underway for treatment of a wide range of disorders such as inflammatory bowel disease, metabolic syndromes and obesity, even neurological conditions due to dysfunction of gut-brain-axis (this thesis), which are all suggested to be linked to dysbiosis of the GI microbiota [92, 93].

The nasopharyngeal microbiome.

Strictly speaking, the nasopharynx is not part of the nasal cavity, but situates posterior to the nasal cavity where the nose blends into the throat. Nevertheless, the nasal cavity and upper airways including the (naso)pharynx are all covered with epithelial mucosa and an underlying lamina propria similar to the earlier discussed GI tract [94]. The nasal- or nasopharynx-associated (mucosa) lymphoid tissue (NALT), comparable to GALT in the gut, is an essential part of the (naso)pharyngeal immunological system, acting a physical barrier protecting the host from viral and bacterial infections [95]. The adenoids (nasopharyngeal tonsils) consist of NALT and constitute the major part of Waldeyer's tonsillar ring in humans, which is an anatomical collective term for the ring-like arrangement of lymphoid tissue in the pharynx [96]. Immunoglobulin IgA plays a vital role as a first line of defence in the host immune response, secreted by IgA-committed B cells that have undergone isotype class switching in the NALT [94]. The five major bacterial constituents of the nasopharyngeal microbiome in human are species of *Staphylococcus*, *Haemophilus*, *Streptococcus*, *Corynebacterium* and *Moraxella*, and in young children its composition is highly variable between seasons [97, 98]. Common bacterial pathogens in the upper airways, especially in children, are *Streptococcus pneumoniae* (pneumococcus), *Haemophilus influenzae*, *Moraxella catarrhalis*, and *Staphylococcus aureus* [99]. Interestingly, these bacteria can also be present as commensals in healthy individuals, raising the yet unresolved question why some microbes can be asymptotically present in one person, and be disease-causing in

another. Viral infections are also commonly found in the respiratory tract, most notably rhinovirus, influenza, coronavirus, adenovirus and respiratory syncytial virus (RSV). Interestingly, there are many reports on interactions between viruses and bacteria in the pathogenesis of respiratory infections, commonly synergising host-pathogen infection processes [100]. One textbook example is the synergism between influenza virus and *S. pneumonia*, where secondary bacterial infection following influenza leads to a significant increase in the mortality rate because of pneumococcal pneumonia. That is what happened during the Spanish flue pandemic in 1918-1919 [101]. Several mechanisms have been proposed by which bacterium-virus interactions can lead to increased secondary infection, such as by altering host epithelial immune responses upon viral infection, thereby upregulating host adhesion proteins and increasing the chance of attachment by bacterial pathogens to the mucosal surfaces [100]. Another mode-of-action is by impairment of components of the immune system by viruses, for example by disruption of epithelial barrier function [102], or by dysfunction or enhanced apoptosis of neutrophils [103]. It is also proposed that specific bacteria can increase susceptibility to a consecutive viral infection (this thesis), such as *H. influenza* which is able to increase expression of surface proteins ICAM-1 and TLR-3 on airway epithelial cells, thereby providing an enhanced chance of binding to these cells for rhinovirus particles [104]. In addition, some viruses can also decrease susceptibility to (secondary) bacterial infection, such as by stimulating the host to produce more AMP such as defensins [100].

The cutaneous microbiome.

Simply put, the human skin is a physical barrier of the body that has one main purpose: to keep the inside in, and the outside out. In addition, the skin functions as an immunological barrier with processes comparable to those as previously described for the gut and nasopharynx: microbial colonization resistance by the skin microbiome, and host immune sensing and surveillance [105]. These immune attributes allow for modulation of skin commensals, killing invading pathogens, and preventing undesirable microbes from infecting the skin when the host acquires skin damage or when cutaneous homeostasis is lost otherwise. The skin is anatomically comprised of several layers with different cells and properties. The uppermost cornified layer (*stratum corneum*) of the epidermis is where most of the skin microorganism reside, and the deeper into the *stratum corneum* the less microbes are present [3]. The microbial density on the skin surface is very low compared to the large intestine, but comparable to quantities found in the small intestine, with estimates of 10^{11} cells in total [2]. The skin layers further down: the *stratum granulosum*, *spinosum* and *basale*, and the lower dermis are mostly considered sterile in healthy / intact skin. With exception of sweat- and sebaceous glands and hair follicles, which are also colonized by microbiota, but lay deeper into the skin [106].

Microbial make-up of skin niches is highly dependent on characteristics based on skin type and location. We distinguish three main physiological skin sites: (i) oily / sebaceous skin sites such as the forehead, upper back and behind the ear; (ii) dry skin sites such as forearm and lower buttocks; and (iii) moist skin sites such as the arm pits, nostrils and the groin. But also acidity (pH) and temperature of the microenvironments are

important drivers for microbial inhabitants. Despite site-to-site compositional variation, common skin commensals typically found on humans are the genera *Corynebacterium*, *Propionibacterium*, *Staphylococcus*, *Micrococcus*, *Actinomyces*, *Streptococcus*, *Prevotella*, amongst others [106-108]. Even though inter-individual variation between healthy volunteers is high, microbial communities are largely stable over time despite exposure of the skin to the external environment [109, 110].

The epidermis largely (>90%) consists of keratinocytes, an epithelial cell type with barrier and host defence functions. The *stratum basale* is the germinative layer from which the cells migrate to the skin surface in approximately 20 days. When cells exit the *stratum basale* they start to differentiate and form the *stratum spinosum*, thereby altering their phenotype and starting to express a distinct set of genes that include early markers of differentiation such as cytokeratin 1 and 10, and involucrin. Further on, the *stratum granulosum* is the last living cell layer, comprised of flattened cells with a granular appearance and distinct gene expression program [111]. Keratinocytes of the *stratum granulosum* enter a program called terminal differentiation, which includes the expression of a large number of genes involved in skin barrier function and host defence. Ultimately, cells excrete large amounts of lipids, form a cross-linked cell envelope and lose their nuclei. These enucleated cornified envelopes form the *stratum corneum*, which is the actual physical skin barrier [112]. Corneocytes contain many natural moisturising factors (NMF), small molecules derived from the breakdown of proteins such as filaggrin. NMF which include small peptides, amino acids and breakdown products derived thereof are key to keep the skin hydrated [113]. Both the *stratum corneum* and the differentiated cell layers of the *stratum granulosum* and *stratum spinosum* are essential to control microbial invasion, by providing a physical barrier and an antimicrobial protein shield. Notorious skin bacteria able to cause (chronic) cutaneous infection are *Staphylococcus aureus*, *Pseudomonas aeruginosa* and *Streptococcus pyogenes*, who can express virulence factors and inhibit wound healing [114, 115]. However, also fungi can cause skin infections, such as *Candida albicans*, certain species of *Aspergillus*, and a group of fungi collectively called dermatophytes (family *Arthrodermataceae*) [116]. Notably, all fungi are potentially disease causing organisms, even the common commensal *Malassezia spp.*, which are the most abundantly occurring fungi on healthy (and diseased) human skin [116-118]. Keratinocytes express many AMP in order to control skin microbiota colonization and infection, comparable to those secreted by gut epithelial cells. Notable example are the *E. coli*-specific psoriasin (S100A7), human β -defensins (also expressed in the gut), with hBD-2 targeting Gram-negative bacteria such as *E. coli* and *P. aeruginosa*, and the earlier discussed hBD-3 [119]. Furthermore, the staphylocidal factor RNase-7, which in addition acts as a broad-spectrum antimicrobial protein against other pathogens such as *E. faecium*, *P. aeruginosa*, *Propionibacterium acnes*, and the fungus *C. albicans* [120]. Other important skin AMP are calprotectin (complex of S100A8/S100A9), the cathelicidin LL37 (CAMP) and lysozyme [78, 119].

Another important cell type in the skin is the Langerhans cell (LC). These are dendritic antigen-presenting immune cells present throughout the skin layers, but mainly in the *stratum spinosum*. Local microbial infection or other changes in homeostasis attracts LC to sites of infection, where they can recognize pathogen-associated molecular patterns

(PAMPs) and take-up microbial antigens [121]. Hereafter, LC travel to regional lymph nodes where they train T helper lymphocytes for subsequent immune responses.

Interestingly, some skin diseases are marked by an imbalance in microbiota in comparison to healthy volunteers, notable examples are atopic dermatitis (AD) and ichthyosis vulgaris (IV). Patients with AD show a strong increase in *Staphylococcus* in their lesional skin [111, 122], whereas IV is marked by a decrease in Gram-positive anaerobe cocci of genera *Finegoldia*, *Anaerococcus*, and *Peptoniphilus* (this thesis). For AD we know that disruption of skin barrier function might be a favorable condition for *S. aureus* in order to adhere to cell surface proteins of corneocytes in the *stratum corneum*, and that *S. aureus* is capable of expressing potent virulence factors leading to infection of deeper skin tissues [111], but for other skin diseases this is less clear. What exactly causes these microbial imbalances, if this accounts for other skin disorders, and whether these relations between microbiota and disease phenotype are causal or consequential remains to be resolved.

6. Modulation of microbiota composition through diet or biotics.

It is clear by now that microbiota are involved in many host processes, either in health and disease. Furthermore, it is widely recognized that diet and lifestyle are important factors for maintaining a healthy, diverse and resilient (gut) microbiome, and consequently to maintain a state favourable for health [123]. Hence, current research focuses on modulation of human-associated microbiota to thereby influence host processes, to the benefit of the host, which is arguably the next step. Human intervention studies with diet or dietary compounds are likely the most direct and effective way of modulating (gut) microbiota. Examples are dietary polyphenols, present in a wide range of plant foods such as wine and black tea [124], short chain fatty acids such as butyrate [125], resistant starch [126], and a broad range of presumed functional foods, many of which simply are rich in dietary fibers or work as antibiotics [127].

Therapeutic application of bacteria or their derivatives

The use of biotics to change microbiome composition and function, and thereby indirectly manipulating host processes, is an emerging and promising field. Probiotics are live (sometimes attenuated) bacteria, whereas postbiotics are non-viable bacterial products or metabolites in order to induce biological effects on the host [128]. Likewise, prebiotics are food ingredients that induce the growth or activity of microorganisms which are considered beneficial for the host, help in nutrient absorption or have (indirect) immunomodulatory effects [129]. Widely studied prebiotics are fermentable dietary fibers such as human milk oligosaccharides (HMO), galacto- (GOS) and fructooligosaccharides (FOS), but also other polysaccharides such as inulin or xylan [130]. Finally, use of antibiotics and the earlier discussed application of fecal transplantation are effective ways to manipulate (gut) microbiota (this thesis). Likewise, the prospect of skin microbiota transplantation is finding its way into research applications (this thesis) [131]. With regard to skin microbiomics, recent research focuses on pre- and probiotics for topical application [132]. Despite the wide focus on pre- and probiotics, examples of convincing and highly replicated biotics studies in humans are scarce, if not, yet unavailable. Nevertheless, a noteworthy example is the recent study in African

infants by Lacroix and co-workers, who show that supplementing GOS to the diet mitigates the adverse effect that iron fortification for prevention of anaemia has on the gut microbiome [125, 133].

7. Visualization and statistical analysis of biomedical study data.

Arguably, an important key step in data analysis is visualization. The most straightforward and classical way to make visualizations is by performing heat map analysis (clustering). Or simply by plotting relative abundances over samples for various taxa or OTU in stacked bar graphs or pie-charts, such as for metagenomics data (e.g. [134]). More recently, novel and both appealing and powerful methods for representation of datasets were developed, interesting examples are by Sundquist *et al.*, [135], Cytoscape software (www.cytoscape.org) [136], and iTOL: interactive tree of life (<http://itol.embl.de>) [137]. Visualization is an effective tool / analysis approach, that can point in the direction of leads such as associations, candidate biomarkers, etc., which can thereafter be further corroborated with conventional statistics. In addition, visualization can provide a straight-forward means for sample quality control by identifying misbehaviour of sample characteristics, such as by (mis)clustering of erroneous samples.

Nowadays, a plethora of different statistical methods and approaches for the analysis of biomedical data is available (Table 2). We distinguish several basic concepts in statistics that are shortly discussed below. First, parametric statistics deals with methods that rely on its data belonging to a particular distribution. Oppositely, when this is not the case, when sample size is low, or when this information is unknown, one has to apply nonparametric alternative statistics (i.e. distribution-free methods) [138]. Biomedical data is generally considered nonparametric, or if unknown is wise to treat as such. Furthermore, descriptive statistics are terms to quantitatively describe or summarize (variables in) a sample or dataset (e.g. average, median, standard deviation, etc.). Inferential statistics, however, deals with extending from what is known about the sample data to what might be true for the total population; for instance, to calculate the probability of an observed effect happening by chance or being dependent on secondary characteristics of the data [138]. Hence, most statistical tests are inferential in practice. We further distinguish supervised and unsupervised statistics [139]. Simply put, supervised means that statistical algorithms infer information from labeled data, whereas with unsupervised statistics unlabeled or 'naive' data is used, the latter which therefore can be considered a more 'unbiased' approach. Examples of supervised statistics are redundancy analysis (RDA) [140, 141], random forest [142], and as applied in Bayesian networks [143] or hidden Markov models (HMM) [144]. Supervised learning is an important hallmark of machine learning, as applied in artificial neural networks and deep learning algorithms. Oppositely, unsupervised statistics are for example methods such as principal component analysis (PCA), and various clustering algorithms such as hierarchical or *k*-means clustering [145, 146].

Table 2. Overview of cross-sectional (A/B) and longitudinal (B/C) data analysis methods (both uni- and multivariate) that are commonly applied in life sciences. Listed are advantages and limitations for each statistical method, and examples of typical research questions.

Method (2A)	Advantages	Limitations	Example of typical research question	Type	Ref.
Student's <i>t</i> -test (non-parametric: Mann-Whitney <i>U</i> test)	<ul style="list-style-type: none"> - Quick and very accessible method - Very powerful for screening of many (simple) contrasts 	<ul style="list-style-type: none"> - Relies on parametric models; use a non-parametric Mann-Whitney <i>U</i> test in case of non-normality - Requires (post-hoc) correction for multiple testing with more than one comparison (variable) 	<ul style="list-style-type: none"> - Do healthy subjects have less numbers of <i>Staphylococcus</i> species on the skin of their inner elbow in comparison to atopic dermatitis subjects? 	uni-variate	[156]
Pearson's test (Chi-square or Fisher's exact test)	<ul style="list-style-type: none"> - Perfect for under- / over-representation analysis - Methods are suitable for application to large datasets 	<ul style="list-style-type: none"> - Preferably applied to (two) nominal groups; meaning of results is difficult to interpret with a large number of ordinal categories - Not suitable for small datasets 	<ul style="list-style-type: none"> - Is there a bacterial gene (functionality) in the microbiome of unaffected skin (inner-elbow) of atopic dermatitis patients that is more abundantly present compared to healthy volunteers? 	uni-variate	[157]
Pearson correlation coefficient (non-parametric rank based: Spearman or Kendall)	<ul style="list-style-type: none"> - Provides a correlation coefficient number between two variables - The correlation <i>p</i>-value can be calculated based on distribution 	<ul style="list-style-type: none"> - Correlations are descriptive statistics, and not necessarily causal; be careful with inference based on correlation - Pearson is sensitive to noise / outliers in your dataset, one can use rank-based alternatives instead 	<ul style="list-style-type: none"> - Does the number of <i>Staphylococcus</i> species on inflamed skin of the inner-elbow of atopic dermatitis patients correlate to the local disease severity score? 	uni-variate / correlation statistics	[158]

Method (2B)	Advantages	Limitations	Example of typical research question	Type	Ref.
Paired Student's t-test (non-parametric: Wilcoxon signed-rank test)	<ul style="list-style-type: none"> - Quick and very accessible method - Perfect for matched pairs data (i.e. repeated measures) - Method can cope with missing data points 	<ul style="list-style-type: none"> - Similar limitations as described for the unpaired Student's t-test 	<ul style="list-style-type: none"> - Do atopic dermatitis subjects with inflammation flare-up in their inner elbow have less numbers of <i>Staphylococcus</i> species on their skin in drug treated (left) versus untreated (right) arms? 	uni-variate / repeated measures	[156]
(M)ANOVA (non-parametric: Kruskal-Wallis)	<ul style="list-style-type: none"> - Quick and very accessible method - Multiple / repeated measurements are allowed (also longitudinal) 	<ul style="list-style-type: none"> - Sensitive to outliers / imbalanced sample groups - Missing values require imputation - Post-hoc analysis of contribution of variable to variance in factor(s) 	<ul style="list-style-type: none"> - I measured the gene expression of several antimicrobial peptides in atopic dermatitis patients. Are these (combined) gene expression profiles significantly different from those in healthy volunteers? 	uni-, multi-variate / repeated measures	[159]
Random Forest (RF) (or its massively parallel variant: Random Jungle)	<ul style="list-style-type: none"> - Perfect for creating predictive models based on the underlying data - Useful for reducing complexity of your dataset 	<ul style="list-style-type: none"> - Requires expert knowledge and specialized software - RF p-values are not straightforward to calculate (requires permutation) 	<ul style="list-style-type: none"> - We have one set of atopic dermatitis-associated <i>Staphylococcus</i> genomes, and another set of <i>Staphylococcus</i> genomes exclusively present on healthy subjects. Can we find one or a combinatorial set of genes that discriminate between health and disease-associated <i>staphylococci</i>? (i.e. gene-trait-matching) 	machine-learning / multi-variate	[142]

Table 2B part 1 of 2

Principal Component Analysis (PCA) (unsupervised)	<ul style="list-style-type: none"> - Very useful method for generating an overview of your data - Provides insight into supplementary variables - Can also be applied on spatial data i.e. principal coordinate analysis (PCoA) 	<ul style="list-style-type: none"> - Requires expert knowledge and specialized software - Strong need to correct for covariates (confounding variables) - PCA does not provide <i>p</i>-values 	<ul style="list-style-type: none"> - For a group of atopic dermatitis patients and healthy controls we have unaffected skin host gene expression data of the inner elbow. - Do we see distinctive profiles (clusters) based on amplitudes of gene expression between patient subjects and healthy controls? - Can we link this to a particular set of genes? Does this change for patients who are faced with a disease flare-up? 	multi-variate	[140, 141]
Redundancy Analysis (RDA) (comparable to canonical correspondence analysis: CCA) (supervised)	<ul style="list-style-type: none"> - Generic and computationally efficient method - Statistics (<i>p</i>-value) by permutation, and no need to correct for multiple measurements (i.e. multiple testing) - Provides insight into supplementary variables - Enables one to (characterize and) correct for any confounding variable in the dataset 	<ul style="list-style-type: none"> - Requires expert knowledge and specialized software - Strong need to correct for covariates (confounding variables) - Explanatory variable(s) needs to be defined - Individual candidate response variables require corroboration with additional univariate statistics methods - Limited applicability to incomplete longitudinal data 	<ul style="list-style-type: none"> - For a group of atopic dermatitis patients and healthy controls we have unaffected skin microbiome data of the inner elbow. Are we able to significantly discriminate between patient subjects and healthy controls (disease state is explanatory variable), based on bacterial genera and abundance that were measured for those volunteers? - Can we link this to a particular (set of) bacterium? - What happens if we add patients samples that were taken upon a disease flare-up? 	multi-variate	[140, 141]

Table 2B part 2 of 2

Method (2C)	Advantages	Limitations	Example of typical research question	Type	Ref.
Multilevel temporal Bayesian networks (MTBN)	<ul style="list-style-type: none"> - Very suitable to gain insight into the correlations between a microbiome and its niche 	<ul style="list-style-type: none"> - Works best if the temporal data has been aligned (true for most longitudinal methods) 	<ul style="list-style-type: none"> - How are characteristics of atopic dermatitis patient groups linked to differences in their skin microbiome and their long term (skin) health outcome? 	longitudinal	[143]
Mixed-effect (Regression) Models (MRM)	<ul style="list-style-type: none"> - Particularly useful for repeated measurements - Handles missing values, i.e. unbalanced data are no problem - Handles non-synchronized data 	<ul style="list-style-type: none"> - Cannot resolve interaction(s) between time-variance and subject-variance - Mixed models are incredibly flexible; proper modelling and biological knowledge is crucial to prevent over- or misinterpretation 	<ul style="list-style-type: none"> - We measured the temporal microbiome (stability) of unaffected skin in the inner-elbow of one atopic dermatitis patient up to disease flare-up and recovery from inflammation. Can we pin-point strain or gene biomarkers in the microbiome that predict disease progression? - How does variation in the skin microbiome relate to variation in other factors present in the niche? - Which correlations exist between the skin microbiome and the conditions in its niche, when the timing of a biological process in the human host is aligned in all subjects? 	longitudinal	[155, 160]
(Generalized) Linear Models (GLM)	<ul style="list-style-type: none"> - Generic / computationally efficient method - Can be extended to generalized estimating equations (GEE) 	<ul style="list-style-type: none"> - Limited applicability to incomplete longitudinal data 	idem. MRM	longitudinal	[155, 160]
Dynamic Time Warping (DTW)	<ul style="list-style-type: none"> - Very relevant for comparing time series - Especially suitable for correcting global (and also local) time delay 	<ul style="list-style-type: none"> - Computationally very expensive 	idem. MRM	longitudinal	[161]
Extended Local Similarity Alignment (eLSA)	<ul style="list-style-type: none"> - Made for time series - Generic and has been applied to microbiota datasets 	<ul style="list-style-type: none"> - Resulting networks might be complex to interpret - Not very suitable for repeated measures experiments 	idem. MRM	longitudinal	[162]

Finally, we distinguish uni- and multivariate statistics [138]. The main difference is that in multivariate data multiple variables (all which could potentially influence study outcome) are analysed simultaneously, whereas in univariate statistics this is limited to only one variable. In multivariate statistics explanatory and confounding variables are important concepts [147]. The former are variables that are possibly predictive of the outcome under study, and therefore also known as independent variables or features. Whereas the latter, covariates, may be variables interacting with other (explanatory) variables, and thereby potentially influencing them. For this reason it is common practice to correct for confounding variables, but is not straight-forward and may require statistical approaches and previous knowledge to recognize these type of variables in your dataset. The fact that with multivariate statistics multiple (potentially interdependable and / or confounding) variables are analysed simultaneously makes the statistical algorithms and methods by which they are typically inquired inherently more complex compared to univariate statistics, requiring more expert knowledge and in case of biomedical data more feedback from clinicians and biologists in order to identify relevant (confounding) variables.

Cross-sectional studies.

Studies where exposure and outcome are determined at the same time-point for each study volunteer, are called cross-sectional studies [148]. Examples of exposures are perturbations, interventions, treatments and medical conditions. These designs typically involve relatively straight-forward study contrasts, yet can have high complexity when various parameters / variables and multiple contrasts are examined in the same study effort. Univariate cross-sectional study designs allow for use of statistical tests like Student's *t*-test for contrasts analysis (or its non-parametric variant: Mann-Whitney *U* test), Pearson's tests for overrepresentation analysis (e.g. Chi-square or Fisher's exact test; and not to be confused with Pearson correlations) or several correlation statistics such as Pearson and Spearman rank (Table 2A). When analysing cross-sectional data involving many variables, multivariate statistical methods are available such as (M) ANOVA, random forest, PCA and RDA (Table 2B). In the end, choice of statistical test primarily depends study design, type of data and on the research question at hand.

Longitudinal and repeated measures studies.

Besides cross-sectional studies we distinguish multiple measures study designs. Large variability exists between individuals on various levels: e.g. genotype, immunological features [149], metabolic profiles [150] and the microbiome [151, 152]. Due to this variability it is difficult to determine commonalities between individuals and to associate microbiome changes with for instance subject parameters. Using multiple measurements of the same individual (e.g. in time / longitudinal studies, or repeated measures study designs) combined with analysis methods enables one to determine these commonalities across individuals and to thereby determine causality of microbiome and health state across different individuals (Figure 3) [153]. Longitudinal studies and their associated longitudinal datasets are becoming common ground as they allow dealing with this inter-individual variability [69]. Challenges in the analysis of these longitudinal data lie in the discovery of patterns across groups of subjects, and also in pinpointing (random) differences from subject to subject [154].

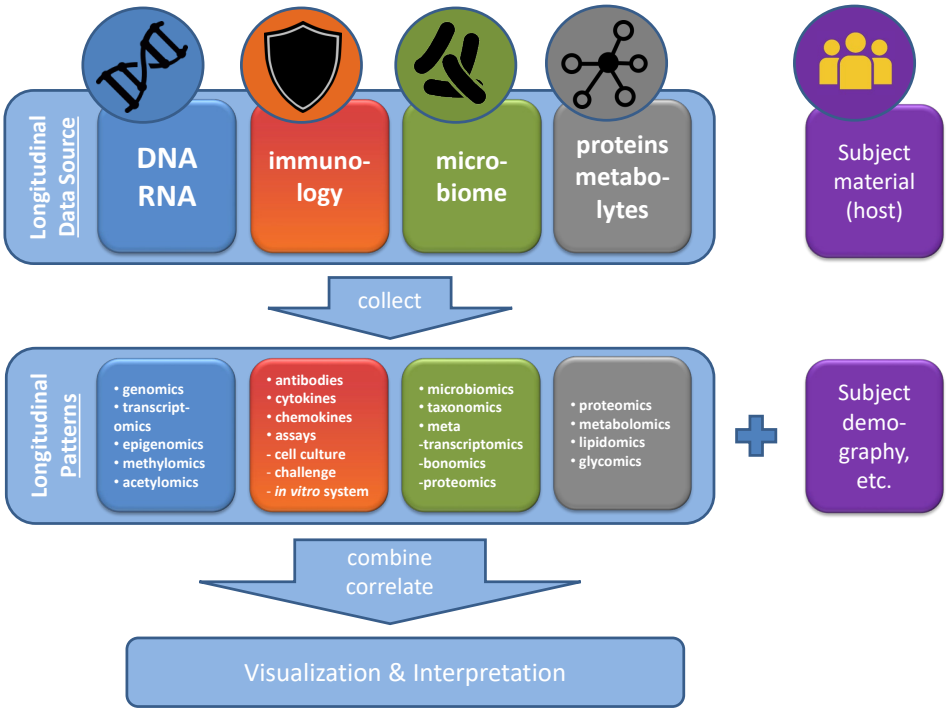


Figure 3. Systems Biology.

The figure represents steps typically involved in an analysis protocol for host / microbiota cross-sectional or longitudinal datasets. The -omics data are first collected from different host sources (*top*), data patterns are extracted, and combined and correlated with available host demographic and life style parameters (*middle*), and subsequently visualized and (biologically) interpreted as guided by expert knowledge (*bottom*). A few analysis methods or (statistical) algorithms that could be applied in such protocol are shown in Table 2.

Different methods have been reported for the analysis of longitudinal or repeated measures (microbiome) data (Table 2B/C). These are the (M)ANOVA, mixed-effect (regression) models (MRM), (generalized) linear models (GLM), dynamic time warping and extended local similarity alignment (eLSA). Furthermore, one can apply a more simple paired Student's t-test, or its non-parametric variant: Wilcoxon signed-rank test. In life-sciences, analysis methods with the use of GLM such as generalized estimating equation (GEE) have also been applied for longitudinal data analysis [155], and penalized and regularized variants of those, and could in principle be extended to microbiology- or host-derived data. Even methods applied for the analysis of time series of gene expression data, such as Bayesian networks can be applied to the analysis of longitudinal datasets [143]. The co-variance structures / correlation structures between genes (linked by the gene-regulatory networks) could be similar to the covariance structures between microbiota members and their interaction with the host via metabolites. Likewise, multivariate methods such as PCA and RDA can in principle also be applied to gain insight into samples that behave similarly (over time or measurement) [140, 141]. A different method is to use random forest classification of samples based on curve-fits of the microbiome / host parameters across time [142]. Of course, after longitudinal data are aligned and patterns are distilled, methods typically applied to cross-sectional studies (e.g. on study contrasts) can likewise be used (Table 2A). In conclusion, biological processes take place in a different pace in every human, when this difference is corrected for with alignment by longitudinal analysis methods, then subsequently identified correlations are statistically stronger.

Decisions on study analysis approaches – what, where and when ?

A statistical analysis method may in principle be applied to any type of data from any kind of biological source or process. In this thesis, mainly: *microbiota* (e.g. microbiomics, metagenomics), *genes* (e.g. genomics, (meta-)transcriptomics), *epigenetics* (e.g. methylomics, histone acetylation patterns) and *protein* data (e.g. proteomics and metabolomics). However, the challenge is to choose the right analysis approach for analysing a particular (longitudinal) dataset. Many methods exist for the analysis of (temporal) patterns in multivariate data. The optimal solution is likely dependent on dataset (type) and research question. The larger the dataset and the more complex the experimental design, the more correlations can be identified making it challenging to pinpoint the relevant ones [152, 154]. But also, for example, a dataset with both microbiome and additional host parameters data requires another methodological approach than a dataset with microbiome data only, as the nature of the data changes and the number of data points hugely increases. Therefore, in the end, the choice for analysis workflow and statistics of a certain study dataset is likely to depend on a combination of (i) what is conventional in the field, (ii) specific study questions and deliverables, (iii) experience and expertise of the bioinformatician, but also (iv) more pragmatic decisions with respect to time-lines.

Thesis summarizing outline

Confident bacterial genome assembly and annotation is key for any downstream genome analysis and for processes / methods requiring genomic content information. For example, in 16S-derived function prediction such as by PICRUSt, or for metagenomics data analysis, such as by MetaPhlAn. Furthermore, bacterial taxonomy assignment is for a great part based on known clusters of conserved genes shared among phylogenetically related bacteria, for which genome information is crucial. Likewise, for mapping metatranscriptomics reads to their respective genomes. For these purposes, in [Chapter 2](#), we explore novel methodologies for improving bacterial (automated) genome annotation processes by combining gene open reading frame and function prediction outcomes of multiple annotation engines. We show that we are able to successfully apply this on a wide set of different bacteria, thereby moderately improving prediction accuracy. Knowing what annotation method works best for bacterial genomes is valuable information, laying the foundation for downstream research applications as outlined in chapters thereafter.

In [Chapter 3](#), we further dive into techniques for bacterial strain tracking and identification for application in biomedical research, and discuss currently established (MLST, 16S) and more recent (SLST, metagenomics) methods for typing of single isolates or complex mixtures of microbiota. Here, we present TaxPhlAn, a novel method and bioinformatics pipeline for SLST-based profiling of bacteria beyond the species-level, which is currently still challenging. We showcase a clinically-relevant case-study of high-resolution *Staphylococcus* profiling on skin of atopic dermatitis patients, and demonstrate that SLST enables profiling of cutaneous *Staphylococcus* members at (sub)species level, is cost-effective and especially by combining with 16S rRNA gene sequencing, and shows higher resolution than current 16S-based sequencing techniques.

In [Chapter 4](#), in close collaboration with the *Laboratory of Experimental Dermatology* (Radboudumc), a novel skin model is described to advance the skin host-microbiota research. This model, based on human callus, allows for bacterial growth of typical skin inhabitants in their natural environment. Application of this *in vitro* skin 'callus' model (in short) for bacterial growth mimics the human *stratum corneum* and thereby opens-up new avenues for straight-forward and costs-effective in-depth screening and studying of skin microbiota-microbiota interactions, for example commensal-pathogen. This skin model additionally allows for testing bacterial growth under a multitude of experimental conditions, in medium-throughput.

In [Chapter 5](#), we further demonstrate application of the aforementioned skin model, where we use it to experimentally corroborate findings from a larger ichthyosis vulgaris (IV) patient cohort about the impact of human filaggrin-null mutations on the skin microbiome. We report that skin on the lower leg of IV patients is marked by significantly less Gram-positive anaerobic cocci (GPAC), a phylogenetically related family of commensal skin bacteria. Interestingly, we are additionally able to *in silico* link the difference in availability of histidine and other natural moisturizing factors (known

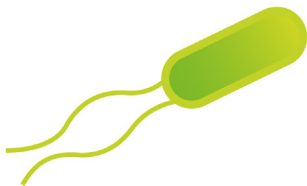
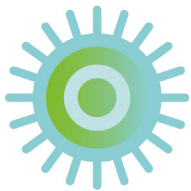
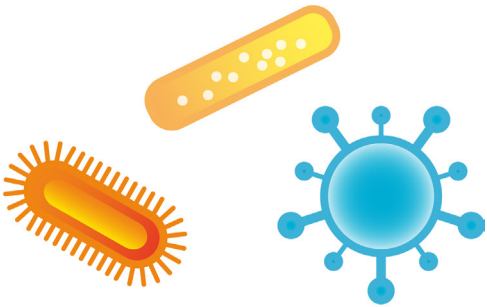
breakdown products of flaggrin) to an overall decreased histidine utilization capacity in the skin microbiome of flaggrin-deficient patients. In particular, we show the GPAC bacterium *Finegoldia magna* to grow less well on callus from flaggrin-deficient (IV) patients, in comparison to flaggrin from wild-type (healthy) volunteers. To our best knowledge, this is currently one of the scarce study examples where is demonstrated that a host genetic factor has a direct effect on the host its microbiome.

In [Chapter 6](#), in collaboration with the *Laboratory of Experimental Rheumatology* (Radboudumc), we investigated gut-associated microbiota in mice. We first demonstrate in IL-1 receptor antagonist (IL-1Ra) KO mice, a model for rheumatoid arthritis, that diversity and composition of the commensal intestinal microbiota is critically maintained by IL-1 signalling. Amongst others, we describe a reduced intestinal microbial diversity and richness, and overrepresented *Helicobacter* and underrepresented *Prevotella* species in IL-1Ra-deficient mice. Most notably, we show that aberrant intestinal microbiota in IL-1Ra-deficient mice specifically potentiate pro-inflammatory IL-17 production by intestinal lamina propria (LP) lymphocytes, and skew the LP T-cell balance in favour of cytotoxic T-helper 17 cells. Interestingly, this effect is transferable to wild-type mice by fecal microbiota transplantation. Furthermore, we intriguingly show that LP Th17 cell expansion and the development of spontaneous autoimmune arthritis in IL-1Ra-deficient mice are attenuated under germ-free conditions. There we demonstrate an important role for gut (commensal) microbiota in inflammation- and autoimmune-related disease processes.

Hereafter, in [Chapter 7](#), in close partnership with the *Department of Psychiatry* (Radboudumc) and *Donders Institute for Brain, Cognition and Behaviour* (Radboud University), we postulate novel ideas involving the gut-brain-axis in humans. In a pilot study about the gut microbiome in attention-deficit / hyperactivity disorder (ADHD) patients, we for the first time describe microbiota differences between ADHD and healthy volunteers. We show that ADHD patients contain more *Bifidobacterium* species in comparison to controls, albeit marginally significant. More importantly, a 16S-based bacterial gene functionality prediction suggests that the difference in the gut microbiome between ADHD and healthy controls is marked by significantly increased levels of predicted cyclohexadienyl dehydratase (CDT), an enzyme responsible for phenylalanine synthesis. Phenylalanine is a precursor for dopamine, and dopamine is known to be strongly involved in ADHD pathogenesis. It therefore is intriguing that we in this chapter are able to demonstrate that increased relative abundance of the CDT functionality is significantly associated with decreased *ventral striatal* fMRI brain activity responses during tests involving reward anticipation. Reward anticipation is a hallmark of ADHD, although the observed effect was found to be independent of ADHD disease status. Based on findings described in this chapter, we propose the idea that the bacterial by-product phenylalanine reaches the brain via the bloodstream, crossing the blood-brain-barrier, and once it is converted into dopamine at the brain site, is potentially able to influence local brain function. Thereby providing fuel for novel research focusing on the gut-brain-axis area, with a special focus on host-microbe interactomics and the implication of gut microbes on brain development and function.

Finally, in [Chapter 8](#), in tandem with the *Laboratory of Pediatric Infectious Diseases* (Radboudumc), we investigated microbiota in the nasopharynx of young infants who suffered from a respiratory syncytial virus (RSV) infection. Evidence is growing that there is cross-talk not only between pathogens (and commensals) and the host immune system, but also that microbe-microbe interactions are important denominators for disease outcome, whether these are between bacteria, fungi or viruses. As such, we report our study results on the severity of RSV infection in the context of nasopharyngeal microbiota in patients, and in comparison to those in healthy infants. We show that nasopharyngeal microbiota composition is significantly different based on CXCL8 levels of the local mucosa, a relevant chemokine that was previously found to be indicative for RSV disease severity. Interestingly, we find abundance of the genus *Haemophilus* as the strongest predictor for CXCL8 levels. Although *Haemophilus* levels did not correlate directly to clinical severity of RSV disease manifestation or its related parameters, we do indeed find that *Haemophilus* is strongly overrepresented in patients compared to healthy infants. These results are appealing clues that further substantiate a potential role for cross-kingdom interaction in disease pathogenesis (either directly, or through host processes), and therewith stress the relevance of focusing on incorporating such analyses for better understanding the mechanisms of disease. This will be particularly relevant in the near future due to increasing accessibility to metagenomics resources, from which cross-kingdom information can more easily be extracted.

In the final chapter of this thesis, [Chapter 9](#), implications of results and conclusions of all presented studies are discussed in short, as well as final considerations and future perspectives are outlined. But more importantly, with exception of Chapter 4, all current chapters have already been published in peer reviewed journals: Therefore, in this chapter, I took the liberty of focussing on what I – as a junior scientist – consider to be my major learning points and challenges. In conclusion, I will discuss what I have come to take to heart, and now envision, when it comes to doing cutting-edge research in the field of microbiomics and host-microbe interactions in human health and disease.



PART I

Bioinformatics Tools and Analyses

CHAPTER 2

OPEN ACCESS

as published in PLoS One, 2013, May 10;8(5):e63523

<https://doi.org/10.1371/journal.pone.0063523>

¹ Centre for Molecular and Biomolecular Informatics, Radboud University Medical Centre, Nijmegen, NL.

² Netherlands Bioinformatics Centre, Nijmegen, the Netherlands.

³ NIZO, Ede, the Netherlands.

⁴ Top Institute Food and Nutrition, Wageningen, the Netherlands.

CHAPTER 2

**COMPANION: REDUCE MANUAL CURATION BY
COMBINING GENE PREDICTIONS FROM MULTIPLE
ANNOTATION ENGINES**

A CASE-STUDY OF START CODON PREDICTION

Thomas H.A. Ederveen ¹

Lex Overmars ^{1,2}

Sacha A.F.T. van Hijum ¹⁻⁴

ABSTRACT

Nowadays, prokaryotic genomes are sequenced faster than the capacity to manually curate gene annotations. Automated genome annotation engines (AGEs) provide users a straight-forward and complete solution for predicting ORF coordinates and function. For many labs, the use of AGEs is therefore essential to decrease the time necessary for annotating a given prokaryotic genome. However, it is not uncommon for AGEs to provide different and sometimes conflicting predictions. Combining multiple AGEs might allow for more accurate predictions. Here we analyzed the *ab initio* open reading frame (ORF) calling performance of different AGEs based on curated genome annotations of eight strains from different bacterial species with GC% ranging from 35 - 52%. We present a case study which demonstrates a novel way of comparative genome annotation, using combinations of AGEs in a pre-defined order (or path) to predict ORF start codons. The order of AGE combinations is from high to low specificity, where the specificity is based on the eight genome annotations. For each AGE combination we are able to derive a so-called projected confidence value, which is the average specificity of ORF start codon prediction based on the eight genomes. The projected confidence enables estimating likeliness of a correct prediction for a particular ORF start codon by a particular AGE combination, pinpointing ORFs notoriously difficult to predict start codons. We correctly predict start codons for 90.5 ± 4.8 % of the genes in a genome (based on the eight genomes) with an accuracy of 81.1 ± 7.6 %. Our consensus-path methodology allows a marked improvement over majority voting (9.7 ± 4.4 %) and with an optimal path ORF start prediction sensitivity is gained while maintaining a high specificity.

Background

The accurate annotation of bacterial genomes is essential to apply sequence data in many (bio)medical research topics such as microbiology, immunology and infectious diseases [163, 164]. It is required for a better understanding of the biology of bacteria as it involves identification of genes and subsequent proteins, regulatory networks and pathways. In practice, genome annotation often starts with the submission of a genome sequence to online annotation services, also named automated genome annotation engines/pipelines, which will be referred to as AGEs throughout this manuscript [165]. The output of these services usually consists of *ab initio* predicted open reading frames (ORFs) with start-, stop positions and function predictions. Start- and stop codon prediction is usually performed by ORF calling software, such as GLIMMER [166, 167], GeneMark [168, 169] or Prodigal [170], implemented in these AGEs. Correctly predicting ORFs is essential; prediction of gene function, ribosomal binding sites, promoter mapping, and subcellular location are all dependent on correct start codon prediction. Subsequent functional annotation of ORFs involves many steps including BLAST-like [171] searches in existing databases such as RefSeq [172], Genbank [173] and SwissProt [174], or hidden Markov model screenings with Pfam [175] or FIGfams [176]. As AGEs consist of different prediction steps and associated parameters ([Supplementary Table S1](#)), they can for a given genome suggest different ORF predictions [177]. AGEs not uncommonly provide incorrect annotation calls [178, 179] (according to our study, roughly 14% - 58% of start codon predictions are incorrect; see below). This begs the question: which AGE to choose for my genome of interest? Next to choosing a particular AGE to annotate a genome of interest, majority voting has been suggested as a method to combine predictions from different ORF prediction algorithms [180-185]. However, one is unable to know which predictions are likely in need for manual curation and which are likely to be correct. Therefore, most predicted ORFs are manually curated for start- and stop codons and gene function [186]. In order to prioritize genes to be manually curated it would therefore be highly advantageous to allocate a level of confidence to every ORF prediction.

Here, we studied the start codon prediction performance of different AGEs based on a total of twelve strains from different bacterial species with widely differing GC%. Eight of these genomes have GC% ranging from 35 - 52% and the remaining four genomes have more extreme GC%. The genome annotations of the twelve well-studied strains have been extensively (manually) curated and are therefore considered to be of high quality. We chose not to incorporate stand-alone ORF calling software (e.g. GLIMMER [166, 167], GeneMark [168, 169] and Prodigal [170]) as this work is meant as a practical case study and therefore focuses on AGEs, as these are commonly used in annotation efforts. As stop codon predictions are only rarely being predicted wrong by AGEs (see below) we present a novel method to combine the result of multiple AGEs in order to more reliably predict start codon locations. Our work-flow uses consensus predictions by specific combinations of AGEs in a particular order (or path). This path was quite conserved for the eight moderate GC% organisms (35 - 52%) under study but less with the more extreme GC% genomes. The order of AGE combinations is from high to low specificity, where the specificity is based on the start codon prediction performance

of these AGE combinations on the individual genome annotations. Based on the eight moderate GC% genomes, this path allows us to correctly predict start codons for $90.5 \pm 4.8\%$ of the genes in a genome with an accuracy of $81.1 \pm 7.6\%$. For each AGE combination we are able to derive a novel so-called projected confidence value, which is the average specificity of ORF start codon prediction based on the eight genomes. This projected confidence value allows pinpointing ORFs of which the start codon is likely notoriously difficult to predict. We hypothesize that the proposed concept can also be applied to the prediction of stop codons and importantly gene function, allowing a researcher to focus resources by manually curating fewer genes.

Results

We studied the ORF start codon predictions by four AGEs (BASys, ISGA, RAST and xBASE; [Table 1](#)) for twelve genomes from well-studied strains of different bacterial species. This set of twelve genomes consists of eight genomes with a moderate GC% (percentage guanine-cytosine; moderate is defined as a range from 35 - 52%): *Escherichia coli* K12 MG1655, *Bacillus subtilis* 168, *Lactobacillus plantarum* WCFS1, *Lactococcus lactis* KF147, *Streptococcus pneumoniae* TIGR4, *Salmonella enterica* subsp. *enterica* serovar Typhi str. Ty2, *Neisseria meningitidis* MC58 and *Haemophilus influenzae* Rd KW20 ([Table 2](#)); and four genomes with a more extreme GC content: *Mycobacterium tuberculosis* H37rv, *Mycoplasma mobile* 163K, *Pseudomonas putida* KT2440 and *Streptomyces coelicolor* A3(2) ([Table 2](#)). In our analysis we evaluated all ORFs that were either predicted by an AGE or that were present in the reference genomes ([Figs. 1 and 2](#) and see [Methods](#)). Below we explore, next to our consensus-path prediction method ([Fig. 3](#)), different alternative approaches to obtain accurate start codon predictions. The results for these alternative approaches are based on a representative set of four moderate GC% genomes: *E. coli* K12 MG1655, *B. subtilis* 168, *L. plantarum* WCFS1, *L. lactis* KF147 ([Table 2](#)). Next, we present the results of our consensus-path approach based on the above-mentioned twelve genomes.

ORF predictions of the four AGEs only partly overlap.

Table 1. Automated genome annotation engines (AGE) used in this study.

The listed AGEs are commonly used pipelines for analysis and annotation of gene function and start- and stop codons.

Engine name (AGE)	Website of AGE	Reference
BASys	http://basys.ca/	Van Domselaar <i>et al.</i> , 2005 [203]
ISGA	http://isga.cgb.indiana.edu/	Hemmerich <i>et al.</i> , 2010 [204]
RAST	http://rast.nmpdr.org/	Aziz <i>et al.</i> , 2008 [205]
xBASE	http://www.xbase.ac.uk/annotation/	Chaudhuri <i>et al.</i> , 2008 [206, 207]

To assess the overlap in ORF start codon predictions by different AGEs we compared the predicted start codons to those in the original annotation for four moderate GC% reference genomes: *B. subtilis* 168, *E. coli* K12 MG1655, *L. lactis* KF147 and *L. plantarum* WCFS1 ([Table 2](#)). These four genomes are assumed to be a fair representation of moderate GC content genomes. Some AGEs better predict ORF start codons than other AGEs ([Supplementary Fig. S1](#)), indicated by the different levels of correct- and

Table 2. Genomes used in this study.

Listed is genome (sequence) data from the twelve reference strains used in this study (genome annotations as downloaded on December 7, 2012). * According to GenBank (<http://www.ncbi.nlm.nih.gov/genbank/>) [11]. ** Totalling both protein-coding and tRNA genes; no pseudogenes were included.

Strain (genome)	GC%	Genome size (Mb)	Gram (+/-)	Sequence date / last update *	NCBI accession number	Number of annotated genes **
<i>Bacillus subtilis</i> 168	44	4.22	+	18-NOV-1997 20-JAN-2012	NC_000964.3	4262
<i>Escherichia coli</i> K12 MG1655	50	4.64	-	16-JAN-1997 11-JAN-2012	NC_000913.2	4235
<i>Haemophilus influenzae</i> Rd KW20	38	1.83	-	28-JUL-1995 15-OCT-2012	NC_000907.1	1715
<i>L. lactis</i> KF147	35	2.60	+	01-DEC-2009 21-NOV-2011	NC_013656.1	2605
<i>Lactobacillus plantarum</i> WCFS1	44	3.31	+	25-JUN-2001 21-NOV-2011	NC_004567.1	3128
<i>Mycoplasma mobile</i> 163K	25	0.77	-	13-APR-2004 01-APR-2010	NC_006908.1	661
<i>Mycobacterium tuberculosis</i> H37rv	66	4.41	-	13-SEP-2001 20-AUG-2012	NC_000962.2	4048
<i>Neisseria meningitidis</i> MC58	52	2.27	-	17-MAR-2000 19-JAN-2012	NC_003112.2	2122
<i>Pseudomonas putida</i> KT2440	61	6.18	-	08-APR-2002 27-SEP-2012	NC_002947.3	5424
<i>Streptomyces coelicolor</i> A3(2)	72	8.67	+	09-MAY-2002 19-JAN-2012	NC_003888.3	7833
<i>Salmonella enterica</i> subsp. <i>enterica</i> serovar Typhi str. Ty2	52	4.79	-	25-SEP-2002 24-OCT-2012	NC_004631.1	4448
<i>Streptococcus pneumoniae</i> TIGR4	40	2.16	+	29-JUN-2001 20-JAN-2012	NC_003028.3	2163

incorrect predicted ORFs (e.g. ISGA consistently has the highest absolute number of correct predictions). This is also illustrated by the different combinations of ORF start codon prediction qualifications assigned for each AGE (e.g. correct-, incorrect-, false positive-, false negative- or N/A prediction), which were found for these four genomes (Fig. 4). For instance, more ORF start codons are correctly predicted by BASys compared to ISGA ($2.6 \pm 0.1\%$ and $1.7 \pm 0.4\%$ respectively). Depending on the genome, either ISGA (83.6% correct predicted ORF start codons for *L. plantarum*) or RAST ($82.7 \pm 2.3\%$ correct predicted ORF start codons for *L. lactis*, *E. coli* and *B. subtilis*) performs best in respect to ORF start codon prediction accuracy (Fig. 4, and Supplementary Fig. S1). Based on the four genomes, about half of the ORF start codons were correctly predicted by all four AGEs. The four AGEs perform inconsistent (Fig. 4) for a considerable fraction of the predicted ORFs (ranging from 40.6% for *L. lactis* KF147 to 57.9% for *B. subtilis* 168). In conclusion, for these four reference genomes, not one AGE will provide the best possible result for all ORFs (Fig. 4). We therefore hypothesize that a combination of AGEs might allow for more accurate prediction of start codon coordinates.

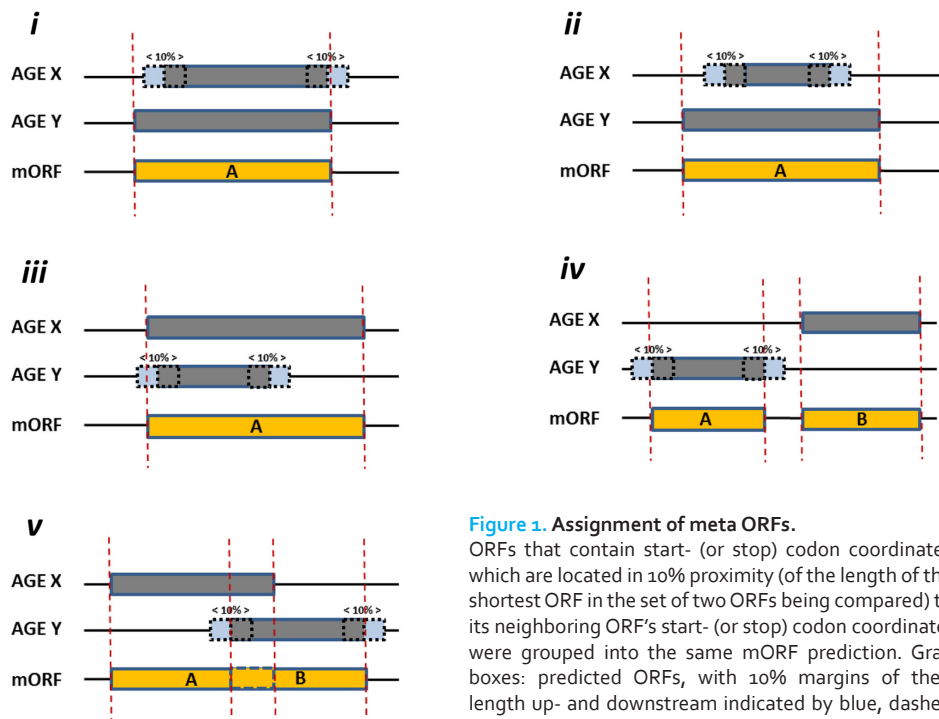


Figure 1. Assignment of meta ORFs.

ORFs that contain start- (or stop) codon coordinates which are located in 10% proximity (of the length of the shortest ORF in the set of two ORFs being compared) to its neighboring ORF's start- (or stop) codon coordinate, were grouped into the same mORF prediction. Gray boxes: predicted ORFs, with 10% margins of their length up- and downstream indicated by blue, dashed boxes. Yellow boxes: reference genome ORFs.

Here, five exemplary mORF allocations are illustrated: (i) The stop codon coordinate of the suggested ORF provided by AGE_Y matches the ORF stop coordinate of that suggested by AGE_X. Therefore, these two ORFs are grouped into the same mORF (mORF A). (ii) The ORF predicted by AGE_X falls within the ORF predicted by AGE_Y. Therefore, these two ORFs are grouped into the same mORF (mORF A). (iii) The start codon predicted by AGE_X falls within the 10% boundary of the start codon predicted by AGE_Y. Therefore, these two ORFs are grouped into the same mORF (mORF A). (iv) The predicted ORF by AGE_X contains start- and stop codons both located outside the 10% boundary to the start- and stop codons of the ORF predicted by AGE_Y. Therefore, these two ORFs are assigned different mORFs (mORFs A and B). (v) Comparable to iv, the predicted ORF by AGE_X contains start- and stop codons located outside the 10% boundary to the start- and stop codons of the ORF predicted by AGE_Y. Even though there is small overlap between these two ORFs (dashed line), they are assigned different mORFs (mORFs A and B) because they exceed the 10% limit. Note that predicted ORFs by the fictive AGEs X and -Y must have the same orientation to be assigned to the same mORF.

Start codon prediction performance by majority voting.

Majority voting allows combining the annotation predictions from different AGEs and potentially results in more reliable predictions [180]. With majority voting, a start codon for a given ORF is based on the fact that it was predicted by most AGEs. The more AGEs are in consensus, the higher the confidence in the majority-voted start codon for that particular ORF. We evaluated the performance of majority voting against the single AGE with the lowest percentage of incorrect start codon predictions (ISGA or RAST, depending on the genome) on the four genomes. Majority voting introduces many more false positive (a predicted ORF was not annotated in the reference genome) ORFs (153 - 397) and false negative (ORF present in the reference genome, but not predicted by AGEs) ORFs (284 - 924) compared to ISGA or RAST (Fig. 5). However, the absolute number of incorrect predictions is also slightly lower with majority voting compared

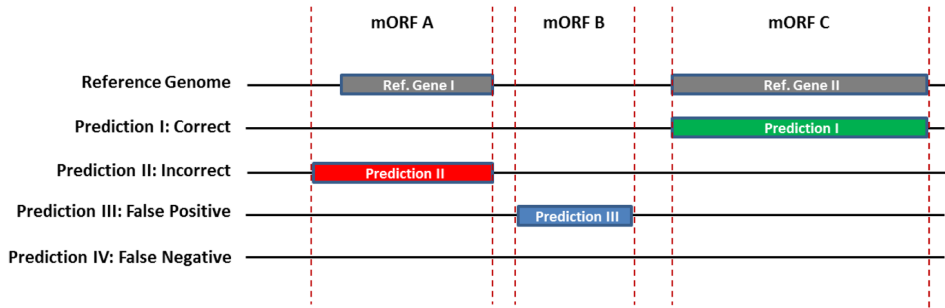


Figure 2. Classification of AGE gene predictions (possible situations).

Gray: reference genes from a bacterial reference genome with which the AGE predictions are compared. Situation I (green): correct AGE prediction with matching coordinates. Situation II (red): “incorrect” predictions with an incorrect start- and/or stop codon, but belonging to the same mORF. Note that the majority of stop codons were correctly assigned ([Supplementary Fig. S1B](#)). Situation III (blue): false positive ORFs predicted by an AGE and not present in the reference genome. Situation IV: false negative predictions were allocated to AGEs that failed to predict a reference ORF.

to ISGA or RAST alone. Likely, certain AGEs introduce inconsistencies in the voting results and thereby prevent correct predictions from being selected, in disfavor of other engines. Because predictions are only considered if a majority is reached, majority voting could lead to fewer predicted ORFs: ORFs for which voting results in a draw are missed. In conclusion, different AGEs may predict different ORFs correctly. As majority voting results in many false positive and false negative predictions, we hypothesize that specific combinations of AGEs possibly achieve better ORF start codon predictions.

Start codon prediction performance by consensus predictions.

Another approach is to trust only the consensus prediction of (combinations of) AGEs. A prediction is only considered if all or a subset of engines predicts the same start codon. We therefore categorized predicted ORFs for which the start codon was determined in consensus for all combinations of two, three or four AGEs. Each of these determined consensus predictions was classified according to the reference genome in one of the following categories: (i) correct, (ii) incorrect, (iii) false negative or (iv) false positive ([Fig. 2](#)). Different combinations of two, three or four AGEs show different subsets of ORF start codons to be correctly predicted in *E. coli* K12 MG1655 ([Supplementary Figs. S3 and S4](#)). For example, the combined consensus prediction of RAST and xBASE has a sensitivity (i.e. the coverage level with respect to the reference genome ORFs, taking into account both incorrectly and the correctly predicted ORFs; see [formula 3](#)) of 77.5% and a specificity (i.e. the percentage correct of predictions; see [formula 2](#)) of 86.0%, while the combination of RAST and BASys has a sensitivity of 58.3% and a specificity of 90.8%. Expectedly, the consensus prediction results of BASys, RAST and xBASE has a decreased sensitivity to 52.4%, but an increased specificity to 92.2%. Apparently, combining multiple AGEs allows predicting fewer ORF starts (lower sensitivity), but allows a more reliable start codon prediction (increased specificity).

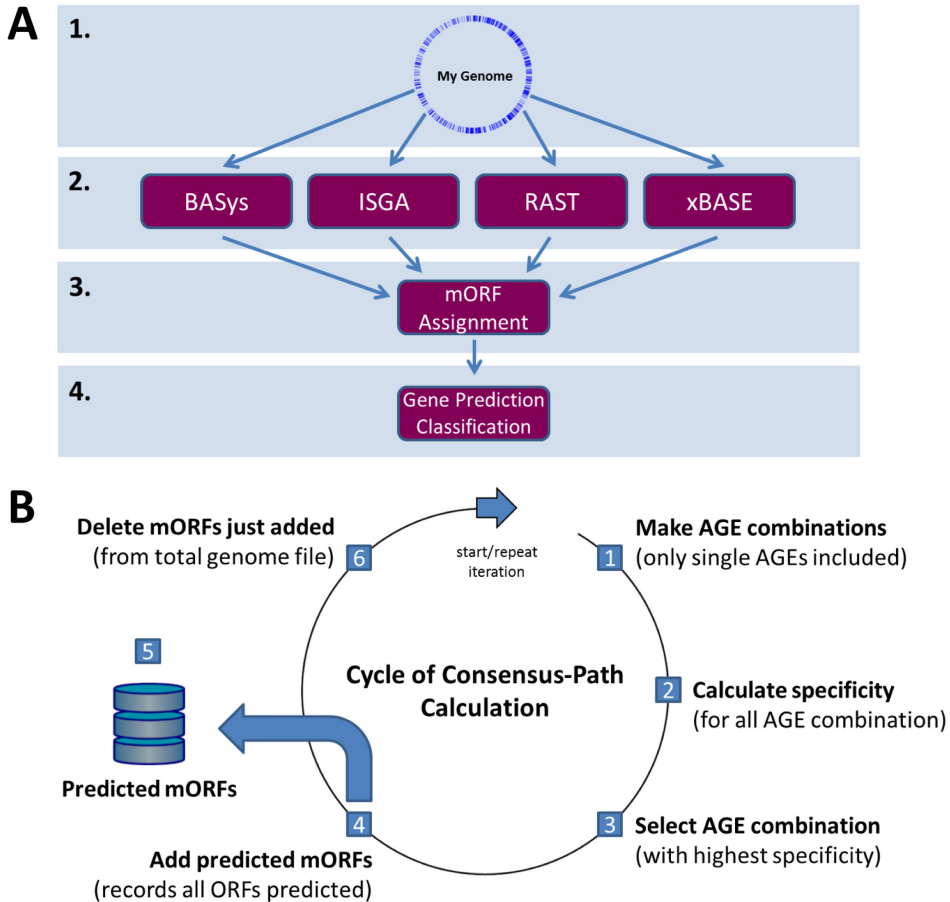


Figure 3. The methodology of comparative annotation and a step-by-step description for the consensus-path.

(A) In our methodology, a single genome of interest (1) is first uploaded to the four different AGEs (2) for ORF prediction and annotation (see also [Table 1](#)). After receiving all predictions from the respective AGEs, a mORF assignment is performed (3), as described in [Figure 1](#). Finally, on this set of mORFs for the genome of interest, the AGE gene predictions are classified (e.g., correct and incorrectly predicted mORFs) (4) as described in [Fig. 2](#). (B) With the mORFs generated as described in A, a consensus-path calculation can be performed to find the sequential (combinations of) AGEs predicting the subset of mORFs under study with the highest specificity. This cycle (or iteration) consists of the following steps: (1) generating all possible AGE combinations (single AGEs only are also included) within the set of engines used (2) calculating for each possible AGE combination (or single AGE) its specificity for that subset of mORFs considered (see [formula 2](#) in the [Methods](#)). AGE combinations selected in the previous round are omitted in subsequent iterations. (3) The AGE combination (or single AGE) generating the highest specificity (hence, lowest error-rate) is selected. (4) These predicted mORFs are added to the (existing) list of predicted mORFs (5), for the genome of interest. (6) mORFs selected in (4) are removed from the mORFs file originally started with. The remaining mORFs are subjected to a next step of selecting the AGE (combination) with the highest specificity (1). This iteration is repeated until for a given genome no new mORFs are added to the prediction results in (4).

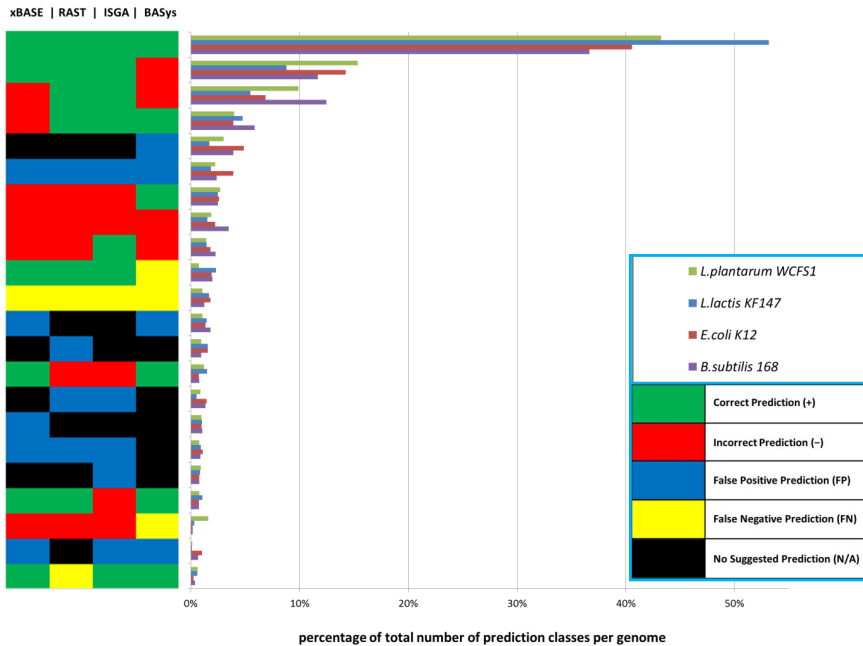


Figure 4. Variation in AGE predictions for four moderate GC% bacterial genomes.

The start codon prediction accuracy by BASys, ISGA, RAST and xBASE is illustrated in this vertical bar-graph for four bacterial reference genomes: *B. subtilis* 168, *E. coli* K12 MG1655, *L. lactis* KF147 and *L. plantarum* WCFS1. On the y-axis, the different classes of predicted ORF starts compared to the respective reference genomes are shown. A black colored box is present only in combination with false positive predictions (blue). It signifies that for these mORFs no prediction data was provided by any of the other AGEs. A total of 82 unique color/prediction classes were defined. They were plotted on the y-axis and a number was assigned according to its prevalence per bacterial genome. These numbers are shown as a bar-graph on the x-axis: as a fraction/percentage of the total number of mORFs available for that genome. In order to reduce the number of classes, those that occurred on average less than 0.50% in the four genomes were removed leaving 22 prediction classes. See for all 82 classes [Supplementary Fig. S2](#).

The combination of all four AGEs has the lowest sensitivity (50.1%) and the highest specificity (94.6%) for the *E. coli* K12 MG1655 data set. Similar observations were made for the other three genomes ([Supplementary Fig. S3](#)): a sensitivity and specificity range of respectively 44.6% and 94.1% for *B. subtilis* 168, 50.5% and 96.5% for *L. lactis* KF147, 49.4% and 96.2% for *L. plantarum* WCFS1 was observed when this consensus prediction of all four AGEs was used.

Interestingly, some combinations of two AGEs provide a higher specificity compared to a combination of three engines; e.g. ISGA, RAST and xBASE (88.9%) versus BASys and ISGA (92.0%) or BASys and RAST (90.8%) ([Supplementary Fig. S3](#)); this trend is observed for *E. coli* K12 MG1655 as well as for the other three genomes ([Supplementary Fig. S3](#)). This leads to the postulate that serially applying consensus predictions of specific combinations of AGEs in order of specificity would allow better overall prediction compared to the alternatives: majority voting, the most reliable AGE only, or consensus predictions based on one AGE combination. As described above,

approximately 50% of the ORF start codons can be correctly predicted by using the consensus of four AGEs. The question remains: how to reliably as possible predict the remaining half of ORF start codons?

A specific path in consensus predictions of different combinations of engines.

In the previous paragraphs it was shown that majority voting has a high false positive rate (Fig. 5). For this reason we are looking for alternative ways to maximize the prediction specificity. Therefore, our objective is to achieve a best-as-possible specificity for start codon prediction based on the twelve genomes. Unfortunately, in order to obtain high specificity we have to accept lower sensitivity, resulting in a lower (re)coverage of ORFs (see also [Supplementary Fig. S3](#)). The consensus prediction of all four AGEs ([Supplementary Figs. S3 and S4](#)) allows accurately predicting half of the actual ORF start codons present in the selected four moderate GC% genomes. If we also want to successfully predict the remaining ORFs with high specificity, they could be predicted with the consensus prediction of different AGE combinations (i.e. using consensus predictions from combinations of two or three AGEs). This process could be done via an iterative algorithm to arrive at an optimal order of consensus predictions. This path of serially applying combinations of AGEs ensures high specificity for prediction of ORF start codons for a genome of interest.

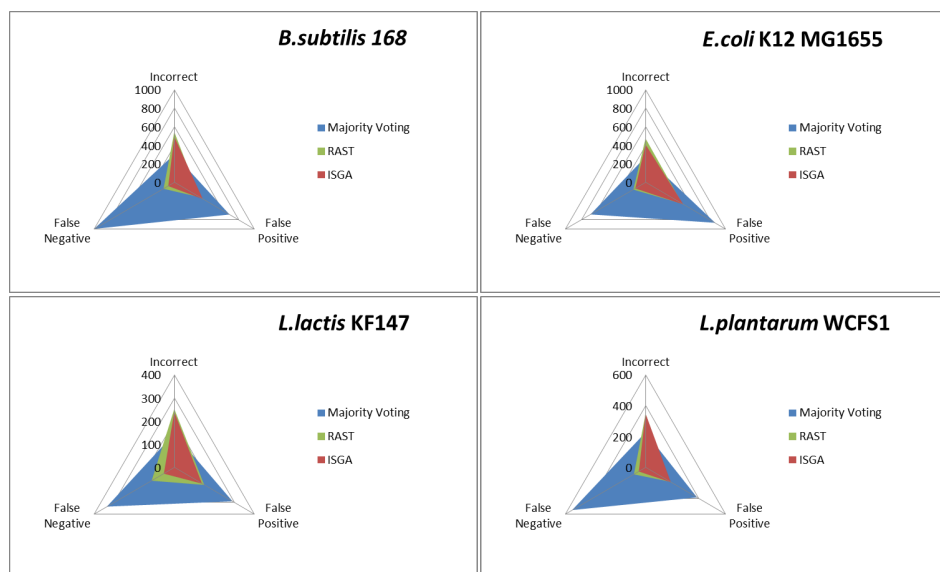


Figure 5. Start codon prediction performance by majority voting versus ISGA and RAST.

This radar-plot illustrates start codon annotation results for four moderate GC% reference genomes *B. subtilis* 168, *E. coli* K12 MG1655, *L. lactis* KF147 and *L. plantarum* WCFS1 determined by majority voting with BASys, ISGA, RAST and xBASE predictions, or by a single AGE. With majority voting a prediction is trusted if more than 50% of the AGEs predict exactly the same start codon coordinate for a given mORF. Predictions were evaluated with the reference genomes ([Table 2](#)). The axis gridlines in all directions are in steps of 100 or 200 ORFs; with incorrect predicted ORF start codons (0°), false positively (FP) and false negatively (FN) predicted ORF start codons (respectively 120° and 240°).

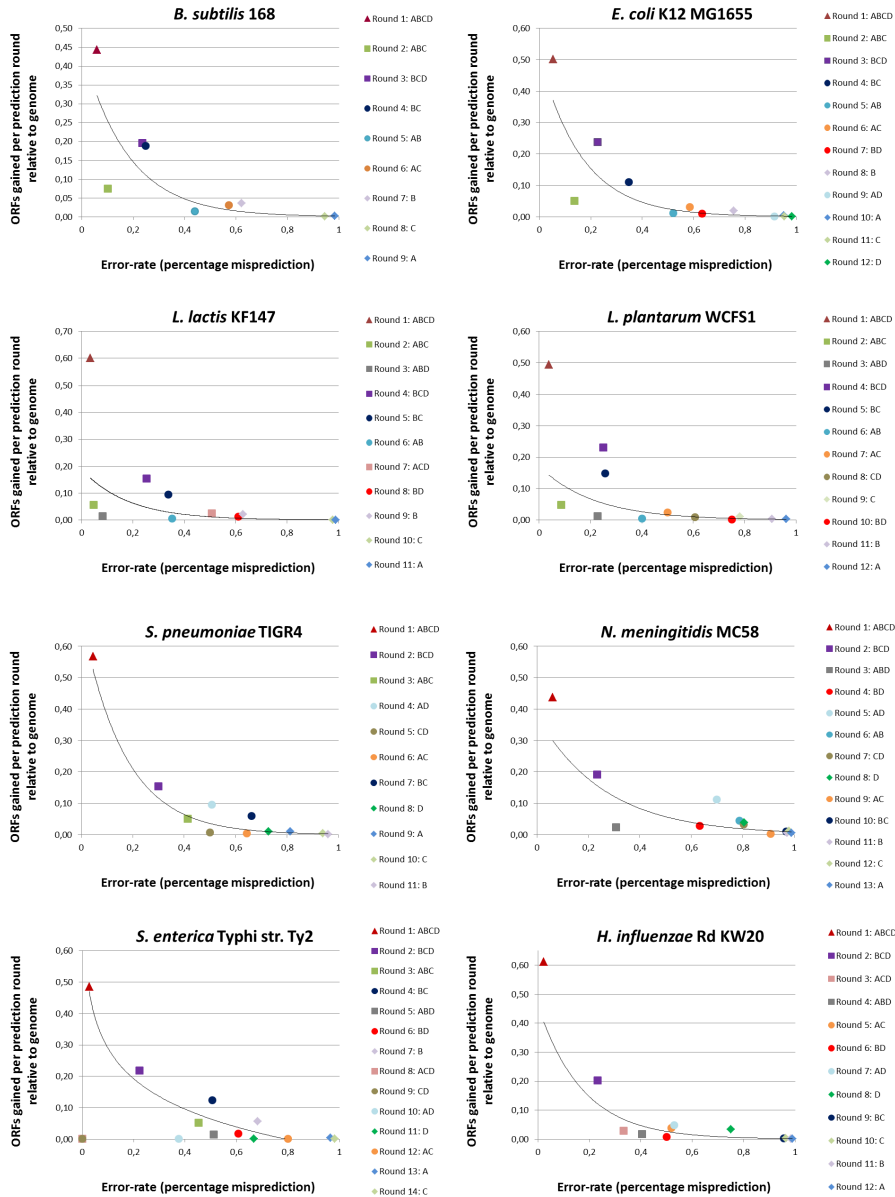


Figure 6. Applying multiple rounds of consensus predictions.

Plotted in these graphs for four moderate GC% prokaryotic reference strains *B. subtilis* 168, *E. coli* K12 MG1655, *L. lactis* KF147 and *L. plantarum* WCFS1 (Table 2) is the error-rate for start codon coordinate prediction (x-axis) versus the new ORFs gained per prediction round (y-axis) for a selected AGE (diamond) or a selected combination of AGEs - two (circle), three (square) or four engines (triangle) - with the lowest calculated error-rate for that concerning round of prediction. ORFs were only taken into account when they were in consensus for their start codon coordinate prediction. Error-rates were calculated as discussed in Methods. A: BASys; B: ISGA; C: RAST and D: xBASE. Note that the trend line is merely for illustrative purposes: it does not signify an actual relation between the data points.

The consensus-path method is explained further in Figure 3 and the Methods. From now on we refer to this optimal order of consensus predictions as consensus-path. From the numbers of correct-, incorrect, and false positive predictions (Fig 4, and Supplementary Figs. S1 and S2), we determined for each AGE and combinations of AGEs (see above) for the twelve reference genomes a sensitivity and error-rate (specificity) in ORF start codon prediction (Supplementary Fig. S3). These error-rates allowed determining for each round of the path an error-rate for each AGE or combination of AGEs (Fig. 3). After each round, the AGE or a combination of AGEs was selected with the highest specificity (Fig. 3). The order of AGE combinations resulting in the highest specificity was quite similar in all eight moderate GC% bacterial genomes (Fig. 6) but less similar for the more extreme GC% genomes (Supplementary Fig. S5).

