Norwegian University of Life Sciences
Faculty of Chemistry, Biotechnology and Food Science

Philosophiae Doctor (PhD)
Thesis 2022:13

Development of Liquid Array Diagnostic (LAD) technology for prediction of human gut microbiota composition and functionality

Bruk av Liquid Array Diagnostics (LAD) som verktøy for analyse av sammensetning og funksjon av tarmens mikrobiota

Pranvera Hiseni

Development of Liquid Array Diagnostic (LAD) technology for prediction of human gut microbiota composition and functionality

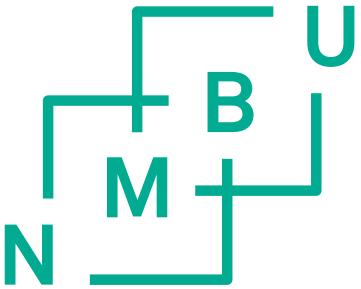
Bruk av Liquid Array Diagnostic (LAD) som verktøy for analyse av sammensetning og funksjon av tarmens mikrobiota

Philosophiae Doctor (PhD) Thesis

Pranvera Hiseni

Norwegian University of Life Sciences
Faculty of Chemistry, Biotechnology and Food Science

Ås (2022)



Thesis number 2022:13
ISSN 1894-6402
ISBN 978-82-575-1889-9

PhD supervisors

Knut Rudi, PhD
Faculty of Chemistry, Biotechnology and Food Science
Norwegian University of Life Sciences
Christian Magnus Falsens vei 18,
1433 Ås, Norway
knut.rudi@nmbu.no

Robert C. Wilson, PhD
Faculty of Applied Ecology, Agricultural Sciences and Biotechnology
Department of Biotechnology
Inland Norway University of Applied Sciences
P.O. Box 400
2418 Elverum, Norway
robert.wilson@inn.no

Lars Snipen, PhD
Faculty of Chemistry, Biotechnology and Food Science
Norwegian University of Life Sciences
Christian Magnus Falsens vei 18,
1433 Ås, Norway
lars.snipen@nmbu.no

Finn Terje Hegge, PhD
Genetic Analysis AS
Kabelgata 8,
0580 Oslo, Norway
fthegge@gmail.com

Kari Furu, PhD
Genetic Analysis AS
Kabelgata 8,
0580 Oslo, Norway
kf@genetic-analysis.com

PhD Evaluation Committee

Reetta Satokari, PhD
Faculty of Medicine
Human Microbiome Research Program
University of Helsinki
P.O. Box 4 (Yliopistonkatu 3)
00014 University of Helsinki, Finland
reetta.satokari@helsinki.fi

Jon Bohlin, PhD
Norwegian Institute of Public Health
Department of Method Development and Analytics
P.O. Box 222 Skøyen
N-0213 Oslo, Norway
jon.bohlin@fhi.no

Harald Carlsen, PhD
Faculty of Chemistry, Biotechnology and Food Science
Norwegian University of Life Sciences
Christian Magnus Falsens vei 18,
1433 Ås, Norway
harald.carlsen@nmbu.no

Acknowledgements

This industrial PhD project was funded by the Norwegian Research Council (project number 283783), for which I am highly appreciative.

My deepest gratitude goes to my academic supervisors, Knut Rudi, Rob Wilson, Lars Snipen, and my supervisors at Genetic Analysis AS (GA), Kari Furu and Finn Terje Hegge. Thank you for your guidance and continuous support. Thank you for inspiring me with your work. I consider myself fortunate to have had the chance to work with you and learn from you.

I am grateful to all my colleagues at GA. Thank you all for providing me with a warm, supporting, and motivating environment during this journey. I am especially thankful to Graceline Tina Kirubakaran for inspiring my work on the 16*S* rRNA gene intragenomic variation. What I initially thought would be a simple answer to a question she posed, led to a valuable finding worthy of journal publication.

I am also appreciative of the MiDiv lab team at NMBU. I am particularly grateful to Inga Leena Angell, who welcomed me from day one, showed me around campus, helped me in the lab, taught me helpful lab techniques, and with whom I had endless inspirational conversations.

This journey would not have been as successful without my husband Kushtrim by my side. Thank you for encouraging me to reach my goals.

In the end, I want to dedicate this thesis to my loving parents. Earning a PhD degree was as much my dream as it was theirs. My father taught me to aim high and trust my skills, while no one better than my mother demonstrated the value of hard work and persistence to me. No written line will ever express my gratitude towards them; however: mam, bab, faleminderit!

Table of Contents

Abbreviations and Definitions

LAD	Liquid Array Diagnostics, a qPCR-compatible method designed to detect multiple gene variants in a single-tube reaction
qPCR	Quantitative Polymerase Chain Reaction
NTC	No Template Control
LP	Labelling probe, an oligonucleotide complementary [only] to a specific target gene, capable of becoming extended with a quencher-labelled ddNTP
RP	Reporter probe, an oligonucleotide complementary to a specific LP, labelled with a fluorophore molecule on the 5'-end.
Quencher	A substance that absorbs the emitted light from a nearby fluorophore
ddNTP	Dideoxynucleoside triphosphate, a synthetic triphosphate nucleoside lacking a 3'-OH.
ddCTP-Q	Quencher-labelled ddCTP
P:B	Propionate-to-butyrate molar ratio
SCFA	Short-Chain Fatty Acid, bacterial fermentation end-product
GC	Gas Chromatography, a method used for separating and analyzing the chemical compounds of a mixture
OTU	Operational Taxonomic Unit. Closely related microorganisms sharing high nucleotide identity (usually 97.5% for 16S rRNA gene sequences)
PLS	Partial Least Squares analysis, a statistical method that reduces the dimensions of multicollinear predictors
LDA	Linear Discriminant Analysis, a classification approach based on features that best separate two or more classes
ANI	Average Nucleotide Identity
HumGut	A comprehensive collection of dereplicated human gut prokaryotic genomes
HumGut_975	The original HumGut collection consisting of >30,000 genome cluster representatives dereplicated at ≥ 97.5% ANI
HumGut_95	A coarser HumGut collection made of >5,000 genomes dereplicated at 95% ANI
MAG	Metagenome-Assembled Genome, a bioinformatically, reference-free, de-novo assembled genome
RefSeq	A Reference database of curated, non-redundant sequences, built by the National Center for Bioinformatic Information (NCBI)

List of papers

Paper I

Pranvera Hiseni, Robert C. Wilson, Ola Storrø, Roar Johnsen, Torbjørn Øien, Knut Rudi. **Liquid array diagnostics: a novel method for rapid detection of microbial communities in single-tube multiplex reactions**. *BioTechniques*, 2019. 66(3): p. 143-149.

Paper II

Pranvera Hiseni, Knut Rudi, Robert C. Wilson, Finn Terje Hegge, Lars Snipen. **HumGut: a comprehensive human gut prokaryotic genomes collection filtered by metagenome data**. *Microbiome*, 2021. 9(1): p. 165.

Paper III

Pranvera Hiseni, Lars Snipen, Robert C. Wilson, Kari Furu, Knut Rudi. **Questioning the quality of 16S rRNA gene sequences derived from human gut metagenome-assembled genomes**. *Frontiers in Microbiology*, 2022. 12:822301.

Paper IV

Pranvera Hiseni, Lars Snipen, Robert C. Wilson, Finn Terje Hegge, Knut Rudi. **Prediction of high fecal propionate-to-butyrate ratios using 16S rRNA-based detection of bacterial groups with Liquid Array Diagnostics**. *Manuscript under review in BioTechniques.*

Abstract

The microbial species residing in the human gut exercise vital functions for the host. They produce different metabolites that are crucial for human wellbeing. A variety of such molecules mediate signalling along the gut-brain axis, regulate host gene expression, develop and maintain intestinal and blood-brain barriers, are involved in lipogenesis and gluconeogenesis, in addition to taking part in a wide range of other functions.

A deviation in the intestinal flora composition is mechanistically linked to various health disorders, including inflammatory bowel disease (IBD), irritable bowel syndrome (IBS), type 2 diabetes, Parkinson's and Alzheimer's disease. Such a deviation, known as *dysbiosis*, represents an unbalanced composition where certain microbial groups are promoted in the expense of others. These species are considered as promising biomarkers, valuable for disease diagnosis, monitoring and treatment. Of particular interest are those markers that can additionally unveil phenotypical characteristics, such as the overall level of short-chain fatty acids (SCFA) in human gut samples. The prospect of discovering additional markers is high, considering that the content of healthy human guts worldwide is not fully characterized.

The field of gut microbiota is at a stage of switching focus to clinically relevant species, particularly to their rapid detection, as a means of offering simple diagnostic solutions with increased availability and accessibility. This affords putting biological findings to practical clinical use, which is often not feasible with current species identification platforms.

With the intention of filling this need, the main aim of this thesis was to develop a targeted approach for rapid gut microbiota testing based on the novel Liquid Array Diagnostics (LAD) technology. LAD is adopted to target 16S rRNA gene sites unique for specific microbial groups. Requiring only commonplace qPCR instrumentation, it can detect up to 30 distinct microbial markers in a single-tube multiplex reaction within a working day. LAD's utility in microbiome studies was validated by testing the prevalence and abundance of 15 microbial markers in 541 samples collected from mothers and their children, as reported in **Paper I**.

Paper II, on the other hand, describes a comprehensive human gut prokaryotic genome collection, HumGut. It was built after screening thousands of human gut metagenome samples, collected from healthy people worldwide, for the presence of any high quality publicly available prokaryote genome. The main rationale for creating it was to enable functional studies through LAD-based 16S targeting.

It was demonstrated that HumGut, as a reference database, aids whole genome sequencing studies by significantly increasing the number of mapped sequencing reads, thus elevating the

potential for an improved taxonomic classification. However, as it is, HumGut exhibits limited practical use for 16S rRNA gene targeted approaches like LAD. This because most of the representative genomes either lack this gene, or the quality of 16S sequences is compromised (addressed in **Paper III**).

Nonetheless, LAD was exploited to infer a segment of human gut microbiota functionality by targeting the 16S rRNA gene. This was performed based on data retrieved from 16S rDNA sequencing and short-chain fatty acid (SCFA) measurements. LAD's value in classifying samples with disturbed SCFA ratios (namely high propionate-to-butyrate ratio) - an indication of functional dysbiosis - is presented in **Paper IV**.

Taken together, this thesis introduces two tools, LAD and HumGut, both pointing at the direction of simplified human gut functional analysis via gut microbial composition detection.

Sammendrag

De mikrobielle artene som bor i menneskets tarm utøver vitale funksjoner for verten. De produserer forskjellige metabolitter avgjørende for menneskers helse. En rekke av disse molekylene deltar i prosesser som signaltransduksjon langs tarm-hjerne-aksen, regulering av genekspresjon, utvikling og vedlikehold av tarm- og blod-hjerne-barrieren, lipogenese og glukoneogenese, samt en rekke andre funksjoner.

Avvik i tarmflorasammensetningen kan knyttes til mange ulike sykdommer og lidelser, inkludert irritabel tarm (IBS), innflammatorisk tarmsykdom (IBD), type -2 diabetes, Parkinsons og Alzheimers sykdom. Slike avvik, kjent som dysbiose, kjennetegnes av at visse mikrobielle grupper fremmes på bekostning av andre.

Disse artene har potensiale som biomarkører, og kan slik være verdifulle for sykdomsdiagnose og behandling. Spesielt lovende er biomarkører i tarm som kan knyttes opp mot phenotypiske trekk, slik som kortkjedede fettsyrer (SCFA). Det antas at enda flere slike arter vil identifiseres i fremtiden, da mikrobiota-komposisjonen i sunne tarmer ikke er fullt karakterisert globalt.

Mikrobiota-feltet er nå på et stadium hvor fokuset endres fra eksplorative studier til identifisering av klinisk relevante arter. Det vil da bli spesielt viktig med metoder som muliggjør rask deteksjon, da dette vil innebære enkle diagnostiske løsninger tilgjengelig for praktisk klinisk bruk, noe som ofte ikke er gjennomførbart med dagens artsidentifikasjonsplattformer.

Hovedmålet med denne oppgaven var å utvikle en målrettet tilnærming for rask tarmmikrobiotatesting basert på det nye Liquid Array Diagnostics (LAD)-prinsippet. LAD er utviklet for å identifisere sekvenser i 16S rRNA-genet som er unike for spesifikke mikrobielle markører. Metoden krever kun et vanlig qPCR-instrument og kan oppdage inntil 30 forskjellige mikrobielle markører i ett enkelt test-rør i løpet av en arbeidsdag. LADs nytteverdi i mikrobiomstudier ble validert ved å teste forekomsten av 15 mikrobielle markører i 541 prøver samlet fra mødre og deres barn, som rapportert i **Artikel I**.

Artikel II beskriver genereringen av en omfattende prokaryot genomsamling av menneskets tarm. Den ble bygget ved å screene tusenvis av metagenom fra tarmprøver samlet inn fra friske mennesker over hele verden. Metagenomene ble screenet for tilstedeværelse av alle offentlig tilgjengelige prokaryote genom. Sekvenser av dårlig kvalitet ble fjernet mens alle andre sekvenser ble samlet i én stor referansedatabase, HumGut. Hovedmålet med å lage denne referansedatabasen var å muliggjøre LAD-baserte funksjonelle studier.

Det ble vist at HumGut fungerer som et nyttig verktøy for full-genoms sekvenseringsstudier ved å øke antallet kartlagte sekvenseringsavlesninger betydelig, da dette gir forbedret taksonomisk

klassifisering. HumGut har imidlertid begrenset nytteverdi for 16S rRNA-baserte metoder som LAD. Dette fordi de fleste genom i samlingen enten mangler dette genet fullstendig, eller har for dårlig kvalitet på 16S-sekvensene (behandlet i **Artikel III**).

Til tross for begrensningene knyttet til 16S rRNA-genet i HumGut, ble LAD benyttet til å utvikle en 16S rDNA-basert test for måling av menneskelig tarmmikrobiotafunksjonalitet. Dette ble utført basert på data hentet fra 16S-sekvensering og målinger av kortkjedede fettsyrer (SCFA). LADs evne til å klassifisere prøver med forstyrret SCFA-forhold (nemlig høyt propionat-til-butyrat-forhold) - en indikasjon på funksjonell dysbiose - er presentert i **Artikel IV**.

Til sammen presenterer denne oppgaven to verktøy, LAD og HumGut, som begge peker i retning av forenklet funksjonell analyse av menneskelig tarm via deteksjon av mikrobiell sammensetning i tarmen.

Introduction

Human bodies are a natural habitat to various microbial species, the majority of which reside in the colon[1]. The composition of gut microbes is highly dynamic throughout different stages of life[2, 3], and distinct among different individuals[4]. Their assemblage supports the physiological needs of the host while being shaped by host genetics[5, 6], delivery mode[7, 8], breastfeeding[9, 10], antibiotic usage[11, 12], diet[13, 14] and lifestyle[15, 16]. Humans and gut microorganisms (gut microbiota) coexist in a tight and lifelong symbiotic relationship, the downfall of which is thought to manifest itself through different health disorders. Such a downfall, termed *dysbiosis*, represents a disbalanced microbial composition mainly associated with diversity loss, overgrowth of pathogenic strains and/or depletion of health-promoting microbes[17].

The connection between health disorders and the gut was postulated by Hippocrates in ancient Greece, more than two millennia ago ('*All disease begins in the gut*')[18]. Yet not much light was thrown onto this field for many centuries to come. The presence of microorganisms ('*animalcules*') in stool samples was reported by Antonie van Leeuwenhoek back in the 17th century[19]. But it was not before the significant findings of Louis Pasteur[20] and Robert Koch[21] in the late 1800's, establishing a causative link between bacteria and infectious diseases, that the field of microbiology would gain momentum. The gut microbiota studies boost we witness today can certainly be attributed to further crucial discoveries made in the mid-20th century.

A successful protocol for culturing anaerobic prokaryotes, brought by Hungate in 1950[22], laid the groundwork for a suitable investigation of obligate anaerobes residing in the gut. This created the possibility of culturing previously uncharacterized intestinal microorganisms, increasing the scientific awareness about their overall richness.

Mid-20th century was also a time when James Reyniers established routines on rearing germ-free rats for successive generations[23], facilitating extensive experimental research in the field. Ever since, studies performed on microbiologically sterile animals or gnotobiotic ones (i.e. animals with fully identified microbial composition), have proven the impact of microbes on host gut morphology[24], vitamin production and deficiency[25, 26], bilirubin degradation[27], drug metabolism[28], brain development[29], complex microbial species interaction[30], etc.

In addition, fecal microbiota transplants from humans to different animal models have demonstrated undisputable links between microbial composition and the onset of various inflammatory systemic diseases. Mice transplanted with feces from individuals suffering from autism spectrum disorder (ASD)[31], multiple sclerosis[31], Parkinson's disease[32], or obesity[33], were

shown to exhibit clinical signs concordant with those observed in diseased humans, opening a whole new chapter of modern medicine.

Dysbiosis as an indicator of disease

Accumulating evidence show a profound association between dysbiosis and a wide range of health disorders, including irritable bowel syndrome (IBS)[34, 35], inflammatory bowel disease (IBD)[36, 37], colon cancer[38, 39], non-alcoholic fatty liver disease (NAFLD)[40, 41], major depression disorder[42, 43], ASD[44-46], Parkinson's disease[47, 48], obesity[49-51], etc.

The state of dysbiosis, characterized by a disbalanced composition of microbial communities, signifies a modified gut architecture mechanistically linked to disease generation and progression[52]. Although a clear definition of a '*balanced microbial community*' is challenging to be composed, accrued data in the field have revealed some important bacterial species thought to act as gatekeepers of human wellbeing, in addition to identifying some key microbes involved in disease pathogenesis. It is, however, important to keep the emphasis on the *balance* as a holistic concept, given that increased levels of beneficial bacteria may also constitute dysbiosis.

For instance, the richness of two of the most abundant and well-studied human gut colonizers, *Akkermansia muciniphila* and *Faecalibacterium prausnitzii*, is commonly linked to a good health status[53-55]. The former is a typical mucin-degrading species[56], while the latter an essential producer of the beneficial butyrate[57]. Many studies have shown that a decrease in either of them is associated with a compromised gut barrier integrity and elevated proinflammatory response, observed in a wide range of systemic diseases[54, 58, 59]. It has been demonstrated that supplementation with *A. muciniphila* helps obese people lose weight, improves their insulin sensitivity and reduces their cholesterol levels[60], pinpointing the involvement of this species in metabolic disease progression and prevention.

Nevertheless, significantly elevated levels of these microorganisms should be viewed with caution. An *Akkermansia muciniphila*-rich dysbiosis has been observed in patients with type 2 diabetes[61], possibly confounded by the usage of metformin drug which was shown to promote this species overgrowth[62]. In addition, increased levels of *F. prausnitzii* were reported in a cohort of obese children[63], proving that the association of these bacteria to a good health status is context-dependent.

A wide range of butyrate-producers like *Blautia faecis*, *Roseburia hominis, Roseburia inulinivorans,* and *Ruminococcus torques* are regarded as important bioindicators as well[64-66]. Butyrate as a fermentation end-product, is the main energy source for gut epithelial cells and a potent epigenetic regulator[67], hence the link between the collective reduction of these species

and IBD and colorectal cancer[64-66]. However, *Ruminococcus torques* has been observed to be enriched in a variety of unrelated cohorts, including autistic children suffering functional gastrointestinal diseases[68] and patients affected from age-related macular degeneration[69], yet again demonstrating the difficulty of characterizing dysbiosis by taking a limited number of markers into account.

Another important group of bacteria strongly involved in maintaining human health is represented by various *Bifidobacteria* and *Lactobacillus* species. Among other functions, they are known to mediate the gut-brain axis through the production of the inhibitory neurotransmitter gamma aminobutyric acid[70]. Their depletion is reported in individuals suffering from depression[71], while concurringly, volunteers supplemented with *B. longum* and *L. helveticus* strains were shown to experience beneficial psychological effects[72].

On the other hand, the abundance of many other bacterial species is known to generally exhibit adverse effects on human health. Such is the case for *Fusobacterium nucleatum*, described for its ability to increase proliferation of colorectal cancer cells, acting as a potential sole marker for disease diagnosis[73, 74]. Similarly, a well-documented indicator of poor health is Proteobacteria overgrowth. This phylum represents a group of facultative anaerobes, prevailing in conditions of gut epithelial oxygenation that is otherwise detrimental to most commensal microorganisms[75]. Their elevated quantities signify a gut barrier dysfunction linked to a high inflammation rate observed in a wide range of disorders[76].

Taken together, these findings show the immense medical value of the gut microbiota composition as a whole in understanding, diagnosing, treating, and preventing a broad list of human illnesses.

Microbial metabolites and human health

As microorganisms break down food particles and other substrates, they produce a complex web of metabolites that either act locally in the gut environment, or reach other body organs through systemic circulation, exercising a direct impact on host physiological reactions[77]. Because of this, exploring the direct relationship between gut microbial metabolites and human health is increasingly gaining attention[78-80]. A deviation of metabolite flow gives biological meaning to dysbiosis as it provides a clear mechanism for the miscommunication between intestinal flora and the host. It also raises the prospect of designing drugs aimed to help compensate for the lack of beneficial molecules or counteract the harmful ones[81-83].

One of the best studied group of metabolites are short-chain fatty acids (SCFAs). These represent bacterial fermentation end-products, known to mediate the gut-brain axis communication[84],

regulate the satiety levels[85], serve as main energy source for colonocytes[86], influence the blood-brain barrier formation[29], act as crucial molecules in lipid metabolism[87], and regulate gluconeogenesis[88], among other functions. In fecal samples collected from healthy subjects, three main SCFAs, acetate, propionate, and butyrate, have a 3:1:1 molar ratio, respectively[89]. However, in samples collected from ill people, the proportions are often deviated, possibly reflecting an overgrowth or depletion of certain functional bacteria groups. Decreased levels of butyrate and/or increased levels of propionate are of special interest. They are steadily being linked to a wide range of diseases, including IBS[90, 91], Alzheimer's disease[92] and high risk of stroke[93].

Other important metabolites produced by gut microbiota include vitamins of groups B and K, substantially contributing to the required vitamin levels for regular cellular functions[94, 95]. The list of metabolites also includes methylamines, products of L-carnitine and choline degradation, known for their association with metabolic and cardiovascular diseases[96, 97]. Specific members of intestinal flora are a great source of essential branched-chain amino acids (leucine, isoleucine, and valine), the increased circulation levels of which have been linked to insulin resistance[98]. Other microorganisms regulate tryptophan metabolism leading to the production of serotonin, an essential neurotransmitter involved in various bodily functions[99]. Gut microbiota is also involved in the production of secondary bile acids, indole, and polyphenol derivatives, each implicated in a variety of human physiological systems[77, 100].

The gut metabolome (i.e. all small molecules produced in the gut) is vastly rich and highly diverse[101], far exceeding the list of metabolites exemplified here. The intricate interactions between gut flora metabolites, other microorganisms in the gut and human body itself are an active subject of modern research, showing signs of promising potential for future pharmacological solutions[81-83].

Techniques for human gut microbiota detection

There exist numerous tools specialized for microbial identification. Besides culturing, traditional techniques employ microscopy, serology and/or biochemical tests, which are aimed at exploring phenotypical traits of the isolated microbes in addition to establishing causality of disease[102]. In contrast to traditional methods, modern techniques are mostly invested in genotyping, especially after recognizing that a large number of microorganisms are challenging - perhaps even impossible to culture[103].

Indeed, the accelerated speed of discovery in microbiota studies came after the elucidation of DNA structure in 1953 by Watson and Crick[104], one of the most important achievements of modern science. The unravelling of genetic code allowed for the development of various

techniques, such as DNA sequencing and Polymerase Chain Reaction (PCR)[105], extensively used in contemporary laboratories for microbial species characterization and detection.

Continuous advancement of Next Generation Sequencing (NGS) techniques in the past 20 years, paired with uninterrupted development and refinement of bioinformatic tools, has made it possible to explore gut microbiomes at a rate never observed before. In line with this, a correspondingly remarkable increase in the annual publication rate of gut microbiota/microbiome/bacteria research is observed in the MEDLINE database accessed through PubMed (**Figure 1**). A similar increase of scientific interest is also observed in research projects related to skin[106], oral[107] and vaginal[108] flora.

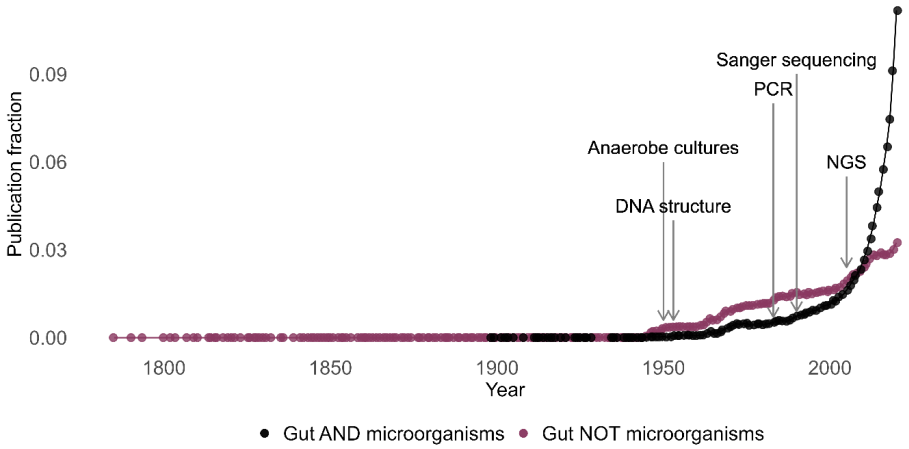


Figure 1. PubMed publication timeline. Years are presented along the x-axis; the y-axis shows the fraction of total publications per category. Black dots (connected with a black line) depict the fraction of publications related to microorganisms and gut; the purple ones show the fraction of publications about gut alone. The former was searched using the following query: ***(gut microbiome) OR (gut microbiota) OR (intestinal microbiota) OR (intestinal microbiome) OR (gut bacteria) OR (intestinal bacteria)****, and the latter with* ***(gut OR intestine) NOT (microbiota) NOT (microbiome) NOT (bacteria)****. The timeline includes and ends with publications from 2020. Key points in time, driving the acceleration of gut microbiome studies, are highlighted (1950: successful cultivation of anaerobic bacteria, 1953: elucidation of DNA structure, 1983: invention of Polymerase Chain Reaction (PCR), 1990: Sanger sequencing, and 2005: Next Generation Sequencing (NGS)).*

In an era of increasingly large amounts of sequencing data, the field is well-equipped to switch to utilizing targeted approaches, shifting the focus to detecting microbial markers of clinical interest only. This serves to reduce expenses, the workload, and the amount of unnecessary and noisy data produced after each diagnostic test, while at the same time elevating the standardization potential by simplifying interpretation of results.

Targeted approaches make use of the existing knowledge about the genetic content of a microorganism of interest. They target DNA sites unique for certain species to infer their presence in a sample. In conventional microbiota studies, the gene encoding 16*S* rRNA represents the most often targeted segment. This gene is omnipresent and highly conserved among prokaryotes, yet contains signature variations for particular taxa[109], rendering it an ideal candidate for targeted methods.

Targeting is typically performed by employing short oligonucleotides (probes), complementary to signature DNA sequences, who generate a type of detectable signal in the presence of their template[110-113].

A wide range of methods, commonly used in contemporary laboratories, are based on signal detection through a qPCR machine[110-113]. Similarly, these methods can be adapted to detection through a more recently developed instrumentation, digital droplet PCR (ddPCR)[114]. They represent excellent tools, capable of detecting and quantifying the amount of target DNA in any environmental sample. However, their multiplexity level is limited mostly to the number of detection channels possessed by the instrumentation, meaning that currently, a typical reaction cannot detect more than a maximum of six target species or groups in a tube (**Table 1**). This presents a severe limitation in a field rich of interacting microorganisms.

Table 1. Characteristics of some of the most commonly used targeted approaches

Method	**Required instrumentation**	**Reporting**	**Multiplexity level***	**Washing step required**
Molecular Beacons®[110]	qPCR	Release of fluorophores from quenching	6	No
KASP[111]	qPCR	Release of fluorophores from quenching	6	No
Taqman®[112]	qPCR	Release of fluorophores from quenching	6	No
EvaGreen®[113]	qPCR	Intercalating dye fluorescing	1	No
SNPlex[115]	Capillary electrophoresis	In-capillary size-dependent fluorescing	96	Yes
GA-map®[34]	Flow cytometer	Fluorescing of bead-coupled probes	100	Yes

To date, the only validated platform for multiplexed microbiota targeted detection is GA-map®[34]. It has a capacity of detecting up to 100 markers in a single tube using a flow cytometer

* Assuming six detection channels for methods requiring a qPCR instrumentation

for detection. However, similar to other methods that allow the detection of multiple targets in a sample, it requires dedicated instrumentation and fragile sample processing[34, 115], rendering routine tests and experiments expensive.

Persisting challenges in the field

Producing worldwide medical solutions presents a major challenge, regardless of rapid developments. That is because there is a huge variability of microbial composition among healthy people to start with[116]. In addition, reconciling disparate conclusions drawn by different studies is often not feasible. Most research groups use their own set of local control samples (against which to compare the species levels), follow a different sample-handling protocol[117], and utilize different bioinformatic tools and reference databases to interpret results[118]. The latter deserves special attention as it poses a unique obstacle. No publicly available database in use was ever thorough enough to be widely accepted as a standard reference genome collection.

Culturing techniques appear to have failed to reveal the full microbial diversity, as indicated by the inexplicably rich data retrieved through shotgun DNA sequencing (metagenomics)[119]. On the other hand, such richness is considered difficult to interpret, as a huge proportion of sequencing reads from the majority of studies typically never map to any members in databases of reference genomes[120] (**Figure 2**).

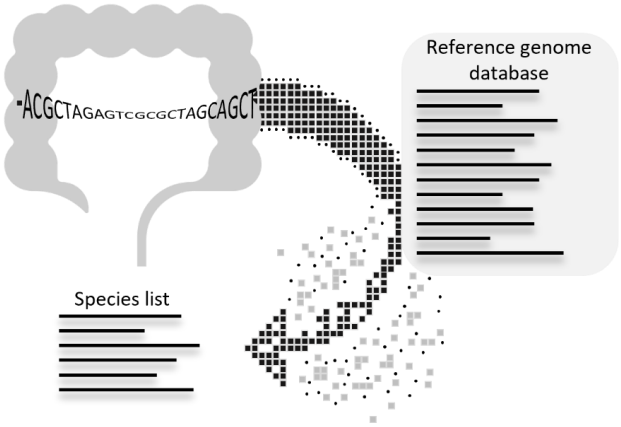


Figure 2. A simple depiction of the taxonomic profiling based on shotgun sequencing. The arrow links the gastrointestinal sample with the results: a list of taxonomically classified strains found in it. For classification to occur, the sequencing reads (the arrow tiles) must query a reference database in search for (a) perfect match(es). The tiles decomposed from the arrow represent reads failing to match a genome in the database, hence becoming discarded from the results for interpretation.

A focused definition of prokaryotes residing in the gut of healthy humans is a fundamental need. Besides merely quelling the curiosity as to what species reside in the intestines of healthy populations, a comprehensive collection of human gut prokaryotic genomes ought to streamline functional studies conducted worldwide, in addition to significantly improving low taxonomic rank classification of sequencing reads[120].

Major advancements in the field of bioinformatics have allowed the reconstruction of reference-free, de-novo assembled genomes: metagenome-assembled genomes (MAGs). This fairly new inventive approach, proposed in 2015 by Hugerth et al., has brought an alternative solution to a wide range of environmental studies, circumventing the need for relying on frequently incomplete reference databases[122].

In recent years, MAGs have also thrown light onto a broad range of hitherto undiscovered human gut microbiota members[120, 123, 124]. Their discovery has enriched the pool of genomes encountered in human intestines worldwide, increasing the prospect of discovering novel biomarkers among them. This has also increased the potential for building a comprehensive collection of human gut prokaryote genomes to be used as a reference database for upcoming studies. However, no efforts have been made towards understanding the prevalence of MAGs in healthy human guts around the world. In addition, their utility is not fully validated. MAGs frequently lack 16S rRNA gene contigs due to difficulties related to their assembly[125], presenting a severe shortcoming in a field where this gene holds the main focus.

Another major challenge in the field is the lack of simple tools for rapid screening of medium-size targets, hindering the translation of biological findings to a clinical setting. We are at a time when simple, robust, and inexpensive tools capable of detecting multiple targets in a single tube are in high demand. An emerging qPCR-compatible method, Liquid Array Diagnostics (LAD), capable of detecting up to 30 targets in a sample, promises an alternative solution in this regard.

Liquid Array Diagnostics principle

Liquid Array Diagnostics (LAD), is a novel, simple genotyping method, based on the single-nucleotide primer extension principle[126]. It can detect multiple microbial targets in a single tube within a working day, while only requiring commonplace instrumentation, namely a qPCR machine (see a general outline of qPCR instrumentation in **Figure 3**).

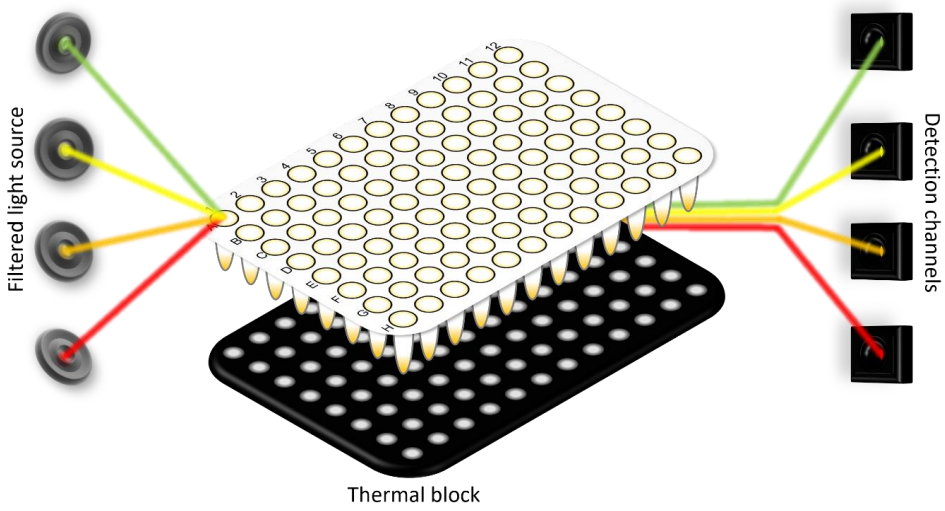


Figure 3. The basics of qPCR instrumentation. A plate holding separate reactions in each well is placed on a thermal block, with the purpose of controlling the temperature of reactions. A filtered source of light passes through each well prior to being captured by its respective optical detection channel. Detection channels register the fluorescence units after each incremental temperature change. Different instruments may contain a different number of detection channels. Usually, the number is between four (CFX Opus 384, Bio-Rad Laboratories, Inc) and six (Rotor Gene Q, Qiagen).

As the name suggests, LAD is performed in a liquid solution. Compared to many existing multiplexed targeted approaches, it does not require separation of probes via the use of a solid phase prior to signal detection, rendering the method uniquely simple.

The technology exploits temperature-dependent fluorescence quenching (the decrease of fluorescence intensity) to detect target probe labelling. A set of oligonucleotides (probes), complementary to signature target sequences, is designed to become labelled with a quencher molecule in the event of hybridization with template DNA. Subsequently, a second set of probes, complementary to the first, are added. These are designed to carry a fluorophore. In the event of target probe labelling, the quencher and the fluorescing moieties come into proximity upon

duplex formation. This allows quenching to occur, reporting thus the presence of designated target(s) (**Figure 4**).

