The human microbiota in early life
Initial colonization and development
Submitted: January 29, 2016
The human microbiota in early life
Initial colonization and development

PhD Thesis

Supervisor:
Professor Søren J. Sørensen, PhD
Assistant professor Waleed Abu Al-Soud, PhD

Section of Microbiology
Institute of Biology
University of Copenhagen

Assessment committee
Associate professor Anders Priéme, PhD (chair)
Department of Biology, University of Copenhagen
Associate professor Henrik Bjørn Nielsen, PhD,
Department of Systems Biology, Technical University of Denmark
M. Pilar Francino, PhD
Department of Genomics and Health, FISABIO-Public Health, Valencia

This PhD thesis has been submitted to the PhD School of Faculty of Science, University of Copenhagen, Denmark, January 29 2016.
Cover Page
Illustrations by Heather Spears of infants with asthma.
Thanks to Hans Bisgaard for allowing me to use his private illustrations.
PREFACE

The research presented in this thesis is the result of my years as PhD student. The work has been carried out at the Section of Microbiology, Department of Biology, Faculty of Science, University of Copenhagen. During my thesis, I worked, for six months in 2014, as part of professor’s William Cookson and professor’s Miriam Moffat group at the National Heart and Lung Institute, Imperial College, London, United Kingdom, on sequence-based identification of Streptococcus (work not included in the thesis).

The Faculty of Science, University of Copenhagen and the Lundbeck Foundation funded this study. Besides the three enclosed manuscripts, part of the work in this thesis has been presented at the Wellcome Trust Scientific Conference: Exploring Human Host-Microbiome Interactions in Health and Disease in Cambridge, UK, 2014 and at the 25th European Congress of Clinical Microbiology and Infectious Diseases (ECCMID) in Copenhagen 2015.

I would like to thank my two supervisors Søren J. Sørensen and Waleed Abu Al-Soud, from the Section of Microbiology. Søren, thank you for giving me the opportunity to be part of this amazing project, for your guidance throughout the project you, and for always supporting, questioning, or challenging me to unlock the full potential of this project. Waleed, thank you for the unwavering support and help to understand and interpret the massive amounts of data in this project.

This project could not have been done without the help and support of the entire Section of Microbiology. Karin Vestberg, thank you for answering my endless stream of questions about the work in the laboratory and for always offering your help, time, and hands. April Cockburn, though you are no longer in Denmark, I owe you my gratitude for the help and for setting up the protocols used in this project. I would like to thank Martin A. Hansen and Asker D. Brejnrod for doing the rough work of turning the raw sequence data into a format I have been able to understand. Michael Roggenbuck, I would like you to thank you for the patience and help you provided when taking me through the process of preparing my first manuscript. I would also like to thank the two master students I have helped supervise, Christina Balle, Jeanne Williams, and Jakob Russel, thank you for your help in the lab and for always keeping me on my toes.

I must thank the rest of the Section of Microbiology for creating a workplace full of smiles and social activities, without you I doubt I would have kept my sanity.
I have to send my thanks to Hans Bisgaard and the COPSAC research group, in particular the microbiome group: Jakob Stokholm, Jonathan Thorsen, Johannes Waage, and Morten A. Rasmussen. Without your work, the high quality of the data and samples, and the constant flow feedback and ideas you have provided, this project would never have reached as far as it has. Hans, I am grateful that you decided to include the Section of Microbiology in the analysis of samples from a cohort as large and unique as COPSAC2010.

I would like to thank William Cookson and Miriam Moffat for inviting me to spend six months at their Department. A special thanks to Michael, Phil, John, Ruth, Hima, Stef, and Spyros for their support and company during my time in London. Thanks to Augustinus Fonden, Oticon Fonden, Knud Højgaards Fond, and Lundbeckfonden for financial support during my time in London.

I would like to thank Karen A. Krogfelt from SSI, for always listening, pushing me to reach further, and for giving me a job in her Lab, even before I finished my bachelor degree. Without that job, I doubt I would have been chosen for this PhD. Thanks to Susanne Schjørring for her tremendous work with the culturing, identification and organization of the samples from the COPSAC2010 cohort. Thanks to her whole group for always providing a breathing space and support whenever needed.

I would like to thank my family and friends for their support, for always believing in me and for making me talk and think of things outside the field of this PhD. Special thanks to the members of Badeklubben af 2012 who have been crazy enough to go swimming with me every winter during my PhD.

This thesis would not have become readable without the help from several volunteer editors. Thank you Annelise, Sten, Sarah, Cecillie, Ulrik, Georgia, and Mom!

Last but certainly not least, I would like to thank my girlfriend Γεωργία Ρούση (Georgia Rousi) for her support and understanding.

Martin Steen Mortensen, Copenhagen 2016
ABSTRACT

The bacteria that colonize the human body, our microbiota, can influence our health, both positively and negatively. The importance and functions of the microbiota in our intestinal tract have been the focus of several research projects and are widely published. However, there are great gaps in our knowledge concerning microbiota composition, development and function in other areas of human body. Lack of knowledge about the microbiota development in the airways is an example of such a deficiency.

The work presented in this PhD thesis is based on the vast sample collection of the COPSAC2010 cohort, with 700 mother-infant pairs. The objectives were to perform a detailed examination of the mothers’ vaginal microbiota, describe the early composition and development of the microbiota in the airways of their infants, and determine whether the infants’ microbiota are affected by that of their mothers or not.

Manuscript I examines the composition and stability of vaginal microbiota, as well as how the mothers’ microbiota contribute to the early bacterial colonization of their infants. In this study, we first confirmed that the vaginal microbiota of the women in the COPSAC2010 cohort represent the already well-defined community state types. In addition, we showed that for most women the vaginal microbiota at week 24 of pregnancy is similar to the microbiota twelve weeks later at week 36. The manuscript also contains a novel description of how mothers’ vaginal microbiota has affected the microbiota of their infants one week after birth. The results show that delivery mode is important for bacterial transfer from mother to infant.

Manuscript II focuses on the microbiota in lower airways aspirates, collected from infants one week, one month and three months after birth. The manuscript explores the differences in the microbiota composition at the three time points, examining as well the time dependent changes of each infant separately. One week after birth, Staphylococcus, traditionally associated with skin microbiota, is dominating the microbiota, but as time passes, bacteria normally found in the airways (e.g. Streptococcus and Moraxella) become increasingly dominant. By defining the core microbiota for each infant, the manuscript shows that 69% of the microbiota, three months after birth, represent bacteria that were present at both one week and one month after birth. Lastly, the manuscript describes how the microbiota can be separated into five distinct pneumotypes: four having a single dominating genus and one without a common defining genus.
The last manuscript, **Manuscript III**, compares the microbiota descriptions obtained by classical identification using culturing and high throughput sequencing of amplified 16S rRNA. Weaknesses and strengths of both methods are presented; the comparison has shown that sequencing gives a more detailed representation, as more bacteria, both types and number, can be identified. However, the sequencing approach used in this study lacks of resolution. The manuscript concludes that sequencing is almost ready for clinical use, as disadvantages of the method can be solved by using a combination of techniques that have been shown to work separately.

The results presented in the manuscripts of this PhD thesis have strengthened current knowledge of our microbiota and have contributed with novel research to improve the understanding of the microbiota development during the early period of life.
RESUME

Bakterier som koloniserer menneskekroppen, mikrobiota, kan påvirke menneskets helbred, både positivt og negativt. Vigtigheden og funktionerne af tarmens mikrobiota har været fokus for mange forskningsprojekter og er beskrevet i mange publikationer, men der er store mangler i vores viden om mikrobiotaens komposition og udvikling andre steder i kroppen. Den begrensede viden om udviklingen af mikrobiotaen i luftvejene er et eksempel på sådanne mangler.

Arbejdet præsenteret i denne Ph.d.-afhandling er baseret på analyse af den omfattende prøvesamling fra COPSAC2010 kohorten, som omfatter 700 moder-barn par. De overordnede mål har været at lave en detaljeret undersøgelse af mødrenes vaginale mikrobiota, beskrive den tidlige sammensætning og udvikling af mikrobiotaen i børnenes luftveje, samt at bestemme om børnenes mikrobiota er påvirket af deres mødres mikrobiota.


mikrobiotaen kan opdeles i fem særskilte pneumotyper: fire domineret af hver sit genus, og en som ikke har et fælles beskrivende genus.

Det sidste manuskript, **Manuskript III**, undersøger hvor godt resultaterne fra to forskellige metoder, klassisk identifikation baseret på dyrkning samt høj-kapacitet sekvensering af amplificeret 16S rRNA, stemmer overens. Styrker og svagheder af begge metoder bliver præsenteret. Sammenligningen viser at sekvensering giver en mere detaljeret beskrivelse, da både flere typer og et større antal bakterier identificeres. Imidlertid, mangler den brugte sekvenseringsmetode den fornødne opløsning. Manuskriptet konkluderer at sekvensering er næsten klar til diagnostisk brug, da svaghederne ved metoden kan løses ved at kombinere flere teknikker som hver for sig løser en del af problemerne.

Resultaterne præsenteret i denne Ph.d.-afhandling har styrket den nuværende viden om menneskets mikrobiota og bidraget med ny forskning til forbedring af forståelsen af mikrobiotaens udvikling i den tidlige periode af menneskets liv.
# TABLE OF CONTENT

PREFACE ........................................................................................................................................... I

ABSTRACT ......................................................................................................................................... III

RESUME ............................................................................................................................................. V

TABLE OF CONTENT ...................................................................................................................... VII

TABLE OF FIGURES ........................................................................................................................ IX

ABBREVIATIONS ............................................................................................................................. XI

1 INTRODUCTION ............................................................................................................................ 1

1.1 Methods for studying the human microbiota .................................................................................. 1
  1.1.1 Culture based identification ........................................................................................................ 1
  1.1.2 PCR based identification ............................................................................................................ 2
  1.1.3 Sequencing ................................................................................................................................ 3
  1.1.4 Analysis of sequencing data ....................................................................................................... 4

1.2 Vaginal microbiota ......................................................................................................................... 5

1.3 Gut microbiota ............................................................................................................................... 7
  1.3.1 Development of the gut microbiota ............................................................................................. 8
  1.3.2 Gut microbiota in health and disease ......................................................................................... 9
  1.3.3 The hygiene hypothesis ........................................................................................................... 10

1.4 Airway microbiota .......................................................................................................................... 11
  1.4.1 Airway microbiota and disease ................................................................................................. 12

1.5 Cohort studies ................................................................................................................................ 14
  1.5.1 Copenhagen prospective studies on asthma in childhood ....................................................... 15

1.6 Objectives ..................................................................................................................................... 16

2 LIST OF MANUSCRIPTS ............................................................................................................. 17
3 SUMMARY OF RESULTS ................................................................. 19

4 DISCUSSION .................................................................................. 23
  4.1 Transfer of vaginal microbiota to infants ..................................... 23
  4.2 The airways microbiota .............................................................. 24
    4.2.1 Defining pneumotypes ....................................................... 26
  4.3 Culturing or sequencing ............................................................ 26
    4.3.1 De-novo OTU clustering of sequencing reads ....................... 27
    4.3.2 Closed reference OTU picking of sequencing reads ............. 27
    4.3.3 Sequencing in a clinical setting ......................................... 28

5 CONCLUSION ................................................................................. 29

6 PERSPECTIVES ............................................................................. 31

7 REFERENCES ............................................................................... 33

8 APPENDIX .................................................................................. 39
  Manuscript I .................................................................................. 41
  Manuscript II ................................................................................ 57
  Manuscript III ............................................................................... 73
TABLE OF FIGURES

Figure 1: Residue variability within the 16S rRNA molecule, showing the mean frequency of the most common residue in a 50 base window. Grey bars indicate the location of the variable regions. Adapted from (Ashelford et al., 2005). ................................................................. 2

Figure 2: Schematic representation of the possible analysis, work-flow and results from studies using approaches based on DNA, RNA, metabolites, proteins or bacterial isolates. The colored area contains the bioinformatics steps. (Weinstock, 2012) ............................................................................................................. 3

Figure 3: Vaginal microbiota fluctuations over 16 weeks. Representation of the vaginal microbiota composition in a non-pregnant, reproductive age, healthy woman. The red dots represent days where she was menstruating. Adapted from (Gajer et al., 2012). ........................................... 6