The consensus-path consisted of using the ORF predictions of five specific combinations of AGEs applied consecutively (Figs. 3 and 6). We limited the number of successive rounds to five as relatively few new ORFs were added in additional rounds (Fig. 6). For instance round 6 added only 152 new ORFs for *E. coli* K12 MG1655 (Fig. 6). Notably, the specificity decreases with each round making the ORF start codon predictions of additional rounds inaccurate (e.g. a 58.6% error-rate for each prediction after round five in *E. coli* K12 MG1655, strongly increasing for the successive rounds) (Fig. 6). Our consensus-path approach enables one to estimate the reliability of prediction, and thus to assess the need for a start codon to be manually curated. An optimal order may work best for a given genome, but it might not work best for another genome. Therefore, we tested if a “conserved” consensus-path was present in the eight moderate GC% genomes (Fig. S6A and S6B).

For the eight moderate GC% genomes the paths appear to be similar (Fig. S6A and S6B). To determine the consensus-path that is most conserved across the eight genomes, we ranked for each genome the AGE combinations based on specificity (Fig. S6A and S6B). A simple formula (see formula 4) was applied. It takes into account the rank, specificity and sensitivity to determine for that AGE (combination) an impact on correctly predicting ORF start codons (Fig. 6; see also Supplementary Fig. S5 for the extreme GC% genomes). The five selected AGE combinations were: BASys-ISGA-RAST-xBASE in the first round, ISGA-RAST-xBASE in the second, BASys-ISGA-RAST in the third, ISGA-RAST in the fourth and BASys-RAST-xBASE in the fifth round. This path was applied to the eight reference genomes (Fig. 7). This resulted in a $9.7 \pm 4.4\%$ gain in precision compared to majority voting and a $1.7 \pm 3.7\%$ gain in precision compared to the single best performing AGE (Fig. 7). Based on the specificity of particular AGE combinations in the five rounds we derive a so-called projected confidence value. It is calculated as the average specificity of a particular AGE combination over the eight genomes (Fig. 6; calculated with formula 2). The overall projected confidence values (i.e. probability of an ORF start codon prediction to be correct) for these AGE combinations were calculated to be: $95.7 \pm 1.4\%$ for a consensus of BASys-ISGA-RAST-xBASE, $75.6 \pm 2.3\%$ for ISGA-RAST-xBASE, $67.6 \pm 26.0\%$ for BASys-ISGA-RAST, $53.1 \pm 19.1\%$ for ISGA-RAST and $42.6 \pm 12.3\%$ for BASys-RAST-xBASE.

Control genome	<i>L. plantarum</i> WCFS1			<i>L. lactis</i> KF147			<i>B. subtilis</i> 168			<i>E. coli</i> K12 MG1655		
Number of genes	3128			2605			4262			4235		
Prediction method	ISGA	MV	Cons. Pr.	ISGA	MV	Cons. Pr.	ISGA	MV	Cons. Pr.	ISGA	MV	Cons. Pr.
Correct ORFs	2728	2351	2650	2313	2122	2221	3691	2956	3519	3692	3266	3556
Incorrect ORFs	350	228	293	240	147	198	494	305	417	410	277	347
False Positive ORFs	189	385	133	138	290	96	355	688	223	470	867	312
Total predicted ORFs	3267	2964	3076	2691	2559	2515	4540	3949	4159	4572	4410	4215
Coverage	98,4%	82,4%	94,1%	98,0%	87,1%	92,9%	98,2%	76,5%	92,4%	96,9%	83,7%	92,2%
Missed ORFs (FN)	1,6%	17,6%	5,9%	2,0%	12,9%	7,1%	1,8%	23,5%	7,6%	3,1%	16,3%	7,8%
Overpredictions (FP)	5,8%	13,0%	4,3%	5,1%	11,3%	3,8%	7,8%	17,4%	5,4%	10,3%	19,7%	7,4%
Correct predicted	83,5%	79,3%	86,2%	86,0%	82,9%	88,3%	81,3%	74,9%	84,6%	80,8%	74,1%	84,4%
Incorrect predicted	16,5%	20,7%	13,8%	14,0%	17,1%	11,7%	18,7%	25,1%	15,4%	19,2%	25,9%	15,6%

Control genome	<i>S. pneumoniae</i> TIGR4			<i>S. enterica</i> Typhi			<i>N. meningitidis</i> MCS8			<i>H. influenzae</i> KW20		
Number of genes	2163			4448			2122			1715		
Prediction method	xBASE	MV	Cons. Pr.	ISGA	MV	Cons. Pr.	xBASE	MV	Cons. Pr.	xBASE	MV	Cons. Pr.
Correct ORFs	1597	1678	1676	3732	3283	3516	1390	1391	1372	1429	1426	1436
Incorrect ORFs	404	175	260	573	416	512	513	285	292	262	136	179
False Positive ORFs	97	410	170	733	1237	427	76	1565	527	30	311	94
Total predicted ORFs	2098	2263	2106	5038	4936	4455	1979	3241	2191	1721	1873	1709
Coverage	92,5%	85,7%	89,5%	96,8%	83,2%	90,6%	89,7%	79,0%	78,4%	98,6%	91,1%	94,2%
Missed ORFs (FN)	7,5%	14,3%	10,5%	3,2%	16,8%	9,4%	10,3%	21,0%	21,6%	1,4%	8,9%	5,8%
Overpredictions (FP)	4,6%	18,1%	8,1%	14,5%	25,1%	9,6%	3,8%	48,3%	24,1%	1,7%	16,6%	5,5%
Correct predicted	76,1%	74,1%	79,6%	74,1%	66,5%	78,9%	70,2%	42,9%	62,6%	83,0%	76,1%	84,0%
Incorrect predicted	23,9%	25,9%	20,4%	25,9%	33,5%	21,1%	29,8%	57,1%	37,4%	17,0%	23,9%	16,0%

Figure 7. Analysis of start codon prediction performance for single engines, majority voting and consensus prediction.

For three methods of processing AGE prediction data: (i) the single, most reliable AGE with highest impact (ISGA or xBASE, this varies between test strains; calculated with [formula 4](#)), (ii) majority voting (MV) and (iii) consensus predictions (Cons. Pr.) are shown for four moderate GC% reference strains from different species: *B. subtilis* 168, *E. coli* K12 MG1655, *L. lactis* KF147 and *L. plantarum* WCFS1. The number of mORFs is indicated which were incorrect-, correct- or false positive predictions according to the reference genomes, by applying each method of prediction. With majority voting a prediction is trusted if more than 50% of the AGEs for a mORF predict exactly the same start codon coordinate for that ORF. With consensus predictions we trust only the in consensus start codon prediction by a combination of AGEs. Also shown are: coverage (percentage correctly and incorrectly predicted ORFs, which are present in the reference genome), the total number of predicted ORFs, the fraction of missed ORFs according to the reference genome (FN; false negatives), and the fraction of over-predicted ORFs (FP; false positives). Bottom row (below the arrow heads): the predicted ORFs and the percentage of correct and incorrect predictions for the three methods.

Consistently mis-predicted ORFs.

Applying any AGE results in ORFs that are either not predicted or incorrectly predicted compared to the reference ORFs. For the four moderate GC% genomes *E. coli* K12 MG1655, *B. subtilis* 168, *L. plantarum* WCFS1, *L. lactis* KF147 we investigated whether these ORFs share common characteristics. Many of the ORFs present in the reference genomes were consistently mis-predicted (i.e. ORF is missed or has an incorrectly predicted start codon) by the AGEs: 208 for *E. coli*, 237 for *B. subtilis*, 103 for *L. plantarum* and 93 for *L. lactis* (Fig. 4, and Supplementary Figs. S1 and S2). If we would be able to understand why these predictions are going wrong, it might be possible to improve current AGEs. We therefore analyzed false negative ORFs that were consistently missed by all four AGEs (93 for *E. coli*, 62 for *B. subtilis*, 37 for *L. plantarum* and 49 for *L. lactis*) and the ORFs with consistently incorrectly predicted start codons by all four AGEs (115 for *E. coli*, 175 for *B. subtilis*, 66 for *L. plantarum* and 44 for *L. lactis*), to determine whether they share characteristics that might explain their incorrect start codon prediction. Smaller reference ORFs (<750 nucleotides; nt) are significantly more often missed by all AGEs (false negatives) (Supplementary Fig. S7; chi-square p -values: <0.0001 for any of these four genomes). This was not the case for ORFs with consistently incorrectly start codons (Supplementary Fig. S7; <750 nt chi-square p -values: *E. coli*: 0.70; *B. subtilis*: 0.07; *L. plantarum*: 0.48; *L. lactis*: 0.36). Possibly, AGEs have a threshold at a specific ORF length which could explain false negative ORF predictions. The same phenomenon is observed for ORFs missed by our consensus-path prediction (i.e. ORFs not called after five rounds of consensus-paths). Compared to the corresponding reference genomes, smaller ORFs (i.e. <750 nt) are significantly more often missed (chi-square p -values <0.0001 for any of these four genomes).

Apart from ORF length, we tested these missed and incorrectly predicted ORFs for overrepresentation in other functional data (predictions) such as protein functionality (COG [187], DAVID database [188, 189], Pfam [175]), and subcellular protein localization (PSORTdb [190]). However, we did not find any significant overrepresentation of these functional annotations in the subset of incorrect predicted ORF start codons, nor in the false negative- and/or positive ORF subset (data not shown).

Discussion

In order to improve start codon prediction in bacterial genomes by current AGEs we present an alternative method to majority voting [180]. We present a consensus-path for the prediction of bacterial ORFs by combining AGEs that could save researchers time, as manual curation of ORF start codons is tedious. As specificity in ORF start prediction is leading in ORF start curation the path is largely based on the specificity of AGEs (combinations). For eight moderate GC% genomes, we observe similar paths allowing us to postulate a generalized consensus-path for moderate GC% genomes. This consensus-path is specific for the four AGEs under study and might change as a result of changes within the ORF prediction procedures used in the AGEs and certainly when other AGEs are considered. Nevertheless, the application of an optimal path allows gaining sensitivity while maintaining a high specificity in ORF start codon

prediction. In our case study, the consensus-path prediction is performed by assessing the consensus predictions of serially applying five AGE combinations: (i) BASys-ISGA-RAST-xBASE, (ii) ISGA-RAST-xBASE, (iii) BASys-ISGA-RAST, (iv) ISGA-RAST and (v) BASys-RAST-xBASE. Compared to majority voting we observe with our consensus path method an increase of 9.7 ± 4.4 % of predicted ORF start codons for which no further manual curation would be required. This equals easily hundreds of genes. Because application of consensus prediction is straight-forward, a researcher could without complex procedures benefit from an increase in annotation quality. Importantly, based on a novel projected confidence value one can determine ORFs for which likely an incorrect start codon prediction has been made. These ORFs can subsequently be targeted for manual curation. In addition, this information could be used to improve current ORF start codon prediction services and tools. Especially ORFs acquired in the fifth round of consensus-path prediction (*E. coli*: 99; *B. subtilis*: 97; *L. plantarum*: 58; *L. lactis*: 63) are notoriously difficult to predict and likely require manual curation.

The results presented in this case study have been achieved with the ORF start codon predictions of web-based genome annotation services that are free to use. Although our conclusions are based on four bacterial AGEs (Table 1) and eight moderate GC% bacterial genomes, we believe that this consensus-path based on the four AGEs can be applied to other moderate GC% organisms. Moreover, the concept of our consensus-path can be applied to other AGEs and other moderate GC% genomes. Compared to majority voting, the consensus prediction of multiple AGEs over multiple rounds of prediction results in 9.7 ± 4.4 % more correct predictions and 12.7 ± 4.8 % less false positive and 6.9 ± 0.5 % less false negative predictions (Fig. 7). Compared to the single, most reliable engines with the highest impact (either ISGA or xBASE) there is a slight gain in correct start codon prediction with 1.7 ± 3.7 % more correct predictions. However, 1.8 ± 2.3 % more false positives and 5.6 ± 1.8 % more false negatives were observed because some ORFs are over- or under predicted by a combination of AGEs but not by another single AGE. In any case, after applying the consensus-path approach, one is still able to supplement the already acquired predictions with those of a single AGE. However, these added predictions will generally be more error-prone and no high level of projected confidence can be assigned to them.

In the coming years, automated (genome) annotation processes will keep continuing to improve to the point that, ideally, barely any manual curation will be necessary. Currently, however, researchers have to be careful with interpreting AGE ORF start predictions. From this study we conclude that every AGE has its own unique strengths and weaknesses, likely related to the underlying tools and protocols used (Supplementary Table S1). Therefore, it could be rewarding to combine AGEs in order to benefit from comparative annotation strategies; and thus to further increase the specificity and sensitivity of ORF predictions for a given genome.

Methods

Genome sequences and annotations

The predicted ORF start codon coordinates were evaluated by validation with ORFs from reference genomes from twelve well-studied strains of different bacterial species. This set of twelve genomes consists of eight moderate GC% genomes: *E. coli* K12 MG1655 (50% GC) [191], *B. subtilis* 168 (44%) [192], *L. plantarum* WCFS1 (44%) [193], *L. lactis* KF147 (35%) [194], *S. pneumoniae* TIGR4 (40%) [195], *S. enterica* subsp. *enterica* serovar Typhi str. Ty2 (52%) [196], *N. meningitidis* MC58 (52%) [197] and *H. influenzae* Rd KW20 (38%) [198]. In addition, four more extreme GC% genomes were analyzed: *M. tuberculosis* H37rv (66% GC) [199], *M. mobile* 163K (25%) [200], *P. putida* KT2440 (61%) [201] and *S. coelicolor* A3(2) (72%) [202] (Table 2).

Automated genome annotation engines (AGEs)

For this study four microbial AGEs were selected: BASys [203], ISGA [204], RAST [205] and xBASE [206, 207] (Table 1). Although there are other excellent online AGEs available, we selected these four based on their relatively short queues and processing time, and on their easily exportable annotation data. These engines are all online, free-of-charge initiatives for non-profitable scientific research. FASTA-files of the reference genomes (Table 2) were uploaded to the engines, and default server settings were applied with exception of the following non-trivial options: for RAST: “automatically fix errors”, “fix frameshifts” and “backfill gaps” were selected; for ISGA: the standard/ suggested “ISGA Prokaryotic Annotation” pipeline of February 2011 was employed. ISGA and xBASE use GLIMMER version 3 for its ORF prediction, BASys uses GLIMMER version 2.1.3. RAST uses a custom ORF caller called “RAST” (which was used in this study); however, RAST also allows the use of GLIMMER version 3.

Comparison of ORF-predictions by defining meta ORFs

For comparing start codon coordinate positions we used in-house Perl scripts for analyzing AGE annotation data. This software package can be downloaded from <https://trac.nbic.nl/companion/> and is open for public use. This tool – which we have named COMPANION: an acronym for comparative genome annotation – evaluates the start- and stop-codon coordinates of (i) ORFs predicted by the different AGEs (Table 1), and (ii) ORFs provided by the reference genome (Table 2). It then groups these ORF coordinates into equivalent ORFs based on position and length of these ORFs. We name these ORF representatives meta ORFs (mORF). ORFs of which the start or stop codon differs less than 10% (of the ORF length of the shortest ORF of the two ORFs being compared) compared to its reference ORF start or stop codon were grouped into the same mORF prediction (Fig. 1). Perl scripts were used to extract from this table for each mORF the start codon for the respective engines. As both predictions and reference genome annotations are merged, for each mORF the predictions can be compared to the (if available) high quality reference ORF start codon annotations. The COMPANION tool was also used for statistical analysis and for determination of a consensus-path as mentioned below.

Statistical analysis on meta ORFs

For each reference genome, the predicted start codons from the various AGEs were matched to those of the corresponding ORFs in the reference genome (Table 2 and Fig. 2). If a specific ORF from the reference genome was not represented by an ORF predicted by an AGE (combination), we defined that ORF start codon prediction to be false negative for those AGEs. If an ORF was predicted by an AGE but it was not present in the reference genome, we defined that ORF as a false positive prediction. When a predicted ORF start codon matched that in the reference genome (hence, start codon predictions from the same mORF; Fig. 1), it was considered a correct and therefore true positive prediction. In case a predicted ORF was present in a reference genome (same mORF) but its start codon did not match that of the reference genome, it was considered an incorrect prediction.

Error-rate was calculated as follows (formula 1):

$$\text{Error rate} = \frac{\sum (\text{Incorrect Predictions} + \text{False Positive Predictions})}{\sum (\text{True Positive Predictions} + \text{Incorrect Predictions} + \text{False Positive Predictions})}$$

Specificity (a number between 0 and 1) was subsequently calculated (formula 2):

$$\text{Specificity} = (1 - \text{Error rate})$$

Sensitivity was calculated as follows (formula 3):

$$\text{Sensitivity} = \frac{\sum \text{True Positive Predictions} + \text{Incorrect Predictions}}{\text{Number of Ref. Genome ORFs}}$$

Determination of the consensus-path

Combining specificity, sensitivity and rank (where rank is the order in which AGE (combinations) are predicted; see Fig. 6, and Supplementary Fig. S5), enables determining the impact of a certain prediction method (i.e. a single AGE, majority voting or a consensus-path) on the correct prediction of subset of ORF start codons under study. To account for AGE (combinations) that perform less on specific genomes, we incorporated the rank of prediction specificity into our formula for determining the *impact* for a prediction method. This was to account for that the order of prediction by an AGE (combination) is crucial in our approach (Fig. 3B). Based on the impact of a particular AGE (combination) for a particular genome, we can calculate the *average impact* over eight moderate GC% genomes of that AGE (combination).

This impact for a prediction method was calculated as follows (formula 4):

$$\text{Impact} = (1 - \text{Specificity}) * \frac{\text{Sensitivity}}{\text{Rank}}$$

This formula 4 enables us to - for each genome - determine an impact value for each AGE (combination). This allows establishing a *general consensus path* (Fig. 3) by taking the highest impact values for AGE (combinations) over the selected eight moderate GC% genomes (Fig. S6A and S6B). This results in the average impact value.

Because we analyze genomes with trusted ORF start annotations (Table 2), we can derive *projected confidence values* for the selected AGE (combinations) part of the consensus-path. These are estimations of the probability of making correct ORF start codon predictions (see formula 2) when applying certain AGE (combinations) to a new genome. Therefore, we are able to assign to each AGE (combination) its own *general projected confidence value*, which is an average of all projected confidence values for an AGE (combination) over the eight moderate GC% genomes.

Supplementary Figures

Figure S1. Variation in AGE ORF start- and stop codon predictions for four moderate GC% bacterial genomes. (PDF)

(online) Figure S1 of online repository at :
<https://doi.org/10.1371/journal.pone.0063523.s001>

Figure S2. Variation in consensus AGE ORF start codon predictions for four moderate GC% bacterial genomes. (PDF)

(online) Figure S2 of online repository at :
<https://doi.org/10.1371/journal.pone.0063523.s002>

Figure S3. AGE annotation prediction specificity and ORF recovery. (PDF)

(online) Figure S3 of online repository at :
<https://doi.org/10.1371/journal.pone.0063523.s003>

Figure S4. Percentage ORFs incorrect when AGEs have a consensus start codon coordinate prediction. (PDF)

(online) Figure S4 of online repository at :
<https://doi.org/10.1371/journal.pone.0063523.s004>

Figure S5. Applying multiple rounds of consensus prediction for four more extreme GC% genomes. (PDF)

(online) Figure S5 of online repository at :
<https://doi.org/10.1371/journal.pone.0063523.s005>

Figure S6. Selecting the most optimal consensus-path. (PDF)

(online) Figure S6 of online repository at :
<https://doi.org/10.1371/journal.pone.0063523.s006>

Figure S7. ORF length bias in false negatively predicted ORFs. (PDF)

([online](#)) Figure S7 of online repository at :
<https://doi.org/10.1371/journal.pone.0063523.s007>

Supplementary Tables

Table S1. Key functionalities in various AGEs. (PDF)

([online](#)) Table S1 of online repository at :
<https://doi.org/10.1371/journal.pone.0063523.s008>

CHAPTER 3

MANUSCRIPT IN SUBMISSION

¹ Center for Molecular and Biomolecular Informatics (CMBI), Radboud Institute for Molecular Life Sciences (RIMLS), Radboud University Medical Center (Radboudumc), Nijmegen, The Netherlands.

² Department of Dermatology, RIMLS, Radboudumc, Nijmegen, The Netherlands.

³ NIZO, Ede, The Netherlands.

CHAPTER 3

TaxPhlAn: A GENERIC WORKFLOW FOR SINGLE LOCUS SEQUENCE TYPING (SLST) DESIGN AND SUBSPECIES CHARACTERIZATION OF MICROBIOTA

A CASE-STUDY OF BEYOND SPECIES-LEVEL PROFILING OF STAPHYLOCOCCUS IN ATOPIC DERMATITIS

Thomas H.A. Ederveen ^{1,3}

Jos P.H. Smits ²

Karima Hajo ¹

Saskia van Schalkwijk ³

Tessa A. Kouwenhoven ²

Sabina Lukovac ³

Michiel Wels ^{1,3}

Ellen H. van den Bogaard ²

Joost Schalkwijk ²

Jos Boekhorst ^{1,3}

Patrick L.J.M. Zeeuwen ²

Sacha A.F.T. van Hijum ^{1,3}

ABSTRACT

We present TaxPhlAn, a new method and bioinformatics pipeline for design and analysis of single-locus sequence typing (SLST) markers to type and profile bacteria beyond the species-level in a complex microbial community background. TaxPhlAn can be applied to any group of phylogenetically-related bacteria, provided that it has at least a dozen sequenced reference genomes available. As TaxPhlAn requires the SLST targets identified to fit the phylogenetic pattern as determined through comprehensive evolutionary reconstruction of input genomes, TaxPhlAn allows for the identification and phylogenetic inference of new biodiversity. Here, we present a clinically relevant case study of high-resolution *Staphylococcus* profiling on skin of atopic dermatitis (AD) patients. We demonstrate that SLST enables profiling of cutaneous *Staphylococcus* members at (sub)species level, is cost-effective and especially by combining with 16S rRNA gene sequencing, and provides higher resolution than current 16S-based sequencing techniques. With the higher discriminative ability provided by our approach, we further show, for the first time, that the presence of *Staphylococcus capitis* on the skin together with *Staphylococcus aureus* associates strongly with AD disease.

Background

Consortia of bacteria are found in many niches and there is increasing evidence of bacterial involvement in health and disease [208]. Bacterial diversity is considerable, and the current challenge lies in determining which bacteria and corresponding functionalities are relevant in a given ecological niche. In order to perform follow-up experiments in *in vitro* or animal models with candidate bacteria that are assumed to be important for a given niche, resolution down to the strain level is desirable [67]. Traditionally, bacterial occurrence is determined through culture-based methods, with subsequent isolate identification by a plethora of available genotyping methods: chemotaxonomy [209], DNA fingerprinting [22, 23], (quantitative) PCR [210], mass spectrometry [211] and genome sequencing [212]. One well-established genotyping method for bacterial strains is multi-locus sequence typing (MLST) [213], which is based on sequence variety in a number of marker core genes revealed by qPCR and Sanger sequencing. However, most of the aforementioned techniques have limited resolution, and placing a novel genotype in its correct phylogenetic context is usually difficult. Notably, single strain full genome sequencing does not suffer from these drawbacks, but is more costly, labor intensive, and again, not easily compatible with high throughput applications [214]. Furthermore, DNA fingerprinting, MLST and genome sequencing do not allow for the measurement of bacterial abundances, and not all bacteria can be cultivated efficiently. Currently, sequencing-based culture-free analysis of complex microbial consortia as a whole can be performed with approaches such as 16S rRNA marker gene sequencing (16S metataxonomics) or shotgun metagenomics [215]. 16S metataxonomics focuses on 16S rRNA genes, which are universally present in all bacteria, is relatively cheap and data analysis is straightforward [216]. Depending on the primer set, 16S allows for confidently profiling most bacteria down to the genus level [40, 41]. Metagenomics sequences in principle all free DNA present in a sample, allowing the classification of sequences at high taxonomical resolution as well as determining functionality present in a microbiome [52, 57]. Recently, computational analysis methods were adopted for strain-level classification of metagenomics sequencing data, such as ConStrains [59], PanPhlAn [60] and StrainPhlAn [61]. However, obtaining sufficient biomaterial for metagenomics as well as generating and analyzing the datasets requires significant resources. Furthermore, aforementioned strain-level classification methods have difficulty with confidently detecting bacterial entities under a relative abundance level threshold of approximately 1% in a metagenomics sample [59]. Marker gene sequencing approaches perfectly allow for profiling bacteria that are below 0.1% relative abundance (i.e. 1 in 10,000 reads). Recently, single locus sequence typing (SLST) has been described for determining down-to strain level identification of *Propionibacterium acnes* human isolates by Scholz *et al.* [25, 217]. Other applications of SLST have been reported for *Lactobacillus plantarum* strain tracking in human gut [26] and industrial biofilms [27], and for *Staphylococcus (aureus)* profiling on skin of atopic dermatitis patients during therapeutic intervention with coal tar (Smits *et al.*, manuscript in preparation).

Until now, an automated bioinformatics pipeline to devise sequence-based SLST screening tools for specific microbes in complex microbial communities is lacking, and currently requires mostly manual searches with a lot of hands-on time. We here present

TaxPhlAn for SLST-based **T**axonomy **P**hylogenetic **A**nalysis: a method and workflow to create and use single locus marker sequences of orthologous genes to profile specific bacterial taxa at and beyond the species level (as illustrated in [Fig. 1](#), based on a toy example with *Pseudomonas*). The TaxPhlAn bioinformatics pipeline finds SLST marker genes based on a set of reference genomes provided by the user in module A of the pipeline ([Supplementary Fig. S1](#)). Candidate SLST regions selected by the TaxPhlAn pipeline balance a trade-off between sequence conservation, which is important for PCR primer design and identification, and sequence variation within single-copy genes shared among all genomes-of-interest to allow for a certain degree of discrimination between genomes. SLST markers are designed to follow the actual phylogeny of the considered strains, and additionally scored based on characteristics such as discriminative resolution, marker length (i.e. amplicon size) and level of conservation of the marker in the taxon-of-interest. TaxPhlAn automatically designs primers for candidate markers, to generate SLST amplicons by PCR for subsequent marker gene sequencing. Finally, next generation sequencing (NGS) data containing SLST marker sequences can be processed in an automated workflow additionally offered by TaxPhlAn in module B of the pipeline ([Supplementary Fig. S2](#)). This oligotyping-based method analyses SLST amplicons sequenced by NGS, and, for each biological sample, calculates the relative abundances of SLST alleles corresponding to representative (sub)species of strains.

TaxPhlAn is available as a Python/Perl command-line application, and is accessible through a pre-configured, plug-and-play Docker virtual machine which is supplied at the Docker Hub repository <https://hub.docker.com/r/ederveen/taxphlan/>. TaxPhlAn is supplied with test datasets, properly documented with an extensive user manual, and can conveniently be applied by anyone with limited bioinformatics knowledge.

The bacterial genus of *Staphylococcus* cannot reliably be profiled with high resolution using conventional 16S marker gene sequencing primers [41], warranting the use of SLST. In this manuscript, we showcase a clinically relevant application of TaxPhlAn by profiling the presence of cutaneous bacteria (with 16S) as well as by targeted *Staphylococcus* species (with SLST) present in atopic dermatitis (AD) versus healthy skin of human volunteers. In short, our results show real-life and cost-effective application of the presented SLST methodology for profiling *Staphylococcus*, and demonstrate significant added value of SLST compared to 16S.

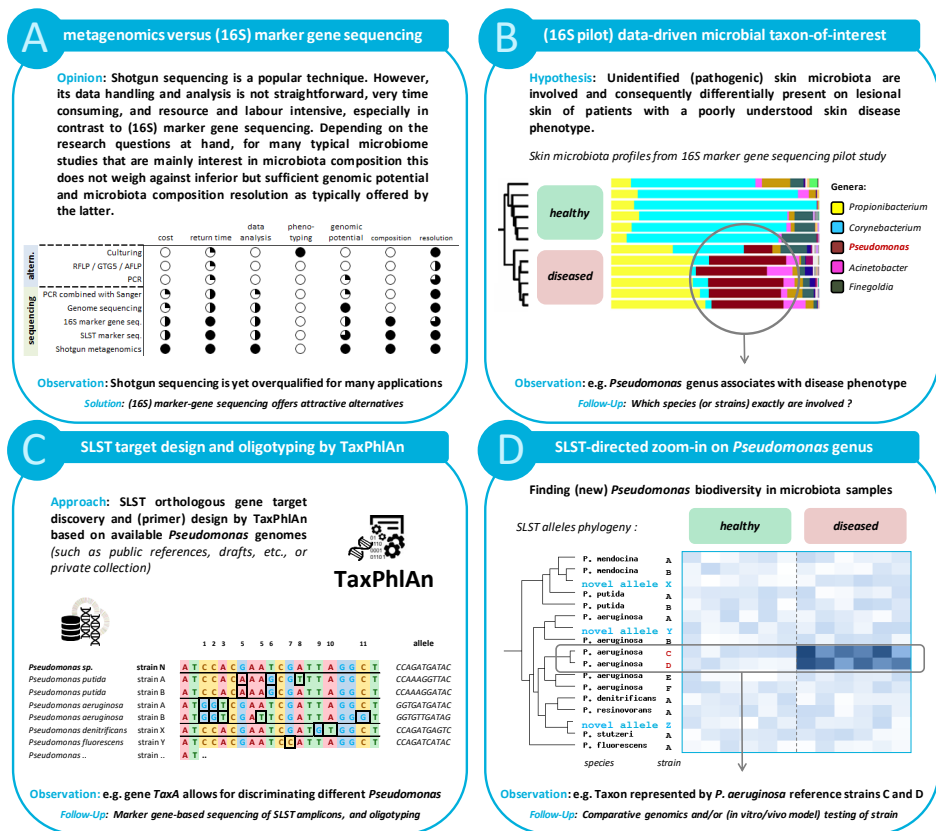


Figure 1. SLST allows for high-resolution discrimination of different bacterial phylotypes, (sub)species and strains.

(1A) Overview of current sequencing-based and alternative methods for microbiota identification and classification, including their (dis)advantages. Table legend as follows: estimation and indication of costs, return time and data analysis complexity; methods that allow for accurate phenotyping; amount of information retrieved with regard to genomic (functional) potential, taxonomic composition and resolution. The symbols represent open versus closed pie charts, meaning: low vs. high, fast vs. slow, simple vs. complex, etc. (1B) Hypothetical dataset with 16S sequencing microbiota data of “healthy” and “diseased” subjects who suffer from an exemplary skin disease. The colored bars represent skin microbiota composition, and is clustered on microbiota profiles. In this hypothetical example, the bacterium *Pseudomonas* associates with skin disease state of the volunteers. (1C) *Pseudomonas* genomes are collected and analyzed with the bioinformatics tool TaxPhlAn in order to search for candidate SLST targets. The goal of TaxPhlAn is to find genetic regions that allow for perfect discrimination of *Pseudomonas* biodiversity. Shown is a hypothetical SLST region (aligned DNA sequences) that allows for discriminating different *Pseudomonas* (sub)species. (1D) Hypothetical example of how the SLST target as identified in 1C could allow for high-resolution *Pseudomonas* typing, and thereby to determine the (sub)species or strain-level bacteria that associate with disease. Panel represents an heatmap with data of identified *Pseudomonas* biodiversity, and their relative abundance as measured for each sample. Available (clinical) isolates of these bacteria can be subsequent candidates for follow-up studies.

Results

We have developed a bioinformatics pipeline that requires reference genomes as input, and then automates the discovery and design of gene targets and corresponding primer sets for SLST-based profiling of specific microbiota in complex communities (TaxPhlAn Module A: [Supplementary Fig. S1](#)). Primers designed by TaxPhlAn Module A can be applied to any biological sample to generate SLST marker gene sequences. We therefore additionally provide a data analysis workflow based on oligotyping to straightforwardly analyze SLST data (TaxPhlAn Module B: [Supplementary Fig. S2](#)). This analysis pipeline allows going from raw SLST marker gene sequencing data to a compositional microbiota-to-sample matrix, with corresponding heatmap visualization of the data and clustering of study samples.

TaxPhlAn performs automated discovery of gene targets for application in SLST.

We validated TaxPhlAn by evaluating four different bacterial taxa at the genus or species level that are typically found on two clinically relevant human niches: skin and gut. These are *Staphylococcus*, *Propionibacterium acnes*, *Bifidobacterium* and *Escherichia / Shigella*, representing both Gram-positive and Gram-negative genomes with variable sizes and GC-content ([Table 1](#), and [Supplementary Tables S1-S4](#) for more detailed information on each individual genome). For each of these four bacterial taxa we were able to successfully pinpoint multiple genic SLST targets able to distinguish (sub) species with significantly higher resolution than currently feasible with alternative high-throughput metataxonomics methods such as Illumina 16S marker gene sequencing on variable regions V1-V2 and V3-V4 ([Table 2](#), and [Supplementary Table S5](#) for primer information, and [Supplementary Tables S6-S9](#) for TaxPhlAn reports).

Table 1. Datasets with bacterial genomes used for TaxPhlAn benchmarking.

For benchmarking we selected four bacterial taxa that are associated with the two clinically relevant human microbial niches of gut and skin. The datasets represent both Gram-positive and Gram-negative bacteria from different taxonomical levels, and with variable genome sizes and GC-content. See [Supplementary Tables S1-S4](#) for more details. SD: standard deviation; Mb: mega base pairs.

Taxon Name	Taxonomy Level of Input	Gram- Staining	Human Niche	# Genomes selected	Average \pm SD Genome Size (Mb)	Average \pm SD Genome GC-content (%)
<i>Bifidobacterium</i>	genus	positive	gut	261	2.30 \pm 0.27	60.1 \pm 2.0
<i>Escherichia / Shigella</i>	supra-genus	negative	gut	200	4.91 \pm 0.35	50.6 \pm 1.1
<i>Propionibacterium acnes</i>	species	positive	skin	123	2.50 \pm 0.03	60.1 \pm 0.1
<i>Staphylococcus</i>	genus	positive	skin	200	2.59 \pm 0.20	33.0 \pm 1.4

Table 2. Performance of a default TaxPhlAn run versus common 16S regions based on the number of unique clusters identified.

The number of reference genomes used on the benchmark can be found in [Table 1](#), for an exact overview of these genomes see [Supplementary Tables S1-S4](#). The number of unique SLST clusters found for each taxon-of-interest is based on the average of the top 10 SLST candidates from a default TaxPhlAn run (see also [Supplementary Tables S6-S9](#)). The number of 16S clusters found for these same taxa is based on primers 27F and 338R for V1-V2, and 357F and 802RV2 for V3-V4 (V for 16S variable region) ([Supplementary Table S5](#)). Methods used for determining the number of clusters are either allele-based with a taxon-specific Shannon diversity index (SDI) threshold (in brackets the average number of informative SNP), or OTU-based with 97% identity threshold (accepted default in the microbiome field [43]). SDI thresholds were set to 0.6, except for *Propionibacterium acnes* for which we set the SDI to 0.2, because we observed less sequence variation on the level of species (for more information about SDI, we refer to the [Supplementary Methods](#)). Values in brackets represent the average number of informative positions for an allele. Coverage is defined as the percentage of genomes identified with the primers tested. SD: standard deviation, OTU: operational taxonomic unit, SDI: Shannon diversity index. * note the very low coverage of V1-V2 primers for *Bifidobacterium*, in contrast to that of the V3-V4 primers.

Taxon Name	# SLST clusters			# 16S V1-V2 clusters			# 16S V3-V4 clusters		
	OTU	SDI	Coverage	Coverage			Coverage		
	Average \pm SD	Average \pm SD	(%)	OTU	SDI	(%)	OTU	SDI	(%)
<i>Bifidobacterium</i>	42 \pm 5	87 \pm 8 (62)	95.8	10 *	12 (56) *	5.0 *	21	50 (26)	96.2%
<i>Escherichia / Shigella</i>	12 \pm 6	29 \pm 10 (11)	97.7	6	16 (8)	71.5	3	0 (0)	74.0%
<i>Propionibacterium acnes</i>	2 \pm 1	7 \pm 2 (16)	99.9	1	0 (0)	88.6	2	0 (0)	96.7%
<i>Staphylococcus</i>	24 \pm 3	51 \pm 6 (55)	91.1	12	41 (18)	94.0	5	9 (4)	100%

Orthologous, single-copy core genes allow for detection and typing of novel biodiversity.

TaxPhlAn relies on single-copy orthologous genes for discovery of SLST candidates ([Supplementary Fig. 1, Phase III](#)), in contrast to targets in intergenic regions, or orthologous genes that are not universally present. The advantage of this key SLST characteristic is that it makes it more likely that novel biodiversity is correctly identified, as we expect the SLST candidate genes to be universally present in the target taxon, and we expect the observed sequence variation in the SLST genes to reflect phylogeny. We tested this assumption *in silico* by running TaxPhlAn with random subsamples of the total genome datasets, for the four representative bacterial taxa as listed previously in [Table 1](#). We undertook jackknifing by random subsampling of n -genomes from the total dataset with a step-size of 12 genomes, and running TaxPhlAn with default pipeline settings on that subset. Hereafter, the top 10 SLST candidates (i.e. primer sets) as reported by the pipeline were subjected to an *in silico* PCR on all available genomes in the entire dataset. The number of total genomes identified with an SLST target designed on a subset of 12 genomes did not improve significantly after addition of more genomes to this training set, nor did the number of unique SLST sequences identified ([Fig. 2](#)).

To confirm that high-scoring SLST targets as reported by the pipeline do indeed follow phylogeny based on full-genome information and hence contain an evolutionary conserved signal, we looked at the prime SLST candidates of each bacterial genome reference dataset. We observed that for every taxon-of-interest the variable region sequences of the prime SLST candidate correlated very well with the actual phylogenetic distances between genomes (based on SNP positions in the bacterial core genomes).

This is illustrated by the unique SLST alleles capable of discriminating between different taxonomical clades, and as supported by their corresponding Spearman correlation values of 0.78 rho on average (for a projection of these alleles to the actual phylogenetic tree, of each prime SLST candidate, see [Supplementary Figs. S3-S6](#); for all predicted SLST targets for the reference runs and their phylogenetic distances correlation values see [Supplementary Tables S6-S9](#)). This data shows that TaxPhlAn allows for accurate phylogenetic placement of known taxa, and for inference of novel biodiversity, even when only a small number of genomes is available for design of an SLST target.

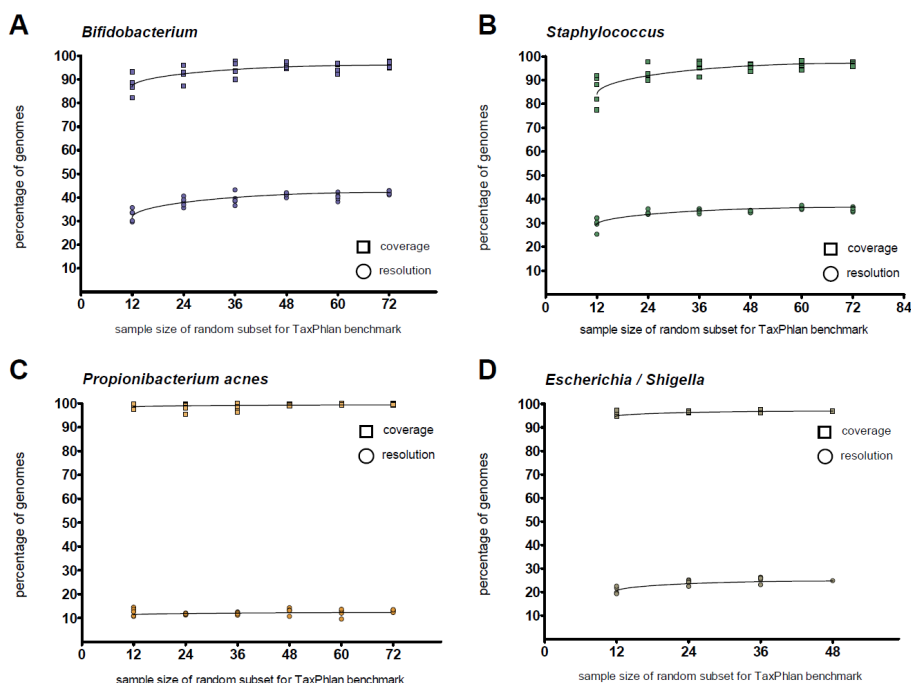


Figure 2. Reducing the number of input genomes for a TaxPhlAn run does not diminish performance of SLST candidates.

We tested the stability, performance, and minimum number of input genomes required for the TaxPhlAn pipeline ([Supplementary Fig. S1](#); Module A) by running the program with random subsets, and with variable sample sizes, of the total benchmark datasets. We determined for each run the number of unique SLST sequences (resolution, i.e. the number of phylotypes that can be distinguished using this amplicon in square symbols) and the total number of genomes for which an *in silico* PCR hit was predicted by PrimerProspector (coverage, in circle symbols). For each run, averages of the top 10 candidates were taken, and each run with each its unique random subset of genomes was plotted in the graph (i.e. 5 replicates). The maximum number of genomes in the benchmark datasets is 261 for *Bifidobacterium* (with datapoints in blue), 200 for *Staphylococcus* (green) and *Escherichia / Shigella* (brown) and 123 for *Propionibacterium acnes* (orange).

TaxPhlAn significantly improves *Staphylococcus* profiling compared to 16S in clinical practice.

The Gram-positive bacterium *Staphylococcus* represents a historically relevant genus in relation to AD. *Staphylococcus aureus* associates strongly with AD disease onset

Typical SLST Study Design Workflow

- case study of *Staphylococcus* involvement in AD -

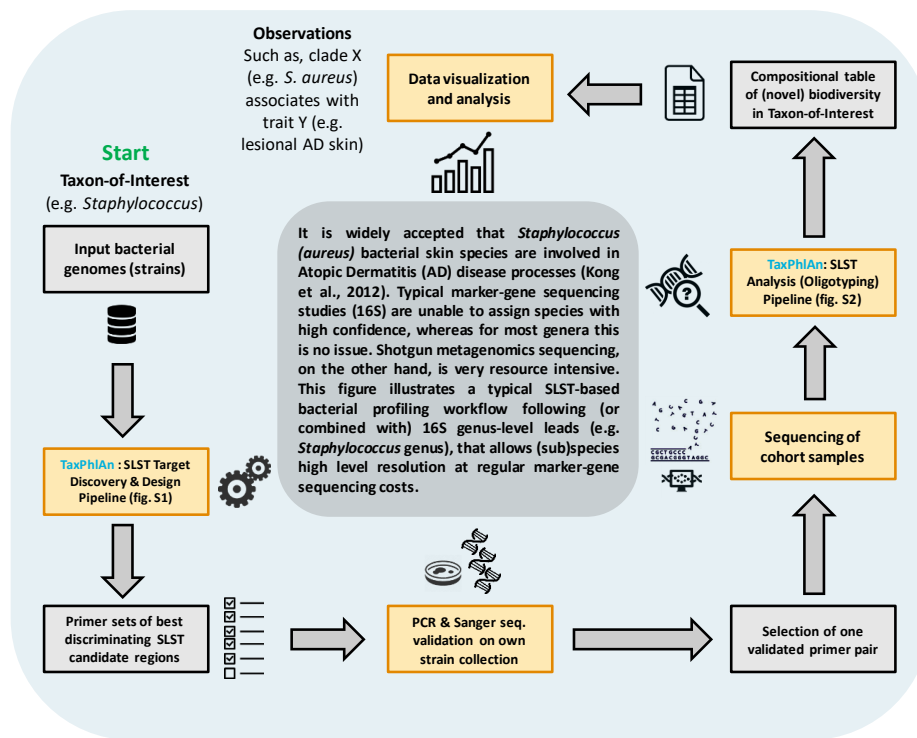


Figure 3. Workflow for SLST typing (graphical abstract of this AD substudy).

This flow chart summarizes our proposed best practices for SLST. For this, we discuss the example of *Staphylococcus* involvement in AD as an SLST use case. It involves the selection of representative genomes for any taxonomy-of-interest, and analyzing them in a straightforward fashion with TaxPhlAn for identification and design of suitable SLST targets for optimal discrimination between input genomes. A small selection of the best-discriminating SLST candidates (i.e. primer pairs for PCR) will be tested in the laboratory by PCR on relevant strain collections. SLST candidates for which primer pairs survive both *in silico* and laboratory selection procedures can in principle be adopted as marker genes for typing of microbiota in next-generation sequencing (NGS) initiatives. Our SLST oligotyping pipeline allows for the analysis of SLST reads from NGS data, including typing of known and identification of unknown biodiversity, and reporting of results in comprehensible format, including automated visualization and clustering of samples. Altogether, this entire process facilitates the potential discovery of study parameters that associate with high-resolution microbiota data. Importantly, in order to “close the experimental circle”, microbiota leads from SLST data can be adopted in an attempt to unravel or further understand underlying biological processes. This is mainly based on pinpointing in high-resolution the most relevant microbial (sub)species, or even strains, for application in follow-up experiments or studies.

and severity [218–221], whereas *S. epidermidis*, amongst others, is considered a skin commensal [222]. However, in general, the 16S rRNA gene does not allow for confident classification of bacteria to the level of species [216]. TaxPhlAn was designed to increase the resolution of bacterial profiling in clinical practice. We therefore compared SLST with 16S rRNA gene sequencing for classification of *Staphylococcus* species. In order to corroborate application and feasibility of the TaxPhlAn SLST method and corresponding proposed experimental workflow (Fig. 3), and to demonstrate clinical relevance, we enrolled from our hospital dermatology clinic a group of patients diagnosed with AD ($n = 5$), and healthy controls (HC) without a history of skin diseases ($n = 9$) (Supplementary Table S10). The skin microbiota on the left and right antecubital fossa (inner elbow) of these human volunteers was obtained by a standardized wet swabbing method, from normal and lesional skin for HC volunteers and AD patients, respectively. Skin bacteria were typed by 16S metataxonomics (V3–V4), and *Staphylococcus* species were determined by SLST marker gene sequencing (Supplementary Table S5 for 16S and SLST primer information). To confirm that we have a representative AD study cohort, we first analyzed the skin microbiota on the level of genus by 16S (Supplementary Table S11 for 16S sequencing read statistics), and found that *Staphylococcus* dominates skin of AD patients at the expense of *Propionibacterium* (Fig. 4A, Supplementary Fig. S9, which is in line with literature [220]).

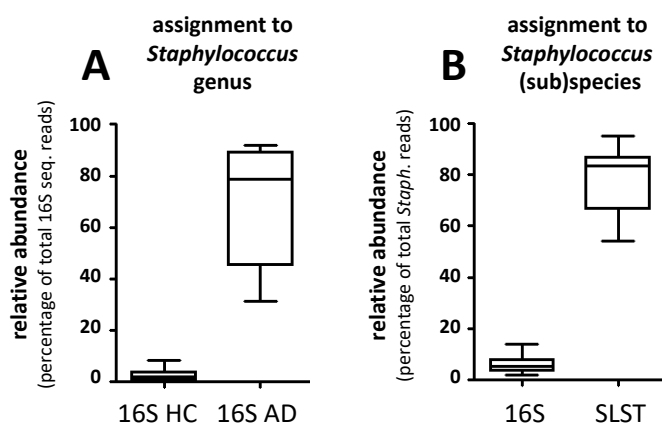


Figure 4. Significantly improved classification rate of *Staphylococcus* (sub)species for SLST compared to 16S.

(4A) Skin 16S V3–V4 sequencing data shows a significant increase for AD patients in *Staphylococcus* genus-level assigned reads in the dataset in comparison to HC ($p = 0.002$) (also see Supplementary Fig. S9), which is in line with literature [220]. (4B) When evaluating the percentage of (sub)species-level assigned reads, that is, number of total *Staphylococcus* reads that could be classified to species level or lower, we observe a significant increased classification rate for SLST in comparison to 16S ($p < 0.0001$). Boxplots as median with interquartile range and whiskers from min to max.

Based on *Staphylococcus* full-length 16S sequences (V1–V9) and clustering of these sequences into OTUs with 97% identity (Supplementary Table S12), no discrimination can be made between clinically relevant species. This analysis of V1–V9 identified only two major clusters, and these do not correlate well with phylogeny, as indicated by their ambiguous distribution within *Staphylococcus* subclades such as *S. aureus* and

S. haemolyticus (Supplementary Table S12). For 16S sub-regions V1-V2 and V3-V4, species of *S. epidermidis* and *S. capitis* cannot be distinguished, albeit that these are in fact very different species (Supplementary Table S12). Likewise, *S. aureus* cannot be distinguished from *S. schweitzeri* and *S. argenteus* species for V1-V2 and V3-V4. Furthermore, *S. aureus* and *S. epidermidis* species can only be discriminated by V1-V2, and not by V3-V4.

By adopting the TaxPhlAn Discovery & Design pipeline on the *Staphylococcus* reference dataset as listed in Table 1 (Supplementary Fig. S1; Module A), we identified an SLST target that allowed for perfect discrimination of *Staphylococcus* species, including multiple sub-species clades. This SLST target, denoted as orthologous group (OG) #1123 (Supplementary Fig. S7), shows a higher resolving power than 16S regions V1-V2 and V3-V4 (Table 2). OG #1123 was selected as our prime SLST candidate after elaborate *in silico* (Supplementary Table S13) and laboratory validations (Supplementary Fig. S8) to prove specificity, and to exclude cross-reactivity with phylogenetically distant bacteria. OG #1123 is predicted to be part of the 30S ribosomal protein S11 (note, this is no rRNA gene), with an on average amplicon sequence length of 381 nucleotides (including primers). We performed marker gene sequencing of OG #1123 on the AD study cohort, and analysis of the obtained sequence data by the TaxPhlAn SLST Oligotyping pipeline (Supplementary Fig. S2; Module B). We found that the SLST target allowed for 77.2% of *Staphylococcus* reads to be assigned to species-level or lower. In comparison to 5.9% for 16S (Fig. 4B), which is a tremendous improvement in *Staphylococcus* classification for SLST over conventional 16S-based sequencing approaches.

TaxPhlAn allows for profiling of *Staphylococcus* phylotypes in complex patient communities.

In further analysis of the *Staphylococcus* SLST data of the AD cohort, we observed a strong and significant increase of allele clusters classified as *S. aureus* in AD patients relative to healthy controls (30.07% AD, 1.43% HC, $p = 0.006$), and a significant decrease of *S. epidermidis* (5.58% AD, 33.34% HC, $p = 0.006$) and *S. haemolyticus* / *S. hominis* (0.12% AD, 12.70% HC, $p = 0.002$) allele clusters in these patients (Fig. 5, and Supplementary Table S14 for the corresponding SLST data). When taking into account the *Staphylococcus* genus-level abundance by 16S, this shift of SLST *S. epidermidis* and *S. haemolyticus* / *S. hominis* clusters between the experimental groups is lost, but not for *S. aureus* (Supplementary Fig. S10, and Supplementary Table S15 for the corresponding SLST data including / corrected for 16S). This is possibly a result of the low relative abundances of *Staphylococcus* in HC samples. Most interestingly, a significant increase of *S. capitis* in AD patients is observed when taking *Staphylococcus* 16S genus-level abundances into account (19.40% AD, 0.47% HC, $p = 0.003$) (Supplementary Fig. S10B).

Redundancy analyses (RDA) of the AD cohort further substantiates that skin microbiota profiles from HC and AD individuals are significantly different, both for 16S- and SLST-based sequencing data (Supplementary Figs. S11-S12; $p = 0.02$). Although RDA on 16S species- and OTU-classified data both indicate that *S. aureus* is crucial for separation of

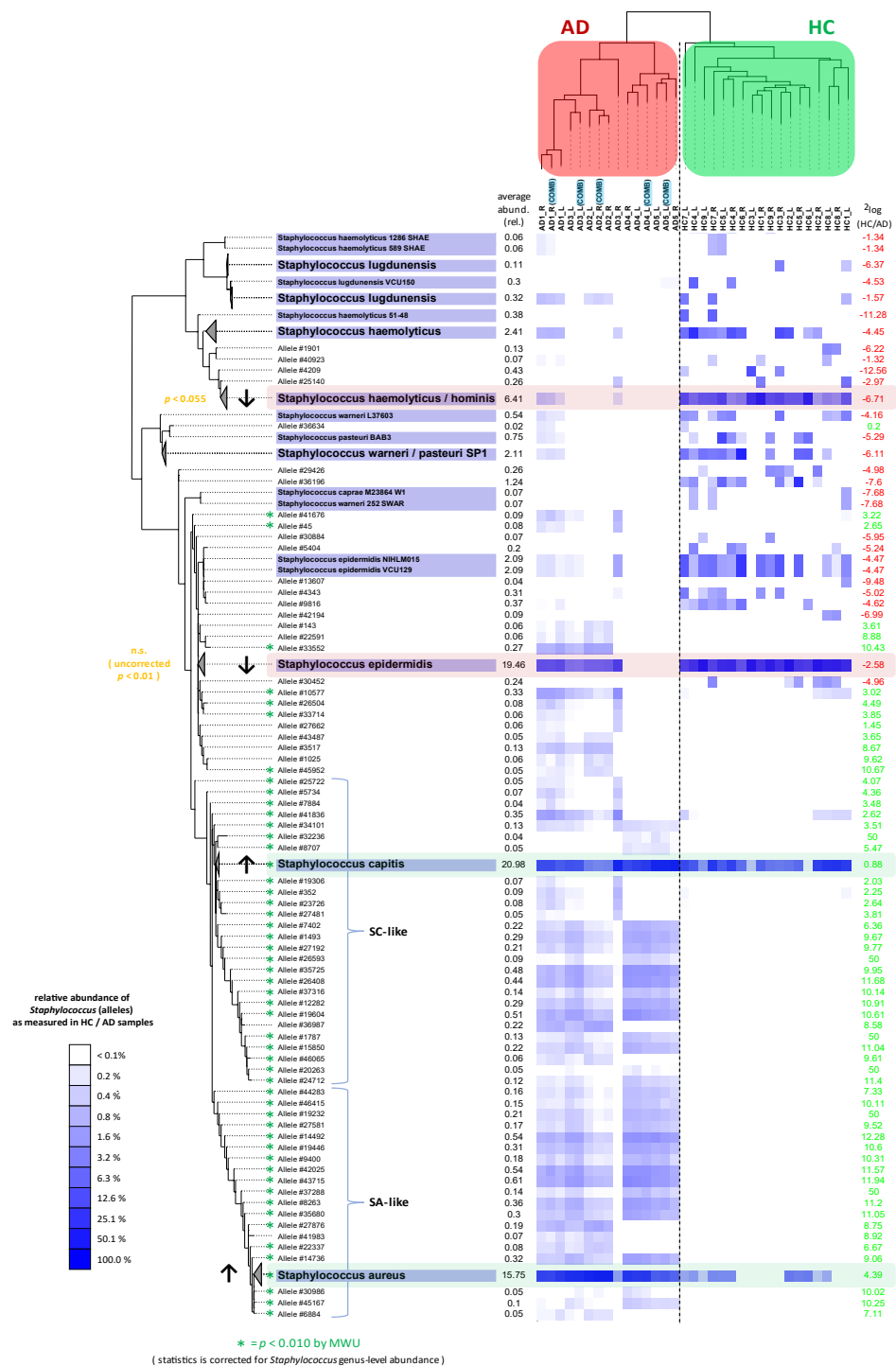


Figure 5. (previous page) *Staphylococcus* SLST data reveals two distinct clusters separating HC and AD samples, typed by *S. aureus*- and *S. capitis*-like species.

Heatmap summarizing the SLST *Staphylococcus* data from AD patients and HC volunteers (Supplementary Table S14). The SLST data was analyzed by the TaxPhlAn Oligotyping pipeline (Module B). The tree on the left is the maximum-likelihood phylogenetic tree of the SLST allele sequences generated with FastTree [230]. The alleles were built with an Shannon diversity index (SDI) threshold of 0.6, resulting in 45 SNP positions of the SLST marker gene sequences. Alleles for which a reference genome was available were named after this reference (e.g. *S. epidermidis* VCU129), instead of having an allele number (e.g. Allele #30452). Multiple references were collapsed into one clade if they shared the same allele sequence (e.g. *S. epidermidis*). The heatmap represents log₁₀-transformed SLST relative abundances. A green asterisk indicates significantly differentially abundances, according to a MWU test, and corrected for *Staphylococcus* genus-level relative abundance derived from 16S data (Supplementary Table S15). Note that the phylogenetic tree was cropped at top and bottom in order to facilitate visualization, leaving out some lowly abundant and statistically insignificant taxa from this view (for a full list see Supplementary Table S14). Sample labels in blue are samples where 16S and SLST marker gene PCR amplicons of the same sample were pooled (technical replicates). SA- and SC-like: *S. aureus*- and *S. capitis*-like. The digital, high-resolution figure is available as PPTX file at : (online) <http://ederveen.science/Thesis/Chapter3/Fig-5.PRESUBMISSION.pptx>

HC and AD subjects (Supplementary Fig. S11A-B, respectively), we observe that 16S simply does not recognize most of the OTUs as *S. aureus* (likewise for *S. epidermidis*). We hereafter combined 16S with SLST data in one analysis, i.e. SLST allele abundances were set relative to the total bacterial population, not relative to only the *Staphylococcus* fraction. Interestingly, this preserves the significant signals that allow for separating HC from AD, and also substantially increases the classification rate of *Staphylococcus* species, thereby showcasing the added value of such approach (Supplementary Fig. S12B). Most notably, the combined RDA shows a great importance for *S. aureus* (22.75% in AD) and *S. capitis* (19.40%) in discriminating between the two study groups, which is in line with the univariate analyses as reported earlier in Supplementary Fig. S10A-B.

Finally, 16S and SLST sequencing efforts were performed separately, however, five 16S technical replicates were taken along in the SLST sequencing run. These technical replicates show highly similar profiles, even though their 16S and SLST amplicons were pooled into one sequencing sample (Fig. 5, and Supplementary Fig. S9; samples indicated in blue are the pooled technical replicates). Likewise, biological replicates cluster very well, indicating that there is little intra-individual difference between sampling arms of the same volunteer.

Discussion

SLST marker gene sequencing has recently been described for determining strain level identification of specific bacteria within a microbiome [217]. Other methods for bacterial typing or profiling, such as species-specific PCR approaches or DNA fingerprinting-based methods for identification usually do not have the desired resolution to pinpoint strains of interest, are not compatible with high-throughput screening and / or are very labor intensive. TaxPhlAn fills the gap between the relatively cheap but coarse-grained 16S rRNA-based analyses and the more expensive, complex and laborious but higher resolution metagenomics sequencing.

TaxPhlAn SLST profiling can be efficiently combined with a 16S amplicon approach,

either following-up on an initial 16S pilot study that generated microbial leads in order to zoom-in on that clade with a higher resolution than feasible by 16S, or directly, by combining 16S and SLST amplicons in one single marker gene sequencing run. Our experimental data show that combining 16S and SLST amplicons for sequencing is practically feasible, can be done without adverse loss of sequencing depth, and without affecting microbiota profiles as obtained by separately sequencing 16S and SLST amplicons. This is particularly relevant, as it enables running SLST experiments in tandem with regular 16S sequencing efforts without additional sequencing costs. SLST sequencing can be multiplexed, for example by combining multiple different SLST targets per sample in a single sequencing initiative, thereby boosting either the discriminatory resolution of SLST (hence, in principle approaching MLST marker gene sequencing), or enabling the profiling of multiple bacterial taxa in the same experiment.

As an additional feature, upon running the TaxPhlAn SLST Discovery & Design pipeline one is able to assign genomes-of-interest (strains) in order to filter on SLST candidates that allow distinguishing these genomes from all other input genomes. This is particularly useful when there is specific need to identify one particular phylotype in your samples, for example when one needs to distinguish sub-species *Bifidobacterium longum subsp. longum* from *Bifidobacterium longum subsp. infantis*. Likewise, one is able to assign negative control genomes to filter on primers of SLST candidates that do not recognize these genomes, for example when one needs to have SLST candidates of which the primers are specific for *Propionibacterium acnes*, but not for other closely related commensals such as *Propionibacterium namnetense*.

One limitation of DNA amplification-based approaches for profiling of specific taxonomic groups is the requirement of sequenced (reference) genomes in order to identify suitable molecular targets. 16S marker gene sequencing does not necessarily require reference genomes to identify new biodiversity, because it is based on the 16S rRNA gene that is universally present in all prokaryotes, and these sequences correlate well with phylogeny. For correct taxonomic classification of 16S marker gene sequencing reads prior knowledge from 16S rRNA gene databases is required, e.g. RDP [36], GreenGenes [37] and SILVA [38]. For SLST these considerations are similar to those of 16S. Even though TaxPhlAn requires a set of references, a dozen of sequenced reference genomes is sufficient to allow for confident inference of phylogenetic information as shown by the benchmarking results presented in this study. Nonetheless, the more reference genomes available for oligotyping analysis, the more confident SLST allele classification and inference of novel biodiversity will be. Also, although we here presented a generic methodological workflow applied to bacteria only, we see no theoretical objections for the characterization of archaea, fungi or lower unicellular eukaryotes with similar SLST approach.

In the current study, we mainly identified taxa of *S. aureus* and *S. capitis*, which were significantly increased in AD, in comparison to taxa of *S. epidermidis* and a cluster of *S. haemolyticus* / *S. hominis*, which were associated with healthy skin. Although these are known *Staphylococcus* species, also many *S. aureus*-like and *S. capitis*-like unknown taxa were identified which were significantly more abundant in AD, for which no

existing reference is available in current databases, but for which we now know their phylogenetic relation to known *Staphylococci*. Importantly, the strong relation of *S. capitis* with AD disease as found by our SLST method has to our knowledge not been reported before; possibly because of the limited resolution of 16S, or the fact that previous in-depth metagenomics studies on AD did not include lesional skin samples [221]. It might therefore be worthwhile pursuing a possible role for *S. capitis* in AD in follow-up studies.

In conclusion, TaxPhlAn provides a method for the automated design of SLST amplicons and the analysis of SLST sequencing data. As TaxPhlAn evaluates regions that are single-copy orthologous genes, the resulting SLST amplicons allow detection of novel variants and placement of these variants in phylogenetic context.

Methods

Skin microbiome sample collection and processing

In advance of study start, medical ethical committee (Commissie Mensgebonden Onderzoek Arnhem-Nijmegen) approval, and individual written informed consent were obtained. The study was performed according to the Declaration of Helsinki principles. Microbiome skin samples were collected by a wet swabbing protocol of the inner elbow (antecubital fossa) skin of human healthy control (HC) volunteers (n = 9) and atopic dermatitis (AD) patients (n = 5). Inclusion criteria (and exclusion, with exception of having AD) as described by Zeeuwen *et al.*, 2012 [3]. All human volunteers were enrolled in the dermatology clinic of the Radboud University Medical Center, Nijmegen, and AD was diagnosed by a trained dermatologist. Sample collection was performed as described previously [3, 40]. In short, a 4 cm² skin area of the inner elbow was swabbed with sterile Catch-All sample collection swabs (Epicentre Biotechnologies, Madison, USA). The swabs were soaked in sterile SCF-1 solution (50mM Tris buffer [pH8], 1 mM EDTA, and 0.5% Tween-20) before sample collection. Mock swabs, only exposed to ambient air, were taken as negative controls. The Mobio Ultraclean Microbial DNA isolation kit (Mobio laboratories, Carlsbad, USA) was used according to the manufacturers protocol to extract microbial DNA, before storing it at -80°C until further use.

16S and SLST PCR chemistry and conditions

Primer sequences used for sequencing of the 16S V₃-V₄ region or by SLST can be found in [Supplementary Table S5](#), and were appended with Illumina adaptor sequences and sample-specific barcodes. PCR protocol with KOD hot start DNA polymerase as follows: 2m 95°C hot start; 35 cycles of 20s 95°C, 10s 61°C, 15s 70°C; 10m 70°C. PCR quality control as follows: agarose gel amplicon sizes were checked, and Sanger sequenced to validate with BLAST to the NCBI database. SLST primer specificity was tested on the following *Staphylococcus* strains: *S. aureus* (ATCC 29213), *S. epidermidis* (ATCC 12228), and clinical isolates of *S. capitis* and *S. hominis*. As negative controls we used the common skin commensals: *Propionibacterium acnes* (ATCC 6919), *Pseudomonas aeruginosa* (ATCC 27853), *F. magna* (ATCC 15794), and a clinical isolate of *C. aurimucosum*.

16S and SLST marker gene sequencing and data pre-processing

PCR 16S and SLST amplicon libraries were generated as described above. For each of the 5 AD patients, one additional pooled sample was taken along, with mixed 16S and SLST amplicons. The libraries were barcoded, multiplexed and sequenced on an Illumina MiSeq machine with paired-end 300 cycles protocol and indexing, by BaseClear B.V. (Leiden, The Netherlands). 16S and SLST sequencing data were generated in separate Illumina runs, but the 16S / SLST pooled samples (technical replicates) were taken along with the SLST run. Illumina sequencing data was quality checked and demultiplexed by BaseClear standards, as detailed in the [Supplementary Methods](#), and FASTQ files were generated. Paired-end reads were assembled into pseudoreads with PEAR [223], with strict assembly settings: quality threshold 30, minimum overlap 35 and p -value 0.0001. On average, only 0.17% of the raw reads could not be assembled (data not shown). Thereafter, for the 16S / SLST pooled samples, the pseudoreads were split to 16S or SLST input FASTA files by a local BLAST to a SLST gene database of the 7078 available *Staphylococcus* genomes: query reads with a hit (by default BLASTn settings from version 2.2.29+) were used for SLST analysis, if not send for 16S analysis by QIIME. This script has been made available as '*split-reads-to-16S-SLST*' in the TaxPhlAn Docker image.

16S marker gene sequencing data analysis

For generation of the 16S-derived taxa-to-sample compositional matrix, a customized Python workflow based on Quantitative Insights Into Microbial Ecology (QIIME version 1.8) [42] was adopted (<http://qiime.org>). Reads were filtered for chimeric sequences using the UCHIME algorithm version 4 [45]. Hierarchical clustering of samples was performed using UPGMA with weighted UniFrac as a distance measure as implemented in QIIME 1.8. Figures resulting from these clustering analyses were generated using the interactive tree of life (iTOL) tool [224]. The Ribosomal Database Project classifier version 2.3 was performed for taxonomic classification of the sequence reads [225]. Alpha diversity metrics (PD whole tree, Chao1, Observed Species and Shannon) were calculated by bootstrapping 6490 reads per sample, and taking the average over 10 trials. For visualization of the differential microbiome, Cytoscape software version 3.4.0 [136] was used together with in-house developed Python scripts for generating the appropriate input data deriving from the QIIME analysis. Note that due to technical limitations in the resolution of 16S marker gene sequencing, OTU (operational taxonomic unit) calling on the level of species should be interpreted with caution.

Selection of representative *Staphylococcus* genomes for TaxPhlAn target discovery

Staphylococcus genomes were downloaded ($n = 7247$) from the NCBI assembled genomes database [226] (<ftp.ncbi.nlm.nih.gov/genomes/genbank/bacteria/>) and the NCBI Traces WGS database (<https://www.ncbi.nlm.nih.gov/Traces/wgs/>). TaxPhlAn requires sub-selection of input genomes due to computational capacity. We selected a diverse subset of those genomes based on their 16S sequences by running a BLAST search with the full 16S gene of *S. aureus subsp. aureus* 71193 to each available genome (*Staphylococcus* 16S reference gene was downloaded from SILVA database [38] in April

2016; accession CP003045). Based on the best scoring BLAST hits, 16S genes were aligned to the aforementioned 16S reference gene by a pairwise global alignment with Needle (version EMBOSS: 6.3.1; default settings) [227]. From these alignments, a 1Kb long 16S region covered by the majority of genomes was selected ($n = 6569$) and 16S OTU clusters were determined using UCLUST (version 1.2.22q) [45] for cluster building with percentage identity set to 99.7%. This yielded 24 (16S) clusters from which we equally selected 200 unique *Staphylococcus* genomes in total. For further details we refer to the [Supplementary Methods](#).

TaxPhlAn (Module A): SLST target discovery and design workflow

The TaxPhlAn *discovery & design* pipeline (Module A) finds SLST markers of single locus orthologous genes to profile specific bacterial taxa up-to and beyond the species level, based on a set of reference genomes provided by the user. It consists of a series of Perl/Python scripts, and has been made available through the supplied Docker as 'TaxPhlAn-SLST-Design-wrapper.py' (wrapper script). For more details concerning the TaxPhlAn *discovery & design* workflow, we refer to the [Supplementary Methods](#). In short, it consists of the following steps ([Supplementary Fig. S1](#)). *Initiation and QC: (Phase I-a)* data preparation based on a configuration file and input genomes; *(Phase I-b)* data preprocessing by input file reformatting and genome annotation. *Orthology: (Phase II)* orthology calculations and phylogenetic analysis on provided genomes. *Target Discovery and Design: (Phase III-a)* selection of core gene clusters for candidate marker genes; *(Phase III-b/c)* prediction of variable regions in candidate core clusters. *Validation and Reporting: (Phase IV)* *in silico* primer design and evaluation of candidate variable regions, and final reporting of results.

TaxPhlAn (Module B): SLST data analysis (oligotyping) workflow

The TaxPhlAn *oligotyping analysis* pipeline (Module B) is used to analyze (and visualize) raw SLST marker gene sequences from raw sequencing data to a compositional microbiota-to-sample matrix ([Supplementary Table S14](#)). It consists of a series of Python scripts, and has been made available through the supplied Docker as 'TaxPhlAn-SLST-Oligotyping-wrapper.py' (wrapper script). For more extensive details concerning the TaxPhlAn *oligotyping analysis* workflow we refer to the [Supplementary Methods](#). In short, it consists of the following summarized steps (see graphical outline in [Supplementary Fig. S2](#)). *Initiation and QC: (Phase 1)* Data preparation with mandatory input and quality control. *Allele building: (Phase II)* Oligotyping by allele building based on Shannon diversity index. *Allele matching: (Phase III-a)* SLST allele matching and scoring; *(Phase III-b)* Phylogenetic analysis of alleles, and data visualization. *Reporting: (Phase IV)* Oligotyping data results and reporting.

Statistics

For the microbiota data in this manuscript, statistical significance between contrasts with regard to taxonomy abundances was tested by a non-parametric (unpaired) Mann-Whitney U (MWU), uncorrected. Statistical tests were performed by custom, in-house Python scripts (SciPy module version 0.17.0; <https://www.scipy.org/>) downstream of QIIME. Redundancy Analysis (RDA) was done using Canoco 5.04 [228]

using default settings of the analysis type “Constrained”. Relative abundance values for taxa were used as response data, and the sample classes as explanatory variables. RDA calculates p -values by permuting the sample classes. Correlations were examined using Spearman’s rank test as performed by custom, in-house R scripts (version 3.2.2; <https://www.r-project.org/>). For any other type of data visualization we adopted GraphPad Prism 5.0 or Microsoft Office Excel 2016. Significances mentioned in figures are as follows: n.s. (not significant), * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Sequencing data availability

SLST and 16S Illumina sequencing data is available for download at the European Nucleotide Archive (ENA) database (<http://www.ebi.ac.uk/ena>) [229] under study accession number PRJEB27442 (or secondary accession number ERP109520). The sequencing data is available in FASTQ-format, including corresponding metadata for each sample. An overview of the samples and metadata can be found in [Supplementary Table S10](#).

TaxPhlAn pipeline distribution

TaxPhlAn is accessible through a pre-configured, plug-and-play Docker virtual machine which is supplied at the Docker Hub repository: <https://hub.docker.com/r/ederveen/taxphlan/>. Directions to SLST test datasets and documentation can be found in the TaxPhlAn Docker home directory upon running the Docker image. For quick and easy install of the TaxPhlAn Docker image please see GitHub at: <https://github.com/ederveen/taxphlan/>.