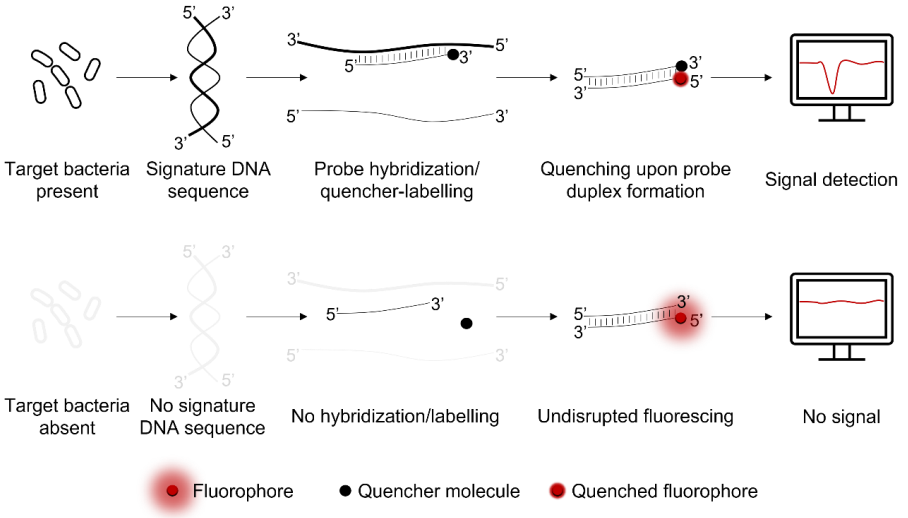


Figure 4. A basic overview of LAD. The top panel depicts probe labelling in the presence of target bacteria sequences. Labelling probes are complementary to target sequences, which serve as a template for single (quencher-labelled) nucleotide extension. In another step, the quencher-labelled probe forms a duplex with a complementary fluorophore-conjugated probe. Due to a proximity between the quencher and the fluorophore, the latter ceases emitting light into the surrounding environment at a temperature dictated by the duplex length, thus reporting the presence of the target. The bottom panel shows a scenario where target bacteria sequences are absent from the reaction. In that case, due to a lack of template, the target probe does not become labelled with a quencher. This, in turn, allows the fluorophore positioned on the complementary probe to continue fluorescing.

Quenching is designed to manifest at specific temperatures, directed by the length and sequence composition of the probe duplexes. Presently, up to five resolvable quenching events can be registered within a single qPCR detection channel. This, combined with six distinct detection channels common to qPCR instrumentation, facilitates detection of up to 30 targets in a single-tube reaction.

Detailed description of the steps related to LAD reactions are presented in the supplementary material.

Aim of the thesis

Immense amounts of sequencing data have revealed important gut microbes associated to human wellbeing, rendering them valuable biomarkers. The field is now prepared to switch the focus to detecting only them, with the intention of reducing the running costs and hands-on time by simplifying lab routines. In addition, detecting an exclusive panel of microbes of high clinical interest reduces the noise (i.e., no data about random microbes with no reported association to human health is generated), streamlining results interpretation. There indeed exist a wide range of targeted methods capable of detecting pre-determined microbial groups. However, they either lack the multiplex capability or they require fragile sample processing in addition to dedicated instrumentation, limiting their utility to specialized labs only.

Considering the demand for easily accessible tools specialized on rapid human gut microbiota diagnostics, the main aim of this thesis was to develop and evaluate LAD, the principles of which make it a promising candidate tool.

The subgoals of the thesis were as follows:

- Build a robust and accurate LAD test for testing human infant gut microbiota.
- Construct a healthy human gut microbial genomes collection for targeted human gut microbiota diagnostics.
- Evaluate the quality of 16S rRNA gene sequence in MAGs.
- Build a LAD assay capable of detecting functional dysbiosis associated to disrupted SCFA levels in fecal samples.

Results and Discussion

This thesis is composed of four distinct sub-studies. Two of them relate to a novel genotyping method, Liquid Array Diagnostics - LAD (**Paper I** and **Paper IV**), and the remaining two to a publicly available collection of human gut prokaryotic genomes, HumGut (**Paper II** and **Paper III**). The four subchapters as listed here correspond to papers constituting the thesis. The first subchapter is extended to include some general information applicable to all LAD-based tests. It also reveals additional unpublished results supporting the utility of LAD for dysbiosis testing.

The published **Paper III** indeed represents a short opinion piece where we swiftly communicate some important findings related to the quality of MAGs. More detailed information is provided in the third subchapter of this section.

I. LAD-based microbiota assays

All LAD-based tests performed during the course of this project used GA-map® CoverAll primers to amplify the 16S rRNA gene. These primers are designed to anneal to conserved primer binding sites, amplifying seven regions (V3 - V9), and yielding amplicons of ca. 1,000-1,200 bp length[34]. A bioinformatic check revealed that they perfectly match >99% of 16S rRNA gene sequences extracted from >6,600 complete RefSeq genomes used to build HumGut (**paper II**). This finding was supported by results generated from wet laboratory work. From 488 genomic DNA (gDNA) templates extracted from different bacteria isolates belonging to different phyla, only 9 failed to yield PCR products (*Lactococcus lactis* subsp. *lactis, Mycobacterium avium* subsp. *paratuberculosis, Mycobacterium avium* subsp. *avium, Bacteroides sp., Corynebacterium jeikeium (13012010), Slackia piriformis, Mycobacterium terrae, Proteus mirabilis* and *Streptomyces lanatus*).

Covering most of the microbial diversity and nucleotide variation, these amplicons served as templates for labelling LAD probes in subsequent steps.

It was proven that the level/degree of probe labelling with a quencher depends on the abundance of complementary target sequences, as presented in the supplementary material, **Figure S6**. However, the potential of utilizing LAD for quantifying targets needs to be investigated and developed further.

Use of LAD for infant gut microbiota composition testing

The utility of LAD in gut microbiome studies was validated by testing 541 mother and infant samples for 15 distinct bacterial markers, as described in **Paper I**. The rationale for validating LAD performance on infant samples was based on the premise that they are well described in the literature[132-135].

The labelling probes were adopted from an already validated GA-map® assay[135], a solid-phase hybridization method based on single nucleotide primer extension. The goal was to reproduce the results retrieved with GA-map® through LAD, by testing samples from the same PACT sample pool (Prevention of Allergy Among Children in Trondheim). Indeed, through two platforms, a decrease of *Staphylococci* from 10-days to 4-month-old infants, and a peak of *Bifidobacteria* at 4-month-old children was observed. Furthermore, LAD signals for a subset of samples (n = 87), corresponded well to 16S rRNA gene Illumina sequencing results.

LAD probe signal measurements showed a distinct, more stable pattern for samples collected from mothers (during their pregnancy) compared to infants and children up to 2 years old (**Figure 5**). This is in accordance with the well-established fact that infant gut microbiota is compositionally different from that of an adult population, in addition to exhibiting a larger interindividual variation[136, 137].

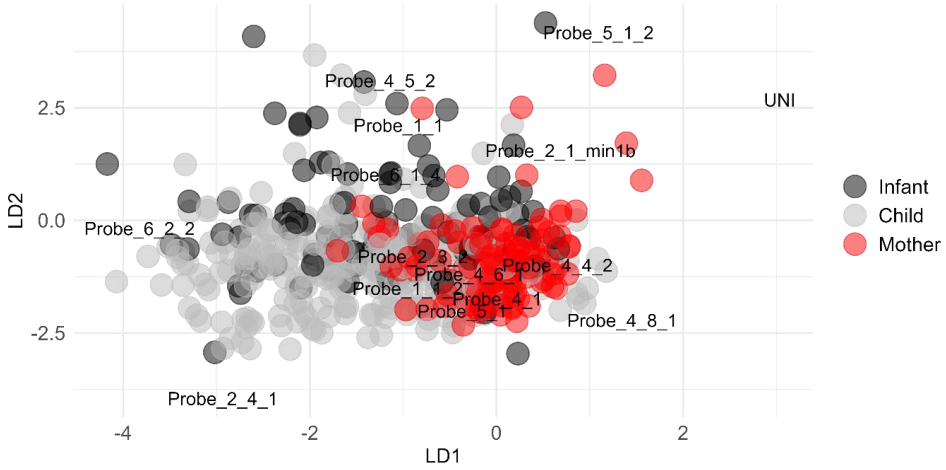


Figure 5. A Linear Discriminant Analysis plot using LAD results for 541 samples as an input. Samples collected from mothers (110 in total) are colored in red, 122 samples collected from infants (2-days to 4-months old) are presented in black, while 309 samples collected from children (1 to 2-years old) are depicted in grey. Samples collected from mothers cluster tighter compared to children and infant samples, which are more spread across the LD1. Probes with the greatest effect on children and infant samples are 6_2_2 and 2_4_1, designed to detect Bifidobacterium breve and a Gamma-proteobacteria subgroup, respectively.

Use of LAD for testing dysbiosis in adults

A LAD-based assay, mimicking the current CE-marked GA-map® Dysbiosis Test[34] was used to test 80 samples biobanked at Genetic Analysis AS (Oslo, Norway, research biobank no. 4071).

Samples were collected from healthy adults (n = 17), *Clostridium difficile*-infected (CDI, n = 15), irritable bowel syndrome (IBS, n = 16), inflammatory bowel disease (IBD, n=17), and diabetes patients (n = 15).

The GA-map® Dysbiosis Test utilizes >48 probes for microbial target detection. As a result, it reports a Dysbiosis Index (DI) score for each sample. The scores are in a scale from 1 to 5, indicating the degree of dysbiosis (DI 1 = normal (normobiosis); DI 5 = a high degree of dysbiosis).

The single-tube LAD assay was built utilizing twenty sequences from the GA-map® probe set with the highest impact on the DI score. This probe subset was designed to detect taxa of various ranks, targeting *Ruminococcus albus* and *R. bromii*, *Faecalibacterium prausnitzii*, *Bacteroides fragilis*, *Ruminococcus gnavus*, *Streptococcus salivarius* ssp. *thermophilus*, *S. sanguinis*, *Akkermansia muciniphila*, *Dialister invisus* and *Megasphaera micronuciformis, Veilonella* spp., *Bacteroides* spp. and *Prevotella* spp., *Bifidobacterium* spp., *Shigella* spp. and *Escherichia* spp., *Parabacteroides* spp., *Alistipes* spp., *Lachnospiraceae*, Clostridia, Bacilli, Firmicutes, Actinobacteria, and Proteobacteria. Due to confidentiality concerns, probe sequences are not revealed here.

The results, depicted in the form of Linear Discriminant Analysis plots (**Figure 6**), showed that this small subset of probes utilized in the LAD platform, effectively served to separate healthy samples from the rest of the cohorts, highlighting the potential of LAD for routine gut microbiota diagnostics.

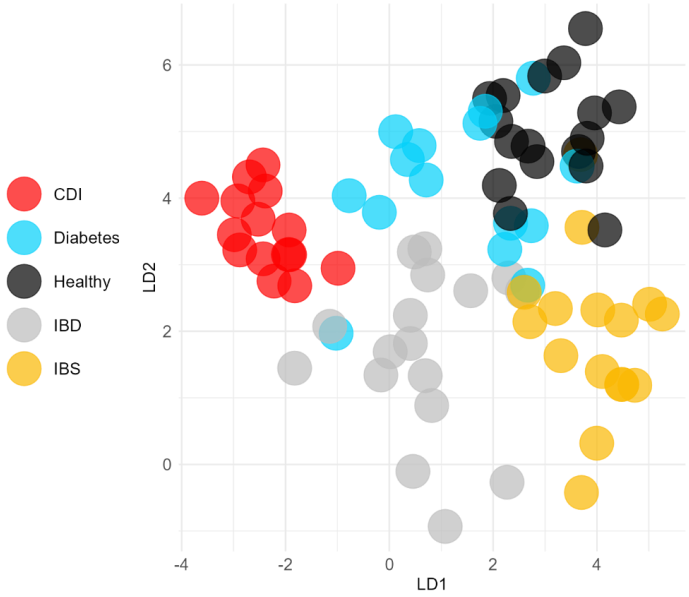


Figure 6. Linear Discriminant Analysis plots based on 20 LAD probe signals for 80 samples: 15 collected from Clostridium difficile infected patients (CDI), 15 from diabetes patients, 17 from healthy adults, 17 from inflammatory bowel disease (IBD), and 16 from irritable bowel syndrome (IBS) patients. Healthy and CDI samples separate with 100% classification accuracy. Same is the case between healthy and IBD. Healthy and IBS samples separate with 93% classification accuracy, i.e., one IBS sample misclassified as healthy and vice-versa. Healthy samples separate from diabetes with 80% accuracy (two diabetes samples classified as healthy, three healthy samples as diabetes). The overall classification accuracy, taking into account the separation between all cohorts, is 84% (13 out of 80 samples misclassified).

II. A comprehensive collection of human gut prokaryote genomes - HumGut

An integral goal of this thesis was building a global comprehensive reference genomes collection, HumGut.

As described in detail in **Paper II**, HumGut was built by screening thousands of human gut metagenomic samples collected from healthy people worldwide. These publicly available samples were screened for the presence of any of the MAGs and human gut isolate genomes stemming from the Unified Human Gastrointestinal Genome (UHGG) collection[138] in addition to any RefSeq prokaryotic genomes. The latter represent non-redundant, curated and annotated genomes by National Center for Biotechnology Information (NCBI)[139].

Only genomes considered to have been found in at least one metagenome were kept in the collection. These were then dereplicated at a 97.5% average nucleotide identity (ANI), keeping the most prevalent genome as a cluster representative. Such a collection ('HumGut_975') consists of >30,000 representative genomes. On another level, HumGut_975 representatives were clustered at a 95% ANI ('HumGut_95'), marking the prokaryotic species-level threshold[140]. This resulted in >5,100 representatives, indicating the presence of at least this many species in healthy human guts worldwide (**Figure 7**).

The approach towards building HumGut is a novel one. It is free from geographical constraints as it considers the gut content of healthy people from various countries around the world. In addition, it keeps as cluster representatives the most prevalent and thereby relevant genomes.

The naming of cluster representative genomes includes a numerical postscript, which indicates their prevalence among the screened metagenomes. For example, HumGut_1, otherwise taxonomically classified as *Bacteroides vulgatus*, represents the most prevalent genome, followed by HumGut_2 (classified as *B. vulgatus*, as well) found as the second most prevalent, and so on. The website hosting HumGut provides a table listing all representative genomes and their occurrence among the 3,545 screened metagenomes. This adds value to HumGut collection, as it aids its users by bringing a wider context to their results.

Besides the proven drastic improvements in classification of shotgun metagenomic reads, identifying and listing human gastrointestinal prokaryote species found in healthy people attains a major milestone for targeted approaches like LAD. It defines the pool of target/non-target sequences which in turn facilitates a streamlined probe design.

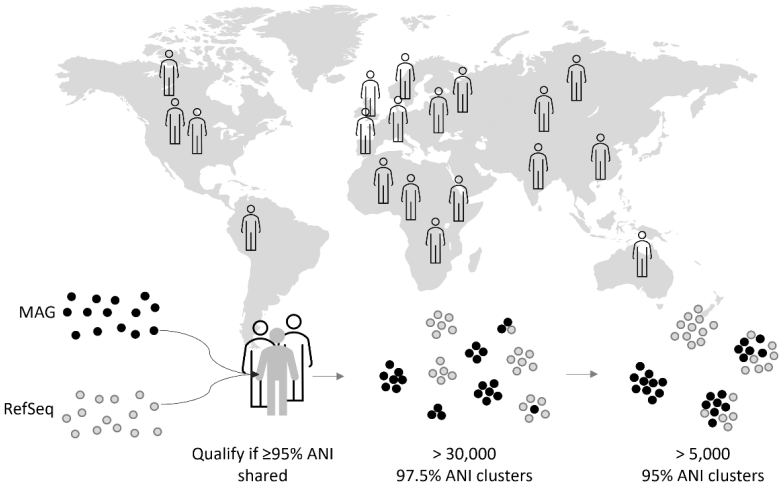


Figure 7. A brief outline of the HumGut building process. Thousands of metagenomic samples collected from healthy people worldwide were screened for the presence of any MAG (black dots) or RefSeq prokaryotic genome (grey dots). Qualified genomes shared ≥ 95% average nucleotide identity (ANI) with at least one metagenome. They were further dereplicated, first at ≥ 97.5% ANI, then at ≥ 95% ANI, generating thus two HumGut collections of various granularities (HumGut_975 and HumGut_95, respectively).

III. HumGut: gaps, limitations, and perspectives

During the course of this thesis, two seemingly unrelated activities were pursued: developing LAD and building HumGut. They are intimately interconnected though, as a comprehensive repository like HumGut aids techniques like LAD by providing the necessary target DNA sequences. However, at this stage, the ultimate connection between the two was impossible to make because an important link was found missing: the 16S rRNA gene.

The vast majority of HumGut_975 (91%) consists of MAGs, which, given the persisting limitations on current technologies, are often not fully completed[125].

The severity of this problem was reflected in our searches for 16S using barrnap (https://github.com/tseemann/barrnap). The 16S rRNA sequences were extracted from 7% of MAGs only, while they were extracted from >93% of other RefSeq genome types (**Figure 8**). From 30,691 HumGut_975 genomes in total, only 4,560 yielded 16S rRNA gene sequences, highlighting the challenge of building a corresponding 16S-HumGut collection.

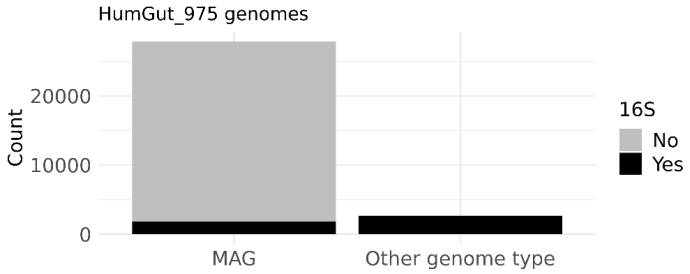


Figure 8. Bar chart depicting the volume of MAGs (left) or other genome types (right: complete RefSeq genomes, chromosomes, contigs, scaffolds, isolates) that yielded 16S rRNA gene sequences using barrnap tool (sections colored in black).

Not all HumGut clusters have similar 16S rRNA gene sequences

As mentioned, HumGut_975 is built of genome cluster representatives. The members of these clusters share at least 97.5% genome-wide average nucleotide identity (ANI), exceeding the species-delineation threshold of 95%[140]. Given the conserved nature of 16S, its identity is expected to be even higher among the same cluster genomes[140-143]. By proxy, 16S sequences extracted from the available HumGut genomes should be nearly identical with those extracted from other genomes belonging in the same cluster (~119,000 genomes from >381,000 used to build HumGut harbored at least one copy of this gene).

However, this was not observed in clusters consisting solely of MAGs (281 clusters in total). A *MUSCLE* multiple sequence alignment of 16S sequences from the same cluster, followed by a

computation of their pairwise distances using a model of DNA evolution (ape package in RStudio[144]), showed that the average distance of all MAGs measurements was 0.074 (SD = 0.095), translated to an average identity of 93% (1 - 0.074 = 0.926). By contrast, an average identity of 99.85% (distance 0.0015, SD = 0.003) was observed in 95 clusters comprised exclusively of complete RefSeq genomes. Only clusters with >5 genomes, each containing at least one 16S copy, were considered for this analysis (**Figure 9**).

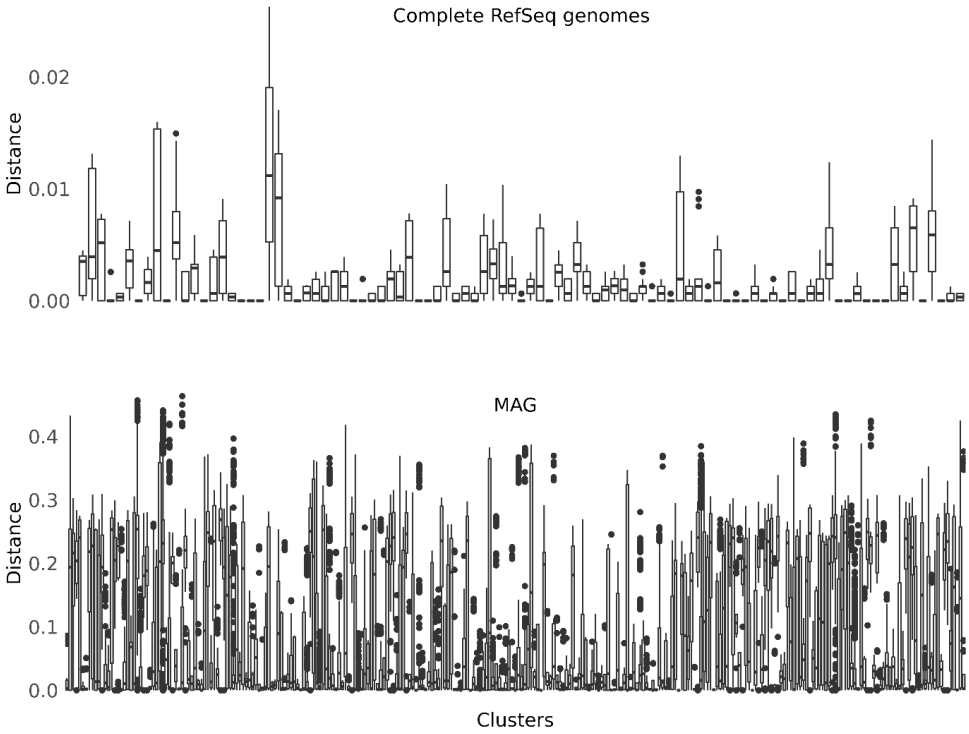


Figure 9. The distance between 16S rRNA gene sequences within HumGut_975 clusters comprised of purely complete RefSeq genomes (95 clusters, top panel) or MAGs (281 clusters below). Each cluster contains more than 5 genomes with at least one 16S sequence. The average distance of all complete RefSeq genome measurements was 0.0015 (SD = 0.003), while for MAGs, the average was 0.074 (SD = 0.095).

The severity of intragenomic 16S heterogeneity is linked to genome type

The variation of copy numbers among different taxa, and the challenges in obtaining the correct copy number from incomplete genomes is well described in the literature[145, 146]. Accordingly, complete RefSeq genomes exhibited the highest average 16S rRNA gene copy number (6.2 copies), while MAGs had 1.1 copies on average (**Figure 10**).

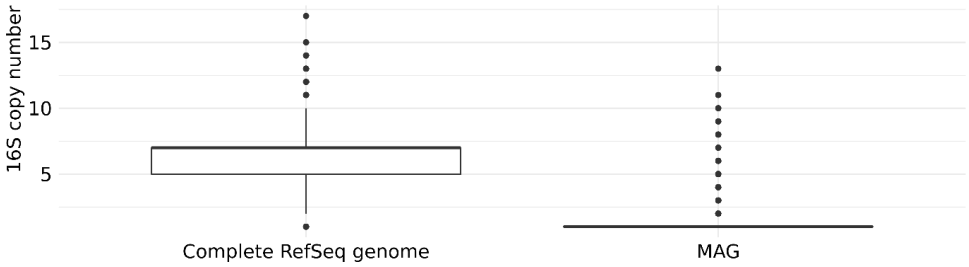


Figure 10. Boxplots showing the distribution of 16S rRNA gene copy numbers (Y-axis) among 6,648 complete RefSeq genomes (left) and 18,709 MAGs (right).

Interestingly, MAGs showed a significantly high intragenomic 16S heterogeneity (**Figure 11**). More than 10% of all MAGs with multiple 16S copies showed a variation in every single position of their sequences, with about 50% of genomes containing copies with a variation (base substitution, insertion, deletion) near position 500. Such a high intragenomic variance degree obscures the true representative sequences and raises doubts about the quality of MAG-extracted 16S rDNA sequences.

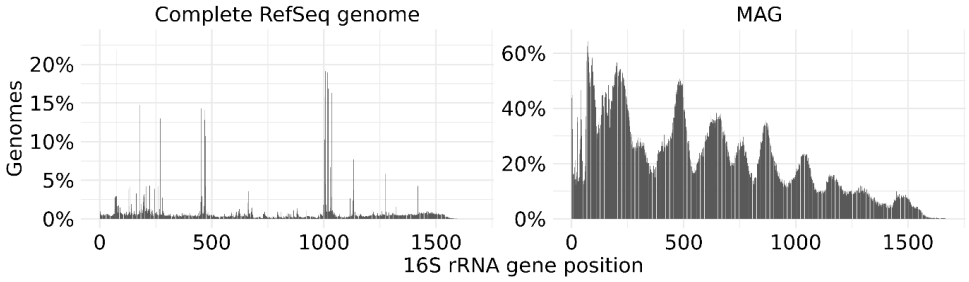


Figure 11. Intragenomic 16S rRNA gene variation in different genome types shown on different panels. The X-axis depicts the 16S gene position. The Y-axis shows the percentage of genomes having some intragenomic variation in the respective position. The percentages were computed considering only genomes with more than one 16S copy (4,984 complete RefSeq genomes and 1,285 MAGs).

Intragenomic variations were also observed in complete RefSeq genomes, considered as the golden standard in terms of their quality. However, most of the variations seem to be linked to specific positions within the gene (positions 75-90, 250-270, 450-470, 1,000-1,030, and 1,130). In line with these findings, similar heterogeneity patterns were also reported by Johnson et al. (2019) in a study of 381 different isolates[147].

An excerpt of the abovementioned gaps is presented in **Paper III**.

RefSeq intragenomic 16S variation is mostly observed in Proteobacteria

About 24% of complete RefSeq genomes harboring multiple 16S copies showed no intragenomic variation of this gene. The remaining showed various degrees of heterogeneity.

The highest diversity was observed in a single 97.5% ANI cluster represented by *Shigella flexneri*. Of all genomes showing variation, >30% belonged to this group (**Figure 12**). More than 70% of all heterogenous genomes for 16S rRNA gene belonged to Proteobacteria phylum, where *Shigella* belongs.

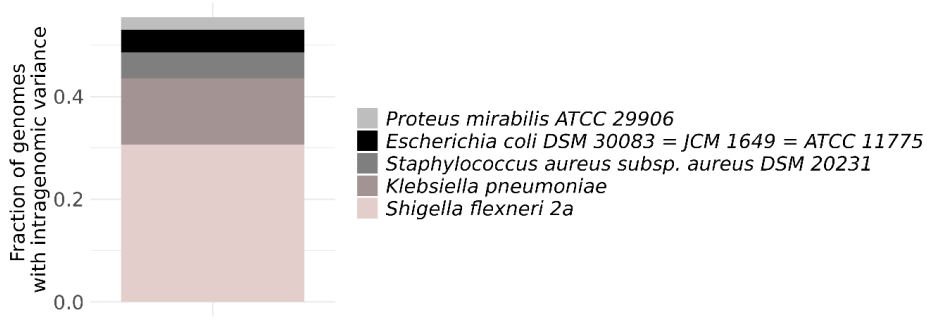


Figure 12. HumGut_975 clusters with most 16S intragenomic heterogeneity. The fraction considers all genomes showing some degree of variation. Most of the variation (>50%) was observed in only 5 different clusters, all belonging to Proteobacterium phylum.

Assuming the observed variations do not reflect sequencing or assembling errors, they all deserve specific attention, especially knowing that many of these genomes resulted highly prevalent among healthy human guts (>80%).

They should be considered when 16S sequencing reads are grouped into operational taxonomic units (OTUs). Using a conventional threshold of 97% identity[148], single genome copies could diverge into more than one OTU, falsely increasing the diversity index of the sample[149]. Similar false results would be produced by reporting amplicon sequence variants (ASVs) instead.

Obtaining this level of information heightens the prospect of building 16S-targetted assays capable of distinguishing *Escherichia coli* and certain *Shigella* species, for example, which has otherwise been considered impossible with current 16S-based molecular diagnostic tools[150].

HumGut clusters could be a resource of species 16S representative sequences

In addition to clustering at 97.5% identity, the genomes used to build HumGut were dereplicated at a species-level threshold (95% ANI)[140], resulting in > 5,100 representatives in total (HumGut_95). The main rationale for performing this was to offer a simpler solution to HumGut

users. Being smaller in size (15.9 GB vs 24.9 GB for HumGut_975), HumGut_95 may be more convenient for metagenome studies where high taxonomic resolution is not essential. Additionally, clustering at 95% ANI was a sensible method of inferring the total number of microbial species in the gut oh healthy humans worldwide.

A further drastic reduction in database size would be possible if only 16S rRNA gene sequences from each genome were included. This would be an ideal solution for 16S amplicon analysis. As already demonstrated, this unfortunately is not possible with the current HumGut versions. However, it is worth considering strategies for the upcoming versions.

As shown, cluster members, and occasionally genomes themselves, contain substantial variations in their 16S copies. This indicates that building a highly sought species-level-16S-HumGut collection is not as straight-forward as extracting copies of this gene from cluster representatives. Instead, one reasonable strategy would be finding the most prevalent sequences within the cluster, as exemplified in **Figure 13**.

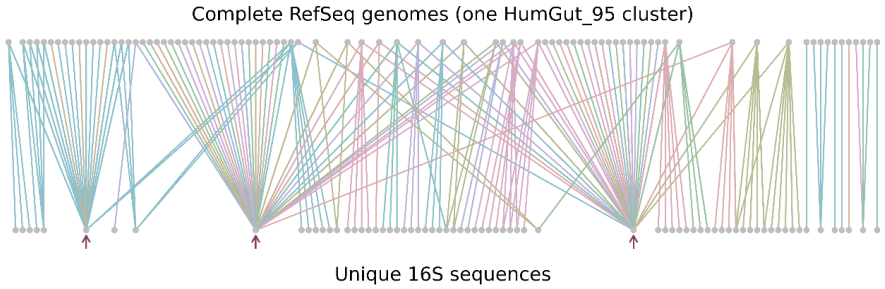


Figure 13. A network of 87 complete RefSeq genomes belonging to a single HumGut_95 cluster (upper points) and their unique 16S rRNA gene sequences (lower points). The connecting lines are colored based on 97.5% nucleotide identity clusters. The arrows emphasize three of the 16S sequences encountered in most of the genomes, which may act as cluster representatives.

Building a collection of such granularity would enhance the specificity of probe design substantially, lifting the performance of targeted approaches like LAD. However, as implied earlier in this chapter, this level of work can only be achieved after technological challenges related to 16S-assembly are overcome.

IV. Linking functions with 16S rRNA gene

The ultimate goal of this thesis was to develop a simple detection method, LAD, for predicting gut microbiota functionality. The focus was directed towards SCFA production. That because human gut bacteria largely affect and maintain the health of the host through these end-products[151], while routines for measuring them are frequently challenging[152, 153].

SCFAs are highly volatile acids, therefore sample processing represents a major bottleneck, prohibiting high-throughput research and clinical utility[152, 153]. On the other hand, inferring SCFA levels based on direct measurements of specialized producers is not viable, given that the short-chain fatty acid production is dependent on complex environmental factors, such as the abundance of cross-feeders. A typical example is that of butyrate-production by *F. prausnitzii*, which is enhanced by interaction with *Bifidobacterium adolescentis*, a species that does not produce this acid[154].

As described in detail in **Paper IV**, we aimed at circumventing the need for a direct measurement of SCFAs, by establishing a LAD test that directly targets the 16S rRNA gene of key bacterial indicator groups. Given the importance of beneficial butyrate, we aimed at predicting samples with abnormally increased P:B with the intention of providing a tool for fast, robust and accurate detection, to be used for high-throughput studies where an increased P:B is expected to act as disease indicator (Parkinson's disease[155], Autism Spectrum Disorder[156], Type 2 Diabetes[157], etc.).

Propionate and butyrate represent fermentation end-products of well separated bacterial groups[158]. Propionate is mainly produced by Bacteroidetes and Negativicutes, while butyrate by Lachnospiraceae and Ruminococcaceae[159]. There are only two known species that, depending on the substrate, can produce both (*Coprococcus catus* and *Roseburia inulinovorans*[158]). On the other hand, both these bacterial groups can additionally produce acetate. Hence, a disruption of propionate-to-butyrate ratios is of special interest, possibly reflecting an unhealthy disbalance between the two major member groups of normal gut flora (and/or their corresponding cross-feeders).

By targeting a limited number of bacteria, not exclusively known to produce butyrate or propionate, LAD was successful at predicting high P:B samples with high accuracy.

Here, a PLS + LDA (Partial Least Squares – Linear Discriminant Analysis) model was employed, for both finding the markers (step I) and predicting the ratio based on LAD results (step II).

In step I, the indicator bacteria were found using the 16S rRNA gene sequencing results as predictors of the P:B (calculated after Gas Chromatography measurements). An eliminator function, reducing the number of selected variables (OTUs), at the expense of a marginal model

performance decrease, was used[161] with the intention of targeting the lowest number of operational taxonomic units (OTUs) while keeping a high prediction accuracy. The design of the probes targeting the intended OTUs, as mentioned in **paper IV**, was performed utilizing the TNTProbe Tool[160], a bioinformatics software developed in house at Genetic Analysis AS.

In step II, LAD probe signals, reporting the abundance of target OTUs, were used as P:B predictors.

A successful 16S rDNA-based assay predicting such a functional outcome marks an important leap towards exploiting the 16S rRNA gene beyond its mere value for taxonomical classification.

Indeed, many tools exist, like PICRUSt[162] or Tax4Fun[163], deemed as highly accurate at predicting functions based on 16S rDNA sequences. However, these tools require reference genomes to which the sequences of this gene are mapped, in order to assess their whole marker gene repertoire. The challenges with obtaining comprehensive reference databases were explained at great lengths in the previous sections of this thesis. In addition, we have shown that butyrate and propionate levels did not correlate well with the abundance of genomes harboring genes responsible of producing such acids. This indicates that no conclusions about the exercised functions (i.e., butyrate and propionate levels) based solely on the presence of these genes can be drawn.

Conclusion and Future Perspectives

The number of gut microbiota-based studies has increased exponentially in recent years. Correspondingly, various tools aimed at aiding the detection of microbial species have been developed. The use of Next Generation Sequencing (NGS) techniques continuously equips researchers with unprecedented amounts of data for each DNA sequencing run. By now, we have collected tremendous amount of information about the presence and abundance of different microbes in various cohorts, revealing a relationship between human wellbeing and a harmonious microbial composition[34-51]. This, in turn, has opened the possibility of utilizing gut microbiota composition as a disease indicator in routine clinical tests, but also as a target for disease treatment and prevention.

Although NGS techniques are excellent tools for gut microbial marker exploration, their sustainability for routine monitoring of microbial composition in a clinical setting is often vulnerable. Besides the running costs and dedicated instrumentation, DNA sequencing methods require multiple dry- and wet lab steps, each introducing a bias in the system[117, 118], resulting in outcomes not straightforwardly commensurable. Furthermore, different methods are prone to a wide range of sequencing errors, producing artifacts that may obscure true sequences and often overestimate their diversity[164, 165].

Routine clinical monitoring of gut biomarkers would make a better use of methods specialized on detecting microbial groups of interest only, bypassing the production of unnecessary data with the aim of increasing the stability of the system by further simplifying it. The aim of this thesis was to develop such a method, Liquid Array Diagnostics (LAD).

Compared to other commonly used qPCR-based detection methods, LAD is characterized by an increased multiplexity level. LAD overcomes the limitation of reporting a single signal within a channel by exploiting a second dimension within the system, the temperature. This is otherwise unachievable using other contemporary methods, such as those based on Molecular Beacons®[110], KASP[111] or Taqman®[112] probes. Additionally, LAD utilizes multiple detection channels simultaneously, otherwise impossible by techniques relying on intercalating dyes, e.g. EvaGreen®[113].

There exist other detection systems allowing a higher multiplexity level than LAD. These methods are mostly based on solid-phase hybridization[34, 166]. However, in contrast to them, LAD does not require a washing step, avoiding a major procedural bottleneck. In addition, bead detection methods require dedicated and expensive instrumentation, like flow cytometers, restricting their utility to specialized laboratories only.

A robust LAD test is characterized by short hands-on time, inexpensive reagents and instrument needed. Requiring only qPCR instrumentation, it is suitable for use in most labs. It can detect up to 30 markers in a single tube, outperforming other contemporary qPCR-instrumentation-based methods. In addition, it has the potential of providing quantitative information, yielding stronger signals for more abundant targets.

Understandably, it can only be utilized for detecting pre-determined targets, not for de-novo biomarker discovery, as probes are designed to anneal to well-described DNA sequences unique to relevant bacteria groups.

The unique trait of targeted approaches like LAD, is their independence from third-party results (for example a reference database) for interpretation. The understanding of what a probe targets may change over time (with the increase of publicly available DNA sequences); however, its signal remains stable for same-type samples, assuming target sequences do not undergo mutation.