Figure 4: Gut microbiota in adults and the development in early life. a) Microbial composition and concentration in the different sections of the gastrointestinal tract show increasing density as well as a shift in composition. b) Factors influencing the gut microbiota development and composition in early life and the expected members of the gut microbiota over time. SFB: segmented filamentous bacteria. (Verdu et al., 2015) ............................................................................................................ 8

Figure 5: Graphical representation of the number of articles published (by year) found in the Pubmed database when searching for "lung microbiome" OR "lung microbiota". The number of articles per year is plotted on the left-hand y-axis and the total number of publications on the right-hand y-axis. Only articles published in the period 2011-2015 are included. ............................. 11

Figure 6: Three factors affect the composition of the airways microbiota: microbial immigration, microbial elimination, and the growth rates of present microbes. In healthy individuals the immigration and elimination is balanced, with very little growth occurring. During acute disease the microbial elimination is compromised and the microbiota composition is determined by growth within the airways. (Dickson et al., 2014b) ........................................................................................................ 13
ABBREVIATIONS

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>BV</td>
<td>Bacterial vaginosis</td>
</tr>
<tr>
<td>CF</td>
<td>Cystic fibrosis</td>
</tr>
<tr>
<td>COPD</td>
<td>Chronic obstructive pulmonary disease</td>
</tr>
<tr>
<td>COPSAC</td>
<td>Copenhagen Prospective studies on asthma in childhood</td>
</tr>
<tr>
<td>CST</td>
<td>Community state type</td>
</tr>
<tr>
<td>DGGE</td>
<td>Denaturing gradient gel electrophoresis</td>
</tr>
<tr>
<td>HMP</td>
<td>Human microbiome project</td>
</tr>
<tr>
<td>HTS</td>
<td>High throughput sequencing</td>
</tr>
<tr>
<td>LAB</td>
<td>Lactic-acid bacteria</td>
</tr>
<tr>
<td>MetaHIT</td>
<td>Metagenomics of Human Intestinal Tract</td>
</tr>
<tr>
<td>OTU</td>
<td>Operational taxonomical unit</td>
</tr>
<tr>
<td>PCR</td>
<td>Polymerase chain reaction</td>
</tr>
<tr>
<td>PT</td>
<td>Pneumotype</td>
</tr>
<tr>
<td>RDP</td>
<td>Ribosomal database project</td>
</tr>
<tr>
<td>SSCP</td>
<td>Single-strand-conformation polymorphism</td>
</tr>
<tr>
<td>T-RFLP</td>
<td>Terminal restriction fragment length polymorphism</td>
</tr>
</tbody>
</table>
INTRODUCTION

1 INTRODUCTION

The human body is host to a huge amount of bacteria, roughly 2 kg in biomass, representing as many bacterial cells as there is human cells in the body (Sender et al., 2016). The term ‘microbiocenosis’ was adopted to characterize the typical microflora (microbiota) of an organ or part of an organ as early as 1958 (Haenel, 1961), indicating a general interest in the study of the human microbiota, focused on the gut microbiota. Joshua Lederberg introduced the term ‘microbiota’ in 2001 to describe “… the ecological community of commensal, symbiotic, and pathogenic microorganisms that literally share our body space…” (Lederberg & McCray, 2001).

In the present work, I will use the terms microbiota and microbiome, as defined by Turnbaugh et al. (2007) where microbiota refers to the microbes and microbiome refers to the collective genepool of the microbiota.

Microbiota are found in all areas of the human body with external access: Our skin, oral cavity, airways, intestinal tract, urinary tract, hair, and genitals. Each of these human body compartments has a distinct microbiota adapted to the physical, physiological and nutritional conditions, e.g. available nutrients, humidity, temperature, pH, and physical stress (Costello et al., 2009).

1.1 Methods for studying the human microbiota

Bacteria have been studied for more than 300 years, since Leeuwenhoek (1677) as the first observed “animalcules”, i.e. bacteria and protozoa, in water by using a microscope. He was also the first to describe the human bacteria, when he studied his own saliva and phlegm (van Leeuwenhoek, 1708). Since that time there has been tremendous progress in the study of bacteria, driven by the development of new methods and technologies. One of the great contributors to this development was Robert Koch, who was the first to publish photos of bacteria (1877), develop solid culture media (1881), and he formulated his postulates for proving infectious properties in bacteria.

1.1.1 Culture based identification

With the solid culture media it was possible to spread a sample on a plate, isolate pure cultures, and then identify the bacteria based on cultural and cellular morphology, nutritional requirements, and cell wall structure. In the 1960s, the interest and progress within the study of
human microbiota leapt forward as result of several groups of scientists, who worked to shape our understanding of the gut microbiota (for references see Savage, 2001). Culturing is still a relevant technique, especially in clinical settings. The method is cheap and enables isolation and identification of pure strains. The disadvantage is that the method is labor intensive, relatively slow, has low reproducibility, and is limited to bacteria that can be cultured.

1.1.2 PCR based identification

The next step in bacterial studies came with the invention of polymerase chain reactions (PCR) (Mullis & Faloona, 1987), which led to the development of improved identification techniques, some of which could be used to describe the bacterial composition of a sample without prior culturing. Most of these novel methods take advantage of the nature of 16S rRNA, which is present in all bacteria. 16S rRNA consists of highly conserved regions and nine variable regions (Figure 1), can be amplified directly from extracted DNA, and the sequence of 16S rRNA from most bacterial species is available in public databases (Woese, 1987).

PCR based DNA Microbial fingerprinting methods can provide a comprehensive assessment of the microbial community without a need for prior culturing and provide an overall profile of the microbial community (microbial diversity). Examples of DNA microbial fingerprinting methods are denaturing gradient gel electrophoresis (DGGE), terminal restriction fragment length polymorphism (T-RFLP) and single-strand-conformation polymorphism (SSCP). DGGE was developed to identify bacteria based on the pattern produced by the melting behavior (GC content and distribution) of the amplified 16S rRNA generated by electrophoresis at high temperature in an acrylamide gel containing gradient concentration of denaturing agents (Muyzer et al., 1993). T-RFLP was developed to profile microbial communities based on the position of a restriction site closest to a labeled end of an amplified gene (Christensen et al., 2004). SSCP uses selective digestion (lambda exonuclease) of amplified 16S rRNA, followed by gel

![Figure 1: Residue variability within the 16S rRNA molecule, showing the mean frequency of the most common residue in a 50 base window. Grey bars indicate the location of the variable regions. Adapted from (Ashelford et al., 2005).](image)
INTRODUCTION

Electrophoresis to create a unique fingerprint of fragment sizes bacterial species (Schwieger & Tebbe, 1998). These methods, in addition to other similar methods, are a huge improvement from culture dependent methods, as they are not limited to culturable bacteria. The main disadvantages with these methods is the difficulty to analyze very complex samples and that only abundant microbial taxa are detected. SSCP cannot separate much more than 10 strains in one sample (Schwieger & Tebbe, 1998), and using DGGE, it is not possible to identify the bacteria in a community as complex as a fecal sample (Bisgaard et al., 2011).

1.1.3 Sequencing

With the development of high throughput sequencing (HTS), 454 pyrosequencing (Ronaghi et al., 1996) and Illumina sequencing (Bentley et al., 2008), the tools available for studying the human microbiota, have taken a new major step forward. For the analysis of bacterial communities, HTS is used in two separate ways; 16S rRNA sequencing uses PCR amplification, of one or more variable regions, in order to identify the microbiota present and their relative abundances. Shotgun sequencing uses sequencing of total DNA to describe the microbiome in order to provide information about the functional potential of the microbiota, i.e. which nutrients the microbiota can process and which molecules it can produce. The advantage of sequencing the amplified 16S rRNA is the depth of the investigation, that even bacteria representing less than 0.01% of the bacterial abundance can be detected with a reasonable number of sequences from each sample at a low cost per sample. Shotgun sequencing on the other hand provides more information, as it provides a view of the structure as well as the functional potential of a community, but despite needing 50 times more sequencing data per sample, the DNA of very low abundant bacteria will still not be covered. The overall workflow and outcome are shown in Figure 2 (Weinstock, 2012).
A comparison of shotgun sequencing and 16S rRNA sequencing has shown that the latter identifies more bacterial species in each sample, that there are significant differences between the results, and that studies using different methods are not directly comparable (Shah et al., 2011).

Currently there are two large consortia working on describing the human microbiome; Metagenomics of Human Intestinal Tract (MetaHIT) and the Human Microbiome Project (HMP). MetaHIT is an EU project aiming to create a reference database for the human intestinal microbiome and relate specific genes to disease (Dusko Ehrlich & MetaHIT consortium, 2010). HMP is an American project aiming to develop a reference set of 3,000 microbial genomes and 16S rRNA gene database generated by sequencing 16S rRNA gene amplicons from 15 and 18 body sites (for men and women, respectively). Their goal is to providing initial answers to the questions of whether there is a "core" microbiome at each site (NIH HMP Working Group et al., 2009).

1.1.4 Analysis of sequencing data

No matter which method is used to sequence a bacterial community, the sequencing itself must be followed by a bioinformatic analysis before any interpretation is possible. Oulas et al. (2015) published an in-depth review of the steps and tools used for both shotgun sequencing and 16S rRNA sequencing.

The first step in processing sequencing data is quality control. This includes quality filtering and trimming of reads, i.e. removing low quality reads or read ends, respectively, and removing chimera sequences. The tools most often used for this is QIIME (Caporaso et al., 2010) and MOOTHUR (Schloss et al., 2009). For 454 sequencing data an initial denoising step is required to filter out low intrinsic errors (Oulas et al., 2015). If more than one sample have been sequenced together, the next step is demultiplexing, where reads are separated into their respective samples, based on their unique one or two 6-8 base barcodes, or indices. In the subsequent steps, the analysis differs depending on the type of sequencing (Figure 2). In the following section, I will focus on 16S rRNA sequencing.

The next step in the analysis is operational taxonomical unit (OTU) picking and taxonomical assignment. 16S rRNA sequencing reads are clustered into OTUs depending on their sequence identity. It is generally accepted that reads with an identity higher than 97% correspond to the
same species (Stackebrandt & Goebel, 1994). While that is for the entire 16S rRNA sequence, a 97% identity threshold is the most commonly used, even when only one or a few variable regions are sequenced. This step is performed using one of three methods; de-novo OTU clustering, closed reference OTU picking, or open reference OTU picking. With de-novo OTU clustering the reads are clustered into OTUs and then assigned taxons based on their representative sequences. With closed-reference OTU picking the sequences are matched against a reference database of known sequences, sequences being insufficiently matched are discarded. The last method is open-reference OTU picking, where all reads are first matched as in closed-reference followed by a de-novo clustering of all discarded reads. There are several open-access databases containing 16S rRNA sequences and the most commonly used are Greengenes (DeSantis et al., 2006), SILVA (Quast et al., 2013), and Ribosomal database project (RDP) (Cole et al., 2014).

The output of OTU picking and taxonomical assignment consists of three tables:

- An OTU table with the number of reads from each OTU in the sequenced samples.
- A taxonomic table with the defined taxonomy for each of the identified OTUs,
- A table containing metadata on the sequenced samples.

These tables are the input for all further analysis of the sequencing results.

1.2 Vaginal microbiota

The vaginal physiology is different from most other parts of the human body as there is no physical movement and the exposure to the outside is minimal. Together with high moisture and stable temperature, the vagina provides an environment very vulnerable to bacterial infection, bacterial vaginosis (BV). BV increases the risk of vaginal yeast infections, sexually transmitted diseases, and urinary tract infections (Reid et al., 2006; Abbai et al., 2015). Studies of the vaginal microbiota have shown clear correlations between increased diversity and risk of bacterial vaginosis (Hummelen et al., 2010). This is contrary to other areas of the human microbiota, as increased bacterial richness and diversity in the gut and upper airway have been associated with improved health (Bisgaard et al., 2011; Chang et al., 2008; Cho & Blaser, 2012). Especially during women’s reproductive years, the vaginal microbiota is highly dynamic. Some changes in structure and composition are caused by external factors such as contraceptive hormone treatment, sexual behavior, use of lubrication, and hygiene practices (Fashemi et al., 2013; Gupta
et al., 2000; Schwebke et al., 1999). The most significant disturbances of the vaginal microbiota follow the menstrual cycle (Eschenbach et al., 2000; Gajer et al., 2012; Hickey et al., 2013), which is characterized by fluctuating levels of the female sex hormones estrogen and progesterone. It has been shown that high levels of female sex hormones increases stability, while menses provides the largest disturbance of the vaginal microbiota (Figure 3) (Ma et al., 2012; Gajer et al., 2012).