Supplementary Methods

([online](#)) Supplementary Methods of online repository at :
<http://ederveen.science/Thesis/Chapter3/Supplementary-Methods.PRESUBMISSION.pdf>

Supplementary Figures

Figure S1. Overview of TaxPhlAn SLST Discovery and Design pipeline steps (Module A). (PDF)

([online](#)) Figure S1 of online repository at :
<http://ederveen.science/Thesis/Chapter3/Supplementary-Figures.PRESUBMISSION.pdf>

Figure S2. Overview of TaxPhlAn SLST Oligotyping pipeline steps (Module B). (PDF)

([online](#)) Figure S2 of online repository at :
<http://ederveen.science/Thesis/Chapter3/Supplementary-Figures.PRESUBMISSION.pdf>

Figures S3-S6. SLST candidates as predicted by TaxPhlAn correlates with the actual phylogenetic distances between genomes. (PDF)

(online) Figures S3-S6 of online repository at :

<http://ederveen.science/Thesis/Chapter3/Supplementary-Figures.PRESUBMISSION.pdf>

Figure S7. *Staphylococcus* full-genome phylogenetic tree projected with 16S clusters and SLST OG #1123 variable regions. (PDF)

(online) Figure S7 of online repository at :

<http://ederveen.science/Thesis/Chapter3/Supplementary-Figures.PRESUBMISSION.pdf>

Figure S8. PCR validation of SLST candidates primer sets with *Staphylococcus* species and common skin commensals. (PDF)

(online) Figure S8 of online repository at :

<http://ederveen.science/Thesis/Chapter3/Supplementary-Figures.PRESUBMISSION.pdf>

Figure S9. Skin microbiota composition in healthy controls and AD patients: *Staphylococcus* dominates skin of AD patients at the expense of *Propionibacterium*. (PDF)

(online) Figure S9 of online repository at :

<http://ederveen.science/Thesis/Chapter3/Supplementary-Figures.PRESUBMISSION.pdf>

Figure S10. *Staphylococcus* species are increased in AD in comparison to HC. (PDF)

(online) Figure S10 of online repository at :

<http://ederveen.science/Thesis/Chapter3/Supplementary-Figures.PRESUBMISSION.pdf>

Figure S11. Strong under-assignment of 16S *Staphylococcus* taxa below the level of genus. (PDF)

(online) Figure S11 of online repository at :

<http://ederveen.science/Thesis/Chapter3/Supplementary-Figures.PRESUBMISSION.pdf>

Figure S12. The below genus-level bacterial entities discriminating HC and AD individuals is mainly explained by *S. aureus* and *S. capitis* species. (PDF)

(online) Figure S12 of online repository at :

<http://ederveen.science/Thesis/Chapter3/Supplementary-Figures.PRESUBMISSION.pdf>

Supplementary Tables

Tables S1-S4. Benchmark datasets genome information and statistics. (PDF)

(online) Tables S1-S4 of online repository at :

<http://ederveen.science/Thesis/Chapter3/Supplementary-Tables.PRESUBMISSION.pdf>

Table S5. 16S and SLST primers used in this study for benchmarking and sequencing. (PDF)

(online) Table S5 of online repository at :

<http://ederveen.science/Thesis/Chapter3/Supplementary-Tables.PRESUBMISSION.pdf>

Tables S6-S9. Benchmark datasets TaxPhlAn run output reports. (PDF)

(online) Tables S6-S9 of online repository at :

<http://ederveen.science/Thesis/Chapter3/Supplementary-Tables.PRESUBMISSION.pdf>

Table S10. Study samples overview Ederveen *et al.* (PDF)

(online) Table S10 of online repository at :

<http://ederveen.science/Thesis/Chapter3/Supplementary-Tables.PRESUBMISSION.pdf>

Table S11. 16S sequencing read statistics and alpha diversity metrics. (PDF)

(online) Table S11 of online repository at :

<http://ederveen.science/Thesis/Chapter3/Supplementary-Tables.PRESUBMISSION.pdf>

Table S12. 16S analysis of *Staphylococcus* by OTU clustering of full-length 16S, V1-2 and V3-4. (PDF)

(online) Table S12 of online repository at :

<http://ederveen.science/Thesis/Chapter3/Supplementary-Tables.PRESUBMISSION.pdf>

Table S13. SLST candidates for *Staphylococcus* profiling. (PDF)

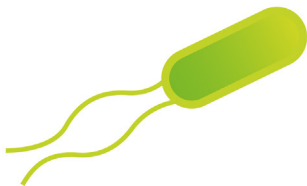
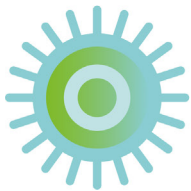
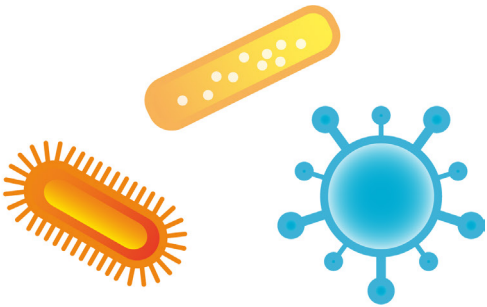
(online) Table S13 of online repository at :

<http://ederveen.science/Thesis/Chapter3/Supplementary-Tables.PRESUBMISSION.pdf>

Tables S14-S15. 16S and SLST compositional matrix. (PDF)

(online) Tables S14-S15 of online repository at :

<http://ederveen.science/Thesis/Chapter3/Supplementary-Tables.PRESUBMISSION.pdf>



PART II

The Skin Microbiome

CHAPTER 4

OPEN ACCESS

as published in *Acta Derm Venereol*, 2016, Nov 2;96(7):873-879

<https://doi.org/10.2340/00015555-2401>

¹ Department of Dermatology, Radboud Institute for Molecular Life Sciences (RIMLS), Radboud University Medical Center (Radboudumc), Nijmegen, The Netherlands.

² Centre for Molecular and Biomolecular Informatics, RILMS, Radboudumc, Nijmegen, The Netherlands.

³ Department of Medical Microbiology, RILMS, Radboudumc, Nijmegen, The Netherlands.

⁴ Department for Health Evidence, Radboud Institute for Molecular Life Sciences (RIMLS) and Radboud Institute for Health Sciences (RIHS), Radboudumc, Nijmegen, The Netherlands.

⁵ NIZO, Ede, The Netherlands.

These authors contributed equally to this work.

CHAPTER 4

AN *IN VITRO* MODEL FOR BACTERIAL GROWTH ON HUMAN *STRATUM CORNEUM*

Danique A. van der Krieken ^{1 #}

Thomas H.A. Ederveen ^{2 #}

Sacha A.F.T. van Hijum ^{2,5}

Patrick A.M. Jansen ¹

Willem J.G. Melchers ³

Paul T.J. Scheepers ⁴

Joost Schalkwijk ¹

Patrick L.J.M. Zeeuwen ¹

ABSTRACT

Recent studies have revealed the diversity and dynamics of the skin microbiome in health and disease, but adequate model systems to study skin microbiota *in vitro* are largely lacking. We developed an *in vitro* system that mimics human stratum corneum and supports bacterial growth. In this model, human callus serves as substrate and nutrient source for bacteria. Bacterial growth of several commensal and pathogenic strains was determined up to one week by counting CFUs or qPCR using strain-specific primers. Subsequently we demonstrated that human skin pathogens could survive amidst a minimal microbiome consisting of two major skin commensals, *S. epidermidis* and *P. acnes*. Finally, we succeeded to inoculate and maintain complete microbiomes, taken from the back of healthy volunteers, on this system. We propose that this model might be used to modulate skin microbiomes *in vitro* and allows testing of pathogens, biological agents and antibiotics, in a medium-throughput format.

Background

Our skin is the physical barrier that protects the interior of the human body from the exterior, and is covered with a complex microbial ecosystem that is in homeostasis with its host [106]. The microbial communities that reside on the skin of our body were recently investigated using culture-independent methods [231, 232]. These microbial communities, termed microbiota, are considered beneficial for our health [233]. Furthermore, it is now widely accepted that disturbance of 'normal' microbial communities, a condition called 'dysbiosis', where homeostatic relations between the host and its microbiota are disturbed, is associated with skin diseases [234]. Skin microbiota have therefore become the subject of studies in a search for markers related to disease onset, progression, and outcome, which might be helpful to develop novel therapies for skin diseases.

Different factors influence the microbiota composition of the skin such as lifestyle, host demographic and environmental characteristics, the use of antibacterial agents, and health status [105, 235-237]. Skin alterations and dysregulated immune responses could cause shifts in microbiota composition [3, 238-241] and on the other hand, skin microbiota can modulate cutaneous immunity [242-244]. Recent studies have shown that the microbial diversity of lesional skin is altered in inflammatory skin diseases like psoriasis and atopic dermatitis [239, 245-247]. However, it is still not known if these shifts in bacterial composition have a causal role in disease or are merely a consequence of disturbed skin homeostasis.

Changes in bacterial diversity, microbiota composition, and host-defence interactions in health and disease can be studied using skin models. For instance, human reconstructed skin models and *ex vivo* skin models have been used to study the efficacy of antibacterial compounds [248-251] or colonization and infection characteristics of certain pathogens [252-256]. However, these models are laborious, expensive, prone to bacterial overgrowth, and their throughput is low.

Under normal skin conditions, most of the bacteria reside in the upper half of the stratum corneum [3] and feed on nutrients derived from the corneocytes, which consist mainly of crosslinked protein and lipids. Here we investigated if human callus could serve as a substitute for stratum corneum to support the growth of skin commensals and pathogens. Our results show that such a model is applicable for investigating bacterial growth, and the assessment of the efficacy of antimicrobial compounds.

Results

Development and validation of the model: callus as a source of nutrients.

In the *in vivo* situation, bacteria live and attach to the dead corneocytes of the outer layers of human skin. We therefore hypothesized that callus would be a natural source of nutrients for microorganism to survive and grow upon. We prepared the model (Fig. 1) in a 24-wells plate and used agar (devoid of nutrients) as a basis for the callus

suspension and to maintain a sufficiently hydrated surface. After drying of the applied callus suspension, a thin layer of dead corneocytes is present on top of the agar, upon which the bacteria are inoculated and cultured. The procedure for collection of the bacteria from the wells and subsequent analysis by determination of colony forming units (CFU) is depicted in [Supplementary Fig. S1](#).

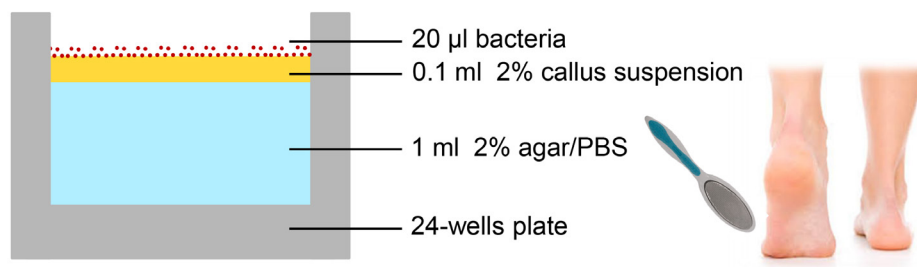


Figure 1. Structure of the human cutaneous microbial ecosystem model.

The model is prepared in a 24-wells plate. 2% agar (in PBS) serves as a basis to place the 2% callus suspension originating from the heel of healthy volunteers. After drying, bacteria are inoculated on top of the dead corneocytes and cultured at 32°C.

Regulation of the proliferation rate of the bacteria: dynamic versus static model.

In the *in vivo* situation, the number of bacteria on healthy human skin remains at a steady-state level, meaning that proliferation rates are low and no overgrowth, as in overt infection, occurs. To mimic these conditions *in vitro*, a static model is required. However, if an infectious situation needs to be mimicked, a more dynamic model is preferred. This could be useful, for instance, when antimicrobial compounds are tested. Therefore we inoculated the model with different numbers of two skin commensals (*S. epidermidis* and *P. acnes*) and two relevant skin pathogens (*S. aureus* and *P. aeruginosa*), cultured them at 32°C, and determined the relation between the starting concentration and the growth of these bacteria. The samples were analyzed on day 1, 4 and 7 by CFU counting ([Fig. 2](#)). At day 1 the CFU/ml was comparable for all samples except for *P. acnes*, regardless of the different starting concentration of the inoculums. It was shown that the bacteria that were applied in a lower concentration proliferated more frequently than those added in higher concentrations. The two highest concentrations of *P. acnes* proliferated 10^2 - 10^3 times from day 0 to day 1, the number of CFU/ml in lower concentrations decreased at every time point ([Fig. 2B](#)). From day 1 to 7, the level of *S. epidermidis* remained steady (10^8 - 10^{10} CFU/ml) for all different starting concentrations ([Fig. 2A](#)), whereas the level of *S. aureus* diminished approximately 10^4 fold in this time period ([Fig. 2C](#)). The highest starting concentrations of *P. aeruginosa* increased at every time point, the lower concentrations remained at a steady level (10^{21} - 10^{22} CFU/ml), whereas the lowest concentrations decreased 10^5 times after day 4 ([Fig. 2D](#)).

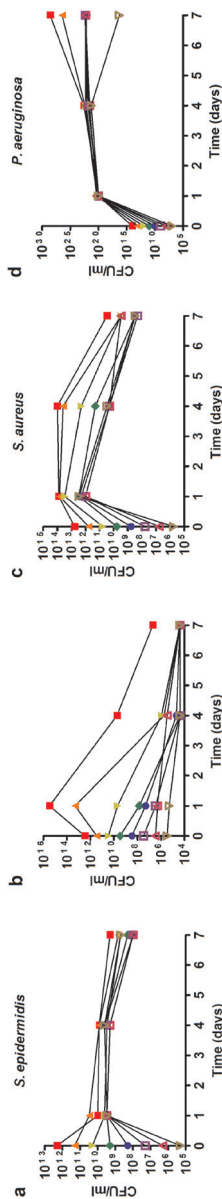


Figure 2. Effect of the inoculum concentration on bacterial survival.

(A) *S. epidermidis*, (B) *P. acnes*, (C) *S. aureus*, and (D) *P. aeruginosa* inoculated on the model in different concentrations ranging from 10^5 to 10^{12} CFU/ml and analysed at day 1, 4 and 7 by CFU counting. Lower starting concentrations of inoculated bacteria resulted in an increased growth at day 1, except for the lower concentrations of *P. acnes*. After the first day, all *S. epidermidis*, *S. aureus* and *P. acnes* concentrations followed a similar growth pattern. The concentrations of *P. aeruginosa* follow the same growth pattern up to day 4, at day 7 the pattern changes depend on the starting concentration. This figure represents one of two separate experiments.

To determine if callus is really indispensable as a source of nutrients for bacteria, *S. epidermidis* and *S. aureus* ($\sim 10^9$ - 10^{10} CFU/ml) were inoculated on the model without callus incorporated. Bacteria still showed a limited ability to proliferate at day 1, however no viable bacteria were detected on the model without callus after three days of culturing ([Supplementary Fig. S2](#)).

Strain-specific qPCR and PMA treatment optimisation.

Counting CFU to analyse bacterial survival can only be applied when single strains are tested on the callus model, as it is difficult to distinguish between different bacterial colonies on blood agar plates. Therefore, strain-specific qPCR primers can be used to analyse mixed bacterial compositions when, for example, testing the efficacy of antimicrobial compounds on bacterial strains. However, during DNA isolation, genomic DNA (gDNA) from non-viable bacteria is also isolated and skews the qPCR data. Propidium monoazide (PMA) is able to penetrate the membrane of non-viable bacteria, and if exposed to light it covalently binds to gDNA. During isolation of gDNA and qPCR analysis, only unbound gDNA from viable cells, can be isolated and amplified [257]. The procedure for bacterial collection from the wells and subsequent analysis by PMA-qPCR analysis is depicted in [Supplementary Fig. S3](#). To optimize the PMA treatment we examined what minimal time of light exposure is still effective against dead bacteria, without being harmful to viable cells. Five min light exposure resulted in a negligible effect on the viable bacteria compared to the no exposure condition ([Supplementary Fig. S4](#)) and that the

PMA procedure is effective to distinguish between viable and non-viable *S. epidermidis*, *P. acnes*, *S. aureus*, *P. aeruginosa* and *S. pyogenes* bacteria ([Supplementary Fig. S5](#)). Subsequently, a ratio of viable and non-viable bacteria was made for all selected strains to validate the PMA treatment. Mixtures of viable, and non-viable bacteria were exposed to PMA and light for 5 min, followed by gDNA isolation and qPCR analysis using Broad Range Universal (BRU) 16S rRNA gene primers ([Supplementary Table S1](#)). Using this PMA-qPCR protocol we could reproduce the viable/non-viable cells ratios as generated in advance ([Fig. 3](#)).

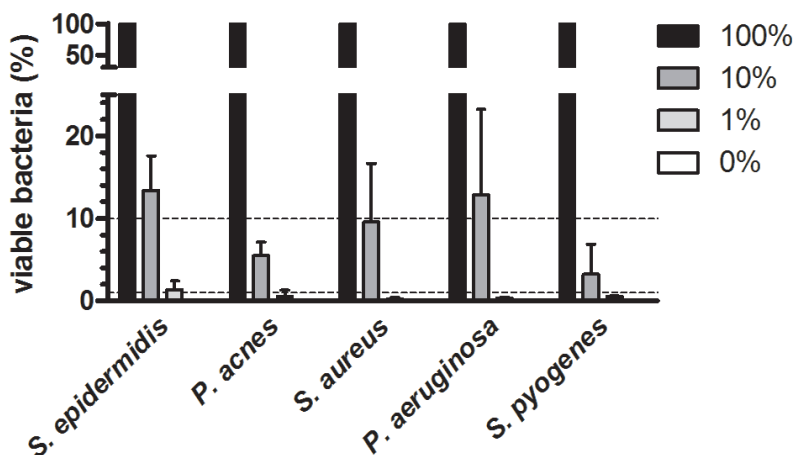


Figure 3. Effectiveness of PMA.

PMA-qPCR analysis on gDNA from 100% viable (black bars), 10% viable (dark grey bars), 1% viable (light greys bars) and 0% viable (white bars) *S. epidermidis*, *P. acnes*, *S. aureus*, *P. aeruginosa* and *S. pyogenes* bacteria. The amount of gDNA detected with PMA-qPCR in the 100% viable group was set at 100%. qPCR was performed with BRU 16S rRNA gene primers. Data represent the mean values \pm SD of two separate experiments.

Establishing a 'minimal' in vitro microbiome on the callus model.

Using the callus model and a strain-specific PMA-qPCR as a read-out, we aimed to create a 'minimal' microbiome *in vitro*. We used *S. epidermidis* and *P. acnes* strains, the two most common and abundant representatives of the Firmicutes and Actinobacteria, to create this 'minimal' microbiome. Both skin commensals were inoculated together on the model (10^8 CFU/ml) and analyzed at day 1 and 7 by PMA-qPCR using strain-specific primers ([Supplementary Table S1](#)). The CFU/ml were calculated based on the Ct values using the calibration curve ([Supplementary Fig. S6](#)). *S. epidermidis* showed the same growth curve ([Fig. 4](#); up on day 1, and down again on day 7) as when cultured as a single species on the model ([Fig. 2A](#)). *P. acnes* also showed the same growth curve in a mixture as cultured as a single species on the model ([Fig. 2B](#)).

Next, we investigated if the addition of a pathogen would have an effect on the minimal microbiome on the model system. We added *S. aureus* (10^8 CFU/ml) to both commensals on the model ([Fig. 4B](#)), but this had no quantifiable effect on the growth of *S. epidermidis* and *P. acnes*. Moreover, *S. aureus* showed the same growth curve as when cultured as a single species on the model. However, *S. aureus* bacteria cultured on the model ([Fig. 2C](#)) grew faster and to higher levels on the first day ($\sim \log_4$ scale), than when they were cultured together with *S. epidermidis* and *P. acnes* ($\sim \log_2$ scale, [Fig. 4B](#)). At day 7, comparable levels of *S. aureus* were observed in both experiments. Addition of *P. aeruginosa* (10^8 CFU/ml) to both commensals did also not lead to growth pattern-related changes of the minimal microbiome ([Fig. 4C](#)), however, *P. aeruginosa*, like *S. aureus*, also showed a decreased proliferation rate when cultured in the presence of *S. epidermidis* and *P. acnes*.

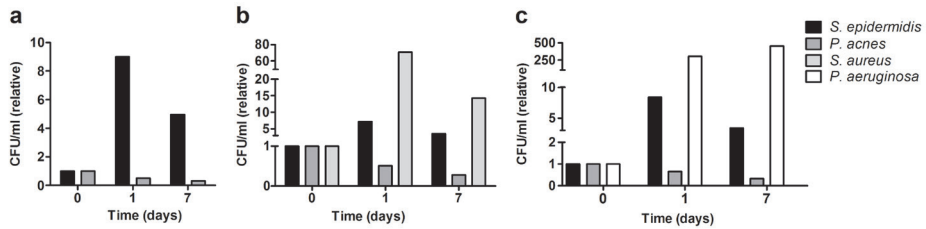


Figure 4. A minimal microbiome and pathogens on the callus model.

(A) *S. epidermidis* and *P. acnes* together form a minimal microbiome that survives for at least one week on the model. (B) Addition of *S. aureus* and (C) *P. aeruginosa* to the minimal microbiome. Both pathogens showed a growth pattern comparable to the strains that were cultured as a single species on the model. No effects of *S. aureus* and *P. aeruginosa* on the growth of *P. acnes* and *S. epidermidis* were found. Values are relative to day 0. qPCR was performed with strain-specific primers. This figure represents one of two separate experiments.

Application of a complete human skin microbiomes on the callus model.

The *in vivo* microbiomes of the lower back of two healthy volunteers (HV) obtained by skin swabbing was cultured for one week on the callus model. The lower back was selected because of the large diversity in bacteria [3]. For analysis, PMA-Illumina sequencing was applied. The *Staphylococcus* genus is the most abundant, other genera present include *Corynebacterium*, *Propionibacterium*, and *Pseudomonas*. The microbiota composition remained fairly constant over the week, except for the increased relative abundance of *Corynebacterium* species in HV1 on day 7 (Fig. 5A). Moreover, phylogenetic diversity (PD) whole tree analysis, a metric for alpha diversity also taking taxonomic distance into account, showed that the bacterial alpha diversity remained constant over the seven days of analysis (Fig. 5B).

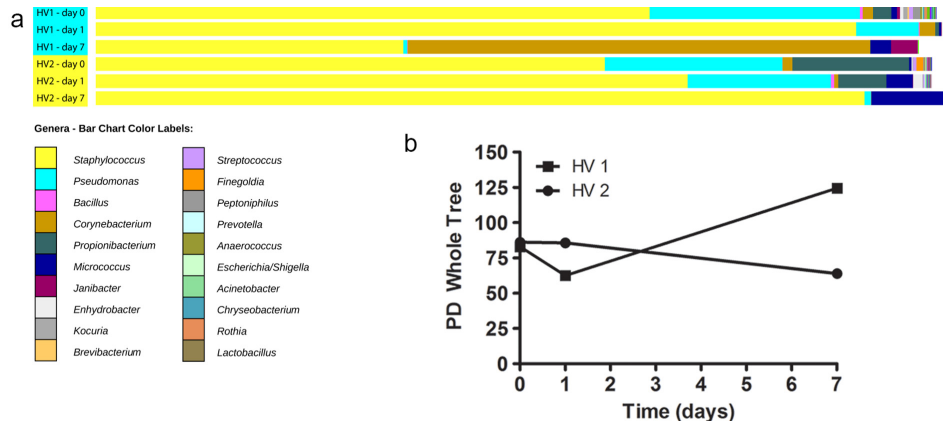


Figure 5. Bacterial composition of human skin microbiomes on the model.

(A) The bacterial composition of *in vivo* microbiomes originating from two healthy volunteers in a one week time period on the model. Samples were treated with PMA before Illumina sequencing. The bacterial composition did not alter significantly at day 1 compared to day 0, however, a change could be observed for HV1 at day 7. (B) Phylogenetic diversity whole tree of the changes in bacterial diversity at day 0, 1 and 7. Bacterial diversity remains relatively constant during one week on the model.

Efficacy of tetracycline on the callus model.

To test if our model is suitable to examine antibacterial properties of compounds, we tested the efficacy of tetracycline. The results obtained with our model were compared to the minimal inhibitory concentration (MIC) found in related experiments performed in Mueller Hinton culture medium (2 $\mu\text{g/ml}$ for *S. aureus* and 64 $\mu\text{g/ml}$ for *S. epidermidis*). We cultured *S. epidermidis* and *S. aureus* in medium and on our model for 24 h together with the skin antibiotic tetracycline. The efficacy of tetracycline was determined by CFU counting. We demonstrated that *S. aureus* is sensitive to tetracycline compared to *S. epidermidis* (Fig. 6A).

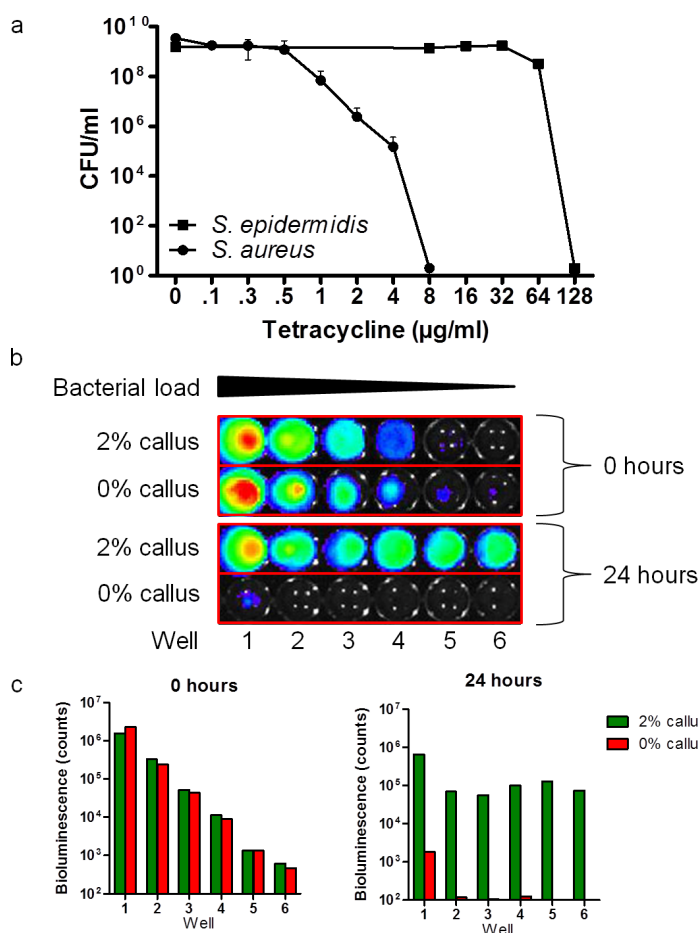


Figure 6. Possible applications and usability of the callus model.

(A) Testing of antimicrobial compounds. *S. epidermidis* and *S. aureus* were exposed for 24 h to tetracycline in the callus model. On the callus model, in accordance to growth in medium, *S. aureus* is more vulnerable to tetracycline compared to *S. epidermidis*. Values represent the mean \pm SD of two experiments. (B) Applicability of bioluminescent bacteria for high-throughput analysis. Bioluminescence of Xen36 measured with IVIS at 0 and 24. Dilutions of 10 were pipetted on the model with and without 2% callus suspension, sample number 1 being the highest concentration of Xen36 on the model. (C) Bioluminescence (counts) of Xen36 in (B), analysed with Living Image 3.1 software.

Bioluminescent bacteria for high-throughput analysis in the callus model.

To create a model suitable for high-throughput screening, we down-scaled our model to 96-wells plates and used bioluminescent *S. aureus* (Xen36) on the model. Different inoculum concentrations of the strain were applied to the model with or without 2% callus suspension added (Figs. 6B and 6C). The different inoculum concentrations of Xen36 all reached the same number of bacteria at day 1, except for the highest inoculum concentration. This trend is similar to the growth of non-luminescent *S. aureus* in 24-wells plates (Fig. 2C); the lower concentrations proliferate more than the higher concentrations. The bacteria show no growth when the model was used without callus.

Discussion

We demonstrate a newly developed stratum corneum model to study skin microbiota *in vitro*. The model is simple, affordable and amenable to high-throughput upscaling. Our model can be applied to study the growth of a single bacterial strain, a mixture of bacteria or a complete human skin microbiome. The PMA-qPCR analysis method ensures that only DNA from viable bacteria is quantified. Moreover, the model can be used to study dynamics of microbial communities e.g. following invasion by pathogenic micro-organisms or application of antimicrobial compounds.

The callus-based skin stratum corneum model has advantages in studying microbial growth compared to human constructed skin equivalents. These cellular models are incubated at 37°C, a temperature that strongly favors bacterial growth. This is, however, not in accordance with the actual temperature of 32°C typically present on exposed skin. The model presented here can be prepared in advance and stored at 4°C. It consists only of three components (agar, callus and bacteria), and no further expensive matrix and medium components are necessary. The bacteria are inoculated and grown on a dry surface and at 32°C. Our model is primarily designed to study the survival and growth of bacteria on the stratum corneum of the skin. When interaction of these bacteria with 'living' epidermal cells needs to be studied, *in vitro* skin equivalents are the preferred model. However, addition of recombinant proteins that play a role in the normal epidermal host defense system to our model might be a good alternative to study the effect of these molecules on the bacterial colonization/composition.

We demonstrated that the callus we used contains sufficient nutrients for the growth of the most relevant strains and even for an entire skin microbiome. From our data, we deduced that callus probably does not contain much sebum components which are present *in vivo* on specific locations of the human skin [258]. Therefore, bacteria, like *P. acnes*, which live on sebaceous areas of the human skin, are able to survive for just one week on our model. Addition of sebaceous components in future experiments might therefore be an option to improve the growth of *P. acnes* on our model. We have also shown that *S. epidermidis* and *S. aureus* are not able to survive for more than one day without callus (Supplementary Fig. S2). However, it is not clear which nutrients from these dead corneocytes are required for bacterial growth. Certainly, lipids in the callus are important factors for the bacteria, as we observed less bacterial proliferation when

delipidized callus was used in the model (data not shown).

We created a 'minimal' microbiome on our model consisting of *S. epidermidis* and *P. acnes*, and added pathogens (*S. aureus* and *P. aeruginosa*) to both of these commensals. As our model is suitable to study the survival/growth of these pathogens, the effects of possible treatments on the pathogen as well as the commensal microbiota could be examined.

Moreover, we demonstrated that a complete *in vivo* skin microbiome can be inoculated on our model. Importantly, the bacterial diversity was relatively stable over a period of 7 days (bacterial alpha diversity remained constant as shown in Fig. 5B). Due to the low inoculation concentration, the *in vivo* microbiome was able to increase some log-scales, but this did not strongly affect the bacterial diversity and composition, except for the number of *Propionibacterium* spp., which decreased in time, probably due to lack of sebum or lipid components in callus as discussed above. Furthermore, we observed the presence of *Pseudomonas* species, which were infrequent to absent in our previous study [3]. However, high levels of *Pseudomonas* on human skin were also reported by others [259]. Our skin microbiome has a high degree of interpersonal variation and therefore it might be coincidental that the two samples in the present study contain higher *Pseudomonas* levels than the individuals in our previous studies. Alternatively, it might be possible that the *Pseudomonas* sequences are derived from contaminations in aqueous solutions from the DNA isolation kit. However, the microbiome samples on day 7 do not contain high levels of *Pseudomonas* species. As the microbial DNA from these samples is isolated with the same kit, contamination is unlikely.

One of the main advantages of this *in vitro* skin stratum corneum model is the ability to control the proliferation rates of bacteria applied to the model. When we lowered the amount of inoculum, the bacteria showed an increased growth at the first day (except for *P. acnes*) (Fig. 2). This results in a more dynamic, infection-like model, suitable to test antibiotics. As an example we used tetracycline to evaluate its antimicrobial properties on the model; the MIC value of tetracycline in medium is 2 µg/ml for *S. aureus* and 64 µg/ml for *S. epidermidis* (not sensitive). Comparable values are found when *S. epidermidis* and *S. aureus* are exposed to tetracycline on the model. We therefore concluded that this model can be used to assess the antibacterial effects of antibiotics and antimicrobial proteins.

Furthermore we down-scaled our model to a 96-wells format, using bioluminescent bacteria. Bioluminescent bacteria (*S. aureus* Xen36 strain) could be detected on the model in a rapid and non-laborious way. The survival trend of the bioluminescent Xen36 strain is similar to the non-luminescent *S. aureus* strain that we used in 24-well plates (Fig. 2C) suggesting that our model can be used as a high-throughput system.

Possible applications of our stratum corneum model include intervention experiments to change disease-related microbiota compositions for example by prebiotics, probiotics or targeted antibiotics. In conclusion, we postulate that the *in vitro* skin stratum

corneum model presented here will contribute to the understanding of diseases linked to bacterial colonization and/or infection.

Methods

Callus collection and preparation

Human callus from the heel of three healthy volunteers was collected using a callus rasp (Ped Eggtm). This callus was mixed, frozen in liquid nitrogen, and subsequently grinded using a Micro Dismembrator U (B. Braun Biotech International, Melsungen, Germany) as previously described [260]. Phosphate buffered saline (PBS; Fresenius Kabi GmbH, Graz, Austria) was added to the callus powder to create a 2% suspension, which was sterilized by exposure to gamma radiation (16.2 kGray/63 h). The sterilized callus suspension was stored at 4°C or -20°C until further use.

Preparation of the *in vitro* skin model

One millilitre or 100 µl of sterile agar (2% in PBS; Becton, Dickinson and Company (BD), Sparks, MD) was added in wells of respectively 24-well and 96-well cell culture plates. On top of this agar, 100 µl (24-well) or 20 µl (96-well) of sterile callus suspension (2% in PBS) was applied. The plates were allowed to dry for 24 h at 37°C, and stored at 4°C until further use. For schematic representation see [Fig. 1](#).

Bacterial cultures

Staphylococcus epidermidis (*S. epidermidis*, ATCC 12228), *Propionibacterium acnes* (*P. acnes*, ATCC 6919), *Staphylococcus aureus* (*S. aureus*, ATCC 29213), *Pseudomonas aeruginosa* (*P. aeruginosa*, ATCC 27853) and *Streptococcus pyogenes* (*S. pyogenes*, ATCC 12344) strains were obtained from the department of Medical Microbiology of the Radboudumc. Bioluminescent *S. aureus* (Xen36) was purchased from Caliper Life Sciences, Boston, MA. Bacteria were inoculated on Columbia agar with 5% sheep blood (BD) overnight (o/n) at 37°C. One single colony of each plate was picked and cultured in Brain Heart Infusion medium (Mediaproducs BV, Groningen, The Netherlands) o/n at 37°C, except for *P. acnes*, which was cultured in thioglycollate medium (BD) under anaerobic conditions for two days at 37°C. Bacterial suspensions were diluted 10 times in medium and allowed to grow for another 3 hours to reach exponential bacterial growth (except for *P. acnes*). Hereafter, the bacteria were collected by centrifugation, washed two times with PBS and finally resuspended in PBS resulting in bacterial concentrations of 10^4 – 10^{12} CFU/ml. Portions of 20 µl or 5 µl of bacteria suspension were added to each well of the callus model (24-wells and 96-wells plate, respectively). When a mixture of bacterial strains was examined on the model, the same amount of every strain was used. The bacteria on the model were incubated at 32°C for different time points up to one week maximal.

CFU counting

The entire model (agar + callus + bacteria) was lifted out of the 24-wells plate and transferred to a 50 ml tube containing 10 ml of PBS ([Supplementary Fig. S1](#)). The tubes

were vortexed at maximum speed for one minute to detach and suspend the bacteria. The aqueous solution containing the bacteria (including some callus particles, but without the agar) was transferred to a new tube. These samples were serially diluted in steps of 10. Ten μl of each dilution was placed on sheep blood agar plates and incubated o/n at 37°C. Next day, colonies visible on the plate were counted for each dilution.

Microbiome samples

Microbiome samples from the lower back of two male volunteers were collected by swabbing the skin with Sample Collection Swabs (Epicentre Biotechnologies, Madison, WI) as previously described [3]. The swabs were collected in PBS, centrifuged for 5 min at 5000 rpm, resuspended in 60 μl of PBS and divided over the model in fractions of 20 μl . The model was incubated at 32°C for one week. The samples were collected for PMA-qPCR analysis and PMA-Illumina sequencing on day 0, 1 and 7.

Antibiotics

Tetracycline (Sigma-Aldrich, Zwijndrecht, The Netherlands) was added in different concentrations to the agar (2% agar in PBS). Callus and bacterial suspensions were added to the model as described above. The bacterial survival was assessed after 24 h of culturing by counting CFU.

Propidium monoazide treatment

Propidium monoazide (PMA) treatment of collected bacteria was performed to eliminate microbial gDNA originating from non-viable cells. Bacteria were collected from the model and finally resuspended in 500 μl of PBS to which 1.25 μl of PMA (20 mM, Biotium, Hayward, CA) was added. These mixtures were incubated for 10 minutes in the dark, and exposed to light for 5 minutes using PhAST Blue equipment (GenIUL, Terrassa, Spain). Afterwards the samples were centrifuged for 5 minutes at 5000 rpm, and the pellet was resuspended in 300 μl Micro Bead solution (MO BIO Laboratories, Carlsbad, CA). To obtain “non-viable” cells, bacteria were heat-killed at 85°C for 45 min. Hereafter, the viability of these bacteria was assessed by culturing on blood agar plates (BD).

Microbial gDNA isolation

Microbial gDNA was extracted using the MO BIO Ultraclean Microbial DNA Isolation Kit (MO BIO Laboratories) with modifications as previously described [3]. Microbial gDNA samples were stored at -20°C until further processing.

Quantitative Polymerase Chain Reaction

Microbial gDNA was used as template for qPCR amplification with SYBR Green using the Bio-Rad CFX Connect apparatus (Bio-Rad, Hercules, CA). Details of qPCR and design of strain-specific primers can be found in [Supplementary Methods](#).

16S amplification prior to sequencing

Microbiota samples derived from skin of the lower back contained small amounts of

microbial gDNA. Therefore, we introduced a pre-amplification step of the V3-V6 regions of the 16S rRNA gene as we did previously [3]. Details of amplicon generation and used primers can be found in [Supplementary Methods](#).

16S metagenomic library preparation and Illumina sequencing

Illumina 16S metagenomic amplicon libraries were generated and sequenced at BaseClear BV (Leiden, The Netherlands). A brief description of the procedure can be found in [Supplementary Methods](#).

Analysis of the Sequencing data

Demultiplexed FASTQ files as provided by BaseClear were first used to generate Illumina paired-end sequence pseudoreads by PEAR [223] using the default settings. The resulting pyrosequencing data were analyzed with a customized QIIME v1.2 [42] workflow, as described previously described [3]. Hierarchical (UPGMA) clustering was performed with weighted UniFrac as its distance measure; figures resulting from these clustering analyses were generated using the interactive tree of life (iTOL) tool. The PD whole tree alpha diversity was calculated by bootstrapping 43,592 reads per sample, and taking the average over four bootstrap trials.

Analysis of bioluminescent Xen36

The bioluminescence of the *S. aureus* Xen36 strain on the model was detected with the In Vivo Imaging System (IVIS, PerkinElmer, Waltham, MA) apparatus, and analysed with Living Image 3.1 software (PerkinElmer, Waltham, MA).

Supplementary Methods

([online](#)) Appendix S1 of online repository at :

http://www.medicaljournals.se/acta/content_files/additional_content/4688S1App.pdf

Supplementary Figures

Figure S1. Bacterial collection and analysis by CFU counting. (PDF)

([online](#)) Figure S1 of online repository at :

http://www.medicaljournals.se/acta/content_files/additional_content/4688S1Fig.pdf

Figure S2. Effect of the presence of callus on bacterial survival. (PDF)

([online](#)) Figure S2 of online repository at :

http://www.medicaljournals.se/acta/content_files/additional_content/4688S2Fig.pdf

Figure S3. Bacterial collection and analysis by qPCR analysis. (PDF)

([online](#)) Figure S3 of online repository at :

http://www.medicaljournals.se/acta/content_files/additional_content/4688S3Fig.pdf

Figure S4. The effect of light exposure on bacterial survival. (PDF)

(online) Figure S4 of online repository at :

http://www.medicaljournals.se/acta/content_files/additional_content/4688S4Fig.pdf

Figure S5. The effectiveness of PMA treatment. (PDF)

(online) Figure S5 of online repository at :

http://www.medicaljournals.se/acta/content_files/additional_content/4688S5Fig.pdf

Figure S6. Correlation between Ct value and CFU/ml. (PDF)

(online) Figure S6 of online repository at :

http://www.medicaljournals.se/acta/content_files/additional_content/4688S6Fig.pdf

Supplementary Tables

Table S1. Primers for qPCR. (PDF)

(online) Table S1 of online repository at :

http://www.medicaljournals.se/acta/content_files/additional_content/4688S1Tab.pdf

CHAPTER 5

OPEN ACCESS

adapted from as published in J Allergy Clin Immunol, 2017, Apr;139(4):1368-1371

<https://doi.org/10.1016/j.jaci.2016.09.017>

¹ Department of Dermatology, Radboud Institute for Molecular Life Sciences (RIMLS), Radboud University Medical Center (Radboudumc), Nijmegen, The Netherlands.

² Center for Molecular and Biomolecular Informatics, RIMLS, Radboudumc, Nijmegen, the Netherlands.

³ Coronel Institute of Occupational Health, Academic Medical Center, Amsterdam, The Netherlands.

⁴ Department of Dermatology, Canisius Wilhelmina Hospital, Nijmegen, The Netherlands.

⁵ Division of Cancer Science, School of Medicine and Division of Biological Chemistry and Drug Development, School of Life Sciences, University of Dundee; A*STAR Institute, Singapore.

⁶ Wageningen University, Host-Microbe Interactomics Group, Wageningen, The Netherlands.

⁷ NIZO, Ede, The Netherlands.

These authors contributed equally to this work.

CHAPTER 5

GRAM-POSITIVE ANAEROBE COCCI ARE UNDERREPRESENTED IN THE MICROBIOME OF FILAGGRIN-DEFICIENT HUMAN SKIN

Patrick L.J.M. Zeeuwen ^{1 #}
Thomas H.A. Ederveen ^{1,2,7 #}
Danique A. van der Krieken ^{1 #}
Hanna Niehues ^{1 #}
Jos Boekhorst ²
Kezic S ³
Hanssen DA ¹
Otero ME ¹
Ivonne M. van Vlijmen-Willems ¹
Diana Rodijk-Olthuis ¹
Falcone D ¹
Ellen H.J. van den Bogaard ¹
Kamsteeg M ¹
de Koning HD ¹
Zeeuwen-Franssen ME ⁴
van Steensel MA ⁵
Michiel Kleerebezem ⁶
Harro M. Timmerman ⁷
Sacha A.F.T. van Hijum ²
Joost Schalkwijk ¹

ABSTRACT

Genetic deficiency or haploinsufficiency of the histidine-rich epidermal protein filaggrin (FLG) is associated with ichthyosis vulgaris (IV) and atopic dermatitis (AD). We investigated if the *FLG* genotype affects the microbiota composition of human skin and the cutaneous host response. *FLG*^{-/-} individuals had a low abundance of proteolytic Gram-positive anaerobic cocci (e.g. *Finnegoldia magna*), which use peptides as nutrient sources. Furthermore, genes involved in histidine utilization were underrepresented in *FLG*^{-/-} microbiota. An *in vitro* survival disadvantage of *F. magna* on FLG-deficient stratum corneum supported the *in vivo* findings. In co-cultures with primary human keratinocytes, *F. magna* induced a stronger antimicrobial peptide response than *Staphylococcus aureus*, whereas *S. aureus* caused a stronger proinflammatory cytokine response than *F. magna*. Our data indicate that a common genetic defect can shape the cutaneous microbiome based on metabolic requirements of certain taxa, and may alter the host response to pathogens.

Background

Next Generation Sequencing technologies and powerful bioinformatics tools have enabled comprehensive analysis of microbiota of human tissues. Large scale studies of the composition of microbial communities on skin of healthy volunteers have revealed that the four most abundant phyla of human skin bacteria are Actinobacteria, Firmicutes, Proteobacteria and Bacteroidetes. In addition, these studies found that the diversity of microorganisms on our skin is much larger than was previously assumed from culture-based methods [106, 232], and that diversity and composition of the skin microbiota strongly depends on the topographical location on the body and has a high degree of interpersonal variation. Nevertheless, the dominant types of bacteria remain relatively stable over time, and specific bacteria are associated with dry, moist and/or sebaceous microenvironments [151, 236, 258]. Microbial communities, genetic host factors and the environmental factors at a particular moment, constitute a complex relationship that is essential for skin barrier homeostasis [234, 245, 261]. The first reports on the human skin microbiota were of healthy volunteers, but more recent studies have also focused on the microbiota of diseased and injured skin [3, 239].

There are numerous monogenic human skin conditions (ichthyoses, blistering skin diseases) that alter the properties of the stratum corneum (SC) in such a way that it could affect the SC microbial ecology, directly or indirectly. However, none of these have been studied at the microbiome level, to date. Here, we selected ichthyosis vulgaris (IV) as a model disease to investigate if genetic polymorphisms resulting in altered SC composition and structure could lead to distinct changes in skin microbiome makeup. IV is caused by loss-of-function (LOF) mutations in the *flaggrin* (*FLG*) gene, which are quite common in the general population [262]. *FLG* LOF mutations are also a strong genetic risk factor for atopic dermatitis (AD) [263]. *FLG* is a histidine-rich structural skin protein that is abundantly expressed in the outer layers of human epidermis, where it acts as a glue-like protein that facilitates dense packing of keratin filaments in terminally differentiating keratinocytes [264]. Furthermore, upon degradation, *FLG* serves as the main source of so-called natural moisturizing factors (NMFs), which includes free amino acids like histidine and metabolites such as urocanic acid (UCA) and pyrrolidone carboxylic acid (PCA), that allow the outermost layers of the SC to remain hydrated [265]. IV is characterized by a dry and scaly skin [266] and is associated with abnormalities in epidermal structure and function [267], but no microbiological abnormalities have been reported.

Here, we analysed the skin microbiota of IV patients of the lower leg, which is typically a location where the ichthyotic skin alterations are most prominent. Our data indicated that Gram-positive anaerobic cocci (GPAC) such as *Finegoldia magna* (*F. magna*), are strongly decreased in the skin microbiome of IV patients. Gene expression analysis of epidermal keratinocytes *in vitro* exposed to *F. magna* or the AD-associated bacterium *Staphylococcus aureus* (*S. aureus*), revealed distinct early host defense responses depending on the microbial stimulation.

Results

Differences in skin microbiota composition between filaggrin deficient individuals and healthy controls.

We recruited 17 patients with the clinical diagnosis of IV (see [Supplementary Table S1](#) for details). Patients were genotyped for the two most common *FLG* mutations in the Western population: R501X and 2282del4. In addition, we performed FLG protein staining on skin biopsies ([Fig. 1](#)).

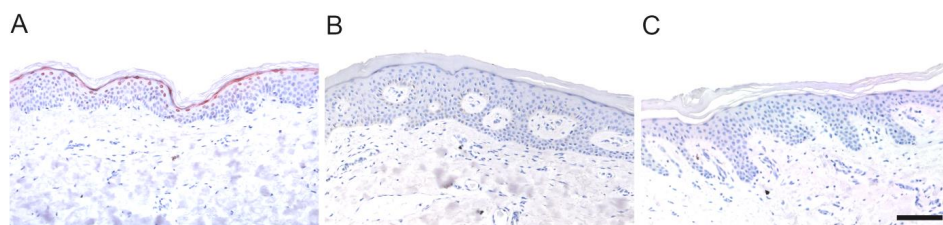


Figure 1. Filaggrin staining of upper buttock skin.

(A) Healthy controls ($FLG^{+/+}$) show filaggrin staining in the stratum granulosum layer of the epidermis. Both keratohyalin granules and nuclei are positive. (B) Skin from IV patients homozygous or compound heterozygous for the *FLG* R501X and 2282del4 mutations ($FLG^{-/-}$) were negative for FLG protein staining. (C) Genotyped patients wherein only one mutation was found (R501X, for $n = 2$), were tested for FLG expression. When negative, the patient still has to have an unknown mutation somewhere else in the *FLG* gene (i.e. R501X?) and consequently was classified in the *FLG* deficient group ($FLG^{-/-}$). Scale bar = 100 μm .

We identified *FLG* mutations in all 17 patients. All homozygous or compound heterozygous patients were negative for FLG protein staining. Notably, four patients with a clinical diagnosis of IV were heterozygous for *FLG* mutations and had residual FLG protein staining in skin biopsies (designated $FLG^{+/-}$). This small group was analysed separately. We analysed the microbiota composition of the 13 $FLG^{-/-}$ patients, the four $FLG^{+/-}$ individuals and 10 healthy $FLG^{+/+}$ controls ([Supplementary Table S1](#)). These groups did not significantly differ regarding age and sex. The lower leg was selected for biophysical measurements, as this is the area where the ichthyotic phenotype (dry and scaly skin) is most evident (15). A large proportion of IV patients has concomitant AD, a disease known to be characterized by microbiome alterations and colonization of the lesional skin by *S. aureus*. However, none of the IV patients in our study had eczematous lesions present at the lower leg, thereby excluding (lesional) AD-associated microbiome alterations as a confounding factor. Biophysical measurements showed increased transepidermal water loss (TEWL), decreased SC hydration, and an equal skin surface pH value in $FLG^{-/-}$ subjects compared to $FLG^{+/-}$ subjects ([Supplementary Fig. S1](#)). Microbiome samples were taken from the same location by swabbing, and subsequently analysed by barcoded 16S marker gene sequencing. In total, 43,137 bacterial 16S rRNA sequences were analysed, resulting in an average of $1,598 \pm 546$ (range 811-3,471) reads per sample ([Supplementary Tables S2 and S3](#)). Taxonomic composition analysis revealed that 99.5% of the operational taxonomical units (OTUs) could be assigned to the phylum level, and 90.6% could be assigned to the genus level

([Supplementary Table S2](#)). In total, four-hundred and forty-seven different OTUs were detected (represented by at least five sequencing reads). Rarefaction curves show that the phylogenetic diversity does not differ between *FLG*^{+/+}, *FLG*^{-/-} and also *FLG*^{+/-} samples ([Supplementary Fig. S2](#)). Overall, no differences in bacterial diversity were found between the samples for all four diversity-metrics tested (see [Supplementary Fig. S3](#) for whole tree phylogenetic diversity, observed OTUs, Shannon, and Chao1). Redundancy Analysis (RDA) revealed a significant effect of FLG deficiency on microbiota composition ([Fig. 2](#); permutation test *p*-value = 0.034).

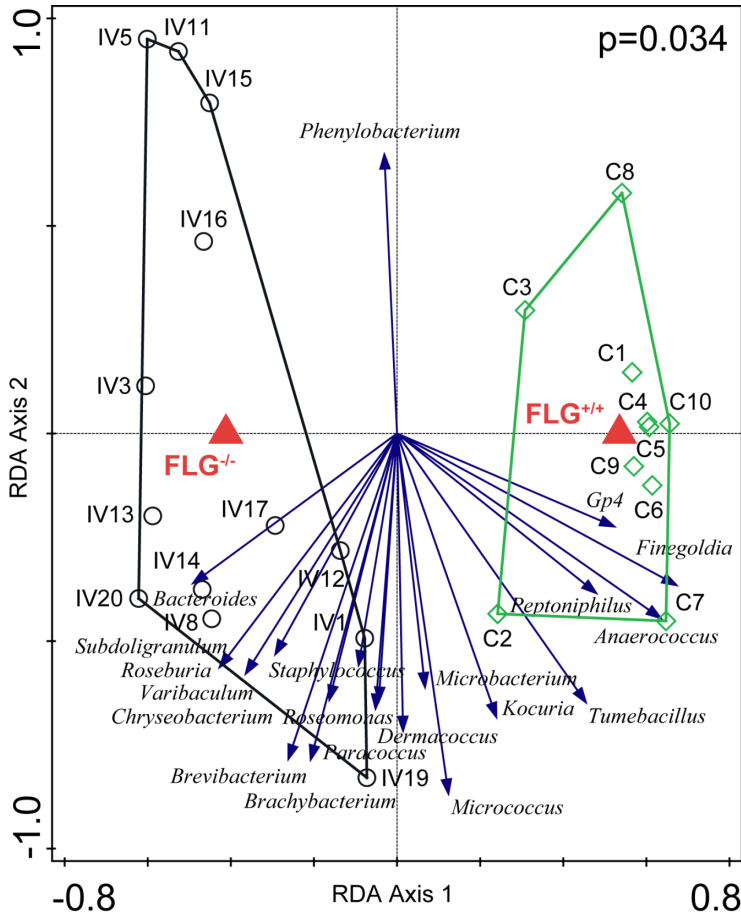


Figure 2. Skin microbiota from *FLG*^{+/+} and *FLG*^{-/-} individuals are significantly different.

The figure shows a Redundancy Analysis (RDA) triplot. Green diamonds and black circles represent *FLG*^{+/+} and *FLG*^{-/-} samples, respectively. Red triangles are the centroids of the corresponding sample groups. The blue arrows are the 20 best-fitting genera (names in *italics*), i.e. best explaining microbiota compositional differences between the *FLG* contrast. The horizontal axis maximizes the variation in sample groups (in contrast to a principal component analysis plot, where the variation between individual samples is maximized). The difference in microbiota is significant (according to a permutation test; *p* = 0.034), i.e. randomly assigning samples to sample groups and repeating the RDA analysis produces a plot where the difference between sample groups is smaller than in the real data.

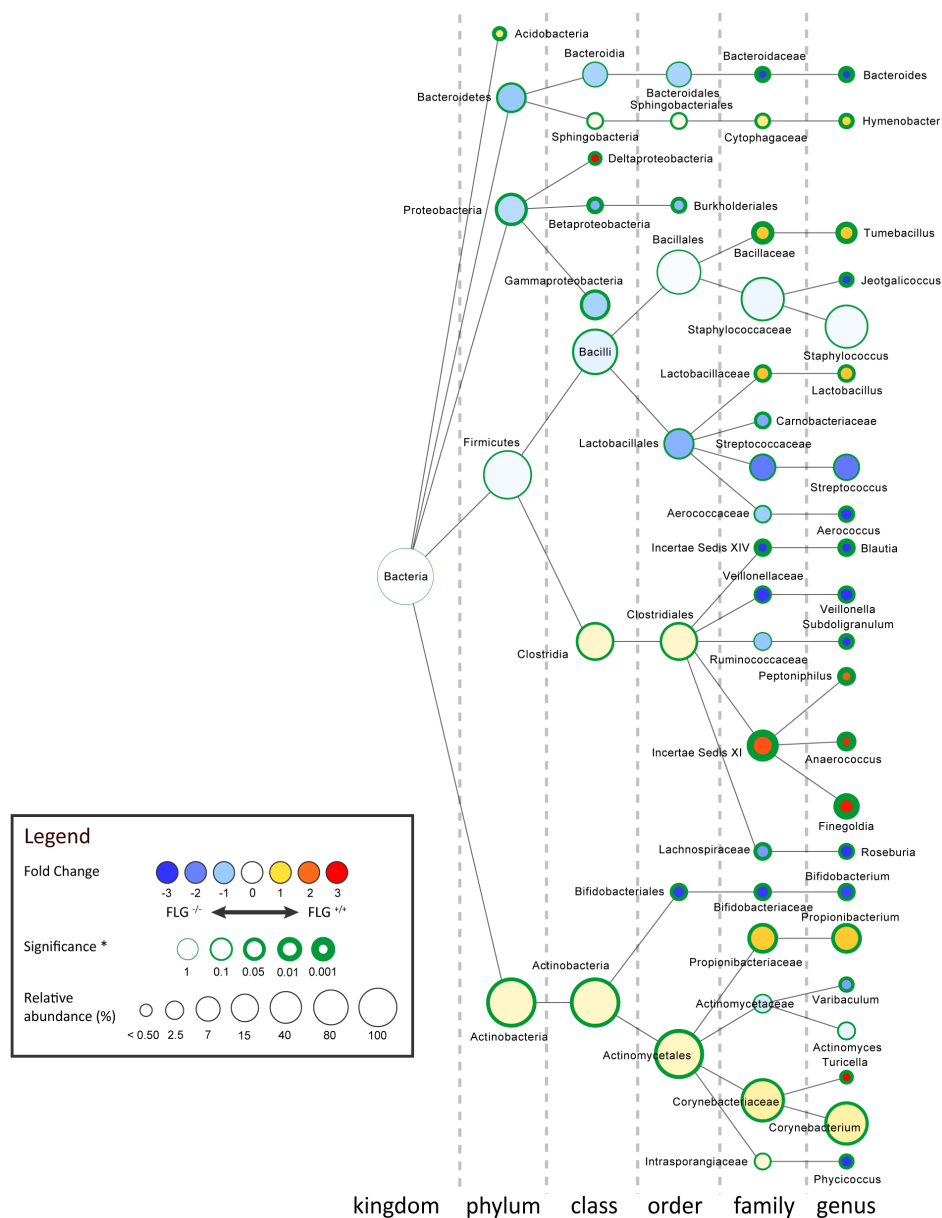


Figure 3. Difference in microbial community composition between FLG^{+/+} and FLG^{-/-} individuals.

Nodes represent taxa, edges link the different taxonomic levels. Node colour indicates the difference in abundance (red is more abundant in FLG^{+/+}, blue in FLG^{-/-}), calculated as the $2\log$ of the ratio of the relative abundance (0 = no difference between genotypes, 1 = twice as abundant in FLG^{+/+}, and so on). The significance is represented by the thickness of the node borders and expressed as the p -value of a Mann-Whitney U test of the samples. The size of a node represents the average relative abundance of a taxon (note that the relation between node-size and abundance is non-linear).

In accordance, we used taxonomy-based OTU classification provided by the Ribosomal Database Project combined with statistical mining to identify significant differences in microbiota composition (Fig. 3 and Supplementary Table S6) [225]. A lower relative abundance of proteolytic GPAC of the family Incertae-Sedis-XI (average 1.7% of the total genera) was found in *FLG*^{-/-} skin compared to *FLG*^{+/+} skin (average 9.3% of the total genera). This family included the genera *Finegoldia*, *Anaerococcus*, and *Peptoniphilus* and all belong to the class Clostridia (phylum Firmicutes), and no other genera were identified within this GPAC family above 0.1% of total reads. Most GPAC are asaccharolytic and use the products of protein degradation as substrates for metabolic energy generation [268]. The average relative quantities of these candidate discriminating genera (Fig. 4) supported a significant difference between the *FLG* deficient and proficient state, or more specifically, a strong decrease by effect of *FLG* protein loss. Interestingly, in *FLG*^{+/-} individuals these three genera tended to be present at intermediate relative abundance quantities, which suggests a *FLG* gene dosage effect. Other highly abundant genera present in the skin microbiome, such as *Staphylococcus*, *Propionibacterium* and *Corynebacterium* or the AD-associated species *S. aureus* did not show significant differences between the genotypes (see Supplementary Table S6), which is in accordance with the lack of eczematous lesions that would be expected to be accompanied by elevated relative abundance levels of *S. aureus*. We therefore focused on members of the GPAC group for further *in silico* analysis and experimental studies.

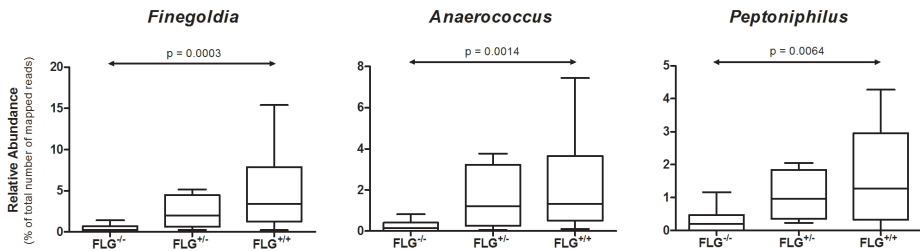


Figure 4. Differences in relative abundance between filaggrin genotypes for the genera *Finegoldia*, *Anaerococcus* and *Peptoniphilus*.

Horizontal double-headed arrows indicate significant differences (Mann-Whitney U test).

Filaggrin deficiency influences stratum corneum NMF content and *in vitro* growth of *Finegoldia magna*.

As the absence of *FLG* in the terminally differentiating keratinocytes can lead to lower levels of peptides and amino acids (NMFs) in the SC as reported for skin of the lower arm [269], we hypothesized that the low abundance of proteolytic GPAC (as indicated in Figs. 3 and 4) could be the result of *FLG* deficiency and decreased availability of its breakdown products. Plantar callus can serve *in vitro* as a sole carbon and nitrogen source for (commensal) microorganisms in order to survive and grow [270]. We investigated the growth of the GPAC type strain *F. magna*, a member of the genus that was most significantly different between *FLG*^{-/-} skin compared to *FLG*^{+/+} skin, in our recently developed *in vitro* system that utilizes human SC for bacterial growth (Chapter 4: Fig.

1), and which can be used to study individual microorganisms or (*ex vivo*) microbial ecology [270]. We found that FLG is present in normal plantar skin, which could therefore be conveniently used as a source of FLG-containing SC (data not shown). We used callus from healthy *FLG*^{+/+} controls and from IV patients homozygous for *FLG* LOF mutations (*FLG*^{-/-}). As the composition of FLG breakdown products in human plantar callus was unknown, we analysed the NMF content of callus from patients and healthy subjects by high-performance liquid chromatography (HPLC). Indeed, we found that patient-derived callus contained significantly lower total NMF levels (Fig. 5A).

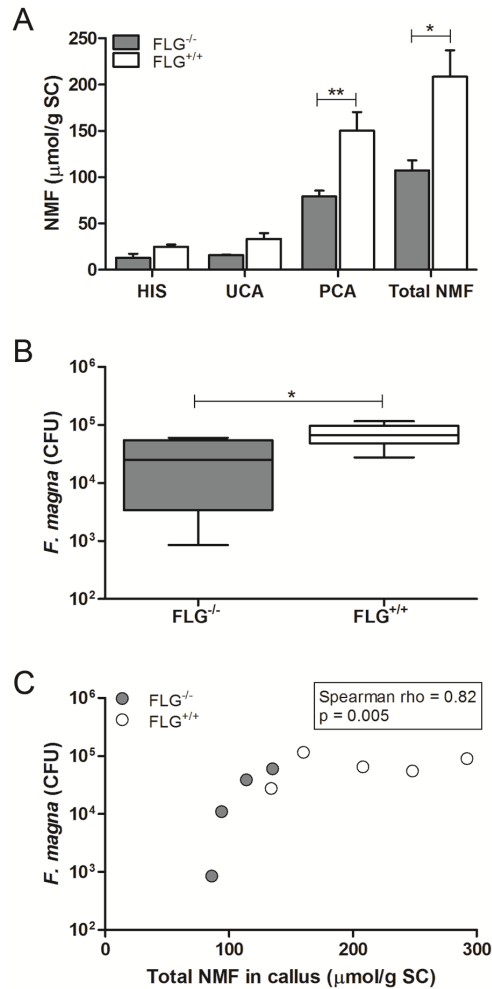


Figure 5. Lower total NMF levels in callus affect *in vitro* growth of *Fingoldia magna* bacteria.

(A) NMF content analysis of callus derived from *FLG*^{-/-} (n = 4) and *FLG*^{+/+} (n = 5) individuals. Data are represented as mean ± SEM. Mann-Whitney U test. **p* < 0.02, ***p* < 0.01 (HIS, *p* = 0.06; UCA, *p* = 0.08). HIS, histidine; PCA, pyrrolidone carboxylic acid, UCA, urocanic acid. (B) *F. magna* growth on the *in vitro* stratum corneum model using callus derived from *FLG*^{-/-} (n = 4) and *FLG*^{+/+} (n = 6) individuals. Data are represented as mean ± SEM. Mann-Whitney U test. **p* < 0.03. (C) Correlation between the growth rate of *F. magna* and the NMF content of *FLG*^{-/-} and *FLG*^{+/+} callus on which the bacteria are grown.

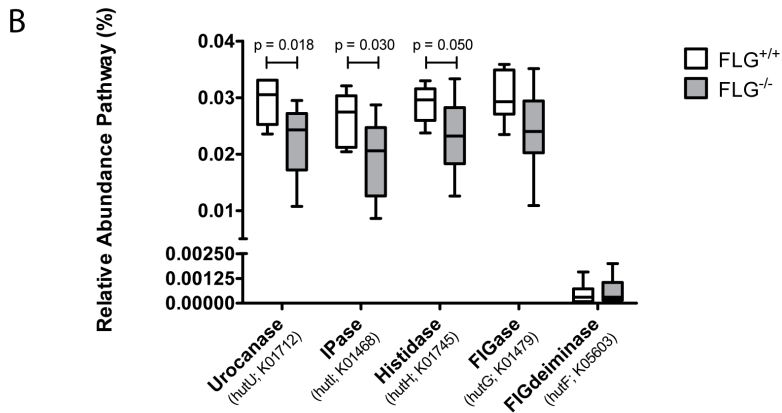
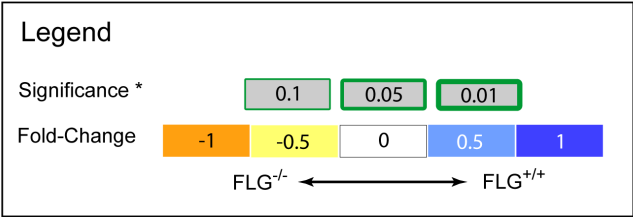
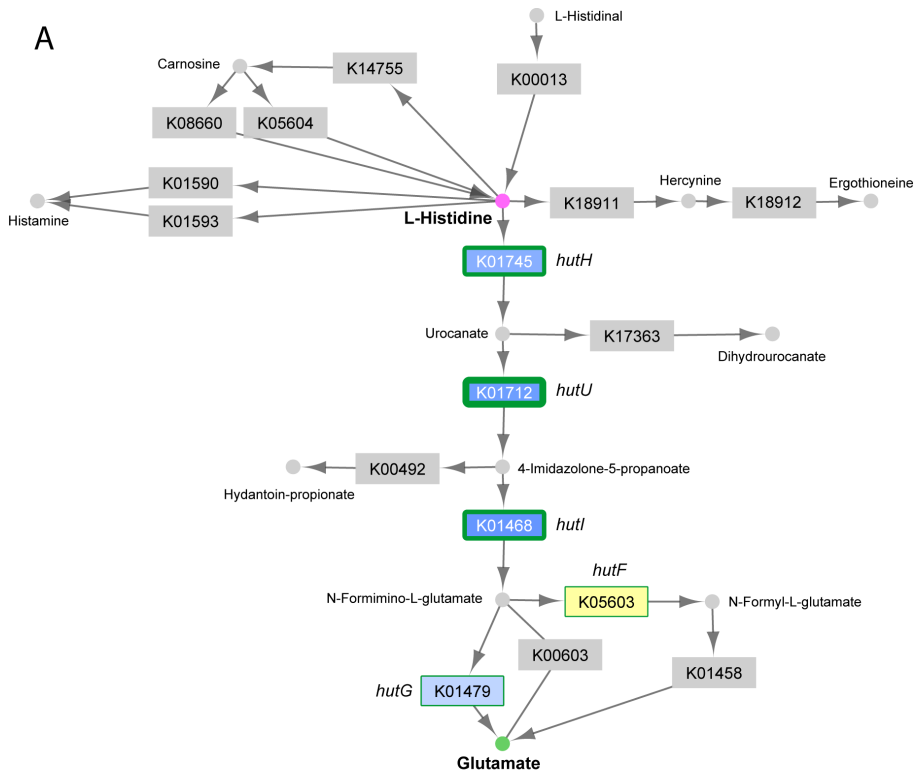
Approximately 2×10^3 *F. magna* bacteria were subsequently inoculated on the *in vitro* SC model and incubated at 32°C under anaerobic conditions. Bacterial growth of *F. magna* was determined after one week by enumerating colony forming units (CFUs). The data showed that callus derived from IV patients deficient for FLG is a less favourable substrate for these bacteria than callus from healthy individuals with normal FLG protein expression (Fig. 5B), as concluded from the lower number of CFUs for *F. magna* when grown on FLG^{-/-} callus. A positive correlation (Spearman rho = 0.82, *p* = 0.005, see Fig. 5C) was found when we plotted the *F. magna* growth of all samples versus the total NMF content (i.e. PCA, UCA and histidine combined).

Hut genes are underrepresented in the microbiota of flaggrin deficient skin.

We tested the potential involvement of microbial reactions utilizing NMF factors histidine, UCA and PCA. According to KEGG (Kyoto Encyclopedia of Genes and Genomes) [271], two orthology groups utilize PCA (KEGG compound name C01879; Pidolic acid): KEGG Orthologs K00682 and K01469. Only K01469 was predicted to be present in the microbiomes. Next to its very low relative abundance (0.0003%, on average) it did not differ significantly between FLG status. The microbial reactions utilizing histidine and UCA are the same, as UCA (urocanate; see Fig. 6) is an intermediate product in the conversion of histidine to glutamate. The human profilaggrin polypeptide has a high histidine content (413/4061 residues; 10.2%) [272]. Histidine is a known carbon source for bacteria that possess the histidine utilization (Hut) pathway, which is highly conserved among bacteria and found with high frequency in most phylogenetic groups within the bacterial domain [273]. This notion, together with the observation of the low abundance of proteolytic GPAC in FLG deficient patients, prompted us to examine the presence of Hut pathway genes in the skin microbiota of IV patients compared to healthy controls (Supplementary Table S7). The Hut operon contains several *hut* genes present as orthologs in a wide range of bacterial species, that allow loss-of-function for conversion of histidine to glutamate by different routes (Fig. 6A) [273].

Figure 6. (next page) Function analysis of the skin microbiota reveals a decreased histidine utilization capacity in FLG deficient patients.

(A) KEGG map of histidine metabolism (snapshot) and the candidate *hut* genes. Nodes represent compounds, the pink node represents our compound of interest: histidine; the green node represents the Hut pathway its end product: glutamate. Boxes represent enzyme functions/reactions, numbers shown inside are K numbers (i.e. KEGG Orthologs). Coloured boxes correspond to *hut* gene candidates that were *a priori* selected for function analysis by PICRUST: blue indicates a decrease in the relative abundance of a *hut* gene as effect of FLG deficiency, and orange indicates an increase in these genes compared to wild type; gray boxes were excluded from analysis. This colour representation of the fold-change difference is calculated as the 2log of the ratio of the relative abundance between FLG wild type and knockout (0 = no difference between genotypes, 1 = twice as abundant in wild type, etcetera). The significance (box border width) is expressed as the *p*-value of a Mann-Whitney U test with Bonferroni's correction for multiple testing. Arrows link the conversion of one compound to another, facilitated by an enzyme function/reaction. (B) Box plots with relative abundances of the candidate *hut* genes for FLG^{+/+} and FLG^{-/-} genotypes. The relative abundance is calculated over all K numbers identified in the total microbiota.



In the first phase, three core *hut* genes encoding a histidase (*hutH*), a urocanase (*hutU*) and an imidazolone-5-propionate hydrolase (IPase, *hutI*) are involved in catabolic conversion of histidine to formimino glutamate (FIG). In the second phase, FIG is converted to glutamate, either directly by the *hutG* encoded FIGase, or indirectly via the intermediate formyl glutamate by the *hutF* encoded FIGdeiminase. These five *hut* genes were chosen as candidates in a contrast analysis by PICRUST (Phylogenetic Investigation of Communities by Reconstruction of Unobserved States), based on functional predictions derived from the 16S marker gene sequencing data [64]. KEGG Orthologs (K numbers) of genes in the Hut pathway were extracted from KEGG. All five *a priori* selected *hut* gene candidates (Supplementary Table S4) were predicted to be present in the microbiota of IV (*FLG*^{-/-} and *FLG*^{+/-}) and healthy controls (*FLG*^{+/+}). Interestingly, contrast analysis of IV versus healthy skin revealed that *hutU*, *hutI* and *hutH* orthologous genes were predicted to be significantly underrepresented in the microbiota of IV patients (Fig. 6B). Furthermore, in an overrepresentation analysis of the *FLG*^{+/+} versus *FLG*^{-/-} contrast, ranking all K numbers on their *p*-value (as determined by Mann-Whitney U test) from low to high resulted in a strong enrichment of the significant differentially abundant candidate *hut* genes K01712, K01468 and K01745 (Fig. 6) in the top list, at ranks 41, 80 and 116 respectively (Supplementary Table S4), out of a total of 5838 identified K numbers. Correspondingly, the non-significantly different *hut* genes K01479 and K05603 rank considerably lower at positions 379 and 4200, respectively. This overrepresentation data corroborates that the observed effect on *hut* genes as a result of the FLG deficiency is not a random one, and is in agreement with the observation that patient derived callus contained lower levels of histidine and its metabolites.

Distinct host defense responses depending on bacterial species.

FLG LOF mutations predispose to AD both in haploinsufficient individuals but even more strongly in homozygous IV patients. A plausible mechanism, but still largely hypothetical, is an effect on skin barrier which would allow for increased exposure to environmental allergens. Other mechanisms have not yet been investigated to our knowledge. As microbial factors such as *S. aureus* colonization have been implicated in AD, we explored possible interactions of the pathogen *S. aureus* and the commensal *F. magna* with epidermal keratinocytes. We investigated the effect of *S. aureus* and *F. magna* on normal human epidermal keratinocytes. We used a submerged monolayer keratinocyte model from 10 *FLG*^{+/-} donors stimulated with viable or heat-killed *S. aureus* and *F. magna* bacteria, and determined the keratinocyte host defense response following 10 hours of exposure to microorganisms. We selected 8 keratinocyte-expressed genes based on previous work or data from literature, which included the cytokines IL-1 β , IL-6 and TNF α , the chemokines IL-8 and CCL20, and the antimicrobial peptides (AMP) hBD2, hBD3 and LL37 (Fig. 7). Viable *F. magna* induced a significantly higher expression of AMPs and IL-1 β compared to *S. aureus*. For viable *S. aureus*, keratinocyte stimulation caused a significantly higher expression of IL-6, IL-8 and CCL20 by these cells than seen for stimulation with viable *F. magna*. Heat-killed *S. aureus* cells were in general less potent stimuli, whereas heat-killed *F. magna* showed equal gene expression levels as viable *F. magna*.