In relation to reference databases, this thesis provides one –HumGut, representing the most comprehensive collection of human gut prokaryote genomes to date. Making a full circle back to explorative sequencing, HumGut is expected to aid metagenomic studies around the world, streamlining biomarker discovery, while LAD is expected to make use of the accumulated knowledge, breaking the circle when acting independently in a clinical setting.

The only other publicly available database of human gut prokaryotic genomes to date is the Unified Human Gastrointestinal Genomes (UHGG) collection[138]. UHGG represents a collection of all genomes (mostly MAGs) derived from sampling human guts. In comparison to HumGut, it contains approximately 500 fewer species-level genomes, reflecting the inclusiveness of our approach.

A specific LAD assay aimed at detecting functional dysbiosis related to an increased fecal propionate-to-butyrate ratio (P:B) is presented here. The current P:B prediction LAD test represents a proof of concept with a great clinical utility potential. Similar tests, focusing on other SCFA levels by targeting other microbial markers can be designed. In addition, a possible HumGut 2.0 collection, comprised of genomes harboring 16S rRNA gene sequences, will elevate our comprehension level regarding the community of gut microbial markers.

A triad between LAD, upcoming HumGut versions, and ever improving bioinformatic tools, presents a great possibility for smart solutions and a greater understanding of human gut microbes.

During this thesis work, the focus was exclusively put on human gut bacteria and short-chain fatty acids. Although bacteria are the most abundant microorganisms residing in the gut, fungi, archaea and viruses deserve similar attention as they are described to equally contribute to human health[167, 168]. As such, a human gut reference genomes collection is not complete without their inclusion. As a future work, HumGut (currently harboring only bacteria and archaea genomes) must consist of genomes belonging to all domains of life. In addition, the same approach to building HumGut may be employed to building collections of genomes from other body sites (HumOral, HumSkin, HumVaginal, conceivably).

The list of functional microbial traits goes well beyond short-chain fatty acid production explored in this thesis. A panel of LAD assays dedicated to different functional disruptions should be feasible to create. Also, designing a LAD test detecting marker species from different life domains must be considered.

Another aspect worth of exploring is a LAD-based oral flora biomarker detection. Studies have shown that there is an association between gut and oral microbiota[169]. Oral dysbiosis is reportedly associated with the same range of diseases typically associated with gut dysbiosis, such as colorectal cancer[170], Alzheimer's disease[171], type 2 diabetes[172], etc., evoking the possibility of targeting saliva microorganisms as a proxy for gut dysbiosis detection. This would allow for easier sampling, further simplifying the method and increasing its availability.

Acknowledging the profound relationship between human health and symbiotic microbial communities, modern medicine is expected to benefit significantly from routine monitoring of their composition - this as a more holistic approach to identifying and treating different systemic diseases.

Literature

1. 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.
2. Mariat, D., et al., *The Firmicutes/Bacteroidetes ratio of the human microbiota changes with age.* BMC Microbiology, 2009. **9**(1): p. 123.
3. O'Toole, P.W. and M.J. Claesson, *Gut microbiota: Changes throughout the lifespan from infancy to elderly.* International Dairy Journal, 2010. **20**(4): p. 281-291.
4. Lozupone, C.A., et al., *Diversity, stability and resilience of the human gut microbiota.* Nature, 2012. **489**(7415): p. 220-230.
5. Bonder, M.J., et al., *The effect of host genetics on the gut microbiome.* Nature Genetics, 2016. **48**(11): p. 1407-1412.
6. Kurilshikov, A., et al., *Host Genetics and Gut Microbiome: Challenges and Perspectives.* Trends in Immunology, 2017. **38**(9): p. 633-647.
7. Reyman, M., et al., *Impact of delivery mode-associated gut microbiota dynamics on health in the first year of life.* Nature Communications, 2019. **10**(1): p. 4997.
8. Mitchell, C.M., et al., *Delivery Mode Affects Stability of Early Infant Gut Microbiota.* Cell Reports Medicine, 2020. **1**(9): p. 100156.
9. Ho, N.T., et al., *Meta-analysis of effects of exclusive breastfeeding on infant gut microbiota across populations.* Nature Communications, 2018. **9**(1): p. 4169.
10. Guo, C., et al., *Breastfeeding restored the gut microbiota in caesarean section infants and lowered the infection risk in early life.* BMC Pediatrics, 2020. **20**(1): p. 532.
11. Panda, S., et al., *Short-Term Effect of Antibiotics on Human Gut Microbiota.* PLOS ONE, 2014. **9**(4): p. e95476.
12. Dethlefsen, L., et al., *The Pervasive Effects of an Antibiotic on the Human Gut Microbiota, as Revealed by Deep 16S rRNA Sequencing.* PLOS Biology, 2008. **6**(11): p. e280.
13. Aguirre, M., et al., *Diet drives quick changes in the metabolic activity and composition of human gut microbiota in a validated in vitro gut model.* Research in Microbiology, 2016. **167**(2): p. 114-125.
14. Chung, W.S.F., et al., *Modulation of the human gut microbiota by dietary fibres occurs at the species level.* BMC Biology, 2016. **14**(1): p. 3.
15. Bressa, C., et al., *Differences in gut microbiota profile between women with active lifestyle and sedentary women.* PLOS ONE, 2017. **12**(2): p. e0171352.
16. Tun, H.M., et al., *Exposure to household furry pets influences the gut microbiota of infants at 3–4 months following various birth scenarios.* Microbiome, 2017. **5**(1): p. 40.
17. Walker, W.A., *Chapter 25 - Dysbiosis*, in *The Microbiota in Gastrointestinal Pathophysiology*, M.H. Floch, Y. Ringel, and W. Allan Walker, Editors. 2017, Academic Press: Boston. p. 227-232.
18. Lyon, L., *'All disease begins in the gut': was Hippocrates right?* Brain, 2018. **141**(3): p. e20-e20.
19. Egerton, F.N., *A History of the Ecological Sciences, Part 19: Leeuwenhoek's Microscopic Natural History.* Bulletin of the Ecological Society of America, 2006. **87**(1): p. 47-58.
20. Bordenave, G., *Louis Pasteur (1822–1895).* Microbes and Infection, 2003. **5**(6): p. 553-560.
21. Münch, R., *Robert Koch.* Microbes and Infection, 2003. **5**(1): p. 69-74.
22. Hungate, R.E., *The anaerobic mesophilic cellulolytic bacteria.* Bacteriological reviews, 1950. **14**(1): p. 1-49.
23. Reyniers, J.A., *Rearing germfree albino rats.* Lobund Reports, 1946. **1**: p. 1.
24. Schaedler, R.W., R. Dubs, and R. Costello, *ASSOCIATION OF GERMFREE MICE WITH BACTERIA ISOLATED FROM NORMAL MICE.* J Exp Med, 1965. **122**(1): p. 77-82.
25. Gustafsson, B.E., *Vitamin K deficiency in germfree rats.* Ann N Y Acad Sci, 1959. **78**: p. 166-74.
26. Sumi, Y., et al., *Vitamin B-6 deficiency in germfree rats.* J Nutr, 1977. **107**(9): p. 1707-14.
27. Gustafsson , B.E. and L.S. Lanke *BILIRUBIN AND UROBILINS IN GERMFREE, EX-GERMFREE, AND CONVENTIONAL RATS.* Journal of Experimental Medicine, 1960. **112**(6): p. 975-981.
28. Peppercorn, M.A. and P. Goldman, *The role of intestinal bacteria in the metabolism of salicylazosulfapyridine.* J Pharmacol Exp Ther, 1972. **181**(3): p. 555-62.
29. Braniste, V., et al., *The gut microbiota influences blood-brain barrier permeability in mice.* 2014. **6**(263): p. 263ra158-263ra158.
30. Gustafsson, B.E., T. Midtvedt, and A. Norman, *METABOLISM OF CHOLIC ACID IN GERMFREE ANIMALS AFTER THE ESTABLISHMENT IN THE INTESTINAL TRACT OF DECONJUGATING AND 7α-DEHYDROXYLATING BACTERIA.* 1968. **72**(3): p. 433-443.
31. Avolio, E., et al., *Modifications of Behavior and Inflammation in Mice Following Transplant with Fecal Microbiota from Children with Autism.* Neuroscience, 2022.
32. Sampson, T.R., et al., *Gut Microbiota Regulate Motor Deficits and Neuroinflammation in a Model of Parkinson's Disease.* Cell, 2016. **167**(6): p. 1469-1480.e12.
33. Ridaura, V.K., et al., *Gut Microbiota from Twins Discordant for Obesity Modulate Metabolism in Mice.* 2013. **341**(6150): p. 1241214.

34. Casen, C., et al., *Deviations in human gut microbiota: a novel diagnostic test for determining dysbiosis in patients with IBS or IBD.* Aliment Pharmacol Ther, 2015. **42**(1): p. 71-83.
35. Masoodi, I., et al., *Microbial dysbiosis in irritable bowel syndrome: A single-center metagenomic study in Saudi Arabia.* JGH Open, 2020. **n/a**(n/a).
36. Baldelli, V., et al., *The Role of Enterobacteriaceae in Gut Microbiota Dysbiosis in Inflammatory Bowel Diseases.* 2021. **9**(4): p. 697.
37. Dey, N., et al., *Association of gut microbiota with post-operative clinical course in Crohn's disease.* BMC gastroenterology, 2013. **13**: p. 131-131.
38. Abed, J., et al., *Colon Cancer-Associated Fusobacterium nucleatum May Originate From the Oral Cavity and Reach Colon Tumors via the Circulatory System.* Frontiers in cellular and infection microbiology, 2020. **10**: p. 400-400.
39. Chen, W., et al., *Human intestinal lumen and mucosa-associated microbiota in patients with colorectal cancer.* PloS one, 2012. **7**(6): p. e39743-e39743.
40. Da Silva, H.E., et al., *Nonalcoholic fatty liver disease is associated with dysbiosis independent of body mass index and insulin resistance.* Scientific Reports, 2018. **8**(1): p. 1466.
41. Jiang, W., et al., *Dysbiosis gut microbiota associated with inflammation and impaired mucosal immune function in intestine of humans with non-alcoholic fatty liver disease.* Scientific reports, 2015. **5**: p. 8096.
42. Huang, Y., et al., *Possible association of Firmicutes in the gut microbiota of patients with major depressive disorder.* Neuropsychiatric disease and treatment, 2018. **14**: p. 3329-3337.
43. Naseribafrouei, A., et al., *Correlation between the human fecal microbiota and depression.* Neurogastroenterology & Motility, 2014. **26**(8): p. 1155-1162.
44. Andreo-Martínez, P., et al., *A Meta-analysis of Gut Microbiota in Children with Autism.* Journal of Autism and Developmental Disorders, 2021.
45. De Angelis, M., et al., *Fecal microbiota and metabolome of children with autism and pervasive developmental disorder not otherwise specified.* PloS one, 2013. **8**(10).
46. Finegold, S.M., *Desulfovibrio species are potentially important in regressive autism.* Medical Hypotheses, 2011. **77**(2): p. 270-274.
47. Petrov, V., et al., *Analysis of gut microbiota in patients with Parkinson's disease.* Bulletin of experimental biology and medicine, 2017. **162**(6): p. 734-737.
48. Pietrucci, D., et al., *Dysbiosis of gut microbiota in a selected population of Parkinson's patients.* Parkinsonism & Related Disorders, 2019. **65**: p. 124-130.
49. Brahe, L.K., et al., *Specific gut microbiota features and metabolic markers in postmenopausal women with obesity.* Nutrition & Diabetes, 2015. **5**(6): p. e159-e159.
50. Fernandes, J., et al., *Adiposity, gut microbiota and faecal short chain fatty acids are linked in adult humans.* Nutrition & diabetes, 2014. **4**(6): p. e121-e121.
51. Kasai, C., et al., *Comparison of the gut microbiota composition between obese and non-obese individuals in a Japanese population, as analyzed by terminal restriction fragment length polymorphism and next-generation sequencing.* BMC gastroenterology, 2015. **15**(1): p. 100.
52. Carding, S., et al., *Dysbiosis of the gut microbiota in disease.* Microbial Ecology in Health and Disease, 2015. **26**(1): p. 26191.
53. Derrien, M., C. Belzer, and W.M. de Vos, *Akkermansia muciniphila and its role in regulating host functions.* Microbial Pathogenesis, 2017. **106**: p. 171-181.
54. Ferreira-Halder, C.V., A.V.d.S. Faria, and S.S. Andrade, *Action and function of Faecalibacterium prausnitzii in health and disease.* Best Practice & Research Clinical Gastroenterology, 2017. **31**(6): p. 643-648.
55. Lopez-Siles, M., et al., *Alterations in the Abundance and Co-occurrence of Akkermansia muciniphila and Faecalibacterium prausnitzii in the Colonic Mucosa of Inflammatory Bowel Disease Subjects.* 2018. **8**.
56. Derrien, M., et al., *Akkermansia muciniphila gen. nov., sp. nov., a human intestinal mucin-degrading bacterium.* 2004. **54**(5): p. 1469-1476.
57. Duncan, S.H., et al., *Growth requirements and fermentation products of Fusobacterium prausnitzii, and a proposal to reclassify it as Faecalibacterium prausnitzii gen. nov., comb. nov.* 2002. **52**(6): p. 2141-2146.
58. Everard, A., et al., *Cross-talk between Akkermansia muciniphila and intestinal epithelium controls diet-induced obesity.* Proc Natl Acad Sci U S A, 2013. **110**(22): p. 9066-71.
59. Dao, M.C., et al., *Akkermansia muciniphila and improved metabolic health during a dietary intervention in obesity: relationship with gut microbiome richness and ecology.* 2016. **65**(3): p. 426-436.
60. Depommier, C., et al., *Supplementation with Akkermansia muciniphila in overweight and obese human volunteers: a proof-of-concept exploratory study.* Nature Medicine, 2019. **25**(7): p. 1096-1103.
61. Qin, J., et al., *A metagenome-wide association study of gut microbiota in type 2 diabetes.* Nature, 2012. **490**(7418): p. 55-60.
62. Lee, H., G. Ko, and M.W. Griffiths, *Effect of Metformin on Metabolic Improvement and Gut Microbiota.* 2014. **80**(19): p. 5935-5943.
63. Balamurugan, R., et al., *Quantitative differences in intestinal Faecalibacterium prausnitzii in obese Indian children.* 2010. **103**(3): p. 335-338.

64. Machiels, K., et al., *A decrease of the butyrate-producing species Roseburia hominis and Faecalibacterium prausnitzii defines dysbiosis in patients with ulcerative colitis.* 2014. **63**(8): p. 1275-1283.
65. Takahashi, K., et al., *Reduced Abundance of Butyrate-Producing Bacteria Species in the Fecal Microbial Community in Crohn's Disease.* Digestion, 2016. **93**(1): p. 59-65.
66. Wu, N., et al., *Dysbiosis Signature of Fecal Microbiota in Colorectal Cancer Patients.* Microbial Ecology, 2013. **66**(2): p. 462-470.
67. HAMER, H.M., et al., *Review article: the role of butyrate on colonic function.* 2008. **27**(2): p. 104-119.
68. Wang, L., et al., *Increased abundance of Sutterella spp. and Ruminococcus torques in feces of children with autism spectrum disorder.* Molecular Autism, 2013. **4**(1): p. 42.
69. Zinkernagel, M.S., et al., *Association of the Intestinal Microbiome with the Development of Neovascular Age-Related Macular Degeneration.* Scientific Reports, 2017. **7**(1): p. 40826.
70. Yunes, R.A., et al., *GABA production and structure of gadB/gadC genes in Lactobacillus and Bifidobacterium strains from human microbiota.* Anaerobe, 2016. **42**: p. 197-204.
71. Aizawa, E., et al., *Possible association of Bifidobacterium and Lactobacillus in the gut microbiota of patients with major depressive disorder.* Journal of Affective Disorders, 2016. **202**: p. 254-257.
72. Messaoudi, M., et al., *Beneficial psychological effects of a probiotic formulation (Lactobacillus helveticus R0052 and Bifidobacterium longum R0175) in healthy human volunteers.* Gut Microbes, 2011. **2**(4): p. 256-261.
73. Flanagan, L., et al., *Fusobacterium nucleatum associates with stages of colorectal neoplasia development, colorectal cancer and disease outcome.* European Journal of Clinical Microbiology & Infectious Diseases, 2014. **33**(8): p. 1381-1390.
74. Yang, Y., et al., *Fusobacterium nucleatum Increases Proliferation of Colorectal Cancer Cells and Tumor Development in Mice by Activating Toll-Like Receptor 4 Signaling to Nuclear Factor−κB, and Up-regulating Expression of MicroRNA-21.* Gastroenterology, 2017. **152**(4): p. 851-866.e24.
75. Litvak, Y., et al., *Dysbiotic Proteobacteria expansion: a microbial signature of epithelial dysfunction.* Current Opinion in Microbiology, 2017. **39**: p. 1-6.
76. Shin, N.-R., T.W. Whon, and J.-W. Bae, *Proteobacteria: microbial signature of dysbiosis in gut microbiota.* Trends in Biotechnology, 2015. **33**(9): p. 496-503.
77. Krautkramer, K.A., J. Fan, and F. Bäckhed, *Gut microbial metabolites as multi-kingdom intermediates.* Nature Reviews Microbiology, 2021. **19**(2): p. 77-94.
78. Chassard, C., et al., *Functional dysbiosis within the gut microbiota of patients with constipated-irritable bowel syndrome.* 2012. **35**(7): p. 828-838.
79. Tanca, A., et al., *Potential and active functions in the gut microbiota of a healthy human cohort.* Microbiome, 2017. **5**(1): p. 79.
80. Armour, C.R., et al., *A Metagenomic Meta-analysis Reveals Functional Signatures of Health and Disease in the Human Gut Microbiome.* 2019. **4**(4): p. e00332-18.
81. Nugent, J.L., et al., *Altered Tissue Metabolites Correlate with Microbial Dysbiosis in Colorectal Adenomas.* Journal of Proteome Research, 2014. **13**(4): p. 1921-1929.
82. Larsen, P.E. and Y. Dai, *Metabolome of human gut microbiome is predictive of host dysbiosis.* GigaScience, 2015. **4**(1).
83. Chen, Y.-Y., et al., *Microbiome–metabolome reveals the contribution of gut–kidney axis on kidney disease.* Journal of Translational Medicine, 2019. **17**(1): p. 5.
84. Dalile, B., et al., *The role of short-chain fatty acids in microbiota–gut–brain communication.* Nature Reviews Gastroenterology & Hepatology, 2019. **16**(8): p. 461-478.
85. Arora, T., R. Sharma, and G. Frost, *Propionate. Anti-obesity and satiety enhancing factor?* Appetite, 2011. **56**(2): p. 511-515.
86. van der Beek, C.M., et al., *Role of short-chain fatty acids in colonic inflammation, carcinogenesis, and mucosal protection and healing.* Nutrition Reviews, 2017. **75**(4): p. 286-305.
87. He, J., et al., *Short-Chain Fatty Acids and Their Association with Signalling Pathways in Inflammation, Glucose and Lipid Metabolism.* 2020. **21**(17): p. 6356.
88. Wong, J.M.W., et al., *Colonic Health: Fermentation and Short Chain Fatty Acids.* 2006. **40**(3): p. 235-243.
89. Parada Venegas, D., et al., *Short Chain Fatty Acids (SCFAs)-Mediated Gut Epithelial and Immune Regulation and Its Relevance for Inflammatory Bowel Diseases.* 2019. **10**(277).
90. Farup, P.G., K. Rudi, and K. Hestad, *Faecal short-chain fatty acids - a diagnostic biomarker for irritable bowel syndrome?* BMC Gastroenterology, 2016. **16**(1): p. 51.
91. Sun, Q., et al., *Alterations in fecal short-chain fatty acids in patients with irritable bowel syndrome: A systematic review and meta-analysis.* Medicine (Baltimore), 2019. **98**(7): p. e14513.
92. Killingsworth, J., D. Sawmiller, and R.D.J.F.i.A.N. Shytle, *Propionate and Alzheimer's Disease.* 2020. **12**: p. 501.
93. Zeng, X., et al., *Higher Risk of Stroke Is Correlated With Increased Opportunistic Pathogen Load and Reduced Levels of Butyrate-Producing Bacteria in the Gut.* 2019. **9**(4).
94. Hill, M.J., *Intestinal flora and endogenous vitamin synthesis.* European Journal of Cancer Prevention, 1997. **6**(2).
95. Rudzki, L., et al., *Gut microbiota-derived vitamins – underrated powers of a multipotent ally in psychiatric health and disease.* Progress in Neuro-Psychopharmacology and Biological Psychiatry, 2021. **107**: p. 110240.

96. Dumas, M.-E., et al., *Metabolic profiling reveals a contribution of gut microbiota to fatty liver phenotype in insulin-resistant mice.* 2006. **103**(33): p. 12511-12516.
97. Wang, Z., et al., *Gut flora metabolism of phosphatidylcholine promotes cardiovascular disease.* Nature, 2011. **472**(7341): p. 57-63.
98. Pedersen, H.K., et al., *Human gut microbes impact host serum metabolome and insulin sensitivity.* Nature, 2016. **535**(7612): p. 376-381.
99. Ridaura, V. and Y. Belkaid, *Gut Microbiota: The Link to Your Second Brain.* Cell, 2015. **161**(2): p. 193-194.
100. Brial, F., et al., *Implication of gut microbiota metabolites in cardiovascular and metabolic diseases.* Cellular and Molecular Life Sciences, 2018. **75**(21): p. 3977-3990.
101. Agus, A., K. Clément, and H. Sokol, *Gut microbiota-derived metabolites as central regulators in metabolic disorders.* 2021. **70**(6): p. 1174-1182.
102. Houpikian, P. and D. Raoult, *Traditional and molecular techniques for the study of emerging bacterial diseases: one laboratory's perspective.* Emerg Infect Dis, 2002. **8**(2): p. 122-31.
103. Sarangi, A.N., A. Goel, and R. Aggarwal, *Methods for Studying Gut Microbiota: A Primer for Physicians.* Journal of Clinical and Experimental Hepatology, 2019. **9**(1): p. 62-73.
104. Watson, J.D. and F.H. Crick. *The structure of DNA*. in *Cold Spring Harbor symposia on quantitative biology*. 1953. Cold Spring Harbor Laboratory Press.
105. Mullis, K.B., *The Polymerase Chain Reaction (Nobel Lecture).* 1994. **33**(12): p. 1209-1213.
106. Byrd, A.L., Y. Belkaid, and J.A.J.N.R.M. Segre, *The human skin microbiome.* 2018. **16**(3): p. 143-155.
107. Kilian, M., et al., *The oral microbiome – an update for oral healthcare professionals.* British Dental Journal, 2016. **221**(10): p. 657-666.
108. Greenbaum, S., et al., *Ecological dynamics of the vaginal microbiome in relation to health and disease.* American Journal of Obstetrics and Gynecology, 2019. **220**(4): p. 324-335.
109. Janda, J.M. and S.L. Abbott, *16S rRNA Gene Sequencing for Bacterial Identification in the Diagnostic Laboratory: Pluses, Perils, and Pitfalls.* 2007. **45**(9): p. 2761-2764.
110. Piatek, A.S., et al., *Molecular beacon sequence analysis for detecting drug resistance in Mycobacterium tuberculosis.* Nature Biotechnology, 1998. **16**(4): p. 359-363.
111. He, C., J. Holme, and J. Anthony, *SNP genotyping: the KASP assay*, in *Crop breeding*. 2014, Springer. p. 75-86.
112. Livak, K.J., et al., *Oligonucleotides with fluorescent dyes at opposite ends provide a quenched probe system useful for detecting PCR product and nucleic acid hybridization.* 1995. **4**(6): p. 357-362.
113. Mao, F., W.-Y. Leung, and X. Xin, *Characterization of EvaGreen and the implication of its physicochemical properties for qPCR applications.* BMC Biotechnology, 2007. **7**(1): p. 76.
114. Hindson, B.J., et al., *High-throughput droplet digital PCR system for absolute quantitation of DNA copy number.* Anal Chem, 2011. **83**(22): p. 8604-10.
115. Tobler, A.R., et al., *The SNPlex genotyping system: a flexible and scalable platform for SNP genotyping.* Journal of biomolecular techniques : JBT, 2005. **16**(4): p. 398-406.
116. Lloyd-Price, J., G. Abu-Ali, and C. Huttenhower, *The healthy human microbiome.* Genome Medicine, 2016. **8**(1): p. 51.
117. Jones, J., et al., *Fecal sample collection methods and time of day impact microbiome composition and short chain fatty acid concentrations.* Scientific Reports, 2021. **11**(1): p. 13964.
118. Allali, I., et al., *A comparison of sequencing platforms and bioinformatics pipelines for compositional analysis of the gut microbiome.* BMC Microbiology, 2017. **17**(1): p. 194.
119. Walker, A.W., et al., *Phylogeny, culturing, and metagenomics of the human gut microbiota.* Trends in Microbiology, 2014. **22**(5): p. 267-274.
120. Forster, S.C., et al., *A human gut bacterial genome and culture collection for improved metagenomic analyses.* Nature Biotechnology, 2019. **37**(2): p. 186-192.
121. Louca, S., et al., *A census-based estimate of Earth's bacterial and archaeal diversity.* 2019. **17**(2): p. e3000106.
122. Hugerth, L.W., et al., *Metagenome-assembled genomes uncover a global brackish microbiome.* Genome Biology, 2015. **16**(1): p. 279.
123. Nayfach, S., et al., *New insights from uncultivated genomes of the global human gut microbiome.* Nature, 2019. **568**(7753): p. 505-510.
124. Almeida, A., et al., *A new genomic blueprint of the human gut microbiota.* Nature, 2019. **568**(7753): p. 499-504.
125. Yuan, C., et al., *Reconstructing 16S rRNA genes in metagenomic data.* Bioinformatics, 2015. **31**(12): p. i35-i43.
126. Hiseni, P., et al., *Liquid array diagnostics: a novel method for rapid detection of microbial communities in single-tube multiplex reactions.* BioTechniques, 2019. **66**(3): p. 143-149.
127. Chakravorty, S., et al., *A detailed analysis of 16S ribosomal RNA gene segments for the diagnosis of pathogenic bacteria.* Journal of microbiological methods, 2007. **69**(2): p. 330-339.
128. Bukin, Y.S., et al., *The effect of 16S rRNA region choice on bacterial community metabarcoding results.* Scientific Data, 2019. **6**(1): p. 190007.
129. Enroth, C.H., et al., *Excess primer degradation by Exo I improves the preparation of 3' cDNA ligation-based sequencing libraries.* Biotechniques, 2019. **67**(3): p. 110-116.
130. Green, M.R. and J.J.C.S.H.P. Sambrook, *Dephosphorylation of DNA fragments with alkaline phosphatase.* 2020. **2020**(8): p. pdb. prot100669.

131. Hiseni, P., *Liquid Array Diagnostics (LAD)-based multiplexed genotyping tests for double-muscling mutations in beef cattle*, in *Applied and Commercial Biotechnology*. 2016, Hedmark University of Applied Sciences: Hamar.
132. Koenig, J.E., et al., *Succession of microbial consortia in the developing infant gut microbiome.* 2011. **108**(supplement_1): p. 4578-4585.
133. Moore, R.E. and S.D. Townsend, *Temporal development of the infant gut microbiome.* 2019. **9**(9): p. 190128.
134. Avershina, E., et al., *Transition from infant- to adult-like gut microbiota.* Environ Microbiol, 2016. **18**(7): p. 2226-36.
135. Vebø, H.C., et al., *Temporal development of the infant gut microbiota in immunoglobulin E-sensitized and nonsensitized children determined by the GA-map infant array.* Clinical and vaccine immunology : CVI, 2011. **18**(8): p. 1326-1335.
136. RodrÍguez, J.M., et al., *The composition of the gut microbiota throughout life, with an emphasis on early life.* Microbial Ecology in Health and Disease, 2015. **26**(1): p. 26050.
137. Avershina, E., et al., *Major faecal microbiota shifts in composition and diversity with age in a geographically restricted cohort of mothers and their children.* FEMS Microbiology Ecology, 2014. **87**(1): p. 280-290.
138. Almeida, A., et al., *A unified catalog of 204,938 reference genomes from the human gut microbiome.* Nature Biotechnology, 2020.
139. Pruitt, K.D., T. Tatusova, and D.R. Maglott, *NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins.* Nucleic acids research, 2005. **33**(Database issue): p. D501-D504.
140. Jain, C., et al., *High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries.* Nature communications, 2018. **9**(1): p. 5114.
141. Hitch, T.C.A., et al., *Automated analysis of genomic sequences facilitates high-throughput and comprehensive description of bacteria.* ISME Communications, 2021. **1**(1): p. 16.
142. Richter, M. and R. Rosselló-Móra, *Shifting the genomic gold standard for the prokaryotic species definition.* 2009. **106**(45): p. 19126-19131.
143. Kim, M., et al., *Towards a taxonomic coherence between average nucleotide identity and 16S rRNA gene sequence similarity for species demarcation of prokaryotes.* 2014. **64**(Pt_2): p. 346-351.
144. Paradis, E. and K. Schliep, *ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R.* Bioinformatics, 2018. **35**(3): p. 526-528.
145. Louca, S., M. Doebeli, and L.W. Parfrey, *Correcting for 16S rRNA gene copy numbers in microbiome surveys remains an unsolved problem.* Microbiome, 2018. **6**(1): p. 41.
146. Perisin, M., et al., *16Stimator: statistical estimation of ribosomal gene copy numbers from draft genome assemblies.* The ISME Journal, 2016. **10**(4): p. 1020-1024.
147. Johnson, J.S., et al., *Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis.* Nature Communications, 2019. **10**(1): p. 5029.
148. Nguyen, N.-P., et al., *A perspective on 16S rRNA operational taxonomic unit clustering using sequence similarity.* npj Biofilms and Microbiomes, 2016. **2**(1): p. 16004.
149. Sun, D.-L., et al., *Intragenomic heterogeneity of 16S rRNA genes causes overestimation of prokaryotic diversity.* Applied and environmental microbiology, 2013. **79**(19): p. 5962-5969.
150. Devanga Ragupathi, N.K., et al., *Accurate differentiation of Escherichia coli and Shigella serogroups: challenges and strategies.* New microbes and new infections, 2017. **21**: p. 58-62.
151. Tan, J., et al., *Chapter Three - The Role of Short-Chain Fatty Acids in Health and Disease*, in *Advances in Immunology*, F.W. Alt, Editor. 2014, Academic Press. p. 91-119.
152. Torii, T., et al., *Measurement of short-chain fatty acids in human faeces using high-performance liquid chromatography: specimen stability.* 2010. **47**(5): p. 447-452.
153. Li, M., et al., *A sensitive method for the quantification of short-chain fatty acids by benzyl chloroformate derivatization combined with GC-MS.* Analyst, 2020. **145**(7): p. 2692-2700.
154. Rios-Covian, D., et al., *Enhanced butyrate formation by cross-feeding between Faecalibacterium prausnitzii and Bifidobacterium adolescentis.* FEMS Microbiology Letters, 2015. **362**(21).
155. Unger, M.M., et al., *Short chain fatty acids and gut microbiota differ between patients with Parkinson's disease and age-matched controls.* Parkinsonism & Related Disorders, 2016. **32**: p. 66-72.
156. Liu, S., et al., *Altered gut microbiota and short chain fatty acids in Chinese children with autism spectrum disorder.* Scientific Reports, 2019. **9**(1): p. 287.
157. Sanna, S., et al., *Causal relationships among the gut microbiome, short-chain fatty acids and metabolic diseases.* Nature Genetics, 2019. **51**(4): p. 600-605.
158. Reichardt, N., et al., *Phylogenetic distribution of three pathways for propionate production within the human gut microbiota.* The ISME Journal, 2014. **8**(6): p. 1323-1335.
159. Louis, P. and H.J. Flint, *Formation of propionate and butyrate by the human colonic microbiota.* Environmental microbiology, 2017. **19**(1): p. 29-41.
160. Thorkildsen, L.T., et al., *Dominant fecal microbiota in newly diagnosed untreated inflammatory bowel disease patients.* Gastroenterology research and practice, 2013. **2013**: p. 636785-636785.
161. Mehmood, T., et al., *A Partial Least Squares based algorithm for parsimonious variable selection.* Algorithms Mol Biol, 2011. **6**(1): p. 27.

162. Langille, M.G.I., et al., *Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences.* Nature Biotechnology, 2013. **31**(9): p. 814-821.
163. Aßhauer, K.P., et al., *Tax4Fun: predicting functional profiles from metagenomic 16S rRNA data.* Bioinformatics, 2015. **31**(17): p. 2882-2884.
164. Avershina, E. and K.J.B.m. Rudi, *Confusion about the species richness of human gut microbiota.* 2015. **6**(5): p. 657-659.
165. Siegwald, L., et al., *Assessment of Common and Emerging Bioinformatics Pipelines for Targeted Metagenomics.* PLOS ONE, 2017. **12**(1): p. e0169563.
166. Schmitt, M., et al., *Bead-Based Multiplex Genotyping of Human Papillomaviruses.* 2006. **44**(2): p. 504-512.
167. Coker, O.O., *Non-bacteria microbiome (virus, fungi, and archaea) in gastrointestinal cancer.* 2022. **37**(2): p. 256-262.
168. Hoffmann, C., et al., *Archaea and Fungi of the Human Gut Microbiome: Correlations with Diet and Bacterial Residents.* PLOS ONE, 2013. **8**(6): p. e66019.
169. Zhang, X., et al., *The oral and gut microbiomes are perturbed in rheumatoid arthritis and partly normalized after treatment.* Nature Medicine, 2015. **21**(8): p. 895-905.
170. Zhang, S., et al., *Human oral microbiome dysbiosis as a novel non-invasive biomarker in detection of colorectal cancer.* Theranostics, 2020. **10**(25): p. 11595-11606.
171. Yang, I., et al., *The oral microbiome and inflammation in mild cognitive impairment.* Experimental Gerontology, 2021. **147**: p. 111273.
172. Long, J., et al., *Association of oral microbiome with type 2 diabetes risk.* 2017. **52**(3): p. 636-643.

Papers I-IV

Paper I

Reports

BioTechniques

Liquid array diagnostics: a novel method for rapid detection of microbial communities in single-tube multiplex reactions

Pranvera Hiseni*[1,3], Robert C Wilson[2], Ola Storrø[4], Roar Johnsen[4], Torbjørn Øien[4] & Knut Rudi[1,2]

ABSTRACT

We present a novel liquid array diagnostics (LAD) method, which enables rapid and inexpensive detection of microbial markers in a single-tube multiplex reaction. We evaluated LAD both on pure cultures, and on infant gut microbiota for a 15-plex reaction. LAD showed more than 80% accuracy of classification and a detection limit lower than 2% of the Illumina reads per sample. The results on the clinical dataset showed that there was a rapid decrease of staphylococci from 10-day- to 4-month-old children, a peak of bifidobacteria at 4 months, and a peak of *Bacteroides* in 2-year-old children, which is in accordance with findings described in the literature. Being able to detect up to 50 biomarkers, LAD is a suitable method for assays where high throughput is essential.

METHOD SUMMARY

Liquid array diagnostics use short DNA duplexes, where one of the oligonucleotides is labeled with a fluorophore and the other, upon the presence of target DNA, becomes labeled with a quencher molecule. The novelty of this method lies in the combination of many duplex melting profiles and several channels of detection on a qPCR instrument, to detect multiple events of fluorescence quenching in a single-tube multiplex reaction.

KEYWORDS:

fluorescence • liquid array • microbiota • qPCR instrument • quenching

*[1]Department of Chemistry, Biotechnology & Food Sciences, Norwegian University of Life Sciences, P.O. Box 5003, 1432 Aas, Norway; [2]Inland Norway University of Applied Sciences, Hamar, Norway; [3]Genetic Analysis AS, Oslo, Norway; [4]Department of Public Health and Nursing, Norwegian University of Science & Technology, Trondheim, Norway; *Author for correspondence: ph@genetic-analysis.com; pranvera.hiseni@nmbu.no*

BioTechniques 66: 143-149 (March 2019) 10.2144/btn-2018-0134

The field of gut microbiota analysis has, until now, been dominated by relatively small-scale explorative studies, with several contradicting findings obscuring the truth in literature [1,2]. We are therefore at a stage where high-throughput, low-cost, targeted approaches are needed in order to generalize knowledge, and to evaluate previous findings. Presently, the GA-map® platform (Genetic Analysis AS) is the only clinically validated tool designated for gut microbiota diagnostics. The GA-map method allows for the faster assessment of the abundance of microbial markers in a sample, compared with NGS techniques [1]. However, it is based on solid-phase hybridization, which creates a bottleneck in sample processing and renders the test relatively expensive.