![Figure 3: Vaginal microbiota fluctuations over 16 weeks. Representation of the vaginal microbiota composition in a non-pregnant, reproductive age, healthy woman. The red dots represent days where she was menstruating. Adapted from (Gajer et al., 2012).](image-url)

A healthy vaginal microbiota is most often dominated by lactic-acid bacteria (LAB) of the genus Lactobacillus (Hyman et al., 2005; Tärnberg et al., 2002; Zhou et al., 2004). Lactobacillus species provide a healthy vaginal microbiota by creating a hostile environment that restricts the invasion of pathogenic bacteria and fungi by secreting lactic acid (causing a low vaginal pH (4.0-4.5)), antimicrobial compounds (e.g. bacteriocins and hydrogen peroxide (H$_2$O$_2$)) and anti-adherence molecules (Hummelen et al., 2010; O’Hanlon et al., 2013; Ravel et al., 2011). Studies of the vaginal microbiota of healthy women during their reproductive years, have described five types of vaginal communities, referred to as community state types (CSTs) (Hyman et al., 2005; Ling et al., 2010; Zhou et al., 2007). Lactobacillus species dominate four of these CSTs, most commonly L. crispatus (CST I), L. gasseri (CST II), L. iners (CST III), or L. jensenii (CST V) (Ravel et al., 2011). CST VI, the last CST, is characterized by lacking Lactobacillus species and the community is composed of facultative and strictly anaerobic bacteria, including Prevotella,
**INTRODUCTION**

*Atopobium vaginae, Gardnerella, Dialister, Sneathia and Megasphaera* (Chaban et al., 2014; Drell et al., 2013; Ravel et al., 2013). The bacteria dominating CST IV have been linked with increased risk of BV (Fredricks et al., 2005; Ling et al., 2013; Verhelst et al., 2004). During pregnancy women experience a high level of circulating estrogen and progesterone, and the vaginal microbiota have been shown to be less diverse and more stable than in non-pregnant, reproductive age women (Hernandez-Rodriguez et al., 2011; Romero et al., 2014; Walther-António et al., 2014). Perturbations to the vaginal microbiota during the course of pregnancy have been implicated in adverse pregnancy outcomes, including preterm labor and low birth weight (Carey & Klebanoff, 2005; Genc et al., 2004; Hillier et al., 1995).

### 1.3 Gut microbiota

The gut microbiota is the most thoroughly studied microbial community in the human body. As fecal samples are accepted as representative of the microbiota in the gut, sampling is simple and noninvasive, and historically the microbiota of the intestines have always been considered important for human health (Haenel, 1961). The bacterial density increases through the gastrointestinal tract, as the content becomes less acidic. From $10^1$ cells/ml in the stomach, the density grows to more than $10^{12}$ cells/ml in the colon (Figure 4a) (Verdu et al., 2015), which is the highest density of any part of the human microbiota. There are approximately $3.6 \times 10^{13}$ bacterial cells in the gut, as many as the number of human cells in our entire body (Sender et al., 2016).

The microbiota of the gut varies depending on general life-style and food culture. In Europeans and Americans, consuming a typical mixed Western diet, the gut microbiota is dominated by the Phyla; Firmicutes ($\sim 60–65\%$), Bacteroidetes ($\sim 20–25\%$), Proteobacteria ($\sim 5–10\%$), and Actinobacteria ($\sim 3\%$), these phyla make up over 97% of the bacterial abundance in the gut (Human Microbiome Project, 2012; Turnbaugh et al., 2009a; Costello et al., 2009; Rosenbaum et al., 2015). Yatsunenko et al. (2012) have shown that a Western gut microbiota is clearly distinguishable from a non-Western gut microbiota, in Malawians and Amerindians. The difference were mostly driven by an increased abundance of *Prevotella* in the non-Western group. This supports findings that Actinobacteria and Bacteroidetes are more dominant in the gut microbiota of African children than in the gut of European children, where Firmicutes and Proteobacteria were more dominant (De Filippo et al., 2010).
Despite the diet-based differences, a study, under the framework of the MetaHIT Project, used metagenomes of fecal samples from 39 individuals (22 from Danish, French, Italian, and Spanish individuals, 13 Japanese, and 4 American) to define three robust microbiome types, based on the clustering pattern of the samples. These microbiome types, called “Enterotypes”, were each characterized by high abundance of one indicator genus: \textit{Bacteroides}, \textit{Prevotella}, and \textit{Ruminococcus}, respectively (Arumugam et al., 2011). Rob Knight’s group has questioned the validity of enterotypes as a concept, as the clear separation of the enterotypes could be an artifact of the data analysis; each indicator genus can be present in a range of abundances and not only highly abundant or absent (Knights et al., 2014).

1.3.1 Development of the gut microbiota

The fecal microbiota is not present from birth, but develops rapidly during the first months and years of life. The initial colonizers are facultative anaerobic bacteria, generally associated with the skin microbiota. As these bacteria use the available oxygen, obligate anaerobic bacteria will populate the gut and within the first week the bacterial density reaches a stable level of $\sim 10^{10}$ cells/g (Figure 4b) (Stark & Lee, 1982; Palmer et al., 2007; Jost et al., 2012; Adlerberth et al.,...
The first period of life, where infants do not eat solid foods, the microbiota is dominated by *Lactobacillus*, *Prevotella* and other *Bifidobacterium* (Biasucci et al., 2008), with *Bifidobacterium* and *Lactobacillus* being more dominant in breast-fed infants, while formula-fed infants have an increased abundance of *Clostridia* and *Bacteroides* (Bergström et al., 2014). The delivery mode affects this very early development, infants born by caesarean section are less colonized by *Bifidobacteria* and *E. coli*, while having a higher prevalence of *Clostridium* and Enterobacteriaceae other than *E. coli* (Penders et al., 2006; Adlerberth et al., 2007). This correlates with the initial bacterial exposure during and after birth, where vaginal born infants are exposed to their mothers vaginal and fecal flora, and infants delivered by caesarean section are initially exposed to the skin microbiota of the individuals that touched and handled them (Dominguez-Bello et al., 2010). Several studies have shown that the differences in microbiota associated with delivery mode persist for at least 6-12 months after birth (Adlerberth et al., 2007; Grönlund et al., 1999).

The next factor influencing the development of the gut microbiota is the introduction of solid food and later weaning. New food provides a change in nutrition for the microbiota, which leads to the introduction of Bacteroidetes. A case study has shown that this change followed directly after the introduction of formula and table foods (Koenig et al., 2011). Another study showed that breast-fed infants have similar microbiota, but after introducing solid foods it is possible to distinguish between European and African children based on their microbiota (De Filippo et al., 2010).

### 1.3.2 Gut microbiota in health and disease

The gut microbiota perform a variety of functions that are important for our health (symbiosis), but it may also affect our health adversely (dysbiosis) and have been connected to several diseases (Ott & Schreiber, 2006; Larsen et al., 2010; reviewed by Tilg & Kaser, 2011). A well-balanced gut microbiota is therefore critical for the human health state; it protects against pathogens, digests and produces nutrients, and trains our immune system (reviewed by Kelly et al., 2007; and Tilg & Kaser, 2011).

The gut microbiota protect against pathogens through competition for nutrients and space and by interacting with the innate and adaptive immune system. Jarchum and Pamer have made a thorough review of the specific mechanisms involved in how the commensal microbiota regulates the innate and adaptive immune system (Jarchum & Pamer, 2011).
Our nutritional uptake is affected by our gut microbiota. Studies have shown that bacterial diversity and the fraction of Bacteroidetes relative to Firmicutes is higher in lean than in obese individuals (Turnbaugh et al., 2009b, 2006). The importance of the gut microbiota’s role in energy harvest and associated diseases was confirmed when Vrieze et al. increased insulin sensitivity in individuals with metabolic syndrome by fecal transplantation from lean donors (Vrieze et al., 2012). Production of short chain fatty acids is another mechanism with which the gut microbiota affect metabolism and energy uptake. Rosenbaum et al. (2015) published an in-depth review of the gut microbiota’s role for energy uptake and obesity.

The early gut microbiota have been associated with the maturation of our immune system and development of allergies later in life, and bacterial diversity within the first year of life has been inversely correlated with the risk of asthma at school age (Wang et al., 2008; Bisgaard et al., 2011). This role of our microbiota in the maturation of the immune system and the risk for immune related diseases, such as asthma and atopy, has been of special interest for the cohort studied by Bisgaard et al. (2013).

1.3.3 The hygiene hypothesis

The idea that our environment determines the risk of allergy was first suggested by David P. Strachan (1989), and he showed an inverse correlation between the number of children in the household at age 11 and hay fever in children at age 11 and 23. In 1998, Wold (1998) revised the hygiene hypothesis based on several studies of the correlation between specific infections and atopy, and she showed that the correlation was more likely due to differences in the gut microbiota. This result has been reproduced in mice, where it was found that lung microbiota in early life promotes tolerance to allergens at a later age (Gollwitzer et al., 2014). The general idea that exposure, or lack thereof, at an early age could be determining for atopic diseases at a later age has since been expanded to include several other diseases (Rautava et al., 2004). The general idea that persons that grow up in a western standard hygienic environment, with limited exposures (bacteria, fungi, parasites, and dirt), which normally provokes either Th1 or Th2 immune responses, risk developing a “lazy immune system” that does not respond appropriately to either allergens or autoantigens (McLachlan, 2003).

Based on data from Danish national registries, Hans Bisgaard’s group analyzed disease incident rates in two million children, born in the period 1977-2012, age 0-15, and found significantly increased risk of asthma, systemic connective tissue disorders, juvenile arthritis, inflammatory
bowel disease, immune deficiencies, and leukemia in children born by caesarean section (Sevelsted et al., 2015). Prescott et al. (1999) showed that lack of immune system maturation within the first year of life leads to development of atopic disease. Together these results indicate that the microbial causes of such diseases will most likely be found in early life.

1.4 Airway microbiota

The airways consist of the upper airways (nasal cavity, nasopharynx and oropharynx (throat), larynx (voice box)) and the lower airways (trachea (windpipe), bronchi (airways), and lungs). Many textbooks and articles state that the lower airways used to be considered sterile in healthy individuals, but this claim is always unsupported and is contrary to studies published in the last 90 years (Gleeson et al., 1997; Huxley et al., 1978; Quinn & Meyer, 1929; Amberson, 1954). Dickson et al. (2015b) have written a thorough review of the conceptual errors and challenges which have led to this belief. In 2010, the first lower airways microbiota study, applying culture independent techniques, found that healthy individuals did have bacteria in their lower airways; the composition differed from their upper airways microbiota, and from lower airways microbiota of individuals with respiratory disease (Hilty et al., 2010). Since then, there has been an increased interest in the lung microbiota and the number of published articles has increased from 2 articles in 2011 to 36 in 2015, totaling 99 articles, including 25 reviews (Figure 5). Many of these have studied the differences in lung microbiota between healthy and unhealthy individuals, with several focusing on Chronic Obstructive Pulmonary Disease (COPD)

![Graphical representation of the number of articles published (by year) found in the Pubmed database when searching for "lung microbiome" OR "lung microbiota". The number of articles per year is plotted on the left-hand y-axis and the total number of publications on the right-hand y-axis. Only articles published in the period 2011-2015 are included.](image)
(Miravitlles & Anzueto, 2015; Pragman et al., 2015; Malhotra & Olsson, 2015; Sze et al., 2014; Pragman et al., 2012) and Cystic Fibrosis (CF) (Caverly et al., 2015; Coburn et al., 2015; Carmody et al., 2015; Stokell et al., 2015; McGuigan & Callaghan, 2015).