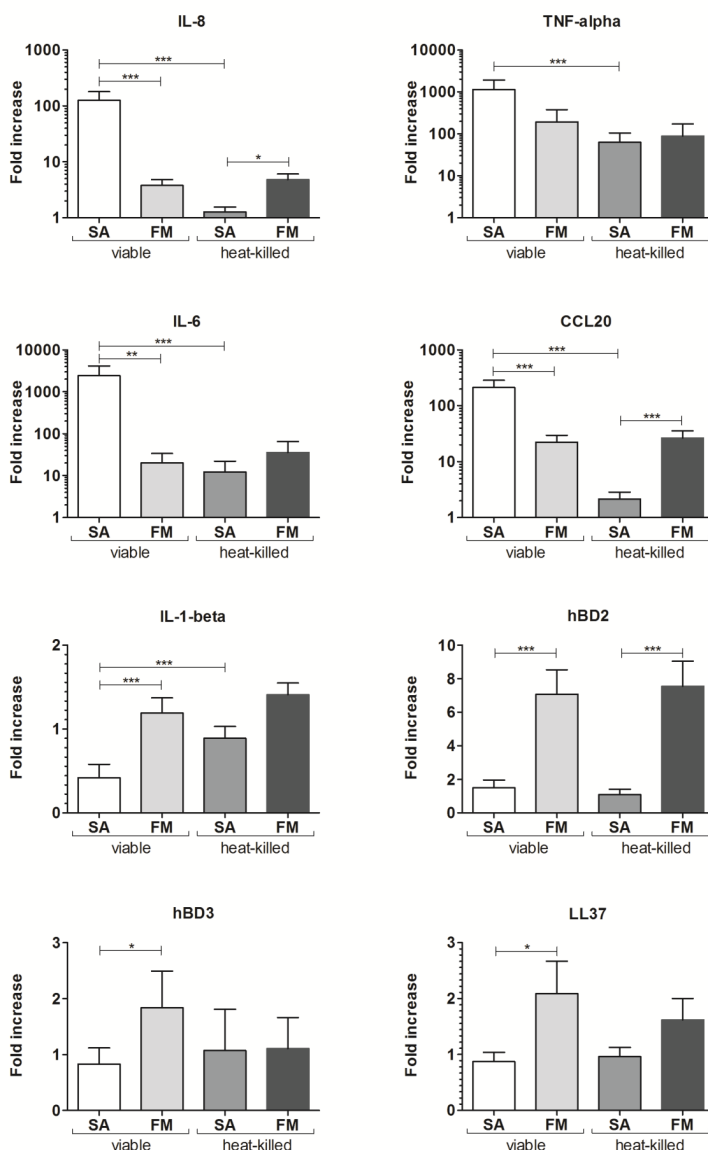


Figure 7. Keratinocyte host defense responses differ between bacterial stimulations.

Submerged keratinocytes (derived from healthy *FLG*^{+/+} controls, $n = 10$) were stimulated for 10 hours with viable and heat-killed *S. aureus* and *F. magna*. Expression of 8 host defense genes was measured by qPCR. For each gene, the control cell culture (not stimulated) was set to 1. The following bacteria-induced expression levels of host defense molecules were significantly different compared to their non-stimulated controls: IL-6, IL-8, TNF α , IL-1 β , and CCL20 levels in control versus viable *S. aureus* stimulation ($p < 0.001$), and hBD2 and CCL20 levels in control versus viable and heat-killed *F. magna* stimulation ($p < 0.001$). All other combinations of control compared to bacterial stimulation (viable and heat-killed) were not significantly different. Significant differences between bacterial species and viable versus heat-killed bacteria are indicated in the graphs. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Data are represented as mean \pm SEM. Repeated measures ANOVA with Bonferroni's post-hoc test on the Δ Ct values of the qPCR data.

Discussion

In this study, we report two novel findings with respect to genetic deficiency of the skin protein *FLG* and cutaneous microbiota. Firstly, we show that *FLG* deficiency is associated with a low relative abundance of proteolytic GPAC, and an underrepresentation of bacteria, not limited to GPAC taxa, capable of utilizing histidine. *In vitro* experiments support the hypothesis that there is a direct mechanistic link between the absence of *FLG* breakdown products and microbiome alterations. Secondly, we show that cultured keratinocytes express proinflammatory cytokines upon viable-microbiota stimulation by *S. aureus*, whereas *F. magna* stimulation revealed a rapid AMP response not seen upon *S. aureus* stimulation.

The composition of human microbiomes is strongly influenced by an array of environmental factors such as diet, lifestyle, antibiotics use, exposure to pathogens and social interactions. In addition to the contribution of environmental factors, evidence is accumulating to support a host genetic component as well. For example, this has been demonstrated for genetic polymorphisms that contribute to an increased predisposition to inflammatory bowel diseases [274] and Crohn's disease in humans [275]. Recently, analysis of host genetic variation data mined from shotgun metagenomic data from the Human Microbiome Project (HMP) revealed a correlation between host genetic variation in immunity-related pathways and microbiome composition [261]. These studies have uncovered human genes and pathways associated with particular genetic backgrounds, although, an unequivocal demonstration of causality still remains challenging in human population-based studies. To date, no human studies are available regarding specific genetic defects of SC composition and structure in relation to cutaneous microbiome composition. Recently, the skin microbiota of wild type mice versus caspase-14 deficient mice was compared [276]. Caspase-14, a protease mainly active in terminally differentiated keratinocytes, is required for the generation of NMFs by processing of *FLG* protein [277]. It was hypothesized that a caspase-14 deficiency reduced the levels of NMFs and could create an altered nutrient environment for commensal bacteria, which was supported by an imbalance of skin-microbiota in caspase-14 deficient mice [276]. However, these mice also displayed increased AMP expression after *Escherichia coli* challenge, which may suggest an alternative explanation for the observed skin microbiota changes in these mice.

Genome-wide approaches that find an association between genetic polymorphisms and microbiome composition should be rigorously controlled for confounding effects of non-genetic factors such as population stratification, which may lead to spurious correlations. Monogenic, Mendelian disorders are easier to study and can be better controlled. Previous studies have documented the strongly abnormal microbiomes of skin lesions in patients with mucocutaneous candidiasis and Hyper-IgE Syndrome (HIES), based on mutations in host defense genes [241, 278]. Deficient Th17 responses lead to uncontrolled growth of pathogens and overt infections. In these cases, there is a causal relationship between a genetic abnormality and infection. In our current study, a potential confounding factor could be the co-occurrence of AD, since heterozygosity for *FLG* null alleles is a known strong risk factor for AD in the general population,

and homozygosity for *FLG* LOF mutations, such as those in IV, is associated with a prevalence of AD of 60-70% in these patients [263]. The lesional skin of AD patients is often colonized by *S. aureus*, both in *FLG* null allele carriers and in *FLG*-proficient AD patients, suggesting that AD, and not *FLG* deficiency *per se* causes the colonization of this bacterium in lesional skin. In our study, although a large proportion of our patient cohort had a history of AD (65%, see [Supplementary Table S1](#)), none of them had active AD lesions in the skin areas that were sampled. This is also supported by the fact that dominance of *Staphylococci* was not observed in any of the samples, as would have been expected for lesional AD skin [239]. An additional RDA, based on genus-level compositional microbiota profiles of IV patients with and without comorbid AD, revealed that both patient groups (IV with/without AD) cannot be confidently separated based on microbiota composition ($p = 0.63$, p -value calculated by randomly permutating the samples' AD status), indicating that our cohort data is not biased by AD status of the patient sample. This may be due to the fact that samples are derived from a location that is not commonly affected by eczema.

The microbiome function prediction strongly supports a metabolic basis for the microbiota compositional changes observed in IV patients, since the shift in bacterial composition in *FLG* deficient relative to wild-type skin could be connected to reduced levels of *hut* genes in the microbiota gene pool. The specific contribution of the Incertae-Sedis-XI family GPAC to the total microbial content of *hut* genes is very low (0.09% for *FLG*^{-/-} and 0.36% for *FLG*^{+/-}, data not shown); which may not be surprising, as the relative abundance of the three candidate GPAC bacteria is relatively low (1.22% and 8.67% for *FLG*^{-/-} and *FLG*^{+/-}, respectively, see [Fig. 4](#)). Notably, the reduced levels of any of the candidate *hut* genes (i.e. *hutH*, *hutI* and *hutU*, see [Fig. 6](#)) did not correlate to the abundance of any single bacterial genus or family (data not shown), illustrating the broad phylogenetic distribution of the *hut* genes within the total microbiota, and exemplifying the importance of the decreased *hut* prevalence associated with *FLG* deficiency. In conclusion, our study reveals that *FLG* deficient skin harbours significantly fewer bacterial taxa capable of utilizing histidine as nutrient (carbon) source, which we postulate to be a consequence of nutrient competition driven by the loss of the histidine-rich protein *FLG*.

Although the primary scope of this study was to investigate a relationship between *FLG* deficiency and cutaneous microbiome alterations, an obvious and highly relevant question would be whether the shift of cutaneous microbiota could play a role in the development of AD in patients that carry *FLG* null alleles. It has to be noted, however, that *FLG* LOF mutations are neither a necessary nor a sufficient condition for the development of AD. This is illustrated by the observation that in Western Europe only 20-40% of AD patients carry *FLG* null alleles and up to 9% of healthy controls are carriers of *FLG* null alleles [279]. Clearly, other genetic causes and exposure to environmental factors (allergens, microbes) could be important in the development of AD [280]. It is, nevertheless, very well possible that microbiota changes due to *FLG* deficiency will affect the cutaneous immune system in such a way that a Th2 response against common environmental allergens is facilitated. Host-microbe and microbe-microbe interactions could alter the local skin microbiota composition, as resident skin

bacteria are able to control colonization by potentially pathogenic microorganisms [281, 282] and can modulate the cutaneous immune system [242-244]. Very recently, an individual's skin microbiota community structure was linked to the ability to clear a bacterial pathogenic infection [283]. Furthermore, commensal microorganisms may produce numerous small molecules with a diverse range of targets that can modulate immune responses, for example, to compete with other bacteria [284].

Our observation that keratinocytes appear to exhibit different cytokine/AMP responses towards *S. aureus* or *F. magna* stimulation is intriguing. Given the known functions and activities of IL-8 (neutrophil chemoattractant), TNF- α (proinflammatory, induction of AMPs) and CCL20 (chemoattractant for various white blood cell types, antimicrobial activity), their strong induction by *S. aureus* (25 to 85-fold) contributes to control of infection. The observed rapid induction of AMPs by *F. magna*, however, suggests that this bacterium may be an important signalling factor to the keratinocytes when the skin barrier is breached. It should be noted that there is not a well-developed stratum corneum in a submerged culture, the bacteria are in direct contact with the keratinocytes. Although *F. magna* is not known as a major taxon of the commensal skin microbiota, we found it to represent on average 8% of all the mapped OTU's of normal lower leg skin, which is not insignificant. Complete or partial absence of *F. magna* could cause an impaired or delayed danger signalling to the keratinocytes in *FLG*^{-/-} or *FLG*^{+/-} individuals. This could, speculatively, be a mechanism that favours *S. aureus* colonization or infection, but clearly requires further investigation.

In conclusion, we have uncovered new and potentially important biological aspects of a very common genetic polymorphism that is associated with a major inflammatory disease. The notion that FLG deficiency or haploinsufficiency may cause skin barrier dysfunction, microbiome alterations and attenuated danger signalling generates new testable hypotheses regarding the disease mechanisms of AD.

Methods

Study approval

All volunteers in this study were selected according to the inclusion/exclusion criteria as approved by a protocol from the *National Institutes of Health* (NIH) Human Microbiome Project. The exact inclusion/exclusion criteria and study procedures that we presented to the volunteers can be found in [Supplementary Methods](#). In advance, medical ethical committee (Commissie Mensgebonden Onderzoek Arnhem-Nijmegen) approval and individual written informed consent were obtained. The study was conducted according to the Declaration of Helsinki principles.

Study participants and skin microbiome samples

Prior to microbiome sample collection, IV patients and healthy controls were all genotyped for *FLG* mutations and checked for FLG protein expression in skin biopsies. Samples were collected from the lower leg and obtained by swabbing 4 cm² skin area using Sterile Catch-All™ Sample Collection Swabs (Epicentre Biotechnologies) soaked

in sterile SCF-1 solution (50 mM Tris buffer [pH8], 1 mM EDTA, and 0.5% Tween-20) as previously described [3]. As negative controls, we took two mock swabs, which were only exposed to ambient air. DNA was extracted from the swabs by using the Mobio Ultraclean Microbial DNA Isolation Kit (Mobio laboratories) with modifications as described previously [3]. DNA samples were stored at -20°C until further processing.

Mutation analysis and immunohistochemistry

Genomic DNA was extracted from saliva samples using the Oragene kit (DNA Genotek Inc.) according to the manufacturer's protocol. *FLG* mutation analysis (for R501X and 2282del4) was performed as described previously [285]. Skin biopsies (3 mm) were immediately fixed in a 10% formalin solution (Baker Mallinckrodt) for 4 hours and subsequently embedded in paraffin. 6 µm paraffin sections were stained with an antibody against flaggrin (1:200, NCL-flaggrin, Novocastra) using an indirect immunoperoxidase technique (Vectastain, Vector Laboratories).

PCR amplification and sample preparation

For amplification of the V3-V6 region of the 16S rRNA gene we used the universal forward and reverse primers [3]. Forward primer: 5'-*CCATCTCATCCCTGCGTGTCTCCGACTCAGNNNNNN***ACTCCTACGGGA GGCAGCAG**-3' (italicized sequence is 454 Life Sciences primer A, bold sequence is the broadly conserved bacterial primer 338F; *NNNNNN* designates the sample-specific 6-base barcode used to tag each PCR product), reverse primer: 5'-*CCTATCCCCTGTGTGCCTTGGCAGTCTCAGCRRACGAGCTGAC* **GAC**-3' (the italicized sequence is 454 Life Sciences primer B, and the bold sequence is the broadly conserved bacterial primer 1061R). We recently described that skin samples contain only small numbers of microorganisms. Therefore, we have introduced a pre-amplification step with the same primers as described above excluding the barcodes and flag sequences. We established that this additional PCR step did not affect the results compared to a single amplification step with bar-coded primers, in other words, we observed no distortion or skewing of the taxa distribution [3]. PCR amplification protocols of step 1 and step 2 were performed as described previously [3]. The purified PCR products were submitted for pyrosequencing of the V3-V4 region of the 16S rRNA gene on the 454 Life Sciences GS-FLX platform using Titanium sequencing chemistry at DNAsion, Charleroi, Belgium.

16S rRNA gene pyrosequencing data analysis

The pyrosequencing data were analysed with a workflow based on QIIME v1.2 [42] and as applied and described previously [3], using pipeline settings recommended in the QIIME v1.2 tutorial, with the following exceptions: (i) sequencing reads were filtered for chimeric sequences using UCHIME [286], (ii) OTU clustering was performed with settings as recommended in the QIIME newsletter of Dec. 17, 2010 (<http://qiime.wordpress.com/2010/12/17/new-default-parameters-for-uclust-otu-pickers/>), using an identity threshold of 97%. In short, (i) raw sequencing reads are quality filtered based on length and quality scores, (ii) reads are demultiplexed (that is, multiplexed reads are assigned to samples based on their nucleotide barcode), (iii) chimeric 16S rRNA

gene sequence reads are identified and removed, (iv) reads are clustered into OTUs, and representative OTU sequences are selected, (v) OTUs are assigned to taxonomy (Greengenes 16S rRNA database version 'October 6, 2010', as default in QIIME v1.2; with QIIME default RDP classifier minimum confidence score of 0.80), and finally, (vi) relative abundance per taxon per sample (or per sample group) is calculated based on (total) number of reads assigned to that taxon. Alpha diversity metrics were calculated as implemented in QIIME v1.2, by bootstrapping 800 reads per sample, and taking the average over four bootstrap trials. Hierarchical clustering of samples was performed using UPGMA with weighted UniFrac as a distance measure, as implemented in QIIME v1.2. The Ribosomal Database Project classifier version 2.0 was applied for taxonomic classification [287]. Visualization of differences in relative abundance of taxa between the study contrast (Fig. 3) was done in Cytoscape [136]. Taxa (i.e. nodes) were included in the visualization if they met the following criteria: (i) all samples together have at least 10 reads assigned to the taxon, and (ii) the sample groups have a fold-difference of at least 0.1 for the taxon, or the taxon has a child (that is, more specific taxonomic classification) meeting the first criterion. The significance of the difference in relative abundance of specific taxa between sample groups was calculated using the Mann-Whitney U test (see section *Statistics*). After pyrosequencing and applying our QC filtering pipeline, we ended up with a small number of 16S sequencing reads for the two air control samples. Based on further (in parallel to the skin samples) clustering and composition analysis, combined with the very low number of 16S sequencing reads retrieved, we concluded that the negative control samples are of expected and sufficient quality (not shown).

Biophysical measurements and instruments

In vivo measurements were performed on the outer lower leg. None of the samples regions had obvious eczematous signs. The published guidelines for TEWL, SC hydration and skin surface pH assessments were followed [288-290]. TEWL was measured with a condenser-chamber method (Aquaflux AF200, Biox Systems), SC hydration with a capacitance-based method (Corneometer CM825, Courage and Khazaka) and skin surface pH with a planar pH electrode and meter (InLab426, Mettler Toledo and Portamess 913 pH, Knick GmbH & Co. KG). Prior to measurements, volunteers acclimatized for at least 15 minutes in a temperature-controlled room ($22 \pm 1^\circ\text{C}$, relative humidity $55 \pm 5\%$) with the body site to be assessed uncovered. Instruments were equilibrated in the same room for at least 30 minutes. The average of three, five and two readings taken in close proximity was calculated for TEWL, SC hydration and skin surface pH, respectively. The tip of the pH electrode was dipped in distilled water prior application on the skin and readings were recorded after stabilizing. The same researcher performed all measurements of skin surface pH, while TEWL and SC hydration were measured by another researcher. Volunteers were not allowed to apply any cream, soap or shower gel on the lower leg skin from the day before the experiments; moreover, they were asked to avoid any contact with water on the test site from 3 hours before the measurements.

***In vitro* model for the human cutaneous microbial ecosystem**

We recently developed an *in vitro* system that mimics human SC for bacterial growth, in which human callus serves as substrate and nutrient source for bacteria [270]. Human callus from the heel of six healthy volunteers (*FLG*^{+/+}) and four *FLG* deficient IV patients (*FLG*^{-/-}) was collected using a callus rasp (Ped Eggtm). This callus was frozen in liquid nitrogen and subsequently grounded using a Micro Dismembrator U (B. Braun Biotech International). Phosphate buffered saline (PBS) was added to the callus powder to create a 2% suspension, which was sterilized by exposure to gamma radiation (16.2 kGray/63 hours). To prepare the model, one mL of sterile agar (2% in PBS) was added in wells of 24-well cell culture plates. On top of this agar, 100 μ L of sterile callus suspension (2% in PBS) was pipetted (Chapter 4: Fig. 1). The plate was allowed to dry for 24 hours at 37°C, and stored at 4°C until further use. Bacteria (*F. magna*) were incubated on the model at 32°C in anaerobic conditions for seven days. The entire model (agar + callus + bacteria) was lifted out of the 24-wells plate and transferred to a 50-mL tube containing 5 mL PBS. The tubes were vortexed at maximum speed for one minute to detach the bacteria from the model. The aqueous solution containing the bacteria (including callus particles, but without the agar) was transferred to a new tube. These samples were serially diluted in steps of 5. Ten μ L of each dilution was placed on sheep blood agar plates and incubated at 37°C in anaerobic conditions. Next days, colonies visible on the plate were counted for each dilution.

***Fingoldia magna* growth on the in vitro stratum corneum model**

Fingoldia magna type strain 2974 (DSM nr. 20470, ATCC nr. 15794) was inoculated on Columbia agar with 5% sheep blood (Dickinson and Company, BD) and grown for four days at 37°C in anaerobic conditions. One single colony of the plate was picked and cultured in Brain Heart Infusion medium (Mediaproducs BV) for four days at 37°C (anaerobic conditions). Hereafter, the bacteria were collected by centrifugation, washed two times with PBS and finally resuspended in PBS resulting in bacterial concentrations of 10⁵ CFU/mL. Portions of 20 μ L bacteria (2x10³ CFU) suspension were added to each well of the SC model (with callus derived from *FLG*^{+/+} and *FLG*^{-/-} individuals as described above). The bacteria on the model were incubated at 32°C in anaerobic conditions for seven days.

Determination of FLG degradation products in callus

The FLG degradation products in callus have been determined by a slightly adapted HPLC method reported previously [291]. 1 mg of callus sample was extracted with 0.5 mL of 25% (w=w) ammonia solution by vigorous shaking using IKA-Vibrax-VXR Model 2200 (IKA-works Inc.). Ammonia extracts were evaporated to dryness, resolved in 0.5 mL of water, and subsequently filtered through a 0.2 mm PVDF membrane filter (Grace Davison Discovery Science). Before HPLC analysis, the aliquots were diluted 1:1 with the mobile phase.

Microbiota derived functional prediction by PICRUST

For each of the sequencing samples, bacterial 16S profiles of the skin microbiota were obtained as described above. Based on these profiles, we predicted the presence of

KEGG Orthologs (K numbers) and subsequent functional and metabolic pathways for IV patients (*FLG*^{-/-} and *FLG*^{+/-}) and healthy controls (*FLG*^{+/+}) using PICRUSt (version 1.0.0) [64]. The Phylogenetic Investigation of Communities by Reconstruction of Unobserved States (PICRUSt) method [64] allow for computationally predicting microbiota (i.e. single microbial entities or, as presented in this manuscript, complete communities) their function potential, based on interring 16S marker-gene sequencing-derived information. PICRUSt requires closed-reference OTU picking in QIIME [42], which was done to the Greengenes [37] reference collection (version 13.5, May 2013). For this step, the same sequencing reads were used as those provided for the QIIME workflow, as described above. Thereafter, for analysis in PICRUSt, the total skin makeup of microbiota for each FLG genotypic group (as derived from QIIME, after OTU clustering and identification) was considered; in other words, no OTUs were sub-selected or excluded from analysis. The default (and author advised) workflow of PICRUSt was followed: (i) OTUs were normalized by dividing each OTU by the known or predicted 16S copy number abundance of its Greengenes reference, (ii) a (metagenome) functional predictions table was generated by multiplying each normalized OTU abundance by each predicted functional trait abundance, which was input for our further downstream analysis (for any additional detail, or for more information about the PICRUSt workflow we refer to the PICRUSt website: <http://picrust.github.io>). The PICRUSt downstream analysis was focused on candidate genes only (KEGG Orthologs; K numbers), by applying in-house Perl scripts that filtered-out the required KEGG Orthologs. The candidate genes were selected from the Hut pathway based on a review [273], and are listed in [Supplementary Table S4](#). For additional calculation of relative abundances of pathways and KEGG Orthologs, and for downstream statistics, in-house Python scripts were used (see section *Statistics*). Based on the KEGG pathway (<http://www.genome.jp/kegg/>) for 'histidine metabolism' (ko00340), a map was constructed in the Cytoscape network interaction visualization program [136] with use of the KEGGScape app/plugin [292] in order to visualize the differential (relative) abundance and metabolic context of candidate *Hut* genes between the study contrast.

Keratinocyte cultures and bacterial stimulations

Primary human keratinocytes obtained from abdominal plastic skin surgery were isolated and expanded according to the Rheinwald-Green protocol [293] and stored in liquid nitrogen. Primary human keratinocytes were cultured in keratinocyte growth medium (KGM bullet kit, Lonza) without antibiotics as described earlier [294]. *Staphylococcus aureus* (ATCC 29213) was obtained from the department of Medical Microbiology of the Radboudumc. *F. magna* (DSM 20470) was purchased from DSMZ, Braunschweig, Germany. Bacteria were inoculated on Columbia agar with 5% sheep blood (Dickinson and Company (BD), Sparks, MD) overnight (*S. aureus*) or for four days (*F. magna*) at 37°C. One single colony of each plate was picked and cultured in Brain Heart Infusion medium (Mediaproducs BV, Groningen, The Netherlands). *F. magna* was cultured under strict anaerobic conditions. *S. aureus* and *F. magna* were collected by centrifugation, washed two times with PBS and finally resuspended in PBS resulting in bacterial concentrations of 10⁷ CFU/mL. A portion of the bacterial suspensions was exposed to 85°C for 25 minutes to heat-kill the bacteria. To determine the amount of *S. aureus* and *F. magna* that was brought on the keratinocyte cultures, bacterial

suspensions were serially diluted in steps of 5. Ten μL of each dilution was placed on sheep blood agar plates and incubated overnight (*S. aureus*) or for four days (*F. magna*) at 37°C in aerobic or anaerobic conditions, respectively. Visible colonies on the plate were counted for each dilution. The number of CFU was calculated: counted CFU \times dilution factor. Confluent submerged keratinocyte cultures (plastic surgery controls, $n = 10$) were inoculated with 3.5×10^5 viable and heat-killed bacteria and incubated for 10 hours. Cells were washed after 6 hours and incubated for another 4 hours in fresh medium before the keratinocytes were collected ($t = 10$ hours). Afterwards, the bacteria were washed away and the keratinocytes were harvested and processed for RNA isolation and qPCR analysis.

RNA isolation and qPCR analysis

RNA isolation, cDNA synthesis and qPCR analysis was performed as described earlier [295]. All primers were designed and used as described previously [296]. Target gene expression was normalized to the expression of the house keeping gene human acidic ribosomal phosphoprotein Po (*RPLPo*). The $\Delta\Delta\text{Ct}$ method was used to calculate relative mRNA expression levels [297] (see [Supplementary Table S5](#) for primer sequences).

Accession numbers

The raw, and unprocessed 16S sequencing reads data is publicly available for download at the European Nucleotide Archive (ENA) database (<http://www.ebi.ac.uk/ena>) under study accession number PRJEB11661 (or secondary accession number ERP013063) [298]. The sequencing data is available in FASTQ-format, including corresponding metadata for each sample.

Statistics

For the microbiota data in this manuscript, statistical significance between contrasts with regard to taxonomy abundances and thereof derived KEGG Orthologs abundances was tested by the non-parametric Mann-Whitney U, corrected with Bonferroni for multiple testing; unless stated otherwise. Statistical tests were performed by custom, in-house R-scripts (version 3.1.2; <https://www.r-project.org/>) or as implemented in SciPy (<https://www.scipy.org/>), downstream of QIIME and PICRUST analyses as described above. Multivariate Redundancy Analysis (RDA) was done using Canoco 5.04 [228] using default settings of the analysis type "Constrained". Relative abundance values for either taxa were used as response data and the sample FLG status as explanatory variable. RDA calculates p -values by permutating the sample FLG status. For all other experiments, statistical significance was tested by a non-parametric Mann-Whitney U test or repeated measures ANOVA with Bonferroni's post-hoc test.

Supplementary Methods

([online](#)) Supplementary Experimental Procedures of online repository text at :
<https://www.sciencedirect.com/science/article/pii/S0091674916311174/>

Supplementary Figures

Figure S1. Biophysical measurements on the lower leg skin. (WORD)

([online](#)) Figure E2 of online repository text at :
<https://www.sciencedirect.com/science/article/pii/S0091674916311174/>

Figure S2. Microbial diversity of samples between different FLG genotypes. (WORD)

([online](#)) Figure E3 of online repository text at :
<https://www.sciencedirect.com/science/article/pii/S0091674916311174/>

Figure S3. Alpha diversity metrics in FLG genotypes. (WORD)

([online](#)) Figure E4 of online repository text at :
<https://www.sciencedirect.com/science/article/pii/S0091674916311174/>

Supplementary Tables

Table S1. Cohort of IV patients and healthy controls. (WORD)

(online) Tables E1 of online repository text at :

<https://www.sciencedirect.com/science/article/pii/S0091674916311174/>

Table S2. Microbiota analysis of the lower leg. (WORD)

(online) Tables E2 of online repository text at :

<https://www.sciencedirect.com/science/article/pii/S0091674916311174/>

Table S3. Read and OTU counts. (WORD)

(online) Tables E3 of online repository text at :

<https://www.sciencedirect.com/science/article/pii/S0091674916311174/>

Table S4. Selection of Hut pathway genes. (WORD)

(online) Tables E4 of online repository text at :

<https://www.sciencedirect.com/science/article/pii/S0091674916311174/>

Table S5. qPCR primers. (WORD)

(online) Table E6 of online repository text at :

<https://www.sciencedirect.com/science/article/pii/S0091674916311174/>

Table S6. (EXCEL)

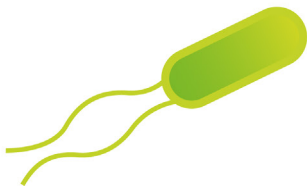
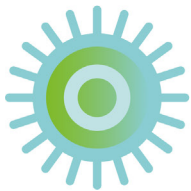
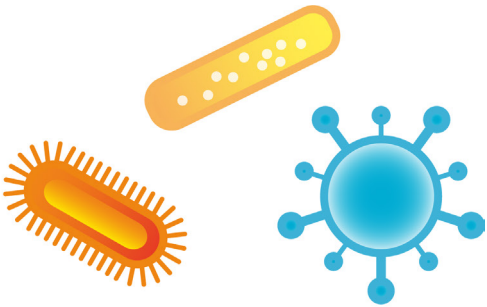
(online) Excel file E1 of online repository at :

<http://www.sciencedirect.com/science/article/pii/S0091674916311174/>

Table S7. (EXCEL)

(online) Excel file E2 of online repository at :

<http://www.sciencedirect.com/science/article/pii/S0091674916311174/>



PART III

The Gut Microbiome

OPEN ACCESS

as published in *Microbiome*, 2017, Jun 23;5(1):63

<https://doi.org/10.1186/s40168-017-0278-2>

¹ Experimental Rheumatology, Radboud Institute for Molecular Life Sciences (RIMLS), Radboud University Medical Center (Radboudumc), Nijmegen, The Netherlands.

² Centre for Molecular and Biomolecular Informatics, RIMLS, Radboudumc, Nijmegen, The Netherlands.

³ NIZO, Ede, The Netherlands.

⁴ Danone Nutricia Research, Utrecht, The Netherlands.

⁵ Laboratory of Microbiology, Wageningen University, The Netherlands.

⁶ Division of Rheumatology, Department of Medicine, New York University School of Medicine, NY, USA.

⁷ Division of Pharmacology, Utrecht Institute for Pharmaceutical Sciences, Utrecht University, NL.

These authors contributed equally to this work.

CHAPTER 6

ABERRANT INTESTINAL MICROBIOTA DUE TO IL-1 RECEPTOR ANTAGONIST DEFICIENCY PROMOTES IL-17- AND TLR4-DEPENDENT ARTHRITIS

Rebecca Rogier^{1 #}

Thomas H.A. Ederveen^{1,2 #}

Jos Boekhorst^{2,3}

Harm Wopereis^{4,5}

Jose U. Scher⁶

Julia Manasson⁶

Sanne Frambach¹

Jan Knol^{4,5}

Johan Garssen^{4,7}

Peter M. van der Kraan¹

Marije I. Koenders¹

Wim B. van den Berg¹

Sacha A.F.T van Hijum^{2,3}

Shahla Abdollahi-Roodsaz^{1,6}

ABSTRACT

BACKGROUND. Perturbation of commensal intestinal microbiota has been associated with several autoimmune diseases. Mice deficient in interleukin-1 receptor antagonist (*IL1rn*^{-/-} mice) spontaneously develop autoimmune arthritis and are susceptible to other autoimmune diseases such as psoriasis, diabetes and encephalomyelitis; however, the mechanisms of increased susceptibility to these autoimmune phenotypes are poorly understood. We investigated the role of interleukin-1 receptor antagonist (IL-1Ra) in regulation of commensal intestinal microbiota, and assessed the involvement of microbiota subsets and innate and adaptive mucosal immune responses that underlie the development of spontaneous arthritis in *IL1rn*^{-/-} mice.

RESULTS. Using high-throughput 16S rRNA gene sequencing, we show that IL-1Ra critically maintains the diversity and regulates the composition of intestinal microbiota in mice. IL-1Ra deficiency reduced the intestinal microbial diversity and richness, and caused specific taxonomic alterations characterized by overrepresented *Helicobacter* and underrepresented *Ruminococcus* and *Prevotella*. Notably, the aberrant intestinal microbiota in *IL1rn*^{-/-} mice specifically potentiated IL-17 production by intestinal lamina propria (LP) lymphocytes and skewed the LP T cell balance in favor of T helper 17 (Th17) cells, an effect transferable to WT mice by fecal microbiota. Importantly, LP Th17 cell expansion and the development of spontaneous autoimmune arthritis in *IL1rn*^{-/-} mice were attenuated under germ-free condition. Selective antibiotic treatment revealed that tobramycin-induced alterations of commensal intestinal microbiota, *i.e.*, reduced *Helicobacter*, *Flexispira*, *Clostridium* and *Dehalobacterium*, suppressed arthritis in *IL1rn*^{-/-} mice. The arthritis phenotype in *IL1rn*^{-/-} mice was previously shown to depend on Toll-like receptor 4 (TLR4). Using ablation of both IL-1Ra and TLR4, we here show that the aberrations in the *IL1rn*^{-/-} microbiota are partly TLR4-dependent. We further identify a role for TLR4 activation in the intestinal lamina propria production of IL-17 and cytokines involved in Th17 differentiation preceding the onset of arthritis.

CONCLUSIONS. These findings identify a critical role for IL1Ra in maintaining the natural diversity and composition of intestinal microbiota, and suggest a role for TLR4 in mucosal Th17 cell induction associated with the development of autoimmune disease in mice.

Background

Interleukin-1 (IL-1) plays a central role in inflammation and immunity [299]. Activation of IL-1 receptor is physiologically controlled by its structural homologue and natural inhibitor, the IL-1 receptor antagonist (IL-1Ra), encoded by the *IL1rn* gene [300]. *IL1rn* knockout (*IL1rn*^{-/-}) mice are susceptible to a variety of autoimmune diseases including arthritis, psoriasis, diabetes and encephalomyelitis [301-305]. This indicates a critical role for IL-1Ra in protection against autoimmunity; however, the mechanisms are poorly understood.

We questioned the role of IL-1Ra in regulation of the intestinal microbiota and the involvement of mucosal immune response as an underlying mechanism for the spontaneous autoimmune arthritis in *IL1rn*^{-/-} mice, which is dependent on T cells and IL-17 [302, 306]. Several studies have associated commensal microbiota with autoimmune disease in mouse models of rheumatoid arthritis (RA), diabetes and multiple sclerosis [307-311]. Importantly, the diversity and the composition of commensal intestinal microbiota are altered in patients with psoriatic and RA compared with healthy individuals [312-316]. One of the most prominent effects of microbiota is to define the balance between the pro-inflammatory CD4⁺ T helper 1 (Th1) and Th17 cells and protective regulatory T (Treg) cells, both at mucosal surfaces and systemically [317-319]. In this context, specific subsets of intestinal microbiota, such as the vancomycin-sensitive segmented filamentous bacteria (SFB), robustly induce differentiation of Th17 cells in small intestine LP (SI-LP) [80, 320]. Th17 cells are considered to play a pathogenic role in a subset of patients with RA by producing proinflammatory mediators, such as IL-17, and inducing osteoclastogenesis [321-325]. Interestingly, SFB colonization has been shown to exacerbate arthritis in K/BxN mice, an autoimmune model of arthritis arising from T cell auto-reactivity to the glycolytic enzyme glucose-6-phosphate isomerase [311, 326]. However, given that SFB were not found in human adults [18, 327], it is important to investigate the involvement of other indigenous microbiota in arthritis.

We previously described that arthritis in *IL1rn*^{-/-} mice is diminished under germ-free (GF) condition [310]. We also showed that *IL1rn*^{-/-} arthritis is dependent on the activation of Toll-like receptor 4 (TLR4), which affected systemic Th17 cell differentiation [310]. Here, we characterized the intestinal microbiota present in autoimmune-prone *IL1rn*^{-/-} mice to clarify the nature of the microbiota that trigger arthritis and the underlying mucosal immune pathways. We also examined the role of TLR4 in the intestinal mucosal immune responses associated with arthritis. Using high-throughput 16S rRNA gene sequencing of fecal microbiota, we demonstrate a critical role for IL-1Ra in maintaining the natural diversity and composition of commensal intestinal microbiota. We show that the aberrant *IL1rn*^{-/-} microbiota increases intestinal Th17 cell differentiation, a phenotype that is transferable to wild-type (WT) mice by the microbiota. We also provide evidence that tobramycin-sensitive indigenous commensal intestinal bacteria contribute to arthritis in *IL1rn*^{-/-} mice and identify a significant role for TLR4 in mucosal induction of IL-1 β and IL-17 prior to the onset of arthritis.

Results

IL-1Ra maintains the biodiversity and richness of commensal intestinal microbiota.

To identify intestinal microbiota associated with arthritis, we sequenced fecal bacterial 16S rRNA genes of *IL1rn*^{-/-} and age- and gender-matched WT control mice. Fecal microbiota were analyzed as an unselected representation of the overall microbial communities in the intestines. Considering differential roles of TLR2 and TLR4 in *IL1rn*^{-/-} arthritis [310], we sequenced samples of *IL1rn*^{-/-} *Tlr2*^{-/-} and *IL1rn*^{-/-} *Tlr4*^{-/-} mice in parallel. The average sequencing depth and total numbers of reads and operational taxonomic units (OTU) per experimental group as well as the hierarchical weighed UniFrac cluster analysis at the genus level are shown in [Supplementary Table S1](#) and [Supplementary Fig. S1](#).

Principal coordinates analysis (PCoA) based on an unweighted UniFrac analysis of intestinal microbiota showed that *IL1rn*^{-/-} microbiota is profoundly different from the WT microbiota ([Fig. 1A](#)). WT and *IL1rn*^{-/-} mice formed clear, separate clusters regardless of the cage or litter of origin ([Fig. 1A](#)). Strikingly, microbial composition of *IL1rn*^{-/-} and *IL1rn*^{-/-} *Tlr2*^{-/-} mice were indistinguishable, while *IL1rn*^{-/-} *Tlr4*^{-/-} mice formed another distinct cluster ([Fig. 1A](#)). To assess the effects of familial transmission and lineage origin versus the effect of the genotype (WT or *IL1rn*^{-/-}), we compared the UniFrac distances within a litter with UniFrac distances across litters of the same genotype as well as the opposite genotype, similar to the study by Ubeda *et al.* This analysis showed that the effect of the lineage origin and litter was limited in our experimental setting, because, as long as the genotype remained the same, the UniFrac distances across different litters were very similar to the UniFrac distances within the litters ([Supplementary Fig. S2A](#)). This was true for both WT and *IL1rn*^{-/-} groups. Importantly, the UniFrac distances were significantly higher when mice from different genotypes were compared, indicating a higher level of dissimilarity ([Supplementary Fig. S2B](#)). Therefore, the effect of the genotype (*IL1rn*-deficiency) on the overall microbiota composition was significantly higher than any litter and cage effect. In addition, *IL1rn*^{-/-} and *IL1rn*^{-/-} *Tlr2*^{-/-} mice showed significantly reduced number of OTUs and loss of microbial diversity based on the Shannon index, the rarefaction curves of phylogenetic distance (PD) whole tree, and the diversity index bootstrapped for the number of retrieved sequences ([Fig. 1B-E](#)). IL-1Ra deficiency also resulted in loss of species richness estimated by Chao index ([Fig. 1F](#)). These effects were fully or partially restored in *IL1rn*^{-/-} *Tlr4*^{-/-} mice ([Fig. 1B-F](#)). Altogether, these data strongly suggest that IL-1Ra plays a critical role in maintaining the intestinal microbial diversity, and that the loss of diversity in *IL1rn*^{-/-} mice partially depends on TLR4.

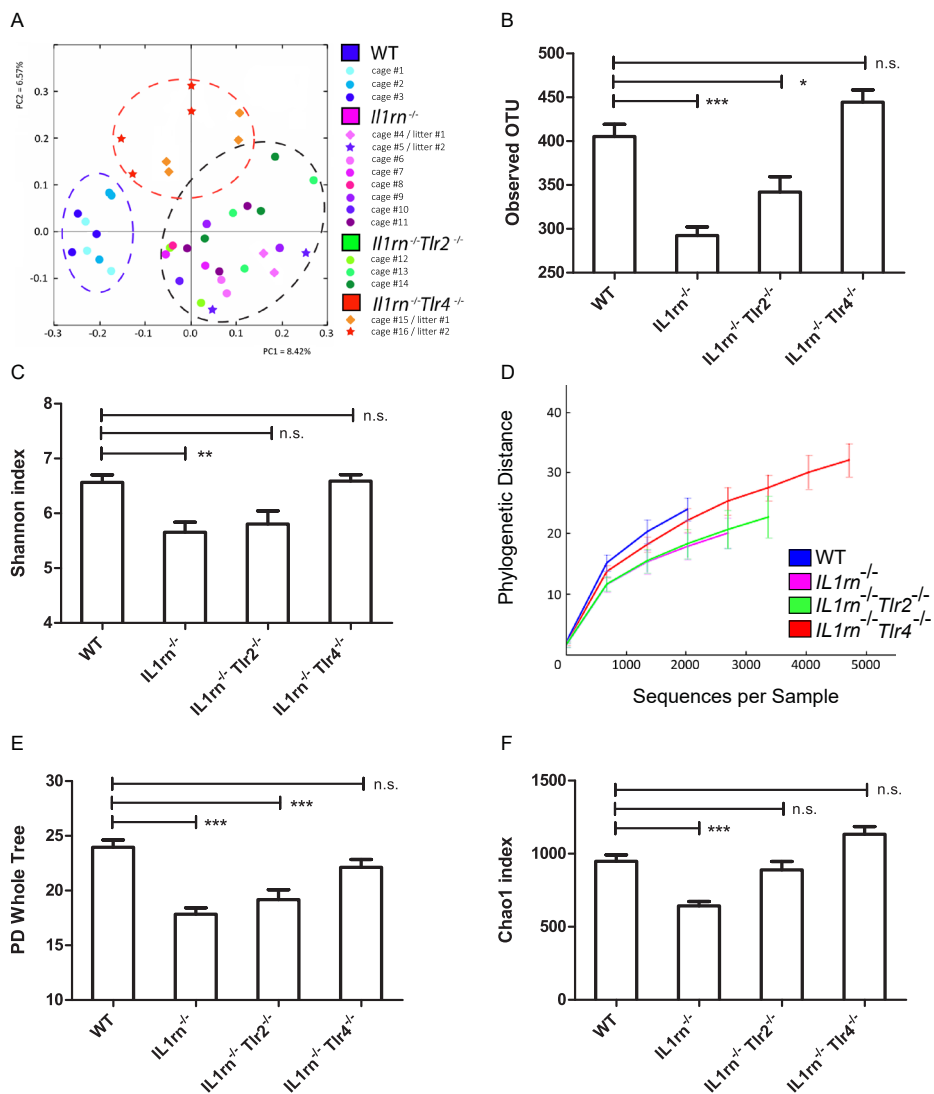


Figure 1. IL-1Ra deficiency skews intestinal microbial composition and reduces its diversity in a TLR4-dependent manner.

(A) Principal coordinates analysis (PCoA) based on an unweighted UniFrac analysis of the intestinal microbial composition where samples of mice from different cages and litters are highlighted with different colors. The position and distance of data points indicates the degree of similarity in terms of presence of bacterial taxonomies. (B) Number of observed operational taxonomic units (OTUs), (C) Shannon index of microbial diversity, (D) alpha diversity rarefaction curves of phylogenetic distance (PD) whole tree, (E) PD whole tree, bootstrapped for 2000 reads per sample, averaged of 4 trials, (F) and Chao index are shown. Data (mean + SEM) represent 16S rRNA gene 454-pyrosequencing analysis of intestinal microbiota of WT ($n = 9$), *IL1rn*^{-/-} ($n = 15$), *IL1rn*^{-/-}*Tlr2*^{-/-} ($n = 8$) and *IL1rn*^{-/-}*Tlr4*^{-/-} ($n = 8$) mice. n.s. = not significant, * $p < 0.05$ and *** $p < 0.001$, by Mann-Whitney U test. See also [Supplementary Fig. 1](#) and [Supplementary Table 1](#).

Specific taxonomic alterations characterize the dysregulated microbiota of autoimmune-prone $IL1rn^{-/-}$ mice.

The phylogenetic tree in [Figure 2](#) summarizes the observed alterations in relative abundances of microbial taxa. Compared with WT microbiota, we found a highly significant overrepresentation of the genus *Helicobacter* ($p = 0.004$, Bonferroni corrected), and a significant underrepresentation of the genus *Prevotella* ($p = 0.008$, Bonferroni corrected) ([Fig. 2 and Supplementary Table S2](#)). In addition, $IL1rn^{-/-}$ intestinal microbial composition was characterized by expansion of *Butyrivimonas*, *Rikenella* and *Streptococcus* by 10, 3.7 and 2.4 folds ($p = 0.0048$, $p = 0.0022$ and $p = 0.0032$, respectively, Bonferroni uncorrected), along with a decrease in *Parasutterella*, *Xylanibacter*, *Ruminococcus* and *Barnesiella* by 10, 6.9, 2.7 and 1.4 folds ($p = 0.040$, $p = 0.0004$, $p = 0.0099$ and $p = 0.0005$, respectively, Bonferroni uncorrected), respectively ([Fig. 2 and Supplementary Table S2](#)). Notably, we were unable to identify any OTUs in our dataset that could be classified as SFB (family *Clostridiaceae*, genus *Candidatus arthromitus*). Moreover, none of the 27 present OTUs assigned to the family *Clostridiaceae* aligned with the known SFB 16S gene sequences in The Ribosomal Database Project [18]. However, SFB were detectable by qPCR in fecal samples of all WT mice and most of the $IL1rn^{-/-}$ mice ([Supplementary Table S3](#)). Although WT mice tended to have slightly more SFB, the level of SFB colonization was not significantly different between the groups ([Supplementary Table S3](#)).

Altogether, these data suggest that multiple yet specific microbial taxa are regulated by the physiologic expression of IL-1Ra. Therefore, a complex set of aberrant microbiota may affect the (mucosal) immune response and contribute to the autoimmune disease in $IL1rn^{-/-}$ mice.

$IL1rn^{-/-}$ intestinal microbiota potentiate IL-17 production by intestinal lamina propria lymphocytes.

To assess the effect of IL-1Ra deficiency on the mucosal T cell response, we cultured enzymatically isolated lamina propria lymphocytes (LPL) *ex vivo* in the presence of PMA and ionomycin. The production of the Th1 signature cytokine IFN γ was low and not altered by the IL-1Ra deficiency ([Fig. 3A](#), gating strategy shown in [Supplementary Fig. S3](#)); however, we observed a marked increase in the production of IL-17 by $IL1rn^{-/-}$ LPLs compared with WT LPLs ([Fig. 3B](#)). Flow cytometry analysis of lamina propria cells of WT and $IL1rn^{-/-}$ mice verified a significant, clear increase of IL-17-producing TCR β^+ CD4 $^+$ cells in $IL1rn^{-/-}$ mice, while TCR β^- cells in LP produced similar amounts of IL-17 in WT and $IL1rn^{-/-}$ mice ([Supplementary Fig. S4](#)). This suggests that Th17 cells, not $\gamma\delta$ T cells, are the source of increased IL-17 production in LP of $IL1rn^{-/-}$ mice. Production of IL-4, IL-6 and TNF α was not affected ([Fig. 3C](#) and data not shown). Interestingly, the production of IL-17 but not IFN γ by lymphocytes in joint-draining lymph nodes (dLN) was significantly increased in $IL1rn^{-/-}$ mice compared with WT mice ([Fig. 3D and E](#)). This was paralleled by a concomitant decrease in the production of the Th2-related cytokine IL-4 in $IL1rn^{-/-}$ mice ([Fig. 3F](#)).

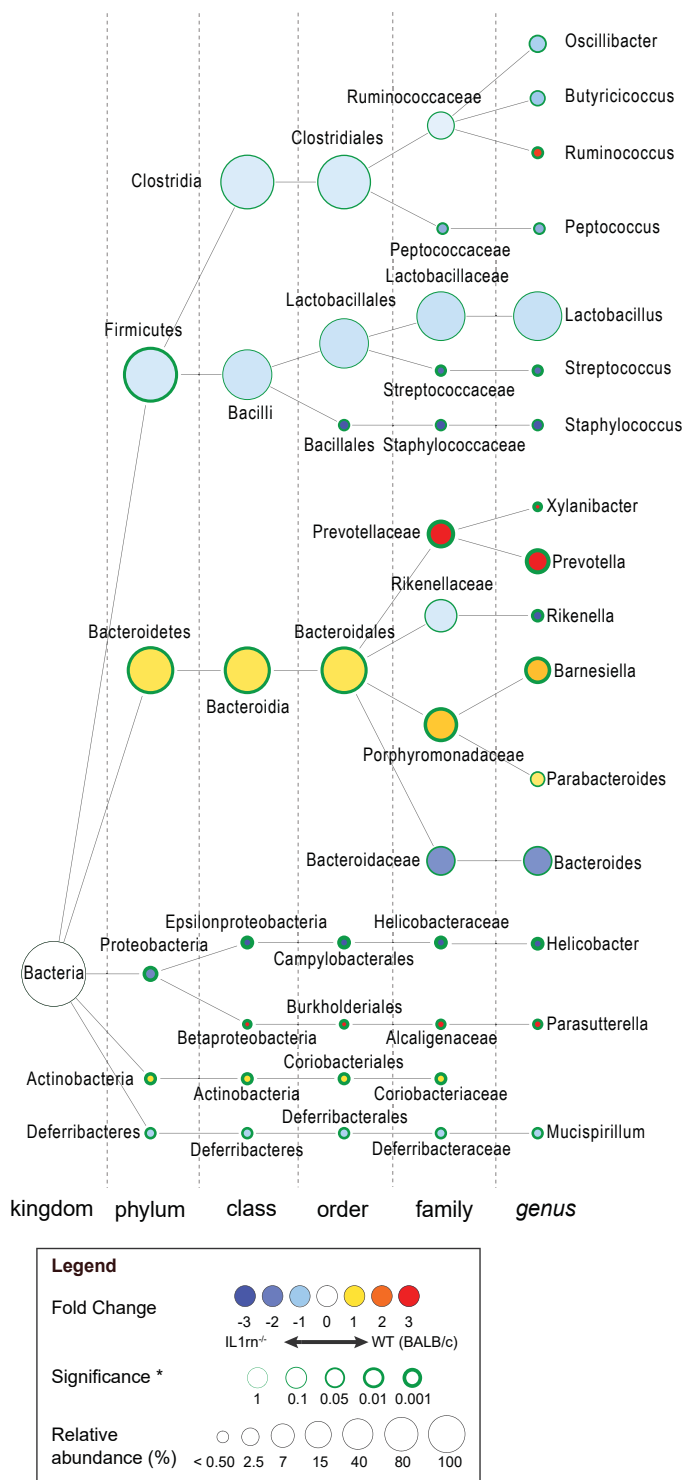


Figure 2. (previous page) IL-1 receptor antagonist controls relative abundance of specific intestinal microbial taxa.

Phylogenetic tree created by Cytoscape software showing specific changes in intestinal microbial community at different taxonomic levels induced by IL-1Ra deficiency. Nodes represent taxa, and the size of each node represents its relative abundance. The color red indicates a decrease and blue represents an increase of relative abundance in *IL1rn*^{-/-} compared with WT mice. The thickness of the green border indicates the degree of the statistical significance by Mann-Whitney U test. See also [Supplementary Table 2](#).

To identify a potential causative relationship between the aberrant microbiota and enhanced mucosal IL-17 production, we transferred *IL1rn*^{-/-} microbiota to WT mice by oral gavage followed by immediate co-housing of the two mouse strains for up to 6 weeks. Transfer of *IL1rn*^{-/-} microbiota clearly potentiated IL-17 production by SI-LP T cells in WT recipients as early as 10 days post fecal transfer and co-housing, without affecting IFN γ and IL-4 ([Fig. 3G-I](#)). This indicated that *IL1rn*^{-/-} intestinal microbiota causes a shift in the LP T cell balance in favor of Th17 cells. However, this was not sufficient for the development of arthritis in WT animals during the 6 weeks follow-up period. This suggests that additional (genetic) susceptibility of the host, as in *IL1rn*^{-/-} mice, is required for the development of arthritis. Furthermore, co-housing with WT mice did not affect the development of arthritis in *IL1rn*^{-/-} mice (not shown).

Potentiated Th17 response and spontaneous arthritis in *IL1rn*^{-/-} mice highly depend on the presence of commensal microbiota.

To determine whether the increase in intestinal Th17 cells and spontaneous arthritis in *IL1rn*^{-/-} mice depends on commensal microbiota, we established GF *IL1rn*^{-/-} mice. Flow cytometry analysis of LPLs showed that germ-free condition had no significant effect on the percentage of Th1 cells while reducing the numbers of Th1 cells in SI-LP ([Fig. 4A and B](#)). In contrast, both the percentage and the number of SI-LP Th17 cells were substantially reduced in GF compared with conventional (CV) mice ([Fig. 4C and D](#)). This strongly suggests that the skewed intestinal T cell balance in *IL1rn*^{-/-} mice is largely microbiota-dependent.

In agreement with IL-17-dependence of *IL1rn*^{-/-} arthritis [302, 306] and in line with our previous observations [310], GF *IL1rn*^{-/-} mice showed a clear sustained protection from arthritis with on average three weeks delay in disease onset ([Fig. 4E](#)). In addition, transfer of conventional *IL1rn*^{-/-} microbiota to GF *IL1rn*^{-/-} mice re-induced arthritis and resulted in a severe disease comparable to that in conventional *IL1rn*^{-/-} mice ([Supplementary Fig. S5](#)). Therefore, *IL1rn*^{-/-} commensal microbiota, although not sufficient to induce arthritis in a WT host, are critical for the full development of arthritis in *IL1rn*^{-/-} mice. Consistently, we also observed a robust reduction of IL-17, but not IFN γ , production in spleen and most notably in joint-adjacent lymph nodes of GF mice ([Fig 4F and G](#)). These effects were accompanied by a significant reciprocal increase in Th2-related cytokines IL-4 and IL-10 as well as IL-2 in spleens of GF mice ([Fig. 4H and Supplementary Fig. S6](#)). These data support modulation of extra-intestinal immune response by intestinal microbiota during arthritis.

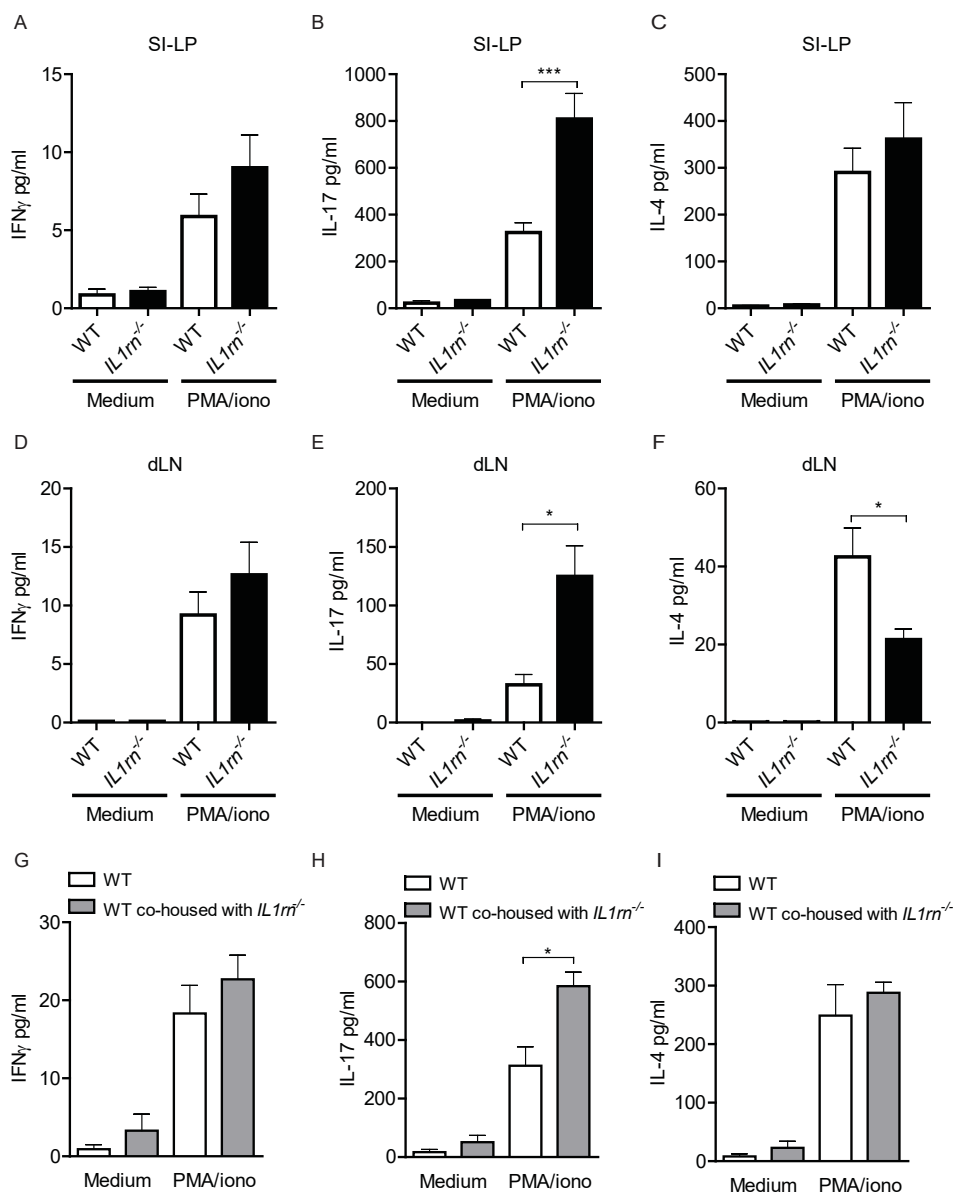


Figure 3. *IL1rn*^{-/-} intestinal microbiota potentiate IL-17 production in intestinal lamina propria and joint-draining lymph nodes.

(A-I) Production of prototypic Th1, Th17 and Th2 cell cytokines (IFN γ , IL-17A and IL-4, respectively) by SI-LP (A-C and G-I) and draining lymph node (dLN) lymphocytes (D-F). Cells were isolated from WT and *IL1rn*^{-/-} mice (A-F), or WT mice transplanted with *IL1rn*^{-/-} feces and co-housed with *IL1rn*^{-/-} mice for 10 days (G-I). Cells were stimulated *ex vivo* with PMA and ionomycin in duplicates for 5 hours, and cytokines were measured by Luminex assay. Data represent mean + SEM of a representative experiment with $n = 5$ (A-C) and $n = 3$ (D-I) mice per group, each stimulated in duplicate. n.s. = not significant, * $p \leq 0.05$ and *** $p \leq 0.001$, by Mann-Whitney U test. See also [Supplementary Fig. 2](#).

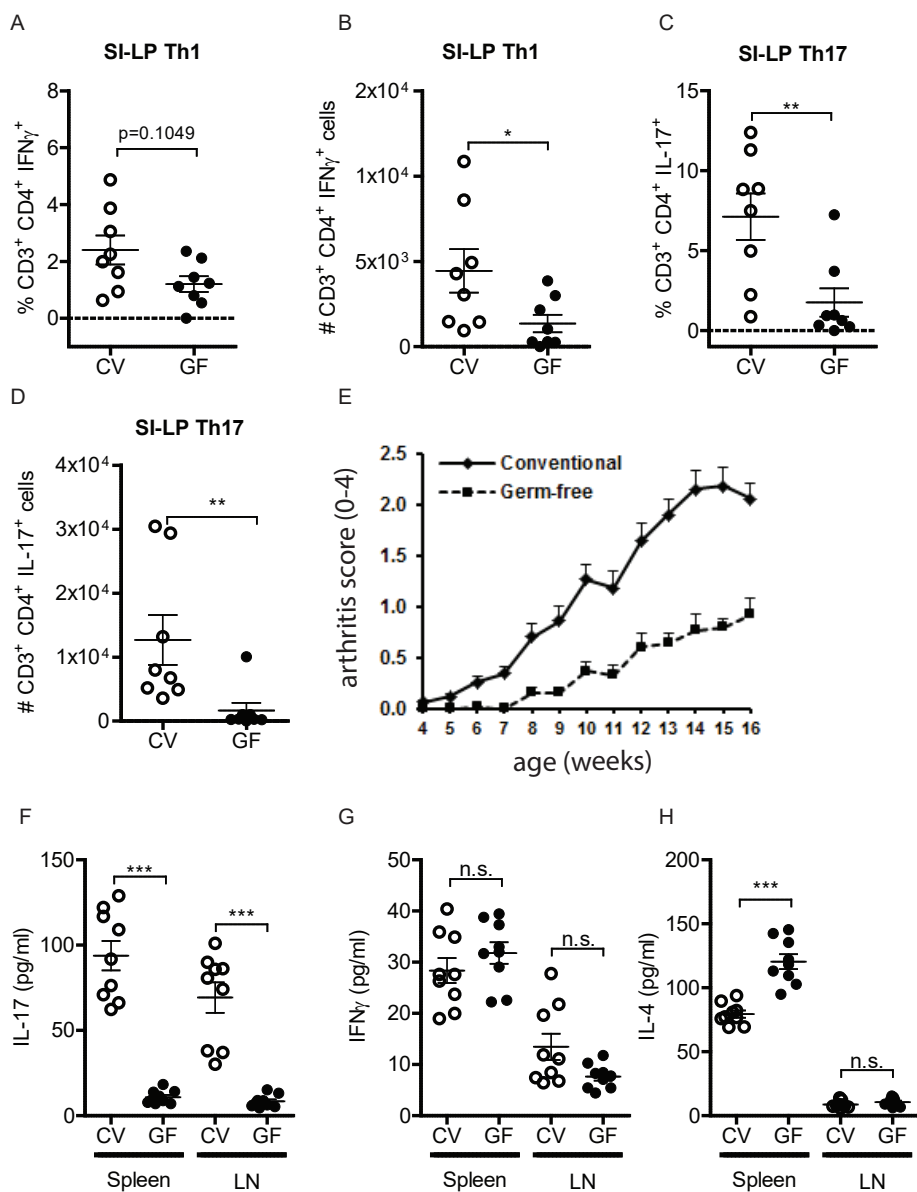


Figure 4. Commensal microbiota drive potentiated Th17 response and spontaneous arthritis in *IL1rn*^{-/-} mice.

(A-D) Frequency and numbers of IFN γ -producing (A,B) and IL-17-producing (C,D) CD3⁺CD4⁺ SI-LP cells. Data are pooled from three independent experiments. (E) Arthritis severity scores of conventional (CV, $n = 12$) and germ-free (GF, $n = 11$) *IL1rn*^{-/-} mice of a representative experiment. Scale 0-2 for each hind paw. Mean + SEM is shown. (F-H) Production of IFN γ , IL-17 and IL-4 upon *ex vivo* stimulation of spleen and lymph node cells from CV and GF mice with PMA and ionomycin for 6 hours, as measured by Luminex assay. $n = 3$ mice per group of each stimulated in triplicate. *n.s.* = not significant, $**p \leq 0.01$ and $***p \leq 0.001$, by Mann-Whitney U test. See [Supplementary Fig. 3](#) for gating strategy.

Tobramycin-induced alteration of intestinal microbiota suppresses arthritis in $IL1rn^{-/-}$ mice.

The lack of microbiota in GF mice is not limited to the intestines. To determine whether intestinal microbiota serve as a relevant trigger for arthritis, we first depleted intestinal microbiota in conventionally-housed mice using a cocktail of metronidazole, neomycin and ampicillin. Treatment of 5-week-old $IL1rn^{-/-}$ mice for only 1 week suppressed arthritis over a sustained period, i.e. 6 weeks after ceasing antibiotics (Fig. 5A). This indicated that abrogation of arthritis in GF mice (Fig. 4C) is not due to an immature immune system and, more importantly, can be reproduced by the sole eradication of intestinal microbiota. Interestingly, colonization of the antibiotic-treated mice with SFB as model organisms inducing SI-LP Th17 cells was sufficient to fully restore arthritis (Fig. 5A).

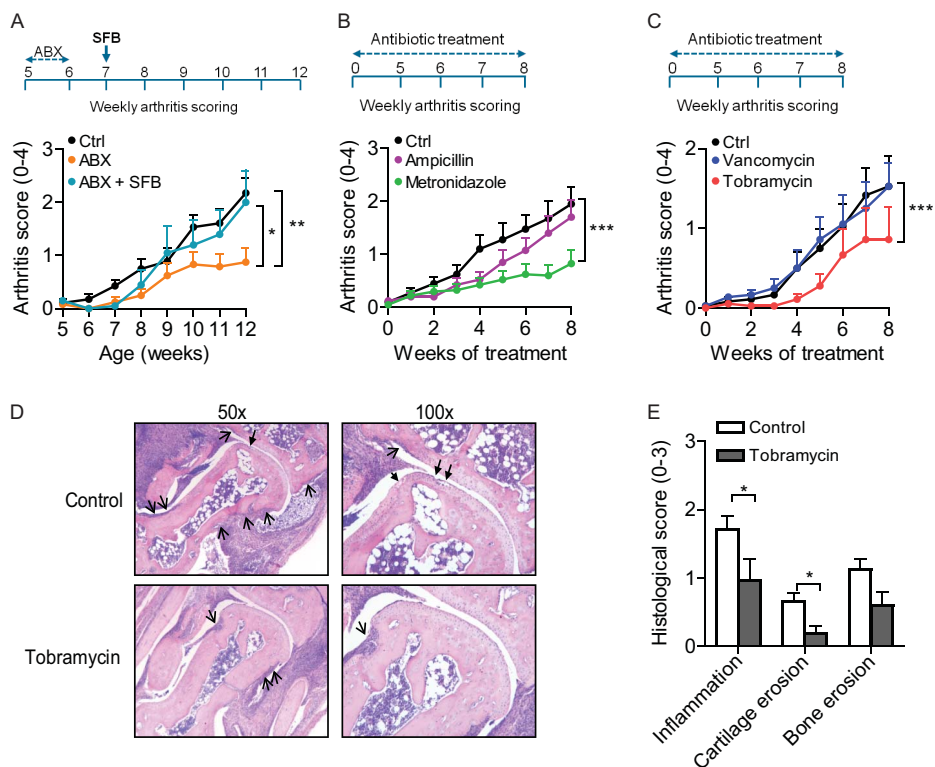


Figure 5. Commensal intestinal anaerobic tobramycin-sensitive microbiota promote arthritis in $IL1rn^{-/-}$ mice.

(A) Arthritis severity scores (0-2 per paw) of $IL1rn^{-/-}$ mice treated with a cocktail of metronidazole, neomycin and ampicillin (ABX) for 7 days (week 5 to 6), followed by re-colonization with SFB (ABX + SFB) one week after ending ABX treatment (at week 7). (B-C) Arthritis severity scores of $IL1rn^{-/-}$ mice either untreated (Ctrl) or treated with the mentioned antibiotic for 8 weeks. Data show mean + SEM of $n = 5-7$ (A) and $n = 10$ (B-C) mice per group. $*p \leq 0.05$, $***p \leq 0.001$, by repeated measures ANOVA with Bonferroni correction for multiple testing. (D) Representative images of ankle joint sections of control and tobramycin-treated mice stained with hematoxylin and eosin illustrating decreased synovial inflammation, cartilage destruction (closed arrows) and bone erosion (open arrows). Original magnification $\times 50$ (left panels) and $\times 100$ (right panels). (E) Histopathologic scores (mean + SEM) of synovial inflammation, cartilage destruction and bone erosion in control and tobramycin-treated $IL1rn^{-/-}$ mice. $n = 9$ per group. $*p \leq 0.05$, by Mann-Whitney U test.

To determine which subset of *IL1rn*^{-/-} microbiota triggers arthritis, we first compared the effects of treatment with ampicillin, broadly targeting aerobic bacteria, and metronidazole, broadly targeting anaerobic bacteria. To our surprise only metronidazole showed efficacy in reducing arthritis severity (Fig. 5B). This suggests involvement of anaerobic bacteria in the progression of arthritis. We next compared the effects of more selective antibiotics tobramycin and vancomycin, the latter of which has been reported to eradicate SFB and inhibit SFB-induced lamina propria Th17 cells and arthritis [80, 311, 328]. These experiments revealed that although SFB were able to exacerbate arthritis in *IL1rn*^{-/-} mice (Fig. 5A), only tobramycin but not vancomycin significantly diminished arthritis (Fig. 5C). To understand the changes in the microbiota induced by tobramycin treatment, we compared 16S rRNA gene sequences of fecal microbiota at the end-point of tobramycin treatment with the baseline microbiota. Among taxa with >0.1% relative abundance, tobramycin treatment resulted in a near-complete elimination of the genera *Helicobacter* and *Flexispira* (both belonging to the family Helicobacteraceae). In addition, a strong and highly significant reduction in the genera *Clostridium* and *Dehalobacterium* was observed (Supplementary Fig. S7 and Table S4). Other changes in the microbiota did not reach the statistical significance after Bonferroni correction for multiple testing. Therefore, tobramycin-induced alterations in these indigenous *IL1rn*^{-/-} microbiota taxa resulted in suppression of arthritis. This was confirmed by histological examination of arthritic joints which showed a significant reduction of synovial inflammation as well as cartilage destruction and a non-significant reduction in bone erosion upon treatment with tobramycin (Fig. 5D and E).

Aberrations of the intestinal microbiota and LP IL-17 production in *IL1rn*^{-/-} mice partly depend on TLR4.

TLR4 plays a major role in recognition of Gram-negative bacteria [329]. We previously showed that *IL1rn*^{-/-} *Tlr4*^{-/-} mice have a marked and sustained reduction of arthritis [310]. Therefore, we assessed whether TLR4 plays a role in alterations of the intestinal microbiota and the induction of LP Th17 cells. A detailed analysis of the intestinal microbiota showed that in addition to the TLR4-dependent loss of microbial diversity in IL-1Ra-deficient mice (Fig. 1B-F), alterations in *Ruminococcus*, *Streptococcus* and *Xylanibacter* were partially dependent on TLR4 and were restored in *IL1rn*^{-/-} *Tlr4*^{-/-} mice (Fig. 6A). Abundance of *Prevotella* was also restored to a statistically significant, yet minor extent (Fig. 6B). In total, 11 out of 44 taxa significantly altered in *IL1rn*^{-/-} mice were normalized toward the WT levels in *IL1rn*^{-/-} *Tlr4*^{-/-} mice (Supplementary Table S2).

We next examined the role of TLR4 in the mucosal T cell response in *IL1rn*^{-/-} mice. Th17 cells require transforming growth factor- β and IL-6, plus IL-1 β in mouse, for initial differentiation, and IL-23 for their functional maturation and pathogenic function [321, 330]. To determine the role of TLR4 in response to *IL1rn*^{-/-} intestinal microbial antigens, we cultured SI-LP mononuclear cells from *IL1rn*^{-/-} and *IL1rn*^{-/-} *Tlr4*^{-/-} mice *ex vivo* with autoclaved *IL1rn*^{-/-} intestinal microbiota. SI-LP mononuclear cells from *IL1rn*^{-/-} *Tlr4*^{-/-} mice produced significantly less IL-1 β (Fig. 7A, $p = 0.0042$). Furthermore, the induction of IL-23 and IL-6 by *IL1rn*^{-/-} fecal microbiota was partly TLR4-dependent (Fig. 7B and C, $p = 0.0014$ and $p = 0.009$, respectively). Reduced cytokine production in *IL1rn*^{-/-} *Tlr4*^{-/-}

mice was not due to an altered composition of mononuclear cells in the SI-LP, because the percentage and abundance of CD11c⁺ MHCII⁺ DCs as well as distinct subsets of CD103⁺ CD11b⁺, CD103⁺ CD11b⁻ and CD11b⁺ CD103⁻ phagocytes were similar between *IL1rn*^{-/-} and *IL1rn*^{-/-} *Tlr4*^{-/-} mice (data not shown). Importantly, stimulation of *IL1rn*^{-/-} LP mononuclear cells with *IL1rn*^{-/-} and *IL1rn*^{-/-} *Tlr4*^{-/-} fecal microbial antigens induced similar concentrations of IL-1 β , IL-23 and IL-6 ([Supplementary Fig. S8](#)). This suggests that the altered composition of microbiota in *IL1rn*^{-/-} *Tlr4*^{-/-} mice as such is not responsible for the lower production of these cytokines. These observations imply a significant role for TLR4 in intestinal production of the cytokines involved in LP Th17 differentiation in *IL1rn*^{-/-} mice.

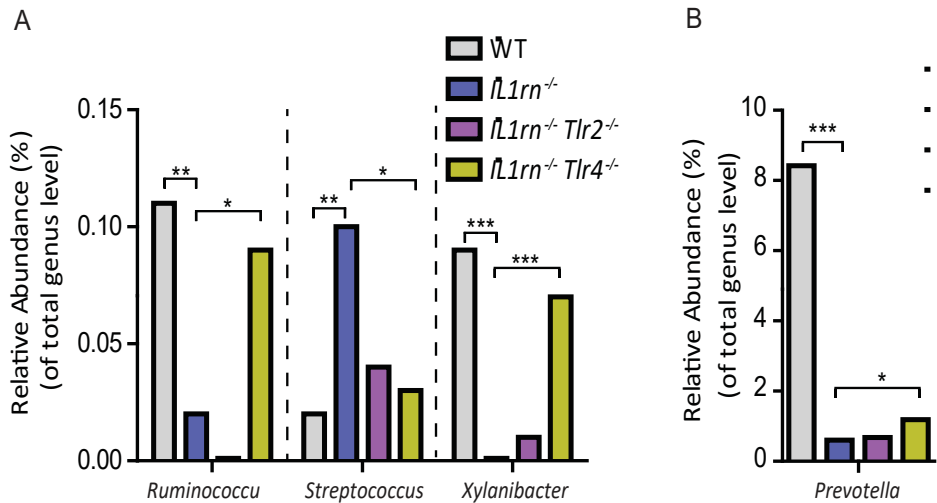


Figure 6. Alteration of specific intestinal microbiota in *IL1rn*^{-/-} mice is partly TLR4 dependent.

(A–B) Relative abundance of *Ruminococcus*, *Streptococcus*, *Xylanibacter* and *Prevotella* in wild-type (WT) ($n = 9$), *IL1rn*^{-/-} ($n = 15$), *IL1rn*^{-/-} *Tlr2*^{-/-} ($n = 8$) and *IL1rn*^{-/-} *Tlr4*^{-/-} ($n = 8$) mice. Data represent relative abundances of these genera obtained by 16S rRNA gene sequencing of the fecal microbiota.