In this article, we present liquid array diagnostics (LAD), a novel approach for detecting bacterial communities using real-time PCR instrumentation. LAD combines single nucleotide primer extension with high-resolution melting (HRM) in the concept of a liquid array. It does not require physical separation of the probes prior to detection, thus avoiding a bottleneck in sample processing and ensuring rapid results at very low running costs. Requiring only a qPCR instrument, it has great potential for use as a routine tool for diagnostics by reporting multiple gut microbial markers in a single-tube multiplex reaction within a working day. A schematic outline of LAD is provided in Figure 1.

We evaluated LAD both on pure cultures, and on infant gut microbiota. The rationale for investigating the infant gut microbiota is that their composition and development are well described by many studies [2–4], and that we can utilize an already designed and validated GA-map probe set [5]. Furthermore, the development of the gut microbiota during infancy is crucial for health later in life. However, large-scale validation studies are required before knowledge about the gut microbiota can be utilized in clinical practice.

We present results demonstrating the sensitivity and specificity of LAD, in addition to exemplifying its utility on a medium-scale clinical cohort.

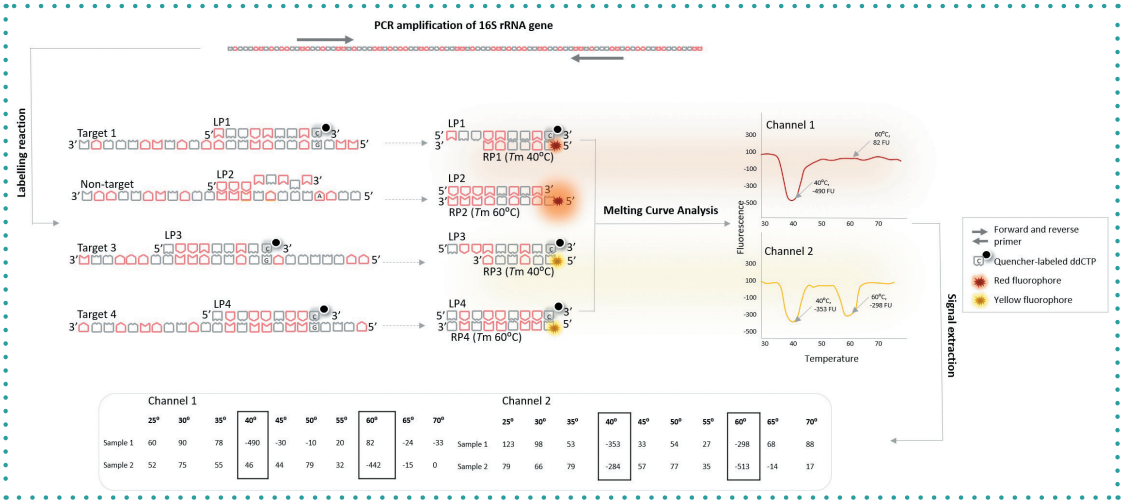
Taken together, LAD is a promising method, filling the need for large-scale gut microbiota validation tools.

MATERIALS & METHODS

Template generation for labeling probes labeling

We used genomic DNA extracted from 18 different bacterial isolates for PCR amplification. These strains represented targets for one or more labeling probes (LP), thus the purpose was to use them for validation of specificity and reproducibility of our assay. The chosen bacteria were: *Gemella sanguinis, Escherichia coli, Salmonella bongori, Salmonella enterica, Salmonella typhimurium, Klebsiella pneumoniae* subsp. *Pneumonae, Streptococcus pyogenes, Streptococcus pneumoniae, Salmonella enterica* subsp. *Enterica, Bacteroides vulgatus, Bacteroides fragilis, Bacteroides dorei, Staphylococcus aureus* subsp. *Aureus, Staphylococcus aureus, Bifidobacterium breve, Bifidobacterium longum, Enterococcus faecalis* and *Streptococcus sanguinis*.

In addition, DNA extracted from 541 PACT (Prevention of Allergy Among Children in Trondheim) study stool samples was utilized in PCRs to generate the templates for LP labeling. These samples were collected from pregnant mothers and their children at up to several post-birth ages. Their ▸



Channel 1	25°	30°	35°	40°	45°	50°	55°	60°	65°	70°
Sample 1	60	90	78	-490	-30	-10	20	82	-24	-33
Sample 2	52	75	55	46	44	79	32	-442	-15	0

Channel 2	25°	30°	35°	40°	45°	50°	55°	60°	65°	70°
Sample 1	123	98	53	-353	33	54	27	-298	68	88
Sample 2	79	66	79	-284	57	77	35	-513	-14	17

Figure 1. An overview of liquid array diagnostics (LAD) method. The initial step includes PCR amplification of 16S rRNA gene, where each LP is targeted. If the target DNA is present, LPs become labeled with a ddCTP conjugated with a quencher molecule. Subsequently, fluorophore-labeled RPs complementary to LPs are added into the solution mix. Upon duplex formation, at a specific melting temperature, the fluorescence of the reporter decreases abruptly. Multiple targets can be detected in a single-tube reaction by combining different duplex melting temperatures and fluorophore colors. In the last step, the derivative fluorescence units (FU) are extracted from each temperature where signals are expected for further data processing.
LP: Labeling probe; RP: Reporter probe.

▶ distribution was as following: 110 were collected from pregnant mothers, 122 from children up to 10 days old, 126 samples from 4-month-old children, 89 samples from 1-year-old children, and 94 from 2-year-old children. We used gDNA that was already extracted. The extraction protocol can be found in the Materials and Methods section of Vebø et *al.* (2011) [5].

A total reaction volume of 25 μl comprising 1 μl bacterial lysate as a source of DNA template, 0.05 U HOT FIREPol® DNA Polymerase, 1X B1 buffer, 2.5 mM $MgCl_2$ (all from Solis Biodyne, Estonia), 0.2 mM dNTPs (Thermo Fisher Scientific, MA, USA), 0.2 μM sense primer (Mangala F-1, 5′-TCCTACGGGAGGCAGCAG-3′), and 0.2 μM antisense primer (16SUR, 5′-3′ CGGTTACCTTGTTACGACTT) was designed to amplify a segment of 16S rRNA gene. PCR amplification was initiated with a period of 15 min at 95°C to activate the DNA polymerase, followed by 30 cycles, each consisting of 30 s denaturation at 95°C, 30 s annealing at 55°C and an 80 s elongation at 72°C performed using an Applied Biosystems Veriti™ Thermal Cycler (Life Technologies, CA, USA). A final elongation step of 7 min at 72°C was also included. The amplified products were treated with 2.4 U of Exonuclease I (ExoI, Biolabs Inc., MA, USA) and 6.4 U of shrimp alkaline phosphatase (USB Corporation, OH, USA) prior to incubation at 37°C for 120 min, and at 80°C for 15 min.

Single nucleotide extension of the LPs

A total reaction volume of 15 μl comprising 5 μl Exo-SAP-treated template (or water as 'no template' control), LPs at a final concentration of 0.1 μM, 0.8 μM ddCTP-ATTO612Q (Jena Biosciences, Germany), 20 μM ddTTP, 1 mM $MgCl_2$, 1XC buffer and 0.25 U HOT TERMIPOL® DNA Polymerase (all from Solis Biodyne, Estonia) was prepared. Labeling reactions were performed using an Applied Biosystems Veriti™ Thermal Cycler, employing an activation step at 95°C for 12 min, followed by 40 cycles, each consisting of 96°C denaturation for 20 s and 60°C annealing/elongation for 40 s.

Melting curve analysis

5′ fluorescently labeled reporter probe(s) (RP) were added to the LP labeling reactions at a final concentration of 0.005 μM each, with the exception of RPs 1_1 RP ROX, 1_2_2 RP ROX, 6_2_2 RP HEX, 6_1_4 RP HEX and 2_4_1 RP FAM, which had a final concentration of 0.02 μM each; reagent S, available from INN (Inland Norway University of Applied Sciences, Norway), was also added to a final concentration of 0.1%. The melting curve analysis was performed using a 7500 Fast qPCR instrument (Applied Biosystems, USA) with the following dissociation steps: 95°C for 15 s, 30°C for 1 min, 95°C for 15 s and 60°C for 15 s. Fluorescence was detected and expressed in dissociation curves as the derivative of the fluorescence versus temperature measurements (dF/dT) versus temperature (Temp.). Positive signals were observed as negative peaks, representing the abrupt, temperature-dependent drop of fluorescence.

Extraction of peaks & determination of positive signals for clinical samples

For the sake of simplicity, all data were multiplied by −1 since originally, positive LAD signals have negative values.

Fluorescence values were extracted from temperature measurements where quenching signals were expected (e.g., the fluorescence value at 67.7°C on HEX channel, where UNI probe was designed to quench). In addition, such values were extracted from 5 no template controls (NTC), with the aim of determining the borderline

Table 1. Probes designed by Genetic Analysis AS for GA-map® array, used as labeling probes by liquid array diagnostics.

Probe identifier	Taxonomic group(s) detected	Probe sequence (5'–3')
1_1	*Bacteroides*	TTGCGGCTCAACCGTAAAATTG
1_2_2	*Bacteroides* (*dorei, fragilis, thetaiotaomicron, vulgatus*)	GCACTCAAGACATCCAGTATCAACTG
2_1_min1b	Gamma-proteobacteria	CAGGTGTAGCGGTGAAATGCGTAGAGAT
2_3_2	Gamma-proteobacteria subgroup	CGGGGATTTCACATCTGA
2_4_1	Gamma-proteobacteria subgroup	TGCCAGTTTCGAATGCAGTT
4_1	Firmicutes (Lactabacillales, *Clostridium perfringens, Staphylococcus*)	CGATCCGAAAACCTTCTTCACT
4_4_2	*Enterococcus, Listeria*	TCCAATGACCCTCCC
4_5_2	*Streptococcus pyogenes*	GATTTTCCACTCCCACCAT
4_6_1	*Streptococcus sanguinis*	CACTCTCACACCCGTT
4_8_1	*Streptococcus pneumoniae, Entrococcus*	CGCGGCGTTGCTCGGTCAGACTT
5_1	Firmucutes (Clostridia, Bacillales, *Enterococcus, Lactobacillus*)	GGACAACGCTTGCCAC
5_1_2	*Staphylococcus*	CGTGGCTTTCTGATTAGGTA
6_1_4	*Bifidobacterium longum*	TGCTTATTCAACGGGTAAACT
6_2	Actinobacteria	CGTAGGCGGTTCGTCGCGT
6_2_2	*Bifidobacterium breve*	CGGTGCTTATTCGAAAGGTACACT
UNI01	16S Universal	CGTATTACCGCGGCTGCTGGCA

separating positive signals from background fluorescence. First, we calculated the distance of the observed positive signals from the mean background fluorescence using a standard Z-score. Following that, the margin separating the signals from the background was assigned to be the mean value of NTC plus two-times its standard deviation ($\mu+2\Sigma$).

However, a different approach was used to assign positive signals for 5_1_2. Considering that there is a tight melting temperature (*T*m) range separating 5_1 from 5_1_2 signals, using the above-mentioned formula would report false-positive signals for *Staphylococcus* (5_1_2 probe) since the fluorescence measurements at 55.8°C, where 5_1_2 is designed to quench, are interferingly high for each sample where 5_1 is truly quenched (50.8°C). Thus, fluorescence values at 55.8°C were extracted from eight random samples where only 5_1 was observed to give signal. The mean value of these samples was added with three standard deviations ($\mu+3\Sigma$), which was used as a margin to separate the bona fide *Staphylococcus* signals. All data values higher than the margins were accepted as positives.

Probe design

The probes, designed by Genetic Analysis AS [5], were used as LPs (Table 1), whereas the RPs were designed to be complementary to the LPs, so that they create duplexes that dissociate at a chosen temperature. Each probe has a code identifier (for example 1_1 for *Bacteroides*), originally used in Vebø *et al.* (2011) [5]. The *T*m of the probes was calculated by the Oligoanalyzer 3.1 web-based bioinformatics tool (Integrated DNA Technologies) and target *T*ms were achieved by varying the length of the RPs.

The reporter probes were designed to anneal to the 3'-end of each respective labeling probe, thus placing the fluorophore, coupled to the terminal 5'nucleotide of the RP, in close physical proximity to the quencher molecule located at the 3' end of the labeled LP. The list of the reporter probes is presented in Table 2.

Comparison of LAD-based results with Illumina sequencing data

87 random PACT samples (34 samples of children up to 10-days old, 15 of 4-month-olds, 15 of 1-year-olds, 12 of 2-year-olds and 11 of pregnant women) were picked to be sequenced with an Illumina MiSeq System (Illumina, CA, USA). The purpose of this step was to confirm the identities of samples and compare them with the results obtained with the LAD assay, by performing *in silico* labeling of the reads. *In silico* labeling was performed by textual mapping of the 'labeled' LPs to the operational taxonomic unit (OTU) DNA sequences retrieved by Illumina, using the Sequence Manipulation Suite: Primer Map tool [6]. All OTUs that were detected by the same probe were grouped together and their number of reads was summed up for each sample. The total number of such reads was then compared with the LAD signal intensity for the said probe. Prior to doing so, LAD data were normalized so that any number below the cut-off value would be equal to zero.

To calculate the specificity and sensitivity, we performed a receiver operator characteristic (ROC) curve analysis (MedCalc Software, Ostend, Belgium), which plots the true positive signals (as determined with LAD) against the false positives for different cut-off points (the number of Illumina reads). This helped find the optimum copy number of target sequences that can be detected using our method. ▸

Table 2. Reporter probe sequences.

Reporter probe	5′–3′ sequence
1_1 RP ROX	/56-ROXN/TTTCAATTTTACGG
1_2_2 RP ROX	/56-ROXN/TTTCAGTTGATACTGG
2_1_min1b RP ROX	/56-ROXN/TATCTCTACGCATTTCACCGCTACA
2_3_2 RP ROX	/56-ROXN/TTTCAGATGTGAAATCCC
4_1 RP CY5	/5CY5/TTTAGTGAAGAAG
4_5_2 RP CY5	/5CY5/TATGGTGGGAGT
4_8_1_RP2_CY5	/5CY5/TAAGTCTGACCGAGCAACGCCGC
4_6_1 RP CY5	/5CY5/TTAACGGGTGTGAGAGTG
2_4_1 RP FAM	/56-FAM/TAACTGCATTC
4_4_2 RP FAM	/56-FAM/TTTGGGAGGGTCAT
5_1 RP FAM	/56-FAM/TTTGTGGCAAGCGTTG
5_1_2 RP FAM	/56-FAM/TTACCTAATCAGAAAGCCACG
6_2 RP HEX	/5HEX/TTTTACGCGACG
6_2_2 RP HEX	/5HEX/TTAGTGTACCTTTCG
6_1_4 RP HEX	/5HEX/TTAGTTTACCCGTTGAAT
UNI01 RP HEX	/5HEX/TTGCCAGCAGCCGCGGTAATACG

Subsequently, for each probe, the numbers of the Illumina reads lower than LAD detection limit were equated to zero, to test the correlation of the positive signals using Spearman's Rho test.

Statistical analysis

Minitab Release 15.1.1.0 (Minitab Inc. 2007) was used to perform Student's t-test to compare the differences on quenching strength (fluorescence mean value) between cohorts. For the sake of illustration, the data were normalized so that the cut-off value equals zero. In addition, the differences regarding the prevalence of positive signals were analyzed by using Pearson's chi-squared test.

RESULTS & DISCUSSION

Optimization of the LAD-based microbiota detection assay

Based on pure cultures, we first adjusted the level of probes present in the reaction in order to achieve high signal-to-noise ratios. A detailed description of the experimental setup used in the evaluation is provided in Supplementary Figures S1, S2 and S3. This process was performed empirically (see supplement for details), resulting in an assay that was capable of reporting 15 distinguishable signals in a one-tube multiplex reaction, consisting of probes reported in Table 1. The signals for each of our probes, besides 6_2 duplex, were at least two standard deviations above the average value of no target reactions, which represented the background noise (Z>2), with a p-value < 0.02 (Table 3).

The initial evaluation of the assay performance was based on comparisons between experimental and theoretical signals, derived from Vebø *et al.* (2011) [5]. This analysis showed that the accuracy and specificity of probes was very high, reporting only the target strains in reactions holding individual bacteria or defined bacteria mixtures (Figure 2).

Comparison of LAD-based microbiota assay with Illumina sequencing

To compare LAD with the output of Illumina sequencing, we sequenced 87 clinical samples, then performed *in silico* labeling of the retrieved sequences for the nine probes covered by the sequencing amplicon. Subsequently, for each probe we compared LAD signals with the number of sequence copies that acted as a template during *in silico* labeling. Specifically, we performed ROC curve analysis for each probe to determine accuracy of classification, and to determine limit of detection for the LAD assay. For most of the probes the accuracy of detection, i.e. the number of correct predictions, was high (>80%). The detection limit for the probes was between 0.2 and 2%, as determined by the percentage of Illumina sequencing reads detected. Furthermore, there was a significant quantitative correlation between Illumina read counts and LAD signals ($p < 0.05$), with Spearman's rho ranging between 0.45 and 0.86 for all the probes (Table 4).

Use of LAD to genotype clinical samples

The verified assay was used to probe the microbiota composition from 541 PACT study fecal samples from infants and their mothers.

The highest number of positive signals was reported for 5_1 and 6_2_2 probe duplexes, designed to detect Firmicutes and *Bifidobacterium breve*, respectively. Overall, the results showed that in terms of prevalence, there is overrepresentation of gammaproteobacteria and *Enterococcus*/*Listeria* in 4-month-old children, *Bacteroides* at 2-years old, *Bifidobacterium* at 4 months and *Staphylococcus* in 10-day-old children (Figure 3).

Table 3. Probe signal-to-noise ratios.

Probe	Average of positive signals (μ_1)	Average of NT signals (μ_2)	NT standard deviation (Σ)	Z-score($(\mu_1 - \mu_2)/\Sigma$)	p-value
6_2	−287.5	70.4	87.9	−4.1	>0.99
UNI01	1704.1	−331.1	58.2	34.9	<0.0002
6_2_2	723.8	−91.2	121.8	6.7	<0.0002
6_1_4	2022.9	−7.2	95.9	21.2	<0.0002
4_1	−537.4	−2599.6	896.6	2.3	0.01
4_5_2	−446.3	−2398.5	826.1	2.3	0.01
4_6_1	−364.9	−2013.9	816.4	2.0	0.02
4_8_1†	N/D			N/D	N/D
1_1	2062.9	−569.1	421.3	6.2	<0.0002
1_2_2	2010.5	−336.4	106.5	22.0	<0.0002
2_3_2	2230.5	−326.5	234.0	10.9	<0.0002
2_1_min1b	1865.0	−473.9	83.4	28.0	<0.0002
2_4_1	117.9	−836.5	365.8	2.6	0.0047
4_4_2	1251.5	−801.7	198.5	10.3	<0.0002
5_1	1355.9	−1021.1	109.9	21.6	<0.0002
5_1_2	2286.7	−515.5	520.8	5.4	<0.0002

†The fluorescence values for 4_8_1 probe duplex were not determinedbecause we lacked the DNA template.
N/D: No data; NT: No template.

The signal strength for *S. pyogenes* and *S. sanguinis* had a peak in 1-year-old children, while the strongest signals for *S. pneumoniae* were in 10-day-old children.

There was no significant change in prevalence for the probe detecting a group of species within Firmicutes (5_1 probe, detecting for Clostridia, Bacillales, *Enterococcus* and *Lactobacillus*); however, the signal strength showed an increase parallel with age. The opposite was observed for the other Firmicutes probe, 4_1 (detecting for *Lactobacillus*, *C. perfringens* and *Staphylococcus*), which had a decrease both on prevalence and signal strength in older children.

Use of LAD for rapid detection of microbial communities

Here we present LAD, a novel technique that combines single-nucleotide-extension of the probes with HRM analysis. Compared with existing tools for microbiome testing, LAD-based tests are simpler to perform, are cheaper as they do not require expensive instrumentation and reagents, and yield results faster, within a working day.

Our method does not require a dedicated instrument that would solely be used for LAD-based tests. It requires real-time PCR instrumentation, which is widely and commonly used in most laboratories. In comparison with other real-time PCR-based approaches, it offers a higher level of multiplexity per well, few reagents and short hands-on time, satisfying the actual need of detecting a relatively low number of markers (<50) in a very large number of samples.

LAD represents a highly reproducible method. Initially, the designed probes undergo a process of validation for their specificity, which ensures that all probes become labeled only when their target is present in the reaction. Further on, each labeled probe is tested to ensure it hybridizes only with its corresponding ▶

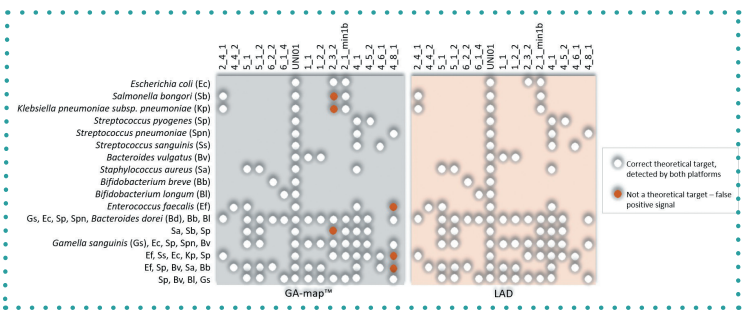


Figure 2. Evaluation of LAD probe accuracy and sensitivity. All signals that were at least two standard deviations away from the background fluorescence were accepted as positives. Tests on individual bacterial strains or defined mixtures of bacteria showed identical results for the correct targets on both platforms, GA-map® (left) and LAD (right). No false-positive signals, reporting nontargets, were registered with LAD.

Figure 3. Signal strength and prevalence of positive signals. Significant differences, that here are depicted with *, were observed between groups for most of the probes.

▸ reporter probe, thus avoiding false-positive signals being generated in the presence of a nontarget probe.

The addition of a synthetic quencher-and-fluorophore-labeled duplex (designated *Tm* and detection channel) into the master mix will provide the basis for a well-to-well data normalization, ensuring reproducibility.

For this study, we chose to adopt probes designed by Genetic Analysis AS [5], considering that their platform, GA-map, is an already validated method based on single nucleotide extension. Results obtained with GA-map served as a reference and allowed us to evaluate the overall performance of LAD. We found highly comparable probe specificities using the two technologies, suggesting the transferability of GA-map probes to LAD detection.

Our results on the clinical dataset show that there is a rapid decrease of staphylococci from 10-day- to 4-month-old children, and a peak of bifidobacteria at the age of 4 months, which is in full accordance with the previous findings made with GA-map [5]. However, we identified a peak of *Bacteroides* in 2-year-old children, whereas Vebø *et al.* (2011) [5] found that *Bacteroides* were overrepresented in 4-month-old children. This may be explained by the fact that we did not test an identical set of samples, since an increase of *Bacteroides* in older children has already been described from many other papers in the literature [7,8].

In addition, we compared our assay with the outcome of Illumina MiSeq sequencing, which demonstrated a high classification accuracy and low detection limit for LAD, providing evidence of its sensitivity. The quantitative comparisons, however, showed some more deviations between the two platforms. Unfortunately, we could not Illumina-sequence the ~1200 bp PCR fragment analyzed with LAD due to the 300-bp limitation in Illumina read-length chemistry, which could potentially explain the differences between the two sets of results.

Numerous gut microbial markers that are linked with many disorders such as obesity [9–11], diabetes [11–13], multiple sclerosis [14,15] or irritable bowel syndrome [1,16] have already been described, but these have not yet been clinically validated in large-scale multicenter studies. With its main advantage of being very cheap, rapid

Table 4. Evaluation of the diagnostic ability of liquid array diagnostics-based tests.

Probe	2_4_1	5_1	5_1_2	1_1	1_1_2	2_3_2	2_1_min1b	4_1	4_8_1
Detection limit (%)	0.8	0.4	1.2	2.4	0.1	1.1	0.4	0.022	0.002
Sensitivity (%)	90	91.9	93.3	84.8	82	85	65.5	62.7	69.2
Specificity (%)	92.1	76.9	95.8	91.7	65.5	83	86.2	85.7	68.9
Spearman's Rho	0.74	0.82	0.86	0.81	0.57	0.72	0.66	0.65	0.45

and simple, in addition to being an accurate method, LAD will offer this possibility.

We acknowledge the limitations of our method regarding systems where the microbiome composition is complex, unpredictable and constantly shifting. Building a LAD assay *de novo* is best conducted in systems with relatively low complexity, where the knowledge regarding the microbiome composition is already described, such as is the case with gut microbiota. A well-defined composition is a prerequisite towards designing targeting probes.

Here, we used 15 different probe duplexes, which were designed to utilize four channels of detection and at least three *T*ms per channel. By using a qPCR machine with six channels of detection and exploiting at least six resolvable *T*ms per channel, the multiplex level can be elevated to at least a 36-plex. Thus, the possibility of multiplexing is limited by the instrument, and not by LAD technology in itself.

In conclusion, we believe LAD will fulfill the need for assays able to detect up to 50 biomarkers, where high throughput is essential. This will particularly relate to human gut microbiota markers related to health and disease.

AUTHOR CONTRIBUTIONS

PH, RW and KR conceived and designed the experiments and drafted the work; OS, RJ and TØ helped with the sample collection. All authors contributed equally to analysis and interpretation of data and critical revision of the drafts. All authors approved the final version to be published.

FINANCIAL & COMPETING INTERESTS DISCLOSURE

This work was financially supported by Norway Research Council through R&D project grant no 283783. The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

No writing assistance was utilized in the production of this manuscript.

OPEN ACCESS

SUPPLEMENTARY DATA

To view the supplementary data that accompany this paper please visit the journal website at: www.future-science.com/doi/suppl/10.2144/btn-2018-0134

REFERENCES

1. Casen C,Vebø HC, Sekelja M *et al.* Deviations in human gut microbiota: a novel diagnostic test for determining dysbiosis in patients with IBS or IBD. *Aliment. Pharmacol. Therap.* 42(1), 71–83 (2015).
2. Milani C, Duranti S, Bottacini F *et al.* The first microbial colonizers of the human gut: composition, activities, and health implications of the infant gut microbiota. *Microbiol. Mol. Biol. Rev.* 81(4), e00036–17 (2017).
3. Dogra S, Sakwinska O, Soh SE *et al.* Dynamics of infant gut microbiota are influenced by delivery mode and gestational duration and are associated with subsequent adiposity. *MBio* 6(1), e02419–14 (2015).
4. Rodríguez JM, Murphy K, Stanton C *et al.* The composition of the gut microbiota throughout life, with an emphasis on early life. *Microbial Ecol. Health Dis.* 26(1), 26050 (2015).
5. Vebø HC, Sekelja M, Nestestog R *et al.* Temporal development of the infant gut microbiota in immunoglobulin E-sensitized and nonsensitized children determined by the GA-map infant array. *Clin. Vaccine Immunol.* 18(8), 1326–1335 (2011).
6. Stothard P. The sequence manipulation suite: JavaScript programs for analyzing and formatting protein and DNA sequences. *Biotechniques* 28(6), 1102–1104 (2000).
7. Arrieta MC, Stiemsma LT, Amenyogbe N, Brown EM, Finlay B. The intestinal microbiome in early life: health and disease. *Front. Immunol.* 5, 427 (2014).
8. Cheng J, Ringel-Kulka T, Heikamp-de Jong I *et al.* Discordant temporal development of bacterial phyla and the emergence of core in the fecal microbiota of young children. *ISME J.* 10(4), 1002–1014 (2016).
9. Guo X, Li S, Zhang J *et al.* Genome sequencing of 39 ***Akkermansia muciniphila*** isolates reveals its population structure, genomic and functional diversity, and global distribution in mammalian gut microbiotas. *BMC Genomics* 18(1), 800 (2017).
10. Gerard P. Gut microbiota and obesity. *Cell Mol. Life Sci.* 73(1), 147–162 (2016).
11. Hartstra AV, Bouter KE, Bäckhed F, Nieuwdorp ML. Insights into the role of the microbiome in obesity and Type 2 diabetes. *Diabetes Care* 38(1), 159–165 (2015).
12. Allin KH, Tremaroli V, Caesar R *et al.* Aberrant intestinal microbiota in individuals with prediabetes. *Diabetologia* 61(4), 810–820 (2018).
13. Delzenne NM, Cani PD, Everard A, Neyrinck AM, Bindels LB. Gut microorganisms as promising targets for the management of Type 2 diabetes. *Diabetologia.* 58(10), 2206–2217 (2015).
14. Chen J, Chia N, Kalari KR *et al.* Multiple sclerosis patients have a distinct gut microbiota compared to healthy controls. *Sci. Rep.* 6, 28484 (2016).
15. Miyake S, Kim S, Suda W *et al.* Dysbiosis in the gut microbiota of patients with multiple sclerosis, with a striking depletion of species belonging to Clostridia XIVa and IV clusters. *PLOS One* (2015).
16. Bennet SMP, Öhman L, Simrén M. Gut microbiota as potential orchestrators of irritable bowel syndrome. *Gut Liver* 9(3), 318–331 (2015).

Paper II

Hiseni et al. Microbiome (2021) 9:165
https://doi.org/10.1186/s40168-021-01114-w

Microbiome

RESEARCH Open Access

HumGut: a comprehensive human gut prokaryotic genomes collection filtered by metagenome data

Pranvera Hiseni[1,3*], Knut Rudi[1,2], Robert C. Wilson[2], Finn Terje Hegge[3] and Lars Snipen[1]

Abstract

Background: A major bottleneck in the use of metagenome sequencing for human gut microbiome studies has been the lack of a comprehensive genome collection to be used as a reference database. Several recent efforts have been made to re-construct genomes from human gut metagenome data, resulting in a huge increase in the number of relevant genomes. In this work, we aimed to create a collection of the most prevalent healthy human gut prokaryotic genomes, to be used as a reference database, including both MAGs from the human gut and ordinary RefSeq genomes.

Results: We screened > 5,700 healthy human gut metagenomes for the containment of > 490,000 publicly available prokaryotic genomes sourced from RefSeq and the recently announced UHGG collection. This resulted in a pool of > 381,000 genomes that were subsequently scored and ranked based on their prevalence in the healthy human metagenomes. The genomes were then clustered at a 97.5% sequence identity resolution, and cluster representatives (30,691 in total) were retained to comprise the HumGut collection. Using the Kraken2 software for classification, we find superior performance in the assignment of metagenomic reads, classifying on average 94.5% of the reads in a metagenome, as opposed to 86% with UHGG and 44% when using standard Kraken2 database. A coarser HumGut collection, consisting of genomes dereplicated at 95% sequence identity—similar to UHGG, classified 88.25% of the reads. HumGut, half the size of standard Kraken2 database and directly comparable to the UHGG size, outperforms them both.

Conclusions: The HumGut collection contains > 30,000 genomes clustered at a 97.5% sequence identity resolution and ranked by human gut prevalence. We demonstrate how metagenomes from IBD-patients map equally well to this collection, indicating this reference is relevant also for studies well outside the metagenome reference set used to obtain HumGut. All data and metadata, as well as helpful code, are available at http://arken.nmbu.no/~larssn/humgut/.

Keywords: Human gut microbiome, Genome collection, Database

* Correspondence: ph@genetic-analysis.com; pranvera.hiseni@nmbu.no
[1]Department of Chemistry, Biotechnology and Food Sciences, Norwegian University of Life Sciences, P.O. Box 5003, 1432 Ås, Norway
[3]Genetic Analysis AS, Kabelgaten 8, 0580 Oslo, Norway
Full list of author information is available at the end of the article

Introduction

Significant efforts have been undertaken to characterize the human gut microbiome, both by microbe isolation and DNA sequencing [1]. A major contribution has also been made by de novo-assembled genomes (Metagenome-Assembled Genomes—MAGs), facilitated by the latest advances in bioinformatics tools [2–6]. As a wrap, a Unified Human Gastrointestinal Genome (UHGG) collection comprised of > 200,000 non-redundant reference genomes was recently announced [7], marking a major milestone in this field.

These studies have laid a solid foundation, identifying a vast variety of genomes encountered in human guts. However, none of them addresses the global prevalence of genomes within healthy people, i.e., providing information about their frequency of occurrence. This knowledge is essential for setting up a collection of human gut-associated prokaryotic genomes that reflects the worldwide healthy human gut microbiome. It is especially important for building custom databases intended to be used for comparative studies in human gastrointestinal microbiome research.

Regionally, studies have shown that the intestinal flora is greatly shaped by the environment [8] and that its composition can be linked to a range of diseases and disorders [9–12]; thus, we are now at a stage where gut microbiota therapeutic interventions are being introduced [13, 14]. However, the lack of a global reference for the intestinal flora in healthy humans represents a bottleneck [15]. This impedes both the understanding of gut microbiota on a worldwide scale and the introduction of large-scale intervention strategies.

The aim of this work was to create a single, comprehensive genome collection of gut microbes associated with healthy humans, called HumGut, as a universal reference for all human gut microbiota studies. We utilized the UHGG collection, mentioned above, along with the NCBI RefSeq genomes. The strategy of building HumGut is outlined in Fig. 1.

HumGut genomes are ranked by their containment in healthy human gut metagenomes collected worldwide. The most commonly encountered genomes (i.e., top-ranked on the list) were selected as taxa representatives during dereplication, securing thus a list of those most relevant.

While it may seem like a relatively simple concept, this work has only become possible with the recent development of bioinformatics tools that allow the swift screening of publicly available human gut metagenomes for the containment of the ever-growing pool of prokaryotic genomes.

Results

Reference metagenomes

More than 5,700 gut metagenome samples collected from healthy people of various ages worldwide were downloaded. These belonged to 72 different BioProjects. To avoid the bias of containing groups of highly similar samples, we computed the MASH distance between metagenomes within each BioProject, then clustered samples with ≥ 95% sequence identity. From each cluster, we only kept the medoid sample, resulting thus in a collection of 3,534 healthy human gut metagenomes (Fig. 2a).

On average, samples within each project shared a 90% sequence identity (D = 0.1), indicating a relatively high degree of similarity between one another. There were some outliers, however. Some infant samples (10 belonging to PRJNA473126 project and 1 to PRJEB6456), 10 samples from a project studying the human gut microbiome of vegetarians, vegans, and omnivores (PRJNA421881), and a sample from a study focusing on microbiome diversity among Cheyenne and Arapaho of Oklahoma (PRJNA299502), showed the highest dissimilarity with at least one other sample from the same project (D = 1) (Fig. 2b).

We wanted to see if samples clustered based on their continent of origin (Fig. 2c). To do so, we computed the average linkage hierarchical clustering of BioProjects. The distance between two BioProjects is the mean pairwise distance between all their samples. Here, we also included a BioProject containing primate gut metagenome samples (n = 95) as an outgroup against which all human BioProjects were compared. The lowest and highest observed average MASH distances (D = 0.05, and D = 0.14, respectively) were between two sets of projects stemming from separate continents each, one from Europe and the other from North America. These observations, together with the mixed distribution of BioProjects in the cluster dendrogram, suggested that the clustering of samples did not heavily depend on continent-of-origin. The primate samples were markedly separated from the rest of the tree, showing an average distance of 0.22 from all other BioProjects.

From genomes to HumGut collection

The majority of genomes stemming from the UHGG collection (99%) and 48% of RefSeq genomes qualified for inclusion in HumGut, resulting thus in a total collection of 381,779 genomes (Fig. 3a). The qualified genomes were contained within at least one reference metagenome. We inferred the containment by computing sequence identity between genomes and metagenomes using MASH screen, and considered a genome as contained when identity was ≥ 0.95.

By applying a rarefaction, we found that the number of new genomes saturated after screening for ca. 1,000 metagenomes, indicating that with > 3,500 metagenomes very few new genomes will be added if screening even more metagenomes from the same population (supplementary material, Figure S1).