The early airway microbiota is acquired directly after birth and is similar to the vaginal flora of the mother or general skin microbiota, depending on delivery mode (Dominguez-Bello et al., 2010). Alike to the gut microbiota, there is a maturation of the microbiota in the upper airways during the first 2.5-3 years of life, leading to microbiota resembling the adult microbiota (Biesbroek et al., 2014; Teo et al., 2015). No study has investigated the microbiota of the lower airways in early life (Dickson et al., 2015b). The lower respiratory tract microbiota differs in composition from its main source community, the upper respiratory tract microbiota (Dickson et al., 2015a; Segal et al., 2013; Morris et al., 2013). The most commonly found phyla in the lower respiratory tract microbiota belong to the phyla Bacteroidetes and Firmicutes, generally from the genera *Veillonella, Prevotella, Streptococcus,* and *Methylobacterium* (Morris et al., 2013; Dickson et al., 2014a, 2015a; Segal et al., 2013).

### 1.4.1 Airway microbiota and disease

To understand the airway microbiota of individuals with respiratory diseases Huffnagle’s group have presented, in several reviews, a model showing the difference between healthy and unhealthy individuals (Figure 6). The general idea is that in healthy lower airways the microbiota are shaped by constant microbial immigration and elimination, with very little growth within the lower airways themselves. When the microbial elimination processes (coughing, mucociliary clearance, and host defences) are impaired, regional growth becomes an increasing contributor to the overall microbiota composition.

Individuals with CF have a genetic mutation that affects their cystic fibrosis conductance regulator protein, a chloride channel with additional regulatory roles, which affects several organ systems (O’Sullivan & Freedman, 2009). I will focus solely on the airways, where the mutation causes poor solubility and aggregation of luminal mucins (Quinton, 2008). The buildup of mucins, which the body cannot clear by itself, provides an environment that is conductive for chronic pulmonary infections (O’Sullivan & Freedman, 2009). The bacterial community in the lungs of CF patients has a low diversity in more advanced stages of the disease (Coburn et al.,
The loss of microbial diversity has been correlated to the cumulative antibiotic use (Zhao et al., 2014). The dominant bacteria in children with CF are Streptococcus, Haemophilus, and Pseudomonas, in adults the microbiota is dominated by Pseudomonas, Burkholderia and Streptococcus (Coburn et al., 2015). The recurring infections are treated with antibiotic therapy, but eventually the lungs are damaged to such a degree that lung transplantation is necessary. A study analyzed the microbiota of explanted lungs from CF patients and found that in 8 of 10 patients Pseudomonas aeruginosa dominated the microbiota, for 6 of them it represented more than 99% of the reads. The lungs of the last two patients were dominated by Burkholderia cepacia complex or Achromobacter xylosoxidans each representing 76% and 92%, respectively (Goddard et al., 2012).

COPD is a progressive disease where lung tissue is broken down and lung function is diminished. Patients with COPD often suffer from increased cough and production of sputum, which as in CF impairs the microbial elimination from the lungs (Decramer et al., 2012).
patients with COPD studies have found that the microbiota differs from healthy controls, and that the degree of change is correlated with the use of inhaled corticosteroids (Pragman et al., 2012).

Asthma appears as a chronic disorder of the airways that is complex and characterized by variable and recurring symptoms, airflow obstruction, bronchial hyperresponsiveness, and an underlying inflammation (Mims, 2015). The disease includes several distinct endotypes with different underlying mechanisms, which confound many studies of asthma as they should be treated differently (Lötvall et al., 2011). Common for all is the reduced microbial elimination, caused by inflammation in the airways. The same phyla are found in the airways of asthmatic individuals compared to healthy individuals, but in different amounts; Proteobacteria are present in higher proportions, whereas the Firmicutes are less abundant (Huang et al., 2011; Hilty et al., 2010).

The airways microbiota in asthmatic individuals differs from the microbiota in untreated individuals (Marri et al., 2013) and individuals treated with macrolide antibiotics (Huang et al., 2011) or corticosteroids (Goleva et al., 2013). So far, the focus has been on changes in the microbiota, as a response to disease and treatment; for diseases such as CF and COPD, this is the only involvement of the microbiota. As the microbiota of the gut affect the risk of developing asthma, the airways microbiota in early life have been similarly correlated with the risk of developing asthma later in life. Teo et al. (2015) have shown that the nasopharyngeal microbiota at 2, 6, and 12 months affects the risk of asthma. He found that asymptomatic *Streptococcus* colonization in early life was a strong asthma predictor. This is supported by data from the two large asthma studies, PARSIFAL (Germany) and GABRIEL (Germany and Switzerland), which showed an inverse correlation between bacterial exposure and risk of asthma (Ege et al., 2011). These findings have been reproduced in animal studies, which showed that the immune response to environmental triggers do not mature in germ-free mice and that tolerance is promoted via the immune regulating transmembrane protein PD-L1 (Gollwitzer et al., 2014).

### 1.5 Cohort studies

Most studies of the human microbiota are cross sectional and descriptive of nature. These studies have provided an understanding of how the microbiota differ between groups of different ages, ethnicities, dietary habits. Studies of correlations between microbiota composition and diseases have generally been retrospective providing correlations, without defining the causal relationship.
between the two. For many diseases, especially immune related diseases, the risk of acquiring them is influenced by exposure at a much earlier time in life.

To study the causes of asthma it is necessary to have large prospective birth cohorts, which register health state and exposures from birth.

1.5.1 Copenhagen prospective studies on asthma in childhood

The clinical research unit Copenhagen Prospective Studies on Asthma in Childhood (COPSAC) aims to develop evidence-based prevention strategies for pediatric asthma. COPSAC was formed to deepen the understanding of asthma, eczema and allergy in young children, aiming to use the acquired knowledge to improve disease prevention, diagnosis and treatment. The group’s first cohort study was the selective COPSAC2000 cohort, which followed 411 infants of asthmatic mothers. Based on their results and experience, they have created a second cohort study, COPSAC2010, which is an unselected birth cohort with 700 mothers and their infants (Bisgaard, 2004; Bisgaard et al., 2013). The group has managed to include experts from a wide range of fields to study factors that likely affect early development and later disease risk. In COPSAC2010, the families have been monitored closely from week 24 of pregnancy, with 2 visits during pregnancy and 9 within 36 months of birth. The group has measured growth, physical activity, indoor air quality, diet, inflammatory markers, immunological markers, and sampled microbiota from feces, airways, as well as mothers’ vagina, and collected a daily diary reporting respiratory and skin-related symptoms (0-3 years), history of the most common childhood infections and medical treatments. Additionally, they have performed a genome-wide genotyping by dense single nucleotide polymorphism arrays (Illumina Omni2.5 BeadChip).
1.6 Objectives

The primary aims of this PhD study were to perform a detailed study of the relationship between mothers’ vaginal microbiota and the early microbiota in the airways and feces of their infants, as well as the early development of the hypopharyngeal microbiota.

Using 16S rRNA sequence analysis of hypopharyngeal, fecal and vaginal samples from the 700 COPSAC2010 children and their mothers, we aimed to investigate these microbial communities and their importance for diseases such as asthma and atopy. The following objectives were pursued:

- To describe mothers’ vaginal microbiota during the last trimester of pregnancy and how that affects the early microbiota of their infants.

- To describe the microbiota present in hypopharyngeal aspirates in early life, and the development of the microbiota within the first three months of life.

- To describe how the description of microbiota differs when using culture dependent and culture independent methods, and how well suited these methods are to provide answers.
LIST OF MANUSCRIPTS

2 LIST OF MANUSCRIPTS


Manuscript II: Martin S. Mortensen, Asker D. Brejnrod, Michael Roggenbuck, Waleed Abu Al-Soud, Karen A. Krogfelt, Jakob Stokholm, Jonathan Thorsen, Johannes Waage, Morten A. Rasmussen, Hans Bisgaard, Søren J. Sørensen. The postnatal development of the hypopharyngeal microbiota. (In preparation) (First author)

SUMMARY OF RESULTS

Manuscript I: Vaginal microbiota in pregnancy and its importance for infant microbiota.

We analyzed the vaginal microbiota of 700 pregnant women at week 24 and week 36 of pregnancy. In addition to a description of the vaginal flora, we also analyzed the influence of the vaginal microbiota at week 36 for their infants’ microbiota one week after delivery. We found that:

- *Lactobacillus* dominates the vaginal flora (78.5%)
- As shown by others, the vaginal community can be separated into 5 community state types (CST)
  - CST IV can be divided into subgroups by whether they are dominated by *Gardnerella*
- The vaginal microbiota were quite stable between week 24 and week 36, 83.5% belonged to the same community state type at both time points.
- 8 OTUs were found to be transferred from mother to infant when vaginally delivered.
  - These were the four dominating *Lactobacillus* spp., the dominant *Gardnerella*, an *Anaerococcus*, and a *Moraxella*.
  - Only two *Lactobacillus* species were correlated between maternal carriers and infants’ fecal microbiota, when the infants were born by caesarean section.

In conclusion, we have confirmed that the vaginal microbiota separate into five distinct CSTs, each with a dominating bacterial species. We have shown that there is a correlation between mothers’ vaginal microbiota and infants’ microbiota at one week of age, if the infant is born vaginally.
Manuscript II: The postnatal development of the hypopharyngeal microbiota.

We analysed the microbiota in hypopharyngeal aspirates from infants at 1 week, 1 month, and 3 months after birth. In this work, we included 1788 samples from 695 infants. We found that:

- The hypopharyngeal microbiota at 1 week after birth were dominated by *Staphylococcus* (49%).
  - As the abundance of *Staphylococcus* decreased (49% → 22% → 10%), there were an increase in abundance of *Streptococcus* (17% → 31%) and *Moraxella* (9% → 13% → 24%).
- 69% of the microbiota at 3 months belonged to OTUs in the infants’ core microbiota.
  - *Streptococcus* was part of the core microbiota in 96% of the infants.
- Five distinct types of microbiota composition, representing five pneumotypes (PTs), were described.
- Infants with PT II or PT V were the least likely to change to another PT, ~20% had these PTs at all three time points.
- The weighted UniFrac distance between samples from subsequent time-points were smaller within infants than between infants.

In conclusion, we described the development of the hypopharyngeal microbiota over 3 months, starting from 1 week after birth. We showed that OTUs present in an infant at 1 week and 1 month represented 69% of the microbiota abundance in the infant 3 months after birth. Overall, we found that the earliest exposure to bacteria shapes the development of the postnatal hypopharyngeal microbiota.
Manuscript III: Sequencing vs. Culturing: Is sequencing ready to replace culturing in a clinical setting?

As both culturing and sequencing data were available for this cohort, we compared the results obtained by the two methods. The aim was to determine if sequencing is ready to replace culturing for diagnostic purposes. We found that:

- Comparing culturing data to sequencing data analyzed by de-novo OTU picking:
  - Bacterial phyla cultured from fecal samples represented ≤50% of the reads classified.
  - In hypopharyngeal samples, more than half of the reads belonged to a genus also identified by culturing.
- Comparing culturing data to sequencing data analyzed by closed reference OTU picking:
  - From 1 to 11 bacterial species were cultured from each sample, the mean was less than 3 identified species per sample.
  - From 1 to 32 bacterial species were found in each sample by sequencing, mean was above 10.
  - Vaginal samples had the lowest richness when sequenced, but the highest richness when cultured.
  - In vaginal samples, less than 50% of all bacteria identified by culturing were found by sequencing.
  - In fecal samples, the percentage of cultured species found by sequencing decreased as the infants got older, whereas the percentage of sequenced bacteria found by culturing increased.

In conclusion, we showed that sequencing is more sensitive than culturing, and by using closed reference OTU picking, we were able to separate most of the cultured species using only their V4 region of 16S rRNA. Based on these findings, we believe that sequencing will be useful for diagnostic use in the near future.
4 DISCUSSION

One of the COPSAC research group’s overall aims is to elucidate any correlation and causality between early gene-environment interactions and later development of asthma and allergies. Part of this has been an effort to increase the understanding of the early bacterial exposure and microbiota development. The focus of this PhD project were sequencing and analyzing the 5411 bacterial samples collected as part of the COPSAC\textsubscript{2010} cohort. In the following part, I will discuss how the methods, results, and conclusion from the three manuscripts provide answers to the overall aims of this PhD thesis.

4.1 Transfer of vaginal microbiota to infants

Several studies have described and analyzed the vaginal microbiota in pregnant women and it is accepted that women have one of five defined community state types. The vaginal microbiota have been correlated with preterm birth and low birth weight (Carey & Klebanoff, 2005; Hillier et al., 1995), household pets (Stokholm et al., 2012), and antibiotic use (Stokholm et al., 2014).