When cultured *ex vivo* with PMA and ionomycin, SI-LP cells from *IL1rn*^{-/-} *Tlr4*^{-/-} mice produced significantly less IL-17 compared with cells from *IL1rn*^{-/-} *Tlr4*^{+/+} mice before the onset of arthritis ($p = 0.0028$; [Fig. 7D](#)). The amount of IFN γ produced in this culture was about 50 folds lower than IL-17 and was significantly reduced in *IL1rn*^{-/-} *Tlr4*^{-/-} mice as well ([Fig. 7E](#)). However, IL-4 levels remained unaffected ([Fig. 7F](#)). LPT cells from *IL1rn*^{-/-} *Tlr4*^{-/-} mice still produced significantly less IL-17 and IFN γ when these mice were co-housed with *IL1rn*^{-/-} (*Tlr4*^{+/+}) mice to transfer the microbiota ([Fig. 7G–I](#)). Therefore, reduced LP IL-17 production in *IL1rn*^{-/-} *Tlr4*^{-/-} mice is a result of the difference in host TLR4 expression rather than altered microbiota in *IL1rn*^{-/-} *Tlr4*^{-/-} versus *IL1rn*^{-/-} mice. Stimulation of LP mononuclear cells of these co-housed mice with fecal microbial antigens confirmed that cells from *IL1rn*^{-/-} *Tlr4*^{-/-} mice produce lower amounts of IL-1 β , IL-23 and IL-6 regardless of stimulation with *IL1rn*^{-/-} or *IL1rn*^{-/-} *Tlr4*^{-/-} microbiota ([Supplementary Fig. S9](#)). These data suggest that TLR4 activation contributes to intestinal LP production of IFN γ and most notably IL-17, and these effects precede the onset of arthritis.

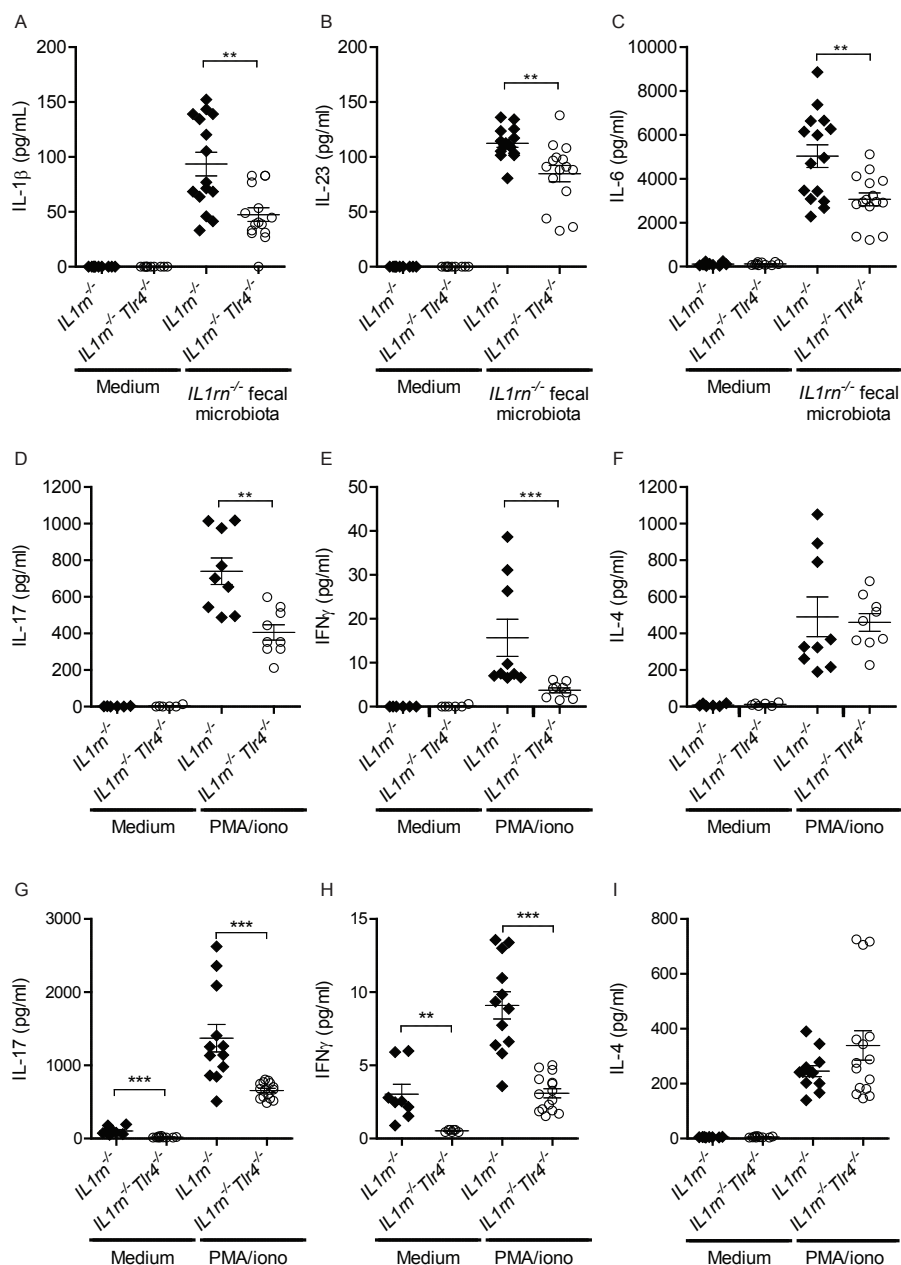


Figure 7. A significant role for TLR4 in intestinal production of cytokines involved in LP Th17 differentiation.

(A-C) Production of IL-1 β , IL-23 and IL-6 by SI-LP mononuclear cells of *IL1m*^{-/-} and *IL1m*^{-/-} *Tlr4*^{-/-} mice cultured in the presence of autoclaved *IL1m*^{-/-} complete fecal microbial antigens for 24 hours. (D-F) Cytokine production by SI-LP lymphocytes of separately-housed *IL1m*^{-/-} and *IL1m*^{-/-} *Tlr4*^{-/-} mice ex vivo stimulated with PMA and ionomycin for 5 hours. (G-I) Cytokine production by SI-LP lymphocytes of *IL1m*^{-/-} and *IL1m*^{-/-} *Tlr4*^{-/-} mice co-housed for 10 days. Cells were stimulated as in D-F. *p ≤ 0.05, **p ≤ 0.01, ***p ≤ 0.001, by Mann-Whitney U.

Importantly, production of IL-17 but not IFN γ in lymph nodes draining the inflamed joints was diminished in *IL1rn^{-/-} Tlr4^{-/-}* mice compared with *IL1rn^{-/-}* mice ([Supplementary Fig. S10](#)). This is in agreement with our previous study showing that TLR4 induces systemic and local IL-17 production and promotes arthritis in *IL1rn^{-/-}* mice [310]. Together these observations suggest an essential role for TLR4 in the induction of intestinal LP IL-17 production associated with extra-intestinal IL-17 levels and the development of arthritis in *IL1rn^{-/-}* mice.

Discussion

The intestinal microbiome has emerged as a key determinant of health and disease. Although advanced sequencing techniques have enabled microbiome profiling in rheumatic patients, study of the underlying mucosal responses and the functional impact on arthritis is limited in human subjects due to the requirement of invasive techniques. The animal studies presented here demonstrate aberrations in intestinal microbiota in mice developing spontaneous autoimmune arthritis, introduce commensal tobramycin-sensitive microbiota as potential triggers for arthritis, and suggest a role for TLR4 activation in mucosal induction of inflammatory pathways including Th17 induction associated with arthritis.

Our study identifies loss of microbial diversity and specific taxonomic alterations in the microbiota of autoimmune-prone *IL1rn^{-/-}* mice. Importantly, loss of intestinal microbial diversity and richness also coincides with human autoimmune diseases such as diabetes, rheumatoid and psoriatic arthritis [312, 314, 331]. Among the microbiota increased in *IL1rn^{-/-}* mice, *Streptococcus* species are known inducers of chronic TLR-mediated arthritis in animal models when injected intra-articularly [332, 333]. Furthermore, a commensal *Helicobacter* (*H. hepaticus*) has been shown to induce IL-23 and mediate T cell-dependent gut inflammation in immunocompromised mice [334]. The decreased *Barnesiella* in *IL1rn^{-/-}* mice is consistent with a previous study associating the abundance of *Barnesiella* with resistance to arthritis in HLA-DRB1*0402 mice [335]. A specific species of *Prevotella*, *P. copri*, is overrepresented in patients with new-onset RA [312], and was recently shown to increase colonic Th17 cells and promote arthritis in SKG mice after co-exposure to the fungal component zymosan [316]. On the other hand, *P. histicola* has been reported to suppress collagen-induced arthritis in transgenic mice expressing RA-susceptibility gene HLA-DQ8 [336]. Therefore, the immunomodulatory effects of the gut microbiota, including *Prevotella*, are species- and sometimes even strain-dependent. Due to inherently limited resolution of 16S rRNA gene sequencing, our data on the abundance of *Prevotella* is limited to the genus level and the exact *Prevotella* species altered in *IL1rn^{-/-}* mice remain unclear. Overall, it is tempting to speculate that complex alterations in several taxa determine the net mucosal response to affect arthritis. It should also be noted that fecal bacterial community structures do not fully mirror the site-specific luminal or mucosa-associated microbiota profiles and were used in this study as a proxy of the gut microbiota of the *IL1rn^{-/-}* mice.

Our previous studies showed that *IL1rn^{-/-} Tlr2^{-/-}* mice develop a more severe arthritis compared with *IL1rn^{-/-}* mice [310]. Given that additional TLR2 deficiency did not affect

the microbiota of *IL1rn*^{-/-} mice to a major extent (Fig. 1), we speculate that severe arthritis in *IL1rn*^{-/-} *Tlr2*^{-/-} mice is due to the altered host immune response, specifically reduced function of Treg cells [310], rather than alteration in the microbiome. However, the data regarding lack of a major influence of TLR2 deficiency on *IL1rn*^{-/-} microbiota should be interpreted with caution due to the absence of littermate *IL1rn*^{-/-} *Tlr2*^{+/+} mice in our studies.

IL1rn^{-/-} mice had specific expansion of intestinal Th17 cells. The pathogenic relevance of IL-17 in the development of arthritis in *IL1rn*^{-/-} mice has been demonstrated before, since both IL-17 gene deficiency and treatment with neutralizing anti-IL-17 antibodies inhibit arthritis [302, 306]. A previous study showed that $\gamma\delta$ T cells rather than Th17 cells represent most IL-17-producing T cells in the inflamed joints of *IL1rn*^{-/-} mice [337]. While $\gamma\delta$ 17 and Th17 cells may have complementary pathogenic roles in the development of *IL1rn*^{-/-} arthritis, our data suggest that IL-17-producing cells located in lamina propria and induced by *IL1rn*^{-/-} intestinal microbiota are TCR β -expressing CD4⁺ Th17 cells (Fig. 4 and Supplementary Fig. S4).

The expansion of LP Th17 cells in *IL1rn*^{-/-} mice was caused by the dysregulated microbiota as confirmed by fecal transfer experiments (Fig. 3G-I). A critical pathogenic link to the spontaneous arthritis was revealed by our germ-free and antibiotic treatment studies (Figs. 4 and 5). Other previous studies which demonstrated the involvement of the gut microbiota in exacerbation of autoimmune arthritis found SFB as the responsible microorganisms. One study showed a role for vancomycin-sensitive microbiota including SFB in the induction of IL-17- and autoantibody-driven arthritis in K/BxN mice [311], and another showed that SFB can lower the activation threshold of self-reactive T cells and promote the differentiation of arthritogenic Th1 cells in a T cell transfer model of arthritis [338]. Our data are the first to demonstrate that although SFB colonization exacerbates arthritis, among the dysregulated indigenous microbiota present in the *IL1rn*^{-/-} mice, those sensitive to tobramycin, i.e., *Helicobacter*, *Flexispira*, *Clostridium* and *Dehalobacterium*, are potential candidates to promote arthritis in a genetically susceptible host. This is relevant for human disease, given that SFB were not found in genome-wide sequences of 263 gut metagenomes of human adults [18, 327].

Our data also provide the first evidence for the involvement of TLR4 in defining the intestinal mucosal T cell phenotype. TLR4 activation of LP mononuclear cells by *IL1rn*^{-/-} microbiota induced IL-1 β , IL-23 and IL-6 (Fig. 7A-C). Microbiota-induced IL-1 β is critical for the development of steady-state Th17 cells in the gut [339]. IL-1 also synergizes with IL-6 and IL-23 to regulate early differentiation of Th17 cells and maintain cytokine expression in effector Th17 cells [340]. It was recently shown that infectious triggers such as influenza lung infection and colitis trigger an IL-1 β -induced Th17 differentiation and promote arthritis induced by KRN transgenic T cells [341]. Interestingly, a subset of human CD14⁺ CD163^{low} lamina propria cells expressing both macrophage and DC markers has been found to express TLR4, produce IL-1 β and IL-6 upon TLR4 stimulation, and induce Th17 differentiation [342]. However, the specific subset of LP phagocytes that orchestrates the phase-dependent TLR4-mediated mucosal response to microbiota in our studies remains to be determined.

Several studies have shown that TLR4 deficiency and systemic inhibition of TLR4 using specific antagonists or neutralizing antibodies can suppress experimental arthritis [310, 343-345]. Importantly, TLR4 is believed to be hyper-responsive in both blood monocyte-derived DCs and CD14⁺ synovial fluid macrophages of RA patients compared with healthy controls [346, 347]. Pathways associated with TLR signaling are upregulated in synovial fluid macrophages of patients with RA. A proinflammatory role for TLR4 during arthritis has previously been widely attributed to TLR4 activation by endogenous damage-associated molecular patterns present in the joint rather than microbial agonists [310, 345, 348, 349]. Our observations suggest that TLR4-mediated modulation of the mucosal immune response in intestinal LP may be another function involving TLR4 in arthritis.

Conclusions

Our study reveals a crucial role for IL-1Ra in regulation of the diversity and the composition of intestinal microbiota and a balanced T cell response in the intestinal LP. We show that the aberrant microbiota in *IL1rn*^{-/-} mice have the capacity to enhance LP Th17 cells which are associated with arthritis, likely via TLR4-induced production of IL-1 β , IL-6 and IL-23. Although *IL1rn*^{-/-} intestinal microbiota do not cause arthritis in a normal (WT) host, these microbiota, in particular tobramycin-sensitive bacteria, contribute to the development of arthritis in *IL1rn*^{-/-} mice. Our data suggest that the interplay between IL-1Ra, intestinal microbiota, TLR4 and mucosal T cells may serve as a potential predisposing or initiating event in the context of autoimmune disease and provide opportunities to control RA.

Methods

Mice

IL1rn^{-/-} mice on a BALB/c background were kindly provided by Dr. M. Nicklin (Sheffield, UK) [350]. WT BALB/c mice were purchased from Harlan, UK. Mice were co-housed in filter-top non-individually-ventilated (non-IVC) cages in the same room in our animal facility for at least 8 weeks prior to feces collection for pyrosequencing. *IL1rn*^{-/-} *Tlr4*^{-/-} mice and their *IL1rn*^{-/-} *Tlr*^{+/+} littermates were generated as described before [310] and used for microbiota sequencing. *IL1rn*^{-/-} *Tlr2*^{-/-} mice were compared to non-littermate *IL1rn*^{-/-} mice in this study.

Microbiota sequencing and data analysis

Fecal bacterial DNA from 15-week-old mice was isolated using phenol-, chloroform-, isoamyl alcohol-based extraction (Sigma). Sequencing was performed by DNA Vision (Charleroi, Belgium) on a Roche 454 GS-FLX System using 16S rRNA gene bar-coded primers targeting the V5-V6 conserved DNA regions (forward primer 784F: 5'-AGGATTAGATACCCTGGTA-3', reverse primer 1061R: 5'-CRRACGAGCTGACGAC-3') [351]. For gene sequence analysis, a customized workflow based on Quantitative Insights Into Microbial Ecology (QIIME version 1.2) was adopted (<http://qiime.org/>) [42]. Settings recommended in QIIME 1.2 tutorial were applied. Additionally, reads were

filtered for chimeric sequences using Chimera Slayer as described before [352]. OTU clustering was performed with settings as recommended by QIIME [353] using an identity threshold of 97%. The Ribosomal Database Project classifier version 2.2 was used for taxonomic classification [287]. Hierarchical clustering of samples was performed using UPGMA with weighted UniFrac as a distance measure as implemented in QIIME 1.2. For statistical analysis and generation of figures, a custom QIIME implemented R-package, SciPy [354] (www.Scipy.org), Graphpad Prism version 5.0, and Microsoft® Office Excel® 2007 were adopted. Presence of SFB was assessed by real-time quantitative PCR (qPCR) on fecal DNA using SFB specific primers as described before [355]. The delta Ct (cycle threshold) value was calculated for SFB-specific rRNA gene relative to the total (conserved) bacterial 16S rRNA genes amplified using universal bacterial primers to correct for the total bacterial DNA input. Data are presented as delta Ct (ΔCt) and relative SFB expression calculated as $2^{-\Delta Ct} \times 10,000$ ([Supplementary Table S3](#)).

Microbiota transfer and co-housing

IL1rn^{-/-} microbiota were transferred to WT mice by oral gavage of 200 μ l of a homogenized *IL1rn^{-/-}* fecal suspension prepared in sterile PBS. Immediately hereafter, the gavaged WT mice were co-housed with *IL1rn^{-/-}* mice in the same individually-ventilated cage for a period of 6 weeks to ensure sustained microbiota transfer by coprophagy. The control WT mice were gavaged with their own fecal suspension and housed separate from *IL1rn^{-/-}* fecal-transplanted mice in IVC cages. To verify microbiota of conventional *IL1rn^{-/-}* mice can trigger arthritis, GF *IL1rn^{-/-}* mice received either 200 μ l of sterile water or 200 μ l fecal suspensions of conventional *IL1rn^{-/-}* mice and were monitored for the development of arthritis for 8 weeks. In some studies, *IL1rn^{-/-}* mice were co-housed with *IL1rn^{-/-}Tlr4^{-/-}* mice for 10 days before analysis of LP T cells.

Antibiotic treatments and reconstitution with SFB

Intestinal microbiota were depleted using a cocktail of metronidazole (Acros Organics), neomycin trisulfate (Sigma) and ampicillin sodium salt (Sigma) (all 1 g/l) provided in drinking water for 1 week. Indicated groups received 200 μ l fecal suspensions of SFB-monocolonized mice by oral gavage 1 week after ceasing antibiotics. For single antibiotic treatments, ampicillin sodium salt (1 g/l), metronidazole (1g/l), vancomycin hydrochloride (0.5 g/l, Fisher Scientific) or tobramycin sulfate (1 g/l, Centrafarm) was added to drinking water for 8 weeks and refreshed once a week. Sucrose (6 g/l) was added to drinking water of all groups including controls during treatments.

Isolation of lamina propria cells

Lamina propria mononuclear cells were isolated from small intestine and colon after removing mesenteric fat and Peyer's patches, followed by incubation with 33 mM EDTA on ice for 30 minutes to remove epithelial cells, and subsequent digestion with 1 mg/ml collagenase-D (Roche) and 10 μ g/ml DNase I (Sigma) at 37 °C for three cycles of 15 minutes. LP lymphocytes were then harvested at the interphase of a 40:80% percoll gradient (Sigma), washed thoroughly and used in culture as described below.

Cell cultures and cytokine measurements

SI-LP mononuclear cells (4×10^5 cells/well), LN cells (2×10^5 cells/well) and splenocytes (1×10^5 cells/well) were cultured in round-bottom 96 well plates in the presence of PMA (50 ng/ml; Sigma) and ionomycin (1 μ g/ml; Sigma) for 5 or 6 hours, as indicated in figure legends. SI-LP cells were also cultured for 24 hours in the presence of *IL1rn*^{-/-} or *IL1rn*^{-/-} *Tlr4*^{-/-} complete microbial antigens (1:200 v/v ratio) prepared by autoclaving the *IL1rn*^{-/-} fecal pellets dissolved in PBS, and then centrifuging the suspension at 2000 rpm for 5 minutes. Cytokine levels in culture supernatants were measured by Luminex using the mouse cytokine/chemokine magnetic bead kit (Milliplex and Bio-Rad).

Flow cytometry

For intracellular cytokine staining, SI-LP cells were incubated with PMA (50 ng/ml; Sigma), ionomycin (1 μ g/ml; Sigma), and Brefeldin A (1 μ l/ml; BD Biosciences) at 37°C for 4 hours. Cells were stained with fixable viability dye Efluor780 (eBioscience), anti-TCR β -FITC (Biolegend) or anti-CD3-PE (BD pharmingen), and anti-CD4-APC (Biolegend), then fixed and permeabilized using fixation/permeabilization buffer (eBioscience) and stained with anti-IL-17-FITC (Biolegend), IL-17-PECy7 (Biolegend) or anti-IFN γ -FITC (BD pharmingen) in permeabilization buffer (eBioscience).

Assessment of arthritis

Severity of arthritis was scored using a previously standardized arbitrary scoring system on a 0-2 scale per paw [310]. Arthritis developed only in ankle joints (maximum score of 4). For histology, total ankle joints were isolated and fixed in 4% formaldehyde for 4 days, thereafter decalcified in 5% formic acid and embedded in paraffin. Tissue sections of 7 μ m were stained using Haematoxylin & Eosin to study synovial inflammation, cartilage destruction and bone erosion. Each parameter was scored on a scale from 0-3 in a blinded manner.

Statistics

Group measures are expressed as mean + SEM. Statistical significance was tested using an unpaired two-tailed Mann-Whitney U test to compare two and ANOVA to compare more groups, with Bonferroni correction for multiple testing when applicable (GraphPad Prism 5.0). Arthritis scores were compared using repeated measures ANOVA with Bonferroni correction. Significance is indicated on figures as follows: *n.s.* (not significant), **p* \leq 0.05, ***p* \leq 0.01, ****p* \leq 0.001.

Availability of data and materials

All 16S rRNA gene sequencing reads data are publicly available at the European Nucleotide Archive (ENA) database (<http://www.ebi.ac.uk/ena>) under study accession number PRJEB7447 (or secondary accession number ERP007176).

Supplementary Figures

Figure S1. Hierarchical clustering of wild-type and *IL1rn*^{-/-} mice based on intestinal microbiota. (WORD)

(online) Figure S1 of Additional file 1 at :
https://doi.org/10.6084/m9.figshare.c.3810208_D1.v2

Figure S2. The impact of lineage origin versus *IL1rn*-deficiency on the overall fecal microbiota composition. (WORD)

(online) Figure S2 of Additional file 1 at :
https://doi.org/10.6084/m9.figshare.c.3810208_D1.v2

Figure S3. Gating strategy. (WORD)

(online) Figure S3 of Additional file 1 at :
https://doi.org/10.6084/m9.figshare.c.3810208_D1.v2

Figure S4. Frequencies and numbers of IL-17-producing cells among TCRβ⁺ and TCRβ⁻ T cell populations with and without CD4 expression. (WORD)

(online) Figure S4 of Additional file 1 at :
https://doi.org/10.6084/m9.figshare.c.3810208_D1.v2

Figure S5. Colonization of germ-free (GF) *IL1rn*^{-/-} mice with fecal microbiota of conventional *IL1rn*^{-/-} mice increases the severity of arthritis. (WORD)

(online) Figure S5 of Additional file 1 at :
https://doi.org/10.6084/m9.figshare.c.3810208_D1.v2

Figure S6. Increased expression of Th2/Treg cytokines in spleens but not popliteal lymph nodes (LN) of germ-free *IL1rn*^{-/-} mice. (WORD)

(online) Figure S6 of Additional file 1 at :
https://doi.org/10.6084/m9.figshare.c.3810208_D1.v2

Figure S7. Effects of 8 weeks oral tobramycin treatment on microbiota of *IL1rn*^{-/-} mice assessed by 16S gene sequencing of fecal bacterial DNA. (WORD)

(online) Figure S7 of Additional file 1 at :
https://doi.org/10.6084/m9.figshare.c.3810208_D1.v2

Figure S8. *IL1rn*^{-/-} and *IL1rn*^{-/-} *Tlr4*^{-/-} microbiota induce similar cytokine response in lamina propria mononuclear cells. (WORD)

(online) Figure S8 of Additional file 1 at :
https://doi.org/10.6084/m9.figshare.c.3810208_D1.v2

Figure S9. Lamina propria mononuclear cells of *IL1rn*^{-/-} *Tlr4*^{-/-} mice co-housed with *IL1rn*^{-/-} mice produce less Th17-inducing cytokines. (WORD)

(online) Figure S9 of Additional file 1 at :

https://doi.org/10.6084/m9.figshare.c.3810208_D1.v2

Figure S10. Decreased IL-17 production in draining lymph nodes of TLR4 deficient mice. (WORD)

(online) Figure S10 of Additional file 1 at :

https://doi.org/10.6084/m9.figshare.c.3810208_D1.v2

Supplementary Tables

Table S1. Sequencing read numbers. (WORD)

(online) Table S1 of Additional file 1 at :

https://doi.org/10.6084/m9.figshare.c.3810208_D1.v2

Table S2. TLR4 deficiency normalizes specific aberrations in *IL1rn*^{-/-} intestinal microbiome towards WT level. (WORD)

(online) Table S2 of Additional file 1 at :

https://doi.org/10.6084/m9.figshare.c.3810208_D1.v2

Table S3. Assessment of the presence of SFB in WT, *IL1rn*^{-/-}, *IL1rn*^{-/-}*Tlr2*^{-/-}, *IL1rn*^{-/-}*Tlr4*^{-/-} mice. (WORD)

(online) Table S3 of Additional file 1 at :

https://doi.org/10.6084/m9.figshare.c.3810208_D1.v2

Table S4. Alterations in fecal microbiota with a relative abundance >0.1% by oral tobramycin, sorted by the abundance at the baseline. (WORD)

(online) Table S4 of Additional file 1 at :

https://doi.org/10.6084/m9.figshare.c.3810208_D1.v2

CHAPTER 7

OPEN ACCESS

as published in PLoS One, 2017, Sep 1;12(9):e0183509

<https://doi.org/10.1371/journal.pone.0183509>

¹ Centre for Cognitive Neuroimaging, Donders Institute for Brain, Cognition and Behaviour, Radboud University, Nijmegen, The Netherlands.

² Center for Molecular and Biomolecular Informatics, Radboud Institute for Molecular Life Sciences (RIMLS), Radboud University Medical Center (Radboudumc), Nijmegen, The Netherlands.

³ Department of Cognitive Neuroscience, Donders Institute for Brain, Cognition and Behaviour, Radboudumc, Nijmegen, The Netherlands.

⁴ NIZO, Ede, The Netherlands.

⁵ Department of Internal Medicine and Radboud Center for Infectious Diseases, Radboudumc, NL.

⁶ Karakter Child and Adolescent Psychiatry University Centre, Nijmegen, The Netherlands.

⁷ Department of Psychiatry, Donders Institute for Brain, Cognition and Behaviour, Radboudumc, NL.

⁸ Department of Human Genetics, Donders Institute for Brain, Cognition and Behaviour, Radboudumc, Nijmegen, The Netherlands.

These authors contributed equally to this work.

¶ These authors also contributed equally to this work.

CHAPTER 7

THE GUT MICROBIOME IN ADHD AND ITS RELATION TO NEURAL REWARD ANTICIPATION

Esther Aarts ^{1 #}

Thomas H.A. Ederveen ^{2 #}

Jilly Naaijen ³

Marcel P. Zwiers ¹

Jos Boekhorst ^{2,4}

Harro M. Timmerman ⁴

Sanne P. Smeekeens ⁵

Mihai G. Netea ⁵

Jan K. Buitelaar ^{3,6}

Barbara Franke ^{7,8}

Sacha A.F.T. van Hijum ^{2,4 ¶}

Alejandro Arias Vasquez ^{3,7,8 ¶}

ABSTRACT

BACKGROUND. Microorganisms in the human intestine (i.e. the gut microbiome) have an increasingly recognized impact on human health, including brain functioning. Attention-deficit/hyperactivity disorder (ADHD) is a neurodevelopmental disorder associated with abnormalities in dopamine neurotransmission and deficits in reward processing and its underlying neuro-circuitry including the ventral striatum. The microbiome might contribute to ADHD etiology via the gut-brain axis. In this pilot study, we investigated potential differences in the microbiome between ADHD cases and undiagnosed controls, as well as its relation to neural reward processing.

METHODS. We used 16S rRNA marker gene sequencing (16S) to identify bacterial taxa and their predicted gene functions in 19 ADHD and 77 control participants. Using functional magnetic resonance imaging (fMRI), we interrogated the effect of observed microbiome differences in neural reward responses in a subset of 28 participants, independent of diagnosis.

RESULTS. For the first time, we describe gut microbial makeup of adolescents and adults diagnosed with ADHD. We found that the relative abundance of several bacterial taxa differed between cases and controls, albeit marginally significant. A nominal increase in the *Bifidobacterium* genus was observed in ADHD cases. In a hypothesis-driven approach, we found that the observed increase was linked to significantly enhanced 16S-based predicted bacterial gene functionality encoding cyclohexadienyl dehydratase in cases relative to controls. This enzyme is involved in the synthesis of phenylalanine, a precursor of dopamine. Increased relative abundance of this functionality was significantly associated with decreased ventral striatal fMRI responses during reward anticipation, independent of ADHD diagnosis and age.

CONCLUSIONS. Our results show increases in gut microbiome predicted function of dopamine precursor synthesis between ADHD cases and controls. This increase in microbiome function relates to decreased neural responses to reward anticipation. Decreased neural reward anticipation constitutes one of the hallmarks of ADHD.

Background

Attention-deficit/hyperactivity disorder (ADHD) is a common neuropsychiatric disorder, characterized by symptoms of inattention and/or impulsivity and hyperactivity. ADHD has been associated with abnormalities in the monoamine neurotransmitter systems dopamine and noradrenaline [356]. Stimulant medication, for example, is highly effective in improving ADHD symptoms by inhibition of re-uptake of dopamine and noradrenaline by their transporters [357]. Moreover, brain functions linked to dopamine processing, such as reward anticipation [358], have been found abnormal in ADHD, as reflected by diminished brain responses in ventral striatum (including nucleus accumbens) in functional magnetic resonance imaging (fMRI) studies [359–362]. ADHD is highly heritable [363, 364], and genetic studies have pointed to a role of dopamine-, noradrenaline-, and serotonin-related genes in ADHD [363], but these studies showed small effects suggesting that environmental factors also play a role in the etiology of ADHD.

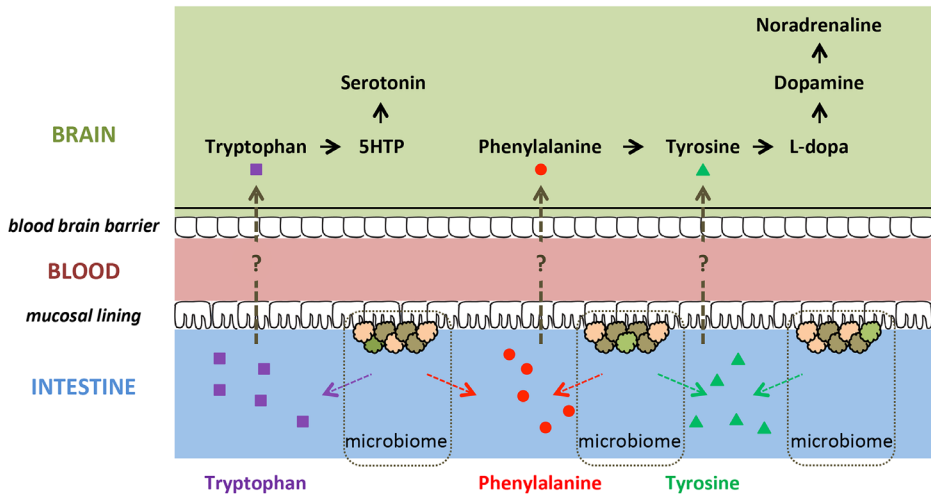


Figure 1. Potential routes in which precursors of monoamines could influence brain functioning.

The large neutral amino acids tryptophan, phenylalanine, and tyrosine, which are absorbed in the intestine [375], are precursors of monoamines. Tryptophan and phenylalanine are essential amino acids, meaning that they cannot be synthesized by the human body itself [378]. 5-HTP = 5-Hydroxytryptophan.

Meta-analyses of non-pharmacologic treatment interventions for ADHD showed that restriction diets for ADHD patients (usually directed at eliminating potential allergens) may lead to a significant reduction in ADHD symptoms, although there is heterogeneity across studies [365–367]. Conceivably, diet might influence behavior and ADHD symptoms by affecting gut microorganisms (i.e. the gut microbiome) [368]. The gut microbiome has an increasingly recognized impact on brain functioning and behavior [369]. One proposed mechanism for the effects of gut microbiota on brain and behavior is through their ability to synthesize neurochemicals and their

precursors that are analogous in structure to those of the host nervous system [370]. Precursors of monoamines involved in ADHD (i.e. dopamine, noradrenaline, serotonin; see above) are produced by several members of the gut microbiota [371-373]. These precursors (i.e. phenylalanine, tyrosine, tryptophan) might be absorbed through the intestinal epithelium, enter the portal circulation [370], and cross the blood-brain barrier; in this way, they could potentially influence host monoamine synthesis (Fig. 1). Consequently, differences in abundance and/or metabolic activity of monoamine precursor-producing inhabitants of the gastrointestinal tract may affect monoamine-related brain functioning and behavior relevant to ADHD. Indeed, a lowered abundance of *Bifidobacterium* in infancy has been associated with increased risk of developing ADHD and Asperger syndrome in childhood in a study focusing on particular microbiota [374]. However, we were not able to identify studies that investigated the (complete) gut microbiome in relation to ADHD, and how differences in microbiome structure might affect brain functioning.

Here, we present the first microbiome study in ADHD patients versus healthy controls. We used bacterial 16S ribosomal RNA (rRNA) marker gene sequencing to characterize microbial communities in adolescents and young adults with ADHD and self-reported healthy controls. We specifically investigated the difference in relative abundance of monoamine precursor-related predicted genes between these microbiomes, and their potential effect on brain activity. We focused on the precursors in the monoamine biosynthesis, as monoamines themselves cannot cross the blood-brain barrier [375], and indeed found overabundance of 16S-based predicted bacterial gene functionality related to phenylalanine synthesis in ADHD. Next, using fMRI, we first replicated the finding of reduced reward anticipation in ventral striatum in ADHD (see above) in a partly overlapping sample of the same cohort. Finally, in the subset of participants with both microbiome and fMRI measurements, we assessed how microbiome function related to neural responses during reward anticipation. We found that reduced reward anticipation in ventral striatum was related to an increase in (predicted) bacterial functionality with regard to production of dopamine's precursor phenylalanine, independent of diagnosis.

Results

Demographics

The ADHD cases and healthy participants did not differ in BMI and gender in the three groups (microbiome, fMRI, and their overlap) (Table 1). No differences in age were present for patients and healthy persons in the fMRI samples, but for the microbiome analyses, controls were older (Table 1). This was due to the fact that the 39 BIG participants were older (33.1 years \pm 17.7 SD). Controls from the NeuroIMAGE II study (21 unaffected siblings [22.3 years \pm 3.7 SD], and 17 healthy controls [19.1 years \pm 3.2 SD]) did not differ from the cases ($t(55)=-1.46$, $p = 0.15$). Consequently, the functional (PICRUST) analyses below were performed with and without the older BIG controls to balance power and homogeneity of the sample.

Table 1. Descriptive characteristics of the study samples.

^a Four sibling pairs were included in the ADHD group, ten sibling pairs and two trio's in the control group. Four ADHD cases had one sibling in the control group. No BMI was available for four control subjects. ^b Four sibling pairs were included in the ADHD group, 13 sibling pairs and six trio's in the control group. Nine ADHD cases had one sibling in the control group; one ADHD case had two siblings in the control group. No BMI was available for two ADHD and three control subjects. Initially, 95 participants performed the reward anticipation task during fMRI. However, four ADHD participants and four control participants were excluded from the fMRI analyses: five were excluded due to excessive (i.e. > 6 mm) movement (three ADHD + two controls), one due to an incomplete data set (ADHD), one due to too many errors (48%, control), and one due to extensive signal drop-out (control). ^c One sibling pair was included in the ADHD group, eight sibling pairs in the control group. One ADHD case had one sibling in the control group. No BMI was available for two control subjects. ^d Differences in gender between the groups were tested with chi-square or Fisher's exact test, as appropriate. All other differences were tested with an independent t-test. ^e No symptoms were available for $n = 39$ control subjects. ^f Metrics for alpha diversity; reads were down-sampled to 1126 reads per sample, average of 4 trials, for the calculation of this diversity metric. N/A: Not Available; BMI: Body Mass Index; OTU: Operational Taxonomic Unit; SD: Standard Deviation; *n.s.* not significant.

	Microbiome Analysis ^a			Imaging Analysis ^b			Microbiome & Imaging Analysis ^c		
	ADHD (n=19)	Controls (n=77)	<i>P</i>	ADHD (n=24)	Controls (n=63)	<i>P</i>	ADHD (n=6)	Controls (n=22)	<i>P</i>
Age in Years (SD)	19.5 (2.5)	27.1 (14.3)	.024	20.3 (3.7)	21.3 (3.4)	<i>n.s.</i>	18.6 (2.5)	21.1 (3.3)	<i>n.s.</i>
BMI (SD)	23.8 (4.1)	23.0 (3.2)	<i>n.s.</i>	22.8 (3.5)	22.7 (2.9)	<i>n.s.</i>	22.1 (4.4)	23.4 (3.7)	<i>n.s.</i>
% Males ^d	68.4	53.2	<i>n.s.</i>	75	61.9	<i>n.s.</i>	66.7	59.1	<i>n.s.</i>
Mean Inattention Symptoms (SD)	6.5 (2.1)	0.7 (1.4) ^e	<.001	6.2 (1.5)	0.5 (1.0)	<.001	6.0 (1.6)	0.7 (1.3)	<.001
Mean Hyperactivity Symptoms (SD)	4.4 (2.1)	0.6 (1.1) ^e	<.001	4.2 (2.4)	0.7 (1.2)	<.001	5.0 (1.4)	0.7 (1.2)	<.001
Mean Total Symptoms (SD)	11.0 (2.9)	1.3 (2.4) ^e	<.001	10.3 (3.0)	1.2 (2.1)	<.001	11.0 (1.8)	1.4 (2.4)	<.001
Mean Number of Reads (SD)	3893 (783)	3760 (1038)	<i>n.s.</i>	N/A	N/A		N/A	N/A	
Total Number of Reads	73,969	289,492		N/A	N/A		N/A	N/A	
→ Assigned at Phylum	73,876 (99.9%)	289,117 (99.9%)	<i>n.s.</i>	N/A	N/A		N/A	N/A	
→ Assigned at Genus	61,657 (83.4%)	230,370 (79.6%)	<i>n.s.</i>	N/A	N/A		N/A	N/A	
Mean Number of OTU (SD)	389 (103)	429 (157)	<i>n.s.</i>	N/A	N/A		N/A	N/A	
Total Number of OTU	7041	33,048		N/A	N/A		N/A	N/A	
Mean Shannon index ^f	5.3	5.2	<i>n.s.</i>	N/A	N/A		N/A	N/A	
Mean Chao index ^f	604.5	579.7	<i>n.s.</i>	N/A	N/A		N/A	N/A	

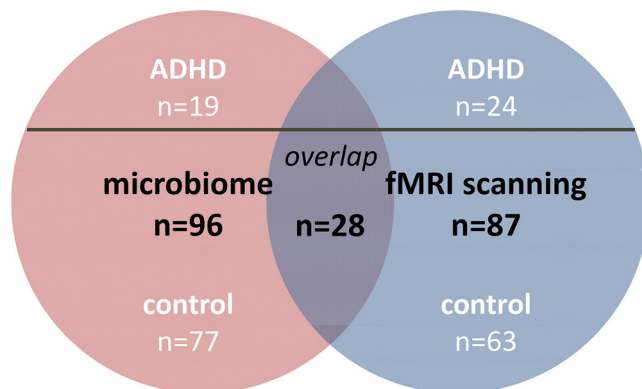


Figure 2. Microbiome sample, fMRI sample, and their overlap.

Gut microbiome taxa

For sample characteristics, diversity metrics and read counts, see [Table S1](#) and [Supplementary Data S1A](#). The overall make-up of the 96 microbiomes consisted of bacteria predominantly from the phyla Firmicutes (77.92%), Actinobacteria (15.68%) and Bacteroidetes (6.05%) ([Fig. 3](#), [Supplementary Fig. S3](#) and [Table S2](#)). We found an increase of Actinobacteria (controls: 14.08% to ADHD: 22.14%; $p = 0.002$, uncorrected), which seemed to occur mainly at the expense of Firmicutes (controls: 79.80% and ADHD: 70.29%; $p = 0.001$, uncorrected), as Bacteroidetes (and other phyla) did not differ significantly (controls: 5.74% to ADHD: 7.29%; $p = 0.166$, uncorrected) in relative abundance between healthy participants and those with ADHD ([Table S2](#); MWU, multiple comparisons corrected p-value threshold = 0.00017). Interestingly, in more taxonomic detail, within the phylum Actinobacteria, the genus *Bifidobacterium* was significantly increased in ADHD cases (controls: 12.66% to ADHD: 20.47%; $p = 0.002$, MWU, uncorrected) ([Fig. 3](#)). *Bifidobacterium* dominates the gut early in life, and slowly decreases in relative abundance during ageing [376]. In order to exclude an age-driven shift in *Bifidobacterium*, we defined an age-matched subsample for our microbiome cohort ([Supplementary Data S1B](#)). Using 15 pairs of ADHD cases and age-matched controls we find a similar rise in *Bifidobacterium* in ADHD-affected individuals (controls: 13.77% to ADHD: 18.90%) ($p = 0.034$, MWU, uncorrected). Furthermore, the order Clostridiales, within the phylum Firmicutes, was found to be decreased in ADHD cases (controls: 77.37% to ADHD: 69.02%; $p = 0.003$, MWU, uncorrected) and best explains the observed drop in Firmicutes, but no specific taxonomical entity belonging to the order Clostridiales was found to be responsible for this effect ([Fig. 3](#)). In conclusion, the genus *Bifidobacterium* shows the most (statistically) strong and taxonomic most specific change as effect of ADHD status.

We selected five candidate taxa with the greatest difference between ADHD and controls to be used in our subsequent analyses of metabolic potential (by PICRUSt): Clostridiales (order), Rikenellaceae (family), Porphyromonadaceae (family), *Bifidobacterium* (genus) and *Eggerthella* (genus).

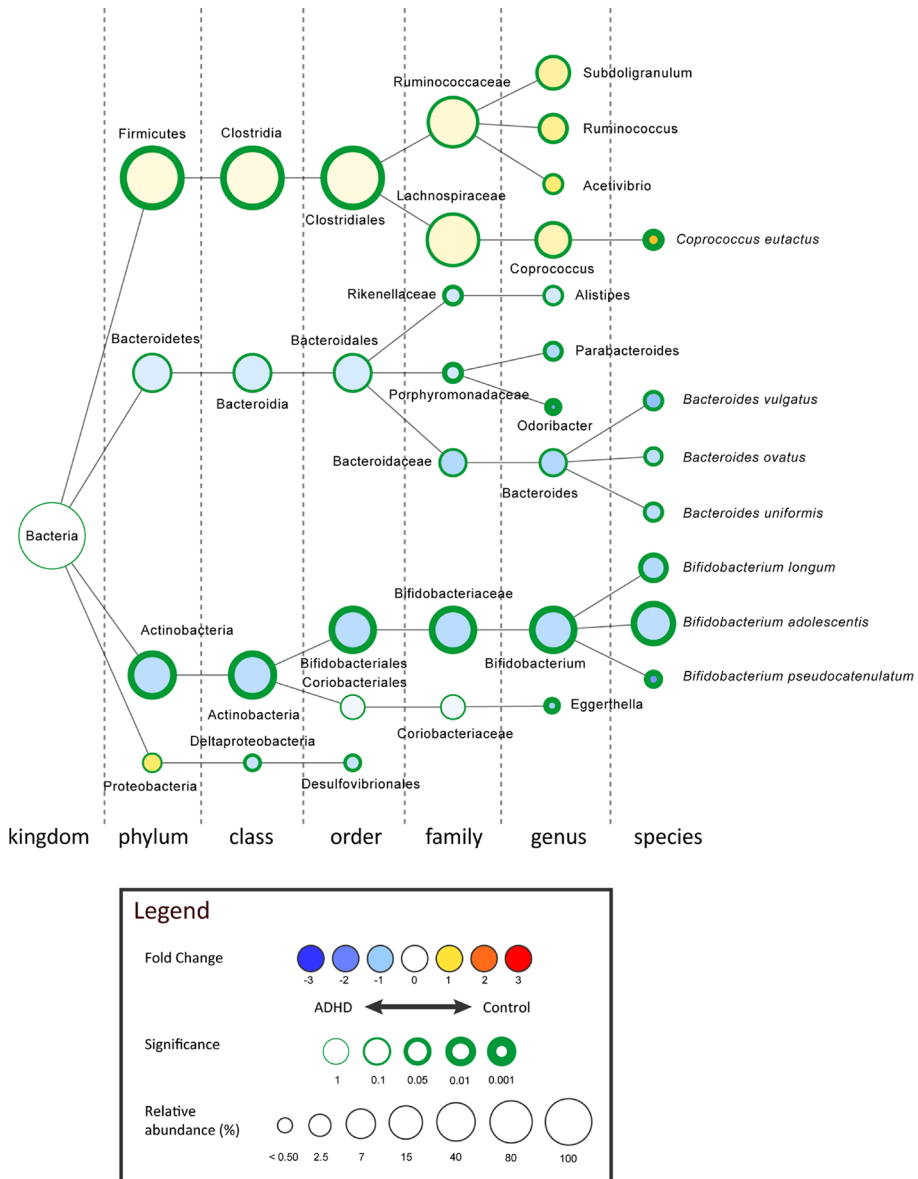


Figure 3. The strongest differentially abundant microbial taxa for ADHD cases (n=19) versus healthy controls (n=77), shown in the graphical Cytoscape visualization [136].

Nodes represent taxa (node size represents average relative abundance, for both experimental groups combined), edges (dashed lines) link the different taxonomic levels. The weighed fold-change (node color) is calculated as the $2\log$ of the ratio of the relative abundance between control and ADHD (0 = no difference between genotypes, 1 = twice as abundant in control, etcetera). In other words: yellow to red indicates an overrepresentation in control, hence an underrepresentation in ADHD, and vice versa for light- to dark blue. The significance (node border width) is expressed as the p-value of a Mann-Whitney U test, uncorrected for multiple comparisons.

Hypothesis-driven analysis of gut microbiome metabolic potential

We predicted bacterial gene function using PICRUSt, focusing on pathways involved in the synthesis of phenylalanine, tyrosine, and tryptophan, which can serve as precursors of human dopamine, noradrenaline and serotonin ([Fig. 1](#)).

Relative abundances of the 17 candidate reactions/enzymes (pathways) that are directly involved in the production of phenylalanine, tyrosine, or tryptophan ([Supplementary Fig. S2](#)) were predicted based on the total gut microbiome of healthy controls and ADHD cases. For these 17 *a priori* selected candidates, 15 K numbers could be identified in the microbiome ([Table S3](#)). One predicted enzyme, cyclohexadienyl dehydratase (CDT; KEGG Ortholog K01713; EC:4.2.1.51), was found to be significantly more abundant (on average, 150% more than in controls) in the microbiome of ADHD cases ($p = 0.038$ by MWU, Bonferroni-corrected for 15 K numbers identified) ([Fig. 4](#) and [Table S3](#)). Moreover, CDT ranked among the top ~1% of a total of 7545 reactions (when we sort our reaction p -values from low to high ([Table S3](#)). Similar results of ADHD cases versus controls were obtained with the sample groups matched for age (i.e. without the older BIG controls), though significance is lost ($p = 0.085$, Bonferroni-corrected). Using logistic regression to control for age (and gender; [Supplementary Data S1C](#)) the results in the age-matched sample (i.e. without the BIG controls) hint to the fact that the change in ADHD for K01713 is a significant effect ($p = 0.024$), but the marginal effect in the complete sample ($p = 0.070$) shows that our results cannot be completely disconnected from the age confounder present in the cohorts studied ([Supplementary Data S1C](#)).

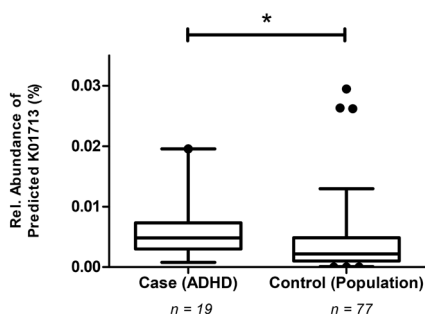


Figure 4. The ADHD microbiome contains significantly increased levels of predicted cyclohexadienyl dehydratase (CDT; KEGG Ortholog K01713; EC:4.2.1.51), responsible for phenylalanine synthesis.

This analysis is based on functional predictions deriving from 16S profiles of the microbiome, as performed by PICRUSt [64]. Box plots represent the relative abundance of predicted CDT, with 5-95% percentile whiskers (dots represent outliers). The significance was tested with a non-parametric MWU (* $p = 0.038$), Bonferroni-corrected for 15 K numbers identified. See also [Supplementary Fig. S2](#).

To assess which of the five taxa differing between cases and controls ([Fig. 3](#)) contributed most to the observed difference in the predicted enzyme CDT, we repeated the functional (PICRUSt) analysis for these candidate taxa only: Clostridiales (order), Rikenellaceae (family), Porphyromonadaceae (family), *Bifidobacterium* (genus) and *Eggerthella* (genus) ([Fig. 3](#)). Differences in relative abundance of the genus *Bifidobacterium* uniquely contributed to the observed differences in the predicted

phenylalanine pathway enzyme CDT in a multiple regression analysis ($p < 0.001$). The same conclusion was drawn from the fact that in the cohort *Bifidobacterium* contributed 99.9% of the predicted CDT (K01713) counts relative to the predicted CDT counts based on PICRUSt analysis of the entire microbiome ($n = 16,340$; [Supplementary Data S1D](#)).

fMRI: effects of reward anticipation and diagnosis

In a partly overlapping sample of the same cohort, we tried to replicate reduced reward anticipation in ADHD versus controls (see [Background](#)). Across the whole sample, reward anticipation (high (15 ct) > low (1 ct) cues) elicited brain responses in the striatum and the occipital, premotor, and frontal cortices ($p_{FWE} < 0.05$, whole-brain, cluster-level correction) in our total fMRI sample ($n = 87$) as well as the sub-sample with microbiome data ($n = 28$) ([Fig. 5A and Supplementary Data S2](#)). Taking the anatomically-defined ventral striatal ROI, we indeed found decreased ventral striatal responses for reward anticipation in patients with ADHD versus controls ($t(85)=2.1$, $p = 0.038$) ([Fig. 5B](#)). This difference was not significant in the sub-sample with microbiome data ($t(26)=0.2$).

fMRI: effects of the microbiome on reward anticipation

Finally, we assessed how the functional microbiome measure found to be different for ADHD relative to controls (i.e. predicted CDT), would relate to neural reward anticipation in the sub-sample with both microbiome and fMRI data. Across the whole sub-sample ($n = 28$), independent of diagnosis, we observed a negative association (whole-brain) of the relative abundance of predicted CDT with reward anticipation responses in bilateral ventral striatum ([Fig. 5C](#)). This effect was also significant in the ventral striatal ROI analysis (standardized beta: -0.48 , $p = 0.032$), when including possible confounding factors (ADHD diagnosis, age, and gender) in the analysis. Also, when adding stimulant use (duration times dose) to control for long-term medication effects, predicted CDT relative abundance was still significantly associated with reward anticipation responses in ventral striatum (standardized beta: -0.42 , $p = 0.048$).

Discussion

This is the first study investigating microbiome differences between ADHD cases and controls. Observed differences in taxa in this exploratory study were strongest for the phylum Actinobacteria, which was more abundant in cases, apparently at the expense of Firmicutes, for which abundance was lower in cases; uncorrected for multiple comparisons. In addition, we describe for the first time effects of a genetically encoded capacity for production of monoamine precursors in the gut microbiome of ADHD: the enzyme CDT, involved in the synthesis of a dopamine precursor (phenylalanine), was predicted to be significantly more abundant in the microbiome of ADHD cases compared to healthy individuals (corrected for multiple comparisons). This effect appeared dependent on age, although significant effects were obtained with age-matched samples. The genus *Bifidobacterium* (within the phylum Actinobacteria) – also more abundant in (age-matched) ADHD versus controls (uncorrected for multiple comparisons) – appeared to be solely responsible for this predicted enzyme CDT increase. Across the sample (i.e. independent of diagnosis), the abundance of

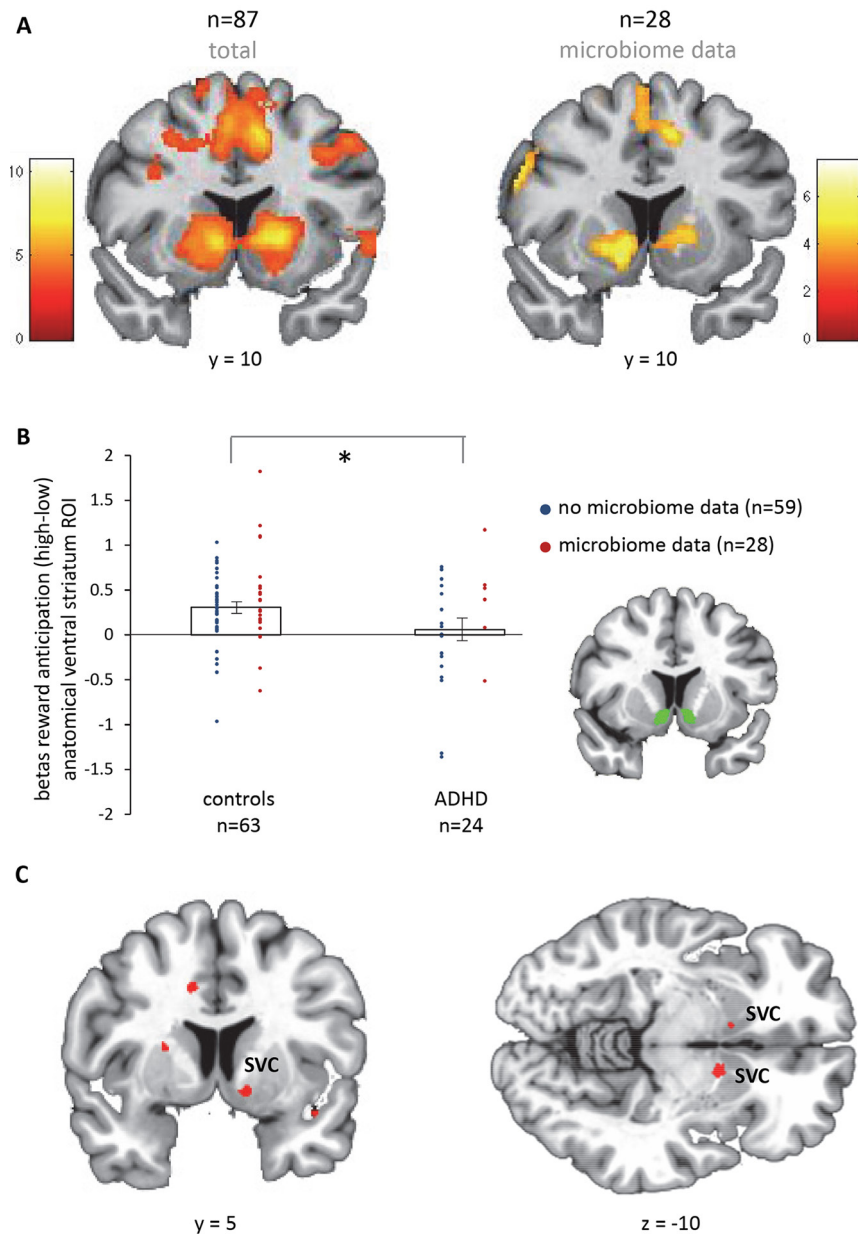


Figure 5. fMRI results.

(A) Main effect of reward anticipation, cluster-level corrected at the whole-brain level ($p_{FWE} < 0.05$). Color bars reflect T-values. (B) Diagnosis effects in the anatomical region of interest (ROI) of the ventral striatum. (C) Negative correlation of the microbiome function CDT (see Fig. 4) with reward anticipation responses across the whole-brain ($n = 28$), intensity threshold at $p < 0.001$ uncorrected ($T = 3.45$). The clusters in bilateral ventral striatum ($x=-11, y=11, z=-9$, cluster size = 8, $p_{FWE, cluster} = 0.024$; $x=11, y=6, z=-11$, cluster size = 2, $p_{FWE, cluster} = 0.036$) are significant after correcting for multiple comparisons across the search volume (cluster-level $p_{FWE} < 0.05$, SVC), i.e. the anatomically defined ventral striatum shown in panel B. SVC = small volume correction. * indicates $p < 0.05$.

predicted CDT genes from the microbiome correlated negatively with bilateral ventral striatal blood oxygenation level-dependent (BOLD) responses for reward anticipation (corrected for multiple comparisons in an anatomically-defined search volume), typically reduced in ADHD [359–362] and replicated here.

Pärtty and colleagues [374] recently used markers for specific bacterial species to show that decreased numbers of *Bifidobacterium* (e.g. *Bifidobacterium longum*) in 3 and 6 months old children (treated with either a *Lactobacillus*-based probiotic or a placebo) predicted ADHD or Asperger syndrome manifestation at age 13 years [374]. At the time of diagnosis (13 years old), they did not observe any significant difference in the assessed taxa, including *Bifidobacterium*, between ADHD and controls. This difference with our marginally significant result of increases in *Bifidobacterium* in ADHD might be explained by (i) our more sensitive method (16S marker gene sequencing versus PCR), and (ii) larger sample size (19 ADHD and 77 controls versus 6 ADHD/Asperger and 69 controls) or (iii) differences introduced by experimental procedures or between cohorts. The observed differences in *Bifidobacterium* genus between controls and adolescents/adults with ADHD (current study) or infants developing ADHD later in life [374] – in opposite direction – require formal replication in longitudinal studies with larger samples. Decreased abundance of *Bifidobacterium* in early infancy versus increased abundance in early adulthood might reflect delayed gut microbiome maturation in ADHD, as *Bifidobacterium* is known to decline with older age [376]. Pärtty and colleagues [374] found that *Bifidobacterium* in infancy predicted ADHD and Asperger syndrome. Changes in microbiome composition have indeed consistently been found in autism spectrum disorder [377]. Although participants with DSM-IV defined Autistic Disorder and other neuropsychiatric disorders were excluded, we cannot be certain that the current findings are specific for ADHD.

Bifidobacterium was responsible for increases in the predicted function of CDT (K01713) – involved in the synthesis of phenylalanine ([Supplementary Fig. S2](#)) – in ADHD cases versus controls. Phenylalanine is an essential amino acid, which cannot be synthesized by humans and has to be absorbed from the gut [378]. Phenylalanine can cross the blood-brain barrier and is the precursor of dopamine and noradrenaline [375] ([Fig. 1](#)). These neurotransmitters are highly affected in ADHD, nevertheless, the exact mechanisms of involvement are still ambiguous [379–381]. Predicted CDT correlated negatively with ventral striatal responses during reward anticipation. Striatal reward anticipation responses are modulated by dopamine and remediated by methylphenidate [358, 382]. Our findings suggest that gut microbial-induced levels of CDT correlate functionally with available levels of the dopamine precursor phenylalanine, which could potentially be a risk factor for disturbed dopamine signaling and reduced brain reward responses. Indeed, high levels of phenylalanine have been linked to ADHD symptoms [383]. However, this association was found in phenylketonuria, a disorder in which phenylalanine cannot be converted to tyrosine, which results in toxic build-up of phenylalanine in the brain. Two other studies, using an overlapping sample, have found decreased blood plasma concentrations of phenylalanine in ADHD [384, 385] and a more recent and larger study did not observe an association between peripheral (i.e. blood and urine) levels of phenylalanine and ADHD [386]. Future studies should link

the presently observed increases in predicted microbiome-derived phenylalanine to peripheral levels of phenylalanine and study how this relates to brain function.

Moreover, the mechanism by which the increased predicted CDT function (and speculatively, increased phenylalanine) would result in decreased striatal BOLD responses for reward anticipation remains to be investigated. Many different routes for a potential causal relationship are possible. Microbial phenylalanine could be absorbed in the blood stream, cross the blood-brain barrier, and influence dopamine synthesis (Fig. 1) positively or negatively (by inhibiting tyrosine hydroxylase) [387]. Alternatively, on the host side, altered blood plasma levels of phenylalanine (or its derivative tyrosine) could have an effect on the synthesis of neuromodulators other than dopamine, e.g. by competing at the blood-brain barrier with tryptophan (precursor of serotonin) or by conversion to trace amines [375, 388, 389]. In addition, yet unknown interactions of multiple bacterial groups involved in the production of neuroactive substances might affect host neurophysiology within the gastrointestinal tract [370]. Nevertheless, our observed relationship between the microbial phenylalanine pathway and neural responses for reward anticipation, known to depend on phenylalanine's derivative dopamine, may argue for a dopaminergic effect at the level of the brain instead of the gut. Future research should confirm this and show specificity for microbial functions, as well as brain regions and associated functions.

Our study should be viewed in the context of its strengths and limitations. Obvious strengths include the 16S microbiome analysis (instead of preselected candidate taxa), larger sample size than previous studies in ADHD [374], and our mechanistic, hypothesis-driven approach in terms of predicted enzymes (function analysis) as well as their link to brain functioning using fMRI. As for limitations, a microbiome shotgun sequencing method preferably combined with a proteomics or metabolomics approach instead of the 16S marker gene method might have allowed us to make more solid claims about microbial biological function. Second, about 25% of our control participants were siblings of ADHD cases (see Table 1), and another sub-sample of the control group did not undergo clinical screening for ADHD (the BIG sample). However, in both cases this would be more likely to cause an underestimation of differences between ADHD cases and controls. Third, the control group was significantly older than the ADHD cases. Consequently, the age confounder in our study hampered us from attributing the significant change in predicted CDT exclusively to ADHD. Nevertheless, our study suggests that the change in ADHD for the CDT enzyme (K01713) is a significant effect, with subject age having an important contribution. This might in part be explained by the explicit nature of ADHD for which it is recognized that its prevalence diminishes with age [390]. Importantly, the gut-microbiome-brain association between predicted CDT and reward anticipation fMRI responses was observed independent of age. Fourth, it is well known that diet affects microbiome structure/stability [368]. In this respect, having a substantial part of the control group consisting of unaffected siblings of ADHD cases (21/77), presumably living in the same household with similar diets, should be viewed as a strength of this study. Importantly, we did not observe any differences in body mass index between the groups. Fifth, our gut-microbiome-brain association was found independent of, or actually controlled for, ADHD diagnosis. Hence, we can

only conclude that functional differences found between ADHD and controls at the microbiome level are related to neural effects *across subjects*. ADHD status-specific conclusions about this gut-brain relation might be made in future studies with larger group sizes. Finally, given the generally observed beneficial effects of *Bifidobacterium* [391, 392], it remains to be resolved in subsequent studies whether our observed effects are a consequence or, perhaps, a compensatory, effect of ADHD status. Considering these uncertainties, our novel functional gut-brain approach provides many leads for new research, but caution should be taken to translate these findings to non-pharmacological intervention strategies in ADHD.

Conclusions

This is the first study demonstrating differences in the gut microbiome between patients with ADHD and healthy individuals, using a comprehensive 16S microbiome analysis, and showing – if anything – an increase in the genus *Bifidobacterium*. This increase was associated with significantly enhanced predicted biosynthesis potential of a dopamine precursor in the gut microbiome of ADHD patients versus controls, which was linked to altered reward anticipation responses in the brain, a neural hallmark of ADHD. With this mechanistic approach, we hypothesize that presumed differences in dopamine precursor production at the gut microbiome level in ADHD might be related to dopamine disturbances at the neural level associated with reduced brain reward responses. This study highlights the importance of investigating the functional effects of microbiome differences in neuropsychiatric disorders.

Methods

Participants

Microbiome Sample

For the microbiome analyses, we included 96 participants, of whom 19 had been diagnosed with ADHD and 77 were healthy ([Fig. 2, Table 1 and Supplementary Methods S1](#)). ADHD cases were derived from the follow-up of the NeuroIMAGE study [393] (NeuroIMAGE II) and were diagnosed based on DSM-IV symptoms using the Schedule for Affective Disorders and Schizophrenia for School-Age Children [394]. The sample of healthy individuals was compiled of two sub-samples: (i) healthy participants ($n = 17$) and unaffected siblings of ADHD probands ($n = 21$) of the ADHD cohort (NeuroIMAGE II project; not necessarily the siblings of the cases in the current sample) and (ii) self-reported healthy volunteers ($n = 39$) of the Brain Imaging Genetics (BIG) study [395].

fMRI sample

For the fMRI analyses, we included 87 participants from the above-mentioned ADHD cohort (NeuroIMAGE II project), of whom 24 had ADHD and 63 were unaffected ([Fig. 2, Table 1 and Supplementary Methods S1A](#)). More participants were unaffected than affected as this was a follow-up study and part of the children with ADHD no longer met the diagnostic criteria in adolescence or adulthood of the current study. Of the 63 controls, 39 participants were unaffected siblings of ADHD probands, and 24

participants were healthy controls.

Out of initial 95 participants who underwent fMRI, eight participants were excluded from analyses: one control due to significant drop-out in the back of the brain, two participants (one ADHD and one control) due to less than 10 correct trials per cell in the factorial design, five participants due to moving more than 5mm (translation; three ADHD and two controls).

For 28 included participants from the fMRI sample, group microbiome data was available: 6 patients with ADHD and 22 controls overlapped between both sample groups ([Fig. 2](#) and [Table 1](#)). Of the overlapping 22 controls, 13 participants were unaffected siblings, and 9 participants were healthy controls. The regression analyses between the microbiome and the fMRI measure in this sub-sample were performed across diagnosis ($n = 28$); see below.

The investigation was carried out in accordance with the latest version of the Declaration of Helsinki. After complete description of the study to the participants, written informed consent was obtained (and from their parents when <18 years old). The study was approved by the regional medical ethics committee (Commissie Mensgebonden Onderzoek: CMO Regio Arnhem Nijmegen, number: NL41950.091.12).

fMRI analysis

Reward anticipation task

Reward anticipation was assessed during fMRI in the context of a rewarded Stroop task ([Supplementary Fig. S1](#)), adapted from a previous study [396], focusing our analyses on reward cues (high > low reward cues). Neural responses to high versus low reward cues in this task reflect motivation for monetary incentives, known to be dependent on dopamine signaling in ventral striatum [358]. Using similar tasks, reward anticipation responses were found to be reduced in patients with ADHD versus controls, and appear negatively correlated with hyperactive-impulsive symptoms in ADHD [for a review, see 361].

MRI data acquisition, preprocessing, and analyses

Whole-brain functional images (multi-echo) and a high-resolution anatomical scan were acquired on a 1.5T MR scanner ([Supplementary Methods S2A](#)). All data were pre-processed and analyzed with SPM8 (<http://www.fil.ion.ucl.ac.uk/spm/>). Echo combination, realignment, slice timing, co-registration, normalization, and spatial smoothing are described in [Supplementary Methods S2B](#).

For each participant, the resulting pre-processed fMRI time-series were analyzed at the first level using an event-related approach in the context of a general linear model, including 24 motion parameters as regressors of non-interest ([Supplementary Methods S2C](#)). We performed one-sample t-tests to assess the main effects of reward anticipation (high (15 ct) > low (1 ct) reward cues) in the total sample ($n = 87$) as well as in the smaller sub-sample with microbiome information ($n = 28$). To further account for motion, we added a summary motion score for every subject in all second level analyses

as covariate of non-interest ([Supplementary Methods S2C](#)). Statistical inference ($p < 0.05$) was performed at the cluster level, correcting for multiple comparisons (Family Wise Error, FWE) over the search volume, i.e. the whole brain. We investigated the effects of diagnosis using a ventral striatum region of interest (ROI), i.e. an anatomically-defined bilateral nucleus accumbens region ([Supplementary Methods S2D](#)).

Microbiome analysis

Microbial faecal DNA extraction

Feecal samples were collected using a standard method consisting in scooping a pea-sized piece of feces and storing it in a 50ml Falcon tube. The sample was then stored at 4°C straight after collection and at -80°C within 24 hours. Faecal genomic DNA from self-collected stool samples was isolated using the DNeasy® Blood and Tissue Kit (Qiagen, Venlo, The Netherlands) as described earlier [397]. The DNA was treated with RNase and eluted in Qiagen elution buffer AE. DNA purity and quantity were checked by spectrophotometry (ND-1000, NanoDrop Technologies, Wilmington, DE, USA).

16S marker gene amplification, sequencing and data acquisition

Preparation of the amplicon pool for pyrosequencing followed established protocols [398], and is described in detail in [Supplementary Methods S3A](#), where additionally sequencing and data acquisition is outlined. In short, hypervariable V3-V4 region of the 16S rRNA gene was sequenced on the 454 Life Sciences GS-FLX platform using Titanium sequencing chemistry (GATC-Biotech, Germany).

Microbiome sequencing data analysis

For gene sequencing analysis, a customized Python workflow based on Quantitative Insights Into Microbial Ecology (QIIME version 1.2) was adopted (<http://qiime.org>) [42] ([Supplementary Methods S3B](#)).

Microbiome-derived function prediction

Based on the generated 16S profiles of the gut microbiome for ADHD and control subjects for the full microbiome cohort, we predicted the presence of Kyoto Encyclopedia of Genes and Genomes (KEGG) Orthologs and subsequent functional and metabolic pathways using PICRUSt ([Supplementary Methods S3C](#)). Based on the KEGG (<http://www.genome.jp/kegg/>) pathway maps of 'Phenylalanine, tyrosine and tryptophan biosynthesis' (k000400, k000360, k000350 and k000380), the 17 candidate reactions/enzymes that directly result in production of phenylalanine, tyrosine, or tryptophan were selected: EC:1.3.1.43; EC:1.3.1.78; EC:1.3.1.79; EC:1.4.1.20; EC:1.4.3.2; EC:1.14.16.1; EC:2.6.1.1; EC:2.6.1.5; EC:2.6.1.57; EC:2.6.1.58; EC:2.6.1.9; EC:4.1.99.1; EC:4.1.99.2; EC:4.2.1.20; EC:4.2.1.51; EC:4.2.1.91 and EC:5.1.1.11 [399] ([Supplementary Fig. S2](#)). From the relative abundances of gene functions predicted by PICRUSt we focused on candidate reactions/enzymes, by applying in-house bioinformatics (Perl-coded) scripts. Relative abundances of reactions/enzymes were calculated ([Supplementary Methods S3C](#)) for each individual sample based on all observed KEGG Orthologs, for each hierarchical functional KEGG level. To determine the contribution of the candidate microbial taxa, i.e. those (marginally) differing between ADHD and

controls, to differences observed in the selected monoamine precursors, the PICRUST analysis was repeated on the subset of candidate taxa only to find the taxa responsible for the observed functional effects.

Microbiome data analysis and statistics

Statistics on the relative abundances of selected monoamine biosynthetic pathways (K numbers: KEGG orthology groups) or taxa between sample groups was performed with SciPy (<http://www.scipy.org>) using a non-parametric Mann-Whitney *U* (MWU) test with Bonferroni correction for (i) multiple comparisons across all microbial taxa in all levels of phylogenetic classification (in analyses of phylogenetic composition), or (ii) across all pathways of the same KEGG hierarchical pathway level (in analyses of predicted functionality), unless stated otherwise. Relative abundance of taxa was correlated with relative abundance of enzymes/reactions (i.e. K numbers) using Spearman correlation. Subsequently, we repeated this analysis with multiple regression (in SPSS, see above) to assess the unique contribution of taxa to the predicted reaction of interest, using the 16S-based taxa as predictors and the enzymes/reactions predicted to be present in those taxa (by PICRUST), as dependent variables. For any additional downstream sequencing-related data analysis, figures, and statistics, Microsoft® Office Excel® 2007 and GraphPad Prism version 5.03 were used.

Data availability

The raw, unprocessed 16S 454-sequencing reads are publicly available for download at the European Nucleotide Archive (ENA) database (<http://www.ebi.ac.uk/ena>) under study accession number PRJEB11512 (or secondary accession number ERP012909) [298]. The sequencing data is available in FASTQ-format, including corresponding metadata for each sample.

Effects of microbiome on reward anticipation (fMRI)

In final analyses, we tested how microbiome function related to brain function. Across the whole sub-sample with both microbiome and fMRI measures available ($n = 28$), we assessed the effects of relative abundance of a candidate predicted microbiome-derived enzyme/reaction as covariate of interest on whole-brain reward anticipation responses (in SPM) with the reward anticipation (high > low) images ($p_{FWE} < 0.05$, cluster-level, small volume: bilateral nucleus accumbens region from the Hammersmith atlas ([Supplementary Methods S2](#)). We also performed multiple linear regression with the ventral striatum ROI betas for reward anticipation (see above) as dependent variable and the microbiome measure as well as ADHD diagnosis, age, gender, and stimulant medication use as predictors ([Supplementary Methods S2](#)).