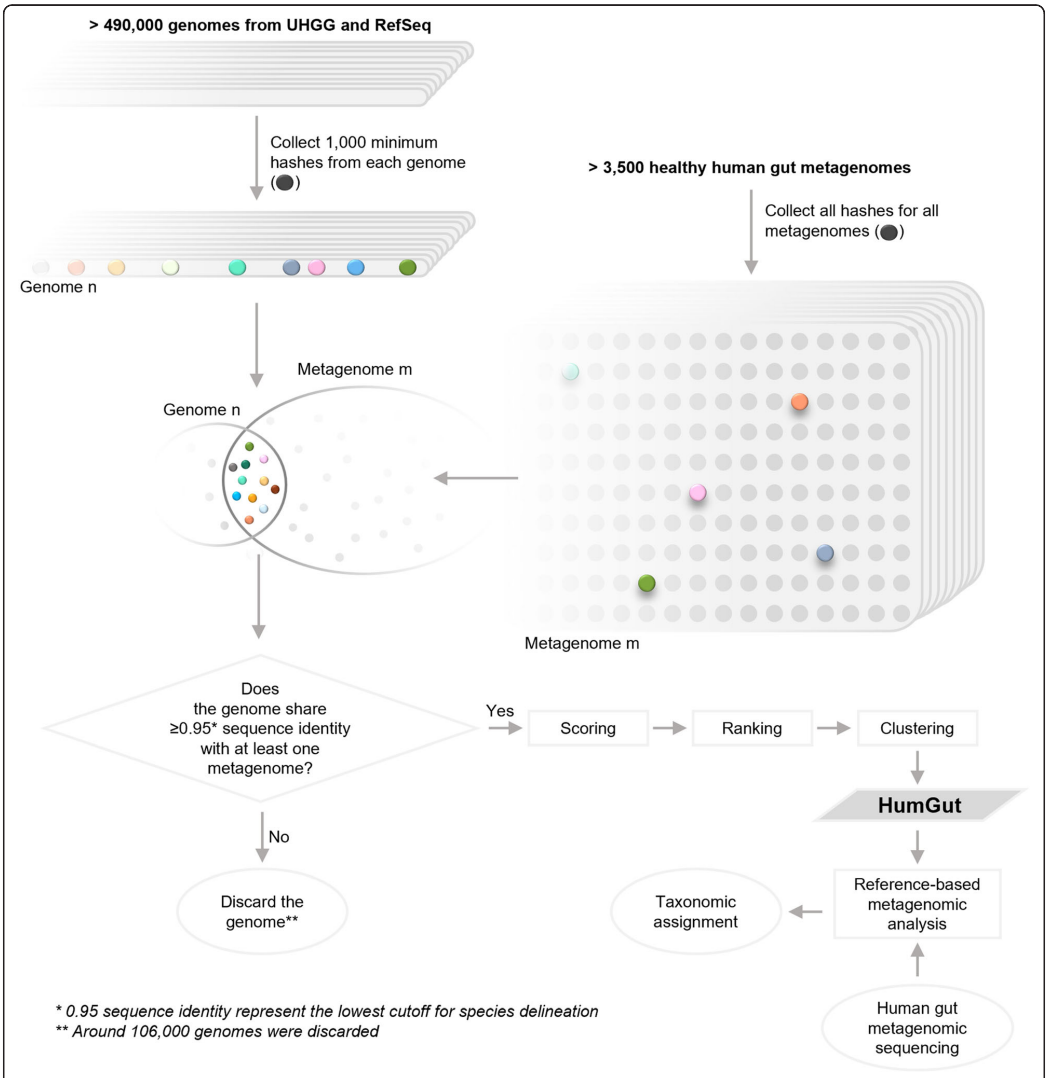


Fig. 1 HumGut overview. HumGut represents a collection of genomes and MAGs contained in 3,534 healthy human gut metagenomes. To be considered as contained, a genome shared at least 0.95 sequence identity with at least one of the metagenomes (inferred by the number of shared hashes). The qualified genomes were scored based on the average sequence identity across all the metagenomes. Next, they were ranked based on their scores: the higher the score, the higher the position on the list. Subsequently, the genomes were clustered based on MASH and fastANI distance (*D*). The top-ranked genome formed a cluster centroid. Around 30,600 clusters were formed applying a $D = 0.025$-threshold. The use of HumGut as a reference set helps the process of taxonomic assignments by drastically reducing the number of unclassified human gut metagenomic reads

The most prevalent genomes, i.e., the genomes contained in most metagenomes, belonged to the genus *Bacteroides*, led by *B. vulgatus* (also known as *Phocaeicola vulgatus*), found in more than 70% of samples. It is worth noting that the UHGG collection contained no genome with this species name. The genomes are named as *Bacteroides dorei* instead. We presume that is related to an earlier GTDB database release used for genome

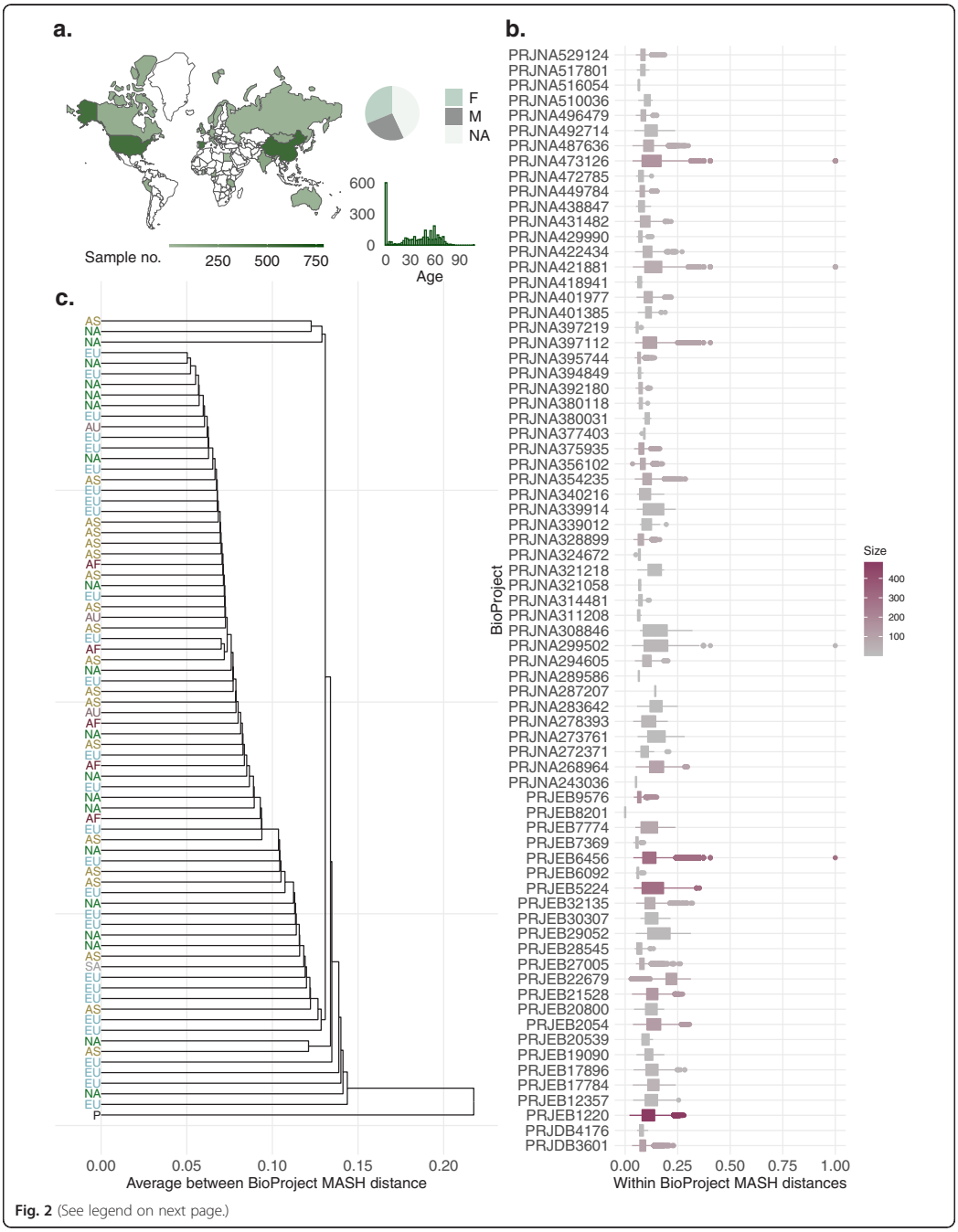


Fig. 2 (See legend on next page.)

(See figure on previous page.)
Fig. 2 An outline of the metagenomes used in this study. **a** The geographical, age, and gender distribution of 3,534 metagenomes collected from healthy people. **b** Boxplots illustrating the distribution of MASH distances between samples within each BioProject. The BioProject accession is used as a label, and the color gradient indicates the size, i.e., the number of samples in each. **c** Average linkage hierarchical clustering of 72 BioProjects containing healthy samples. BioProjects containing samples from different continents are presented separately. Labels indicate the continent of origin: EU—Europe, AS—Asia, NA—North America, AS—Australia, AF—Africa, SA—South America, and P stands for Primates. Except for the single primate BioProject (BioSample), each BioProject is listed in colored font according to the continent from which it originates. No severe clustering of samples based on origin is detected

taxonomic classifications by Almeida et al. (GTDB-Tk v0.3.1; database release 04-RS89) [7]. In the current version of GTDB, the species *Phocaeicola vulgatus* is listed.

We performed clustering of genomes based on sequence similarity using the top-ranked genome as a cluster centroid for each iteration. We initially applied an ANI threshold of 97.5% to compile a HumGut collection of highest resolution (HumGut_97.5). This collection resulted in 30,691 genomes with ≥ 50% genome completeness and ≤ 5% contamination. They were all given a GTDB-Tk taxonomic annotation [16] as well as an NCBI taxonomy assignment.

These genomes were subsequently clustered further to form a coarser collection at 95% identity, the HumGut_95 with 5,170 genomes. This corresponds roughly to species resolution [17].

Looking into genome sources, we found that 9% of HumGut_95 clusters were RefSeq-only genomes (Fig. 3c). These genomes, 756 in total, clustered into 460 HumGut_95 clusters, belonged to 125 different genera.

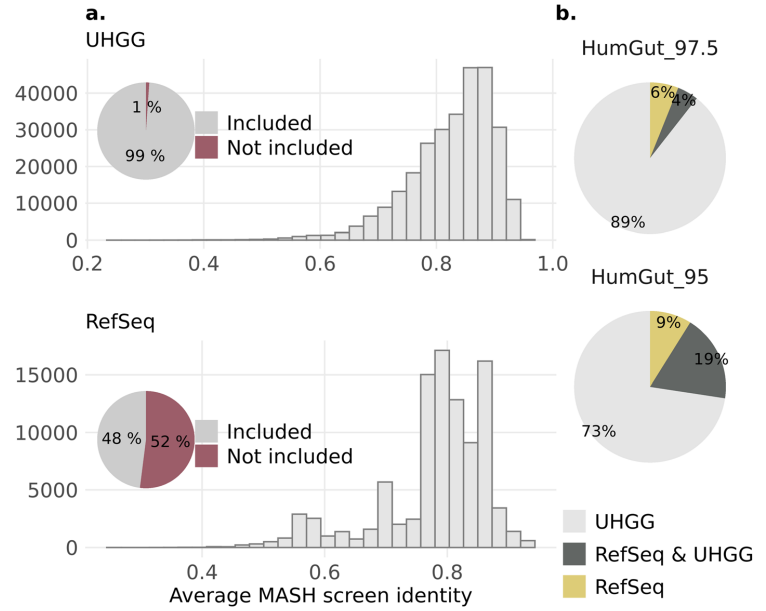


Fig. 3 An overview of the genomes used to build HumGut. **a** The pie charts show the proportion of genomes from each collection (UHGG above, RefSeq below) included in HumGut. To qualify for HumGut inclusion, genomes had to have at least 0.95 MASH screen identity with at least one healthy metagenome, as did most of the UHGG and half of the RefSeq genomes. Histograms show the distribution of the mean identity shared between the qualified genomes and healthy metagenomes. A high average identity means that the qualified genome has been found contained in most of the screened samples. **b** The genome sources for HumGut clusters. The upper pie chart shows data for 30,691 clusters belonging to HumGut_97.5 (genomes grouped based on 97.5% genome sequence identity); the bottom one presents data for 5,170 HumGut_95 clusters (95% sequence identity—species level threshold). The majority of clusters in both HumGut collections are comprised of only UHGG genomes, while 6% and 9% of the clusters consist of only RefSeq genomes (HumGut_97.5 and HumGut_95, respectively)

Most of the genomes (299 in total) belonged to various *Streptococcus* species.

HumGut genome clusters

Not all species-level clusters were equally diverse, that is, not all of them encompassed a similar number of HumGut_97.5 clusters. The majority of HumGut_95 clusters (3,009 out of 5,100) consisted of a single HumGut_97.5 cluster. On the other hand, the most diverse HumGut_95 cluster was one built of 533 different HumGut_97.5 clusters, all named as *Agathobacter rectalis* with GTDB taxonomy (*[Eubacterium] rectale ATCC 33656* with NCBI). It was followed by a group of 495 clusters of 97.5% sequence identity, consisting of various *Collinsella*-related species names, and a HumGut_95 cluster comprised of 400 different HumGut_97.5 clusters, all GTDB-named as *UBA11524 sp000437595*, and NCBI-named as *Faecalibacterium* sp. *CAG:74*.

Regarding taxonomy, many genomes were not given species names by GTDB, rather they were named after the genus, family, order, or class they belong to. Similarly, the NCBI taxonomic annotations for many genomes resulted in ambiguous names not specific to species, such as for example *uncultured bacterium* or *Firmicutes bacterium*. This contributed greatly in a discrepancy between the total number of species-level clusters (5,170 clusters in HumGut_95) and the total number of distinct cluster names (3,310 GTDB names, 1,716 NCBI names).

There were also many species-level clusters that shared the same species name. This was especially the case with various *Collinsella* clusters, where 81 different GTDB *Collinsella* species gave name to 7 different clusters each, on average. Comparably, 19 NCBI *Collinsella* species were seen in 44 different clusters on average.

Classifying the metagenome reads

We used the HumGut collection at both resolutions, in addition to the UHGG (species-level collection, containing 4,644 genomes) and the standard Kraken2 database, to classify the metagenomic reads from the 3,534 downloaded samples. On average, there were 56% unclassified reads when using the standard Kraken2 database, while the average dropped substantially when any one of the HumGut or the UHGG collection was utilized (UHGG = 14.1%, Humgut_95 = 11.7%, and HumGut_97.5 = 5.4%, Fig. 4a).

In comparison to the UHGG, both HumGut collections performed better. HumGut_95, a collection of species-level representatives—much like the UHGG collection—classified on average 2.3% more reads than the latter. With HumGut_97.5 as a custom database, this increased by 8.7%, marking a significant increase in recognized reads, with an obvious potential for improved classification accuracy.

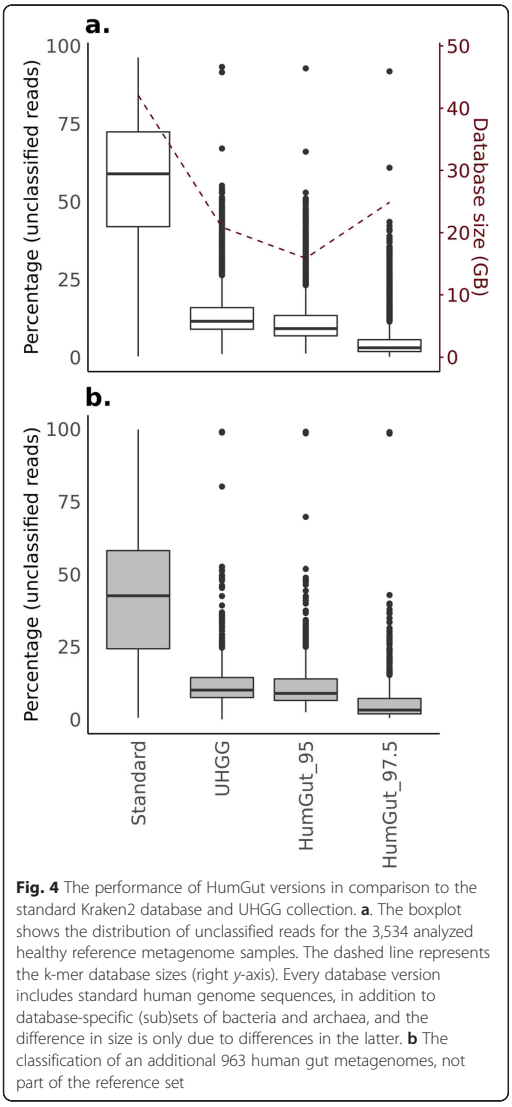


Fig. 4 The performance of HumGut versions in comparison to the standard Kraken2 database and UHGG collection. **a**. The boxplot shows the distribution of unclassified reads for the 3,534 analyzed healthy reference metagenome samples. The dashed line represents the k-mer database sizes (right *y*-axis). Every database version includes standard human genome sequences, in addition to database-specific (sub)sets of bacteria and archaea, and the difference in size is only due to differences in the latter. **b** The classification of an additional 963 human gut metagenomes, not part of the reference set

Both HumGut k-mer databases were smaller than the standard Kraken2 database of k-mers, necessitating reduced computer memory to perform the analyses. The lowest memory was required by the HumGut_95 database (Standard = 42.1 GB, UHGG = 20.9, HumGut_95 = 15.9 GB, HumGut_97.5 = 24.9).

Analysis of an additional 963 gut metagenome samples (collected from people suffering from IBD), not part of

the reference set, showed similar results regarding the number of classified reads: 42.3% unclassified reads on average when the Standard database was used, dropping to 12.5%—UHGG, 11.8%—HumGut_95, and 6.2% with HumGut_97.5 usage (Fig. 4b).

In comparison to UHGG, > 92% of samples from both datasets individually (healthy and IBD), had a higher number of classified reads with HumGut_95.

In addition to classification with Kraken2, we mapped the reads of 72 random healthy samples (one sample from each BioProject) using Bowtie2. We wanted to have an approximation of how well the results from a full-length-alignment approach corresponded to those of a k-mer-based algorithm. For this example, we only built UHGG and HumGut_95 indexes. On average, 20.5% of the reads were left unmapped with UHGG, and 17.1% with HumGut_95 (Supplementary material, Figure S2). That is an increase of 8.3% and 7.5% for UHGG and HumGut_95 correspondingly, compared to the results retrieved with Kraken2 for the same samples.

Taxa abundances

We used the KrakenUniq as a means of identifying false positive classifications, and removing them from the Kraken2 reports. We then used the Bracken software on the modified Kraken2 results, to re-estimate species abundance in the classified human gut metagenomes. These tasks were performed using HumGut_97.5 and GTDB taxonomy.

The results showed that, on average, healthy adults contained 202 species, people diagnosed with IBD, 145, and infants, 79 species. The overall species number distribution is presented in Fig. 5a.

In total, 52 species were found present in > 70% of healthy adult samples, led by *Agathobacter rectalis*, *Blautia_A sp900066165*, *Bacteroides uniformis*, *KLE1615 sp900066985*, *Agathobaculum butyriciproducens*, and *Fusicatenibacter saccharivorans*, discovered in > 90% of healthy adult samples, representing a core community of healthy adult human gut microbiota (Fig. 5b). A complete hierarchical linkage of samples, computed based on the abundance of these top 52 prevalent species, showed that African and South American (coming exclusively from Peru) metagenomes were more distant from the rest, while two species were not encountered at all in South American samples (*Alistipes onderdonkii* and *Lawsonibacter assacharolyticus*). In addition, these samples clustered more distantly from the others on a PCA plot (built based on the readcounts from all species), as depicted on Fig. 5c. The PCA loadings show that *Prevotella* species were more abundant in South American and African samples. In contrast, the *Alistipes* and *Bacteroides* species and lay on the opposite side of the plot, indicating a negative correlation to the former.

Infant samples separated from the adult samples as well. They are represented with crosses instead of dots on the PCA ordination plot, positioned on the leftmost part of the graph along PC1 axis. The loading plot shows that *Escherichia coli* species exercise the largest effect on samples positioned there. The most prevalent bacterium in infants was *Bifidobacterium longum* (68%), followed by *E. coli* (64%).

Bacteroides vulgatus, which, after screening the metagenomes using the MASH screen software, was the species of the top scoring genome, was no longer the most prevalent species among healthy human guts when classifying with all HumGut genomes. This was due to a lower diversity among *B. vulgatus* genomes, compared to *Agathobacter rectalis*. The genomes belonging to the former grouped into 2 species-level clusters (D = 0.05), while the latter resulted in 16 such groups. It is worth noting that we found the top *B. vulgatus* genome present in 2,536 healthy samples using MASH screen, and we found this species present in 2,537 healthy samples using Kraken2-KrakenUniq-Bracken classification tools. These almost identical results, obtained by two different sets of tools, increase confidence in the trustworthiness of these findings.

We also investigated the prevalence of species that only had RefSeq as a genome source in our collection. *Streptococcus sanguinis* was found present in 73% of all samples (healthy infants and adults, and IBD), followed by *Flavonifractor sp002161085*, *Escherichia sp005843885*, *Streptococcus sp001587175*, *Pauljensenia sp000466265*, *Flavonifractor sp002161215*, *Actinomyces naeslundii*, *Raoultella terrigena*, and *Mediterraneibacter sp900120155* (found in 10-36% of samples).

Discussion

The HumGut collection contains 30,691 genomes (HumGut_97.5), with a subset of 5,170 genomes clustered at 95% sequence identity (HumGut_95). The criterion for including a genome in HumGut was its prevalence in healthy human gut metagenomes.

Both HumGut versions showed superior performance in terms of assigned reads compared to the standard Kraken2 database, while demanding far less computational resources, as presented in Fig. 4. In addition, the species-level HumGut mapped more reads than UHGG when Bowtie2 was tested in a small subset of healthy samples. We consider this to be a strong argument in favor of HumGut's comprehensiveness and utility. Classifying a record-high proportion of classified reads per sample, HumGut aids the accuracy of taxonomic classification, which in turn facilitates a next-generation exploration of the human gut microbiome.

The vast majority of UHGG genomes qualified for inclusion in HumGut, as shown in Fig. 3a. However, in

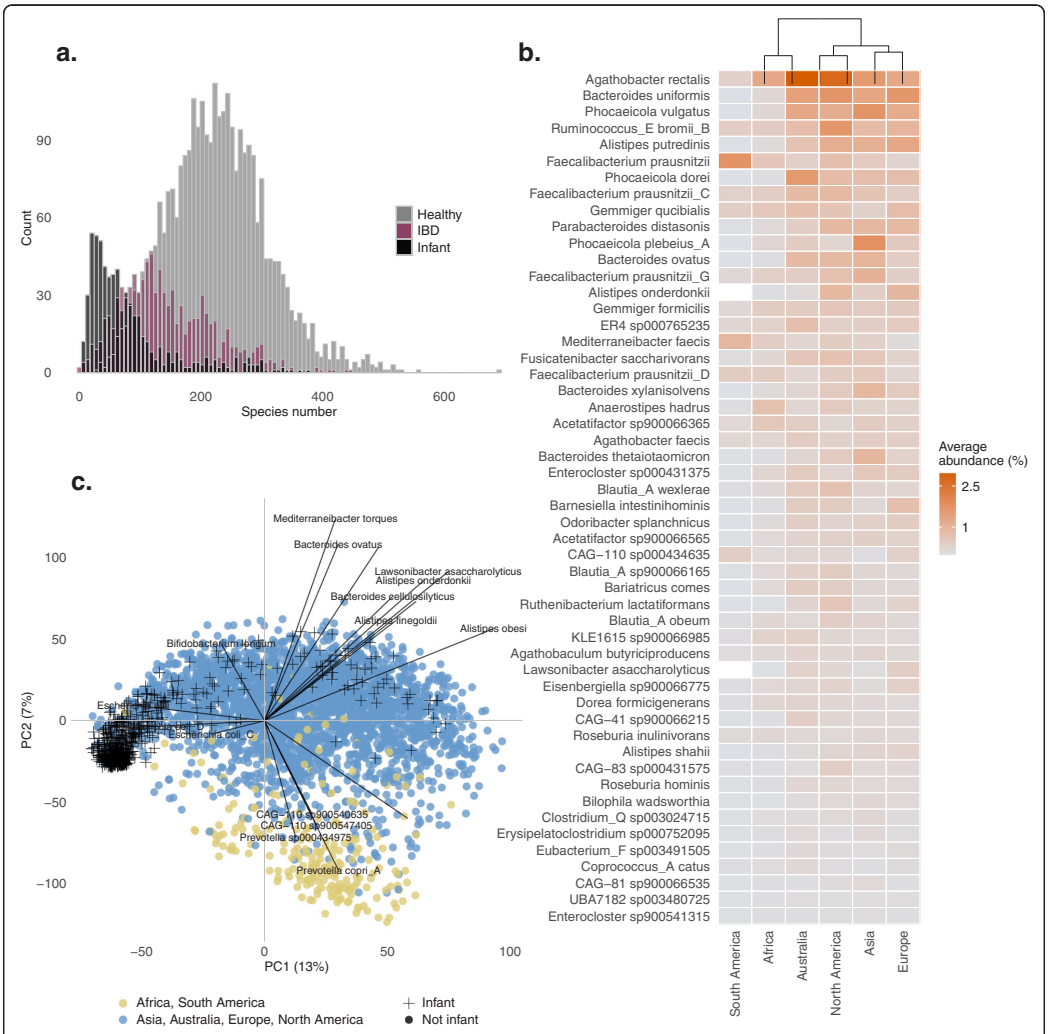


Fig. 5 Taxa analysis. **a** Species number distribution between three cohorts: healthy infants, healthy adults, and IBD samples. **b** Heatplot listing 52 bacterial species found in > 70% healthy adult samples. The tiles are colored based on the average abundance of species in samples from the same continent (*Y*-axis). On top of the graph, a dendrogram shows the complete linkage between samples from separate continents. **c** PCA score plot and loading plot. Each dot on the PCA ordination plot represents a metagenome sample. They are colored based on continent of origin. African and South American samples separate more from samples coming from Asia, Australia, Europe, and North America. The superimposed text represents some of the species (loadings) with the strongest effect along the PC2 axis

comparison to the UHGG collection, HumGut holds the advantage of containing more relevant human gut prokaryotic genomes in its pool, reflected by the additional RefSeq genomes that showed no sequence similarity with the qualified UHGG genomes, forming separate clusters of 95% sequence identity (Fig. 3b). An example of its utility is the discovery of *Streptococcus sanguinis* in > 70% of all metagenome samples, which would otherwise be impossible using the UHGG collection as a custom Kraken2 database. Also, the identification of one of the most prevalent species in human guts, *Bacteroides vulgatus*, would have been mistaken for *Bacteroides dorei*.

HumGut sets were built after ranking the genomes based on their prevalence among metagenomes and

using the top-ranked ones as cluster representatives. This has ensured that the collections only contain genomes highly relevant to healthy human guts worldwide. Comparing the HumGut_95 collection to the UHGG collection (same resolution) shows that more metagenomic reads are classified for the former. Additionally, its set of unique k-mers is 24% smaller in size than the UHGG. This indicates the UHGG contains a higher genomic diversity, requiring memory which is not really needed for successful read classification. These are rare genomes found in the occasional human gut metagenome, but with low prevalence.

HumGut can serve as a global reference for bacteria inhabiting the gut of healthy humans, highlighting its importance for future gut microbiome studies and is available for download (http://arken.nmbu.no/~larssn/humgut/).

Our analysis showed that the diversity of gut samples across the world is not profoundly affected by geography (Fig. 2); therefore, having a global genome collection like HumGut is reasonable.

However, we acknowledge that such a finding may be confounded by the shared similarity of lifestyle choices across people whose metagenomes were analyzed here.

We found 50 bacterial species present in more than 70% of the samples, regardless of the country of origin. This group of species, led by *Agathobacter rectalis*, represents what we think is the core human gut bacterial community (Fig. 5b). We must, however, cautiously refer to *A. rectalis* as the most prevalent/abundant species found in human gut samples. That because we found this species to be highly diverse in sequence identity. In our collection, we have 16 different species-level clusters, and more than 530 clusters of 97.5% sequence identity with this name.

We discovered that, on average, healthy adults contain around 60 bacterial species more than IBD subjects, and around 120 species more than healthy infants (Fig. 5a). A low microbiome complexity among the latter two groups is well documented in literature [18–22].

Although we found a great homogeneity of top prevalent species among healthy adults, our analysis showed that samples originating from Africa and South America were rich in *Prevotella* and poor in *Bacteroides*, which made them cluster in our principal component analysis, as depicted in Fig. 5c. A *Prevotella-Bacteroides* antagonism and their correlation to lifestyle and diet have long been described in literature [23, 24]. Our results are, therefore, consistent with these findings.

We have demonstrated that HumGut is useful in research that goes beyond studying healthy subjects, exemplified by the equally high number of classified metagenomic reads collected from IBD subjects.

A challenge that remains is the nomenclature of species in our genome collection. There is a profound inconsistency between the total number of species-level clusters and the total number of names annotating them (a ratio of 1.5:1 with GDTB-based annotation, and 3:1 with NCBI names). We believe that as long as not all names reflect species individuality, it will be difficult to truly explore the composition differences between various cohorts, in addition to posing a challenge in studies linking functions to species. On our website, we have prepared files for building a custom Kraken2 database where all HumGut clusters also have been given artificial "taxonomy IDs," making it possible to classify to clusters instead of taxa. We note that the decision regarding which version the HumGut collection to employ depends on users' computational resources as well as the level of taxonomic resolution required.

On another note, it is important to emphasize that the microbiome composition results presented here are all based on k-mer-based methods. It remains to be seen how well these results compare to those from whole-read-based alignment methods.

As future work, we will also extend our approach to more disease-associated genomes and metagenomes, in addition to screening for gut genomes that will expectedly be published in the future.

Conclusion

We believe that by using HumGut as a reference, the scientific community will be one step closer to method standardization sorely needed in the field of human gut microbiome analysis, and that the discovery of potential microbiome markers will be facilitated with higher certainty in less time and computational resources.

Methods

Human gut reference metagenomes

A set of publicly available human gut metagenome samples was collected and used for ranking all genomes in the search for human gut relevant ones. A text search for all human gut microbiome samples at the Sequence Read Archive (NCBI/SRA, https://www.ncbi.nlm.nih.gov/sra) was performed. The list of hits was manually curated, keeping only samples with > 1,000,000 reads and annotated as healthy individuals. NCBI/BioProject accessions of these projects were used to locate the same data in the European Nucleotide Archive (EMBL-EBI/ENA, https://www.ebi.ac.uk/ena), from which all samples were downloaded as compressed fastq-files, using the Aspera download system (https://www.ibm.com/products/aspera). This resulted in 5,737 healthy metagenomes (samples) covering 74 BioProjects. For many BioProjects, some samples tended to be very similar to each other, presumably due to samples collected from

individuals sharing the same lifestyle, geographical sub-population, genetics, or other factors that may affect the human gut microbiome. To avoid too much bias in the direction of such heavily sampled sub-populations, samples from the same BioProject were clustered. From each metagenome sample, a MinHash sketch of 10,000 k-mers was computed using the MASH software [25], discarding singleton k-mers (21-mers). Based on these sketches the MASH distances between all pairs of samples were calculated. A MASH distance close to zero means two samples are very similar, sharing most of their k-mers. Next, hierarchical clustering with complete linkage was computed, and samples were partitioned at a 0.05 distance threshold, resulting in clusters with "diameters" no larger than this chosen threshold. The medoid sample from each cluster, i.e., the one with the minimum sum of distances to all members of the cluster, was retained as the reference sample representing its cluster. This resulted in the set of 3,534 healthy metagenome samples. Below, we refer to this collection as *MetHealthy*.

The same procedure was utilized to collect another set of metagenomes from patients diagnosed with Inflammatory Bowel Disease (ulcerative colitis, or Crohn's disease). From initially 2,064 metagenomes, the clustering resulted in a collection of 963 metagenomes covering 14 BioProjects. This is the *MetIBD* collection. Finally, a set of 95 samples containing gut metagenome data from primates was collected and used as an outgroup in a comparison of the human gut metagenomes. The metagenomes' metadata is included in the Supplementary Table 1.

Genome collections

The main source was the recently published Unified Human Gut Genomes (UHGG) collection, containing 286,997 genomes exclusively related to human guts: http://ftp.ebi.ac.uk/pub/databases/metagenomics/mgnify_genomes/human-gut/v1.0/all_genomes/. The other source was NCBI/Genome, the RefSeq repository at ftp://ftp.ncbi.nlm.nih.gov/genomes/refseq/bacteria/ and ftp://ftp.ncbi.nlm.nih.gov/genomes/refseq/archaea/. At the time of writing, ~204,000 genomes were downloaded from this site.

Metadata about the genomes considered and qualified for HumGut are presented in Supplementary Table 2.

Genome ranking

Only metagenomes collected from healthy individuals, *MetHealthy*, were used in this step. For all genomes, the MASH software was again used to compute sketches of 1,000 k-mers, including singletons [26]. The MASH screen compares the sketched genome hashes to all hashes of a metagenome, and, based on the shared number of them, estimates the genome sequence identity I to the metagenome. Given that $I = 0.95$ (95% identity) is regarded as a species delineation for whole-genome comparisons [17], it was used as a soft threshold to determine if a genome was contained in a metagenome. Genomes meeting this threshold for at least one of the *MetHealthy* metagenomes were qualified for further processing. Then the average I value across all *MetHealthy* metagenomes was computed for each genome, and this prevalence-score was used to rank them. The genome with the highest prevalence-score was considered the most prevalent among the *MetHealthy* samples, and thereby the best candidate to be found in any healthy human gut. This resulted in a list of genomes ranked by their prevalence in healthy human guts.

Genome clustering

Many ranked genomes were very similar, some even identical. Due to errors introduced in sequencing and genome assembly, it made sense to group genomes and use one member from each group as a representative genome. Even without any technical errors, a lower meaningful resolution in terms of whole genome differences was expected, i.e., genomes differing in only a small fraction of their bases should be considered identical.

The clustering of the genomes was performed in two steps, like the procedure used in the dRep software [27], but in a greedy way based on the ranking of the genomes. The huge number of genomes (hundreds of thousands) made it extremely computationally expensive to compute all-versus-all distances. The greedy algorithm starts by using the top ranked genome as a cluster centroid, and then assigns all other genomes to the same cluster if they are within a chosen distance D from this centroid. Next, these clustered genomes are removed from the list, and the procedure is repeated, always using the top ranked genome as centroid.

The whole-genome distance between the centroid and all other genomes was computed by the fastANI software [17]. However, despite its name, these computations are slow in comparison to the ones obtained by the MASH software. The latter is, however, less accurate, especially for fragmented genomes. Thus, we used MASH-distances to make a first filtering of genomes for each centroid, only computing fastANI distances for those who were close enough to have a reasonable chance of belonging to the same cluster. For a given fastANI distance threshold D, we first used a MASH distance threshold $D_{mash} >> D$ to reduce the search space. In supplementary material, Figure S3, we show some results guiding the choice of D_{mash} for a given D.

A distance threshold of $D = 0.05$ is regarded as a rough estimate of a species, i.e., all genomes within a

species are within this fastANI distance from each other [16, 17]. This threshold was also used to arrive at the 4,644 genomes extracted from the UHGG collection and presented at the MGnify website. However, given shotgun data, a larger resolution should be possible, at least for some taxa. For this reason, we started out with a threshold D = 0.025, i.e., half the "species radius." An even higher resolution was tested (D = 0.01), but the computational burden increases vastly as we approach 100% identity between genomes. It is also our experience that genomes more than ~98% identical are very difficult to separate, given today's sequencing technologies [28]. However, the genomes found at D = 0.025 (HumGut_97.5) were also again clustered at D = 0.05 (HumGut_95) giving two resolutions of the genome collection.

The taxonomic annotation of the genomes was performed with GTDB toolkit (GTDB-Tk, release 05-RS95, https://github.com/Ecogenomics/GTDBTk) [16], and in the genome metadata tables we provide on our website, we made efforts to also list the corresponding NCBI Taxonomy names and ID's for all genomes.