The mothers’ vaginal microbiota have yet to be directly correlated to disease development in their infants, but some studies have shown data where the vaginal microbiota are expected to be part of a causal relationship. Use of antibiotics during pregnancy increases the risk of asthma development (Stensballe et al., 2013) and delivery by caesarean section increases the risk of several diseases, for which microbiota are important to risk and disease development (Sevelsted et al., 2015); both studies were retrospective database studies lacking microbiota data. While no direct effect of mothers’ vaginal microbiota has been shown so far. The findings indicate that mother’s vaginal microbiota could have an effect on disease outcome. A small study, ten mother-infant pairs, has shown that microbiota sampled from infants directly after birth are dependent on delivery mode and vaginally born infants could be paired with their mothers based on microbiota similarities (Dominguez-Bello et al., 2010).

We found that, when vaginally born, six \textit{Lactobacillus} species and two other OTUs were present more often in the infants’ fecal samples or hypopharyngeal aspirates one week after birth, if they had been present in the mothers’ vaginal sample at week 36 of pregnancy. In our data the transfer from mother to infant is not as clear as in the results presented by Dominguez-Bello \textit{et al.} (2010). This is mainly due to the huge difference in study design, as Dominguez-Bello \textit{et al.}
have samples only ten mother-infant pairs (5 vaginal deliveries); we have 650 mother-infant pairs (519 vaginal deliveries). In addition, Dominguez-Bello et al. collected samples from just prior to birth and five minutes after birth (within hours); we have collected samples three weeks prior to birth and one week after birth (four weeks apart). High similarity of the microbiotas should be expected, directly after birth, as bacterial exposure are limited to have taken place during birth. As part of the COPSAC2010 cohort, it was attempted to collected vaginal swabs from a subset of the mothers during labour (after the water broke and before giving birth). This would have provided data to determine if the week transfer from mother to infant was due to an altered vaginal microbiota before birth or if the microbiota simply does not colonize the infants during birth. As this was initiated late in the recruiting phase of the cohort, only 55 samples were collected, in a non-standardized manner, and have consequently been omitted from the analysis.

The bacteria transferred from mothers to infants were not present in high abundance and have not been found to affect the overall composition of the infants’ microbiota. We still have not correlated the microbiota to clinical outcomes, so whether this transfer is significant has yet to be determined.

4.2 The airways microbiota

The microbiota of the lower airways are notoriously difficult to sample, as the only access is through either the nose or mouth; Both are areas with a significantly more dense microbiota than the lower airways (Dickson et al., 2015a). Charlson et al. (2011) collected ten distinct samples from the airways of six individuals, including swaps from the upper airways, bronchial lavages, and bronchial brushing. Their results showed that the lower airways microbiota largely reflect the upper airways microbiota and that there are bacteria in the lungs of healthy individuals. This overlap between the upper and lower airways means that it is not feasible to remove upper airways associated microbiota from the data during analysis.

When sampling the airways of infants, ethical concerns as well as their small size prevent sampling by bronchial lavage or bronchial brushing. In the COPSAC2010 cohort, it was decided to collect hypopharyngeal aspirates, at the beginning of lower airways, by provoking a cough and then aspirating (Bisgaard et al., 2013). Even though the samples taken are believed to represent the lower airways, it would have improved the strength of results if we could have compared the microbiota composition with a sample from the upper airways.
Only few studies have investigated the microbiota in healthy airways within the first year of life; Cardenas et al. (2012) found difference in the oropharyngeal microbiota from 24 healthy and 24 non-infectious wheezing infants (average age 10.2 months). Bisgaard et al. (2007) found a correlation between microbiota in hypopharyngeal aspirates at 1 month (321 infants) and 12 months (231 infants) and risk of asthma at 5 years of age. Teo et al. (2015) correlated bacterial and viral presence in the nasopharynx at 2, 6, and 12 months in 234 infants with risk of developing allergic sensitization and asthma later in life. The size of the current cohort and the number of samples, 1788 samples from 700 infants, are significantly higher than in prior studies. Even if only including infants with high quality samples at all three time points are included (438 infants), our cohort still have a higher number of infants and samples. In addition, we have sampled the microbiota earlier and within a shorter timespan (1 week – 3 months) and have also analyzed the development of the microbiota in each individual (Manuscript II).

We found Staphylococcus representing almost half of the microbiota one week after birth and then steadily decreasing over time, as the abundance of genera normally associated with airways microbiota increased. An early presence of skin associated bacteria confirms the trend shown for the nasopharyngeal microbiota, where 80% of the healthy infants dominated by Staphylococcus were less than 12 weeks old (Teo et al., 2015). This development could signify that the present microbiota is just bacteria not yet eliminated and not a specific lower airways microbiota. We have shown that 69% of the microbial abundance three months after birth represents OTUs, which were present in the same individual, both in the first (one week) and second (one month) sample. Thus, we showed that while there are significant changes, the composition of the later microbiota is dependent on which bacteria were present already as early as one week after birth.

In addition, we analyzed how much the airways’ microbiota changed over time; for each infant we calculated the ratio between the weighted UniFrac distance within infant’s samples, from first to second time point, and the median distance, from the infant’s sample, from first time point, to other infants’ samples, from second time point. The median of these ratios were significantly less than 1 meaning less difference in microbiota between the infants’ own samples. The microbiota of samples from the infant were not so similar that we could pair infants based on their microbiota at each time point, but there was a significant trend of them being less different from each other than from other infants samples.
4.2.1 Defining pneumotypes

In the vaginal microbiota five distinct CSTs have been defined, based on the microbiota composition. For the gut microbiota three enterotypes have been defined, each dominated by a specific genus (Arumugam et al., 2011). Both CSTs and enterotypes have been described as stable over time, with enterotypes having been linked to long term dietary habits (Wu et al., 2011) and shown to be stable through a 6-month dietary intervention (Roager et al., 2014).

We have identified five clusters for the airways samples, referred to as pneumotypes (PTs). Similar to the vaginal CSTs, four have a dominating genus (Staphylococcus (PT I), Streptococcus (PT II), Moraxella (PT III), and Corynebacterium (PT IV)) and one lacks a common defining genus (PT V). The PTs have low stability compared to CSTs and enterotypes; only 9% of the infants had the same PT at all three time-points. However, this may be due to the age of the individuals at the time the samples were collected, both CSTs and enterotypes are defined in adults, whereas we have defined the PTs in infants within 3 months of birth. We do see an increase in stability over the second time span, even though the time between sampling is almost three times longer than between the first and second sample. The dominance of PT I (dominated by Staphylococcus) one week after birth indicates that, for more than half of the infants, a typical airways microbiota might not be present yet, and for the infants that were not dominated by Staphylococcus at one week, 16% had the same PT at all time-points. Comparing results from other studies, there is a large overlap with the microbiome profile groups presented by Teo et al. (2015), unfortunately they have not analyzed the temporal changes of the samples from each infant.

4.3 Culturing or sequencing

The samples from the COPSAC2010 cohort have been analyzed by both culturing and sequencing, culturing was performed qualitatively to identify the bacterial species in the samples, and 16S rRNA sequencing was performed to determine the bacterial composition of the samples, providing information about the bacteria present and their relative abundances. Both methods have some more or less well defined advantages and disadvantages. As the name indicates, culturing is limited to identification of bacteria that can be cultivated using the chosen settings. The method allows the identification of specific bacterial strains and specific antibiotic resistance profiling. Bacteria with a slow growth rate or low abundance are at risk of being overgrown by faster growing or more abundant bacteria, which reduces the number of species...
that can be identified per sample. For 16S rRNA sequencing there is no cultivation step, this means that even unculturable bacteria can be identified. In addition, the method provides a measure of the relative abundance of the bacteria in the sample. Using HTS it is possible to sequence thousands of reads from each sample and thereby allow for identification of bacteria present in very low abundances. The disadvantage using HTS is the introduction of PCR bias, as no PCR primer can amplify the 16S rRNA from all bacterial species, in addition only a fragment of the 16S rRNA is sequenced, which means that the resolution often is limited to genus level. Another consideration when using HTS it that bacteria are identified based on rRNA content and the method cannot distinguish between viable and dead bacterial cells.

### 4.3.1 De-novo OTU clustering of sequencing reads

We have compared the results from culturing and sequencing in an attempt to determine if sequencing could replace culturing for diagnostic purposes. For culturing, three selective media, under two different incubation settings, were used, and for sequencing, the V4 region of 16S rRNA was amplified and sequenced on Illumina MiSeq, reading at 250 bp paired-end reads, with a minimum of 2,000 reads per sample (Manuscript III).

Using de-novo OTU clustering at 97% identity, we could taxonomically identify most OTUs to genus level, some e.g. Enterobacteriaceae only to family level. In our first comparison, we analyzed the abundance of sequence reads in each sample that matched bacteria, which had been cultured from the same sample, at all taxonomical levels from phylum to genus. It was clear from the results that there were significant differences between the quantity of matching abundance in the three compartments and at the three time-points. Using de-novo OTU picking is not feasible for a clinical setting, the resolution of the sequencing data is not sufficient, and the difference in mean matching abundances between the compartments is not surprising, as it is known that only a fraction of the gut microbiota are culturable under the conditions (aerobic or microaerophilic) used (Simon & Gorbach, 1984).

### 4.3.2 Closed reference OTU picking of sequencing reads

To increase the resolution of the sequencing data, we used closed reference OTU picking against a limited reference database. Based on the 16S sequences of type strains from the ribosomal database project (RDP), we created a reference database containing sequences of the V4 region from bacteria cultured in this project. We matched reads with 100% sequence identity, to be able to increase the separation of the species. From the 140 bacteria we had 103 unique sequences; 8
Staphylococcus species, including S. aureus, had identical sequences, and 9 species, from 5 genera, of Enterobacteriaceae had identical sequences. In this approach, we have ignored the possibility of sequencing errors, which can occur in Illumina MiSeq data (Schirmer et al., 2015). Sequencing errors causing no match with our reference database are not a problem, as it is reasonable to expect that several reads represent each species and having errors in all of them would be unlikely. The problem is that many species have near identical sequences and we might get false positives because of sequencing errors. If known samples had been analyzed, it would be simple to calculate false positives, false negatives, and the accuracy of each method, but here we compare the results from two methods without a golden standard for comparison. We decided to follow the approach of Rhoads et al. (2012), where true negatives are disregarded and only the positive predictive values are evaluated. On average three out of four bacteria identified by culturing, were also confirmed by sequencing, while less than one in seven bacteria identified by sequencing had been cultured as well. That sequencing is more sensitive than culturing have been presented in prior studies (Rhoads et al., 2012) and can be explained by the ability to identify more bacteria per sample by sequencing (Manuscript III).

4.3.3 Sequencing in a clinical setting

According to the results presented in Manuscript III, sequencing is not ready for use in a clinical setting; as the resolution is insufficient, the 16S rRNA sequences of all species are not known, and the method still depends on too much practical work for the data analysis.

In the work presented, we have amplified and sequenced V4 only, as the Illumina MiSeq platform was not yet able to provide the reads long enough to include both V3 and V4 in sufficiently high quality. We have shown, in silico, that if both V3 and V4 had been sequenced, we would have been able to separate all the cultured Enterobacteriaceae species and the Staphylococcus species with identical sequences to S. aureus. The 16S rRNA sequence of culturable bacteria can easily be sequenced to solve the lack of certain type strains in the available databases. The last problem is the need for manual interactions to perform the data treatment and analysis. To solve that, it will be necessary to create an automated pipeline where library preparation is a 2-3 step process and where the direct output from sequencing is a list of high quality matches to a curated reference database of clinical relevant bacteria.
The COPSAC\textsubscript{2010} cohort is very comprehensive, concerning the number of mother-infant pairs, sample types and frequencies, metadata collection, and the diversity of the research performed. This study have included microbiota analysis of two vaginal samples from mothers, taken during pregnancy, three fecal samples, collected within the first year after birth, and three airways samples collected from more infants, an earlier time-point, sampling deeper in the airways, and three samples within a shorter timeframe than any other study of the microbiota in airways of healthy infants.

Based on sample collection and metadata, the present study has provided a deeper understanding of the development of the human microbiota within the first three months of life.