Supplementary Methods

Methods S1: Cohorts. (PDF)

([online](#)) Methods S1 of online repository at :

<http://ederveen.science/Thesis/Chapter7/Supplementary-Methods.PRESUBMISSION.pdf>

Methods S2: fMRI parameters and analyses. (PDF)

([online](#)) Methods S2 of online repository at :

<http://ederveen.science/Thesis/Chapter7/Supplementary-Methods.PRESUBMISSION.pdf>

Methods S3: Microbiome sequencing and analyses. (PDF)

([online](#)) Methods S3 of online repository at :

<http://ederveen.science/Thesis/Chapter7/Supplementary-Methods.PRESUBMISSION.pdf>

Supplementary Data

Data S1: Additional information microbiome. (PDF)

([online](#)) Data S1 of online repository at :

<http://ederveen.science/Thesis/Chapter7/Supplementary-Data.PRESUBMISSION.pdf>

Data S2: fMRI main effects of reward anticipation. (PDF)

([online](#)) Data S2 of online repository at :

<http://ederveen.science/Thesis/Chapter7/Supplementary-Data.PRESUBMISSION.pdf>

Supplementary Figures

Figure S1. Reward anticipation task. (PDF)

(online) Figure S1 of online repository at :

<http://ederveen.science/Thesis/Chapter7/Supplementary-Figures.PRESUBMISSION.pdf>

Figure S2. Monoamine precursor biosynthesis pathways. (PDF)

(online) Figure S2 of online repository at :

<http://ederveen.science/Thesis/Chapter7/Supplementary-Figures.PRESUBMISSION.pdf>

Figure S3. Intestinal microbiome composition. (PDF)

(online) Figure S3 of online repository at :

<http://ederveen.science/Thesis/Chapter7/Supplementary-Figures.PRESUBMISSION.pdf>

Supplementary Tables

Table S1. Microbiome descriptives. (EXCEL)

(online) Table S1 of online repository at :

<https://doi.org/10.1371/journal.pone.0183509.s001>

Table S2. Relative abundance taxa. (EXCEL)

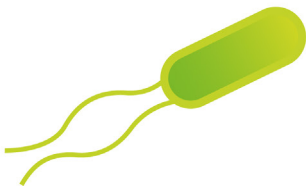
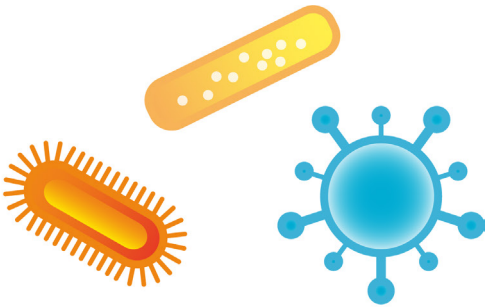
(online) Table S2 of online repository at :

<https://doi.org/10.1371/journal.pone.0183509.s002>

Table S3. Relative abundance pathways. (EXCEL)

(online) Table S3 of online repository at :

<https://doi.org/10.1371/journal.pone.0183509.s003>



PART IV

The Nasal Microbiome

CHAPTER 8

OPEN ACCESS

as published in *Microbiome*, 2018, Jan 11;6(1):10

<https://doi.org/10.1186/s40168-017-0395-y>

¹ Center for Molecular and Biomolecular Informatics, Radboud Institute for Molecular Life Sciences (RIMLS), Radboud University Medical Center (Radboudumc), Nijmegen, The Netherlands.

² Laboratory of Pediatric Infectious Diseases, Radboud Center for Infectious Diseases, Radboudumc, Nijmegen, The Netherlands.

³ Centre for Infectious Disease Control, National Institute for Public Health and the Environment, Bilthoven, The Netherlands.

⁴ NIZO, Ede, The Netherlands.

These authors contributed equally to this work.

¶ These authors also contributed equally to this work.

CHAPTER 8

***HAEMOPHILUS* IS OVERREPRESENTED IN THE NASOPHARYNX OF INFANTS HOSPITALIZED WITH RSV INFECTION AND ASSOCIATED WITH INCREASED VIRAL LOAD AND ENHANCED MUCOSAL CXCL8 RESPONSES**

Thomas H.A. Ederveen ^{1 #}

Gerben Ferwerda ^{2 #}

Inge M. Ahout ²

Marloes Vissers ³

Ronald de Groot ²

Jos Boekhorst ^{1,4}

Harro M. Timmerman ⁴

Martijn A. Huynen ¹

Sacha A.F.T. van Hijum ^{1,4 ¶}

Marien I. de Jonge ^{2 ¶}

ABSTRACT

BACKGROUND. While almost all infants are infected with respiratory syncytial virus (RSV) before the age of two years, only a small percentage develops severe disease. Previous studies suggest that the nasopharyngeal microbiome affects disease development. We therefore studied the effect of the nasopharyngeal microbiome on viral load and mucosal cytokine responses, two important factors influencing the pathophysiology of RSV disease. To determine the relation between (i) the microbiome of the upper respiratory tract, (ii) viral load and (iii) host mucosal inflammation during an RSV infection, nasopharyngeal microbiota profiles of RSV infected infants (<6 months) with different levels of disease severity and age-matched healthy controls were determined by 16S rRNA marker gene sequencing. The viral load was measured using qPCR. Nasopharyngeal CCL5, CXCL10, MMP9, IL6 and CXCL8 levels were determined with ELISA.

RESULTS. Viral load in nasopharyngeal aspirates of patients associates significantly to total nasopharyngeal microbiota composition. Healthy infants (n = 21) and RSV patients (n = 54) display very distinct microbial patterns, primarily characterized by a loss in commensals like *Veillonella* and overrepresentation of opportunistic organisms like *Haemophilus* and *Achromobacter* in RSV infected individuals. Furthermore, nasopharyngeal microbiota profiles are significantly different based on CXCL8 levels. CXCL8 is a chemokine that was previously found to be indicative for disease severity and for which we find *Haemophilus* abundance as the strongest predictor for CXCL8 levels.

CONCLUSIONS. The nasopharyngeal microbiota in young infants with RSV infection is marked by an overrepresentation of the genus *Haemophilus*. We present that this bacterium is associated with viral load and mucosal CXCL8 responses, both which are involved in RSV disease pathogenesis.

Background

Respiratory syncytial virus (RSV) is a major cause of respiratory tract infections in young children that frequently leads to hospitalization [400]. The clinical symptoms vary from upper respiratory tract infection, to severe bronchiolitis with respiratory insufficiency for which mechanical ventilation is needed. Despite the high disease burden of RSV in young children, no antivirals, vaccines or other targeted treatments for bronchiolitis have proven to be beneficial yet, and therefore only supportive care is currently recommended [400-402]. Prematurity, young age (<6 months) and the presence of siblings or daycare attendance are important risk factors for severe RSV disease requiring hospitalization and respiratory support, indicating that development of the immune system as well as environmental factors play a role in the pathogenesis of RSV bronchiolitis [403-405].

To date, it is thought that the host response to the virus contributes most prominently to the key features of bronchiolitis, marked by swelling of the mucosa, secretion of mucus and eventually obstruction of the smaller airways [405-408]. In addition, high viral load in the lungs and nasopharynx have been found in children with severe infection, suggesting that viral replication also contributes to increased pathology [409-411]. During the inflammatory response induced by the virus, cytokines and chemokines are released and attract circulating leukocytes to the site of infection. As recently reviewed by Russel *et al*, CXCL-8 and CXL-10 have been associated with immune pathology, whereas CCL-5 in general is associated with a more beneficial immune response [406]. It seems that a balanced inflammatory response with influx of immune cells is crucial, as indicated by near absence of cytotoxic CD8 cells in fatal cases of RSV bronchiolitis, and the presence of massive influx of neutrophils in the lumen of the airways of most severe infections [412, 413]. The importance of the inflammatory response in the development of severe disease is further underlined by the association of the level of circulating immune cells and secreted cytokines with disease severity [414].

Respiratory mucosal surfaces are colonized directly after birth with bacteria from the mother, and the composition changes dramatically during the first years of life depending on genetics and environmental factors [98, 415, 416]. These commensal bacteria release components such as lipopolysaccharides and peptidoglycans that have been shown to pass the epithelial barrier of the mucosal lining under non-inflammatory conditions [417]. This translocation of bacterial products increases during inflammation [418, 419]. Induction of the initial innate immune response in epithelial cells and immune cells by RSV and the secretion of cytokines and chemokines should therefore be considered in the presence of these bacterial products [420].

Findings from the MARC-35 bronchiolitis cohort by Mansbach and coworkers, a prospective study of 1016 infants below one year of age hospitalized with bronchiolitis, underpin the importance of bacterial colonization of the respiratory tract during viral infection [421]. They show that depending on infection with RSV and rhinovirus, different compositions of nasopharyngeal microbiota are found, mainly higher levels of *Streptococcus* and *Haemophilus/Moraxella*, respectively. Steenhuijsen Piters and co-

workers recently showed that clusters characterized by the dominance of *Haemophilus (influenzae)* and *Streptococcus* were positively associated with RSV infection and RSV related hospitalization, based on nasopharyngeal microbiota clusters that were stratified prior to analyses [422]. In addition, they reported enhanced expression of host genes linked to inflammation and immune signaling in patients with the two microbiota clusters (*Haemophilus* and *Streptococcus* dominated), suggesting that interactions between RSV and specific components of the nasopharyngeal microbiota modulate the host immune response, potentially driving clinical disease severity. These findings are further supported by the association of specific metabolic profiles in respiratory samples with severe bronchiolitis, which could be linked to the composition of the microbiome [423]. In this study, we investigate the association of the nasopharyngeal microbiota, without prior stratification on microbiota types, to RSV load, nasopharyngeal cytokine responses and the association with disease severity.

Results

RSV study cohort and baseline characteristics.

We retrospectively selected a cohort of 54 infant patients younger than six months of age, which were hospitalized with an RSV infection, and 21 age-matched healthy infants (Fig. 1A, Table 1, and Supplementary Table S1A). Patients were stratified based on severity of RSV disease using clinically defined parameters as follows: mild disease included children without hypoxemia (n = 9); moderate disease included children receiving supplemental oxygen (n = 27); severe disease included children requiring mechanical ventilation (n = 18) (Fig. 1B). We collected nasopharyngeal aspirate (NPA) samples from a subset of RSV patients (n = 25) 4-6 weeks after hospital discharge, enabling us to evaluate mucosal immune responses and the microbiome after recovery of disease (n = 2, 16 and 7, for mild, moderate and severe, respectively). NPA samples were collected in which bacterial composition, viral load and host immune responses were measured to allow for integrated analysis (Fig. 1C).

To account for potential confounding variables, we screened our cohort on various parameters such as gender, age, (birth) weight and RSV viral load (Table 1, and Supplementary Fig. S1). Infant age (and gender to a minor extent) were found to be unevenly distributed over our cohort, which is likely a consequence of age being a known risk factor for developing (severe) RSV infection [402]. Birth weight, although not significantly different between study groups, did show a trend towards being increased in RSV patients in comparison to healthy infants. Antibiotics use was evenly distributed over the RSV patient severity stratifications (Table 1), none of the healthy infants used antibiotics at time of enrollment. Altogether, this prompted us to correct for age, gender and birth weight effects in multivariate analyses throughout this study. We also assessed viral co-infections, for which we found that only coronavirus (11%) and rhinovirus (20%) had a high prevalence. However, patients with these co-infections appeared to be randomly distributed in the disease severity stratifications

Table 1. Descriptive characteristics of the study samples (n = 75).

Data are presented as average + standard deviation (SD) for continuous variables, and as percentages for categorical variables. Numbers listed in brackets: SD. The antibiotics use was at the time of infection, and breastfeeding duration was not specified. # note that due to technical limitations in the resolution of 16S marker gene sequencing, OTU (Operational Taxonomic Unit) calling on the level of species should be interpreted with caution. † alpha diversity metrics.

No. participants (n = 75)	Control n = 21	Mild n = 9	Moderate n = 27	Severe n = 18	Recovery n = 25
Age (days)	89 (37)	88 (29)	75 (50)	37 (21)	109 (48)
Hospital Stay (days)	0 (0)	3 (3)	8 (5)	14 (6)	0 (0)
RSV viral load (Ct)	50.0 (0)	22.4 (3.3)	27.0 (5.4)	23.6 (3.4)	50.0 (0)
Weight (gram)	N/A	14674 (21111)	5516 (2002)	3874 (931)	5375 (2333)
Birth weight (gram)	2624 (768)	3323 (618)	3271 (656)	2867 (738)	3116 (677)
% Antibiotics	0%	22%	7%	11%	8%
% Breastfed	38%	67%	56%	72%	76%
% Males	24%	56%	52%	44%	44%
Mean Number of Reads	27449 (4464)	24916 (4826)	21787 (5776)	28866 (32974)	22658 (5208)
Total Number of Reads	5.76E+05	2.24E+05	5.88E+05	5.20E+05	5.66E+05
Assigned at Family	99.72% (0.01%)	99.88% (0.01%)	99.90% (0.01%)	99.94% (0.01%)	99.91% (0.01%)
Assigned at Genus	94.72% (0.07%)	97.91% (0.03%)	98.98% (0.02%)	98.15% (0.03%)	97.89% (0.04%)
Assigned at Species #	21.22% (0.19%)	8.07% (0.06%)	10.25% (0.09%)	10.27% (0.09%)	8.43% (0.17%)
Mean Number of OTUs	203 (56)	167 (44)	139 (38)	150 (38)	160 (50)
Total Number of OTUs	4270	1499	3756	2705	3999
Mean Chao index †	176.8 (54.4)	151.7 (42.6)	131.1 (39.4)	137.5 (17.8)	150.2 (53.5)
Mean Shannon index †	3.1 (1.1)	2.8 (0.7)	2.4 (0.8)	2.7 (0.7)	2.4 (0.8)
Mean PDWT index †	6.0 (1.7)	5.6 (1.1)	5.1 (1.6)	4.9 (0.9)	6.1 (2.0)

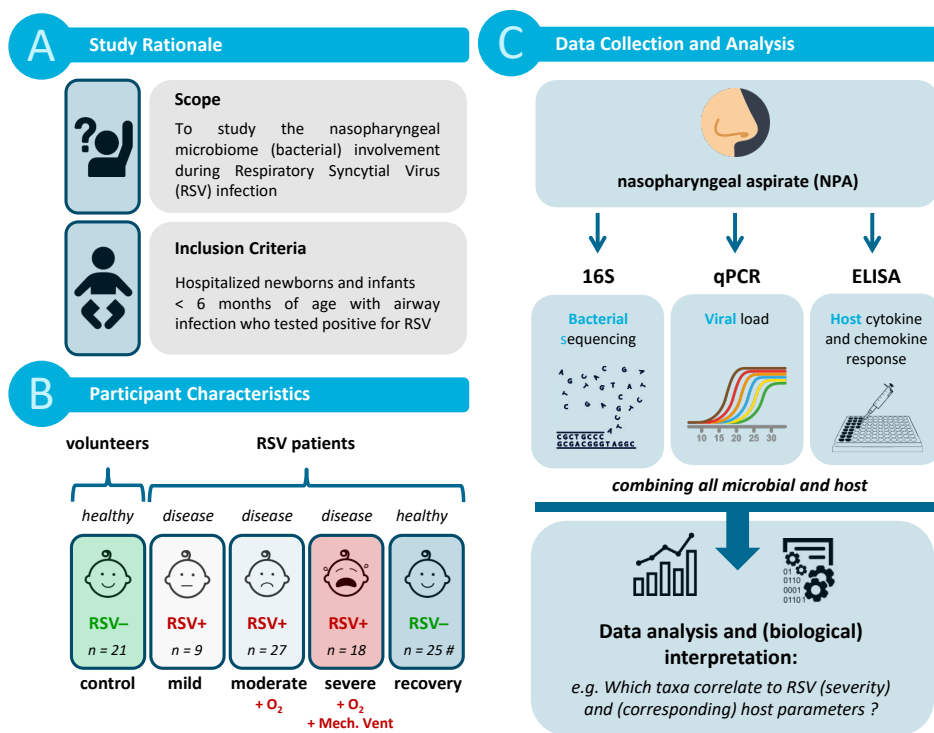


Figure 1. Study design.

Graphical summary of the study focus, design and analysis, as adopted in this manuscript.

Nasopharyngeal microbiota composition.

16S rRNA marker gene Illumina sequencing (16S) of the V3-V4 region was used to study the consortia of bacteria present in the NPA samples. By adopting a customized QIIME-based workflow for analysis of sequencing reads, we were able to classify 97.6% of sequencing reads with confidence to the genus-level, and study-wide identified a total of 156 unique genera ([Supplementary Tables S2 and S3](#)). Sufficient sequencing reads and OTU (Operational Taxonomic Unit) counts for each sample were realized with an average of $25k \pm 15k$ SD reads and 162 ± 51 SD OTUs, respectively ([Table 1, and Supplementary Table S2](#)). We identified *Haemophilus* (30.5% relative abundance on average; healthy and RSV-infected individuals combined), *Streptococcus* (29.4%), *Moraxella* (10.7%), *Corynebacterium* (6.89%) and *Staphylococcus* (2.9%) as the main genera (and known to be typically present) in the nasopharynx ([Fig. 2](#)) [422]. In addition, we found minor levels of *Prevotella* (4.2%), *Achromobacter* (3.6%), *Neisseria* (2.2%) and *Veillonella* (1.4%) representing additional genera ([Fig. 2](#)).

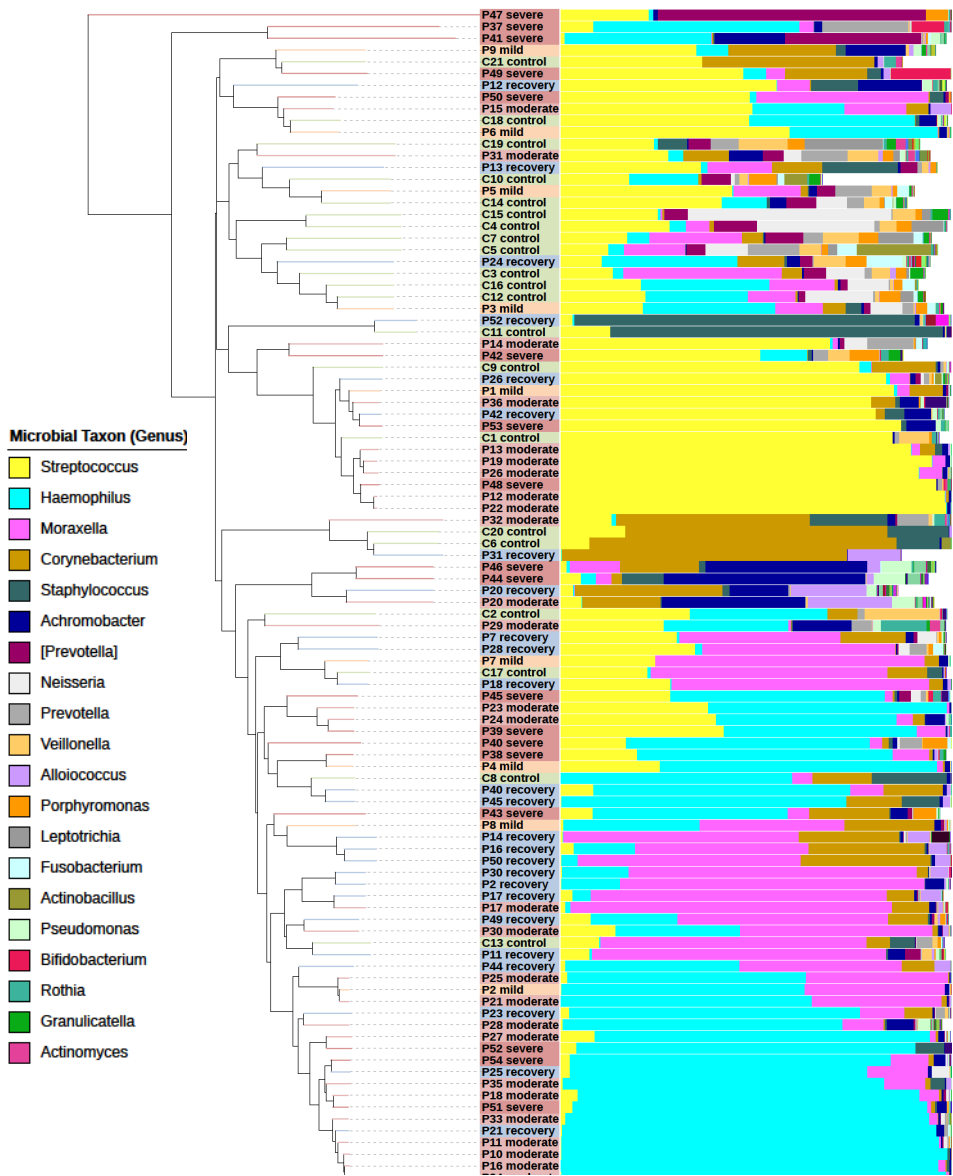


Figure 2. Nasopharyngeal microbiota composition in healthy and RSV-infected infants. Each leaf of the tree represents a single sample. Samples were clustered based on beta diversity ("between-sample distance"), using weighted UniFrac as a distance measure and hierarchical UPGMA as a clustering method. Vertical bars show the relative abundance microbiota composition on the genus level (reads that could not be classified up to this level are in white). The 20 most dominant genera are shown in the legend. Colored sample labels represent sample classes: mild (orange), moderate (light red) or severe (dark red) disease, healthy control (green) and recovery (blue) samples. The figure was generated with the interactive tree of life (iTOL) program.

RSV disease, viral load and CXCL8 levels can be explained by nasopharyngeal microbial make-up.

To examine the microbial involvement in RSV disease processes, redundancy analysis (RDA) was performed and showed a strong and significant separation of RSV patients and healthy controls based on NPA microbial make-up (**Fig. 3A**; $p = 0.001$). This difference in contrast is mainly characterized by *Haemophilus* and *Achromobacter*, not by *Agrobacterium*, as this bacterium is only represented by 3 RSV samples with $<0.01\%$ relative abundance each. Both viral load and host immune responses are hypothesized to play a major role in the disease processes, through which microbiota could potentially (directly or indirectly) assert their effect on the pathophysiology and the course of disease. Therefore, we further focused on differences in microbiota content between different viral loads and host cytokines and chemokines that were previously described to be relevant in RSV disease (CCL5, CXCL8/IL8, CXCL10, IL6 and MMP9) [414, 424, 425].

We found that RSV viral load (as determined with qPCR, Ct) in nasopharyngeal aspirates of patients was associated significantly with total nasopharyngeal microbiota composition (**Fig. 3B**; RDA $p = 0.036$), although no statistically different viral load

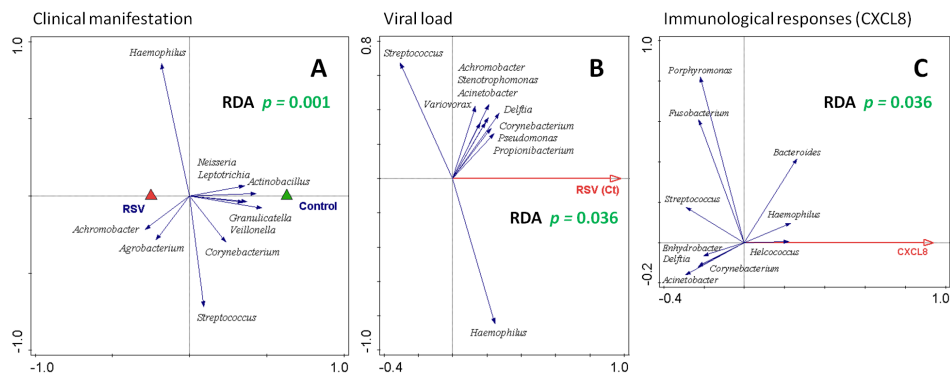


Figure 3. RSV disease, viral load and CXCL8 levels in healthy and RSV-infected individuals can be explained by nasopharyngeal microbial make-up (genus-level).

Redundancy analysis (RDA) biplots are shown. Nasopharyngeal genus-level microbiota from healthy and RSV-infected individuals are significantly different, irrespective of disease severity (**A**) (according to a permutation test; p -value = 0.001). Triangles are the centroids of the study sample groups: RSV (red) and healthy control (green). RDA of RSV-infected individuals (healthy and recovery samples were excluded from analysis) shows that nasopharyngeal genus-level microbiota can significantly be separated based on viral load (**B**) (Ct threshold plotted: higher Ct number corresponds to higher number of PCR cycles before confident virus detection; hence lower viral load; p -value = 0.036). For **A** and **B**: The blue arrows are the 10 best-fitting genera (names in italic), which are the genera best explaining microbiota compositional differences between disease status (**A**) or different levels of RSV virus (**B**) as plotted on the horizontal axis. RDA of healthy and RSV-infected individuals shows that nasopharyngeal genus-level microbiota can significantly be separated based on levels of CXCL8 (**C**) (p -value = 0.036; log transformation was set to 1000). The first component (horizontal axis) is optimized to explain CXCL8 level based on microbiota relative abundances (concentration of CXCL8 in pg/ μ l). Correspondingly, the blue arrows are the genera (names in italic) explaining at least 5% of this variation. RDA were corrected for age, gender and birth weight. See [Supplementary Fig. S2](#) for similar analysis on the OTU-level.

between the patient RSV severity stratifications could be detected ([Supplementary Fig. S1B](#)). However, because the response variables of the data in the RDA are also strongly oriented in the vertical direction, this indicates that separation based on genus-level microbiota is not primarily driven by the viral load contrast.

Furthermore, RDA of healthy and RSV-infected individuals showed that nasopharyngeal genus-level microbiota can significantly be separated based on levels of CXCL8 ([Fig. 3C](#); RDA $p = 0.036$) but not for any of the other candidate markers (data not shown). Strikingly, in this RDA, *Haemophilus* is among the most important genera contributing to the observed separation, as its arrow points to CXCL8 in a strong horizontal direction. *Helcococcus*, although its arrow is also strongly horizontally oriented, has a very low abundance (0.01% on average).

The above described genus-level RDA analyses ([Fig. 3](#)) were repeated with higher-resolution OTU-level microbiota data (clustering on 97.0%), and on this level of taxonomical detail, similar statistical significances and microbial associations were observed ([Supplementary Fig. S2](#)). Mainly, OTUs classified to genus or species of *Haemophilus* were found to be important in driving these outcomes, and *Streptococcus* to a lesser extent. However, we did lose significance for the association of CXCL8 with microbiota on the level of OTU in comparison to genus ($p = 0.036$ to $p = 0.089$; [Supplementary Fig. S2C](#)), which might in part be governed by birth weight, as removal of birth weight as confounding factor resulted in restoring of statistical significance as observed at the level of genus ($p = 0.02$; data not shown). Interestingly, OTU-level analysis of microbiota associated with RSV viral load provided increased insight with regard to *Haemophilus*-classified OTUs, as these best explained association with viral load, but not other OTUs in the top 10 response variables ([Supplementary Fig. S2B](#)). Furthermore, the above described association of *Helcococcus* with levels of CXCL8 was not observed on the level of OTU ([Supplementary Fig. S2C](#)), and instead, was likewise best explain by OTUs classified as *Haemophilus* taxa. In conclusion, *Haemophilus* seems to be a strong correlate to the outcome for clinical manifestation of RSV disease, viral load and CXCL8 immune response.

RSV disease is marked with increased populations of *Haemophilus* and *Achromobacter*.

Further evaluation of microbial nasopharyngeal (alpha) diversity between study groups suggested that RSV infection is characterized by decreased species richness ($p = 0.006$; MWU, corrected), which is more pronounced in moderate and severe disease compared to controls ($p = 0.016$ and $p = 0.024$) ([Supplementary Fig. S3A,C](#)). Regarding species diversity, no significant differences associated with RSV infected infants or severity of disease were found ([Supplementary Fig. S3B,D](#)). A trend in reduction of species diversity in RSV patients was observed here ($p = 0.098$), and this species diversity, interestingly, is partly restored after disease recovery ($p = 0.060$, uncorrected; [Supplementary Fig. S3B](#)). Furthermore, microbial make-up of the nasopharynx of RSV-infected compared to healthy individuals was characterized by a strong and significant overrepresentation of *Haemophilus* ([Fig. 4](#)), increasing from 11.8% to 37.8% relative abundance ($p = 0.011$; MWU, corrected) and (with lower abundance) of *Achromobacter* ([Supplementary Fig.](#)

S4A; increasing from 0.7% to 4.7%; $p = 0.001$). After recovery from RSV disease, based on a paired sample analysis, the level of *Haemophilus* was restored in these infants, and was found to be similar to the level of healthy volunteers (**Fig. 4B**). In contrast, a significant underrepresentation of *Veillonella* and *Leptotrichia* was observed during RSV infection (**Supplementary Fig. S4B-C**), which however might be a consequence of the increase of the aforementioned bacteria. On a side note, *Moraxella* appeared

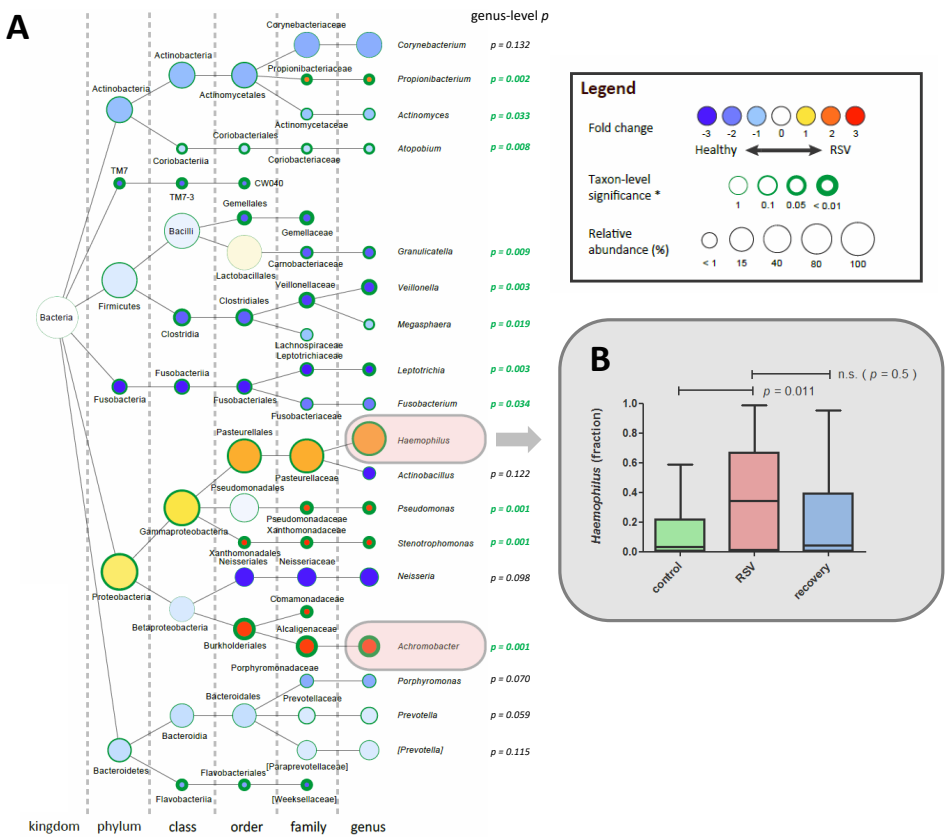


Figure 4. Difference in nasopharyngeal microbial community composition between healthy and RSV-infected individuals is strongly characterized by an overrepresentation of *Haemophilus*. The strongest differentially abundant microbial taxa for healthy versus RSV-infected individuals are shown in a graphical Cytoscape visualization (**A**) [136]. Nodes represent taxa (node size represents average relative abundance (i.e. dominance) for both experimental groups combined), edges (dashed lines) link the different taxonomic levels. The weighed fold-change (node color) is calculated as the $^2\log$ of the ratio of the relative abundance between healthy and RSV (0 = no difference between disease state, 1 = twice as abundant in RSV, etcetera). So, yellow to red indicates an overrepresentation during RSV infection, hence an underrepresentation in healthy infants, and vice versa for light to dark blue. The significance (node border width) is expressed as the p -value of a Mann-Whitney U test, FDR-corrected for multiple testing. The genus-level p -values are listed on the right of the genera nodes. We observe a strong and significant overrepresentation of *Haemophilus* genus during RSV infection ($p = 0.011$) (**B**) and of *Achromobacter* ($p = 0.001$) (**Supplementary Fig. S4A**). For recovery versus RSV samples, significance was determined using Wilcoxon signed rank test.

to be primarily present in recovery samples from RSV infected individuals (Fig. 2, and Supplementary Fig. S5). In conclusion, RSV infection caused a strong microbial perturbation characterized by presence of *Haemophilus* and *Achromobacter*.

Mucosal IL6 and CXCL8 responses correlate with clinical RSV disease severity, but not with a specific microbiota profile.

Local host response measurements in the NPA indicated that cytokines and chemokines are either negatively (CCL5, CXCL10 and MMP9; Fig. 5A) or positively (CXCL8 and IL6; Fig. 5B) correlating with RSV disease severity. This was statistically confirmed for IL6 and CXCL8 by Spearman correlation ($p = 0.0009$ / $\rho = 0.39$ and $p = 0.0001$ / $\rho = 0.51$, respectively; uncorrected), although between RSV sample groups only we did not find significant differences with ANOVA. As severity measure based on oxygen and mechanical ventilation requirement is a non-continuous value, we validated the above numbers by correlating with duration of hospital stay, which is an accepted measure of RSV severity, and found similar correlations ($p = 0.0005$ / $\rho = 0.41$ and $p = 0.0003$ / $\rho = 0.46$, for IL6 and CXCL8, respectively; uncorrected). Furthermore, paired analysis with recovery samples indicated that the IL6 and CXCL8 cytokine responses were restored to normal levels after recovery from RSV disease, as the responses after recovery were

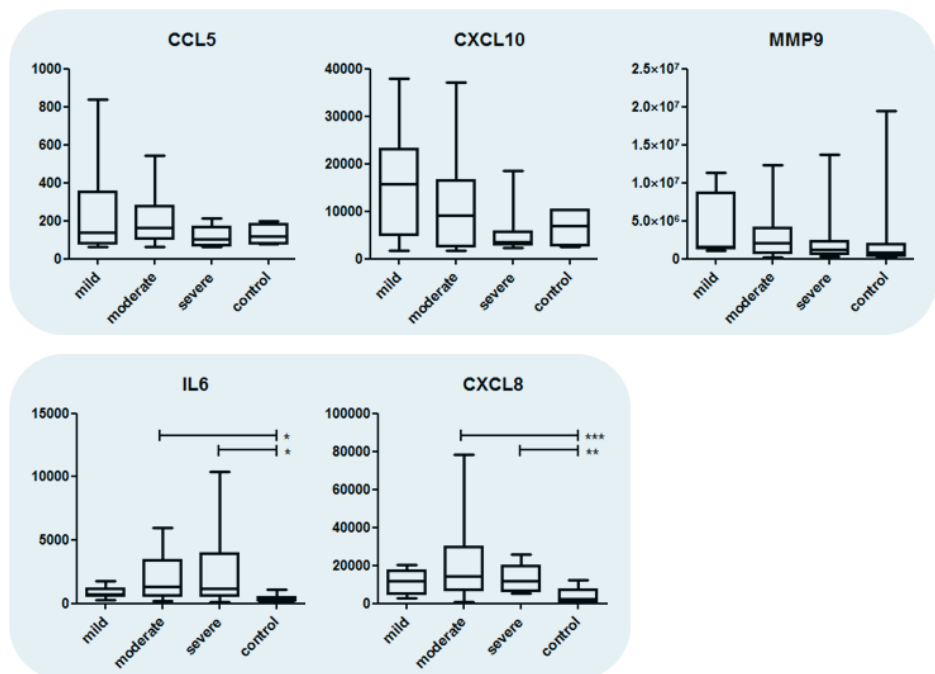


Figure 5. Chemokine and cytokine levels during RSV infection.

Chemokine and cytokine levels in nasopharyngeal aspirates of healthy, RSV-infected individuals with different disease severities (A-B). In general, host responses are observed to be negatively (A) or positively (B) correlated with RSV severity. Data is presented in pg/ml. Statistics in these plots by Kruskal-Wallis one-way ANOVA, with Dunn's correction for multiple testing.

Significances as follows: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

significantly lower than for their corresponding sample upon disease (IL6 $p = 0.02$; CXCL8 $p = 0.005$; [Supplementary Fig. S6](#)).

In line with our findings as described above, RDA indicated that healthy infants and RSV infected patients do indeed display very distinct microbial patterns ([Fig. 6A](#); RDA $p = 0.007$). However, this RDA separation seemed to be primarily driven by disease state rather than disease severity, as an additional RDA on the subset of RSV patients alone did neither yield convincing nor significant separation of RSV severity groups (data not shown; RDA $p = 0.75$). Correspondingly, levels of *Haemophilus* did not differ significantly between patient groups with distinct disease severities ([Fig. 6B](#)); although in comparison to control subjects, *Haemophilus* levels were higher in moderate and severe patients than in mild.

Interestingly, based on an extended *in silico* analysis of the dominant *Haemophilus*-level classified OTUs detected in this study ($>0.01\%$ average relative abundance, $n = 10$), we conclude that *Haemophilus* data described in this study predominantly belongs to the species *Haemophilus influenzae* ([Supplementary Fig. S7](#)). The top 10 most abundant

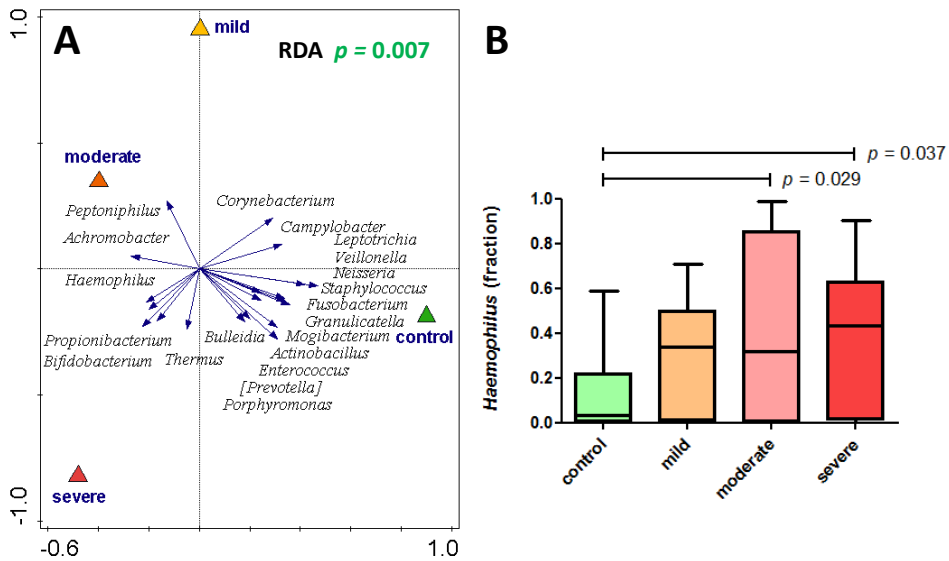


Figure 6. Nasopharyngeal microbiota composition allows for separation of healthy and RSV-infected individuals.

Nasopharyngeal genus-level microbiota from healthy and RSV-infected individuals are significantly different (A). The figure shows a redundancy analysis (RDA) biplot. Triangles are the centroids of the study sample groups: mild (yellow), moderate (orange) and severe (red) RSV, and healthy control (green). The blue arrows are the 20 best-fitting bacterial genera (names in italic), i.e. taxa that best explain the differences between the sample groups. The horizontal axis maximizes the variation in sample groups (in contrast to a principal component analysis plot, where the variation between individual samples is maximized). In RDA, samples (also) separated in the vertical direction indicates that this separation is (also) driven by other factors than the primary contrast, such as by individuality. The difference in microbiota is significant (according to a permutation test; $p = 0.007$). We observe a strong and significant overrepresentation of *Haemophilus* genus in RSV ($p = 0.011$; MWU, FDR-corrected) especially in moderate and severe RSV infections (B), and of *Achromobacter* ($p = 0.001$) ([Supplementary Fig. S4A](#)).

Haemophilus-classified OTUs together comprise 99.65% of the reads belonging to all 361 detected *Haemophilus* genus OTUs (Supplementary Table S4). Of these 10 OTUs, one could confidently be assigned to *H. parainfluenzae*, and comprises 0.25% of all reads in this study, and two other OTUs were assigned to an unknown cluster (0.31%). The remaining seven OTUs, the major fraction of *Haemophilus*-assigned reads, could confidently be assigned to *H. influenzae* (27.7% of all reads in this study). Other candidate taxa involved in the RDA separation, such as *Achromobacter*, *Veillonella* and *Leptotrichia* were not highly abundant in the overall microbiome and did not correlate with disease severity (Supplementary Fig. S4). In conclusion, RSV is marked by a strong increase in *Haemophilus* and to lesser extent *Achromobacter*, but RSV disease severity cannot be related to microbiota composition.

Discussion

Understanding why some children develop severe bronchiolitis while most children experience a upper respiratory tract infection upon RSV infection remains essential and needs to be answered to improve the care of RSV infected children in the future. Where several previous studies focused on the microbial content involved in bronchiolitis, most notably by Hasegawa and coworkers, our study exclusively focused on RSV-implicated bronchiolitis [426, 427]. RSV disease severity is a multifactorial problem, in which the viral load and the inflammatory response are important drivers of disease, although this is mainly true in previously healthy children whose airways are normal [405, 406, 411]. An important question this study tried to answer is whether nasopharyngeal microbiome composition relates to local viral load and exerts an influence on mucosal immune responses. Viral load and mucosal immune responses are thought to directly impact disease severity, and therefore it is difficult to disentangle these direct and indirect effects of the microbiome on disease outcome. The alpha diversity of the microbiota from patients with an acute RSV infection showed a decrease in species richness but not in diversity compared to healthy infants. This means that the number of different species reduces while there is no difference in the distribution of the numbers of each species. A reduced alpha diversity of the upper respiratory tract has also been shown in a longitudinal cohort study of symptomatic rhinovirus infections in infants (<1 year) [428] and after administration of an intranasal live-attenuated influenza vaccine in adults [429]. Both studies were performed prospectively and show that lower alpha diversity is caused by viral infection, and is not a prerequisite for viral infection. In another observational study, comparing the microbiome during RSV and rhinovirus infections, it was shown that there might be an association with different nasopharyngeal microbiome profiles, although this has not been confirmed by others yet [430]. This suggests that the type of alteration of the microbiome might be specific for RSV infection, but also could be a more general effect of immune system perturbation. Yet, it is uncertain whether the observed microbiome changes in our study are cause or effect of RSV infection.

A multivariate redundancy analysis (RDA) could separate healthy controls from RSV infected infants, but could not separate microbiomes based on the level of disease severity. In the study of Steenhuijsen Piters *et al.*, the authors concluded that the

microbiota composition might affect the clinical severity, but we were unable to reproduce these findings, perhaps because different RSV severity measures were applied. Nevertheless, a strong and significant overrepresentation of *Haemophilus* was found in RSV infected children, especially in severe and moderate disease. In a recent study, Rosas-Salazar and coworkers observed a similar increase of *Haemophilus* in infants with acute RSV infection compared to healthy controls, but also of *Streptococcus* and *Moraxella* [431]. In line with these and our own findings, Hasegawa and coworkers reported that infants with RSV infection and *Haemophilus*-dominant profiles had higher odds of intensive care use than RSV infected infants with profiles dominated by other bacteria, such as *Moraxella* [427]. In the MARC-35 bronchiolitis cohort by Mansbach and coworkers, lower levels of *Haemophilus/Moraxella* were found in the nasopharynx of RSV infected infants in comparison to rhinovirus, however, this study did not include healthy infants [421]. Although in our studies *Achromobacter* shows a much lower abundance than *Haemophilus*, it appeared to be also dominant in RSV infected infants. Apart from a few studies describing the association of *Achromobacter xylosoxidans* with pneumonia in newborns and young infants, and chronic infection in cystic fibrosis patients, not much is known about this species [432-434]. Much more *in vitro* and *in vivo* evidence is present with regard to increased susceptibility for viral infections and increased inflammatory responses associated with *Haemophilus* colonization and infection [435-439]. However, to the best of our knowledge this is the first time that it is shown that the composition of the microbial colonization in the nasopharynx is associated with a higher RSV load, with *Haemophilus (influenzae)* as most prominent candidate contributing to these differences (Supplementary Fig. S2B). Interaction of *Haemophilus* with epithelial cells may lead to the suppression of anti-viral immune pathways allowing increased replication. An alternative explanation would be that RSV facilitates the colonization of *Haemophilus*. However, no evidence for any of the two hypotheses can be found in the current literature. The *Helcococcus* genus, another lead based on our reported association with levels of CXCL8 in the NPA (Fig. 3C), is not likely to have a great impact on its niche, as it is very lowly abundant (0.01% on average), is only represented by 3 RSV patients and the association was not observed at the level of OTUs. The increase in IL6 and CXCL-8 responses in RSV patients colonized with *Haemophilus* as found in our study corroborates a recent *in vitro* study by Gulraiz *et al.* 2015. They showed that that release of IL6 and CXCL-8 after RSV infection, but not rhinovirus infection, was synergistically increased in *Haemophilus influenzae* pre-treated human bronchial epithelial cells [437], suggesting that the interaction between *H. influenzae*, CXCL8 and RSV may be specific for RSV infections. It should finally be noted that other cytokine responses than the ones studied here might be important for RSV pathogenesis, such as T-helper 1, -2 and -regulatory cell type cytokines [440]. Their potential involvement with the nasopharyngeal microbiome during RSV pathogenesis can therefore not be confirmed within the current study.

Interestingly, Steenhuijsen Piters *et al.*, recently reported an enhanced (non-significant) CXCL8 gene expression response, measured in a whole blood transcriptome analysis, that was associated with nasopharyngeal microbiomes that were dominated by *Haemophilus* [422]. In our study, we are able to show a significant association between nasopharyngeal CXCL8 levels and RSV infection, and even with disease severity.

Although presence of *Haemophilus* was related to the levels of CXCL8, we could not confirm a direct relation between the composition of the microbiome and disease severity. A possible explanation could be that although we stratified the patients on severity level, all patients included in our study were hospitalized and therefore severely ill. The differences in disease severity might have been too subtle to correlate them to microbial composition. In contrast to the study from Steenhuijsen Piters *et al.*, we chose to follow a more unbiased approach and therefore we did not stratify for 'nasotypes' on forehand. This may have led to the fact that we could not confirm a relation between *Streptococcus* and disease severity. In a previous study, we even found a reversed relation between *Streptococcus pneumoniae* and disease severity, although this was based on qPCR data and not on 16S sequencing [423, 441, 442]. It should also be noted that although we rigorously corrected our data analyses for a number of potential confounders (i.e. age, gender and birth weight), we did neither have socioeconomic information of our cohort, nor on mode-of-delivery, which can be seen as limitation of this study.

In conclusion, although no association of the nasopharyngeal microbiota to disease severity was found, we show that RSV infection affects the microbiota composition. *Haemophilus*-dominated profiles were in part associated with an increased viral load and increased IL6 and CXCL-8 responses on the genus level of, and these effects were even stronger on the level of OTU (*Haemophilus influenzae*). Upon recovery, *Moraxella* appears to thrive in the previously RSV-perturbed microbiomes. A better understanding of the mechanisms behind the influence of the microbiota on these host-virus processes is needed and should be the focus of future research. Improved insight in microbiome effects on RSV pathogenesis might pave the way for new preventive and therapeutic strategies to reduce the burden by RSV disease.

Conclusions

Severity of RSV infection in infants is determined by several factors. A growing body of evidence suggests that the nasopharyngeal microbiome may play an important role and influence both the local immune response as well as viral load. Interactions between RSV, local mucosal nasopharyngeal microbial content and the plethora of host factors involved are not well enough understood to explain severity of disease. This retrospective study using a well-characterized cohort of young children with RSV infection and age-matched healthy controls shows for the first time that composition of nasopharyngeal microbiota associates with CXCL8 levels in RSV patients. CXCL8 is an important chemokine that correlated with RSV disease severity. *Haemophilus* was identified as the most important genus associating with the amplitude of the CXCL8 response. Host-microbe interactions are increasingly recognized as important factors in determining outcome of host processes. Our findings show a strong interdependency between the nasopharyngeal microbiota and the mucosal immune response, potentially influencing severity of disease. These data contribute to a better understanding of the importance of commensal microbiota in respiratory health and disease.

Methods

Study design

This study was performed in two hospitals in Nijmegen, Radboud University Medical Center and Canisius Wilhelmina Ziekenhuis (CWZ). From the area of Nijmegen, children younger than 2 years of age with laboratory confirmed RSV infections were prospectively included during three consecutive winter seasons (2010/2011, 2011/2012 and 2012/2013), if they were hospitalized to the pediatric ward or intensive care unit (PICU) [443]. Written informed consent was obtained from all parents. Because the peak incidence of (severe) bronchiolitis is below the age of 6 months, and to limit the variation in age-related effects on microbiome and inflammatory response, we only included the children younger than 6 months with a PCR confirmed RSV infection. Healthy age-matched controls admitted to the hospital for surgery, who needed an elective inguinal hernia correction and had no signs of respiratory infection and a negative PCR for RSV, were included. We obtained permission from the medical ethical commission of the Radboud University Medical Center to collect control samples during one year, resulting in 21 samples as used in this study. Patients younger than 6 months with PCR confirmed RSV-positive bronchiolitis were selected and divided into three groups. Children without hypoxemia were classified as 'mildly ill', 'moderately ill' children received supplemental oxygen, while 'severely ill' children required mechanical ventilation. Patients with congenital heart or lung disease, immunodeficiency or glucocorticoid use were excluded. Within 24h after admission, a nasopharyngeal aspirate (NPA) was collected and parents from hospitalized children were asked for permission to collect a second NPA sample 4-6 weeks after admission (recovery). The final study cohort as reported here existed of $n = 21$ healthy infants, and $n = 9$ mild (2), $n = 27$ moderate (16) and $n = 18$ severe (7) patients (recovery samples in brackets). For more details on cohort design, available demographics and sample characteristics we refer to [Supplementary Table S1A](#).

Nasopharyngeal aspirate collection and diagnostics

The nasopharyngeal aspirates (NPA) were collected by introducing a catheter into the nasopharyngeal cavity. For viral diagnostics samples were analyzed by multiplex PCR, quantifying 15 different viral pathogens: influenza virus types A and B, coronavirus 229E and OC43, human bocavirus, enterovirus, adenovirus, parechovirus, PIV types 1-4, human metapneumovirus, rhinovirus (RV), and RSV, as previously described [444]. See [Supplementary Table S1B](#) for co-infection information of the samples. IL-6, CXCL8, CXCL10, CCL5 and MMP9 levels were measured by ELISA as described in [Supplementary Methods](#).

16S rRNA gene amplification prior to sequencing

Bacterial DNA extraction and quantification was performed as previously described [97] with some modifications as reported in [Supplementary Methods](#). To generate the PCR amplicon libraries, sample-specific barcoded amplicons for the V3-V4 hypervariable region of the small subunit ribosomal RNA 16S genes were generated using a two-step PCR. 10-25 ng genomic (g)DNA was used as template for the first PCR with a total

volume of 50 µl using the 341F (5'-CCT ACG GGN GGC WGC AG-3') and the 785R (5'-GAC TAC HVG GGT ATC TAA TCC-3') primers appended with Illumina adaptor sequences.

16S rRNA marker gene sequencing

Illumina 16S rRNA amplicon libraries were generated and sequenced at BaseClear BV (Leiden, The Netherlands) on an Illumina MiSeq paired-end 300 system. Sequencing data analysis was performed by QIIME and is more elaborately described in [Supplementary Methods](#). The Ribosomal Database Project [225] classifier version 2.3 was performed for taxonomic classification of the sequence reads. Alpha diversity metrics (PD whole tree, Chao1, Observed Species and Shannon) were calculated by bootstrapping 4822 reads per sample, and taking the average over 10 trials. Figures resulting from QIIME clustering analyses were generated using the interactive tree of life (iTOL) tool [445]. For visualization of the differential microbiome, Cytoscape software version 3.1.3 was used [136].

Statistics

For the microbiota data, statistical significance between contrasts with regard to taxonomy abundances was tested by a non-parametric (unpaired) Mann-Whitney U (MWU) test, corrected with False Discovery Rate (FDR) for multiple testing; unless stated otherwise. Multivariate Redundancy Analysis (RDA) and Principal Component Analysis (PCA) was done using Canoco 5.04 [228]. For all other experimental data (i.e. protein measurements, metadata, etc.), statistical significance was tested likewise using a non-parametric Kruskal-Wallis one-way ANOVA, with Dunn's correction for multiple testing (GraphPad Prism 5.0); unless stated otherwise.

For a more detailed description of this Methods section, see [Supplementary Methods](#).

Supplementary Methods

([online](#)) Additional file 6 of online repository at :
<https://doi.org/10.6084/m9.figshare.5782695.v1>

Supplementary Figures

Figure S1. Potential confounding characteristics of this study. (WORD)

(online) Figure S1 of Additional file 2 at :

<https://doi.org/10.6084/m9.figshare.5782626.v1>

Figure S2. RSV disease and viral load in healthy and RSV-infected individuals can be explained by nasopharyngeal microbial make-up (OTU-level). (WORD)

(online) Figure S2 of Additional file 2 at :

<https://doi.org/10.6084/m9.figshare.5782626.v1>

Figure S3. Species richness, but not diversity, is reduced in RSV-infected infants. (WORD)

(online) Figure S3 of Additional file 2 at :

<https://doi.org/10.6084/m9.figshare.5782626.v1>

Figure S4. Difference in *Achromobacter*, *Veillonella* and *Leptotrichia* abundance between healthy and RSV-infected individuals with different disease severities. (WORD)

(online) Figure S4 of Additional file 2 at :

<https://doi.org/10.6084/m9.figshare.5782626.v1>

Figure S5. RSV-infected individuals with mild disease symptoms display an 'intermediate' nasopharyngeal microbiota compositional profile in comparison to moderate/severe disease and their recovery controls. (WORD)

(online) Figure S5 of Additional file 2 at :

<https://doi.org/10.6084/m9.figshare.5782626.v1>

Figure S6. Chemokine and cytokine levels during RSV infection and upon recovery. (WORD)

(online) Figure S6 of Additional file 2 at :

<https://doi.org/10.6084/m9.figshare.5782626.v1>

Figure S7. Comparison of OTUs from current study to *Haemophilus* reference species shows that *Haemophilus*-classified OTUs are predominantly belonging to *Haemophilus influenzae* species.

(online) Figure S7 of Additional file 2 at :

<https://doi.org/10.6084/m9.figshare.5782626.v1>

Supplementary Tables

Table S1: Participants metadata and characteristics. (EXCEL)

(online) Additional file 1 of online repository at :
<https://doi.org/10.6084/m9.figshare.5782608.v1>

Table S2: Samples sequencing data numbers and metrics. (EXCEL)

(online) Additional file 3 of online repository at :
<https://doi.org/10.6084/m9.figshare.5782638.v1>

Table S3: Samples microbiota compositional table. (EXCEL)

(online) Additional file 4 of online repository at :
<https://doi.org/10.6084/m9.figshare.5782659.v1>

Table S4: Samples OTU table. (EXCEL)

(online) Additional file 5 of online repository at :
<https://doi.org/10.6084/m9.figshare.5782677.v1>

CHAPTER 9

GENERAL CONCLUSION AND DISCUSSION

General conclusion and discussion

In my view, for each current thesis chapter, we have had ample time and space to discuss research results and implications for those specific topics, either between all research partners involved and in conjunction with the respective journals where the research results have been published. The results of these processes have accordingly been reported in the discussion section of each chapter, here in this thesis. Therefore, I propose to here, in this discussion of my dissertation, to instead focus on what I have come across during my PhD candidature in terms of science, bioinformatics and the academia. In addition, to share what I consider to be my major learning points and challenges, and the many ideas resulting from this. Hence, the current chapter will be a 'meta discussion' on those research-related topics. In particular, I will focus on microbiota sequencing, data visualization, data science and ethics, the importance of experimental design and the role of the bioinformatician therein, data analysis strategies, and will finally come to conclusions with ideas about strategies for going from association to causation, because this last is an important challenge to tackle if we together want to further the field of biomedical research.

1. The importance of data visualization.

In order to yield significant biological insight from study data, it is crucial to have an in-depth understanding of the study questions, experimental design and conditions. Although data analysis is for its major part comprised of data handling, data processing, quality control and statistical testing, visualization is a vital part of this process. During my PhD studies, I realized that visualization of data may even be the most important step in data analysis, mainly because visualization is an effective tool to perceive patterns in the data. To find biologically relevant patterns, however, is a challenge and requires biological knowledge.

The bioinformatician is typically faced with a tremendous amount of (different types of) data, study parameters, and different (biological) questions. Being able to efficiently and effectively combine and visualize the data and its characteristics requires creativity, insight and experience. Visualization allows one to perceive patterns in the data, even those invisible with conventional statistics. Patterns observed through data visualization – *e.g. by clustering of data points, co-appearance, over- or underrepresentation of variables, colouring changes over time or contrast, etc.* – can provide leads for subsequent analysis or for corroboration by statistical methods. Hence, every question and every lead results in a set of follow-up questions, for science in general, but also for the process of data analysis. Ideally, follow-up experimental studies contain elements from the previous, elaborating on the former, an iterative process once referred to as 'knitting' by one of my first supervisors in the academia. I strongly believe this process is also true for data analysis, and is strongly guided by data visualization.

2. Application of marker gene sequencing on the verge of a 'Shotgun' era.

Unquestionably, in this thesis, metagenomics (shotgun) would have in many ways been a more favourable technique for solving the research questions at hand than

the currently applied marker gene sequencing such as 16S and SLST. More precisely, shotgun allows for higher resolution taxonomy assignment, and more importantly, allows for mining whole genome genetic information for the most abundant microbes, and thus for predicting the complete microbiome its function potential. Nonetheless, our reasons for choosing marker gene sequencing over shotgun was a single and pragmatic one: sequencing costs. Shotgun is currently roughly five to ten times more expensive than marker gene sequencing ([Chapter 1: Table 1](#)). In our view, the main reasons to choose shotgun over marker gene sequencing is when there are specific research questions with regard to, for example: microbial growth and metabolism, presence of virulence factors and antibiotics resistance. There are however certain additional challenges for application of shotgun: (i) much more sequencing data will be generated, making analysis of shotgun data computationally more expensive; (ii) shotgun data requires alternative analysis pipelines and methodologies than for marker gene sequencing, something which is not straight-forward and implementation of which requires dedicated investment by the involved research groups; (iii) the detection-limit for identification of microbes is higher for shotgun because whole genomes are being sequenced in comparison to a single gene for marker gene sequencing (roughly >1% to >0.01% relative abundance, respectively); and (iv), shotgun requires more DNA for sequencing compared to marker gene sequencing, something which is a challenge for example for skin which typically yields only low amounts of microbial DNA. On the other hand, 16S does allow for a 'poor man's' prediction of the microbiome its function potential by means of previously discussed methods such as PICRUSt *et al.*, as was successfully applied in [Chapters 4 and 6](#) of this thesis. Nevertheless, 16S function inference-based methods are inferior to shotgun, and I do acknowledge that with shotgun we would likely have been able to get more out of our data than presented in current thesis. For instance, the observed decrease in GPAC bacteria on filaggrin-deficient skin, as described in [Chapter 4](#), would have allowed us to mine the specific genetic content of these bacteria in order to search for (histidine utilization) genes unique to this phylogenetic clade, in order to explain their underrepresentation in IV patients. Likewise, in [Chapter 6](#), we could have specifically searched for complete genes and pathways involved in production dopamine precursors. In both examples, of [Chapters 4 and 6](#), metagenomics data would likely have allowed us to corroborate our current findings by a more targeted analysis. In [Chapters 5 and 7](#) we have likewise adopted 16S sequencing. However, I believe that it would not have changed our results and conclusions, although it could have allowed for identification of additional biomarkers that are now being missed. Ideally, in due time, when sequencing costs drop, replication studies with shotgun need to be performed, preferably on larger cohorts, to validate and extend the findings reported in this thesis.

3. Ethics and accountability in data sciences.

When it comes to bioinformatics, data analysis and statistics, I believe that, and I have first-hand experienced during my work on this thesis, that in order to generate (novel) biological knowledge and insight based the underlying data, it is important that multiple lines of evidence point in the same direction. For example, based on different visualization approaches, types of metrics, modes of statistics, etc. In other words, one observation in your data is usually not sufficient to draw strong and convincing

conclusions. This, for me, has been an eye-opener: a simple yet effective manner to cope with issues and concerns I personally experienced with regard to questions like: “*Why did you use this statistical test and not that alternative?*”, “*On what terms did you decide it to be warranted to exclude these samples from analysis?*”, “*Why do you put any value in a (borderline) statistically significant, yet multiple-testing corrected insignificant effect?*”, “*Why did you look at this sample contrast and not that?*” or “*Why can you conclude your samples to be of sufficient quality whereas you’ve detected (low levels of) notorious contamination-associated microbes?*”. When such concerns are raised by colleagues, one can account for most of these, if the initial decisions were not based on a single observation.

In conclusion, it all boils down to being able to (i) explain what you observe in your data, (ii) making sure multiple lines of evidence point in the same direction of your conclusion(s), and (iii) when great care and effort has been undertaken, to openly discuss and explain all decisions that were made in the process. Alternatively, something can be statistically significant; yet, if it doesn’t make any sense, or if you can’t explain it, or if it does not fit in your running hypothesis / story, then why put value in (prominently) reporting it? Of course, the most unexpected findings can sometimes lead to dogma-changing theories; still, such observations can never stand on their own and should be backed by substantial additional evidence. For me, data analysis has its limitations: if I don’t (biologically) understand (single observations in) my data, the patterns, the observed effects, etc., then I cannot use them. I do not see the point of endless listings of (microbial) abundances, statistical outcomes, etc., when there is no clear story behind it. At best, I make sure to provide them as supplementary tables for data browsing and mining, for future alternative hypotheses for me or others, and to prevent data from getting lost. When adhering to the previous, I strongly believe this scientifically justifies decisions being made at any time (for instance, see our reply to Meisel *et al.*, 2017 [40]), and to confidently enter ethical discussions whenever they may arise.

4. The importance of experimental design.

A proper experimental design and analysis strategy goes hand in hand with the right biological question. Therefore, on beforehand, it is necessary to agree on: (i) the cohort size, (ii) parameter distribution between the subjects and the expected value diversity across subjects, (iii) what parameters should be measured, and (iv) across how many time-points or repeated measures. The experimental design and analysis methods applied to its data dictate the statistical power. In our experience, in some cases an analysis approach is *ad-hoc* applied to the data generated on an ongoing or already finished study, at which point study design cannot straight-forwardly be changed. Although this has been a challenge, we have therefore always strived to setup study design in collaboration with our partners in order to maximize the potential of uncovering meaningful patterns in the (microbiome) data.

5. The role of the bioinformatician in biomedical sciences.

In our experience, it has been difficult to involve bioinformaticians in early research phases. Unfortunately, we experienced (and still do) that many research groups start

searching for data analysts once experimental data has already been generated. This widely spread and persistent culture of *in retrospect* involving bioinformaticians and data scientists should change to early participation of all researchers, if we want to maximize value and impact of research projects.

I believe this can only be achieved if on the one hand bioinformaticians do not isolate themselves in their 'dry lab' offices, but instead physically join workspace with 'wet lab' scientists. The latter, on their turn, should learn to understand the importance of defining study data characteristics and approaches in early collaboration with their data analysis partners. Consequently, I believe there is still work to be done to improve and optimize this, and this involves cooperation from both sides. Although I believe the described problem is still commonly present, I am carefully optimistic to see that this is slowly shifting for the better, especially as demonstrated by collaborations reported in this thesis.

Furthermore, in biomedical sciences, it is very important that bioinformaticians start getting involved in research of their specific interest similarly as scientists at the laboratory do. What do I mean by this? In my personal experience with numerous bioinformaticians (from conferences, symposia, collaborations, etc.), many see themselves not as (biomedical) scientists, but more as service providers, or developers of bioinformatics tools, statistics and methods, etc. Although I would not dispute that these are important aspects to advance the specific fields and the general scientific community, bioinformaticians in the academia should not forget that they are trained to be researchers like any other, but applying *in silico* rather than wet lab techniques for data collection, analysis and reporting. If not, their function is better described as technical support, software development, etc., instead of an independent (not meaning isolated, uncooperative) and dynamic scientist with knowledge beyond its computer terminal. I believe this last should be the role of a bioinformaticians, although, obviously, there is nothing wrong with a supportive role in a team. Furthermore, bioinformatics software development such as data analysis tools, methods, pipelines, etcetera, is similar to techniques which are developed on the wet lab: i.e. technologies to advance research, for example *in silico* or animal models, plethora of DNA, RNA, protein or cell measuring equipment, etcetera. In the end, what matters is not development of such techniques, but how these can be adopted to advance the scientific field. Unfortunately, I encountered many bioinformaticians who see bioinformatics as a general goal, rather than a means to do research. I believe this distinction discriminates between being a (biomedical) scientific bioinformatician or a software engineer / statistics consultant, although, obviously, there is no right or wrong with regard to any of these job descriptions. Nevertheless, I do believe, in other words, that *bioinformatics should be a means to an end, not an end goal in itself*.

On the other side of the spectrum, laboratory scientists need to appreciate the role of bioinformaticians beyond that of 'data crunchers' or 'informaticians'. To realize that they are expected to understand the biology of the problem, in order to connect their data analysis outcomes to underlying biology, whether this data is experimentally or *in silico*-derived. This concept is analogous to wet lab scientists who should be

able to extract meaning from their experimental data. Of course, it is a challenge, if not very difficult, for bioinformaticians to possess enough background knowledge on multiple fields. Mainly because their expertise and application more easily bridge different research fields of focus than for wet lab scientists. For example, microbiomics-related questions can be applied to the field of inflammation and immunity, but also to cancer research, food and nutrition, host-microbe or microbe-microbe interactions, environmental biology, etc. Obviously, this is not limited to bioinformatics, and is analogous to scientists involved in technology platforms of, for example, mass-spectrometry (proteomics), flow-cytometry (cell biology) and high-throughput screening assays (drug discovery). Although data deriving from such techniques, where large amounts of data are generated, can in principle all be labelled as bioinformatics. Nonetheless, for this reason, I believe it is wise for a bioinformatician to focus his / her research on one field to narrow-down the required background knowledge to a workable amount. Such as in this thesis: microbiomics and host-microbe interactions in skin, oral and gut health and disease, with particular interest in inflammation and immunity.

6. Hypothesis-based versus hypothesis-free data analysis strategy.

In the last paragraphs, I shortly discussed the large amount of data that is typically being generated nowadays, and the difficulties involved. When dealing with this type of data, it is considered 'good practice' to correct for the number of statistical tests that are being performed in an analysis: the so called multiple testing correction. Common examples of multiple testing correction methods are by Bonferroni or False Discovery Rate (FDR), the latter also known as the Benjamini–Hochberg procedure. For aforementioned reasons, methods like these were by default applied in this thesis, most notably in [Chapters 5-8](#). However, multiple testing correction can be a challenge for analysis of datasets with a large number of different measurements or variables, such as for metataxonomics, metagenomics or (meta)transcriptomics data. Illustratively, say we have a dataset with 10,000 variables, for example pathway reactions, and we statistically test every pathway reaction for being differentially abundant in a given contrast of two sample groups, for example healthy versus diseased. Then, according to Bonferroni, one has to divide the p -value statistical significance threshold ($\alpha = 0.05$) by that same number, thus yielding a threshold of $(0.05 / 10,000)$ of $p < 5 * 10^{-6}$. This statistical threshold can in theory only be reached with very large sample groups, which in a biomedical setting is an extremely laborious and costly endeavor. One way to cope with this notorious problem is by *not* looking at *all* variables in your dataset (hypothesis-free), but to only look at those candidate variables that fit your hypothesis (hypothesis-based). For example, by only looking at one or several specific pathways-of-interest. By this simple yet effective approach, one circumvents large numbers for multiple testing correction, thereby making the analysis strategy more accessible. In this thesis, we adopted such hypothesis-based analysis approach in a similar data setting, that of 16S-based function prediction (by PICRUST), which typically yields thousands of unique pathways or elements thereof (see [Chapters 5 and 7](#)). In these chapters, by focusing on single pathways-of-interest only, we were able to successfully demonstrate that specific reactions of those pathways were specifically associated with the phenotype under study.

7. The corroboration of study findings – from association to causation.

In my view, the holy grail in any biomedical study project is to make a translation from data-derived association to experimentally-validated causality. Or in other words, to go from (experimental) data to biological (mechanistic) understanding. **Figure 1** summarizes a schematic typical workflow / study approach in order to go from association to causality, and ideally functional application. This conceptual workflow has a specific focus on skin (micro)biology (**Figure 1** panels *i-ii*), and is partly based on experience deriving from this thesis (especially **Chapters 3 and 4**). For this, a strong interaction between the experimental and computational groups is crucial in order to pinpoint promising candidates from the data at hand (**Figure 1** panels *iii-v*). Furthermore, (a few of) these candidates should be tested in *in vitro* or *in vivo* follow-up experiments or studies in order to ensure relevance of the project findings (**Figure 1** panels *vi-vii*). This requires a project team with the right mix of expertise, as exemplified by work in this thesis. I strongly believe that studies do not stop at mere association, and that strategies like presented here should ultimately result in mechanistic understanding, and ideally thereof derived functional application (**Figure 1** panel *viii*). Play time for bioinformaticians is over: it is less and less accepted in higher impact journal to report only data-derived experimental correlations. Instead, mechanistic understanding of correlations in the experimental data is highly desired. I foresee this is where the medical biology field is leading for and what is by default demanded in the near future.

Future perspective

I believe it is the role of the bioinformatician to be involved in devising project experimental designs from start, and to think about data analysis approaches before study data has been generated. Furthermore, he / she has to monitor data in each step and process of analysis, look for abnormalities in the data (quality control), and make sure that intermediate data is of expected composition and quality. Finally, the bioinformatician should know when / where and why to use what statistical test, procedure and methodology. It is important to adhere to the above considerations, now, but especially in the future, where data being generated will only continue to increase, and will also be more often combined with other types of data. Such integration of multi-omics data is already being applied to various large cohorts such as Lifelines, where data from different sources of the same individuals are collected (over time) (<https://www.lifelines.nl/>). Examples of such data, are different microbiome niches, genetics, RNA and protein profiles, output from *in vitro* cell, enzyme or challenge assays, blood biomarkers, and demographical parameters such as geographic information, ethnic background, gender, age, etcetera. In the end, it is up to the bioinformatician to make sense out of this enormous collection of information, and more importantly, to connect relevant data patterns to biological processes in order to learn. For this reason, I foresee that in the near future bioinformaticians will have a more prominent role in life sciences as central hubs to connect wet lab experimental data to novel biological understanding of processes under study.

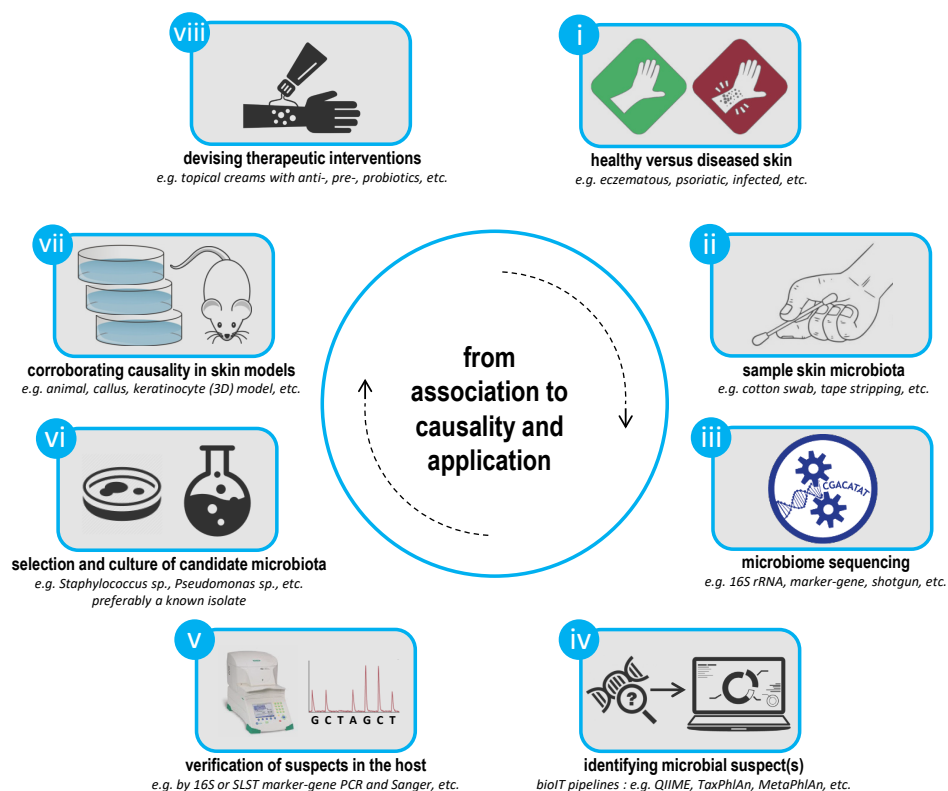


Figure 1. “Studies do not stop at association”.

Schematic typical workflow / study approach in order to potentially go from association to causality, with specific focus to skin (micro)biology. (i) Healthy versus affected skin is evaluated, with suspected involvement of the microbiome as causative driver. (ii) Skin samples are collected through standard protocols, and (iii) are sequenced by a suitable platform depending on research question and study budget. (iv) Microbial suspects are identified by available data analysis pipelines, and (v) their specific presence and (differential) abundance are validated in the host by alternative (conventional) methods. Thereafter, (vi) candidates are selected and cultured for (vii) corroboration of microbiota-associated effects of initial study findings by relevant in vitro or in vivo (disease) models. Finally, if applicable, (viii) functional applications could potentially be devised from study findings (for example organisms, proteins, compounds, protease inhibitors, etc.), ultimately leading to novel therapeutic interventions for treatment of skin disease.