All UHGG genomes were already checked for completeness and contamination [7]. The completeness and contamination of RefSeq genomes was performed using CheckM (https://ecogenomics.github.io/CheckM/) [29]. The handful genomes not having > 50% completeness and < 5% contamination were discarded. All qualified genomes had a genome quality score ≥ 50 (completeness − 5×contamination).

Metagenome classifications

The Kraken2 software was used for classifying reads from the metagenome samples [30]. To see the effects of using a different database, the standard Kraken2-database was compared by custom databases built from the 4,644 UHGG genomes at the MGnify website as well as our HumGut collections. In all custom databases, the standard Kraken2 library for the human genome was also included, since host contamination is quite normal in metagenome data. All classifications were performed using default settings in Kraken2.

Since Kraken2, like most other software for taxonomic classification, uses the Lowest Common Ancestor (LCA) approach, many reads are assigned to ranks high up in the taxonomy. The Bracken software [31] has been designed to re-estimate the abundances at some fixed rank, by distributing reads from higher ranks into the lower rank, based on conditional probabilities estimated from the database content. A Bracken database (100-mers) was created for HumGut_97.5 database and used to re-estimate all abundances at the species rank.

If counting all listed taxa, regardless of low readcounts, the Kraken2 is known to produce many false positives [32], i.e., list taxa as present when they are in fact not. The KrakenUniq software has been developed to handle this problem [32]. We ran it to classify the metagenome reads for both healthy and IBD metagenomes. The overall results from both Kraken2 and KrakenUniq tools were similar, but KrakenUniq also reports the number of unique k-mers in each genome covered by the reads. On the other hand, only Kraken2 reports are compatible for running the Bracken software. Since we were interested in both—that is finding the true positive identifications, and their estimated abundances—we combined the two approaches. For each sample, we found from the KrakenUniq report a k-mer count threshold, following the authors recommendations (2,000 unique k-mers per 1,000,000 sequencing reads depth) [32]. Taxa falling below this threshold were given zero read counts in the corresponding modified Kraken2 reports. We then ran Bracken on these modified Kraken2 reports.

Additionally, we tested 72 random healthy samples, each belonging to a distinct BioProject, using Bowtie2—a full-length sequence aligner (HumGut_95 and UHGG reference databases only) [33].

A principal component analysis was conducted on the matrix of species readcounts for all metagenome samples, after the following transformation: a pseudo-count of 10 was added to all species before using Aitchison's centered log-ratio transform [34, 35] to remove the unit-sum constraint otherwise affecting a PCA of such data.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s40168-021-01114-w.

Additional file 1. Figure S1. Rarefaction curves for healthy (left panel) and IBD metagenomes (right panel), showing that the number of new expected genomes flattens after screening ca. 1,000 metagenomes.

Additional file 2. Figure S2. Mapping of 72 samples using Bowtie2. Y-axis shows the percentage of unmapped reads when any of the two reference index databases was used (UHGG, or HumGut_95).

Additional file 3 Figure S3. MASH and fastANI distances. **a.** A plot of ca. 20,000 genome distances computed with both fastANI (x-axis) and MASH (y-axis). fastANI distances tend to be a little smaller than MASH distances, they however have a substantial variance. **b.** The rationale behind using 0.08, and 0.1 MASH distance thresholds (vertical dashed lines) for HumGut clustering algorithm. The vast majority of fastANI distances < 0.025 have a MASH distance < 0.08 and genomes with fastANI < 0.05 have a MASH distance < 0.1. When clustering, the distance between all genomes was first computed using MASH, then only genomes with distances below the abovementioned thresholds were included to speed up fastANI computations.

Additional file 4. Table S1. Metagenenomes metadata. Table S2. Genomes metadata.

Acknowledgements

Bioinformatics/data analysis was performed using resources at the Orion HPC at NMBU.

Authors' contributions

LS conceived the study. LS and PH worked out the technical aspects of the paper. All authors discussed and interpreted the results. PH wrote the article

with equal input from all authors. The authors read and approved the final manuscript.

Funding
This work was financially supported by Norway Research Council through R&D project grant no. 283783, 248792, and 301364.

Availability of data and materials
The HumGut genome collection and all associated metadata can be found at http://arken.nmbu.no/~larssn/humgut/. This also includes files and recipes for building Kraken2 databases using these genomes.

Declarations

Ethics approval and consent to participate
Not applicable

Consent for publication
Not applicable

Competing interests
Both PH and FTH are employed at Genetic Analysis AS, but all authors agree this fact does not represent a conflict of interest in the context of our manuscript.

Author details
[1]Department of Chemistry, Biotechnology and Food Sciences, Norwegian University of Life Sciences, P.O. Box 5003, 1432 Ås, Norway. [2]Department of Biotechnology, Inland Norway University of Applied Sciences, 2318 Hamar, Norway. [3]Genetic Analysis AS, Kabelgaten 8, 0580 Oslo, Norway.

Received: 22 February 2021 Accepted: 18 June 2021
Published online: 31 July 2021

References

1. Methé BA, et al. A framework for human microbiome research. Nature. 2012;486(7402):215–21.
2. Forster SC, Kumar N, Anonye BO, Almeida A, Viciani E, Stares MD, et al. A human gut bacterial genome and culture collection for improved metagenomic analyses. Nat Biotechnol. 2019;37(2):186–92. https://doi.org/10.1038/s41587-018-0009-7.
3. Almeida A, Mitchell AL, Boland M, Forster SC, Gloor GB, Tarkowska A, et al. A new genomic blueprint of the human gut microbiota. Nature. 2019;568(7753):499–504. https://doi.org/10.1038/s41586-019-0965-1.
4. Zou Y, Xue W, Luo G, Deng Z, Qin P, Guo R, et al. 1,520 reference genomes from cultivated human gut bacteria enable functional microbiome analyses. Nat Biotechnol. 2019;37(2):179–85. https://doi.org/10.1038/s41587-018-0008-8.
5. Nayfach S, Shi ZJ, Seshadri R, Pollard KS, Kyrpides NC. New insights from uncultivated genomes of the global human gut microbiome. Nature. 2019;568(7753):505–10. https://doi.org/10.1038/s41586-019-1058-x.
6. Pasolli E, et al. Extensive unexplored human microbiome diversity revealed by over 150,000 genomes from metagenomes spanning age, geography, and lifestyle. Cell. 2019;176(3):649–662.e20.
7. Almeida A, et al. A unified catalog of 204,938 reference genomes from the human gut microbiome. Nat Biotechnol. 2020.
8. Rothschild D, Weissbrod O, Barkan E, Kurilshikov A, Korem T, Zeevi D, et al. Environment dominates over host genetics in shaping human gut microbiota. Nature. 2018;555(7695):210–5. https://doi.org/10.1038/nature25973.
9. Lozupone CA, Stombaugh JI, Gordon JI, Jansson JK, Knight R. Diversity, stability and resilience of the human gut microbiota. Nature. 2012;489(7415):220–30. https://doi.org/10.1038/nature11550.
10. Shin N-R, Whon TW, Bae J-W. Proteobacteria: microbial signature of dysbiosis in gut microbiota. Trends Biotechnol. 2015;33(9):496–503. https://doi.org/10.1016/j.tibtech.2015.06.011.
11. Halfvarson J, Brislawn CJ, Lamendella R, Vázquez-Baeza Y, Walters WA, Bramer LM, et al. Dynamics of the human gut microbiome in inflammatory bowel disease. Nat Microbiol. 2017;2(5):17004. https://doi.org/10.1038/nmicrobiol.2017.4.
12. Rajilić-Stojanović M, et al. Global and deep molecular analysis of microbiota signatures in fecal samples from patients with irritable bowel syndrome. Gastroenterology. 2011;141(5):1792–801.
13. Cotillard A, Kennedy SP, Kong LC, Prifti E, Pons N, le Chatelier E, et al. Dietary intervention impact on gut microbial gene richness. Nature. 2013;500(7464):585–8. https://doi.org/10.1038/nature12480.
14. Wallace TC, Guarner F, Madsen K, Cabana MD, Gibson G, Hentges E, et al. Human gut microbiota and its relationship to health and disease. Nutr Rev. 2011;69(7):392–403. https://doi.org/10.1111/j.1753-4887.2011.00402.x.
15. McBurney MI, Davis C, Fraser CM, Schneeman BO, Huttenhower C, Verbeke K, et al. Establishing what constitutes a healthy human gut microbiome: state of the science, regulatory considerations, and future directions. J Nutr. 2019;149(11):1882–95. https://doi.org/10.1093/jn/nxz154.
16. Chaumeil P-A, et al. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. Bioinformatics. 2019;36(6):1925–7.
17. Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT, Aluru S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. Nat Commun. 2018;9(1):5114. https://doi.org/10.1038/s41467-018-07641-9.
18. Manichanh C, et al. Reduced diversity of faecal microbiota in Crohn's disease revealed by a metagenomic approach. Gut. 2006;55(2):205–11.
19. Manichanh C, Borruel N, Casellas F, Guarner F. The gut microbiota in IBD. Nat Rev Gastroenterol Hepatol. 2012;9(10):599–608. https://doi.org/10.1038/nrgastro.2012.152.
20. Matsuoka K, Kanai T. The gut microbiota and inflammatory bowel disease. Semin Immunopathol. 2015;37(1):47–55. https://doi.org/10.1007/s00281-014-0454-4.
21. Rodrĺguez JM, et al. The composition of the gut microbiota throughout life, with an emphasis on early life. Microb Ecol Health Dis. 2015;26(1):26050.
22. Moore RE, Townsend SD. Temporal development of the infant gut microbiome. Open Biol. 2019;9(9):190128.
23. Gorvitovskaia A, Holmes SP, Huse SM. Interpreting Prevotella and Bacteroides as biomarkers of diet and lifestyle. Microbiome. 2016;4(1):15. https://doi.org/10.1186/s40168-016-0160-7.
24. Hjorth MF, Blædel T, Bendtsen LQ, Lorenzen JK, Holm JB, Kiilerich P, et al. Prevotella-to-Bacteroides ratio predicts body weight and fat loss success on 24-week diets varying in macronutrient composition and dietary fiber: results from a post-hoc analysis. Int J Obes. 2019;43(1):149–57. https://doi.org/10.1038/s41366-018-0093-2.
25. Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, et al. Mash: fast genome and metagenome distance estimation using MinHash. Genome Biol. 2016;17(1):132. https://doi.org/10.1186/s13059-016-0997-x.
26. Ondov BD, et al. Mash Screen: high-throughput sequence containment estimation for genome discovery. BioRxiv. 2019:557314.
27. Olm MR, Brown CT, Brooks B, Banfield JF. dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. ISME J. 2017;11(12):2864–8. https://doi.org/10.1038/ismej.2017.126.
28. Snipen L, et al. Reduced metagenome sequencing for strain-resolution taxonomic profiles. Microbiome. 2021.
29. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res. 2015;25(7):1043–55. https://doi.org/10.1101/gr.186072.114.
30. Wood DE, Lu J, Langmead B. Improved metagenomic analysis with Kraken 2. Genome Biol. 2019;20(1):257. https://doi.org/10.1186/s13059-019-1891-0.
31. Lu J, Breitwieser FP, Thielen P, Salzberg SL. Bracken: estimating species abundance in metagenomics data. PeerJ Comp Sci. 2017;3:e104. https://doi.org/10.7717/peerj-cs.104.
32. Breitwieser FP, Baker DN, Salzberg SL. KrakenUniq: confident and fast metagenomics classification using unique k-mer counts. Genome Biol. 2018;19(1):198. https://doi.org/10.1186/s13059-018-1568-0.
33. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9(4):357–9. https://doi.org/10.1038/nmeth.1923.
34. Weiss S, Xu ZZ, Peddada S, Amir A, Bittinger K, Gonzalez A, et al. Normalization and microbial differential abundance strategies depend upon data characteristics. Microbiome. 2017;5(1):27. https://doi.org/10.1186/s40168-017-0237-y.
35. Aitchison J. The statistical analysis of compositional data. J R Stat Soc Ser B Methodol. 1982;44(2):139–60.

Publisher's Note

Paper III

frontiers in Microbiology

OPINION
published: 04 February 2022
doi: 10.3389/fmicb.2021.822301

Questioning the Quality of 16S rRNA Gene Sequences Derived From Human Gut Metagenome-Assembled Genomes

Pranvera Hiseni[1,2]*, *Lars Snipen*[1], *Robert C. Wilson*[3], *Kari Furu*[2] *and Knut Rudi*[1,3]

[1] Department of Chemistry, Biotechnology and Food Sciences, Norwegian University of Life Sciences, Ås, Norway, [2] Genetic Analysis AS, Oslo, Norway, [3] Department of Biotechnology, Faculty of Applied Ecology, Agricultural Sciences and Biotechnology, Inland Norway University of Applied Sciences, Hamar, Norway

Keywords: 16S rRNA, metagenome assembled genome (MAG), metagenome analyses, human gut microbiome, prokaryotic genome

OPEN ACCESS

Edited and reviewed by:
Franck Carbonero, Washington State University Health Sciences Spokane, United States

***Correspondence:**
Pranvera Hiseni
ph@genetic-analysis.com

Specialty section:
This article was submitted to Microorganisms in Vertebrate Digestive Systems, a section of the journal Frontiers in Microbiology

Received: *25 November 2021*
Accepted: *28 December 2021*
Published: *04 February 2022*

Citation:
Hiseni P, Snipen L, Wilson RC, Furu K and Rudi K (2022) Questioning the Quality of 16S rRNA Gene Sequences Derived From Human Gut Metagenome-Assembled Genomes. Front. Microbiol. 12:822301. doi: 10.3389/fmicb.2021.822301

The recent introduction of metagenome-assembled genomes (MAGs) has marked a major milestone in the human gut microbiome field (Almeida et al., 2019; Nayfach et al., 2019; Pasolli et al., 2019). Such reference-free, *de novo*-assembled genomes (Hugerth et al., 2015) have revealed a wide range of hitherto uncultured microbial species in human gut samples.

The significance of MAGs in unraveling human gut microbial diversity was supported by their overwhelming representation in a comprehensive human gut prokaryotic collection filtered by metagenome data dereplicated at 97.5% average nucleotide identity (ANI) (Hiseni et al., 2021). More than 90% of the collection consists of MAGs, while the rest of the collection mainly comprises RefSeq genomes (**Figure 1A**).

A great challenge related to MAGs is their lack of 16S rRNA sequences. Skewed species abundance, high 16S sequence similarity, and high volumes of short-reads data cause major difficulties for assembling the sequences of this gene (Yuan et al., 2015), frequently rendering these genomes incomplete.

A barrnap search (https://github.com/tseemann/barrnap) revealed that from >270,000 qualified MAGs, only 7% yielded 16S sequences, while this gene was found in 93% of >106,000 other genome types. MAGs positive for 16S had a significantly lower copy number compared to complete RefSeq genomes (**Figure 1B**; top panel) and substantially higher intragenomic variance (**Figure 1B**; bottom panel). Challenges in obtaining multiple 16S copies from incomplete genomes are well-described in the literature (Perisin et al., 2016; Louca et al., 2018); however, to exacerbate the problem, their enormous intragenomic heterogeneity renders their overall quality questionable.

A multiple sequence alignment of 16S rDNA sequences extracted from members of identical 97.5% ANI clusters, followed by the computation of their distance [*ape* package in RStudio (Paradis and Schliep, 2018)], has revealed that clusters consisting purely of MAGs share on average 93% identity, as contrasted by 99.8% average 16S sequence identity in clusters made of pure, complete RefSeq genomes (**Figure 1C**).

Considering that 16S is a highly conserved gene, its identity among same-cluster genomes was expected to be higher than the threshold used for dereplicating them (>97.5%; Kim et al., 2014; Jain et al., 2018). The excessive 16S divergence among MAG-only clusters raises red flags, potentially reflecting issues related to their assembly, as previously reported (Nelson et al., 2020; Meziti et al., 2021).

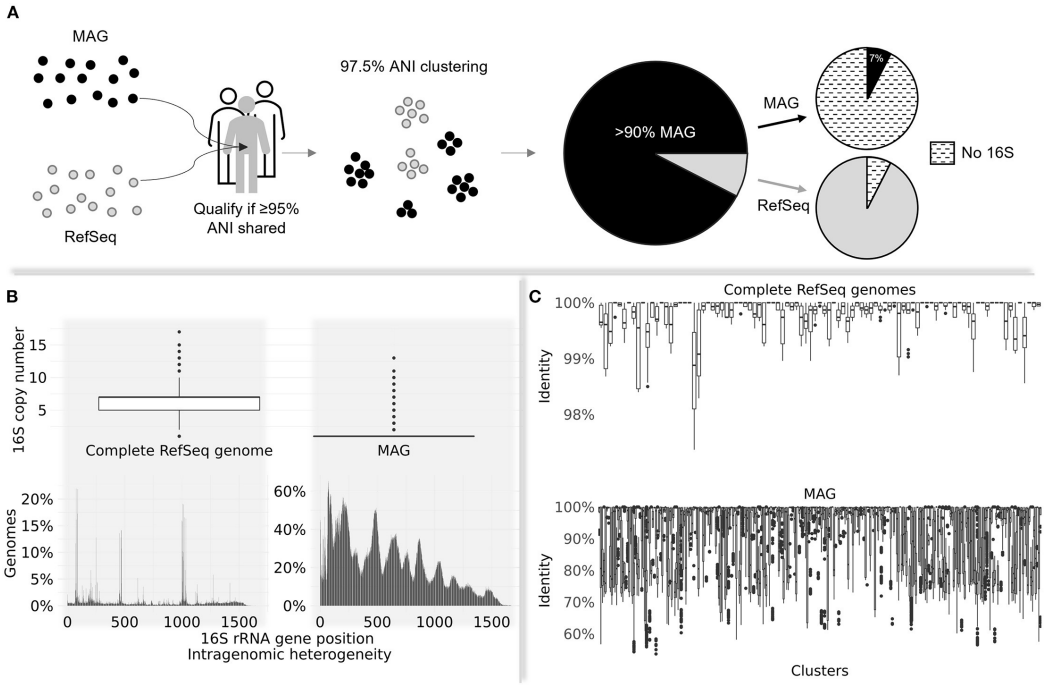


FIGURE 1 | (A) The process of filtering human gut-derived MAGs and RefSeq prokaryotic genomes against a pool of >3,500 non-redundant healthy human gut metagenomes. Only genomes sharing ≥95% average nucleotide identity (ANI)—a conventional threshold marking species delineation (Jain et al., 2018)—were kept for further processing. The qualified genomes dereplicated at 97.5% ANI were mostly represented by MAGs (>90%). Only 7% of MAGs harbored detectable 16S rRNA gene sequences, while the opposite was observed in RefSeq genomes (7% lacked detectable 16S). **(B)** The distribution of 16S copy numbers on complete RefSeq genomes vs. MAGs (upper panel); the intragenomic 16S rRNA gene heterogeneity on genomes with multiple 16S copies for the same groups (bottom panel). MAGs are associated with increased intragenomic variability across all positions compared to RefSeq genomes. **(C)** The average nucleotide identity of 16S sequences belonging to the same 97.5% ANI cluster. Each boxplot refers to one cluster. The upper panel depicts clusters made of pure complete RefSeq genomes, while the bottom panel shows the distribution of shared identities on clusters entirely comprising MAGs. RefSeq-derived 16S sequences within same clusters show high identity (average of 99.8%); MAG clusters contain highly variable 16S sequences, with an average identity of 93%.

All MAGs studied here were >95% complete with <5% contamination, a conventional criterion marking their high quality. Given the extreme importance of the 16S gene in microbial taxonomy and ecology, it seems unacceptable that MAGs can be labeled as such and at the same time contain low-quality information about this single most important gene that links the re-constructed genomes to the huge body of 16S-based microbiota studies conducted worldwide.

Furthermore, the acceptance of poor 16S rDNA quality in MAGs currently excludes a majority in the microbial research community that does not have the economic or computational resources to perform large-scale shotgun sequencing.

AUTHOR CONTRIBUTIONS

KR and PH conceived the idea. PH wrote the manuscript with an equal input from all authors. All authors discussed and interpreted the findings. All authors contributed to the article and approved the submitted version.

FUNDING

This work was financially supported by Norway Research Council, a Norwegian government agency funding research and innovation, through R&D project grant nos. 283783, 248792, and 301364.

REFERENCES

Almeida, A., Mitchell, A. L., Boland, M., Forster, S. C., Gloor, G. B., Tarkowska, A., et al. (2019). A new genomic blueprint of the human gut microbiota. *Nature* 568, 499–504. doi: 10.1038/s41586-019-0965-1

Hiseni, P., Rudi, K., Wilson, R. C., Hegge, F. T., and Snipen, L. (2021). HumGut: a comprehensive human gut prokaryotic genomes collection filtered by metagenome data. *Microbiome* 9:165. doi: 10.1186/s40168-021-01114-w

Hugerth, L. W., Larsson, J., Alneberg, J., Lindh, M. V., Legrand, C., Pinhassi, J., et al. (2015). Metagenome-assembled genomes uncover a global brackish microbiome. *Genome Biol.* 16,279. doi: 10.1186/s13059-015-0834-7

Jain, C., Rodriguez-R, L. M., Phillippy, A. M., Konstantinidis, K. T., and Aluru, S. (2018). High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat. Commun.* 9,5114. doi: 10.1038/s41467-018-07641-9

Kim, M., Oh, H.-S., Park, S.-C., and Chun, J. (2014). Towards a taxonomic coherence between average nucleotide identity and 16S rRNA gene sequence similarity for species demarcation of prokaryotes. *Int. J. Syst. Evol. Microbiol.* 64, 346–351. doi: 10.1099/ijs.0.059774-0

Louca, S., Doebeli, M., and Parfrey, L. W. (2018). Correcting for 16S rRNA gene copy numbers in microbiome surveys remains an unsolved problem. *Microbiome* 6,41. doi: 10.1186/s40168-018-0420-9

Meziti, A., Rodriguez,-R., L. M., Hatt, J. K., Peña-Gonzalez, A., Levy, K., et al. (2021). The reliability of metagenome-assembled genomes (MAGs) in representing natural populations: insights from comparing MAGs against isolate genomes derived from the same fecal sample. *Appl. Environ. Microbiol.* 87, e02593–e02520. doi: 10.1128/AEM.02593-20

Nayfach, S., Shi, Z. J., Seshadri, R., Pollard, K. S., and Kyrpides, N. C. (2019). New insights from uncultivated genomes of the global human gut microbiome. *Nature* 568, 505–510. doi: 10.1038/s41586-019-1058-x

Nelson, W. C., Tully, B. J., and Mobberley, J. M. (2020). Biases in genome reconstruction from metagenomic data. *PeerJ* 8,e10119. doi: 10.7717/peerj.10119

Paradis, E., and Schliep, K. (2018). ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* 35, 526–528. doi: 10.1093/bioinformatics/bty633

Pasolli, E., Asnicar, F., Manara, S., Zolfo, M., Karcher, N., Armanini, F., et al. (2019). Extensive unexplored human microbiome diversity revealed by over 150,000 genomes from metagenomes spanning age, geography, and lifestyle. *Cell.* 176, 649–662.e620. doi: 10.1016/j.cell.2019.01.001

Perisin, M., Vetter, M., Gilbert, J. A., and Bergelson, J. (2016). 16Stimator: statistical estimation of ribosomal gene copy numbers from draft genome assemblies. *ISME J.* 10, 1020–1024. doi: 10.1038/ismej.2015.161

Yuan, C., Lei, J., Cole, J., and Sun, Y. (2015). Reconstructing 16S rRNA genes in metagenomic data. *Bioinformatics* 31, i35–i43. doi: 10.1093/bioinformatics/btv231

Conflict of Interest: PH and KF were employed by company Genetic Analysis AS.

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

Paper IV

Prediction of high fecal propionate-to-butyrate ratios using 16S rRNA-based detection of bacterial groups with Liquid Array Diagnostics

- **Short running title:** Propionate-to-butyrate ratio prediction with LAD
- **Author names:** Pranvera Hiseni[1, 2], Lars Snipen[2], Robert C. Wilson[3], Finn Terje Hegge[1], Knut Rudi[2,3]

- **Author affiliations:**
[1]Genetic Analysis AS, Kabelgata 8, 0580 Oslo, Norway
[2]Department of Chemistry, Biotechnology and Food Sciences, Norwegian University of Life Sciences, P.O. Box 5003, 1432 Aas, Norway
[3]Inland Norway University of Applied Sciences, Hamar, Norway

Corresponding author details: Pranvera Hiseni, Department of Chemistry, Biotechnology and Food Sciences, Norwegian University of Life Sciences, P.O. Box 5003, 1432 Aas, Norway; Genetic Analysis AS, Kabelgata 8, 0580 Oslo, Norway; ph@genetic-analysis.com, pranvera.hiseni@nmbu.no; +47 48 644 835.

- **Financial disclosure:** PH is employed by Genetic Analysis AS, a licensee of the LAD method.

- **Acknowledgements:** The authors are grateful to the Norwegian Research Council for supporting this work through R&D grant no. 283783.

- **Information pertaining to writing assistance:** N/A

- **Ethical disclosure:** The collection and handling of fecal samples was performed in accordance with the Norwegian Regional Committee for Medical and Health Research Ethics (reference no. 2010/3209).

- **Author Contributions:** The idea was conceived by KR. PH planned and conducted the lab work. All authors discussed and helped interpreting the results. PH wrote the article with an equal input from all authors.

- **Data sharing statement:** PacBio and Illumina Whole-Genome Shotgun sequencing data are deposited at Sequence Read Archive (PRJNA820539).

Word count: 4378
Figure number: 6
Table number: 2

Abstract

Short-chain fatty acids (SCFAs) represent the main fermentation end-products of intestinal microbiome. They exercise various functions, affecting and maintaining the overall health status of humans. The ratio between butyrate and propionate (P:B), is particularly important. However, it remains a challenge to adopt SCFA detection techniques in clinical settings, due to the volatile nature of these acids.

Here, we aimed to estimate SCFA information indirectly, through a novel, simple qPCR-compatible assay (Liquid Array Diagnostics - LAD), targeting a limited number of microbiome markers.

Utilizing 15 LAD probes to target microbiome markers selected by a PLS + LDA model, the classes (normal vs. high P:B) were best separated at a threshold of 2.6, yielding a prediction accuracy of 96%.

Keywords: Propionate, butyrate, SCFA, gut microbiome, qPCR, LAD

Introduction

The human gut microbiome maintains the health of the host through the fermentation of non-digestible carbohydrates that escape small intestinal digestion and absorption[1]. The end-products of this fermentation - short-chain fatty acids (SCFAs) - serve as the main energy source for colonocytes[2], enhance the intestinal epithelial barrier[3], regulate mucus production[4], modulate inflammatory responses[5], induce apoptosis in colon cancer cells[6], regulate blood pressure[7], mediate gut-brain cross-talk[8], play a crucial regulatory role in glucose homeostasis[9], regulate lipid metabolism and adjusts satiety levels[10], among other effects.

It is estimated that, in healthy adult populations, the three major SCFAs, acetic-, propionic-, and butyric acid, accumulate in a 3:1:1 molar ratio[11-13]. Deviation of such proportions, with a significant decrease of butyrate levels, has been observed in people consuming a diet high in protein and low in carbohydrates[14]. Butyrate production is solely dependent on the intake of non-digestible fiber, while the major propionate-producers, like Bacteroidetes, metabolize peptides as well, thus rendering propionate levels unaltered[15]. Lower butyrate levels have also been linked to a slower fecal transition time and both are associated with a higher colonic pH, which, in turn, promotes the production of propionate[16]. A low-pH environment protects against the overgrowth of pathogens[17], thus, in this context, an increase of the propionate-to-butyrate ratio may indicate a vulnerable gastrointestinal state.

A deviant ratio in favor of propionate was proposed to act as a diagnostic marker for Irritable Bowel Syndrome (IBS)[18]. Increased levels of this acid (but not butyrate) were also reported in overweight and obese people[11], individuals with an increased risk for type 2 diabetes (T2D)[19], patients with Alzheimer's Disease (AD)[20] and those with Non-Alcoholic

Fatty Liver Disease (NAFLD)[21]. Additionally, a reduced butyrate concentration (but not propionate) was observed in people with a high risk of stroke[22].

While the evidence linking disproportionally low levels of butyrate and/or high levels of propionate with various diseases is expanding, routine diagnostic measurement for SCFA content remains challenging, mainly due to their high volatility and a complex sample clean-up procedure[23, 24].

Here, we aimed to infer SCFA levels by targeting a limited number of key bacteria using a novel qPCR-instrumentation-compatible method, Liquid Array Diagnostics (LAD)[25], circumventing the need to utilize Gas Chromatography-based methods. A LAD test targets variable regions within the 16S rRNA gene and allows the detection of up to 25 bacterial markers in a single-tube.

We focused on the propionate-to-butyrate ratio (P:B), a single variable with the potential of providing an indication of functional dysbiosis in clinical samples. The analytical strategy followed in this work is outlined in **Figure 1**.

Results and discussion

Identification of taxonomic biomarkers for propionate-to-butyrate ratio

We examined the microbiome composition (PacBio sequencing of the 16S rRNA gene) and the SCFA content of 93 adult fecal samples. We were interested in finding potential associations between different members of the microbiome and levels of propionate and butyrate, so that we could build a simple, predictive LAD test.

We computed the correlation between the CLR-transformed OTU readcounts and propionate and butyrate relative abundances. Only OTUs with > 0.2 or < -0.2 correlation ($p < 0.05$) were considered for further analysis. In total, 65 OTUs correlated to propionate levels (39 positively, 26 negatively correlated), and 62 correlated to butyrate (28 had a positive correlation, 34 a negative one). Out of these, 11 correlated to both butyrate and propionate, albeit in opposite directions.

A simplified network of SCFA/OTU relationships is presented in **Figure 2**.

A BLAST search was performed using OTU sequences as queries. Among the OTUs positively correlated to butyrate, we found some that shared high sequence identity with typical butyrate-producers, like *Fecalibacterium prausnitzii*[26, 27] (cor = 0.21, $p < 0.05$), *Agathobaculum butyriciproducens*[28] (cor = 0.23, $p < 0.05$), and *Coprococcus catus*[29] (cor = 0.21, $p < 0.05$). However, we also found a positive relationship between butyrate and the readcount of sequences sharing high identity with *Lactobacillus acidophilus* (cor = 0.22, $p < 0.05$), *Fusicatenibacter saccharivorans* (cor = 0.33, $p < 0.005$) and *Blautia wexlerae* (cor = 0.26, $p < 0.05$) - species not known to produce this acid[30-32]. Furthermore, *Dysosmobacter welbionis* (cor = -0.25, $p < 0.05$) and *Flavonifractor plautii* (cor = -0.3, $p < 0.005$), both butyrate-producers[33, 34], exhibited a negative correlation with the relative

abundance of butyrate. Similarly, propionate levels did not exclusively correlate to well-described propionate-producers.

Aware of this complex outcome, we decided to build a model based on a binary classification system, i.e., classifying samples as having a high or normal acid level. Aiming for a simple method, we chose to detect and classify samples based on a single variable that infers information about both acid concentrations, the propionate-to-butyrate ratio (P:B). Seeking to classify samples based on this ratio makes sense biologically, given that, in healthy adults, the molar ratio between propionate and butyrate is nearly 1.0[11-13]. Understanding the role of butyrate in maintaining human health[2, 6, 35-37], our goal was to detect samples where their levels are depleted – inferred by a deviant ratio in favor of propionate (i.e., P:B ratio >> 1.0).

We computed P:B from GC data for all samples. We then built a PLS + LDA model using OTU readcounts as predictors and aimed to find the ratio that best separated the two groups (normal vs. high ratio), while selecting a reasonably small number of OTUs to act as markers. We found the best separation to be at a ratio of 2.5 using only 37 OTUs as targets. These OTUs did not exclusively represent propionate- and butyrate-producers. The leave-one-out cross-validated model showed 90% sensitivity and a specificity of 98%.

(A table with GC measurements for each sample is presented in supplementary material, ***Table S1****, while a list of all OTUs correlated with propionate and/or butyrate is presented in* ***Table S2***).

Validation of the prediction model using a LAD-based test

We designed 21 LAD probes to cover all 37 OTUs selected by the PLS + LDA model, with the intention of converting the dry-lab results into a routine molecular diagnostic tool for

147 classification. Six of the probes failed to produce signal, so they were removed from the
148 assay. The remaining LAD probes were used to analyze 71 random samples in total, 9 of
149 which were not PacBio-sequenced. The performance of LAD probes is presented in
150 supplementary material, **Figure S1**.

151 Our primary model, which we used to select the OTU targets from PacBio sequencing
152 results, had generated the best P:B threshold at 2.5. When 15 LAD probe signals were used
153 as an input, the best separation - yielding the highest model prediction accuracy (leave-one-
154 out cross-validated), was observed at 2.6 (**Figure 3a**).

155 To assure that a high P:B (≥ 2.6) implied increased levels of propionate at the
156 expense of butyrate (not acetate), we computed the average levels of these acids within the
157 different groups. The average butyrate concentration for the normal-ratio group was 20%,
158 while the same for the high-ratio group was 7.2%. Samples with a normal P:B had on
159 average a propionate level of 16.6%, whereas samples high in such a ratio had a level of
160 29.8% (boxplots in **Figure 3c**).

161 The average sample P:B was 1.29 (median 0.92, minimum value 0.24 and maximum
162 9.4; the distribution is presented in **Figure 3b**). The median absolute deviation (MAD) was
163 0.54. A stringent way of finding outliers in positively skewed data is adding 3×MAD to the
164 median value[38]. Applying this formula, all samples with a P:B > 2.54 represented outliers
165 (0.92 + 3×0.54), which corresponded well with the threshold to which our model was
166 sensitive.

167 From 9 samples with a ratio ≥ 2.6, the algorithm correctly classified 7, and missed 2,
168 while from 62 samples with a ratio < 2.6, 61 were classified as such (**Figure 4**). The positive
169 predictive value showed that for any sample classified as having a “high ratio,” the chances

for that sample indeed having a ratio > 2.6 was 87.5%. The negative predictive value was 97%.

All 9 samples which were not PacBio-sequenced, and therefore not included in the initial model of selecting OTU markers, were correctly classified (all normal ratio).

Converted to a butyrate:propionate ratio, the border was at a range of 0.34 – 0.38. The model performed identically, correctly classifying 7 out of 9 samples with B:P ratio < 0.35, and 61 out of 62 as 'high' on the basis of a greater ratio.

We do not possess clinical details about the individuals whose samples we tested, and that may present a limitation for this study. It would be of particular interest to learn if these people suffer from health conditions for which high propionate and/or low butyrate has been reported. Nevertheless, we screened the metadata of 130 samples used by Zeng et al. (2019)[22], where significantly increased propionate levels were reportedly associated with a high risk of stroke. We found that on average, people with a low risk of stroke had a P:B < 2.6, while significantly higher P:B ratios were observed in people with medium and high risk of stroke (average P:B of 2.04, 3.22 and 2.84, for low, medium and high risk, respectively; $p < 0.05$).

A P:B border of ~2.6 was revealed to us using two different approaches. It represents a boundary separating normal samples from biological outliers in terms of both SCFA and microbiome composition. It could very well reflect an important biological threshold with a direct implication on the etiology of complex diseases.

Functional and strain resolution associations with the propionate-to-butyrate ratio

We chose to further analyze 23 randomly chosen samples of various P:Bs (17 normal, 6 high) by performing whole-genome shotgun sequencing, in an attempt to further explore the

193 biological differences between the two classes. On average, samples with a normal P:B
194 displayed 205 species, while samples with a high ratio harbored 10 fewer species,
195 suggesting a lower diversity in the latter. However, this difference did not exhibit an
196 acceptable significance level ($p > 0.1$).

197 Looking deeper into the composition, we found that high-ratio samples were
198 significantly richer in *Escherichia coli*, *Phocaeicola dorei* (a known propionate-producer,
199 formerly named as *Bacteroides dorei[39]*), *Enterocloster sp001517625* (named as
200 *Clostridium bouchedurhonense* at NCBI), *Blautia_A sp900066165*, and *Anaerotruncus*
201 *colihominis* (butyrate-producer[40]). There was also a tendency for lower *Fecalibacterium*
202 *praustnizii_C* (butyrate-producer[26, 27]) and *Eisenbergiella sp900066775*, and higher
203 *Akkermansia muciniphila* (propionate-producer[41]) ($p < 0.1$) (**Figure 5**).