We have shown that part of the vaginal microbiota, which is well defined and stable during pregnancy. We have found that vaginal microbiota is transferred from mother to infant during vaginal delivery and affect the infants’ microbiota one week after birth.

More importantly, we have analyzed the microbiota from hypopharyngeal aspirations taken one week, one month, and three months after birth. We have shown that the lower airways microbiota develop into five pneumotypes of increasing stability and that, with exception of one, are dominated by one specific genus. Our results improve the understanding of the development of our airways microbiota, which has been lacking until now.

As all samples had been analyzed using both classical culturing and 16S rRNA sequencing, we compared the results from both analysis. We have shown that sequencing is close to ready for use in a clinical setting. The method has a higher sensitivity for identifying less abundant bacteria than culturing, but the technique and workflow are still too complicated, and require too much manual bioinformatic interaction, for clinical use. With dedicated work to sequence 16S rRNA from species missing in the databases, optimization and automation of the workflow, as well as sequencing of a larger 16S rRNA fragment than just V4, sequencing would be a feasible solution to replace classical culturing and identification methods for analysis of clinical samples.
6 PERSPECTIVES

The work and results presented in this PhD thesis provide a big step towards understanding how our microbiota develop in early life, but it have raised as many questions as have been answered. The first question now is how the microbiota correlate with all the information and measurements collected in the COPSAC 2010 cohort; does mothers diet affect the microbiota, is there correlations between the infants’ immune profiles and their microbiota, how do the microbiota in early life affect the risk of disease development later in life?

All answers to such questions will help us understand if there is such a thing as a healthy microbiota composition, or if development, as a process, is important for maturation of our immune system. The next step is to investigate the mechanisms by which the microbiota interact with itself and our body. Shotgun sequencing of the samples can provide such mechanistic information and is a step that we are hoping to take within the nearest future.

Lastly, it would be very interesting to use the knowledge of the causative effects of the early microbiota to develop methods or treatments to control the early microbiota development.
REFERENCES


REFERENCES


REFERENCES


## APPENDIX

<table>
<thead>
<tr>
<th>Manuscript I</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABSTRACT</td>
<td>41</td>
</tr>
<tr>
<td>BACKGROUND</td>
<td>41</td>
</tr>
<tr>
<td>RESULTS</td>
<td>44</td>
</tr>
<tr>
<td>DISCUSSION</td>
<td>47</td>
</tr>
<tr>
<td>CONCLUSIONS</td>
<td>48</td>
</tr>
<tr>
<td>METHODS</td>
<td>49</td>
</tr>
<tr>
<td>AUTHOR CONTRIBUTIONS</td>
<td>52</td>
</tr>
<tr>
<td>REFERENCES</td>
<td>52</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Manuscript II</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABSTRACT</td>
<td>57</td>
</tr>
<tr>
<td>INTRODUCTION</td>
<td>57</td>
</tr>
<tr>
<td>RESULTS</td>
<td>58</td>
</tr>
<tr>
<td>DISCUSSION</td>
<td>59</td>
</tr>
<tr>
<td>METHODS</td>
<td>62</td>
</tr>
<tr>
<td>REFERENCES</td>
<td>64</td>
</tr>
<tr>
<td>ACKNOWLEDGEMENTS</td>
<td>66</td>
</tr>
<tr>
<td>AUTHOR CONTRIBUTIONS</td>
<td>68</td>
</tr>
<tr>
<td>COMPETING FINANCIAL INTERESTS</td>
<td>68</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Manuscript III</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABSTRACT</td>
<td>73</td>
</tr>
<tr>
<td>BACKGROUND</td>
<td>74</td>
</tr>
<tr>
<td>RESULTS AND DISCUSSION</td>
<td>76</td>
</tr>
<tr>
<td>CONCLUSIONS</td>
<td>81</td>
</tr>
<tr>
<td>METHODS</td>
<td>81</td>
</tr>
<tr>
<td>REFERENCES</td>
<td>84</td>
</tr>
<tr>
<td>SUPPLEMENTARY MATERIAL</td>
<td>85</td>
</tr>
</tbody>
</table>
Manuscript I

Stability of Vaginal microbiota during pregnancy and its importance for early infant microbiota

Martin S. Mortensen¹, Asker D. Brejnrod¹, Christina Balle¹, Waleed Abu Al-Soud¹, Jakob Stokholm², Jonathan Thorsen², Johannes Waage², Morten A. Rasmussen²+⁴, Karen A. Krogfelt³, Hans Bisgaard², Søren J. Sørensen¹.

1) Section of Microbiology, Department of Biology, University of Copenhagen, 2100 Copenhagen, Denmark
2) Copenhagen Prospective Studies on Asthma in Childhood, Faculty of Health Sciences, University of Copenhagen, Copenhagen University Hospital Gentofte, Denmark
3) Microbiology and Infection Control, Statens Serum Institut, Artillerivej 5, 2300 Copenhagen S, Denmark
4) Department of Food Science, Faculty of Science, University of Copenhagen, 1958 Frederiksberg C, Denmark.

ABSTRACT

The vaginal microbiota of pregnant women might be important for the early development of their infants microbiota development. Studies have described the vaginal microbiota of pregnant women, and defined five distinct community state types, each dominated by specific bacteria. With exception of a single limited study the importance of the vaginal microbiota have not been investigated. We have used Miseq sequencing of 16S rRNA gene amplicons to characterize the vaginal microbiota, at week 24 and 36 of their pregnancy, of 700 women from the COPSAC2010 cohort and their infants’ microbiota from fecal samples and hypopharyngeal aspirations one week after birth. We show that the women presents the known community state types and that only one out of six changes community state type. More interestingly we have shown 8 bacterial species that are transferred from mother to infant, only if delivered vaginally.

Keywords: Pregnancy, vaginal microbiome, Lactobacillus, community stability, longitudinal sampling, women’s health, human microbiome, ecology, Gardnerella.

BACKGROUND

The human microbiota is the collection of microorganisms, which naturally inhabits exposed body surfaces and cavities. The human body is colonized by 10¹³-10¹⁴ microbes corresponding to
more than 10 times the total number of cells in the body (Ley et al., 2006). These microorganisms live in a symbiotic balance with the human host and a healthy microbiota is a prerequisite for normal homeostasis.

The vaginal community differs from the rest of the human microbiome. The physiology of the vagina differs from most other sites on the human body as there is no physical movement through the area as in the oral cavity, upper airways or intestinal tract and because it is not an exposed area as the skin. The vaginal microbiota is mostly dominated by lactic-acid producing species of the genera Lactobacillus (Hyman et al., 2005; Tärnberg et al., 2002; Zhou et al., 2004). Lactobacillus species are believed to maintain a healthy and balanced vaginal microbiota by creating a hostile environment that restricts the invasion of pathogenic bacteria and fungi. They do so by secreting antimicrobial compounds such as bacteriocins, hydrogen peroxide (H₂O₂) and anti-adherence molecules, and by maintaining a low vaginal pH (4.0-4.5) through their production of lactic acid (O’Hanlon et al., 2013; Ravel et al., 2011; Hummelen et al., 2010). Increased bacterial richness and diversity in the gut and upper airway microbiota, have been associated with health (Bisgaard et al., 2011; Chang et al., 2008; Cho & Blaser, 2012; Pflughoeft & Versalovic, 2012). However, for the vaginal microbiota, studies have shown clear correlations between increased diversity and risk of bacterial vaginosis (Hummelen et al., 2010).

In several studies describing the vaginal microbiota of healthy women during their reproductive years, five to six types of vaginal communities, referred to as community state types (CSTs), have been described (Hyman et al., 2005; Ling et al., 2010; Zhou et al., 2007). Four of these CSTs are mainly dominated by Lactobacillus species, most commonly L. crispatus (CST I), L. gasseri (CST II), L. iners (CST III), or L. jensenii (CST V) (Ravel et al., 2011). The fifth community state type, CST VI, has been characterized by a significant lack of Lactobacillus species and the presence of facultative and strictly anaerobic bacteria, including Prevotella, Atopobium vaginae, Gardnerella vaginalis, Dialister, Sneathia and Megasphaera (Chaban et al., 2014; Drell et al., 2013; Ravel et al., 2013). These bacterial species have been linked with the condition bacterial vaginosis (BV) (Verhelst et al., 2004; Ling et al., 2013; Fredricks et al., 2005) characterized by vaginal discharge and fishy odor. The lack of physical movement, the moisture, and the warmth are factors that make the vaginal environment very vulnerable to BV and increase the risk of acquiring urinary tract infections, vaginal yeast infections and sexually transmitted infections (Reid et al., 2006; Abbai et al., 2015).
During a woman’s reproductive years, the vaginal microbiota is highly dynamic and changes in structure and composition in response to external factors such as hormonal contraception, sexual behaviors, use of lubrication, and hygiene practices (Schwebke et al., 1999; Fashemi et al., 2013; Gupta et al., 2000). Additionally, the vaginal microbial community is highly affected by the menstrual cycle (Eschenbach et al., 2000; Gajer et al., 2012; Hickey et al., 2013), which is characterized by fluctuating levels of the female sex hormones estrogen and progesterone in addition to discharge of blood and mucosal tissue from the inner lining of the uterus. It has been shown that high levels of female sex hormones increases stability, while menses provides the largest disturbance of the vaginal microbiota (Ma et al., 2012; Gajer et al., 2012).

A woman experiences the highest levels of circulating estrogen and progesterone during pregnancy, and the vaginal microbiota of women with uncomplicated pregnancies have been shown to differ from that of non-pregnant, reproductive age women in both composition and stability with the vaginal microbiota of pregnant women being less diverse and more stable (Hernandez-Rodriguez et al., 2011; Romero et al., 2014; Walther-António et al., 2014). Perturbations to the vaginal microbiome during the course of pregnancy have been implicated in adverse pregnancy outcomes, including preterm delivery and low birth weight (Genc et al., 2004; Hillier et al., 1995; Carey & Klebanoff, 2005).

For decades, culture-based analyses have been the main approach for characterizing bacterial communities. These classic, culture-based studies have provided us with important information on the microbes inhabiting the human body, including the vagina, and contributed to the increasing understanding of the roles bacteria play in human health and disease. Yet, cultivation techniques are labor-intensive, time-consuming and many bacteria are difficult to culture (Suau et al., 1999). The advancements within genomic sequencing technology and bioinformatics tools have revolutionized the way we analyze complex microbial communities (Human Microbiome Project, 2012; Aas et al., 2005; Qin et al., 2012).

The purpose of this study was to characterize the temporal microbiota composition and diversity of the vaginal microbiota of 700 women followed longitudinally over the course of pregnancy and to investigate how the vaginal microbiota can influence the development of the intestinal tract and airways microbiota, of their infants at 1 week of age. The bacteria were identified by classifying 16S rRNA gene sequences using high-throughput amplicon sequencing.
RESULTS

Composition. Our study has been based on amplicon sequencing of the V4 region of 16S rRNA gene, which has limited the resolution to differentiate between *Lactobacillus* species, as a result we have several OTUs which aligned equally well with more than one species. In the case here, we see a big influence from *L. crispatus*/*L. acidophilus*, yet published studies on the vaginal microbiota concur that *L. crispatus*, in contrast to *L. acidophilus*, constitute an important part of the vaginal microbiota. Therefore, we will refer to *L. crispatus*/*L. acidophilus* solely as *L. crispatus*. Based on this reasoning we will also refer to *L. gasseri/johnsonii*, solely as *L. gasseri*.

The vaginal microbiota at gestational week 24 and 36 was dominated by bacteria belonging to the three phyla Firmicutes (85.0%), Acinobacteria (11.8%) and Proteobacteria (2.0%). The *Lactobacillus* genus represented 78.5% of the vaginal bacteria, *Gardnerella* 8.7%, *Enterobacteriaceae family* 1.5%, and 1.4% from *Atopobium* as well as other Lactobacillales, 1.5% *Bifidobacterium*, and 1.0% *Enterococcus*. The genera *Megasphaera*, *Streptococcus*, *Prevotella*, *Staphylococcus*, and *Dialister* were represented at a lower extent (0.38–0.83%). The most abundant lactobacilli were *L. crispatus* (33.3%), *L. iners* (28.6%), *L. gasseri* (10.7%), and *L. jensenii* (4.9%). A large proportion of the women had a vaginal microbiota predominantly consisting of *Lactobacillus*, whereas a smaller group of women presented with a more diverse vaginal microbiota with higher proportions of the anaerobic bacteria *Gardnerella*, *Dialister*, *Bifidobacterium*, *Prevotella*, *Megasphaera*, *Atopobium*, *Aerococcus*, and *Sneathia*.