Final considerations

This thesis summarizes our work on microbiome dynamics in health and disease. We have seen how different body niches bring about different types of microbes, and how different types of diseases and perturbations influence composition and function of these microbiome-associated microbes and components thereof derived.

As a bioinformatician: be involved in the full course of the project, from head to tail. Try to be creative, and have a healthy amount of paranoia. Be aware of the implications of different steps in your methods and data handling on your intermediate data and (final) outcomes, why you chose to apply or discard them, etcetera. Always question and double-check your own data and procedures, for instance, does data-in equal data-out, can you explain all numbers and output, etcetera. Be well-informed, combined with an attitude of healthy criticism towards data and procedures of others, because scientific debate is the main driver of high quality research in any group. In the end, trust your own data and analyses, and if you can, adhere to the multiple lines of evidence concept. As final words, the best fitting description I could find for the definition of bioinformatics is as follows: "*The bioinformatics is an interdisciplinary field that develops methods and software tools for understanding biological data. As an interdisciplinary field of science, bioinformatics combines computer science, statistics, mathematics, and engineering to analyze and interpret biological data*" (source Wikipedia.org). In my view, especially the "*.. and interpret biological data.*" is an important notion that we should not forget to adhere to as bioinformaticians. This step will ultimately generate new hypotheses, which over time translate into new biological knowledge, insight and mechanistic understanding, thereby driving novel biomedical and technical applications.

CHAPTER 10

LIST OF REFERENCES

List of references

(in order of citation)

1. Marchesi, J.R. and J. Ravel, *The vocabulary of microbiome research: a proposal*. Microbiome, 2015. **3**(1): p. 31.
2. Sender, R., S. Fuchs, and R. Milo, *Revised Estimates for the Number of Human and Bacteria Cells in the Body*. PLOS Biology, 2016. **14**(8): p. e1002533.
3. Zeeuwen, P.L., et al., *Microbiome dynamics of human epidermis following skin barrier disruption*. Genome Biol, 2012. **13**(11): p. R101.
4. Baeshen, N.A., et al., *Cell factories for insulin production*. Microb Cell Fact, 2014. **13**: p. 141.
5. Lin, Z., et al., *Metabolic engineering of Escherichia coli for the production of riboflavin*. Microb Cell Fact, 2014. **13**: p. 104.
6. Samazan, F., et al., *Production, secretion and purification of a correctly folded staphylococcal antigen in Lactococcus lactis*. Microb Cell Fact, 2015. **14**: p. 104.
7. Wang, C., et al., *Viable and culturable populations of Saccharomyces cerevisiae, Hanseniaspora uvarum and Starmerella bacillaris (synonym Candida zemplinina) during Barbera must fermentation*. Food Res Int, 2015. **78**: p. 195-200.
8. Abe, K., et al., *Impact of Aspergillus oryzae genomics on industrial production of metabolites*. Mycopathologia, 2006. **162**(3): p. 143-53.
9. Alkema, W., et al., *Microbial bioinformatics for food safety and production*. Brief Bioinform, 2016. **17**(2): p. 283-92.
10. Wels, M., et al., *Draft Genome Sequence of Streptococcus thermophilus C106, a Dairy Isolate from an Artisanal Cheese Produced in the Countryside of Ireland*. Genome Announc, 2015. **3**(6).
11. Plengvidhya, V., et al., *DNA fingerprinting of lactic acid bacteria in sauerkraut fermentations*. Appl Environ Microbiol, 2007. **73**(23): p. 7697-702.
12. Suzuki, K., M. Koyanagi, and H. Yamashita, *Genetic characterization of non-spoilage variant isolated from beer-spoilage Lactobacillus brevis ABBC45*. J Appl Microbiol, 2004. **96**(5): p. 946-53.
13. Sanders, J.W., et al., *Biodiversity of spoilage lactobacilli: phenotypic characterisation*. Food Microbiol, 2015. **45**(Pt A): p. 34-44.
14. Blum, H.E., *The human microbiome*. Adv Med Sci, 2017. **62**(2): p. 414-420.
15. Wanat, K.A., et al., *Bedside diagnostics in dermatology: Viral, bacterial, and fungal infections*. J Am Acad Dermatol, 2017. **77**(2): p. 197-218.
16. Iverson, W.G. and N.F. Millis, *A method for the detection of starch hydrolysis by bacteria*. J Appl Bacteriol, 1974. **37**(3): p. 443-6.
17. Clarke, P.H., *Hydrogen sulphide production by bacteria*. J Gen Microbiol, 1953. **8**(3): p. 397-407.
18. Qin, J., et al., *A human gut microbial gene catalogue established by metagenomic sequencing*. Nature, 2010. **464**(7285): p. 59-65.
19. Francis, O.E., et al., *Pathoscope: species identification and strain attribution with unassembled sequencing data*. Genome Res, 2013. **23**(10): p. 1721-9.
20. Segata, N., et al., *PhyloPhlAn is a new method for improved phylogenetic and taxonomic placement of microbes*. Nature communications, 2013. **4**: p. 2304-2304.
21. Parizad, E.G., E.G. Parizad, and A. Valizadeh, *The Application of Pulsed Field Gel Electrophoresis in Clinical Studies*. J Clin Diagn Res, 2016. **10**(1): p. De01-4.
22. Lindstedt, B.A., *Multiple-locus variable number tandem repeats analysis for genetic fingerprinting of pathogenic bacteria*. Electrophoresis, 2005. **26**(13): p. 2567-82.
23. Braem, G., et al., *(GTG)₅-PCR fingerprinting for the classification and identification of coagulase-negative Staphylococcus species from bovine milk and teat apices: a comparison of type strains and field isolates*. Vet Microbiol, 2011. **147**(1-2): p. 67-74.
24. Frickmann, H., et al., *Fluorescence in situ hybridization (FISH) in the microbiological diagnostic routine laboratory: a review*. Crit Rev Microbiol, 2017. **43**(3): p. 263-293.
25. Scholz, C.F., et al., *A novel high-resolution single locus sequence typing scheme for mixed populations of Propionibacterium acnes in vivo*. PLoS One, 2014. **9**(8): p. e104199.
26. van Bokhorst-van de Veen, H., et al., *Congruent strain specific intestinal persistence of Lactobacillus plantarum in an intestine-mimicking in vitro system and in human volunteers*. PLoS One, 2012. **7**(9): p. e44588.
27. Fernández Ramírez, M.D., *Characterisation of Lactobacillus plantarum single and multi-strain*

- biofilms. 2016, Wageningen University: Wageningen.
28. Peterson, J., et al., *The NIH Human Microbiome Project*. Genome Res, 2009. **19**(12): p. 2317-23.
29. website. *NIH Human Microbiome Project (HMP)*.
Available from: <http://hmpdacc.org/>.
30. website. *Human Gut Microbiome Initiative (HGMI)*.
Available from: <http://genome.wustl.edu/projects/detail/human-gut-microbiome/>.
31. website. *American Gut*.
Available from: <http://humanfoodproject.com/americangut/>.
32. website. *METAgonomics of the Human Intestinal Tract (MetaHIT)*.
Available from: <http://www.metahit.eu/>.
33. website. *Earth Microbiome Project (EMP)*.
Available from: <http://www.earthmicrobiome.org/>.
34. website. *International Census of Marine Microbes (ICoMM)*.
Available from: <http://icomm.mbl.edu/>.
35. website. *TerraGenome*.
Available from: <http://www.terragenome.org/>.
36. Cole, J.R., et al., *Ribosomal Database Project: data and tools for high throughput rRNA analysis*. Nucleic Acids Res, 2014. **42**(Database issue): p. D633-42.
37. DeSantis, T.Z., et al., *Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB*. Appl Environ Microbiol, 2006. **72**(7): p. 5069-72.
38. Pruesse, E., et al., *SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB*. Nucleic Acids Research, 2007. **35**(21): p. 7188-7196.
39. Baker, G.C., J.J. Smith, and D.A. Cowan, *Review and re-analysis of domain-specific 16S primers*. J Microbiol Methods, 2003. **55**(3): p. 541-55.
40. Zeeuwen, P.L.J.M., et al., *Reply to Meisel et al*. Journal of Investigative Dermatology, 2017. **137**(4): p. 961-962.
41. Meisel, J.S., et al., *Skin Microbiome Surveys Are Strongly Influenced by Experimental Design*. Journal of Investigative Dermatology, 2016. **136**(5): p. 947-956.
42. Caporaso, J.G., et al., *QIIME allows analysis of high-throughput community sequencing data*. Nat Methods, 2010. **7**(5): p. 335-336.
43. Kuczynski, J., et al., *Using QIIME to analyze 16S rRNA gene sequences from microbial communities*. Curr Protoc Bioinformatics, 2011. **Chapter 10**: p. Unit 10.7.
44. Navas-Molina, J.A., et al., *Advancing our understanding of the human microbiome using QIIME*. Methods Enzymol, 2013. **531**: p. 371-444.
45. Edgar, R.C., *Search and clustering orders of magnitude faster than BLAST*. Bioinformatics, 2010. **26**(19): p. 2460-1.
46. Schloss, P.D., et al., *Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities*. Appl Environ Microbiol, 2009. **75**(23): p. 7537-41.
47. Gardes, M. and T.D. Bruns, *ITS primers with enhanced specificity for basidiomycetes--application to the identification of mycorrhizae and rusts*. Mol Ecol, 1993. **2**(2): p. 113-8.
48. Abarenkov, K., et al., *The UNITE database for molecular identification of fungi – recent updates and future perspectives*. New Phytologist, 2010. **186**(2): p. 281-285.
49. Deurenberg, R.H., et al., *Application of next generation sequencing in clinical microbiology and infection prevention*. Journal of Biotechnology, 2017. **243**: p. 16-24.
50. Coughlan, L.M., et al., *Biotechnological applications of functional metagenomics in the food and pharmaceutical industries*. Frontiers in Microbiology, 2015. **6**: p. 672.
51. Luo, C., et al., *Individual genome assembly from complex community short-read metagenomic datasets*. Isme j, 2012. **6**(4): p. 898-901.
52. Segata, N., et al., *Metagenomic microbial community profiling using unique clade-specific marker genes*. Nat Methods, 2012. **9**(8): p. 811-4.
53. Huson, D.H., et al., *MEGAN analysis of metagenomic data*. Genome Res, 2007. **17**(3): p. 377-86.
54. Huson, D.H. and N. Weber, *Microbial community analysis using MEGAN*. Methods Enzymol, 2013. **531**: p. 465-85.
55. Huson, D.H., et al., *MEGAN Community Edition - Interactive Exploration and Analysis of Large-Scale Microbiome Sequencing Data*. PLoS Comput Biol, 2016. **12**(6): p. e1004957.
56. Wu, M. and J.A. Eisen, *A simple, fast, and accurate method of phylogenomic inference*. Genome Biol,

2008. **9**(10): p. R151.

57. Wu, M. and A.J. Scott, *Phylogenomic analysis of bacterial and archaeal sequences with AMPHORA2*. *Bioinformatics*, 2012. **28**(7): p. 1033-4.
58. Sunagawa, S., et al., *Metagenomic species profiling using universal phylogenetic marker genes*. *Nat Meth*, 2013. **10**(12): p. 1196-1199.
59. Luo, C., et al., *ConStrains identifies microbial strains in metagenomic datasets*. *Nat Biotechnol*, 2015. **33**(10): p. 1045-52.
60. Scholz, M., et al., *Strain-level microbial epidemiology and population genomics from shotgun metagenomics*. *Nat Methods*, 2016. **13**(5): p. 435-8.
61. Truong, D.T., et al., *Microbial strain-level population structure and genetic diversity from metagenomes*. *Genome Res*, 2017. **27**(4): p. 626-638.
62. Meyer, F., et al., *The metagenomics RAST server – a public resource for the automatic phylogenetic and functional analysis of metagenomes*. *BMC Bioinformatics*, 2008. **9**: p. 386-386.
63. Abubucker, S., et al., *Metabolic reconstruction for metagenomic data and its application to the human microbiome*. *PLoS Comput Biol*, 2012. **8**(6): p. e1002358.
64. Langille, M.G., et al., *Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences*. *Nat Biotechnol*, 2013. **31**(9): p. 814-21.
65. Jun, S.-R., et al., *PanFP: pangenome-based functional profiles for microbial communities*. *BMC Research Notes*, 2015. **8**: p. 479.
66. Asshauer, K.P., et al., *Tax4Fun: predicting functional profiles from metagenomic 16S rRNA data*. *Bioinformatics*, 2015. **31**(17): p. 2882-4.
67. Kaas, R.S., et al., *Estimating variation within the genes and inferring the phylogeny of 186 sequenced diverse *Escherichia coli* genomes*. *BMC Genomics*, 2012. **13**: p. 577.
68. Chen, H. and W. Jiang, *Application of high-throughput sequencing in understanding human oral microbiome related with health and disease*. *Frontiers in Microbiology*, 2014. **5**: p. 508.
69. Gajer, P., et al., *Temporal dynamics of the human vaginal microbiota*. *Sci Transl Med*, 2012. **4**(132): p. 132ra52.
70. Dickson, R.P., J.R. Erb-Downward, and G.B. Huffnagle, *Homeostasis and its disruption in the lung microbiome*. *American Journal of Physiology - Lung Cellular and Molecular Physiology*, 2015. **309**(10): p. L1047-L1055.
71. Whittaker, R.H., *Evolution and Measurement of Species Diversity*. *Taxon*, 1972. **21**(2/3): p. 213-251.
72. Gerritsen, J., et al., *Intestinal microbiota in human health and disease: the impact of probiotics*. *Genes & Nutrition*, 2011. **6**(3): p. 209-240.
73. Nardone, G. and D. Compare, *The human gastric microbiota: Is it time to rethink the pathogenesis of stomach diseases?* *United European Gastroenterology Journal*, 2015. **3**(3): p. 255-260.
74. Jung, C., J.-P. Hugot, and F. Barreau, *Peyer's Patches: The Immune Sensors of the Intestine*. *International Journal of Inflammation*, 2010. **2010**: p. 823710.
75. Shi, N., et al., *Interaction between the gut microbiome and mucosal immune system*. *Military Medical Research*, 2017. **4**: p. 14.
76. Bemark, M., P. Boysen, and N.Y. Lycke, *Induction of gut IgA production through T cell-dependent and T cell-independent pathways*. *Annals of the New York Academy of Sciences*, 2012. **1247**(1): p. 97-116.
77. Kim, S., A. Covington, and E.G. Pamer, *The intestinal microbiota: Antibiotics, colonization resistance, and enteric pathogens*. *Immunol Rev*, 2017. **279**(1): p. 90-105.
78. Gallo, R.L. and L.V. Hooper, *Epithelial antimicrobial defence of the skin and intestine*. *Nature reviews. Immunology*, 2012. **12**(7): p. 503-516.
79. Marcinkiewicz, M. and S. Majewski, *The role of antimicrobial peptides in chronic inflammatory skin diseases*. *Advances in Dermatology and Allergy/Postępy Dermatologii i Alergologii*, 2016. **33**(1): p. 6-12.
80. Ivanov, II, et al., *Induction of intestinal Th17 cells by segmented filamentous bacteria*. *Cell*, 2009. **139**(3): p. 485-98.
81. Rogier, R., M.I. Koenders, and S. Abdollahi-Roodsaz, *Toll-like receptor mediated modulation of T cell response by commensal intestinal microbiota as a trigger for autoimmune arthritis*. *J Immunol Res*, 2015. **2015**: p. 527696.
82. Kosiewicz, M.M., et al., *Relationship between gut microbiota and development of T cell associated disease*. *FEBS Letters*, 2014. **588**(22): p. 4195-4206.
83. Ochoa-Reparaz, J., et al., *A polysaccharide from the human commensal *Bacteroides fragilis* protects*

- against CNS demyelinating disease. *Mucosal Immunol*, 2010. **3**(5): p. 487-95.
84. Marteau, P., et al., *Comparative Study of Bacterial Groups within the Human Cecal and Fecal Microbiota*. Applied and Environmental Microbiology, 2001. **67**(10): p. 4939-4942.
 85. Moen, B., et al., *Effect of Dietary Fibers on Cecal Microbiota and Intestinal Tumorigenesis in Azoxymethane Treated A/J Min/+ Mice*. PLoS ONE, 2016. **11**(5): p. e0155402.
 86. Li, D., et al., *Microbial Biogeography and Core Microbiota of the Rat Digestive Tract*. Scientific Reports, 2017. **7**: p. 45840.
 87. Becker, D., et al., *Novel orally swallowable IntelliCap((R)) device to quantify regional drug absorption in human GI tract using diltiazem as model drug*. AAPS PharmSciTech, 2014. **15**(6): p. 1490-7.
 88. Lawley, T.D. and A.W. Walker, *Intestinal colonization resistance*. Immunology, 2013. **138**(1): p. 1-11.
 89. Kolling, G., M. Wu, and R. Guerrant, *Enteric pathogens through life stages*. Frontiers in Cellular and Infection Microbiology, 2012. **2**(114).
 90. Forslund, K., et al., *Country-specific antibiotic use practices impact the human gut resistome*. Genome Res, 2013. **23**(7): p. 1163-9.
 91. Gough, E., H. Shaikh, and A.R. Manges, *Systematic review of intestinal microbiota transplantation (fecal bacteriotherapy) for recurrent Clostridium difficile infection*. Clin Infect Dis, 2011. **53**(10): p. 994-1002.
 92. Millan, B., M. Laffin, and K. Madsen, *Fecal Microbiota Transplantation: Beyond Clostridium difficile*. Current Infectious Disease Reports, 2017. **19**(9): p. 31.
 93. Konturek, P.C., et al., *Emerging role of fecal microbiota therapy in the treatment of gastrointestinal and extra-gastrointestinal diseases*. J Physiol Pharmacol, 2015. **66**(4): p. 483-91.
 94. Shikina, T., et al., *IgA Class Switch Occurs in the Organized Nasopharynx- and Gut-Associated Lymphoid Tissue, but Not in the Diffuse Lamina Propria of Airways and Gut*. The Journal of Immunology, 2004. **172**(10): p. 6259.
 95. Brandtzaeg, P., *Potential of nasopharynx-associated lymphoid tissue for vaccine responses in the airways*. Am J Respir Crit Care Med, 2011. **183**(12): p. 1595-604.
 96. Ogasawara, N., et al., *Epithelial barrier and antigen uptake in lymphoepithelium of human adenoids*. Acta Otolaryngol, 2011. **131**(2): p. 116-23.
 97. Cremers, A.J.H., et al., *The adult nasopharyngeal microbiome as a determinant of pneumococcal acquisition*. Microbiome, 2014. **2**: p. 44.
 98. Bogaert, D., et al., *Variability and diversity of nasopharyngeal microbiota in children: a metagenomic analysis*. PLoS One, 2011. **6**(2): p. e17035.
 99. Watson, K., et al., *Upper respiratory tract bacterial carriage in Aboriginal and non-Aboriginal children in a semi-arid area of Western Australia*. Pediatr Infect Dis J, 2006. **25**(9): p. 782-90.
 100. Bosch, A.A.T.M., et al., *Viral and Bacterial Interactions in the Upper Respiratory Tract*. PLOS Pathogens, 2013. **9**(1): p. e1003057.
 101. McCullers, J.A., *Insights into the Interaction between Influenza Virus and Pneumococcus*. Clinical Microbiology Reviews, 2006. **19**(3): p. 571-582.
 102. Sajjan, U., et al., *Rhinovirus disrupts the barrier function of polarized airway epithelial cells*. Am J Respir Crit Care Med, 2008. **178**(12): p. 1271-81.
 103. McNamee, L.A. and A.G. Harmsen, *Both influenza-induced neutrophil dysfunction and neutrophil-independent mechanisms contribute to increased susceptibility to a secondary Streptococcus pneumoniae infection*. Infect Immun, 2006. **74**(12): p. 6707-21.
 104. Sajjan, U.S., et al., *H. influenzae potentiates airway epithelial cell responses to rhinovirus by increasing ICAM-1 and TLR3 expression*. Faseb j, 2006. **20**(12): p. 2121-3.
 105. Belkaid, Y. and J.A. Segre, *Dialogue between skin microbiota and immunity*. Science, 2014. **346**(6212): p. 954-959.
 106. Grice, E.A. and J.A. Segre, *The skin microbiome*. Nat Rev Microbiol, 2011. **9**(4): p. 244-253.
 107. Kong, H.H. and J.A. Segre, *Skin Microbiome: Looking Back to Move Forward*. The Journal of investigative dermatology, 2012. **132**(3 0 2): p. 933-939.
 108. Grice, E.A. and J.A. Segre, *The human microbiome: our second genome*. Annu Rev Genomics Hum Genet, 2012. **13**: p. 151-70.
 109. Oh, J., et al., *Temporal Stability of the Human Skin Microbiome*. Cell, 2016. **165**(4): p. 854-66.
 110. Oh, J., et al., *Biogeography and individuality shape function in the human skin metagenome*. Nature, 2014. **514**(7520): p. 59-64.
 111. Bitschar, K., et al., *Keratinocytes as sensors and central players in the immune defense against Staphylococcus aureus in the skin*. J Dermatol Sci, 2017. **87**(3): p. 215-220.

112. Eckhart, L., et al., *Cell death by cornification*. Biochim Biophys Acta, 2013. **1833**(12): p. 3471-3480.
113. Verdier-Sévrain, S. and F. Bonté, *Skin hydration: a review on its molecular mechanisms*. Journal of Cosmetic Dermatology, 2007. **6**(2): p. 75-82.
114. Fiedler, T., T. Koller, and B. Kreikemeyer, *Streptococcus pyogenes biofilms-formation, biology, and clinical relevance*. Front Cell Infect Microbiol, 2015. **5**: p. 15.
115. Serra, R., et al., *Chronic wound infections: the role of Pseudomonas aeruginosa and Staphylococcus aureus*. Expert Rev Anti Infect Ther, 2015. **13**(5): p. 605-13.
116. de Hoog, S., et al., *Skin Fungi from Colonization to Infection*. Microbiol Spectr, 2017. **5**(4).
117. Findley, K., et al., *Human Skin Fungal Diversity*. Nature, 2013. **498**(7454): p. 367-370.
118. Zhang, E., et al., *Characterization of the skin fungal microbiota in patients with atopic dermatitis and in healthy subjects*. Microbiol Immunol, 2011. **55**(9): p. 625-32.
119. Schroder, J.M., *Antimicrobial peptides in healthy skin and atopic dermatitis*. Allergol Int, 2011. **60**(1): p. 17-24.
120. Harder, J. and J.M. Schroder, *RNase 7, a novel innate immune defense antimicrobial protein of healthy human skin*. J Biol Chem, 2002. **277**(48): p. 46779-84.
121. Atmatzidis, D.H., W.C. Lambert, and M.W. Lambert, *Langerhans cell: exciting developments in health and disease*. J Eur Acad Dermatol Venereol, 2017.
122. Tett, A., et al., *Unexplored diversity and strain-level structure of the skin microbiome associated with psoriasis*. npj Biofilms and Microbiomes, 2017. **3**(1): p. 14.
123. Conlon, M.A. and A.R. Bird, *The impact of diet and lifestyle on gut microbiota and human health*. Nutrients, 2014. **7**(1): p. 17-44.
124. Duenas, M., et al., *A survey of modulation of gut microbiota by dietary polyphenols*. Biomed Res Int, 2015. **2015**: p. 850902.
125. Lacroix, C., T. de Wouters, and C. Chassard, *Integrated multi-scale strategies to investigate nutritional compounds and their effect on the gut microbiota*. Curr Opin Biotechnol, 2015. **32**: p. 149-55.
126. Walker, A.W., et al., *Dominant and diet-responsive groups of bacteria within the human colonic microbiota*. Isme j, 2011. **5**(2): p. 220-30.
127. Xiao, S., et al., *A gut microbiota-targeted dietary intervention for amelioration of chronic inflammation underlying metabolic syndrome*. FEMS Microbiol Ecol, 2014. **87**(2): p. 357-67.
128. Vieira, A.T., C. Fukumori, and C.M. Ferreira, *New insights into therapeutic strategies for gut microbiota modulation in inflammatory diseases*. Clin Transl Immunology, 2016. **5**(6): p. e87.
129. Markowiak, P. and K. Slizewska, *Effects of Probiotics, Prebiotics, and Synbiotics on Human Health*. Nutrients, 2017. **9**(9).
130. Hutkins, R.W., et al., *Prebiotics: why definitions matter*. Current opinion in biotechnology, 2016. **37**: p. 1-7.
131. Myles, I.A., et al., *Transplantation of human skin microbiota in models of atopic dermatitis*. JCI Insight, 2016. **1**(10).
132. Maguire, M. and G. Maguire, *The role of microbiota, and probiotics and prebiotics in skin health*. Arch Dermatol Res, 2017. **309**(6): p. 411-421.
133. Paganini, D., et al., *Prebiotic galacto-oligosaccharides mitigate the adverse effects of iron fortification on the gut microbiome: a randomised controlled study in Kenyan infants*. Gut, 2017.
134. Schloss, P.D., et al., *The dynamics of a family's gut microbiota reveal variations on a theme*. Microbiome, 2014. **2**(1): p. 25.
135. Sundquist, A., et al., *Bacterial flora-typing with targeted, chip-based Pyrosequencing*. BMC Microbiol, 2007. **7**: p. 108.
136. Shannon, P., et al., *Cytoscape: a software environment for integrated models of biomolecular interaction networks*. Genome Res, 2003. **13**(11): p. 2498-504.
137. Letunic, I. and P. Bork, *Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy*. Nucleic Acids Res, 2011. **39**(Web Server issue): p. W475-8.
138. Denis, D.J., *Applied Univariate, Bivariate, and Multivariate Statistics*. 2016.
139. Gris, K.V., J.-P. Coutu, and D. Gris, *Supervised and Unsupervised Learning Technology in the Study of Rodent Behavior*. Frontiers in Behavioral Neuroscience, 2017. **11**: p. 141.
140. Warton, D.I., S.T. Wright, and Y. Wang, *Distance-based multivariate analyses confound location and dispersion effects*. Methods in Ecology and Evolution, 2012. **3**(1): p. 89-101.
141. van den Wollenberg, A.L., *Redundancy analysis an alternative for canonical correlation analysis*. Psychometrika, 1977. **42**(2): p. 207-219.

142. Touw, W.G., et al., *Data mining in the Life Sciences with Random Forest: a walk in the park or lost in the jungle?* Brief Bioinform, 2013. **14**(3): p. 315-26.
143. Lappenschaar, M., et al., *Multilevel temporal Bayesian networks can model longitudinal change in multimorbidity.* J Clin Epidemiol, 2013. **66**(12): p. 1405-16.
144. Alves, J.M., et al., *GenSeed-HMM: A Tool for Progressive Assembly Using Profile HMMs as Seeds and its Application in Alpavirinae Viral Discovery from Metagenomic Data.* Front Microbiol, 2016. **7**: p. 269.
145. Graham, E.D., J.F. Heidelberg, and B.J. Tully, *BinSanity: unsupervised clustering of environmental microbial assemblies using coverage and affinity propagation.* PeerJ, 2017. **5**: p. e3035.
146. Bø, T.H., B. Dysvik, and I. Jonassen, *LSimpute: accurate estimation of missing values in microarray data with least squares methods.* Nucleic Acids Research, 2004. **32**(3): p. e34-e34.
147. Williamson, E.J. and A. Forbes, *Introduction to propensity scores.* Respirology, 2014. **19**(5): p. 625-635.
148. Pandis, N., *Cross-sectional studies.* Am J Orthod Dentofacial Orthop, 2014. **146**(1): p. 127-9.
149. Fairfax, B.P., et al., *Innate immune activity conditions the effect of regulatory variants upon monocyte gene expression.* Science, 2014. **343**(6175): p. 1246949.
150. Rhee, E.P., et al., *A genome-wide association study of the human metabolome in a community-based cohort.* Cell Metab, 2013. **18**(1): p. 130-43.
151. Costello, E.K., et al., *Bacterial community variation in human body habitats across space and time.* Science, 2009. **326**(5960): p. 1694-7.
152. Ding, T. and P.D. Schloss, *Dynamics and associations of microbial community types across the human body.* Nature, 2014. **509**(7500): p. 357-60.
153. Gevers, D., et al., *The treatment-naïve microbiome in new-onset Crohn's disease.* Cell Host Microbe, 2014. **15**(3): p. 382-392.
154. Liu, C., T.P. Cripe, and M.O. Kim, *Statistical issues in longitudinal data analysis for treatment efficacy studies in the biomedical sciences.* Mol Ther, 2010. **18**(9): p. 1724-30.
155. Nakai, M. and W. Ke, *Statistical Models for Longitudinal Data Analysis.* Applied Mathematical Sciences, 2009. **Vol. 3**(no. 40): p. 1979 - 1989.
156. Parab, S. and S. Bhalerao, *Choosing statistical test.* International Journal of Ayurveda Research, 2010. **1**(3): p. 187-191.
157. McHugh, M.L., *The Chi-square test of independence.* Biochemia Medica, 2013. **23**(2): p. 143-149.
158. Zou, K.H., K. Tuncali, and S.G. Silverman, *Correlation and simple linear regression.* Radiology, 2003. **227**(3): p. 617-22.
159. Everitt, B.S., *Analysis of longitudinal data. Beyond MANOVA.* Br J Psychiatry, 1998. **172**: p. 7-10.
160. Caruana, E.J., et al., *Longitudinal studies.* Journal of Thoracic Disease, 2015. **7**(11): p. E537-E540.
161. Lemire, D., *Faster retrieval with a two-pass dynamic-time-warping lower bound.* Pattern Recognition, 2009. **42**(9): p. 2169-2180.
162. Xia, L.C., et al., *Extended local similarity analysis (eLSA) of microbial community and other time series data with replicates.* BMC Syst Biol, 2011. **5 Suppl 2**: p. S15.
163. MacLean, D., J.D.G. Jones, and D.J. Studholme, *Application of 'next-generation' sequencing technologies to microbial genetics.* Nature Reviews Microbiology, 2009. **7**(4): p. 287-296.
164. Hall, N., *Advanced sequencing technologies and their wider impact in microbiology.* Journal of Experimental Biology, 2007. **210**(9): p. 1518-1525.
165. Siezen, R.J. and S. van Hijum, *Genome (re-)annotation and open-source annotation pipelines.* Microbial Biotechnology, 2010. **3**(4): p. 362-369.
166. Delcher, A.L., et al., *Improved microbial gene identification with GLIMMER.* Nucleic Acids Research, 1999. **27**(23): p. 4636-4641.
167. Delcher, A.L., et al., *Identifying bacterial genes and endosymbiont DNA with Glimmer.* Bioinformatics, 2007. **23**(6): p. 673-679.
168. Besemer, J. and M. Borodovsky, *GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses.* Nucleic Acids Research, 2005. **33**: p. W451-W454.
169. Besemer, J., A. Lomsadze, and M. Borodovsky, *GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions.* Nucleic Acids Research, 2001. **29**(12): p. 2607-2618.
170. Hyatt, D., et al., *Prodigal: prokaryotic gene recognition and translation initiation site identification.* BMC Bioinformatics, 2010. **11**: p. 119.
171. Altschul, S.F., et al., *Basic Local Alignment Search Tool.* Journal of Molecular Biology, 1990. **215**(3):

p. 403-410.

172. Pruitt, K.D., et al., *NCBI Reference Sequences: current status, policy and new initiatives*. Nucleic Acids Research, 2009. **37**: p. D32-D36.
173. Benson, D.A., et al., *GenBank*. Nucleic Acids Research, 2000. **28**(1): p. 15-18.
174. Bairoch, A., et al., *The universal protein resource (UniProt)*. Nucleic Acids Research, 2005. **33**: p. D154-D159.
175. Punta, M., et al., *The Pfam protein families database*. Nucleic Acids Research, 2012. **40**(D1): p. D290-D301.
176. Meyer, F., R. Overbeek, and A. Rodriguez, *FIGfams: yet another set of protein families*. Nucleic Acids Research, 2009. **37**(20): p. 6643-6654.
177. Bakke, P., et al., *Evaluation of Three Automated Genome Annotations for Halorhabdus utahensis*. Plos One, 2009. **4**(7).
178. Bocs, S., A. Danchin, and C. Medigue, *Re-annotation of genome microbial CoDing-Sequences: finding new genes and inaccurately annotated genes*. Bmc Bioinformatics, 2002. **3**: p. 5.
179. Brenner, S.E., *Errors in genome annotation*. Trends in Genetics, 1999. **15**(4): p. 132-133.
180. Wall, M.E., et al., *Genome Majority Vote Improves Gene Predictions*. Plos Computational Biology, 2011. **7**(11).
181. Yok, N.G. and G.L. Rosen, *Combining gene prediction methods to improve metagenomic gene annotation*. Bmc Bioinformatics, 2011. **12**: p. 20.
182. Yok, N. and G. Rosen, *Benchmarking of gene prediction programs for metagenomic data*. Conference proceedings : ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Conference, 2010. **2010**: p. 6190-3.
183. Shah, S.P., et al., *GeneComber: combining outputs of gene prediction programs for improved results*. Bioinformatics, 2003. **19**(10): p. 1296-1297.
184. Yada, T., et al., *DIGIT: a novel gene finding program by combining gene-finders*. Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing, 2003: p. 375-87.
185. Pavlovic, V., A. Garg, and S. Kasif, *A Bayesian framework for combining gene predictions*. Bioinformatics, 2002. **18**(1): p. 19-27.
186. Richardson, E.J. and M. Watson, *The automatic annotation of bacterial genomes*. Brief Bioinform, 2013. **14**(1): p. 1-12.
187. Tatusov, R.L., E.V. Koonin, and D.J. Lipman, *A genomic perspective on protein families*. Science, 1997. **278**(5338): p. 631-637.
188. Dennis, G., Jr., et al., *DAVID: Database for Annotation, Visualization, and Integrated Discovery*. Genome biology, 2003. **4**(5): p. P3.
189. Huang, D.W., B.T. Sherman, and R.A. Lempicki, *Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources*. Nature Protocols, 2009. **4**(1): p. 44-57.
190. Yu, N.Y., et al., *PSORTdb-an expanded, auto-updated, user-friendly protein subcellular localization database for Bacteria and Archaea*. Nucleic Acids Research, 2011. **39**: p. D241-D244.
191. Riley, M., et al., *Escherichia coli K-12: a cooperatively developed annotation snapshot - 2005*. Nucleic Acids Research, 2006. **34**(1): p. 1-9.
192. Barbe, V., et al., *From a consortium sequence to a unified sequence: the Bacillus subtilis 168 reference genome a decade later*. Microbiology-Sgm, 2009. **155**: p. 1758-1775.
193. Siezen, R.J., et al., *Complete Resequencing and Reannotation of the Lactobacillus plantarum WCFS1 Genome*. Journal of Bacteriology, 2012. **194**(1): p. 195-196.
194. Siezen, R.J., et al., *Genome-scale genotype-phenotype matching of two Lactococcus lactis isolates from plants identifies mechanisms of adaptation to the plant niche*. Applied and Environmental Microbiology, 2008. **74**(2): p. 424-436.
195. Tettelin, H., et al., *Complete genome sequence of a virulent isolate of Streptococcus pneumoniae*. Science, 2001. **293**(5529): p. 498-506.
196. Deng, W., et al., *Comparative genomics of Salmonella enterica serovar typhi strains Ty2 and CT18*. Journal of Bacteriology, 2003. **185**(7): p. 2330-2337.
197. Tettelin, H., et al., *Complete genome sequence of Neisseria meningitidis serogroup B strain MC58*. Science, 2000. **287**(5459): p. 1809-1815.
198. Fleischmann, R.D., et al., *Whole-Genome Random Sequencing and Assembly of Haemophilus Influenzae RD*. Science, 1995. **269**(5223): p. 496-512.
199. Cole, S.T., et al., *Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence (vol 393, pg 537, 1998)*. Nature, 1998. **396**(6707): p. 190-198.

200. Jaffe, J.D., et al., *The complete genome and proteome of Mycoplasma mobile*. Genome Research, 2004. **14**(8): p. 1447-1461.
201. Nelson, K.E., et al., *Complete genome sequence and comparative analysis of the metabolically versatile Pseudomonas putida KT2440*. Environmental Microbiology, 2002. **4**(12): p. 799-808.
202. Redenbach, M., et al., *A set of ordered cosmids and a detailed genetic and physical map for the 8 Mb Streptomyces coelicolor A3(2) chromosome*. Molecular Microbiology, 1996. **21**(1): p. 77-96.
203. van Domselaar, G.H., et al., *BASys: a web server for automated bacterial genome annotation*. Nucleic Acids Research, 2005. **33**: p. W455-W459.
204. Hemmerich, C., et al., *An Ergatis-based prokaryotic genome annotation web server*. Bioinformatics, 2010. **26**(8): p. 1122-1124.
205. Aziz, R.K., et al., *The RAST server: Rapid annotations using subsystems technology*. BMC Genomics, 2008. **9**: p. 75.
206. Chaudhuri, R.R. and M.J. Pallen, *xBASE, a collection of online databases for bacterial comparative genomics*. Nucleic Acids Research, 2006. **34**: p. D335-D337.
207. Chaudhuri, R.R., et al., *xBASE2: a comprehensive resource for comparative bacterial genomics*. Nucleic Acids Research, 2008. **36**: p. D543-D546.
208. Belkaid, Y. and T. Hand, *Role of the Microbiota in Immunity and inflammation*. Cell, 2014. **157**(1): p. 121-141.
209. Pechar, R., et al., *Bifidobacterium apri sp. nov., a thermophilic actinobacterium isolated from the digestive tract of wild pigs (Sus scrofa)*. Int J Syst Evol Microbiol, 2017. **67**(7): p. 2349-2356.
210. Wei, S., et al., *Molecular discrimination of Bacillus cereus group species in foods (lettuce, spinach, and kimba) using quantitative real-time PCR targeting groEL and gyrB*. Microb Pathog, 2018. **115**: p. 312-320.
211. Bulane, A. and A. Hoosen, *Use of matrix-assisted laser desorption/ionisation-time of flight mass spectrometry analyser in a diagnostic microbiology laboratory in a developing country*. Afr J Lab Med, 2017. **6**(1): p. 598.
212. Wels, M., et al., *Draft Genome Sequences of 11 Lactococcus lactis subsp. cremoris Strains*. Genome Announc, 2017. **5**(11).
213. Adamiak, P., et al., *Effectiveness of the standard and an alternative set of Streptococcus pneumoniae multi locus sequence typing primers*. BMC Microbiology, 2014. **14**: p. 143-143.
214. Quainoo, S., et al., *Whole-Genome Sequencing of Bacterial Pathogens: the Future of Nosocomial Outbreak Analysis*. Clin Microbiol Rev, 2017. **30**(4): p. 1015-1063.
215. Ranjan, R., et al., *Analysis of the microbiome: Advantages of whole genome shotgun versus 16S amplicon sequencing*. Biochemical and Biophysical Research Communications, 2016. **469**(4): p. 967-977.
216. Sanschagrin, S. and E. Yergeau, *Next-generation Sequencing of 16S Ribosomal RNA Gene Amplicons*. Journal of Visualized Experiments : JoVE, 2014(90): p. 51709.
217. Scholz, C.F. and A. Jensen, *Development of a Single Locus Sequence Typing (SLST) Scheme for Typing Bacterial Species Directly from Complex Communities*. Methods Mol Biol, 2017. **1535**: p. 97-107.
218. Leyden, J.J., R.R. Marples, and A.M. Kligman, *Staphylococcus aureus in the lesions of atopic dermatitis*. Br J Dermatol, 1974. **90**(5): p. 525-30.
219. Higaki, S., et al., *Comparative study of staphylococci from the skin of atopic dermatitis patients and from healthy subjects*. Int J Dermatol, 1999. **38**(4): p. 265-9.
220. Kong, H.H., et al., *Temporal shifts in the skin microbiome associated with disease flares and treatment in children with atopic dermatitis*. Genome Res, 2012. **22**(5): p. 850-9.
221. Chng, K.R., et al., *Whole metagenome profiling reveals skin microbiome-dependent susceptibility to atopic dermatitis flare*. Nat Microbiol, 2016. **1**(9): p. 16106.
222. Kloos, W.E. and M.S. Musselwhite, *Distribution and persistence of Staphylococcus and Micrococcus species and other aerobic bacteria on human skin*. Appl Microbiol, 1975. **30**(3): p. 381-5.
223. Zhang, J., et al., *PEAR: a fast and accurate Illumina Paired-End reAd mergeR*. Bioinformatics, 2014. **30**(5): p. 614-20.
224. Letunic, I. and P. Bork, *Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees*. Nucleic Acids Res, 2016. **44**(W1): p. W242-5.
225. Wang, Q., et al., *Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy*. Appl Environ Microbiol, 2007. **73**(16): p. 5261-7.
226. Kitts, P.A., et al., *Assembly: a resource for assembled genomes at NCBI*. Nucleic Acids Res, 2016.

44(D1): p. D73-80.

227. Rice, P., I. Longden, and A. Bleasby, *EMBOSS: the European Molecular Biology Open Software Suite*. Trends Genet, 2000. **16**(6): p. 276-7.
228. Braak, C.J.F.t. and P. Smilauer, *Canoco reference manual and user's guide: software for ordination, version 5.0*. 2012, Ithaca USA: Microcomputer Power.
229. Leinonen, R., et al., *The European Nucleotide Archive*. Nucleic Acids Res, 2011. **39**(Database issue): p. D28-31.
230. Price, M.N., P.S. Dehal, and A.P. Arkin, *FastTree: computing large minimum evolution trees with profiles instead of a distance matrix*. Mol Biol Evol, 2009. **26**(7): p. 1641-50.
231. Human Microbiome Project, C., *A framework for human microbiome research*. Nature, 2012. **486**(7402): p. 215-21.
232. Human Microbiome Project, C., *Structure, function and diversity of the healthy human microbiome*. Nature, 2012. **486**(7402): p. 207-14.
233. Blaser, M., *Antibiotic overuse: Stop the killing of beneficial bacteria*. Nature, 2011. **476**(7361): p. 393-4.
234. Sanford, J.A. and R.L. Gallo, *Functions of the skin microbiota in health and disease*. Semin Immunol, 2013. **25**(5): p. 370-7.
235. Findley, K. and E.A. Grice, *The Skin Microbiome: A Focus on Pathogens and Their Association with Skin Disease*. PLoS Pathog, 2014. **10**(11): p. e1004436.
236. Ursell, L.K., et al., *The interpersonal and intrapersonal diversity of human-associated microbiota in key body sites*. J Allergy Clin Immunol, 2012. **129**(5): p. 1204-8.
237. Rosenthal, M., et al., *Skin microbiota: microbial community structure and its potential association with health and disease*. Infect Genet Evol, 2011. **11**(5): p. 839-848.
238. Zeeuwen, P.L., et al., *Microbiome and skin diseases*. Curr Opin Allergy Clin Immunol, 2013. **13**(5): p. 514-20.
239. Kong, H.H., et al., *Temporal shifts in the skin microbiome associated with disease flares and treatment in children with atopic dermatitis*. Genome Res, 2012. **22**(5): p. 850-9.
240. Gallo, R.L. and T. Nakatsuji, *Microbial symbiosis with the innate immune defense system of the skin*. J Invest Dermatol, 2011. **131**(10): p. 1974-1980.
241. Smeekens, S.P., et al., *Skin microbiome imbalance in patients with STAT1/STAT3 defects impairs innate host defense responses*. J Innate Immun, 2014. **6**(3): p. 253-62.
242. Wanke, I., et al., *Skin commensals amplify the innate immune response to pathogens by activation of distinct signaling pathways*. J Invest Dermatol, 2011. **131**(2): p. 382-390.
243. Naik, S., et al., *Compartmentalized control of skin immunity by resident commensals*. Science, 2012. **337**(6098): p. 1115-9.
244. Lai, Y., et al., *Commensal bacteria regulate Toll-like receptor 3-dependent inflammation after skin injury*. Nat Med, 2009. **15**(12): p. 1377-82.
245. Cho, I. and M.J. Blaser, *The human microbiome: at the interface of health and disease*. Nat Rev Genet, 2012. **13**(4): p. 260-270.
246. Gao, Z., et al., *Substantial alterations of the cutaneous bacterial biota in psoriatic lesions*. Plos One, 2008. **3**(7): p. e2719.
247. Fahlen, A., et al., *Comparison of bacterial microbiota in skin biopsies from normal and psoriatic skin*. Arch Dermatol Res, 2012. **304**(1): p. 15-22.
248. Edelblute, C.M., et al., *Human platelet gel supernatant inactivates opportunistic wound pathogens on skin*. Platelets, 2015. **26**(1): p. 13-16.
249. Nataraj, N., et al., *Synthesis and anti-staphylococcal activity of TiO₂ nanoparticles and nanowires in ex vivo porcine skin model*. J Biomed Nanotechnol, 2014. **10**(5): p. 864-70.
250. Abdelaziz, A.A., et al., *Optimization of niosomes for enhanced antibacterial activity and reduced bacterial resistance: in vitro and in vivo evaluation*. Expert Opin Drug Deliv, 2015. **12**(2): p. 163-180.
251. van Drongelen, V., et al., *Reduced filaggrin expression is accompanied by increased Staphylococcus aureus colonization of epidermal skin models*. Clin Exp Allergy, 2014. **44**(12): p. 1515-1524.
252. Popov, L., et al., *Three-dimensional human skin models to understand Staphylococcus aureus skin colonization and infection*. Front Immunol, 2014. **5**.
253. de Breij, A., et al., *Three-dimensional human skin equivalent as a tool to study Acinetobacter baumannii colonization*. Antimicrob Agents Chemother, 2012. **56**(5): p. 2459-64.
254. Duckney, P., et al., *The role of the skin barrier in modulating the effects of common skin microbial*

- species on the inflammation, differentiation and proliferation status of epidermal keratinocytes*. BMC Res Notes, 2013. **6**: p. 474.
255. Naik, S., et al., *Commensal-dendritic-cell interaction specifies a unique protective skin immune signature*. Nature, 2015.
 256. Holland, D.B., et al., *Microbial colonization of an in vitro model of a tissue engineered human skin equivalent--a novel approach*. FEMS Microbiol Lett, 2008. **279**(1): p. 110-5.
 257. Nocker, A., C.Y. Cheung, and A.K. Camper, *Comparison of propidium monoazide with ethidium monoazide for differentiation of live vs. dead bacteria by selective removal of DNA from dead cells*. J Microbiol Methods, 2006. **67**(2): p. 310-20.
 258. Grice, E.A., et al., *Topographical and temporal diversity of the human skin microbiome*. Science, 2009. **324**(5931): p. 1190-2.
 259. Grice, E.A., et al., *A diversity profile of the human skin microbiota*. Genome Res, 2008. **18**(7): p. 1043-1050.
 260. Zeeuwen, P.L., et al., *The cystatin M/E-cathepsin L balance is essential for tissue homeostasis in epidermis, hair follicles, and cornea*. FASEB J, 2010. **24**(10): p. 3744-3755.
 261. Blekhman, R., et al., *Host genetic variation impacts microbiome composition across human body sites*. Genome Biol, 2015. **16**: p. 191.
 262. Smith, F.J., et al., *Loss-of-function mutations in the gene encoding filaggrin cause ichthyosis vulgaris*. Nat Genet, 2006. **38**(3): p. 337-42.
 263. Palmer, C.N., et al., *Common loss-of-function variants of the epidermal barrier protein filaggrin are a major predisposing factor for atopic dermatitis*. Nat Genet, 2006. **38**(4): p. 441-6.
 264. Manabe, M., et al., *Interaction of filaggrin with keratin filaments during advanced stages of normal human epidermal differentiation and in ichthyosis vulgaris*. Differentiation, 1991. **48**(1): p. 43-50.
 265. Harding, C.R., S. Aho, and C.A. Bosko, *Filaggrin - revisited*. Int J Cosmet Sci, 2013. **35**(5): p. 412-23.
 266. Irvine, A.D., W.H. McLean, and D.Y. Leung, *Filaggrin mutations associated with skin and allergic diseases*. N Engl J Med, 2011. **365**(14): p. 1315-27.
 267. Gruber, R., et al., *Filaggrin genotype in ichthyosis vulgaris predicts abnormalities in epidermal structure and function*. Am J Pathol, 2011. **178**(5): p. 2252-63.
 268. Murphy, E.C. and I.M. Frick, *Gram-positive anaerobic cocci--commensals and opportunistic pathogens*. FEMS Microbiol Rev, 2013. **37**(4): p. 520-53.
 269. Kezic, S., et al., *Loss-of-function mutations in the filaggrin gene lead to reduced level of natural moisturizing factor in the stratum corneum*. J Invest Dermatol, 2008. **128**(8): p. 2117-9.
 270. van der Krieken, D.A., et al., *An In vitro Model for Bacterial Growth on Human Stratum Corneum*. Acta Derm Venereol, 2016. **96**(7): p. 873-879.
 271. Kanehisa, M., et al., *The KEGG databases at GenomeNet*. Nucleic Acids Res, 2002. **30**(1): p. 42-6.
 272. Kezic, S., et al., *Levels of filaggrin degradation products are influenced by both filaggrin genotype and atopic dermatitis severity*. Allergy, 2011. **66**(7): p. 934-40.
 273. Bender, R.A., *Regulation of the histidine utilization (hut) system in bacteria*. Microbiol Mol Biol Rev, 2012. **76**(3): p. 565-84.
 274. Knights, D., et al., *Complex host genetics influence the microbiome in inflammatory bowel disease*. Genome Med, 2014. **6**(12): p. 107.
 275. Tong, M., et al., *Reprogramming of gut microbiome energy metabolism by the FUT2 Crohn's disease risk polymorphism*. ISME J, 2014. **8**(11): p. 2193-206.
 276. Kubica, M., et al., *The skin microbiome of caspase-14-deficient mice shows mild dysbiosis*. Exp Dermatol, 2014. **23**(8): p. 561-7.
 277. Denecker, G., et al., *Caspase-14 protects against epidermal UVB photodamage and water loss*. Nat Cell Biol, 2007. **9**(6): p. 666-74.
 278. Oh, J., et al., *The altered landscape of the human skin microbiome in patients with primary immunodeficiencies*. Genome Res, 2013. **23**(12): p. 2103-14.
 279. O'Regan, G.M., et al., *Filaggrin in atopic dermatitis*. J Allergy Clin Immunol, 2008. **122**(4): p. 689-93.
 280. Leung, D.Y., *Our evolving understanding of the functional role of filaggrin in atopic dermatitis*. J Allergy Clin Immunol, 2009. **124**(3): p. 494-5.
 281. Iwase, T., et al., *Staphylococcus epidermidis Esp inhibits Staphylococcus aureus biofilm formation and nasal colonization*. Nature, 2010. **465**(7296): p. 346-9.
 282. Shu, M., et al., *Fermentation of Propionibacterium acnes, a commensal bacterium in the human skin*

microbiome, as skin probiotics against methicillin-resistant *Staphylococcus aureus*. Plos One, 2013. **8**(2): p. e55380.

283. van Rensburg, J.J., et al., *The Human Skin Microbiome Associates with the Outcome of and Is Influenced by Bacterial Infection*. MBio, 2015. **6**(5): p. e01315-15.
284. Donia, M.S. and M.A. Fischbach, *HUMAN MICROBIOTA. Small molecules from the human microbiota*. Science, 2015. **349**(6246): p. 1254766.
285. van den Bogaard, E.H., et al., *Coal tar induces AHR-dependent skin barrier repair in atopic dermatitis*. J Clin Invest, 2013. **123**(2): p. 917-27.
286. Edgar, R.C., et al., *UCHIME improves sensitivity and speed of chimera detection*. Bioinformatics, 2011. **27**(16): p. 2194-200.
287. Cole, J.R., et al., *The Ribosomal Database Project: improved alignments and new tools for rRNA analysis*. Nucleic Acids Res, 2009. **37**(Database issue): p. D141-5.
288. Rogiers, V. and E. Group, *EEMCO guidance for the assessment of transepidermal water loss in cosmetic sciences*. Skin Pharmacol Appl Skin Physiol, 2001. **14**(2): p. 117-28.
289. Parra, J.L., M. Paye, and E. Group, *EEMCO guidance for the in vivo assessment of skin surface pH*. Skin Pharmacol Appl Skin Physiol, 2003. **16**(3): p. 188-202.
290. Berardesca, E., C. European Group for Efficacy Measurements on, and P. Other Topical, *EEMCO guidance for the assessment of stratum corneum hydration: electrical methods*. Skin Res Technol, 1997. **3**(2): p. 126-32.
291. Dapic, I., et al., *Evaluation of an HPLC Method for the Determination of Natural Moisturizing Factors in the Human Stratum Corneum*. Analytical Letters, 2013. **46**(14): p. 2133-2144.
292. Nishida, K., et al., *KEGGscape: a Cytoscape app for pathway data integration*. F1000Res, 2014. **3**: p. 144.
293. Rheinwald, J.G. and H. Green, *Serial cultivation of strains of human epidermal keratinocytes: the formation of keratinizing colonies from single cells*. Cell, 1975. **6**(3): p. 331-43.
294. Nygaard, U.H., et al., *Antibiotics in cell culture: friend or foe? Suppression of keratinocyte growth and differentiation in monolayer cultures and 3D skin models*. Exp Dermatol, 2015. **24**(12): p. 964-5.
295. Bergboer, J.G., et al., *Psoriasis risk genes of the late cornified envelope-3 group are distinctly expressed compared with genes of other LCE groups*. Am J Pathol, 2011. **178**(4): p. 1470-7.
296. Zeeuwen, P.L., et al., *Genetically programmed differences in epidermal host defense between psoriasis and atopic dermatitis patients*. Plos One, 2008. **3**(6): p. e2301.
297. Livak, K.J. and T.D. Schmittgen, *Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method*. Methods, 2001. **25**(4): p. 402-8.
298. Pakseresht, N., et al., *Assembly information services in the European Nucleotide Archive*. Nucleic Acids Res, 2014. **42**(Database issue): p. D38-43.
299. Joosten, L.A., M.G. Netea, and C.A. Dinarello, *Interleukin-1beta in innate inflammation, autophagy and immunity*. Semin Immunol, 2013. **25**(6): p. 416-24.
300. Carter, D.B., et al., *Purification, cloning, expression and biological characterization of an interleukin-1 receptor antagonist protein*. Nature, 1990. **344**(6267): p. 633-8.
301. Horai, R., et al., *Development of chronic inflammatory arthropathy resembling rheumatoid arthritis in interleukin 1 receptor antagonist-deficient mice*. J Exp Med, 2000. **191**(2): p. 313-20.
302. Nakae, S., et al., *IL-17 production from activated T cells is required for the spontaneous development of destructive arthritis in mice deficient in IL-1 receptor antagonist*. Proc Natl Acad Sci U S A, 2003. **100**(10): p. 5986-90.
303. Shepherd, J., M.C. Little, and M.J. Nicklin, *Psoriasis-like cutaneous inflammation in mice lacking interleukin-1 receptor antagonist*. J Invest Dermatol, 2004. **122**(3): p. 665-9.
304. Matsuki, T., et al., *IL-1 plays an important role in lipid metabolism by regulating insulin levels under physiological conditions*. J Exp Med, 2003. **198**(6): p. 877-88.
305. Matsuki, T., et al., *Abnormal T cell activation caused by the imbalance of the IL-1/IL-1R antagonist system is responsible for the development of experimental autoimmune encephalomyelitis*. Int Immunol, 2006. **18**(2): p. 399-407.
306. Koenders, M.I., et al., *Interleukin-1 drives pathogenic Th17 cells during spontaneous arthritis in interleukin-1 receptor antagonist-deficient mice*. Arthritis Rheum, 2008. **58**(11): p. 3461-70.
307. Ochoa-Reparaz, J., et al., *Role of gut commensal microflora in the development of experimental autoimmune encephalomyelitis*. J Immunol, 2009. **183**(10): p. 6041-50.
308. Berer, K., et al., *Commensal microbiota and myelin autoantigen cooperate to trigger autoimmune*

- demyelination. *Nature*, 2011. **479**(7374): p. 538-41.
309. Kriegel, M.A., et al., *Naturally transmitted segmented filamentous bacteria segregate with diabetes protection in nonobese diabetic mice*. *Proc Natl Acad Sci U S A*, 2011. **108**(28): p. 11548-53.
 310. Abdollahi-Roodsaz, S., et al., *Stimulation of TLR2 and TLR4 differentially skews the balance of T cells in a mouse model of arthritis*. *J Clin Invest*, 2008. **118**(1): p. 205-16.
 311. Wu, H.J., et al., *Gut-residing segmented filamentous bacteria drive autoimmune arthritis via T helper 17 cells*. *Immunity*, 2010. **32**(6): p. 815-27.
 312. Scher, J.U., et al., *Expansion of intestinal Prevotella copri correlates with enhanced susceptibility to arthritis*. *Elife*, 2013. **2**: p. e01202.
 313. Zhang, X., et al., *The oral and gut microbiomes are perturbed in rheumatoid arthritis and partly normalized after treatment*. *Nat Med*, 2015. **21**(8): p. 895-905.
 314. Scher, J.U., et al., *Decreased bacterial diversity characterizes the altered gut microbiota in patients with psoriatic arthritis, resembling dysbiosis in inflammatory bowel disease*. *Arthritis Rheumatol*, 2015. **67**(1): p. 128-39.
 315. Chen, J., et al., *An expansion of rare lineage intestinal microbes characterizes rheumatoid arthritis*. *Genome Med*, 2016. **8**(1): p. 43.
 316. Maeda, Y., et al., *Dysbiosis contributes to arthritis development via activation of autoreactive T cells in the intestine*. *Arthritis Rheumatol*, 2016.
 317. Hooper, L.V., D.R. Littman, and A.J. Macpherson, *Interactions between the microbiota and the immune system*. *Science*, 2012. **336**(6086): p. 1268-73.
 318. Atarashi, K., et al., *Treg induction by a rationally selected mixture of Clostridia strains from the human microbiota*. *Nature*, 2013. **500**(7461): p. 232-6.
 319. Round, J.L., et al., *The Toll-like receptor 2 pathway establishes colonization by a commensal of the human microbiota*. *Science*, 2011. **332**(6032): p. 974-7.
 320. Gaboriau-Routhiau, V., et al., *The key role of segmented filamentous bacteria in the coordinated maturation of gut helper T cell responses*. *Immunity*, 2009. **31**(4): p. 677-89.
 321. van den Berg, W.B. and I.B. McInnes, *Th17 cells and IL-17 α -focus on immunopathogenesis and immunotherapeutics*. *Semin Arthritis Rheum*, 2013. **43**(2): p. 158-70.
 322. Lubberts, E., *The IL-23-IL-17 axis in inflammatory arthritis*. *Nat Rev Rheumatol*, 2015. **11**(7): p. 415-29.
 323. Genovese, M.C., et al., *A phase II randomized study of subcutaneous ixekizumab, an anti-interleukin-17 monoclonal antibody, in rheumatoid arthritis patients who were naive to biologic agents or had an inadequate response to tumor necrosis factor inhibitors*. *Arthritis Rheumatol*, 2014. **66**(7): p. 1693-704.
 324. Genovese, M.C., et al., *One-year efficacy and safety results of secukinumab in patients with rheumatoid arthritis: phase II, dose-finding, double-blind, randomized, placebo-controlled study*. *J Rheumatol*, 2014. **41**(3): p. 414-21.
 325. Burmester, G.R., et al., *Association of HLA-DRB1 alleles with clinical responses to the anti-interleukin-17A monoclonal antibody secukinumab in active rheumatoid arthritis*. *Rheumatology (Oxford)*, 2016. **55**(1): p. 49-55.
 326. Block, K.E., et al., *Gut Microbiota Regulates K/BxN Autoimmune Arthritis through Follicular Helper T but Not Th17 Cells*. *J Immunol*, 2016. **196**(4): p. 1550-7.
 327. Sczesnak, A., et al., *The genome of th17 cell-inducing segmented filamentous bacteria reveals extensive auxotrophy and adaptations to the intestinal environment*. *Cell Host Microbe*, 2011. **10**(3): p. 260-72.
 328. Ivanov, I.I., et al., *Specific microbiota direct the differentiation of IL-17-producing T-helper cells in the mucosa of the small intestine*. *Cell Host Microbe*, 2008. **4**(4): p. 337-49.
 329. Takeuchi, O., et al., *Differential roles of TLR2 and TLR4 in recognition of gram-negative and gram-positive bacterial cell wall components*. *Immunity*, 1999. **11**(4): p. 443-51.
 330. Zuniga, L.A., et al., *Th17 cell development: from the cradle to the grave*. *Immunol Rev*, 2013. **252**(1): p. 78-88.
 331. Giongo, A., et al., *Toward defining the autoimmune microbiome for type 1 diabetes*. *ISME J*, 2011. **5**(1): p. 82-91.
 332. Joosten, L.A., et al., *T cell dependence of chronic destructive murine arthritis induced by repeated local activation of Toll-like receptor-driven pathways: crucial role of both interleukin-1 β and interleukin-17*. *Arthritis Rheum*, 2008. **58**(1): p. 98-108.

333. Abdollahi-Roodsaz, S., et al., *Shift from toll-like receptor 2 (TLR-2) toward TLR-4 dependency in the erosive stage of chronic streptococcal cell wall arthritis coincident with TLR-4-mediated interleukin-17 production*. Arthritis Rheum, 2008. **58**(12): p. 3753-64.
334. Kullberg, M.C., et al., *IL-23 plays a key role in Helicobacter hepaticus-induced T cell-dependent colitis*. J Exp Med, 2006. **203**(11): p. 2485-94.
335. Gomez, A., et al., *Loss of sex and age driven differences in the gut microbiome characterize arthritis-susceptible 0401 mice but not arthritis-resistant 0402 mice*. Plos One, 2012. **7**(4): p. e36095.
336. Luckey, et al. (2012) *Commensal Gut-Derived Bacteria As Therapy for Systemic Autoimmune Disease*. Abstracts of the American College of Rheumatology/Association of Rheumatology Health Professionals Annual Scientific Meeting **64**, DOI: DOI: 10.1002/art.39368.
337. Akitsu, A., et al., *IL-1 receptor antagonist-deficient mice develop autoimmune arthritis due to intrinsic activation of IL-17-producing CCR2(+)Vgamma6(+)gamma delta T cells*. Nat Commun, 2015. **6**: p. 7464.
338. Chappert, P., et al., *Specific gut commensal flora locally alters T cell tuning to endogenous ligands*. Immunity, 2013. **38**(6): p. 1198-210.
339. Shaw, M.H., et al., *Microbiota-induced IL-1beta, but not IL-6, is critical for the development of steady-state Th17 cells in the intestine*. J Exp Med, 2012. **209**(2): p. 251-8.
340. Chung, Y., et al., *Critical regulation of early Th17 cell differentiation by interleukin-1 signaling*. Immunity, 2009. **30**(4): p. 576-87.
341. Chevalier, N., et al., *Avenues to autoimmune arthritis triggered by diverse remote inflammatory challenges*. J Autoimmun, 2016. **73**: p. 120-9.
342. Ogino, T., et al., *Increased Th17-inducing activity of CD14+ CD163 low myeloid cells in intestinal lamina propria of patients with Crohn's disease*. Gastroenterology, 2013. **145**(6): p. 1380-91.e1.
343. Komai-Koma, M., et al., *Anti-Toll-like receptor 2 and 4 antibodies suppress inflammatory response in mice*. Immunology, 2014. **143**(3): p. 354-62.
344. van den Brand, B.T., et al., *Toll-like receptor 4 in bone marrow-derived cells as well as tissue-resident cells participate in aggravating autoimmune destructive arthritis*. Ann Rheum Dis, 2013. **72**(8): p. 1407-15.
345. Pierer, M., et al., *Toll-like receptor 4 is involved in inflammatory and joint destructive pathways in collagen-induced arthritis in DBA1J mice*. Plos One, 2011. **6**(8): p. e23539.
346. Huang, Q., et al., *Increased macrophage activation mediated through toll-like receptors in rheumatoid arthritis*. Arthritis Rheum, 2007. **56**(7): p. 2192-201.
347. Roelofs, M.F., et al., *The expression of toll-like receptors 3 and 7 in rheumatoid arthritis synovium is increased and costimulation of toll-like receptors 3, 4, and 7/8 results in synergistic cytokine production by dendritic cells*. Arthritis Rheum, 2005. **52**(8): p. 2313-22.
348. Midwood, K., et al., *Tenascin-C is an endogenous activator of Toll-like receptor 4 that is essential for maintaining inflammation in arthritic joint disease*. Nat Med, 2009. **15**(7): p. 774-80.
349. Abdollahi-Roodsaz, S., F.A. van de Loo, and W.B. van den Berg, *Trapped in a vicious loop: Toll-like receptors sustain the spontaneous cytokine production by rheumatoid synovium*. Arthritis Res Ther, 2011. **13**(2): p. 105.
350. Nicklin, M.J., et al., *Arterial inflammation in mice lacking the interleukin 1 receptor antagonist gene*. J Exp Med, 2000. **191**(2): p. 303-12.
351. Andersson, A.F., et al., *Comparative analysis of human gut microbiota by barcoded pyrosequencing*. Plos One, 2008. **3**(7): p. e2836.
352. Haas, B.J., et al., *Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons*. Genome Res, 2011. **21**(3): p. 494-504.
353. QIIME. *New default parameters for uclust OTU pickers*. 2010 December 17, 2010 [cited 2016 May 25]; Available from: <https://qiime.wordpress.com/2010/12/17/new-default-parameters-for-uclust-otu-pickers/>.
354. SciPy.org. *SciPy open-source software*. 2012 25.05.16; Available from: <http://www.scipy.org/>.
355. Barman, M., et al., *Enteric salmonellosis disrupts the microbial ecology of the murine gastrointestinal tract*. Infect Immun, 2008. **76**(3): p. 907-15.
356. Faraone, S.V., et al., *Attention-deficit/hyperactivity disorder*. Nature reviews disease primers, 2015: p. 15020.
357. Faraone, S.V. and S.J. Glatt, *A comparison of the efficacy of medications for adult attention-deficit/hyperactivity disorder using meta-analysis of effect sizes*. J Clin Psychiatry, 2010. **71**(6): p. 754-63.

358. Knutson, B. and S.E. Gibbs, *Linking nucleus accumbens dopamine and blood oxygenation*. Psychopharmacology (Berl), 2007. **191**(3): p. 813-22.
359. Scheres, A., et al., *Ventral striatal hypo-responsiveness during reward anticipation in attention-deficit/hyperactivity disorder*. Biol Psychiatry, 2007. **61**(5): p. 720-4.
360. Stroble, A., et al., *Reward anticipation and outcomes in adult males with attention-deficit/hyperactivity disorder*. Neuroimage, 2008. **39**(3): p. 966-72.
361. Plichta, M.M. and A. Scheres, *Ventral-striatal responsiveness during reward anticipation in ADHD and its relation to trait impulsivity in the healthy population: a meta-analytic review of the fMRI literature*. Neurosci Biobehav Rev, 2014. **38**: p. 125-34.
362. Hoogman, M., et al., *Nitric oxide synthase genotype modulation of impulsivity and ventral striatal activity in adult ADHD patients and healthy comparison subjects*. Am J Psychiatry, 2011. **168**(10): p. 1099-106.
363. Faraone, S.V., et al., *Molecular genetics of attention-deficit/hyperactivity disorder*. Biol Psychiatry, 2005. **57**(11): p. 1313-23.
364. Franke, B., et al., *The genetics of attention deficit/hyperactivity disorder in adults, a review*. Mol Psychiatry, 2012. **17**(10): p. 960-87.
365. Nigg, J.T., et al., *Meta-analysis of attention-deficit/hyperactivity disorder or attention-deficit/hyperactivity disorder symptoms, restriction diet, and synthetic food color additives*. J Am Acad Child Adolesc Psychiatry, 2012. **51**(1): p. 86-97 e8.
366. Sonuga-Barke, E.J., et al., *Nonpharmacological interventions for ADHD: systematic review and meta-analyses of randomized controlled trials of dietary and psychological treatments*. Am J Psychiatry, 2013. **170**(3): p. 275-89.
367. Pelsser, L.M., et al., *Effects of a restricted elimination diet on the behaviour of children with attention-deficit hyperactivity disorder (INCA study): a randomised controlled trial*. Lancet, 2011. **377**(9764): p. 494-503.
368. David, L.A., et al., *Diet rapidly and reproducibly alters the human gut microbiome*. Nature, 2014. **505**(7484): p. 559-63.
369. Cryan, J.F. and T.G. Dinan, *Mind-altering microorganisms: the impact of the gut microbiota on brain and behaviour*. Nat Rev Neurosci, 2012. **13**(10): p. 701-12.
370. Lyte, M., *Microbial endocrinology in the microbiome-gut-brain axis: how bacterial production and utilization of neurochemicals influence behavior*. PLoS Pathog, 2013. **9**(11): p. e1003726.
371. Desbonnet, L., et al., *The probiotic Bifidobacteria infantis: An assessment of potential antidepressant properties in the rat*. J Psychiatr Res, 2008. **43**(2): p. 164-74.
372. Clayton, T.A., *Metabolic differences underlying two distinct rat urinary phenotypes, a suggested role for gut microbial metabolism of phenylalanine and a possible connection to autism*. FEBS Lett, 2012. **586**(7): p. 956-61.
373. Gertsman, I., et al., *Perturbations of tyrosine metabolism promote the indolepyruvate pathway via tryptophan in host and microbiome*. Mol Genet Metab, 2015. **114**(3): p. 431-7.
374. Partty, A., et al., *A possible link between early probiotic intervention and the risk of neuropsychiatric disorders later in childhood: a randomized trial*. Pediatr Res, 2015. **77**(6): p. 823-8.
375. Kandel, E.R., J.H. Schwartz, and T.M. Jessell, *Principles of neural science*. 2000: McGraw-Hill companies
376. Ottman, N., et al., *The function of our microbiota: who is out there and what do they do?* Front Cell Infect Microbiol, 2012. **2**: p. 104.
377. Vuong, H.E. and E.Y. Hsiao, *Emerging Roles for the Gut Microbiome in Autism Spectrum Disorder*. Biol Psychiatry, 2017. **81**(5): p. 411-423.
378. Medicine, I.o., *Dietary Reference Intakes: The Essential Guide to Nutrient Requirements*, ed. J.J. Otten, J.P. Hellwig, and L.D. Meyers. 2006, Washington, DC: The National Academies Press. 1344.
379. Biederman, J. and T. Spencer, *Attention-deficit/hyperactivity disorder (ADHD) as a noradrenergic disorder*. Biol Psychiatry, 1999. **46**(9): p. 1234-42.
380. Staller, J.A. and S.V. Faraone, *Targeting the dopamine system in the treatment of attention-deficit/hyperactivity disorder*. Expert Rev Neurother, 2007. **7**(4): p. 351-62.
381. Gizer, I.R., C. Ficks, and I.D. Waldman, *Candidate gene studies of ADHD: a meta-analytic review*. Hum Genet, 2009. **126**(1): p. 51-90.
382. Aarts, E., et al., *Reward modulation of cognitive function in adult attention-deficit/hyperactivity disorder: a pilot study on the role of striatal dopamine*. Behav Pharmacol, 2015. **26**(1 and 2): p. 227-

240.