204 Our test is designed to detect both *E. coli* and *F. prausnitzii*, who have commonly
205 been found to act as markers in a wide range of diseases[42-44].

206 Next, we used the sequencing reads to search for genes related to propionate and
207 butyrate production using Diamond software[45]. No linear relationship was found between
208 them, as presented in **Figure 6**.

209 This finding complemented well the ones retrieved with PacBio sequencing, where
210 the majority of OTUs correlated to either propionate or butyrate were not known to be
211 producers of such acids. Furthermore, possessing the ability to produce an acid did not
212 necessarily translate into a positive relationship with the product itself, as was the case with
213 *Dysosmobacter welbionis* and *Flavonifractor plautii* – both butyrate-producers, with a
214 relative abundance found in negative correlation with butyrate levels. (The latter was found
215 in positive correlation with propionate and is a target of our assay.)

Yet again, these results suggested that the levels of SCFA in fecal samples cannot be inferred by quantifying known acid-producers alone, presumably due to complex cross-feeding mechanisms involved[46]. For example, we believe that the inclusion of *Bifidobacterium adolescentis* (lactate- and acetate-producer) as a target of our test is tightly related to cross-feeding between this bacterium and well-described butyrate-producers[47, 48].

Clinical relevance

Currently, it seems like the most relevant clinical application of the P:B would be related to neurogenerative diseases, such as Alzheimer's[20] and Parkinson's disease[49]. A contributing cause for neurodegenerative diseases in the elderly is their reduced ability to metabolize propionate through decreased Methylmalonyl-CoA mutase activity[50]. This leads to the potential accumulation of toxic methylmalonic acid, which has been associated with decreased cognitive function in old people[51]. On the other hand, it has been shown that butyrate, a histone deacetylase inhibitor, can act as a therapeutic agent by reducing levels of abnormally deposited brain amyloid-β[52, 53]. Therefore, a potential clinical utility of the P:B measurement could be related to dietary advice in elderly in order to minimize potential harmful effects of propionate by increasing the beneficial butyrate.

There are also other diseases and disorders that can potentially be linked to a high P:B. Association with a significant propionate increase or butyrate decrease has been reported for these ailments, listed in **Table 2**. The most pronounced association related to P:B for these studies, is an increase in the propionate-to-butyrate ratio for IBS patients[18]. In addition to being a biomarker, there could also be a causality between the P:B and IBS

238 severity. Thus, this ratio could also potentially have a utility in treatment of these patients
239 through e.g. dietary advice.

240 Given the complex sample clean-up and preparation procedure, combined with a
241 high evaporation nature of acids, the SCFA measurement using today's technology remains
242 a challenging task[23, 24, 54]. That is why the accumulated knowledge in the field continues
243 to be derived from fragmented, small-scale studies, hardly standardized across laboratories.

244 The lack of robust methods for use in clinical settings creates a gap between state-
245 of-the-art knowledge in the field and its practical utility and application. A simple molecular
246 diagnostics method, like the LAD test presented here, allows inexpensive, high-throughput
247 screening of fecal samples, bridging this gap. The major benefits of LAD in a clinical setting
248 are related to simplicity and cost, in addition to detecting the microorganisms underlying
249 the P:B, which in turn can be used in therapeutics.

250 Our approach offers a solution for at least two problems. First, it focuses on the ratio
251 between propionate and butyrate, ignoring their absolute values which are known to
252 fluctuate based on the time of day of sample collection and processing[55]. Second, it
253 circumvents the need to measure SCFA levels – it utilizes a robust molecular diagnostic
254 system instead.

255 We offer an indirect way of detecting both propionate and butyrate levels,
256 identifying biological outliers - samples with highest propionate and/or lowest butyrate
257 proportions.

258 Conclusion

259 Here, we present a novel qPCR-instrumentation-compatible, single-tube multiplex test that
260 predicts samples with increased proportions of propionate in the expense of butyrate.

261 Circumventing the need to directly measure the SCFA content in fecal samples, a robust and
262 simple test like this will enable high-throughput analysis and regular monitoring of
263 functional dysbiosis in the gut.

Materials and Methods

Fecal samples and gDNA extraction

In total, 115 anonymized adult fecal samples, biobanked at Genetic Analysis AS (Oslo, Norway, research biobank no. 4071), were used for this study. Samples were collected and anonymized in accordance with the Norwegian Regional Committee for Medical and Health Research Ethics ruling (reference no. 2010/3209).

All fecal samples were stored at -40 °C prior to gDNA extraction or Gas Chromatography sample prep. The gDNA of all fecal samples was extracted using a MagMidi LGC kit (LGC Biosearch™ Technologies), following the steps suggested by the manufacturer. Genomic DNA extracts were further analyzed with LAD, PacBio SMRT[56], or Whole Genome Shotgun sequenced (Illumina).

Measurement of SCFA content with Gas Chromatography (GC)

The SCFA content of 115 fecal samples was measured with GC (Trace™ 1310 with autosampler, ThermoFisher Scientific™). Fecal samples were diluted in water (1:10) in a total volume of 1,500 µl, then homogenized for 2 X 40 seconds at 1,800 rpm using a Fastprep®-96 (MP Biomedicals). After a gentle spin, 300 µl of supernatant were transferred to a new tube, where 300 µl of internal standard was added. The internal standard consisted of 0.4% formic acid and 2 mM 2-methylvaleric acid. The samples were centrifuged at 13,000 rpm for 10 minutes. Subsequently, 300 µl of supernatant were transferred to spin columns (0.2 µm filters) and centrifuged at 10,000 rpm for 5 minutes. The solution that passed the membrane was transferred to GC vials for SCFA measurement. An internal standard (1 mM 2-methylvaleric acid), was used as a reference for sample-to-sample normalization of results. In total, 9 samples did not pass quality control by failing to produce

a measurement on acetic acid. Given that this acid is the most volatile one, its depletion was taken as an indication that the samples were compromised, therefore they were discarded from further processing. In addition, a sample erroneously handled during laboratory work was removed. The P:B of 105 remaining samples was computed and used for further data analyses.

PacBio Sequencing of 16S rRNA gene

Ninety-five samples randomly selected from 115 with SCFA contents determined by GC, were sent for PacBio Sequencing (Full-Length 16S Amplification SMRTbell® Library Preparation and Sequencing) at Norwegian Sequencing Centre (http://www.sequencing.uio.no). The first round of amplification was performed using our in-house 16S primers (GA-map® Forward primer 5'-TCCTACGGGAGGCAGCAG-3', GA-map® Reverse primer 5'-CGGTTACCTTGTTACGACTT-3', both protected by the US20110104692 A1 patent) tailed with universal sequences, as recommended in the PacBio protocol. The reads sharing at least 0.97 sequence identity were clustered into OTUs using the open source metagenomics tool VSEARCH[57]. The OTU readcounts were Center Log-Ratio (CLR) transformed[58] (after the addition of one pseudo-readcount) prior to further processing. Two of the 95 samples sent for sequencing did not pass the GC criteria (no measured acetate), hence were discarded from the downstream analysis.

Identification of bacterial targets through PLS + LDA modelling

CLR-transformed OTU readcounts from 93 samples were used as input for a PLS + LDA algorithm, with the aim of selecting variables (OTUs) to act as markers for classifying normal vs. high P:B samples [59]. Our aim was to correctly identify and classify the samples with the highest ratios, as they represent the deviation from the norm. We allowed the border

between the two types of samples (normal vs. high ratio) to go as low as possible without losing model prediction accuracy. The highest accuracy was reached at a P:B border of 2.5, with 37 OTUs acting as predictors, spanning 15 dimensions (leave-one-out cross-validated: sensitivity = 90%, specificity = 99%, positive prediction rate = 90%, negative prediction rate = 99%). These OTUs were subsequently considered as targets for LAD assay development.

Probe design for Liquid Array Diagnosis (LAD)

Eight-mer sequences containing a C at their 3' ends, shared only between 16S in-silico amplicons of target organisms, were computed using the in-house TNTProbeTool[60]. These sequences were considered as the 3' end segments of potential LAD labelling probes (LP). Probes had to have a minimum melting temperature (*T*m) of 70 °C (computed by the nearest-neighbor method) hybridizing to the target group, and a maximum *T*m of 30 °C hybridizing to a non-target group.

The final LP sequences did not contain the 3' end Cs. In this way, the presence of the corresponding bacterial target would ensure they become extended with a quencher-labelled ddCTP.

A reverse-complementary reporter probe (RP) was designed for each of the labelling probes. The RPs were designed with a fluorophore tag on their 5' ends, ensuring proximity to the quencher in duplexes harboring a 3´-labelled, RP-complementary LP. Duplexes containing the same fluorophore were designed with varying lengths to produce distinct temperature-dependent signals on the same qPCR channel of detection. The quenching effect of a longer duplex is observed as a dissociation curve with a higher *T*m. The DNA duplex *T*ms were calculated using the web-based OligoAnalyzer tool Tool^{TM} 3.1, also based on the nearest-neighbor method (Integrated DNA Technologies).

BLAST searches with each OTU sequence as query were performed to infer their taxonomy (blastn, nucleotide collection nt database).

Initially, 21 probes were designed, covering all 37 OTUs. However, six of the probes, targeting 11 OTUs (*Coprococcus eutactus, Alistipes indistinctus, Bacteroides eggerthii, [Clostridium] spiroforme, Ruthenibacterium lactatiformans, [Clostridium] glycyrrhizinilyticum*), failed to produce a signal, therefore they were excluded from the assay.

Due to sequence similarity between F_3_1 and R_12_1 LPs, designed to detect *Dorea longicatena*, and *Fusicatenibacter saccharivorans*, respectively, it was impossible to keep them in a single test tube, otherwise we would risk producing double signals when only one target was present. We therefore split the test into two tubes and divided the number of probes between them proportionally (8 probes in group 1, 7 probes in group 2). We used ROX_12_1 RP as a reporter probe for both LPs.

A list of final probes, their *T*ms, and their target species, respectively, are presented in **Table 1.**

Generation of templates for LAD labeling reaction

Genomic DNA from 71 available samples was PCR amplified. The SCFA content of all those samples had been measured, however 9 of them were not additionally PacBio sequenced. Each PCR reaction, with a total volume of 25 µl consisted of 5 µl bacterial lysate, 3.75 U of HOT FIREPOL® DNA Polymerase, 1 X B1 buffer, 2.5 mM $MgCl_2$ (all from Solis Biodyne, cat. no. 01-02-00500), 0.2 mM dNTPs (Thermo Fisher Scientific, cat. no. 18427088), 0.2 µM in-house primers (GA-map® Forward primer 5'-TCCTACGGGAGGCAGCAG-3', GA-map® Reverse primer 5'-CGGTTACCTTGTTACGACTT-3', both protected by the US20110104692 A1 patent). The

amplification was carried out using an Applied Biosystems Veriti ™ Thermal Cycler (Life Technologies), with an initiation period of 15 min at 95 °C, followed by 30 cycles of 30 s denaturation at 95 °C, 30 s annealing at 55 °C and 80 s of elongation at 72 °C, ending with a final step of elongation at 72 °C for 7 min. The PCR products were then treated with 2.7 U of Exonuclease I (ExoI, New England Biolabs Inc., cat. no. M0293L) and 7.36 U of Shrimp Alkaline Phosphatase (rSAP, New England BioLabs Inc., cat. no. M0371L) and set for incubation at 37 °C for 10 min, followed by 15 min at 80 °C to inactivate the enzymes.

Single nucleotide extension of LPs and melting curve analysis - LAD

Ten ml of ExoI-SAP treated PCR products (14.5-25.6 ng/µl) were used as templates for LP labelling. The labelling reaction was also comprised of LPs in 0.1 µM final concentration (biomers.net), 1 X Buffer C, 1 mM final concentration $MgCl_2$, 7.5 U Hot TERMIpol® DNA polymerase (all from Solis Biodyne, cat. no. 01-06-00500), and 0.96 µM ddCTP-DYQ660 (Jena Bioscience, cat. no. NU-850-660Q). The reaction was performed in a PCR instrument, with an initiation step at 95 °C for 12 minutes, followed by 40 cycles of denaturing (96 °C for 20 seconds) and annealing/elongation (68 °C for 40 seconds).

Following labelling, a mixture of RPs in a 0.01 µM final concentration for each RP and 5 mM $MgCl_2$ were added to the reactions. Reagent S, available from INN (Inland Norway University of Applied Sciences, Norway), was also added to a final concentration of 1 %. The melting curve analysis (31 °C to 85 °C) was performed using a CFX96 qPCR instrument (Bio-Rad Laboratories).

The extraction of peaks and the determination of positive signals was performed similarly as described in the material and methods section of Hiseni et al. (2019)[25], with a slight modification. Prior to extracting the signals, the fluorescence measurements within each

channel were centered with the purpose of minimizing the range of measurements across wells at any given temperature. Then the baseline within each channel was corrected (flattened) by subtracting the centered values of each sample with the average "No template control" - centered values. As an ultimate step, for group 1 samples only, a further correction of FAM and CY5 baselines was performed by subtracting the values of each other (FAM = FAM-CY5 and CY5 = CY5-FAM).

Bioinformatic evaluation of probe specificity

OTU sequences (PacBio) were used as subjects to check for sequences complementary to 3' C-labelled probes. A search for the occurrence of probes, allowing one mismatch anywhere along the sequence (excluding the probe 3'-C) was performed. The intention of this step was to prove that probes precisely targeted the intended bacteria. Sequences resulting positive in the containment of probes were considered as "labelling templates". The readcounts of all such sequences were considered as in-silico signals, which were then used to compute the correlation with real LAD signals.

Whole Genome Shotgun (WGS) Sequencing

In total, 24 samples were sent for WGS sequencing at Norwegian Sequencing Center. One of the samples failed the GC quality check (no measured acetate), hence was removed from further analysis.

The libraries for the remaining 23 samples were prepared using a Nextera™ DNA Flex Library Preparation kit (Illumina), following its manufacturer-recommended protocol. Samples had different SCFA levels, spanning well across the P:B values.

Processing of WGS Sequencing results

Diamond software[45] was used to search for genes related to propionate and butyrate into raw WGS sequencing reads. For propionate, we searched for the genes Methylmalonyl-CoA decarboxylase, alpha-subunit (*mmdA*), Lactoyl-CoA dehydratase subunit alpha (*lcdA*), and CoA-dependent propionaldehyde dehydrogenase (*pduP*) (markers for succinate, acrylate and propanediol pathways, respectively [61]). For butyrate, the process involved searching for butyryl-CoA:acetate CoA transferase and butyrate kinase. For each read, only the hit with the highest bit score per pathway was kept (e-value <= 1e-05). Next, for each sample the number of reads that got a hit with one of the genes was counted, then grouped and summed according to the SCFAs to which they were related. After normalizing for the query sequence size and sequencing depth, the total number of hits related to propionate and butyrate was compared to the relative abundance of these acids.

The taxonomic assignment of the sequencing reads was performed using a combination of Kraken2[62], KrakenUniq[63] and Bracken[64], as described in Hiseni et al. (2021)[65], using HumGut_975[65] as a custom database.

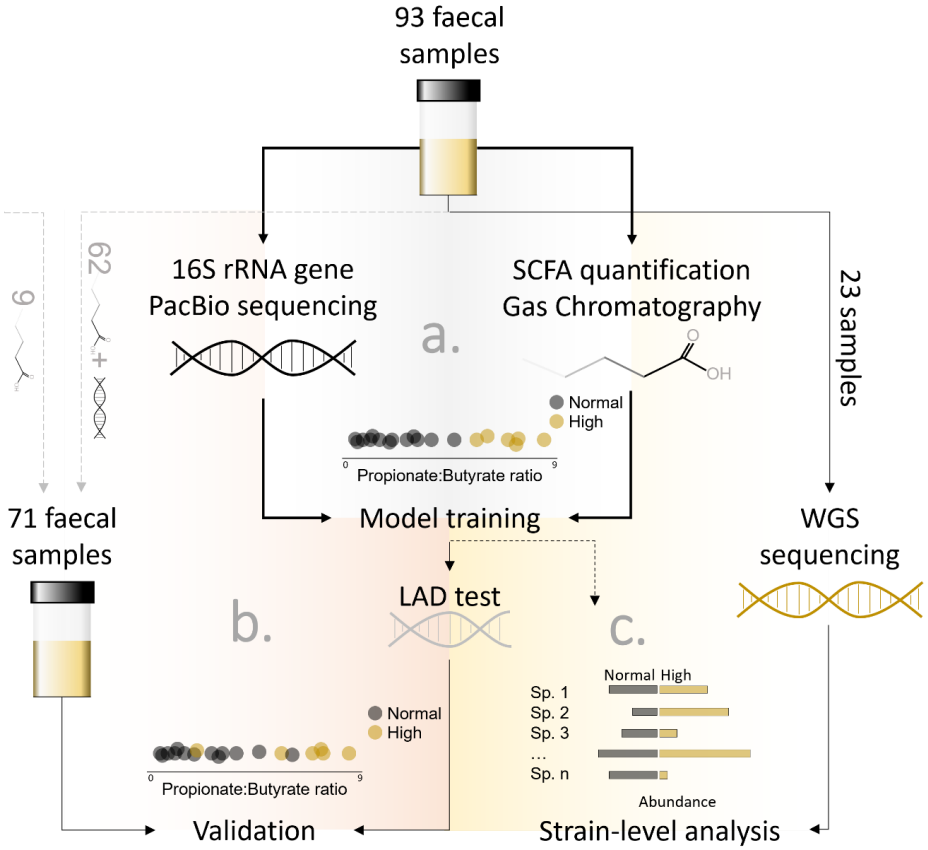
Future Perspective

Assays targeting 16S rRNA gene will become widely used for microbial functional analysis.

Executive summary

- Healthy adult fecal propionate and butyrate levels are expected to be equimolar.
- An increased propionate-to-butyrate ratio (P:B) has been linked to several health disorders.
- Measurement of SCFA levels is challenging due to their highly volatile nature, presenting a major bottleneck for high-throughput studies.
- The challenges related to SCFA measurements create a gap between the acquired knowledge in the field and its clinical utility.
- This article presents a method for predicting and classifying samples with significantly elevated P:B by directly targeting predictor bacteria, circumventing thus the need to measure SCFA levels.
- The method is based on a Liquid Array Diagnostics (LAD) assay, a qPCR-compatible test capable of detecting multiple targets in a single-tube multiplex reaction.
- The test predicting high P:B samples showed 78% sensitivity and 98% specificity (leave-one-out cross-validated)
- The assay presented here has the potential to be utilized in high-throughput studies, validating the reported findings in the literature, in addition to serving as a robust screening tool for routine diagnostics.

436 **Figures**



437

438 **Figure 1.** Building and validation of a propionate:butyrate ratio (P:B) prediction model. **a**.
439 Identification of taxonomic biomarkers for P:B In this step, 93 fecal samples were analyzed
440 for both their taxonomic composition (16S rRNA gene sequencing with PacBio SMRT
441 technology) and SCFA content (Gas Chromatography). A PLS + LDA model was built,
442 selecting a limited number of OTUs to act as predictors of normal vs. high P:B **b**. Validation
443 of the prediction model using a LAD-based test. In total, 71 fecal samples, 9 of which were
444 not PacBio sequenced, were tested with a set of LAD probes, designed to target OTUs
445 selected by the PLS + LDA model in the previous step. **c**. Functional and strain resolution
446 associations with P:B.

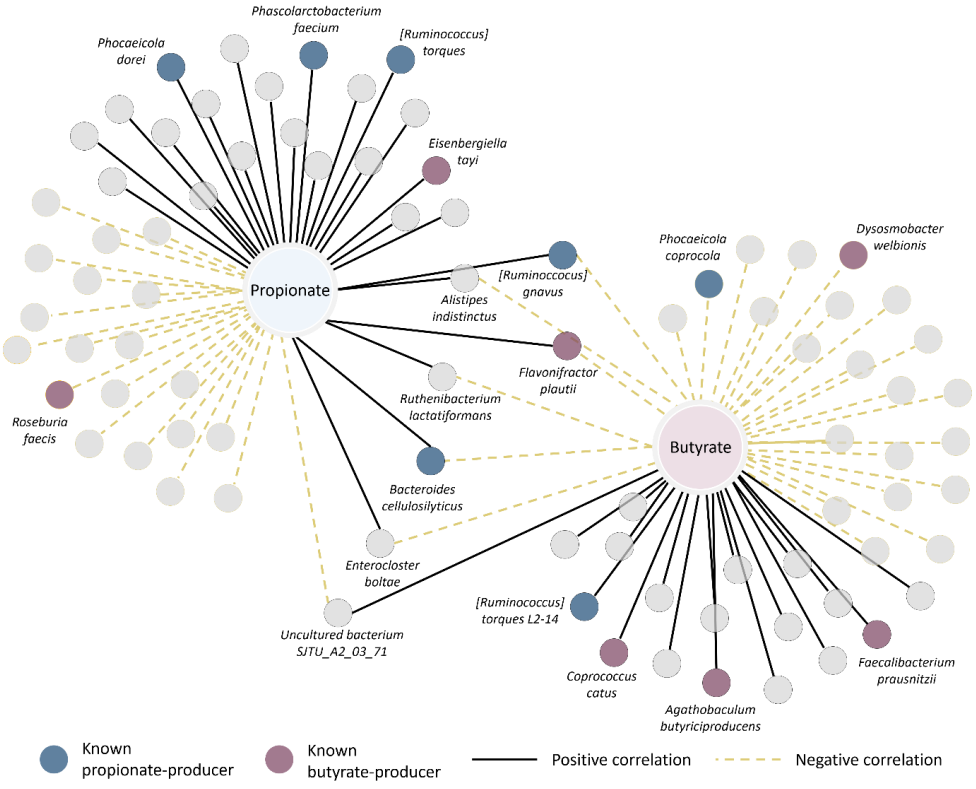
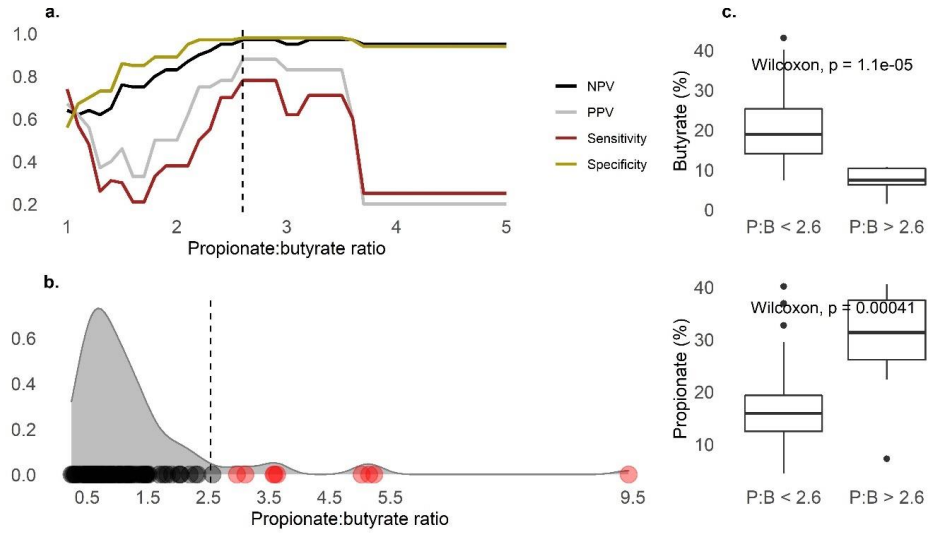
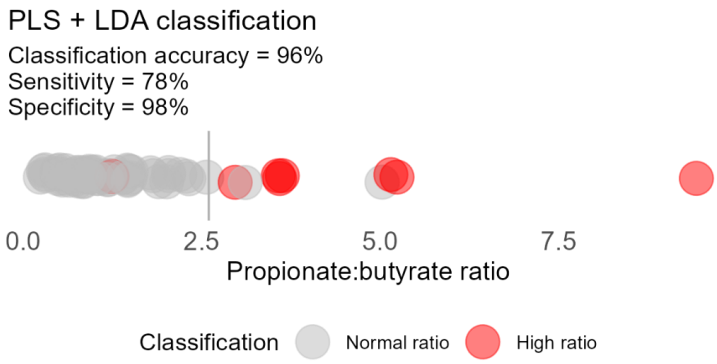


Figure 2. A network graph depicting the relationship between propionate and butyrate with significantly correlated OTUs. Similar OTUs were grouped together after checking for their taxonomy with BLAST. The blue nodes represent species that are known as propionate-producers, while the purple ones represent well-described butyrate-producers. Positive correlations between the SCFA and OTUs are presented with black edges (lines), while the negative ones are depicted in yellow dashed lines.



454

455 **Figure 3. a**. Diagnostic testing accuracy of various P:B thresholds using LAD signals as
456 predictors. Different thresholds were tested to determine the border between high and
457 normal ratio samples for further PLS + LDA classification. The dashed vertical line depicts the
458 2.6 ratio, the lowest ratio to yield the highest sensitivity, specificity, negative predictive
459 value (NPV) and positive predictive value (PPV). **b**. Ratio density among 71 tested samples.
460 Most of samples (three quartiles) had a ratio < 1.5 while the median ratio was 0.9. The
461 dashed vertical line at 2.54 separates the outliers from the data (Median + 3 × Median
462 Absolute Deviation). Dots along the X-axis show measured ratios for each sample, colored
463 based on classes they belong to according to the PLS + LDA model: black = normal P:B, red =
464 high P:B. **c**. Boxplots showing the difference in distribution of butyrate (upper panel) and
465 propionate (lower panel) levels across two different groups of samples (normal ratio = P:B <
466 2.6, high ratio = P:B > 2.6).



467

468 **Figure 4.** The PLS + LDA model prediction accuracy with LAD probe signals used as an input.
469 The vertical line placed at 2.6 marks the borderline between normal and high P:B. The
470 positioning of each dot shows the real sample ratio (x-axis), while the dot color indicates the
471 classification by the model: grey = normal, red = high. Most (7 out of 9) samples were
472 correctly classified as having a high ratio (red dots on the right side of the 2.6 border). Only
473 one normal-ratio sample was miss-classified as a high-ratio one (the red dot on the left).

474

475 **Figure 5.** Top 50 species with the greatest differences in on-average-abundance between
476 groups (normal vs. high P:B samples). Grey circles indicate the average abundance on the
477 normal-ratio samples, burgundy represents the abundance of samples with a high-ratio. The
478 circle size shows the percentage of samples within the group where the bacteria was found
479 present. The arrows point towards samples with a high ratio, marking the tendency of
480 increased (facing right) or decreased (facing left) species abundance in them. The dot and

star symbols indicate the significant differences (Wilcoxon p-value: *** = p < 0.005, * = p < 0.05, • = p < 0.1).

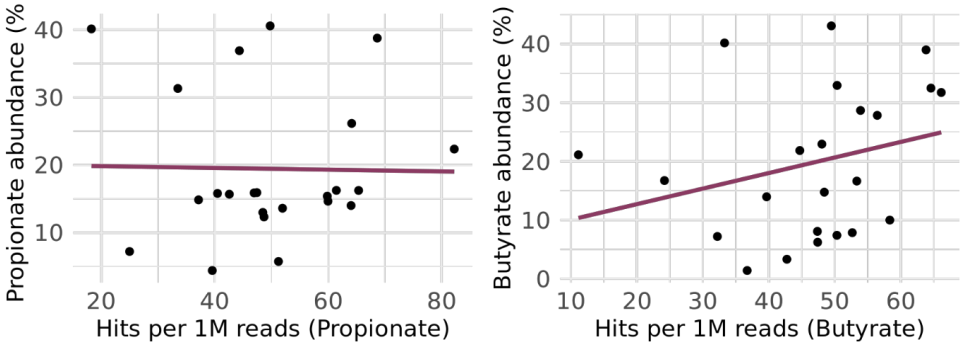
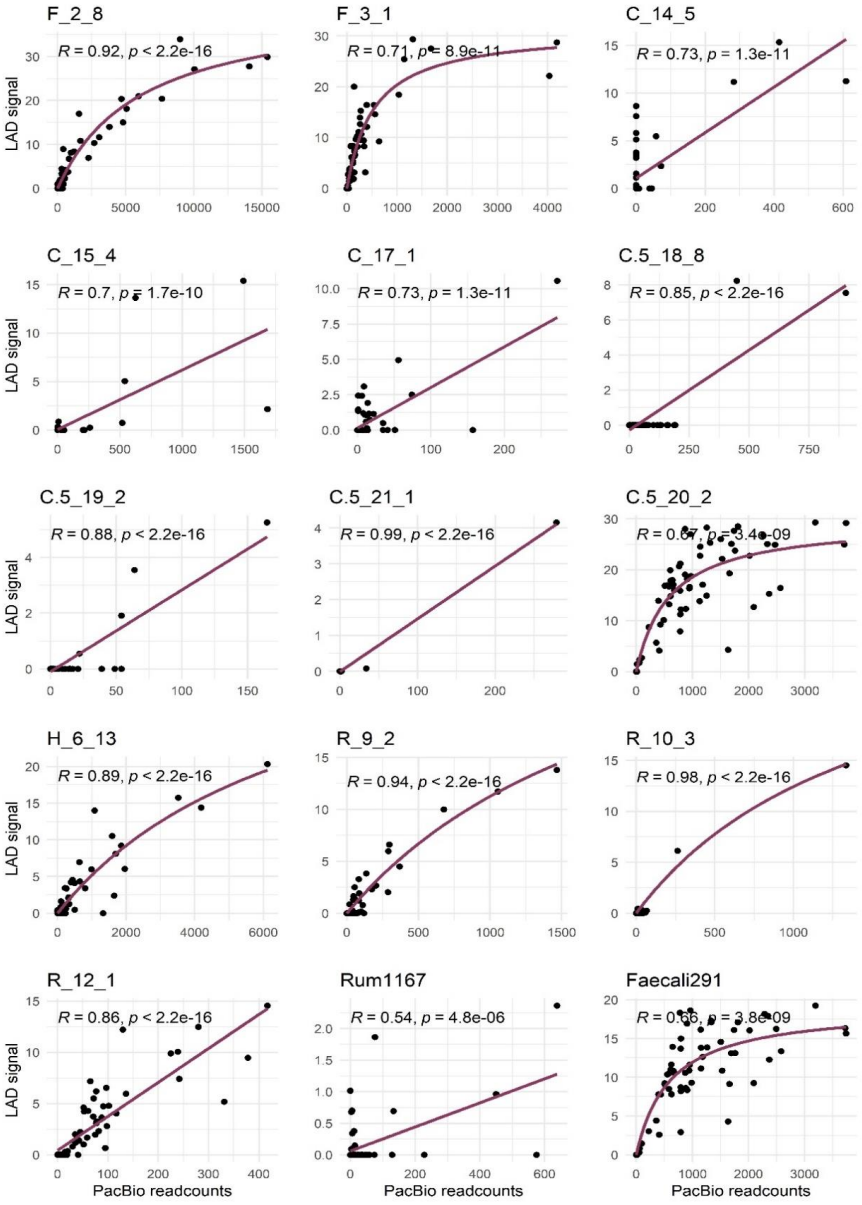


Figure 6. The relationship between the propionate and butyrate relative abundances with the corresponding number of reads that got a hit (highest bit score, e-value <= 1e-05) with respective marker genes. The number of hits was normalized after considering the sequence length of queries and sequencing depth.



488

489 **Figure S1.** LAD signals (x-axis) plotted against PacBio readcounts (y-axis) for in-silico targets.
490 The results for the 62 sequenced samples were used to perform in-silico labelling of the
491 probes, to confirm their specificity. Single mismatches were allowed. All OTUs sequences
492 that matched with probe sequences were considered as potential templates for bona-fide
493 labelling.

494 **Tables**

495 **Table 1.** The Liquid Array Diagnostics test probe list[1]

LP name	LP sequence (5'-3')	RP name	RP sequence (5'-3')	Tm (°C)	Target species	Group
F_2_8	GCTACACACGTGCTACAATGGCGCATA	FAM_2_8	tTATGCGCCATTG	43.7	Escherichia coli	1
F_3_1	CGGGACTGCATTTGGAACTGCTGAG	ROX_12_1	tAGCCAGACAGTTTCCA ATGCAGTCCCA	52.9	Dorea longicatena	1
H_6_13	GGTGGATGCTGGATGTGGGGAC	HEX_6_13	ttGTCCCCACATCC	45.5	Bifidobacterium adolescentis	2
R_9_2	CCGGGACTGCTTTGGAAACTATGCAG	ROX_9_2	ttCTGCATACTTTCCAAA GC	43	Coprococcus comes	2
R_10_3	GGAGCGTAGAAGGCATTGCAAGC	ROX_10_3	ttGCTTGCAATGCCTTC	52.8	Blautia sp. Marseille-P3313	2
R_12_1	TGGGACTGCATTGGAAACTGTCTGGCT	ROX_12_1	tAGCCAGACAGTTTCCA ATGCAGTCCCA	69.3	Fusicatenibacter saccharivorans	2
C_14_5	CCCGTCACTCCATGAGAGTTGGAGATAC	CY5_14_5	ttGTATCTCCAACTCTC	45.5	Uncultured bacterium clone AP07S.190	1
C_15_4	CCGTACTGGCTCTGGAAACTGTTCAG	CY5_15_4	ttCTGAACAGTTTCCAGA GC	55.4	Holdemanella biformis	1
C_17_1	GGCCACACACGTACTACAATGGTGGTT AA	CY5_17_1	tTTAACCACCATTGTAGT ACGTGTGTGG	64.6	Flavonifractor plautii, Flintibacter sp. KGMB00164	1
C.5_18_8	TGGAAGCCGGGAGTACCTGAAG	CY5.5_18_8	ttCTTCAGGTACcCCC	35.9	Barnesiella sp. strain mt172, Barnesiella sp. strain mt155	1
C.5_19_2	CGCGAGGGGGAGCAAAACTGGAAAA	CY5.5_19_2	tTTTTCCAGTTTTGC	44.8	Uncultured bacterium isolate DGGE gel band RB1-25	1
C.5_20_2	GCGGACTACTGGGCACCAA	CY5.5_20_2	tTTGGTGCCCAGTAGTC	55.2	Fecalibacterium prausnitzii, Uncultured bacteria clones: 2-002-f10, A3-213, and TS3_a01c08	2
C.5_21_1	GGAAGCGACTGGGCAACCAGAAG	CY5.5_21_1	ttCTTCTGGTTGCCCAGT CGCTTC	64.9	Uncultured organism clone ELU0116-T290-S-NI_000152,	1
Fecali291	TTGCTCCACCTCGCGGTCTTGCTTCTCTT TGTTTAA	Fecali291 CY5	TTAAACAAAGAGAAGCA AGACCGCGAGGTGGAG CAA	72.2 ºC	Fecalibacterium prausnitzii, [Eubacterium] siraeum V10Sc8a, Ruminococcaceae bacterium strain MT139, Uncultured bacteria clones: PB1_aai26e05, C3-2 16S, A3-213, TS3_a01c08, SJTU_A2_03_71, and A5_016	2
Rum1167	CACTCTAGCCTGACAGTT	Rum1167 CY5	AACTGTCAGGCTAG	47.1 ºC	[Ruminococcus] gnavus, Uncultured bacteria clones: SJTU_G_10_25, Cadhufec15ml, and CFT114H1	2

496

497

498

499

500

501

502

[1] The lower-case t nucleotides in RP sequences represent the 5'-end T-tails. These were introduced with the purpose of securing physical distance between the fluorophore and Gs (either within the RP sequence itself, or the Gs in the 3' end of the complementary LP sequence). Gs are known to have an intrinsic quenching property.