Community state types. We sought to cluster the vaginal microbiota into distinct community types based on their similarity in OTU composition. The vaginal microbiota clustered into six CSTs (Figure 1); four dominated by *Lactobacillus* species (*L. crispatus*, *L. gasseri*, *L. iners*, or *L. jensenii*), one dominated by *Gardnerella*, and one, lacking a significant dominating OTU or genus, with an increased abundance of anaerobic and strictly anaerobic bacteria. These groups were not associated with specific time points during pregnancy as illustrated by the annotation bar in the heatmap depicting the sampling times (gestational week 24 or 36, respectively) (Figure 2). We named the CSTs in accordance with published studies (Ravel et al., 2011; MacIntyre et al., 2015), where: CST I is dominated by *L. crispatus*, CST II by *L. gasseri*, CST III by *L. iners*, CST IV-b by anaerobic bacteria, CST IV-c by *Gardnerella*, and CST V by *L. jensenii*.

Stability of microbiota within each mother. We have investigated the stability of the vaginal microbiota, within each woman, in two ways; first by analyzing the likelihood of a change of CST from gestational week 24 to 36, and secondly by comparing the distance (Jensen-Shannon
divergence) between the women’s own samples, compared to the distances to other women’s samples. For these two analyses, only women sampled twice were included.

**Figure 1:** Clustering of vaginal samples. NMDS plot of the weighted UniFrac distances between the vaginal samples. Samples are colored by their CST and the ellipses contain 75% of the samples from the respective CSTs. A) shows NMDS axes 1 and 4, B) shows the NMDS axes 3 and 5.

**Figure 2:** Heatmap of the most abundant species and OTUs. The samples in the plot have been sorted by their CST. The bars above the heatmap show how they separate into five, six, or seven clusters, and the lower line indicates at which time point each sample was taken.
We saw that 83.5% of the women had the same CST at week 24 and week 36, with CST IV-b (68.3%) and CST V (71.7%) being significant less stable ($\chi^2$-test, p-value = 0.001) (figure 3).

The distances between the samples were calculated as the ratio of the distance from a woman’s week 24 sample to her week 36 sample and the median distance from her week 24 sample to all other women’s week 36 samples. The median ratio was found to be 0.159, indicating that the distance between a woman’s samples was around one sixth of the median distance to other women’s samples, which was confirmed to be significant using a permutation test ($p < 10^{-4}$).

**Figure 3:** Alluvial plot of the women’s CST at week 24 and week 36.
For each woman a line connects their CST at week 24 with their CST at week 36.

Transfer of the microbiota. To analyze the importance of the vaginal microbiota on infant bacterial colonization, we looked at transfer of microbiota from mother to infant. To investigate this, we looked at which bacteria were present in both the maternal week 36 vaginal samples and
either airway or fecal samples from the offspring at 1 week of age. We found 8 OTUs, which were more often present in the infant if the bacteria were present vaginally in the mother (Table 1). For seven the correlation was significant to the infants’ fecal samples if the infant had been born vaginally, whereas only two, L. crispatus and Gardnerella_OTU13, were correlated to the infants airways samples. For the infants born by caesarean section, we found just two OTUs, L. gasseri and L. crispatus, to be more abundant in the fecal samples and none in the airways samples. This shows that the vaginal flora is important for the early bacterial exposure of the infants and their initial colonization.

DISCUSSION

The present study has characterized the vaginal microbiota of 695 Danish women during pregnancy. In concordance with previous North American and European studies (MacIntyre et al., 2015; Ravel et al., 2013; Romero et al., 2014), our results demonstrate that there are five major community state types of the vaginal microbiome, one of which can be further divided into subgroups. The vaginal microbiota were most commonly dominated by a single Lactobacillus species: Lactobacillus crispatus (CST I), L. gasseri (CST II), L. iners (CST III), and L. jensenii (CST V). We observed one CST IV subgroup in which a single Gardnerella OTU, CST IV-c, dominated the samples. Lastly, we saw the heterogeneous CST IV subgroup IV-b which were not dominated by a single OTU, but rather a wide range of anaerobic bacteria from the genera Gardnerella, Dialister, Bifidobacterium, Prevotella, Megasphaera, Atopobium, Aerococcus, and Sneathia (Chaban et al., 2014; Drell et al., 2013; Ravel et al., 2013).

We observed high levels of vaginal microbiota stability, which is consistent with previous studies of the vaginal microbiota during pregnancy (MacIntyre et al., 2015). Sex steroid
hormones are believed to play a major role in driving the composition and stability of the vaginal microbiota (Gajer et al., 2012), causing large variation depending on the time in the menstrual cycle (Ma et al., 2012). The larger stability, and small change over time, is expected when considering the lack of menses and high level of sex steroid hormones during the time of pregnancy, which have been sampled in this study.

When analyzing the vaginal microbiota’s importance for the early colonization of the airways and gut we compared the presence of the OTUs in the vaginal samples from gestational week 36 and the infants’ 1-week samples. We observed that the transfer of bacteria was more likely when the infants were delivered vaginally compared to when they were born by caesarean section. As far as we know, only one other study (Dominguez-Bello et al., 2010) has investigated how delivery mode affects the transfer of bacteria from mother to infant and they have shown that it was possible to separate the infants by delivery mode when analyzing bacterial samples taken from the infants within minutes of delivery. We show that even though there is a huge difference between vaginal, fecal, and airways samples, we can still see an effect of transfer from mother to infant in samples taken at 1 week of age, especially when the infants were delivered vaginally. This shows that the vaginal microbiota is important, not only for pregnancy outcomes (Kim et al., 2009; Vitali et al., 2012), but also for the formation of the infants’ microbiota. The microbiota in early life has been shown to affect risk of diseases later in life (Wang et al., 2008; Bisgaard et al., 2011), and that we have shown that the delivery mode affects the transfer of microbiota from mother to infant. This transfer of bacteria could be one of the mechanisms through which caesarean section increases the risk of several diseases later in life (Sevelsted et al., 2015).

CONCLUSIONS

Based on the participants in the COPSAC2010 cohort we have studies the community structure and stability of 700 women. Our findings confirms the present knowledge and the defining CSTs of the vaginal microbiota, including the low variability; five out of six women present the same CST at both time points. In addition, we show that the transfer of microbiota from mother to infant differ significantly between infants born vaginally and infants born by caesarean section and that members of the mothers vaginal microbiota can be identified in the infants fecal samples, and to a lesser extend airways samples, at least one week after birth. The proof that
caesarean section disrupt the transfer of microbiota from mother to infant can explain part of the lasting health effects associated with caesarean section.

METHODS

**Ethics.** This study followed the principles of the Declaration of Helsinki, and was approved by the Ethics Committee for Copenhagen (The Danish National Committee on Health Research Ethics) (H-B-2008-093) and the Danish Data Protection Agency (2008-41-2599). Written informed consent was obtained from all participants. The study is reported in accordance with the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) guidelines (von Elm et al., 2007).

**Study population.** Copenhagen Prospective Studies on Asthma in Childhood 2010 (COPSAC<sub>2010</sub>) is an ongoing Danish cohort study of 700 unselected children and their families followed prospectively from pregnancy week 24 in a protocol designed from the first COPSAC birth cohort (COPSAC<sub>2000</sub>) (Bisgaard, 2004). Exclusion criteria were gestational age below week 26; maternal daily intake of more than 600 IU vitamin D during pregnancy; or having any endocrine, heart, or kidney disorders.

**Sample collection.** Vaginal samples from the symptom-free women at pregnancy week 24 and 36 were collected from the posterior fornix of the vagina using flocked swabs (ESWAB regular, SSI Diagnostica, Hillerød, Denmark) (Stokholm et al., 2012). Hypopharyngeal aspirates, at 1 week of age, were collected with a soft suction catheter passed through the nose into the hypopharynx as previously described in detail (Bisgaard et al., 2007). Fecal samples were collected in sterile plastic containers at 1 week of age, stored at 4°C until they were transported (within 24 hours) to Statens Serum Institute (Copenhagen, Denmark). Each sample was mixed on arrival with 1 mL of 10% vol/vol glycerol broth (SSI, Copenhagen, Denmark) and frozen at -80 °C until further processing. 2,670 samples were collected and initially included.

**DNA extraction.** Genomic DNA was extracted from the mothers’ and infants’ samples using the PowerMag® Soil DNA Isolation Kit optimized for epMotion® (MO-BIO Laboratories, Inc., Carlsberg, CA, US) using the epMotion® robotic platform model (Eppendorf) according to the manufacturer’s protocol with the following alterations to the workflow: 150-250 µl of the samples were added to the 96-well bead plate containing 750 µl bead/RNase A Solution and 60 µl lysis solution. Centrifugation steps were performed at 3220xRCF for 9 minutes. Removal of
enzymatic inhibitors and DNA purification was performed as described by the manufacturer. Finally, the DNA was eluted with 100 µl Tris buffer (10mM, pH 7.5). DNA concentrations were determined using the Quant-iT™ PicoGreen® quantification system (Life Technologies, CA, US). Extracted DNA was stored at -20°C.

**16S amplicon sequencing.** The 16S rRNA gene amplification procedure was divided into two PCR steps. In the first PCR reaction, the hypervariable V4 region of the 16S rRNA gene was amplified using the modified broad range primers 515F (5’-GTGCCAGCMGCCGCGGTAA-3’) and 806R (5’-GGACTACHVGGGTWTCTAAT-3’)(Neefs & De Wachter, 1990; Yu et al., 2005; Sundberg et al., 2013). Amplification was performed in 96-well microtiter plates with a reaction mixture consisting of 1X AccuPrime PCR Buffer II, 0.6 U AccuPrime Taq DNA Polymerase (Invitrogen, Life technologies, CA, US), 0.5 µM primer 515F, 0.5 µM primer 806R, and 2.0 µl template DNA, giving a total volume of 20.0 µl per sample. Reactions were run in a 2720 thermal cycler (Applied Biosystems®, Life Technologies, CA, US) according to the following cycling program: 2 minutes of denaturation at 94°C, followed by 30 cycles of 20 seconds at 94°C (denaturing), 30 seconds at 56°C (annealing) and 40 seconds at 68°C (elongation), with a final extension at 68°C for 5 minutes. For each plate, a negative template-free control and a positive control containing 2.0 µl DNA from a known bacterial mock community (1.0 ng/µl; HM-782D, BEI Resources, VA, US) were included. The PCR products were quantified using the Quant-iT™ PicoGreen® quantification system (Life Technologies, CA, US) and samples with a concentration above 6.0 ng/µl were diluted to approximately 3.0-6.0 ng/µl prior to further analysis.

Sequencing primers and adaptors were added to the amplicon products in the second PCR step as follows: 2.0 µl of the diluted amplicons were mixed with a reaction solution consisting of 1X AccuPrime PCR Buffer II, 0.6U AccuPrime Taq DNA Polymerase (Invitrogen, Life technologies, CA, US) and 0.5 µM fusion forward and 0.5 µM fusion reverse primer (total volume 20 µl). The PCR was run according to the cycling program above except with a reduced cycling number of 15. The amplification products were purified with Agencourt AMPure XP Beads (Beckman Coulter Genomics, MA, US) according to the manufacturer’s specifications using 0.7X volume beads and quantified as described above. Equimolar amounts of the amplification products were pooled together in a single tube. The pooled DNA samples were concentrated using the DNA Clean & Concentrator™-5 Kit (Zymo Research, Irvine, CA, US) according to the manufacturer’s instructions. The concentration of the pooled libraries was
determined using the Quant-iT™ High-Sensitivity DNA Assay Kit (Life Technologies) following the specifications of the manufacturer. Amplicon sequencing was performed on the Illumina MiSeq System (Illumina Inc., CA, US). For each run, a 1.0% PhiX internal control was included. All reagents used were from the MiSeq Reagent Kits v2 (Illumina Inc., CA, US). Automated cluster generation and 250 paired-end sequencing with dual-index reads were performed. The sequencing output was generated as demultiplexed fastq-files for downstream analysis. Up to 192 samples were sequenced per run.