383. Antshel, K.M. and S.E. Waisbren, *Developmental timing of exposure to elevated levels of phenylalanine is associated with ADHD symptom expression*. J Abnorm Child Psychol, 2003. **31**(6): p. 565-74.
384. Baker, G.B., et al., *Phenylethylaminergic mechanisms in attention-deficit disorder*. Biol Psychiatry, 1991. **29**(1): p. 15-22.
385. Bornstein, R.A., et al., *Plasma amino acids in attention deficit disorder*. Psychiatry Res, 1990. **33**(3): p. 301-6.
386. Bergwerff, C.E., et al., *No Tryptophan, Tyrosine and Phenylalanine Abnormalities in Children with Attention-Deficit/Hyperactivity Disorder*. Plos One, 2016. **11**(3): p. e0151100.
387. Stevenson, M. and N. McNaughton, *A comparison of phenylketonuria with attention deficit hyperactivity disorder: do markedly different aetiologies deliver common phenotypes?* Brain Res Bull, 2013. **99**: p. 63-83.
388. Broadley, K.J., *The vascular effects of trace amines and amphetamines*. Pharmacol Ther, 2010. **125**(3): p. 363-75.
389. Fernstrom, J.D., *Large neutral amino acids: dietary effects on brain neurochemistry and function*. Amino Acids, 2013. **45**(3): p. 419-30.
390. Biederman, J., E. Mick, and S.V. Faraone, *Age-dependent decline of symptoms of attention deficit hyperactivity disorder: impact of remission definition and symptom type*. Am J Psychiatry, 2000. **157**(5): p. 816-8.
391. Arbolea, S., et al., *Gut Bifidobacteria Populations in Human Health and Aging*. Front Microbiol, 2016. **7**: p. 1204.
392. Yang, C., et al., *Bifidobacterium in the gut microbiota confer resilience to chronic social defeat stress in mice*. Sci Rep, 2017. **7**: p. 45942.
393. von Rhein, D., et al., *The NeurolMAGE study: a prospective phenotypic, cognitive, genetic and MRI study in children with attention-deficit/hyperactivity disorder. Design and descriptives*. Eur Child Adolesc Psychiatry, 2015. **24**(3): p. 265-81.
394. Kaufman, J., et al., *Schedule for Affective Disorders and Schizophrenia for School-Age Children-Present and Lifetime Version (K-SADS-PL): initial reliability and validity data*. J Am Acad Child Adolesc Psychiatry, 1997. **36**(7): p. 980-8.
395. Franke, B., et al., *Genetic variation in CACNA1C, a gene associated with bipolar disorder, influences brainstem rather than gray matter volume in healthy individuals*. Biol Psychiatry, 2010. **68**(6): p. 586-8.
396. Aarts, E., et al., *Dopamine and the cognitive downside of a promised bonus*. Psychol Sci, 2014. **25**(4): p. 1003-9.
397. Steegenga, W.T., et al., *Sexually dimorphic characteristics of the small intestine and colon of prepubescent C57BL/6 mice*. Biol Sex Differ, 2014. **5**: p. 11.
398. Jaeggi, T., et al., *Iron fortification adversely affects the gut microbiome, increases pathogen abundance and induces intestinal inflammation in Kenyan infants*. Gut, 2015. **64**(5): p. 731-42.
399. Kanehisa, M. and S. Goto, *KEGG: kyoto encyclopedia of genes and genomes*. Nucleic Acids Res, 2000. **28**(1): p. 27-30.
400. Nair, H., et al., *Global burden of acute lower respiratory infections due to respiratory syncytial virus in young children: a systematic review and meta-analysis*. Lancet, 2010. **375**(9725): p. 1545-55.
401. Hall, C.B., et al., *The burden of respiratory syncytial virus infection in young children*. N Engl J Med, 2009. **360**(6): p. 588-98.
402. Meissner, H.C., *Viral Bronchiolitis in Children*. New England Journal of Medicine, 2016. **374**(1): p. 62-72.
403. Openshaw, P.J. and J.S. Tregoning, *Immune responses and disease enhancement during respiratory syncytial virus infection*. Clin Microbiol Rev, 2005. **18**(3): p. 541-55.
404. Simoes, E.A., *Environmental and demographic risk factors for respiratory syncytial virus lower respiratory tract disease*. J Pediatr, 2003. **143**(5 Suppl): p. S118-26.
405. Tregoning, J.S. and J. Schwarze, *Respiratory viral infections in infants: causes, clinical symptoms, virology, and immunology*. Clin Microbiol Rev, 2010. **23**(1): p. 74-98.
406. Russell, C.D., et al., *The Human Immune Response to Respiratory Syncytial Virus Infection*. Clin Microbiol Rev, 2017. **30**(2): p. 481-502.
407. Openshaw, P.J.M., et al., *Protective and Harmful Immunity to RSV Infection*. Annu Rev Immunol, 2017. **35**: p. 501-532.

408. Collins, P.L. and B.S. Graham, *Viral and host factors in human respiratory syncytial virus pathogenesis*. J Virol, 2008. **82**(5): p. 2040-55.
409. El Saleeby, C.M., et al., *Respiratory syncytial virus load, viral dynamics, and disease severity in previously healthy naturally infected children*. J Infect Dis, 2011. **204**(7): p. 996-1002.
410. Houben, M.L., et al., *Disease severity and viral load are correlated in infants with primary respiratory syncytial virus infection in the community*. J Med Virol, 2010. **82**(7): p. 1266-71.
411. Hasegawa, K., et al., *Respiratory syncytial virus genomic load and disease severity among children hospitalized with bronchiolitis: multicenter cohort studies in the United States and Finland*. J Infect Dis, 2015. **211**(10): p. 1550-9.
412. Welliver, T.P., et al., *Severe human lower respiratory tract illness caused by respiratory syncytial virus and influenza virus is characterized by the absence of pulmonary cytotoxic lymphocyte responses*. J Infect Dis, 2007. **195**(8): p. 1126-36.
413. Cortjens, B., et al., *Neutrophil extracellular traps cause airway obstruction during respiratory syncytial virus disease*. J Pathol, 2016. **238**(3): p. 401-11.
414. Brand, H.K., et al., *CD4+ T-cell counts and interleukin-8 and CCL-5 plasma concentrations discriminate disease severity in children with RSV infection*. Pediatr Res, 2013. **73**(2): p. 187-93.
415. Johnson, C.L. and J. Versalovic, *The human microbiome and its potential importance to pediatrics*. Pediatrics, 2012. **129**(5): p. 950-60.
416. Yatsunenko, T., et al., *Human gut microbiome viewed across age and geography*. Nature, 2012. **486**(7402): p. 222-7.
417. Clarke, T.B., et al., *Recognition of peptidoglycan from the microbiota by Nod1 enhances systemic innate immunity*. Nat Med, 2010. **16**(2): p. 228-31.
418. Clarke, T.B., et al., *Invasive bacterial pathogens exploit TLR-mediated downregulation of tight junction components to facilitate translocation across the epithelium*. Cell Host Microbe, 2011. **9**(5): p. 404-14.
419. Brechley, J.M., et al., *Microbial translocation is a cause of systemic immune activation in chronic HIV infection*. Nat Med, 2006. **12**(12): p. 1365-71.
420. Vissers, M., R. de Groot, and G. Ferwerda, *Severe viral respiratory infections: are bugs bugging?* Mucosal Immunol, 2014. **7**(2): p. 227-38.
421. Mansbach, J.M., et al., *Respiratory syncytial virus and rhinovirus severe bronchiolitis are associated with distinct nasopharyngeal microbiota*. J Allergy Clin Immunol, 2016. **137**(6): p. 1909-1913.e4.
422. de Steenhuijsen Piter, W.A., et al., *Nasopharyngeal Microbiota, Host Transcriptome, and Disease Severity in Children with Respiratory Syncytial Virus Infection*. Am J Respir Crit Care Med, 2016. **194**(9): p. 1104-1115.
423. Stewart, C.J., et al., *Associations of Nasopharyngeal Metabolome and Microbiome with Severity among Infants with Bronchiolitis. A Multiomic Analysis*. Am J Respir Crit Care Med, 2017. **196**(7): p. 882-891.
424. Roe, M.F., et al., *Changes in helper lymphocyte chemokine receptor expression and elevation of IP-10 during acute respiratory syncytial virus infection in infants*. Pediatr Allergy Immunol, 2011. **22**(2): p. 229-34.
425. Brand, K.H., et al., *Use of MMP-8 and MMP-9 to assess disease severity in children with viral lower respiratory tract infections*. J Med Virol, 2012. **84**(9): p. 1471-80.
426. Hasegawa, K., et al., *Nasal Airway Microbiota Profile and Severe Bronchiolitis in Infants: A Case-control Study*. Pediatr Infect Dis J, 2017. **36**(11): p. 1044-1051.
427. Hasegawa, K., et al., *Association of nasopharyngeal microbiota profiles with bronchiolitis severity in infants hospitalised for bronchiolitis*. Eur Respir J, 2016. **48**(5): p. 1329-1339.
428. Korten, I., et al., *Interactions of Respiratory Viruses and the Nasal Microbiota during the First Year of Life in Healthy Infants*. mSphere, 2016. **1**(6).
429. Tarabichi, Y., et al., *The administration of intranasal live attenuated influenza vaccine induces changes in the nasal microbiota and nasal epithelium gene expression profiles*. Microbiome, 2015. **3**: p. 74.
430. Rosas-Salazar, C., et al., *Differences in the Nasopharyngeal Microbiome During Acute Respiratory Tract Infection With Human Rhinovirus and Respiratory Syncytial Virus in Infancy*. J Infect Dis, 2016. **214**(12): p. 1924-1928.
431. Rosas-Salazar, C., et al., *Nasopharyngeal Microbiome in Respiratory Syncytial Virus Resembles Profile Associated with Increased Childhood Asthma Risk*. Am J Respir Crit Care Med, 2016. **193**(10): p. 1180-1183.

432. Lu, W., et al., *Microfloral diversity in the lower respiratory tracts of neonates with bacterial infectious pneumonia combined with ventilator-associated pneumonia*. Mol Med Rep, 2016. **14**(6): p. 5223-5230.
433. Lambiase, A., et al., *Achromobacter xylosoxidans respiratory tract infection in cystic fibrosis patients*. Eur J Clin Microbiol Infect Dis, 2011. **30**(8): p. 973-80.
434. Rajan, S. and L. Saiman, *Pulmonary infections in patients with cystic fibrosis*. Semin Respir Infect, 2002. **17**(1): p. 47-56.
435. Teo, S.M., et al., *The infant nasopharyngeal microbiome impacts severity of lower respiratory infection and risk of asthma development*. Cell Host Microbe, 2015. **17**(5): p. 704-15.
436. Franz, A., et al., *Correlation of viral load of respiratory pathogens and co-infections with disease severity in children hospitalized for lower respiratory tract infection*. J Clin Virol, 2010. **48**(4): p. 239-45.
437. Gulraiz, F., et al., *Haemophilus influenzae increases the susceptibility and inflammatory response of airway epithelial cells to viral infections*. FASEB J, 2015. **29**(3): p. 849-58.
438. Bellinghausen, C., et al., *Exposure to common respiratory bacteria alters the airway epithelial response to subsequent viral infection*. Respir Res, 2016. **17**(1): p. 68.
439. Langereis, J.D. and M.I. de Jonge, *Invasive Disease Caused by Nontypeable Haemophilus influenzae*. Emerg Infect Dis, 2015. **21**(10): p. 1711-8.
440. Christiaansen, A.F., et al., *Altered Treg and cytokine responses in RSV-infected infants*. Pediatr Res, 2016. **80**(5): p. 702-709.
441. Hasegawa, K., et al., *The relationship between nasopharyngeal CCL5 and microbiota on disease severity among infants with bronchiolitis*. Allergy, 2017. **72**(11): p. 1796-1800.
442. Vissers, M., et al., *High pneumococcal density correlates with more mucosal inflammation and reduced respiratory syncytial virus disease severity in infants*. BMC Infect Dis, 2016. **16**: p. 129.
443. Jong, V.L., et al., *Transcriptome assists prognosis of disease severity in respiratory syncytial virus infected infants*. Sci Rep, 2016. **6**: p. 36603.
444. Templeton, K.E., et al., *Rapid and sensitive method using multiplex real-time PCR for diagnosis of infections by influenza A and influenza B viruses, respiratory syncytial virus, and parainfluenza viruses 1, 2, 3, and 4*. J Clin Microbiol, 2004. **42**(4): p. 1564-9.
445. Letunic, I. and P. Bork, *Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation*. Bioinformatics, 2007. **23**(1): p. 127-8.

CHAPTER 11

**NEDERLANDSE SAMENVATTING
(DUTCH SUMMARY)**

Samenvatting van dit proefschrift

(Dutch summary)

De wereld om ons heen is bezaaid met microscopisch klein leven. De diversiteit van deze voornamelijk eencellige *micro-organismen*, ook wel *microbiota* genoemd, is enorm groot en bestaat onder meer uit bacteriën, schimmels en gisten, maar ook uit complexere eencellige organismen zoals amoeben en algen. De woonomgeving van deze microbiota bevindt vrijwel overal op aarde, bijvoorbeeld in het zoet- en zoutwater van meren en zeeën, in de aardbodem, en op bomen, planten en dieren, inclusief op en in ons eigen lichaam. De natuurlijke leefomgeving van microbiota, ook wel het *microbioom* genoemd, is een zeer complex geheel en exacte definities hiervan leidt tot veel debat in de literatuur. Kortgezegd is een microbioom een beschrijving van een geografische locatie waar microbiota aanwezig is (bijvoorbeeld op de huid, oftewel het *huidmicrobioom*), tezamen met hun erfelijke materiaal (DNA, genen, etc.) en alle componenten afkomstig van hun cellen (eiwitten, vetten, bouwstoffen, etc.), en inclusief aanwezige niet-levende omgevingsfactoren zoals temperatuur, vochtigheid, zuurtegraad en beschikbare voedingsstoffen. Tot slot worden de veelal aanwezige virussen en bacteriofagen (virussen specifiek voor bacteriën) ook tot het microbioom gerekend, ofschoon zij zelf technisch gezien geen microbiota zijn.

In algemene zin worden bacteriën en schimmels door het bredere publiek (toch nog) voornamelijk geassocieerd met ziekte en 'enge beestjes', ofschoon dit langzaamaan aan het veranderen is omdat men om zich heen steeds meer de industriële en medische toepassingen en potentie van microbiota ziet. Hierbij kun je denken aan moderne toepassingen van eeuwenoude vergistingsprocessen (bier, wijn, sojasaus, zuurkool, etc.), maar denk ook aan de zuivelindustrie (kaas, yoghurt, kefir, etc.) en het gebruik van microbiota als kleine celfabrieken voor de productie van bijvoorbeeld medicijnen, vaccines, enzymen of specifieke stoffen. Binnen het wetenschappelijke domein geniet het menselijk microbioom de laatste jaren enorm veel aandacht in relatie tot ziekte en gezondheid, zo ook als onderwerp van dit proefschrift. Dat komt door de toenemende hoeveelheid bewijs en onderzoek wat laat zien dat een 'gezond' microbioom beschermt tegen ziekte, en dat een 'ongezond' microbioom het risico op ontwikkeling van ziekte vergroot. Vaak spreekt men van een *dysbiose* als er sprake is van een 'ongezond' microbioom, iets wat simpel gezegd op te vatten is als een ontregeling van het microbioom. Ofschoon in deze context de exacte definities van 'gezond en ongezond' in de praktijk haast nog niet te definiëren zijn. Grotendeels is dit te wijten aan de enorme verschillen in microbiota tussen individuen, maar ook tussen lichaamslocaties en zelfs tijdstippen van meten. Er zijn slechts enkele voorbeelden in de literatuur bekend waarbij er maar één duidelijke speler in het microbioom verantwoordelijk is voor of gerelateerd aan een bepaald ziektebeeld, en waarbij dit zo zwart-en-wit is, zoals bij eczeem waarbij de bacterie *Staphylococcus aureus* een belangrijke rol speelt ([hfst 3](#)). Een algeheel beeld is wel dat een meer divers microbiota (wanneer er meer verschillende typen van micro-organismen aanwezig zijn) een betere bescherming lijkt te bieden tegen dysbiose ten gevolge van veranderingen in de omgeving.

Er is nog volop onderzoek nodig om beter begrip te krijgen van de manier waarop en mate waarin microbiota een rol spelen bij gezondheidsprocessen, en om hier in de toekomst bijvoorbeeld farmaceutisch op in te kunnen grijpen. Ondanks het veelal negatieve karakter van microbiota in de maatschappij valt het op dat men zich ook daar steeds meer bewust wordt van het belang van microbiota ten aanzien van gezondheid, ofschoon in de praktijk evenals nog onduidelijk is wat dit precies omvat. Interessant is ook de opkomst van de probiotica, en de toenemende mate waarin dit populair begint te worden. Een probioticum betreft het toedienen van levende bacteriën die bij zouden moeten dragen aan je gezondheid. Vaak gaat het hierbij om melkzuurbacteriën zoals *Lactobacillus* of om andere (darm)bacteriën zoals de welbekende *Bifidobacterium*. Een ander mooi voorbeeld is de recente ontwikkeling ten aanzien van poeptransplantaties, welke met veel medisch succes worden toegepast als therapie voor *Clostridium difficile* infecties. Analooq aan de bloedbank lopen momenteel verschillende initiatieven in Nederland voor het opbouwen van een poepbank.

Het doel van dit proefschrift was om meer inzicht te verschaffen in mechanismen van microbiota interacties met de omgeving. Om informatie te verschaffen over dergelijke mechanismen is het allereerst zaak om microbiota uit de omgeving van interesse in kaart te brengen. Dit werd van oudsher gedaan door het kweken van microbiota op speciaal met voedingsstoffen verrijkt kweekmedium. Echter lang niet alle microbiota zijn te kweken, en je brengt met *kweektechnieken* maar een fractie van de “natuurlijke” populatie in kaart. Voor het daadwerkelijk meten van de totale microbiële populatie in een microbioom gebruikt men tegenwoordig *sequencing technieken* waarbij het erfelijk materiaal (DNA) van microbiota wordt gemeten en geanalyseerd.

In **hfst 2** van deze dissertatie gaan we in op methoden voor het analyseren van bacterieel DNA om voorspellingen te kunnen doen over de aanwezige *genen* in een *bacterieel genoom* (een genoom is de totale erfelijke informatie van een organisme). Door het slim combineren van bestaande detectiemethoden voor bacteriële genen laten we zien dat je het karakteriseren van genen binnen een bacterieel genoom kunt verbeteren. Er bestaan grofweg twee verschillende methoden voor het *sequencen van DNA*, en dus voor het in kaart brengen van microbiota, één op basis van markergenen, en één door het sequencen van het totale microbiële genomen.

In **hfst 3** laten we een nieuwe methode zien die we zelf hebben ontwikkeld. Deze methode is analoog aan de methode voor het meten van *markergenen* en stelt ons in staat om bacteriën te meten in een microbioom met hogere resolutie dan momenteel mogelijk is met de standaard methode. Deze eerste twee onderzoekshoofdstukken van dit proefschrift zijn *bioinformatisch* van aard, zijn methodologisch gedreven, en zijn bedoeld als fundering voor het vervolgonderzoek zoals beschreven in de hoofdstukken daarna.

In **hfst 4** beschrijven we een nieuw model voor het kweken van huidbacteriën op basis van menselijk eelt. Zoals eerder genoemd zijn niet alle bacteriën te kweken in een laboratorium op kweekmedium, en ook voor huidbacteriën is dit lastig. Interessant genoeg, door het toevoegen van verpulverd en gesteriliseerd eelt van de voeten aan

het kweekmedium laten we zien dat we in staat zijn om specifieke huidbacteriën te kunnen laten groeien op een vergelijkbare manier als dat we van nature zien groeien op de huid. Dit model noemen we het *callusmodel* (callus is eelt).

In **hfst 5** laten we zien dat mensen met de huidaandoening *ichtyosis vulgaris* (ook wel vissenschubziekte genoemd) andere bacteriën met zich meedragen op de huid dan gezonde mensen. In dit hoofdstuk passen we het callusmodel toe in ons onderzoek, en ontdekken we dat dit verschil in microbiota op de huid van *ichtyosis vulgaris* patiënten waarschijnlijk het gevolg is van een verminderde hoeveelheid (kleine) eiwitten die de huid van nature soepel en gesmeerd houden, en dat dit komt omdat deze eiwitten als voedselbron worden gebruikt door een specifieke groep van huidbacteriën, namelijk de Gram-positieve anaerobe cocci (GPAC), zoals de kandidaatbacterie *Finegoldia*.

In **hfst 6** verleggen we onze aandacht van huidbacteriën naar *darmbacteriën*, en onderzoeken we hun relatie met de ziekte *reumatoïde artritis*. In speciale *kiemvrije muizen* (dit zijn muizen die geen microbiota met zich meedragen dus volledig steriel zijn) laten we zien dat artritis zich minder snel ontwikkelt, en minder snel verergerd dan in muizen die gekoloniseerd zijn met natuurlijke microbiota. In dit hoofdstuk proberen we tevens te achterhalen welke specifieke populatie van bacteriën hier een rol in zou kunnen spelen, en laten we zien dat de groep van Gram-negatieve anaerobe bacteriën, bijvoorbeeld de kandidaatbacterie *Helicobacter*, een belangrijke bijdrage lijkt te leveren aan het artritis ziekteproces in muizen.

In **hfst 7** kijken we naar verschillen in de darmbacteriën van jongvolwassenen met en zonder *ADHD* (dit keer in mensen, niet in muizen), en onderzochten we of eventuele verschillen iets te maken kan hebben met de manier waarop de hersenen van beide groepen functioneren. Het onderzoeks idee klinkt misschien wat vreemd, maar is zeker niet uit de lucht gegrepen. *ADHD* is een stoornis in de *hersenenontwikkeling*. De stoornis hangt samen met afwijkingen in de *dopamine verwerking*, de *beloningsmechanismen* en de onderliggende *neurologische bedrading*. Via de zogenaamde *hersen-darm-as* kunnen die twee organen elkaar beïnvloeden, en zou het microbiom dus potentieel vanuit de darm invloed kunnen uitoefenen op de hersenen, en vice versa. De onderzoeksresultaten laten zien dat in de darmen van vrijwilligers met *ADHD* méér bacteriën zitten die een stof produceren die kan worden omgezet in *dopamine*, dan bij gezonde vrijwilligers. Bij de mensen bij wie ook een *hersenscan* was gemaakt zien we dat mensen met en zonder *ADHD* die meer van die bacteriën in hun darmen hebben inderdaad minder activiteit in de *beloningsgebieden* van hun hersenen vertonen, een essentieel kenmerk van mensen met *ADHD*. Hierdoor hebben we de hypothese opgeworpen dat die bacteriële stof via het bloed in de hersenen terecht kan komen om daar, na omzetting in dopamine, de hersenenfunctie te beïnvloeden.

In ons laatste onderzoekshoofdstuk **hfst 8** hebben we *RS-virusinfecties* bij *pasgeboren kinderen* onderzocht. Vrijwel alle kinderen maken een *RS-virusinfectie* door voor hun tweede levensjaar. In de meeste gevallen leidt dit tot een milde *luchtweginfectie* zonder veel complicaties, vergelijkbaar met een ordinaire *verkoudheid*. Echter, een klein percentage van de kinderen ontwikkelt ernstige complicaties ten gevolge van *RS*

waarbij in sommige gevallen zelfs ziekenhuisopname noodzakelijk is met intensieve zorg. Waarom bepaalde kinderen zo extreem reageren op een RS-virusinfectie wordt nog niet voldoende begrepen, en zodoende hebben wij gekeken naar de mogelijke invloed van *natuurlijke bacteriën* in de *neus/keelholte* op *immuunprocessen* tijdens een dergelijke luchtweginfectie. Onze belangrijkste bevinding is dat kinderen die een RS-virusinfectie doormaken veel meer van de bacterie *Haemophilus* bij zich dragen. Deze bacterie wordt als opportuun beschouwd, ofschoon we dit niet direct kunnen koppelen aan ernst van verloop van deze RS-virusinfecties. Verder vinden we een verband van de samenstelling van het lokale microbioom met de hoeveelheid aanwezige RS-virusdeeltjes, en met een specifieke immunologische signaalstof, waarbij in beide gevallen opnieuw *Haemophilus* een belangrijke rol speelt.

In de discussie van deze dissertatie gaan we in op het belang van experimentele validatie van onderzoeksresultaten, vooral wanneer deze voornamelijk met bioinformatische methoden zijn gevonden. Hierbij kun je denken aan laboratoriummodellen om hypothesen te testen, zoals bijvoorbeeld in de eerder genoemde huid- en diermodellen. Het ultieme doel hierbij is om associaties uit onderzoeken naar oorzakelijke verbanden te vertalen. Verder besteden we aandacht aan onderzoeksethiek en de rol van de bioinformaticus in onderzoek, het belang van de onderzoeksopzet en datavisualisatie, en bespreken we hypothese-gedreven onderzoek versus 'grasduinen' oftewel ongericht onderzoek. In de toekomst zal de behoefte en noodzaak om onderzoeksresultaten te vertalen naar concrete ideeën en toepassingen enkel toenemen. Zo is het mechanistisch begrip van ziekteprocessen van enorm belang om met vernieuwende therapeutische toepassingen te komen, en de bioinformatica zal hier zeker een belangrijke rol in gaan spelen.

ADDENDUM

About the author
Dankwoord
Research data stewardship and accessibility (FAIR)
List of publications
List of abbreviations
notes

About the author

(curriculum vitae)

Thomas (Tom) Hendrikus Antonius Ederveen was born on June 13th 1987, in the Radboud hospital in Nijmegen. He grew up in Wijchen next to Nijmegen, and there he finished his high school (HAVO) at the Maaswaal College in 2004.

Thereafter, he started with a Bachelor of Applied Life Sciences at the HAN University where he followed training as a Lab Technician (i.e. *HLO: Hoger Laboratorium Onderwijs*) with focus on biological and medical research, and majoring in biochemistry. During this education he got the opportunity to do a six month internship at Schering-Plough (Oss, NL) at the Department of Pharmaceuticals in the group of dr. Joop Waterval, studying analytical methods for stability monitoring of biologicals. In addition, he followed another half-year internship at Intervet Animal Health (Boxmeer, NL), Bacteriological R&D, in the group of dr. Paul Vermeij, with a project on bacterial protein expression systems for vaccine antigen production where, in retrospect, his first real interest in microbiota and molecular biology was sparked.

After his bachelor, in search for a more mechanistic understanding of biological processes, Tom started the Master Medical Biology at the Radboud University in 2009 where he received his Master of Science degree in 2012. During that time he did a 9 month internship and literature study at the Molecular Biology department of the Radboud University under supervision of dr. Colin Logie, and this led to his first scientific publication, on the human histone H3 protein complement. During that time Tom got his first taste of the field of bioinformatics, which excited him and resulted in searching for a training in bioinformatics. Finally, in his graduation year, Tom was kindly invited to the Bacterial Genomics group of dr. Sacha van Hijum at the CMBI bioinformatics department of prof. Gert Vriend (Radboudumc), where he stayed for almost a year. During that time he gained a basic understanding of bioinformatics working on comparative genome annotation in prokaryotes, and this ultimately led to another scientific paper in 2013 (part of this thesis).

Being trained mainly as a wet-lab scientist Tom initially decided to continue his career in the laboratory environment, and was enrolled in a PhD studentship at the department of Experimental Rheumatology in 2012. Nevertheless, after one and a half year Tom switched to the aforementioned Bacterial Genomics group of Sacha van Hijum where he had more scientific opportunities and support with regard to bioinformatics, and continued his PhD there. His efforts at Experimental Rheumatology resulted in a publication on the gut microbiome in relation to arthritis (part of this thesis). During the next years Tom developed his computational skills applying them in the microbiome field, with many collaborations in the Radboudumc and outside, as reflected in this thesis. He strongly believes that enabling relevant and high-quality research requires to work in a multidisciplinary environment, and in close collaboration with (clinical) researchers from other research fields and departments (especially for computationally-oriented researchers).

In 2015, Tom had a secondment at NIZO health and food research (Ede, NL) for a year. The goal of this exchange was to (i) learn from one another, to (ii) provide some support in NIZO contract research projects, and (iii) to enable fruitful collaborations between Radboudumc and NIZO in the dynamic field of microbiota and health. Tom still visits NIZO weekly.

In 2018, Tom worked at the laboratory of Dermatology of prof. Joost Schalkwijk (Radboudumc) for a year, after several successful projects on human skin microbiota. This enabled him to fully focus on that topic, and to support the department in (sequencing) data analysis.

Tom currently works at the CMBI department as a post-doc in an European consortium focusing on the gut-brain axis involvement of intestinal bacteria in early life. In addition, he has a role in the Bioinformatics Radboud Technology Center (RTC) and supports many Radboudumc projects in the area of data analysis, sequencing and microbiomics.

Thomas H.A. Ederveen
website : <http://ederveen.science>

Dankwoord

(Acknowledgements)

Promoveren doe je niet alleen. Voor promotie is onderzoek nodig, en dat kan tegenwoordig haast niet meer zonder een divers en multidisciplinair team aan onderzoekers, inclusief de ondersteunende staf en afdelingen. Ik heb het geluk en genoeg gehad om samen te kunnen werken met veel fantastische mensen van verschillende afdelingen, binnen en buiten het Radboudumc. Ik heb van eenieder enorm veel geleerd, volop waardevolle input op mijn projecten en advies ontvangen tijdens mijn promotietraject. Promoveren (onderzoek) is helaas ook een relatief langzaam proces, een les in geduld, lange adem hebben en omgaan met tegenslag. Ook op die momenten heb ik enorm veel (!) support en steun om mij heen ontvangen tijdens dit traject, van collega's, vrienden en familie. Ik ben iedereen daarvoor enorm dankbaar, ik heb een fantastische tijd gehad in het Radboudumc en hoop dit nog lang door te kunnen zetten, want er is nog volop werk aan de winkel op prachtige reeds lopende onderzoeksprojecten, en op nieuwe initiatieven die we momenteel aan het opstarten zijn.

Beste **Sacha**, dit proefschrift zou er absoluut niet zijn geweest ware het niet voor jouw onuitputtelijke inspanningen, maar vooral ook door jouw vertrouwen in mij. Om te beginnen toen ik jaren geleden, in 2011 alweer, bij jou aanklopte voor een stage in de bioinformatica (waarin ik destijds nog nul achtergrond had). Ik had interesse in de bioinformatica, maar ik had twijfels of ik daar als medisch bioloog zomaar in zou kunnen rollen. Jij had wel oor naar iemand met een biologische mind-set voor de juiste interpretatie van biologische data, en de bioinformatica en het programmeren was volgens jou geen "rocket science" dus jij voorzag weinig problemen. En dat is gebleken. Na een fantastisch jaar in jouw groep heb ik mij de basisprincipes van de bioinformatica eigen kunnen maken, die ruimte heb je me gegeven. Ook toen ik twee jaar later weer aan je deur stond heb je me voor een aantal maanden in je groep opgenomen, ondanks dat de middelen eigenlijk nauwelijks beschikbaar waren. De kans die je mij op dat moment hebt gegeven is, terugkijkend, de kick-start geweest voor mijn huidige carrière in het microbiom onderzoeksveld, en daarvoor ben ik je enorm dankbaar. In die korte tijd is het ons gelukt om voldoende projecten binnen te halen om mijn tijd in jouw groep te prolongeren. En zie hier waar dat uiteindelijk toe heeft geleid. In de jaren die volgde hebben we heel wat moois opgebouwd en spannende projecten tot een succesvol einde gebracht (zie deze thesis). De vrijheid en ruimte die je me ten alle tijden hebt gegeven in combinatie met jouw directe stijl van coaching is iets waarbij ik het beste tot mijn recht kom in dit werk, en dat heb jij altijd heerlijk aangevoeld. Het is dan ook tot mijn diepe spijt dat je afscheid hebt moeten nemen van het Radboudumc, het is wrang dat het ophouden van jouw aanstelling niet te voorkomen was, terwijl je dit zo lang bij mij hebt kunnen afwenden. Daarnaast is de laatste twee jaar is een lastige periode voor je geweest, maar toch ben je tot het eindstation betrokken geweest met de afronding van deze thesis en al haar hoofdstukken, en dat siert je. Onwijs bedankt voor de fijne en spannende maar ook soms turbulente tijd samen, onze vele mooie gesprekken en wetenschappelijke en filosofische discussies. Dat laatste is zeker iets wat ik mis. Ik wens

je het allerbeste en ben er van overtuigd dat je met jouw ervaring en persoonlijkheid op een nieuwe plek weer iets fantastisch kunt opzetten.

Helaas waren er ook momenten die enigszins eenzaam waren, niet zozeer door gebrek aan mensen om me heen, maar wel doordat de zeer specifieke niche waarin ik onderzoek deed het inhoudelijk reflecteren vaak lastig maakte. **Sacha, Jos** (Boekhorst) en **Michiel**, ik denk dat jullie op die momenten letterlijk de enige waren die volledig begrepen waar ik mee bezig was en met wie ik van gedachten kon wisselen. Jullie gezamenlijke kennis van microbiota, genetica en bioinformatica is ongeëvenaard en het was mij een genoegen om jaren met jullie samen te hebben kunnen werken. Ik durf wel te zeggen dat ik mijn huidige kennis van zaken in het microbiom / bioinformatica veld voor het grootste deel aan jullie te danken heb. Ondanks alle ondersteuning zijn er ook momenten waarop je grotendeels op jezelf bent aangewezen bij het maken van beslissingen in onderzoek, niet door gebrekkige coaching maar omdat je simpelweg niet alle details continu kunt afstemmen. Data analyse is constant beslissingen maken op basis van wat je observeert. Door jullie heb ik geleerd op mij eigen handelingen te vertrouwen en om te data voor zich te laten spreken. Desalniettemin kon ik jullie altijd direct mailen of bellen als de behoefte daar was om mijn gedachten op een rijtje te zetten, voor lastige vragen, of om bij jullie te testen of mijn ideeën hout sneden: waarvoor mijn dank op die momenten en voor jullie torenhoge geduld!

In het bijzonder, **Jos** (Boekhorst), jouw onuitputtelijker enthousiasme voor onderzoek heeft mij altijd gestimuleerd. Dit gecombineerd met jouw bijzonder sterk analytisch vermogen zorgde ervoor dat ik vaak ongeremd met jou kon sparren en schakelen over wild uiteenlopende onderwerpen binnen ons onderzoeksveld (en met regelmaat sneller dat ik aankon haha), en heeft daarom zonder twijfel enorm veel bijgedragen aan niet alleen de resultaten van dit proefschrift, maar ook aan mijn begrip van bioinformatica en biomedisch onderzoek. En niet te vergeten dat velen bioinformatische tools en methoden die gebruikt zijn in onze onderzoeken door jou zijn geïmplementeerd of ontwikkeld. Ik heb enorm veel van je geleerd en dat doe ik nog elke week, dat waardeer ik zeer, en hoop nog lang het genoegen te hebben om op deze manier met jou te kunnen samenwerken!

Moeilijke momenten zijn er zeker geweest, ik denk dat de meeste promovendi uiteindelijk wel ergens tegenaan lopen. Voor mij waren lastige momenten in eerste instantie de terugkerende mate van onzekerheid rondom een aanstelling om mijn promotie af te kunnen ronden. Wonderbaarlijk genoeg is dit altijd goed gekomen door jouw inspanningen **Sacha**, en hebben we gezamenlijk altijd genoeg projecten binnen kunnen halen, of zijn we op andere manieren creatief genoeg geweest om continuïteit te garanderen, bijvoorbeeld door de samenwerkingen met NIZO. Ik wil in deze ook zeker **Barbara** en **Gert** bedanken, omdat jullie je heel sterk hebben gemaakt voor mijn voortgang op de afdeling, en meer dan eens! Immers, een afdeling die je steunt en achter je staat is zeer belangrijk om het beste uit je werk te kunnen halen. **Barbara**, jij bent voor mij de enige echt zekere factor op het CMBI gebleken. Ik kon altijd bij jou tot vermoedens toe binnen lopen met de zoveelste vraag of willekeurig verhaal over wat dan ook, en je maakte altijd tijd voor mij om mij direct te helpen of simpelweg

aan te horen. Ondanks dat ik ongetwijfeld zelden de eerste was met een willekeurige vraag die dag, en ook ondanks dat er toch regelmatig zaken langs mij heen gaan die gewoon netjes zijn medegedeeld op de afdeling (mijn familie en vrienden schijnen dit te herkennen ...). Niet alleen jouw dagelijkse ondersteuning en kennis van reilen en zeilen van de afdeling en management daarvan heeft mij enorm geholpen, ook je bijdrage aan de gezelligheid op de afdeling is enorm. Vooral ook in de afgelopen periode van extreme leegloop op de afdeling heb ik veel plezier gehad van jouw aanwezigheid om ons heen. **Gert**, ik heb je leren kennen als een warm persoon die volledig achter zijn mensen staat zolang je zelf je uiterste best doet om er het beste van te maken. Iedereen heeft bij jou altijd een kans, en iedereen wordt bij jou gelijk behandeld, zo ook ik toen ik in 2011 begon op jouw afdeling zonder enig bioinformatisch begrip of kennis van eiwitten (dat laatste heb ik nog steeds niet eigenlijk, sorry!). Ondanks dat we inhoudelijk vrijwel nooit hebben samengewerkt, en we daarom wat minder directe interactie hadden vanwege de afdelingsstructuur (verschil in onderzoeksfocus), heb ik wel gemerkt dat je voor me klaar stond. Vooral de laatste twee jaar na vertrek van Sacha merkte ik dat je zijn rol stiekem overnam en regelmatig bij mij binnen liep om te kijken of alles goed ging, en of ik nog wat nodig had, en om mij indien nodig te voorzien van advies en ondersteuning. Dat gaf mij een gevoel van betrokkenheid en heb ik enorm gewaardeerd, bedankt!

Beste **Karima**, wat heb ik een ontzettend leuke werktijd gehad in die paar jaar dat je bij ons in de BAMICS groep zat! We waren op dat moment als groep misschien niet zo groot meer, maar met Sacha, met z'n drieën, hebben we in mijn ervaring toch een zeer productieve en leuke tijd gehad. Ik heb je leren kennen als een ontzettend warm persoon, altijd vrolijk en positief. Het was dan ook zeer aangenaam om bij je op kantoor te zitten, en mis nog steeds zo nu en dan onze mooie gesprekken waar ik met veel plezier op terug kijk, evenals je heerlijke baksels! Je bent bijzonder gedreven in je werk en een zeer vaardige programmeur, ik ben dan ook dankbaar dat je mij regelmatig hebt kunnen helpen met code/Unix zaken, en vooral toen ik - eindelijk (!) - overstapte van Perl naar Python (ofschoon die Perl *regex* is toch wel fijn hè!). Ondanks een turbulent einde bij ons in de groep wat ons allemaal destijds best heeft aangegrepen, ben ik blij dat het uiteindelijk goed is gekomen, en je een mooie bioinformatica positie hebt gevonden in Utrecht. Veel succes en geluk in je verdere privéleven en wetenschappelijke carrière!

Fredrick, *my brother from another mother! : -) You as my office roommate at the CMBI has been a pleasure. I enjoyed your company and our sharing of bioinformatics expertise. I'm glad you have learned me how to properly pronounce "Zebra" and how almost every Dutchman is doing that wrong! It's a funny fact that I will probably never forget. But most memorable of course was your defense, with me together with your brother as your paranympths. Thank again for that honor and nice experience, glad to know that you are doing great in Australia, and for knowing to have a friend at the other end of the world.*

Beste **Arthur**, jouw support en bijdrage aan het CMBI is van vitaal belang gebleken om onze systemen vlekkeloos draaiende te houden, je bent voor mij de stille kracht van de afdeling als het gaat om technische zaken en ICT. Bedankt voor je hulp op tal van zaken,

waarbij je me altijd op onvermoeibare wijze verder hebt gelopen. Het maakt niet uit of dit nu vragen of problemen zijn met software, servers, hardware of zelfs printers: jij denkt altijd goed mee en komt stevast met een geschikte oplossing.

Beste **Peter-Bram**, je maakt nog net het staartje mee van mijn promotietijd, want we kennen elkaar nog maar kort. Maar ik kan nu al kan zeggen dat jouw komst als nieuwe afdelingshoofd van het CMBI een goede impuls heeft gegeven aan de afdeling. Je voltrekt je als de nieuwe bindende factor van de afdeling, en we zijn sinds tijden weer flink aan het groeien in groeps grootte. Het is fijn dat je de noodzaak ziet van samenwerken tussen verschillende afdelingen en disciplines, en dat je er zichtbaar actief naar handelt om dat ook te verwezenlijken. En dit alles zonder bioinformatica weg te zetten als enkel een service provider. Bedankt voor een mooie start, en ik kijk er naar uit om met z'n allen aan de slag te gaan!

Beste **Lex**, toen ik als student begon in de BAMICS groep was jij al bezig met je PhD. Je bent altijd bijzonder behulpzaam en geduldig geweest met mijn vragen en (programmeer)problemen, ondanks dat ik officieel niet bij jou stage liep. Tevens hebben wij samen met Sacha destijds een mooi paper geschreven o.b.v. mijn stage bij de BAMICS, wat nu onderdeel is van deze thesis. Dank voor al je hulp! Toen ik zelf aan mij PhD begon zat jij alweer in de eindfase en kort erna vertrok je naar Amsterdam voor je postdoc in het lab van Gerard Muyzer, toevallig genoeg nu commissielid in mijn promotie. Het is me veel waard dat ik naast het wetenschappelijk inhoudelijke aan jou en de andere BioIT Boys (zie hieronder) een mooie vriendschap heb overgehouden! Een bijzonder mooi jaar toegewenst aan jou en Emily, en het allerbeste!

Beste **Tim**, tijdens jouw CMBI tijd hebben we helaas niet echt samengewerkt op dezelfde projecten, maar ik heb je later wel leren kennen als een bijzonder vaardige (bio)informaticus, en als vriend ben je altijd zeer betrokken en bijzonder behulpzaam, iemand die altijd klaar staat voor zijn vrienden. Ik geniet nog altijd van onze etentjes en gesprekken. Bedankt voor een luisterend oor, je eindeloze belangstelling en je (bioinformatica) adviezen, welke mij zeker geregeld hebben geholpen in het op een rijtje krijgen van mijn eigen gedachten.

Dan nu aandacht voor de **BioIT Boys**, te weten: **Lex, Joep, Tim, Tom, Rob, Martin en Eugène**, met behoorlijk wat oud-CMBI'ers (welke lezer kan ze nog aanwijzen?). Beste mannen, wat hebben we ondertussen veel memorabele momenten meegemaakt in binnen en buitenland, wat begon op het BBC congres in Luxemburg in 2011 is uitgegroeid tot een jaarlijks evenement opzoek naar, zoals Lex het mooi verwoordde: de meest premium C-locatie van Europa (en sinds afgelopen jaar zelfs daarbuiten). Ofschoon het tegenwoordig wel veel te veel luxe is (helemaal niet erg trouwens, haha), maar dat er nog vele edities mogen volgen! Joep, het doet ons pijn dat je ons gaat verlaten voor Nieuw Zeeland, maar we begrijpen het, want jij hebt immers "alles over voor de wetenschap"! ; -) En jongens (lees: Tim), nogmaals sorry voor de ondertussen velen spelletjesavonden die ik heb verziekt met mijn onbewuste trollen, het is een gave ik weet het. Hoe dan ook, het doet me elke keer weer deugd om met jullie samen te zijn, bedankt!

Er zijn nog velen andere (oud) **CMBI** collega's die hebben bijgedragen aan mijn prettige verblijf op deze afdeling de afgelopen jaren tijdens mijn promotie. Hierbij wil ik nog specifiek noemen enkele oud BAMICS teamleden: **Lex, Aldert, Juma, Lennart, Christof, Bernadette, Leonardo, Tom, Victor, Tilman** en **Roland**. En van de andere CMBI groepen wil ik bedanken **Martijn, Bas, Hanka, Laurens, Jordy, Coos, Carolien, Lisette, Selma, Josh, Daniel** en **Dei**. *Thanks, also, to the Greek students that I've coached during my PhD: Nikos and Nestor!* En iedereen die ik hierbij nog vergeet. Bedankt voor de gezelligheid en goede (wetenschappelijke) sfeer op de afdeling!

Ik heb altijd veel plezier en genoeg gehaald uit zoveel mogelijk samenwerken met andere onderzoekers, binnen en buiten onze afdeling, en zelfs geregeld buiten het Radboudumc. Ondanks dat je er soms haast niet aan ontkomt als wetenschapper om op een soort van onderzoekseiland te zitten (zonder nu een semantische discussie te beginnen), is het vooral voor mij als bioinformaticus een absolute must om met wet-lab onderzoekers samen te werken (gelukkig maar, want dat maakt mijn werk heel veel meer aangenaam!).

Beste collega's van afdeling **Reumatologie**. Beste "Team Cytokine": **Marije, Rebecca, Shahla, Miranda, Birgitte, Sija** en **Debbie**: Bedankt voor een korte doch leuke en leerzame tijd bij jullie in de groep! We morgen er trots op zijn dat onze gezamenlijke inzet uiteindelijk heeft geleid tot een prachtig verhaal in een mooi tijdschrift (onderdeel van deze thesis). Ik wens eenieder van jullie veel succes en geluk in jullie verdere carrière en privéleven. Ook dank aan alle andere collega's en collega-OIO's van de afdeling die hebben bijgedragen aan een mooi jaar voor mij! En niet te vergeten, **Alex** en **Mike** van het CDL dierenlaboratorium, bedankt voor jullie beide enthousiasme en inzet bij de verzorging van de (kiemvrije) dieren!

*Dear **New** and **Victor**. Even though our publication on dog skin microbiota did not make it as one of the chapters in this thesis, I have been working on it for almost my entire PhD. It has been a great collaboration with many ups and downs, and I'm proud that you have been able to finish the study together with our support, as of the start of this year, with a nice publication. Congratulations again, and in particular to New: I wish you all the best with your own defense and with the next steps in your career!*

*Dear **Esther** and **Alejandro**. We have had a great collaboration on microbiota involvement in ADHD through the gut-brain axis, which after an enormous effort from us all has led to a nice scientific publication (part of this thesis) and even a small press release. I've learned a lot from the both of you, not only on the specific topics, but also from the way of working with people from multiple disciplines on one focused project. It has been a great pleasure working together. Many thanks, and I'm sure our paths will cross more often in future Radboudumc projects, to which I'm already looking forward!*

Beste **Marien** en **Gerben**. Bedankt voor de meer dan aangename samenwerking op het RSV project waarin we de betrokkenheid van lokale commensalen bacteriën hebben onderzocht ten tijde van een virale luchtweginfectie in pasgeborene (onderdeel van deze thesis). Ik vond het een spannend en interessant project, mede door het unieke cohort

en de achterliggende immunologie, maar temeer vanwege jullie enorme enthousiasme en waardering in onze samenwerking! Ik heb microbiologische en immunologisch veel geleerd van de velen werkinhoudelijke gesprekken, brainstorm sessies en koffieuurtjes, waarvoor heel erg bedankt! Ik vind het dan ook zeer leuk om op dit moment nog steeds actief met jullie samen te werken en bij te dragen aan jullie projecten, en ik ben er van overtuigd dat dit met dezelfde energie zal blijven gebeuren als dat ik gewend ben van jullie. **Marien**, ik wil jou nog in het bijzonder bedanken voor jouw rol als mijn mentor tijdens mijn promotie, en voor je steun en advies in een soms lastige en turbulente tijd na het vertrek van Sacha. Jouw enthousiasme en positivisme is enorm aanstekelijk en motiverend, en heeft me veel plezier gebracht in mijn samenwerking met jullie, bedankt!

Beste **Joost** en **Patrick** (Zeeuwen). Het is alweer 5 jaar geleden dat ik begon met een spannende samenwerking met jullie afdeling over huid microbiota bij verschillende dermatologische aandoeningen. Vlak daarvoor hadden jullie net het eerste huidmicrobioom paper de deur uit i.s.m. BAMICS en het NIZO, dus ik viel met mijn neus in de boter want al het methodologische was al uitgezocht waardoor ik mij volop met de biologie bezig kon houden. We hebben ondertussen verschillende mooie stukken gepubliceerd, waarvan er drie hier in deze thesis zijn opgenomen en een aantal liggen nog op de stapel c.q. bij de tijdschriften. Wat mij betreft een output om met z'n allen trots op te zijn. De samenwerking met jullie heeft mij altijd enorm veel energie en plezier gebracht, en ik kan mij niet herinneren dat er ooit problemen zijn geweest of dat er afspraken niet (op tijd) zijn nagekomen, wat voor mij een bevestiging is van jullie buitengewoon prettige "modus operandi". Wat ik vooral heb weten te waarderen is jullie diepe focus en toewijding waardoor jullie (wetenschappelijke) output hoog blijft: alles wat potentieel ook maar een beetje publiceerbaar is wordt uiteindelijk gebruikt en niets blijft op de plank liggen, en dat zonder op biologisch-vernieuwend vlak in te boeten. Dat is een bijzondere kwaliteit in ons werkveld, en ik vermoed dat dit deels komt omdat jullie als gehele afdeling bijzonder sterk onderling samenwerken en betrokken zijn bij elkaars werk, zonder enig gevoel van onderlinge competitie, en dat betaald zich uit. Verder zijn jullie zeer prettig in de omgang, en staan sterk achter jullie mensen, ik heb mij altijd welkom en vertrouwd gevoeld, en daarvoor ben ik jullie zeer dankbaar! Beste **Joost**, ik wil jou nog in het bijzonder bedanken voor jouw vertrouwen in mij, dat je mij een jaar in-huis nam op je afdeling, en dat je zonder enige twijfel wilde fungeren als mijn promotor, ik was hier sterk door vereerd. Ik heb veel van je geleerd, en je betrokkenheid bij mijn werk evenals bij mij als persoon heb ik zeer gewaardeerd. Heel veel geluk, en geniet van je welverdiende emeritaat!

Ik heb het afgelopen jaar dat ik bij **lab Derma** "gedetacheerd" zat een fantastisch leuke tijd gehad! Jullie zijn een enorm fijne, kundige en gezellige afdeling waar ik mij van begin af aan meer dan welkom heb gevoeld. Wel opvallend is dat ik meteen "op mijn woorden moest letten" (ofschoon ik niet de indruk heb dat me dit ooit gelukt is haha), want één misstap op een onbewaakt moment en je wordt genadeloos afgestraft met de meest idiote (woord)grappen. Echter, na een week doe je er zelf aan mee en nu mis ik het zelfs een beetje : -) Beste **Patrick Z., Jos S., Ellen, Hanna, Joost, Gijs, Diana, Noa, Patrick J., Danique, Elkie, Ivonne, Piet** en natuurlijk niet te vergeten de "**Backoffice Girls**": Bedankt voor jullie meedenken en betrokkenheid, en voor een zeer aangename tijd waar ik met veel plezier op terug kijk!

Dank ook aan mijn roomies op de "**Achterste kamer**" tijdens mijn jaar op jullie lab: **Jos, Patrick J., Danique** en **Noa**: Bedankt voor een leuke tijd, het was meer dan gezellig, met regelmaat te gezellig! : -) Onze geïmproviseerde airco, met ijs van het lab in een piepschuim bak met ventilator, werkte verrassend goed, maar wat de warmte betreft ben ik blij de komende zomer weer terug te zijn op het CMBI haha.

Dan ben ik nu eindelijk aangekomen bij het beste paard van stal de "**Mooiboy**s"! Beste **Jos** "de boss" van Oss (Ja, Heesch, OK ..), aka Jos "de frikandellenkoning" Smits. Best **Wilbert** "Willie" Bouwman. Wat hebben we sinds onze Master samen een geweldige tijd meegemaakt! Diep vereerd ben ik dan ook, dat jullie mijn paranimfen willen zijn. En dat terwijl ik jullie hier nooit officieel voor heb gevraagd, want zoiets schijnt alleen geldig te zijn onder het genot van een door mij aangeboden speciaalbiertje, aldus de jongens. Wie weet komt het daar ooit nog van, net als naar de dierentuin gaan : -) Ondanks de onuitputtelijke hoeveelheid aan anekdotes als het om ons drieën gaat moet ik het kort houden, want dit dankwoord loopt al aardig uit de hand met mijn wollige schrijfstijl. Wat ik nog wel wil noemen, welke het meest memorabel is voor mij, en ik verwacht ook voor jullie, zijn onze ontelbare avonden na werk bij Camelot op de Grote Markt, de velen stempelkaarten die wij daar vol hebben gemaakt is wellicht ongeëvenaard, evenals de velen varkenssaté en Camelot-burgers die we daar hebben besteld (ze hadden mij best kunnen sponsoren met dit drukwerk denken jullie niet?). De spelletjes die daarbij op tafel kwamen mis ik overigens niet, de Da Vinci Code in het bijzonder .. Bij jullie is naast al het dollen ook altijd ruimte voor serieuzere stof, bijvoorbeeld over ons werk/onderzoek, waar wij door onze gezamenlijke achtergrond en interesses niet over uitgepraat raken. Best jongens het is mij een waar genoegen om met jullie op te trekken! Bedankt voor jullie gezelligheid, steun, en voor altijd een luisterend oor. Jullie onvoorwaardelijke kameraadschap is mij enorm veel waard! En veel liefs aan jullie **Joyce** en **Alinda**, want achter elk succesvolle paranimf staat natuurlijk een sterke vrouw.

Beste **Jos** (Smits), jou wil ik ook nog specifiek noemen in de context van collega-OIO. Ik vond het echt mega tof om met jou samen te werken op gezamenlijke projecten! In het bijzonder, jouw grote bijdrage aan het hoofdproject van mijn PhD (onderdeel van deze thesis) was erg belangrijk om tot een succesvolle validatie van de methode te komen. Heel erg bedankt voor je harde werk en inzet, het was echt leuk om met een goede vriend op deze manier samen te kunnen werken! Maar het absoluut hoogtepunt was natuurlijk afgelopen mei met "Mooiboy on Tour" naar Orlando (Florida) voor het

IID congres! En dan bedoel ik uiteraard vooral onze trip naar Miami en de Keys, want laten we eerlijk zijn, dat congres was maar matig ; -) Jos, bedankt voor alles, en we zijn voorlopig nog bezig met het peer-review proces van onze beide papers, en tegelijkertijd liggen er gelukkig alweer nieuwe projecten en ideeën op de stapel! En ondertussen ben je zelf ook bijna bij de eindstreep, veel succes met de afronding van jouw eigen PhD!

Beste mensen van het **NIZO**, ik heb altijd (en nog steeds!) een bijzonder leuke en leerzame tijd gehad bij jullie op de werkvloer in Ede. Het feit dat ik wekelijks een kijkje in jullie keuken kon nemen, en met jullie kon sparren en van jullie expertise kon leren heb ik altijd als een voorrecht ervaren. Jullie bekleden een vrij unieke niche in Nederland op gebied van microbiologisch onderzoek, en zijn een wetenschappelijk sterk team om trots op te zijn, hartelijk dank voor alles! Een aantal personen heb ik hierboven al genoemd, maar specifiek wil ik nog noemen de (oud) collega's van het Microbiome team: **Sacha, Sabina, Jos, Michiel, Wynand, Harro, Kim, Erik, Guus, Saskia**, en natuurlijk alle anderen van de Health afdeling waar ik tijdens de lunchmeetings en het wekelijks koffieoverleg veel van heb geleerd en plezier aan heb beleefd. NIZO'ers bedankt!

Beste **Martijn** en **Real**. Ik weet dat ik vooral het laatste jaar heel wat heb ingeleverd op het gebied van onze trainingen, maar verwacht dit spoedig weer op te kunnen pakken. Dank voor jullie regelmatige belangstelling in mijn werk, jullie algehele support vooral de afgelopen jaren, en ik zie er naar uit om met jullie weer op regelmatige basis lekker te chillen in Arnhem!

Lieve **OP** vrienden: **Leon, Dave, Kim, Rob, Erik, Lianne, Nick, Anne, Mark, Gislene, Elwin, Annefloor, Edwin** en **Michelle**! Ik denk dat de meeste van jullie weinig benul hebben wat ik nu precies heb uitgevoerd in het Radboudumc de laatste jaren, maar aan gebrek aan interesse lag het in ieder geval niet. Hopelijk kan mijn lekenpraatje jullie straks wat duidelijkheid bieden over mijn onderzoek van de afgelopen jaren, maar dat doet er verder ook weinig toe. Wat er wel toe doet is dat ik het elke keer weer bijzonder vind dat we het al zo lang met elkaar uithouden : -) en dat we er altijd weer een geweldige tijd op nahouden als we elkaar zien en spreken. Ik heb aan jullie een onwijs fijne groep vrienden om op terug te vallen, vrienden: bedankt!

Beste **Dave**, de eerste paar jaar van mijn PhD dronken we vrijwel elke week met de middagpauze koffie in het studentencafé. Ik heb daar nog altijd goede herinneringen aan, en vond het altijd fijn je bij te praten over van alles, ook over wat mij bezig hield op het werk. Gelukkig doen we dat nog steeds, ofschoon vanwege de afstand nu wat minder regelmatig. Bedankt maat, voor je nuchtere en relativerende kijk op de wereld, je velen adviezen en voor een luisterend oor! Beste **Mark** (Thielen), jij was denk ik de enige die echt begreep wat een promotietraject precies behelst, bedankt voor de fascinerende gesprekken die we elke keer weer hebben, over eigenlijk alles wat ons bezig houdt. Succes met de afronding van je eigen proefschrift, erg fijn te weten dat jij ook bijna bij de eindstreep bent! Beste **Erik, Nick, Rob, Leon**, met jullie is het eigenlijk altijd dolle pret maar we kunnen ook serieus zijn, bedankt voor de fijne momenten samen, voor jullie belangstelling in mij (mijn werk), en bovendien voor jullie warme

kameraadschap! Ik geniet elke keer weer enorm van de etentjes die we met regelmaat hebben, op naar de volgende!

Lieve **Stijn** en **Sarah**! Ik vind het heerlijk dat ik geregeld in Bennekom bij jullie op bezoek kan komen, bijvoorbeeld na werktijd bij NIZO, ik geniet er sowieso altijd enorm van om jullie te zien en om bij te praten. Beste Stijn, wij gaan al jaren terug (!), en heb bij jou dan ook meestal aan twee woorden genoeg, bedankt voor altijd een luisterend oor, je vele adviezen, en ook je onuitputtelijke wijsheden ; -) Ondanks dat je uiteraard niet volledig kan begrijpen wat ik werkinhoudelijk precies doe (wat maakt het ook uit), is het mij altijd wel duidelijk dat je op abstracter niveau begrijpt wat ik ervaar, en weet je dat perfect te vertalen naar je eigen referentiekaders. Maat, bedankt voor alle steun en begrip wat ik ten alle tijden van jou ontvang, het is me veel waard en elke keer weer een waar genoegen jou te zien en spreken!

Beste **Bas** (Meeuwissen) en **Priscilla**. Ook wij gaan al lang terug, en ik vind het een voorrecht om nog altijd goede vrienden aan jullie te hebben. Bas, tof dat als wij elkaar spreken je altijd zo betrokken en geïnteresseerd bent in mij en mijn werk. Bedankt en tot op de volgende borrel/koffie!

Beste **Mark** (Taris), je bent de laatste broeder die ik nog graag bedank! Opmerkelijk dat we elkaar bijna dagelijks spreken via *Teh Interwebs* en maar weinig IRL. Maar misschien juist daarom dat we altijd zo goed op te hoogte zijn van elkaar. En ook fijn dat we op die manier direct onze sores dan wel onze successen kunnen delen, wat in beide gevallen vrij therapeutische werkt moet ik bekennen. Maar gelukkig is er ook altijd ruimte voor random nonsens! Jouw geduld en steun is enorm, en ook ten tijden van deze PhD heb ik veel aan onze gesprekken gehad, en er zeer veel plezier aan beleefd, bedankt bro! Boks. T.

Tot slot, wil ik mijn familie hier " in het zonnetje zetten ". Het beste voor het laatste! Lieve Papa en Mama, zonder jullie onvoorwaardelijke steun en toeverlaat was dit alles echt aanzienlijk veel lastiger geweest. Jullie hebben mij altijd heel veel zorgen uit handen genomen waardoor ik mij volledig heb kunnen focussen op mijn studie, en later op het onderzoek. Als klein voorbeeld de ontelbare avonden waarin ik " in de flow " zat met analyses, programmeren of schrijven, waarbij ik in alle rust kon doorwerken om vervolgens bij jullie nog een opzijgelegde avondmaal te krijgen halverwege de avond. Jullie advies, steun en geduld is mij heel veel waard, evenals het feit dat ik altijd lekker bij jullie kan ventileren / reflecteren, waar ik nogal veel behoefte aan heb : -) Alles bij elkaar heeft me door soms lastige tijden geloodst. Als ik er doorheen zat kon ik me bij jullie weer opladen en een emotionele oppepper krijgen, en vaak was een luisterend oor voldoende maar dan waren jullie er wel altijd voor mij. Gezondheid, vrienden en familie is het allerbelangrijkste in het leven. Jullie zijn voor mij één van die belangrijke pijlers. Ik houd van jullie! Veel liefs van een trotse zoon!

Lieve familie Bolder en Ederveen, bedankt voor jullie belangstelling in mij, er is mij vaak gevraagd door jullie hoe ik er voor stond met mijn werk / onderzoek. Het is dan eindelijk zo ver, en ik vind het geweldig dat jullie er straks allemaal bij zijn op de promotiedag, dat betekend veel voor mij!

Lieve broeders en zusters, lieve Ellen, Koen en Lisa, en die andere twee jongemannen Frank en Andries, jullie allemaal heel erg bedankt voor de onuitputtelijke steun en belangstelling. Ik kan niets beters wensen dan jullie altijd dicht om me heen. Uit een groot gezin komen is voor mij echt een zegen : -)

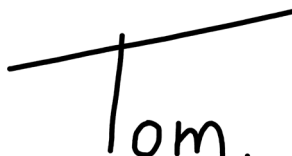
Koen, ik wil jou nog in het bijzonder bedanken! Want dankzij jou ligt er nu een prachtig proefschrift in onze handen die is vormgegeven door jou, speciaal voor mij! Zowel de cover als het binnenwerk / huisstijl heb jij zelf gedaan. Dat maakt het extra bijzonder voor mij. Beste Koen, daar ben ik enorm mee vereerd, je hebt er veel werk aan gehad en het resultaat mag er zijn! Dit proefschrift is daardoor iets waar ik nu dan ook extra trots op ben!

Aan iedereen die ik hier per abuis ben vergeten: Mea culpa ..

Hopelijk heb ik in deze sectie jullie aandacht wèl vast kunnen houden, maar dat betwijfel ik als ik teruglees wat voor " wall-of-text " ik hier wederom heb gegenereerd : -)

TL;DR : Iedereen bedankt!!

Cheers,
Thomas



Tom.

Research data stewardship and accessibility (FAIR)

DISTRIBUTION OF SOFTWARE,

Bioinformatics tools and pipelines resulting from this thesis :

COMPANION : <https://trac.nbic.nl/companion/>
TaxPhIA : <https://hub.docker.com/r/ederveen/taxphian/>

Other or related software available at **GitHub** : <https://github.com/ederveen/>

DATA AVAILABILITY,

The available 16S marker gene sequencing data and main characteristics, as used in this thesis.

Data used in Chapter	Host	Disease	Sample source	Location	Number of samples	Platform	Marker-gene	16S region	ENA identifier
CHPT. 3	human	atopic dermatitis	skin	inner elbow	142 16S (34 SLST)	Illumina MiSeq	16S / SLST on <i>Staphylococcus</i>	V3-V4	PRJEB27442
CHPT. 5	human	ichthyosis vulgaris	skin	lower leg	27	Roche 454	16S	V3-V4	PRJEB11661
CHPT. 6	mouse	arthritis (model)	intestine	feces (proxy)	40	Roche 454	16S	V5-V6	PRJEB7447
CHPT. 7	human	ADHD	intestine	feces (proxy)	96	Roche 454	16S	V3-V4	PRJEB11512
CHPT. 8	human	RSV-infection	naso-pharynx	nasal lavage	100	Illumina MiSeq	16S	V3-V4	PRJEB20811

List of publications

(sorted by date)

Chermprapai S., Ederveen T.H.A., Broere F., Broens E.M., Schlotter Y.M., van Schalkwijk S., Boekhorst J., van Hijum S.A.F.T., Rutten V.P.M.G., *The bacterial and fungal microbiome of the skin of healthy dogs and dogs with atopic dermatitis and the impact of topical antimicrobial therapy, an exploratory study*. **Vet Microbiol**, 2019, Feb;229:90-99

Ederveen T.H.A.*, Ferwerda G.*, Ahout I.M., Vissers M., de Groot R., Boekhorst J., Timmerman H.M., Huynen M.A., van Hijum S.A.F.T.*, de Jonge M.I.*, *Haemophilus is overrepresented in the nasopharynx of infants hospitalized with RSV infection and associated with increased viral load and enhanced mucosal CXCL8 responses*. **Microbiome**, 2018, Jan;6(1):10

Aarts E.*, Ederveen T.H.A.*, Naaijen J., Zwijs M.P., Boekhorst J., Timmerman H.M., Smeekens S.P., Netea M.G., Buitelaar J.K., Franke B., van Hijum S.A.F.T.*, Arias Vasquez A.*, *Gut microbiome in ADHD and its relation to neural reward anticipation*. **PLoS One**, 2017, Sep;12(9):e0183509

Rogier R.*, Ederveen T.H.A.*, Boekhorst J., Wopereis H., Scher J.U., Manasson J., Frambach S.J.C.M., Knol J., Garssen J., van der Kraan P.M., Koenders M.I., van den Berg W.B., van Hijum S.A.F.T., Abdollahi-Roodsaz S., *Aberrant intestinal microbiota due to IL-1 receptor antagonist deficiency promotes IL-17- and TLR4-dependent arthritis*. **Microbiome**, 2017, Jun;5(1):63

Zeeuwen P.L., Boekhorst J., Ederveen T.H.A., Kleerebezem M., Schalkwijk J., van Hijum S.A.F.T., Timmerman H.M., *Reply to Meisel et al. : Skin microbiome surveys are strongly influenced by experimental design*. **J Invest Dermatol**, 2017, Apr;137(4):961-962

Zeeuwen P.L.*, Ederveen T.H.A.*, van der Krieken D.A.*, Niehues H.*, Boekhorst J., Kezic S., Hanssen D.A., Otero M.E., van Vlijmen-Willems I.M., Rodijk-Olthuis D., Falcone D., van den Bogaard E.H., Kamsteeg M., de Koning H.D., Zeeuwen-Franssen M.E., van Steensel M.A., Kleerebezem M., Timmerman H.M., van Hijum S.A.F.T., Schalkwijk J., *Gram-Positive Anaerobe Cocci (GPAC) are Underrepresented in the Microbiome of Filaggrin-Deficient Human Skin*. **J Allergy Clin Immunol**, 2017, Apr;139(4):1368-1371

van der Krieken D.A.*, Ederveen T.H.A.*, van Hijum S.A.F.T., Jansen P.A., Melchers W.J., Scheepers P.T., Schalkwijk J., Zeeuwen P.L., *An In vitro Model for Bacterial Growth on Human Stratum Corneum*. **Acta Derm Venereol**, 2016, Nov;96(7):873-879

Ederveen T.H.A., Overmars L., van Hijum S.A.F.T., *Reduce Manual Curation by Combining Gene Predictions from Multiple Annotation Engines, a Case Study of Start Codon Prediction*. **PLoS One**, 2013, May;8(5):e63523

Ederveen T.H.A., Mandemaker I.K., Logie C., *The Human Histone H3 Complement Anno 2011*. **Biochim Biophys Acta**, 2011, Oct;1809(10):577-66

* equal contributions

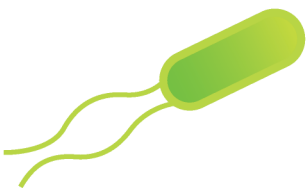
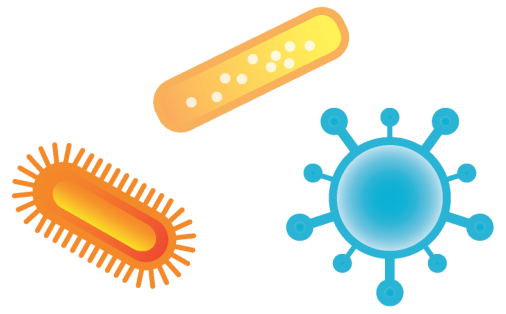
List of abbreviations

(sorted alphabetically)

- 16S**: the 16S ribosomal RNA (rRNA) gene used for metataxonomics analysis;
5-HTP: 5-hydroxytryptophan;
AD: atopic dermatitis;
ADHD: attention-deficit / hyperactivity disorder;
AGE: automated genome annotation engine;
AMP: antimicrobial protein;
ANOVA: analysis of variance;
APC: allophycocyanin;
ATCC: American Type Culture Collection;
BF: Bonferroni (multiple-testing correction);
BIG: Brain Imaging Genetics study;
BMI: body mass index;
BOLD: blood oxygenation level-dependent;
BRU: broad range universal;
CAMP: see *LL37*;
CCD: (Centrale) Commissie Dierproeven;
CCL: (C-C motif) chemokine ligand (CCL5, CCL20, a.o.);
CD: cluster of differentiation;
CD: Crohn's disease;
cdNA: copy DNA;
CDT: cyclohexadienyl dehydratase;
CFU: colony forming units;
Chao1: Chao1 species richness measure;
CMO: (Centrale) Commissie Mensgebonden Onderzoek;
COMPANION: bioinformatics tool for COMPArative geNome annotatIOn;
CR: conserved region;
CXCL: (C-X-C motif) chemokine ligand (CXCL8, CXCL10, a.o.);
DAMP: damage-associated molecular pattern;
DEC: dierexperimentencommissie;
DKO: double knock-out;
DM: Distmat (from EMBOSS), to create a distance matrix from multiple alignments;
DSM: Diagnostic and Statistical Manual of Mental Disorders;
DSMZ: Deutsche Sammlung von Mikroorganismen und Zellkulturen (German Collection of Microorganisms and Cell Cultures);
EDTA: ethylenediaminetetraacetic acid;
ELISA: enzyme-linked immunosorbent assay;
ENA: European Nucleotide Archive;
FACS: fluorescence-activated cell sorting;
FB: FACS buffer, see *FACS*;
FCS: foetal calve serum;
FDR: false discovery rate (multiple-testing correction);
FIG: formimino glutamate;
FITC: fluorescein isothiocyanate;
FLG: filaggrin;
fMRI: functional magnetic resonance imaging;
FoxP3: forkhead box P3;
FWE: family-wise error rate;
gDNA: genomic DNA;
GM-CSF: granulocyte-macrophage colony-stimulating factor;
GPAC: Gram-positive anaerobic cocci;
hBD: human beta-defensin (hBD2, hBD3, a.o.);
HC: healthy control;
HIES: Hyper-IgE Syndrome;
HMP: Human Microbiome Project;
hut: histidine utilization gene pathway;
hutF: gene encoding encoding FIGdeiminase;
hutG: gene encoding FIGase;
hutH: gene encoding histidase;
hutI: gene encoding IPase;
hutU: gene encoding urocanase;
HV: healthy volunteer;
IBD: inflammable bowel disease;
IFN: interferon (IFN a.o.);
IL: interleukin (IL-1, IL-6, IL8, a.o.);
IL-1R: interleukin-1 receptor;
IL-1Ra: interleukin-1 receptor antagonist;
IL1rn: gene encoding IL-1Ra;
ILC: innate lymphoid cell;
IPase: imidazolone-5-propionate hydrolase, see *hutI*;
iTOL: interactive tree of life, bioinformatics tool for visualization of phylogenetic trees;
IV: ichthyosis vulgaris;
IVIS: *in vivo* imaging system;
KEGG: Kyoto Encyclopedia of Genes and Genomes;
K-number: KEGG ortholog identifier, see *KEGG*;
KO: KEGG ortholog, see *KEGG*;
KO: knock-out;
KOD: *Thermococcus kodakaraensis*;
K-SADS: Kiddie Schedule for Affective Disorders and Schizophrenia;
LC: Langerhans cell;
LL37: cathelicidin antimicrobial peptide LL37, see *CAMP*;
LOF: loss-of-function mutations;
LP: lamina propria;
LPL: lamina propria lymphocytes;
Mb: mega base pairs;
MCL: Markov cluster algorithm;
MIC: minimal inhibitory concentration;
ML: maximum likelihood;
MLST: multi-locus sequence typing;
MMP: matrix metalloproteinase (MMP9, a.o.);

mORF : meta open reading frame;	SLST : single-locus sequence typing;
MR : magnetic resonance;	SNP : single nucleotide polymorphism;
mRNA : messenger RNA;	SPM : statistical parametric mapping, software for analysis of brain imaging data sequences;
MWU : Mann-Whitney U statistical test;	SVC : small volume correction;
NA : not available / applicable (N/A);	TaxPhlAn : bioinformatics pipeline for TAXonomy PhyLogenetic ANalysis;
NGS : next-generation sequencing;	Tbet : T-box transcription factor TBX21;
NIH : National Institutes of Health;	TEWL : transepidermal water loss;
NIZO : Nederlands Instituut voor Zuivelonderzoek;	Th : T helper;
NLRP6 : nucleotide-binding oligomerization domain, leucine-rich repeat and pyrin domain containing 6;	TLR : Toll-like receptor;
NMF : natural moisturizing factor;	TNF : tumor necrosis factor (TNF α , a.o.);
NOD : non-obese diabetic;	Treg : T-regulatory;
NPA : nasopharyngeal aspirate;	UCA : urocanic acid;
o/n : overnight;	UMC : University Medical Center;
OG : orthologous group;	UPGMA : unweighted pair group method with arithmetic mean, a hierarchical clustering method;
ORF : open reading frame;	V1-V2 : variable regions V1 and V2 of the 16S rRNA gene, idem. V3-4 and V1-9 ;
OTU : operational taxonomic unit;	VR : variable region;
P/I/G : PMA (phorbol myristate acetate), ionomycin, brefeldin A;	WT : wild-type.
P/I : PMA (phorbol myristate acetate), ionomycin;	
PAMP : pathogen-associated molecular pattern;	
PBS : phosphate buffered saline;	
PC : principal component;	
PCA : principal component analysis;	
PCA : pyrrolidone carboxylic acid;	
PCoA : principal coordinate analysis;	
PCR : Polymerase chain reaction;	
PD : phylogenetic diversity;	
PDWT : phylogenetic distance whole tree, alpha diversity metric;	
PE : phycoerythrin;	
PICRUSt : Phylogenetic Investigation of Communities by Reconstruction of Unobserved States;	
PICU : pediatric intensive care unit;	
PMA : propidium monoazide;	
QIIME : quantitative insights into microbial ecology;	
qPCR : quantitative (real-time) PCR, see <i>PCR</i> ;	
RDA : redundancy analysis;	
RDP : Ribosomal Database Project;	
RFD : Robinson-Foulds distance;	
ROI : region of interest;	
RORyt : retinoic acid receptor-related orphan receptor gamma t;	
RPMI : Roswell Park Memorial Institute medium;	
rRNA : ribosomal RNA;	
RSV : respiratory syncytial virus;	
RT : response time;	
RUMC : Radboud University Medical Center;	
RV : rhinovirus;	
SA : <i>Staphylococcus aureus</i> ;	
SC : <i>Staphylococcus capitis</i> ;	
SC : stratum corneum;	
SD : standard deviation;	
SDI : Shannon diversity index;	
SEM : standard error of the mean;	
SFB : segmented filamentous bacteria;	
SI : small intestine;	

notes



WEB



EDERVEEN.SCIENCE

ISBN 978-94-93118-09-6