503 **Table 2.** Studies associating diseases with an increase/decrease of SCFA fecal levels (high
504 P:B)

Reference	Health disorder	Individuals tested	Significant change compared to controls
Schwiertz et al. (2010)[11]	Obesity	30 lean 35 overweight 33 obese	↑Total SCFA ↑Propionate
Sanna et al. (2019)[19]	Type 2 diabetes	952 participants from LifeLines Deep cohort	↑Propionate
Rau et al. (2018)[21]	Nonalcoholic fatty liver disease (NAFLD)	27 healthy 32 NAFLD	↑Acetate ↑Propionate
Farup et al. (2016)[18]	Irritable bowel Syndrome (IBS)	25 healthy 25 IBS	↓Butyrate
Zeng et al. (2019)[22]	Stroke	51 low risk of stroke 54 medium risk 36 high risk	↓Butyrate
Liu et al. (2019)[66]	Autism spectrum disorder (ASD)	20 healthy 30 ASD	↓Acetate ↓Butyrate ↑Valerate
Wang et al. (2019)[67]	Chronic kidney disease (CKD)	61 healthy 128 CKD	↓Butyrate
Strati et al. (2016)[68]	Rett syndrome	29 healthy 50 RTT	↑Total SCFA ↑Propionate ↑Isovalerate ↑Isobutyrate
Tana et al. (2010)[69]	IBS	26 healthy 26 IBS	↑Total SCFA ↑Acetate ↑Propionate
Unger et al. (2016)[49]	Parkinson's disease (PD)	34 healthy 34 PD	↓Total SCFA ↓Butyrate

505

Literature

1. Morrison DJ, Preston T. Formation of short chain fatty acids by the gut microbiota and their impact on human metabolism. *Gut microbes* 7(3), 189-200 (2016).
2. Donohoe DR, Garge N, Zhang X *et al*. The microbiome and butyrate regulate energy metabolism and autophagy in the mammalian colon. *Cell metabolism* 13(5), 517-526 (2011).
3. Wang H-B, Wang P-Y, Wang X, Wan Y-L, Liu Y-C. Butyrate Enhances Intestinal Epithelial Barrier Function via Up-Regulation of Tight Junction Protein Claudin-1 Transcription. *Digestive Diseases and Sciences* 57(12), 3126-3135 (2012).
4. Willemsen LEM, Koetsier MA, Van Deventer SJH, Van Tol EaF. Short chain fatty acids stimulate epithelial mucin 2 expression through differential effects on prostaglandin E₁ and E₂ production by intestinal myofibroblasts. 52(10), 1442-1447 (2003).
5. Chen J, Vitetta L. Inflammation-Modulating Effect of Butyrate in the Prevention of Colon Cancer by Dietary Fiber. *Clinical Colorectal Cancer* 17(3), e541-e544 (2018).
6. Encarnação JC, Abrantes AM, Pires AS, Botelho MF. Revisit dietary fiber on colorectal cancer: butyrate and its role on prevention and treatment. *Cancer metastasis reviews* 34(3), 465-478 (2015).
7. Miyamoto J, Kasubuchi M, Nakajima A, Irie J, Itoh H, Kimura I. The role of short-chain fatty acid on blood pressure regulation. 25(5), 379-383 (2016).
8. Dalile B, Van Oudenhove L, Vervliet B, Verbeke K. The role of short-chain fatty acids in microbiota–gut–brain communication. *Nature Reviews Gastroenterology & Hepatology* 16(8), 461-478 (2019).
9. Ulven T. Short-chain free fatty acid receptors FFA2/GPR43 and FFA3/GPR41 as new potential therapeutic targets. *Frontiers in endocrinology* 3 111 (2012).
10. Chambers ES, Viardot A, Psichas A *et al*. Effects of targeted delivery of propionate to the human colon on appetite regulation, body weight maintenance and adiposity in overweight adults. *Gut* 64(11), 1744-1754 (2015).
11. Schwiertz A, Taras D, Schäfer K *et al*. Microbiota and SCFA in Lean and Overweight Healthy Subjects. 18(1), 190-195 (2010).
12. Cook, Sellin. Review article: short chain fatty acids in health and disease. 12(6), 499-507 (1998).
13. Fernandes J, Su W, Rahat-Rozenbloom S, Wolever TMS, Comelli EM. Adiposity, gut microbiota and faecal short chain fatty acids are linked in adult humans. *Nutrition & diabetes* 4(6), e121-e121 (2014).
14. Duncan SH, Belenguer A, Holtrop G, Johnstone AM, Flint HJ, Lobley GE. Reduced dietary intake of carbohydrates by obese subjects results in decreased concentrations of butyrate and butyrate-producing bacteria in feces. *Appl Environ Microbiol* 73(4), 1073-1078 (2007).
15. Russell WR, Gratz SW, Duncan SH *et al*. High-protein, reduced-carbohydrate weight-loss diets promote metabolite profiles likely to be detrimental to colonic health. 93(5), 1062-1072 (2011).
16. Walker AW, Duncan SH, Leitch ECM, Child MW, Flint HJ. pH and Peptide Supply Can Radically Alter Bacterial Populations and Short-Chain Fatty Acid Ratios within Microbial Communities from the Human Colon. 71(7), 3692-3700 (2005).
17. Duar RM, Kyle D, Casaburi G. Colonization Resistance in the Infant Gut: The Role of B. infantis in Reducing pH and Preventing Pathogen Growth. 9(2), 7 (2020).
18. Farup PG, Rudi K, Hestad K. Faecal short-chain fatty acids - a diagnostic biomarker for irritable bowel syndrome? *BMC Gastroenterology* 16(1), 51 (2016).*

19. Sanna S, Van Zuydam NR, Mahajan A *et al*. Causal relationships among the gut microbiome, short-chain fatty acids and metabolic diseases. *Nature Genetics* 51(4), 600-605 (2019).*
20. Killingsworth J, Sawmiller D, Shytle RDJFIaN. Propionate and Alzheimer's Disease. 12 501 (2020).
21. Rau M, Rehman A, Dittrich M *et al*. Fecal SCFAs and SCFA-producing bacteria in gut microbiome of human NAFLD as a putative link to systemic T-cell activation and advanced disease. 6(10), 1496-1507 (2018).*
22. Zeng X, Gao X, Peng Y *et al*. Higher Risk of Stroke Is Correlated With Increased Opportunistic Pathogen Load and Reduced Levels of Butyrate-Producing Bacteria in the Gut. 9(4), (2019).*
23. Torii T, Kanemitsu K, Wada T, Itoh S, Kinugawa K, Hagiwara A. Measurement of short-chain fatty acids in human faeces using high-performance liquid chromatography: specimen stability. 47(5), 447-452 (2010).
24. Li M, Zhu R, Song X, Wang Z, Weng H, Liang J. A sensitive method for the quantification of short-chain fatty acids by benzyl chloroformate derivatization combined with GC-MS. *Analyst* 145(7), 2692-2700 (2020).
25. Hiseni P, Wilson RC, Storrø O, Johnsen R, Øien T, Rudi K. Liquid array diagnostics: a novel method for rapid detection of microbial communities in single-tube multiplex reactions. *BioTechniques* 66(3), 143-149 (2019).
26. Duncan SH, Hold GL, Harmsen HJM, Stewart CS, Flint HJ. Growth requirements and fermentation products of Fusobacterium prausnitzii, and a proposal to reclassify it as Faecalibacterium prausnitzii gen. nov., comb. nov. *Int J Syst Evol Microbiol* 52(Pt 6), 2141-2146 (2002).
27. Lopez-Siles M, Khan TM, Duncan SH, Harmsen HJM, Garcia-Gil LJ, Flint HJ. Cultured representatives of two major phylogroups of human colonic Faecalibacterium prausnitzii can utilize pectin, uronic acids, and host-derived substrates for growth. *Appl Environ Microbiol* 78(2), 420-428 (2012).
28. Ahn S, Jin TE, Chang DH *et al*. Agathobaculum butyriciproducens gen. nov. sp. nov., a strict anaerobic, butyrate-producing gut bacterium isolated from human faeces and reclassification of Eubacterium desmolans as Agathobaculum desmolans comb. nov. *Int J Syst Evol Microbiol* 66(9), 3656-3661 (2016).
29. Louis P, Flint HJ. Formation of propionate and butyrate by the human colonic microbiota. *Environmental microbiology* 19(1), 29-41 (2017).
30. Zhao R, Sun J, Mo H, Zhu YJWJOM, Biotechnology. Analysis of functional properties of Lactobacillus acidophilus. 23(2), 195-200 (2007).
31. Takada T, Kurakawa T, Tsuji H, Nomoto K. Fusicatenibacter saccharivorans gen. nov., sp. nov., isolated from human faeces. *Int J Syst Evol Microbiol* 63(Pt 10), 3691-3696 (2013).
32. Liu C, Finegold SM, Song Y, Lawson PA. Reclassification of Clostridium coccoides, Ruminococcus hansenii, Ruminococcus hydrogenotrophicus, Ruminococcus luti, Ruminococcus productus and Ruminococcus schinkii as Blautia coccoides gen. nov., comb. nov., Blautia hansenii comb. nov., Blautia hydrogenotrophica comb. nov., Blautia luti comb. nov., Blautia producta comb. nov., Blautia schinkii comb. nov. and description of Blautia wexlerae sp. nov., isolated from human faeces. 58(8), 1896-1902 (2008).
33. Le Roy T, Moens De Hase E, Van Hul M *et al*. Dysosmobacter welbionis is a newly isolated human commensal bacterium preventing diet-induced obesity and metabolic disorders in mice. doi:10.1136/gutjnl-2020-323778 %J Gut gutjnl-2020-323778 (2021).
34. Carlier J-P, Bedora-Faure M, Apos, Ouas G, Alauzet C, Mory F. Proposal to unify Clostridium orbiscindens Winter et al. 1991 and Eubacterium plautii (Séguin 1928) Hofstad and Aasjord 1982, with description of Flavonifractor plautii gen. nov., comb. nov., and reassignment of Bacteroides capillosus to Pseudoflavonifractor capillosus gen. nov., comb. nov. 60(3), 585-590 (2010).

35. Takahashi K, Nishida A, Fujimoto T *et al*. Reduced Abundance of Butyrate-Producing Bacteria Species in the Fecal Microbial Community in Crohn's Disease. *Digestion* 93(1), 59-65 (2016).
36. Laserna-Mendieta EJ, Clooney AG, Carretero-Gomez JF *et al*. Determinants of Reduced Genetic Capacity for Butyrate Synthesis by the Gut Microbiome in Crohn's Disease and Ulcerative Colitis. *Journal of Crohn's & colitis* 12(2), 204-216 (2018).
37. Pozuelo M, Panda S, Santiago A *et al*. Reduction of butyrate- and methane-producing microorganisms in patients with Irritable Bowel Syndrome. *Scientific Reports* 5(1), 12693 (2015).
38. Leys C, Ley C, Klein O, Bernard P, Licata LJJOESP. Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. 49(4), 764-766 (2013).
39. Gutiérrez N, Garrido D. Species Deletions from Microbiome Consortia Reveal Key Metabolic Interactions between Gut Microbes. *mSystems* 4(4), e00185-00119 (2019).
40. Lawson PA, Song Y, Liu C *et al*. Anaerotruncus colihominis gen. nov., sp. nov., from human faeces. *Int J Syst Evol Microbiol* 54(Pt 2), 413-417 (2004).
41. De Vos WM. Microbe Profile: Akkermansia muciniphila: a conserved intestinal symbiont that acts as the gatekeeper of our mucosa. 163(5), 646-648 (2017).
42. Khan AA, Khan Z, Malik A *et al*. Colorectal cancer-inflammatory bowel disease nexus and felony of Escherichia coli. *Life Sciences* 180 60-67 (2017).
43. Lopez-Siles M, Martinez-Medina M, Busquets D *et al*. Mucosa-associated Faecalibacterium prausnitzii and Escherichia coli co-abundance can distinguish Irritable Bowel Syndrome and Inflammatory Bowel Disease phenotypes. *International Journal of Medical Microbiology* 304(3), 464-475 (2014).
44. Machiels K, Joossens M, Sabino J *et al*. A decrease of the butyrate-producing species Roseburia hominis and Faecalibacterium prausnitzii defines dysbiosis in patients with ulcerative colitis. *Gut* 63(8), 1275 (2014).
45. Buchfink B, Reuter K, Drost H-G. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nature Methods* 18(4), 366-368 (2021).
46. Ríos-Covián D, Ruas-Madiedo P, Margolles A, Gueimonde M, De Los Reyes-Gavilán CG, Salazar N. Intestinal Short Chain Fatty Acids and their Link with Diet and Human Health. 7(185), (2016).
47. Belenguer A, Duncan SH, Calder AG *et al*. Two Routes of Metabolic Cross-Feeding between Bifidobacterium adolescentis and Butyrate-Producing Anaerobes from the Human Gut. 72(5), 3593-3599 (2006).
48. Rios-Covian D, Gueimonde M, Duncan SH, Flint HJ, De Los Reyes-Gavilan CG. Enhanced butyrate formation by cross-feeding between Faecalibacterium prausnitzii and Bifidobacterium adolescentis. *FEMS Microbiology Letters* 362(21), (2015).
49. Unger MM, Spiegel J, Dillmann K-U *et al*. Short chain fatty acids and gut microbiota differ between patients with Parkinson's disease and age-matched controls. *Parkinsonism & Related Disorders* 32 66-72 (2016).*
50. Elliot JM. Propionate metabolism and vitamin B12. In: *Digestive Physiology and Metabolism in Ruminants: Proceedings of the 5th International Symposium on Ruminant Physiology, held at Clermont — Ferrand, on 3rd–7th September, 1979*, Ruckebusch Y,Thivend P (Ed.^(Eds).Springer Netherlands Dordrecht 485-503 (1980).
51. Mccracken C, Hudson P, Ellis R, Mccaddon A. Methylmalonic acid and cognitive function in the Medical Research Council Cognitive Function and Ageing Study. *Am J Clin Nutr* 84(6), 1406-1411 (2006).
52. Fernando W, Martins IJ, Morici M *et al*. Sodium Butyrate Reduces Brain Amyloid-β Levels and Improves Cognitive Memory Performance in an Alzheimer's Disease Transgenic Mouse Model at an Early Disease Stage. *Journal of Alzheimer's disease : JAD* 74(1), 91-99 (2020).

53. Govindarajan N, Agis-Balboa RC, Walter J, Sananbenesi F, Fischer A. Sodium Butyrate Improves Memory Function in an Alzheimer's Disease Mouse Model When Administered at an Advanced Stage of Disease Progression. *Journal of Alzheimer's Disease* 26 187-197 (2011).
54. García-Villalba R, Giménez-Bastida JA, García-Conesa MT, Tomás-Barberán FA, Carlos Espín J, Larrosa M. Alternative method for gas chromatography-mass spectrometry analysis of short-chain fatty acids in faecal samples. 35(15), 1906-1913 (2012).
55. Jones J, Reinke SN, Ali A, Palmer DJ, Christophersen CT. Fecal sample collection methods and time of day impact microbiome composition and short chain fatty acid concentrations. *Scientific Reports* 11(1), 13964 (2021).
56. Rhoads A, Au KF. PacBio Sequencing and Its Applications. *Genomics, Proteomics & Bioinformatics* 13(5), 278-289 (2015).
57. Rognes T, Flouri T, Nichols B, Quince C, Mahé F. VSEARCH: a versatile open source tool for metagenomics. *PeerJ* 4 e2584-e2584 (2016).
58. Aitchison J. The statistical analysis of compositional data. *Journal of the Royal Statistical Society: Series B (Methodological)* 44(2), 139-160 (1982).
59. Mehmood T, Martens H, Sæbø S, Warringer J, Snipen L. A Partial Least Squares based algorithm for parsimonious variable selection. *Algorithms for molecular biology : AMB* 6(1), 27 (2011).
60. Eggesbø M, Moen B, Peddada S *et al*. Development of gut microbiota in infants not exposed to medical interventions. *APMIS* 119(1), 17-35 (2011).
61. Reichardt N, Duncan SH, Young P *et al*. Phylogenetic distribution of three pathways for propionate production within the human gut microbiota. *The ISME Journal* 8(6), 1323-1335 (2014).
62. Wood DE, Lu J, Langmead B. Improved metagenomic analysis with Kraken 2. *Genome biology* 20(1), 257 (2019).
63. Breitwieser FP, Baker DN, Salzberg SL. KrakenUniq: confident and fast metagenomics classification using unique k-mer counts. *Genome Biology* 19(1), 198 (2018).
64. Lu J, Breitwieser FP, Thielen P, Salzberg SL. Bracken: estimating species abundance in metagenomics data. *PeerJ Computer Science* 3 e104 (2017).
65. Hiseni P, Rudi K, Wilson RC, Hegge FT, Snipen L. HumGut: a comprehensive human gut prokaryotic genomes collection filtered by metagenome data. *Microbiome* 9(1), 165 (2021).
66. Liu S, Li E, Sun Z *et al*. Altered gut microbiota and short chain fatty acids in Chinese children with autism spectrum disorder. *Scientific Reports* 9(1), 287 (2019).*
67. Wang S, Lv D, Jiang S *et al*. Quantitative reduction in short-chain fatty acids, especially butyrate, contributes to the progression of chronic kidney disease. *Clinical Science* 133(17), 1857-1870 (2019).*
68. Strati F, Cavalieri D, Albanese D *et al*. Altered gut microbiota in Rett syndrome. *Microbiome* 4(1), 41 (2016).*
69. Tana C, Umesaki Y, Imaoka A, Handa T, Kanazawa M, Fukudo S. Altered profiles of intestinal microbiota and organic acids may be the origin of symptoms in irritable bowel syndrome. 22(5), 512-e115 (2010).*

Reference annotations

*References of interest, reporting a link between elevated propionate and/or reduced butyrate in samples collected from people suffering from various diseases.

Supplement

The four sections presented below contain detailed information about the four major steps performed during a LAD-based test.

i.Polymerase Chain Reaction, primer and dNTP degradation

The initial step, not exclusively related to LAD, ensures amplification of a relevant target genetic sequence, conventionally 16S rDNA. Most of the sequence variations discriminating closely-related species are located in at least one of the nine hypervariable regions of this gene (V1–V9)[127, 128]. These are bordered and interrupted by phylogenetically conserved sequences, typically serving as templates for universal PCR primers (**Figure S1**).

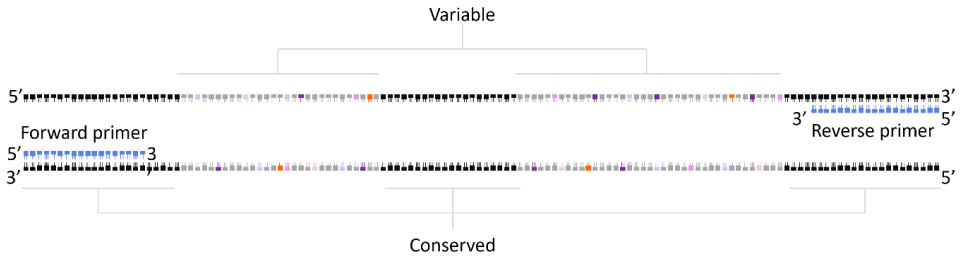


Figure S1. A simplified outline of 16S rRNA gene amplification. With the intention of amplifying this gene from as many different bacteria, the primers are designed to anneal to conserved regions, flanking the variable ones. Depending on the variable regions of interest, primers can be designed to target different flanking conserved regions.

LAD reactions require PCR products free of extendable primers and stable dNTPs. For this reason, after PCR, reactions are treated with a single stranded DNA exonuclease (Exo I) and a thermolabile phosphatase, e.g., shrimp alkaline phosphatase (SAP). The former degrades the remaining PCR primers[129], the latter removes the phosphate groups from dNTPs[130].

ii.Labelling

This marks the initiation of LAD-specific steps of the protocol. Here, following heat-inactivation of the exonuclease and phosphatase enzymes, the ExoSAP-treated PCR products are used as templates for single-nucleotide extension of labelling probes (LPs). LPs are short oligonucleotides (usually 15-30 nucleotides) complementary to signature variable regions. The complementarity ensures specific binding of LPs solely to target bacterial sequences, conditioning their extension (labelling) with a single, quencher-coupled ddNTP. The absence of specific targets precludes extension of corresponding, complementary LPs, due to their poor binding to other 16S rRNA

variant sequences (**Figure S2**). Multiple LPs, targeting multiple bacteria groups, can become labelled simultaneously in a single-tube multiplex reaction.

The specificity of labelling is also ensured by providing the reaction with a single nucleotide type (for example ddCTP). This nucleotide must be complementary to the base opposite and adjacent to the labelling probe 3'-end. This single DNA template position acts as the ultimate marker based on which the probe succeeds or fails at becoming extended; a G-nucleotide at this position on the template ensures extension of the LP with the labelling nucleotide ddCTP.

Dideoxyucleotides (ddNTPs) are employed for labelling since they lack the 3'-OH group on the deoxyribose, preventing further extension once they are incorporated into the probe. In addition, they must carry a dark quencher label. Quenchers are molecules that absorb the emitted light from nearby fluorophores. In summary, the selective labelling of probes with a quencher is the essence of the LAD method.

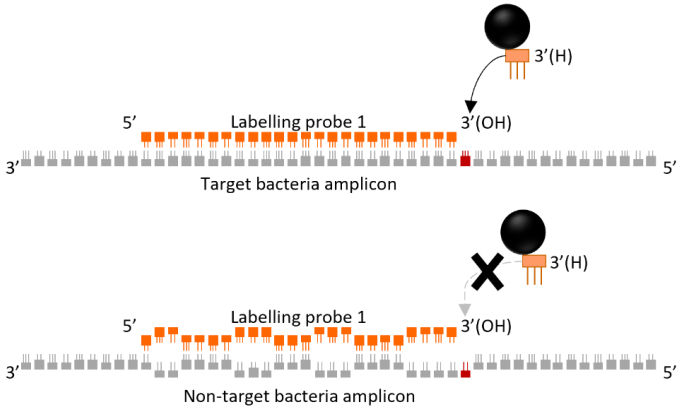


Figure S2. A representation of a LAD labelling reaction. A set of labelling probes (LPs) are added to the ExoSAP-treated PCR product. LPs are designed to be complementary to unique sites within the target bacteria 16S rRNA gene. Only in the presence of targets serving as stable labelling templates can these LPs extend (top example). The extension is designed to occur with a single nucleotide, a quencher-labelled dideoxynucleoside triphosphate (ddNTP). Quencher molecules (presented with a black mark) absorb light from nearby fluorophores, a feature exploited in subsequent steps. For a successful labelling, the type of quencher-labelled ddNTP present in the reaction must be complementary to the base right opposite and downstream the 3'-end of the LP (depicted in red along the template sequence).

Overcoming DNA polymerase infidelity

It was mentioned that ensuring a proper template-directed probe labelling requires that the reaction contains a ddNTP complementary to the base immediately downstream of the LP 3'-end on the complementary strand. This way, even if the probe were to conveniently anneal to a non-

target gene variant, the lack of the complementary template nucleotide would assure that the probe remained unlabeled.

However, the fidelity of modified DNA polymerases used for labelling is often compromised. For example, Hot TERMIpol® DNA Polymerase (Solis Biodyne, Estonia), allows the incorporation of ddGTP opposite a T, and vice-versa, when the complementary ddNTP is absent from the reaction[131]. This issue can be resolved by adding three types of unlabeled ddNTPs (for example ddUTP, ddCTP, ddATP) in addition to the quencher-labelled ddNTP (Q-ddGTP).

iii. Reporting

Labelling the probes with a quencher only gains meaning when complementary fluorophore-coupled probes are added to the reaction. After the completion of the labelling cycles, a separate set of probes, complementary to LPs, carrying a fluorescing molecule on their 5'-ends, are added. The DNA complementarity facilitates juxtaposition of the quencher molecule (3'-end of the labelled LP) with the fluorophore (5'-end of the RP). A graphic depiction of this step is presented in **Figure S3**.

An active DNA polymerase (used for the preceding labelling reaction) frequently acts as a sequence-specific quencher[131]. For this reason, a polymerase-denaturing agent such as sodium dodecyl sulphate (SDS), sarkosyl or heparin are added into the RP mix. This ensures the integrity of the test, eliminating any possible quenching caused from polymerase-related effects.

Apart from their sequences, the difference between various RPs reporting different targets depends on their lengths and the fluorophores they carry. The length *usually* determines the melting temperature (*T*m) of the RP-LP duplex, while the fluorophore defines the channel of detection in a qPCR machine. Registering as many as five quenching events per channel (from 30°C to 70°C, a resolvable signal at 10°C increments) while employing six detection channels renders it possible to report up to 30 targets simultaneously. Up until now, LAD has been proven to generate 20 distinguishable signals utilizing four detection channels (unpublished data).

After the addition of RPs, the reaction is subjected to a melting curve analysis in a qPCR machine. The machine captures the fluorescence values for each channel separately while steadily increasing the temperature of the reaction (usually a 0.5°C increase after each pause of 5 seconds).

All duplexes are expected to have formed and remain stable at room temperature. As the temperature increases, the shorter duplexes dissociate first, while longer, more stable duplexes, dissociate at higher temperatures. RPs in a quencher-labelled duplex start fluorescing after RP-LP duplex separation. In a raw measurements graph, this is viewed as an increase in fluorescence

with an increase in temperature, with distinct duplex *T*ms marking significant rises in fluorescence. In a graph presenting the negative value of the change in fluorescence per temperature (-dF/d*T*), this gets translated into distinct declines of fluorescence with negative 'peaks' at certain *T*ms. The latter graphs intuitively present quenching as a maximum loss of registered fluorescence at a given temperature.

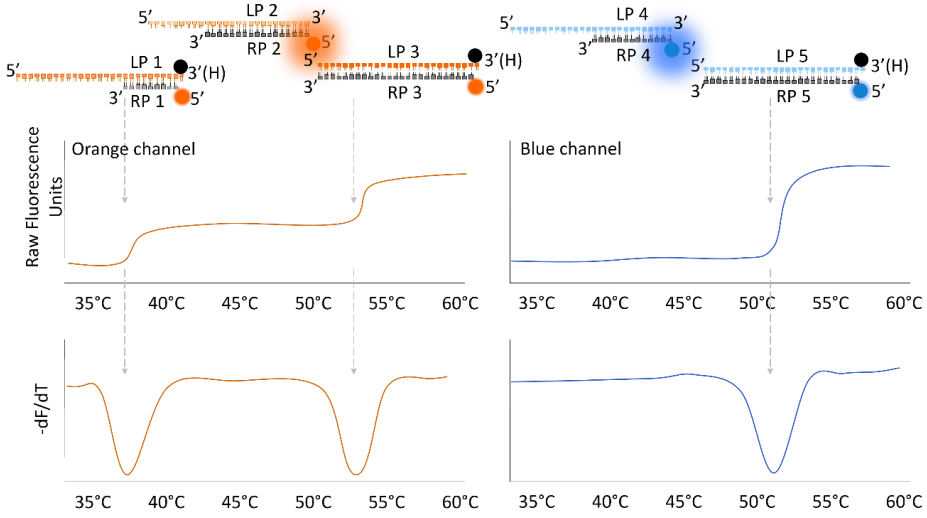


Figure S3. A set of reporter probes (RPs), complementary to LPs is added into the reaction. RPs are coupled to a fluorophore on their 5'-ends, assuring a close physical proximity with the quencher molecule upon LP-RP duplex formation. Quenching of fluorophores can only occur if the complementary LPs have been labelled in the preceding step (labelling probes 1, 3 and 5). RPs differ in length, producing duplexes of diverse sizes for each target. Duplexes of different lengths typically exhibit different melting temperatures (Tm), i.e., shorter duplexes dissociate with lower Tms; higher Tms are required to dissociate long and stable duplexes. Quenching events are displayed as an abrupt increase of fluorescence in raw measurements (top graphs), but as negative peaks in graphs built of negative values of change in fluorescence per temperature (-dF/dT, bottom graphs). The presence of each target produces such a derivative decline at a characteristic Tm in a designated detection channel (for example Escherichia coli – 37 °C in orange channel).

A simplified flow linking the presence/absence of target bacteria with the reporting (or not) of a signal can be viewed as following:

Target bacteria present → Labelled LP → Quenched RP → Temperature-dependent Quenching response (signal)

Target bacteria absent → Unlabeled LP → Fluorescing RP → No Quenching response (no signal)

For calibration and signal extraction purposes, each test requires a set of no template controls (NTC, ≥3 wells), where water instead of gDNA is added into the PCR reaction.

iv.Processing

The processing of signals starts with the collection of -dF/d*T* measurements for all channels from the qPCR instrument. Because of the high noise-to-signal ratio, processing is fundamental. Extracting signals from unprocessed -dF/d*T* curves may often lead to false positives/negatives, mainly because there is a major fluctuation of fluorescence values between wells.

There are two important steps to be performed prior to signal extraction:

- Centering the fluorescence measurements of all wells (samples) within a channel. This with the purpose of minimizing the range of measurements across wells at any given temperature. This step is performed by subtracting the sample mean from each measurement, separately for each well (exemplified in **Table S1**).

  Table S1. An example of centering the derivative data

Temp.	Sample 1	Sample 2	NTC 1	Centered sample 1 (Sample 1 – Mean)	Centered sample 2	Centered NTC 1
30°C	20	60	10	20 - 12.5 = 7.5	60 - 52.5 = 7.5	10 – 2.5 = 7.5
35°C	15	55	4.5	15 - 12.5 = 2.5	55 - 52.5 = 2.5	4.5 – 2.5 = 2
40°C	10	50	0.5	10 - 12.5 = -2.5	50 - 52.5 = -2.5	0.5 – 2.5 = -2
…	…	….		…	…	…
80°C	5	45	-5	5 - 12.5 = -7.5	45 - 52.5 = -7.5	-5 – 2.5 = -7.5
Mean	**12.5**	**52.5**	**2.5**			

- Correcting/flattening the baseline within each channel by subtracting the centered values of each sample with the average NTC centered values (**Table S2**).

  Table S2. Flattening the baseline following the example presented on Table 1

Temperature	Centered NTC 1	Centered NTC 2	Centered NTC 3	Centered NTC mean	Centered sample 1	Corrected sample 1 (Cent. sample 1 – Cent. NTC mean)
30°C	7.5	7	8	**7.5**	7.5	0
35°C	2	3	2.5	**2.5**	2.5	0
40°C	-2	-2.5	-3	**-2.5**	-2.5	0
…						
80°C	-7.5	-7.5	-7.5	**-7.5**	-7.5	0

The effects of processing the data using a real example can be viewed in **Figure S4**.

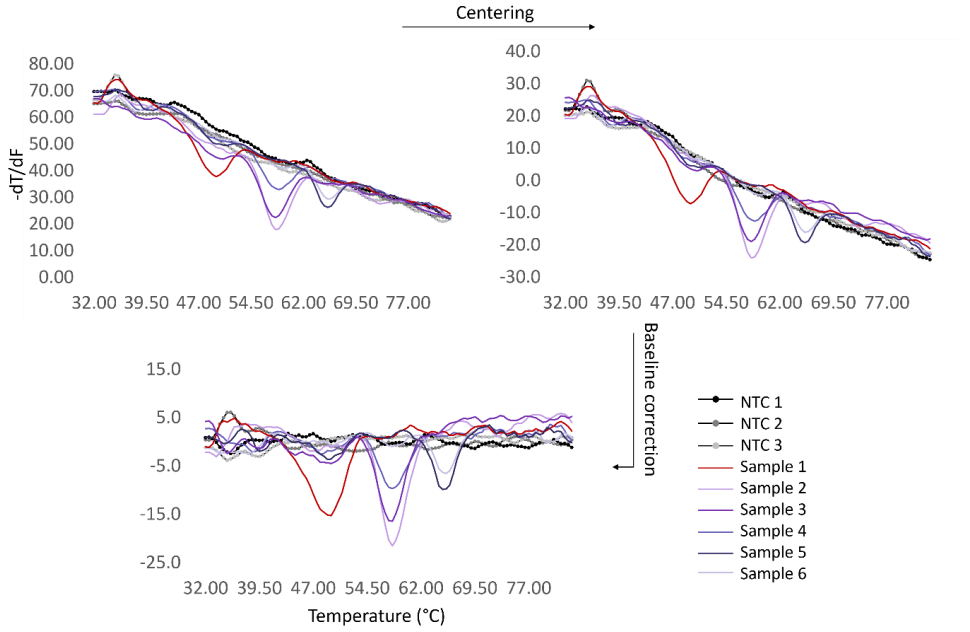


Figure S4. Processing of derivative results. The top left chart depicts a real example of unprocessed derivative curves observed in three no template controls (NTCs) and six samples. After centering the samples, the chart slightly changes form, containing less noise (right panel). The bottom graph shows the results of baseline correction. Here the baseline is flattened, allowing for a better qualitative assessment of the three quenching events (49°C, 58°C and 66°C).

The table harboring processed values is used further for the extraction of the signals. Data at *T*ms where a quenching event is expected to occur are used for further processing (for example measurements at 37°C on orange channel, designated to report the presence of *E. coli*). Average processed NTC values are used to construct a border between noise and signal, using the standard deviation. The threshold is marked at 3 × standard deviations below the mean. Any measurement beyond his threshold is considered to report the presence of the target bacteria (**Figure S5**).

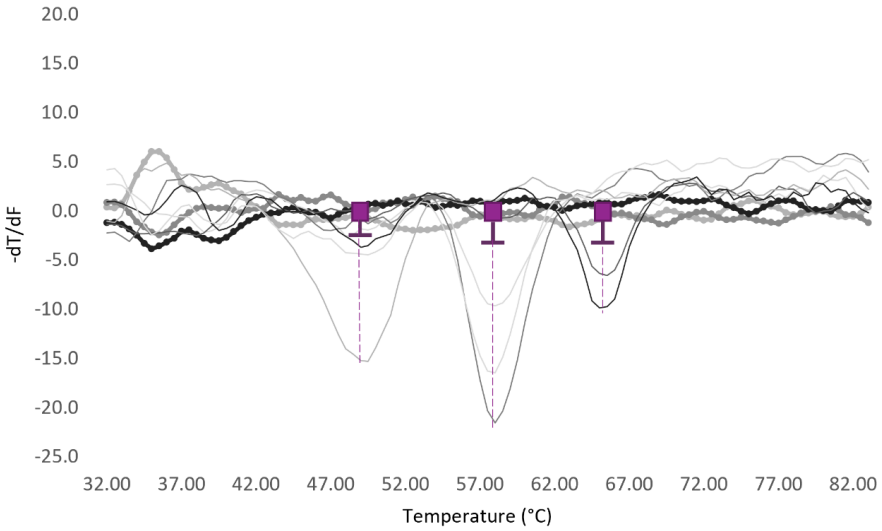


Figure S5. A visual representation of signal extraction. The boxes mark the range of NTC measurements in the Tm where quenching events are expected to occur. The inverted T symbol depicts the threshold below which all measurements are accepted as signals (curves meeting the dashed vertical lines). The threshold is found by subtracting 3 standard deviations from the mean of NTC measurements.

The stronger the quenching, the more negative the derivative fluorescence values are. It may be considered as more intuitive to invert the values by multiplying them with -1. This way, the stronger the signals the higher their value.

Quantitative LAD

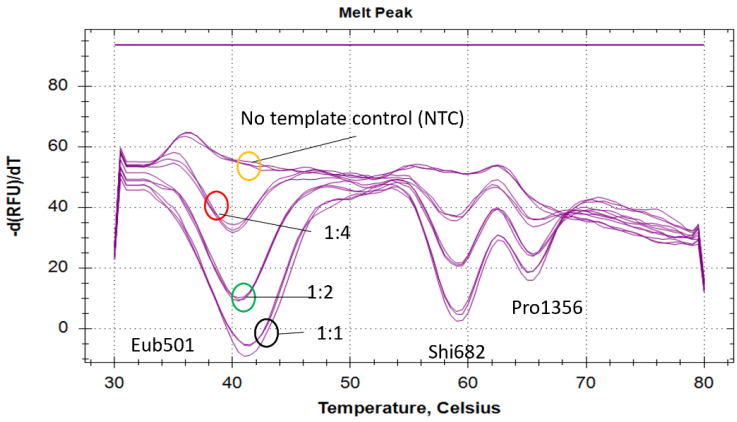


Figure S6. Unprocessed melting curves for four different types of triplicate reactions (1:1 PCR products, 1:2 diluted PCR products, 1:4 diluted PCR products and No Template Control (NTC) reaction). With a decrease of PCR product concentration (lower template abundance) the signal strength becomes weaker for three of the probes presented here (Eub501, Shi682 and Pro1356). No signal is observed in NTC reactions.

ISBN: 978-82-575-1889-9
ISSN: 1894-6402

Norwegian University
of Life Sciences

Postboks 5003
NO-1432 Ås, Norway
+47 67 23 00 00
www.nmbu.no