**Bioinformatics analysis.** Fastq-files were demultiplexed by the MiSeq Controller Software, trimmed for amplification primers, diversity spacers, and sequencing adapters using biopieces (Hansen, 2015), mate-paired using usearch v7.0.1090 (Edgar, 2010) and filtered using usearch -maxee 0.5. UPARSE (Edgar, 2013) was used for OTU clustering as recommended, in particular removing singletons after dereplication. Chimera checking was performed with usearch against the gold database (Haas et al., 2011) as recommend. Representative sequences were classified using Mothur v.1.25.0 wang() function (Schloss et al., 2009) at 0.8 confidence threshold. Qiime wrappers for PyNAST (Caporaso et al., 2010a), FastTree (Price et al., 2009), and filter_alignment.py (Caporaso et al., 2010b) were used to construct a phylogenetic tree. Alignments were built against the 2011 version of Greengenes (McDonald et al., 2012) and filtered using --allowed_gap_frac 0.999999 and --threshold 3.0.

For data treatment and analysis we used the open source statistical program ‘R’ (R Core Team, 2015), predominantly the R-package “phyloseq” (McMurdie & Holmes, 2013).

**Rarefaction of sample reads.** Samples with less than 2,000 sequences were excluded. 2,359 samples were included containing, on average, over 32,000 sequences per sample, representing 3,934 distinct OTUs. To avoid bias due to sampling depth, we removed the difference by randomly subsampling the OTU table at even sequencing depth of 2,000 observations. This dataset was used when calculating Jensen-Shannon divergence, as the method is sensitive to bias due to sequencing depth.

**Clustering analysis.** Clustering analysis was performed using hierarchical clustering, and the optimal number of clusters were chosen based on multiple cluster validation techniques (Handl et al., 2005): average silhouette width (Rousseeuw, 1987), pearson gamma index (Halkidi et al., 2001), dunn (Dunn, 1974), Calinski and Harabasz index, as well as comparison with similarity to CST presented in prior studies.
Comparison of distances to own and other women’s samples. The distance to each woman’s own sample was divided by the median distance to samples from other women. Following this we found the median of these ratios (W). To test the statistical significance, we performed 57,957 permutations by randomly assigning which column contained the within sample and gathered these ratios in a list (B). The p-value was calculated as: \( \text{sum}(W < B) / \text{length}(B) \).

Transfer of microbiota. To determine if the microbiota were transferred from mother to infant we used a binomial test, where the ratio of infants having the bacteria when the mother also had it were compared to the ratio of infants having the bacteria without their mother having had the bacteria. To do this we used binom.test in R and adjusted the p-values using Benjamini & Hochberg false discovery rate correction.

AUTHOR CONTRIBUTIONS

MSM performed DNA extraction, sequencing, and is the main author of this paper. ADB performed the initial bioinformatics analysis. MSM performed the microbiota analysis under supervision of WAA. JS sampled the infants. JS, JT, JW, and MAR helped interpret the data. This project was conceived and designed by HB, SJS, and KAK. All the authors have read and understood the manuscript.

REFERENCES


The postnatal development of the hypopharyngeal microbiota

Martin S. Mortensen¹, Asker D. Brejnrod¹, Michael Roggenbuck¹, Waleed Abu Al-Soud¹, Karen A. Krogfelt³, Jakob Stokholm², Jonathan Thorsen², Johannes Waage², Morten A. Rasmussen²+⁴, Hans Bisgaard², Søren J. Sørensen¹”.

¹Section of Microbiology, Department of Biology, University of Copenhagen, 2100 Copenhagen, Denmark
²Copenhagen Prospective Studies on Asthma in Childhood, Faculty of Health Sciences, University of Copenhagen, Copenhagen University Hospital Gentofte, Denmark
³Microbiology and Infection Control, Statens Serum Institut, Artillerivej 5, 2300 Copenhagen S, Denmark
⁴Department of Food Science, Faculty of Science, University of Copenhagen, 1958 Frederiksberg C, Denmark.

* Correspondence and requests for materials should be addressed to S.J.S. (sjs@bio.ku.dk)

ABSTRACT

The lower airways are among the few places in the human body where the immune system interacts closely with a microbial community. However, knowledge about the early microbial development is still limited. Recently studies suggest a critical role of early airway microbiota in development of immune mediated diseases in later life. We investigated hypopharyngeal microbiota development in the unselected COPSAC2010 cohort of 700 infants, by comprehensively analysing hypopharyngeal microbiota aspirates from one week, one month, and three months of age. We demonstrate five distinct community pneumotypes. Four of these are dominated by the genera Staphylococcus, Streptococcus, Moraxella, and Corynebacterium. Furthermore, we show a temporal change in pneumotype suggesting a rapid maturation of the hypopharyngeal microbiota. Despite an overall maturation, infants’ microbiota are more similar to their own, than to others, over time. Collectively, these findings indicate trajectories in the early development of the hypopharyngeal microbiota.
INTRODUCTION

No surfaces on the human body can escape colonization by microbes; this is particularly true for the large surface area of the gut.

The human gut microbiota have been extensively studied\(^1\). It has been shown that the gut is initially, in newborn infants, colonized by facultative anaerobe skin bacteria, followed by better adapted obligate anaerobe bacteria, when a sufficiently anaerobic environment have been formed\(^2,3\). During the first years of life the gut microbiota adapts to changing diet and is over time shaped into specific enterotypes, each representing a defined microbiota composition\(^4\). The gut microbiome has been shown to influence maturation of the immune system and the development of various immune mediated diseases\(^5-8\).

The human lungs actually have a larger surface area than the gut\(^9,10\), and provide, similar to the gut, intimate contact with host immune system. However, healthy lungs have traditionally been considered sterile and only colonized by bacteria during infections. In addition, the lower airways are difficult to access, which makes sampling invasive and uncomfortable. Therefore, less is known about the pulmonary microbiota, where questions pertaining to community stability and development are still unanswered. Recent studies have however found correlations between specific compositions of the airway microbiota and diseases such as chronic obstructive pulmonary disease, cystic fibrosis and asthma\(^11-13\). We have previously demonstrated that indeed 22% of neonates were asymptomatically carrying *Haemophilus influenzae*, *Streptococcus Pneumoniae* and *Moraxella Catarrhalis*. These children exhibited 4-5 fold increased risk of developing asthma later in childhood\(^14\). Teo et al. studied the infant nasopharyngeal microbiota and found also that certain bacteria correlated to acute respiratory infections and the risk of allergic sensitization and chronic wheeze\(^15\).

More knowledge on the development of the airway microbiota in early life is a fundamental necessity for understanding the interactions between the microbiota and the developing immune system, which are expected to have a huge impact on later health\(^16\).

The present study is based on the prospective COPSAC\(_{2010}\) mother-child birth cohort\(^17\). Here, the lower airway microbiota were sampled by aspirations from the hypopharynx of healthy infants at one week, one month and three months of age. We included 1,788 samples from 695 infants to, for the first time, explore the hypopharyngeal microbiota development, whether distinct microbial compositions exist in earliest life, and how stable they are over time.
RESULTS

Composition of hypopharyngeal microbiota over time. To investigate the hypopharyngeal microbiota we sequenced the V4 region of the 16S rRNA phylogenetic marker gene. The most frequent phyla found were Firmicutes (61% of reads), Proteobacteria (30%), Actinobacteria (6%), and Bacteroidetes (2%) (Supplementary Fig. 1). We observed a temporal increase in diversity and a change in the composition, as shown in Supplementary Fig. 1: The genus *Streptococcus* almost doubled between one week and one month and then maintained that abundance at three months (17%, 31%, and 29% of reads, respectively). *Staphylococcus* dominated at one week, but decreased more than 50% between each of the following time-points (49%, 22%, and 10%). The abundance of *Moraxella* increased between each time-point, from 9% at one week, to 13% at one month, and to 24% at three months.

Core microbiota in each infant. To test whether the hypopharynx is colonized by a distinct microbiota or solely contain microbes originating from micro-aspirations or contamination, we investigated the stability of the microbiota composition and the abundance of OTUs present in infants at all time-points (438 infants, Supplementary Table 1). When we defined an infant’s individual core microbiota as the OTUs present in all three samples from that infant. Each infant’s core microbiota consisted of 1-16 OTUs, most commonly *Streptococcus_OTU4*, with the five most abundant genera dominating the ten most common OTUs (Supplementary Table 2). Comparing the median abundance of the core microbiota at each time point (1 week: 90%, 1 month: 83%, and 3 months: 69%), we found a significant decrease in the median core microbiota (Wilcoxon P<0.001), reflecting increased diversity as the hypopharyngeal microbiota develops. However, OTUs already present at 1 week represented 69% of the microbiota abundance at 3 months, indicating an early establishment of a permanent resident hypopharyngeal microbiota.

Characterization of pneumotypes. Having confirmed the existence of a distinct hypopharyngeal microbiota we wanted to investigate if the children can be grouped by their hypopharyngeal microbiota in ways similar to what have been shown for the gut microbiota. Distance-based clustering revealed that the samples were optimally described by five clusters, as shown in Fig. 1, henceforth referred to as pneumotypes (PTs), based on the average silhouette width metric (Supplementary Fig. 2). As an average silhouette index below 0.5 suggest a weak
clustering, we confirmed that no other 5 cluster grouping had lower within cluster distances using a permutation test (\(P < 10^{-6}\)), thus confirming that our clustering is the optimal representation of 5 clusters, as shown in Fig. 1.

Characterizing the main microbial constituents of each PT using the indicator species approach\(^{18}\), we found clear indicator OTUs for 4 PTs belonging to the genera *Staphylococcus* (PT I), *Streptococcus* (PT II), *Moraxella* (PT III), and *Corynebacterium* (PT IV), respectively, at all time-points, whereas no indicator OTUs were found for the last PT (PT V) (Supplementary Table 3, Fig. 2).

**Figure 1**: Pneumotype clusters separated in NMDS plots using 4 axes. Microbial clustering on the basis of weighted UniFrac distances (visualized by non-metric multidimensional scaling (NMDS)), with ellipses encircling 75% of samples from each PT. (a) PT I, PT II, and PT III are separated on NMDS axes 1 and 2, with PT IV and PT V overlapping in the center of the plot. (b) On axes 3 and 4 PT IV separates from the other PTs. Axes minimum and maximum limits were fixed to exclude six outlying samples; the coordinates of these samples can be found in Supplementary Table 6.

**Time dependency of pneumotypes.** There was a significant change in the distribution of samples in the five PTs over time (\(\chi^2\)-test, \(P < 10^{-15}\); Fig. 3, Supplementary Table 4). The number of infants with PT I decreased while PT II, PT III, and PT V became more abundant. Despite these large-scale changes over time, the PTs were correlated over the two time spans, 1 week vs. 1 month and 1 month vs. 3 months (\(P = 0.006\) and \(P < 10^{-6}\), respectively). Additionally, finding the same PT at 1 month and 3 months (34% of infants) was more likely than at 1 week and 1
Figure 2: Abundance of dominant genera shows the difference between pneumotypes. Bar plot showing the abundance indicator genera in each sample, separated by time. (a) 1 week samples, (b) 1 month samples, (c) 3 month samples. The samples are sorted by PT and within the PT, by the abundance of the PTs indicator genus.

Figure 3: Dynamics of the infants’ pneumotype shown by an alluvial plot. Alluvial plot showing which PT each infant presents over time, including the 438 infants with three samples (Supplementary Table 4). Each infant is represented by a line connecting their PT at 1 week, through their PT at 1 month, to their PT at 3 months. The first part of the lines, from 1 week to 1 month, colored by their PT at 1 week, and the second part of the lines, from 1 month to 3 months, colored by their PT at 1 month.
month (27%) ($\chi^2$-test, $P = 0.02$). Of the 438 infants with samples from all time-points (Supplementary Table 1), infants maintaining the same PT at all time points, were significantly more frequent than what would have been expected if PTs were not correlated over time ($\chi^2$-test, $P < 3 \cdot 10^{-16}$). This was also true when testing each PT separately, except for the less well-defined PT IV (Supplementary Table 5).

The changes in PTs over time show that the maturation of the hypopharyngeal microbiota follows a trajectory, while still being partially dependent on the initial colonization found at one week.

Comparison of distance within and between infants. To confirm that this maturation of the microbiota was not solely a time dependent change, we analyzed the ratio of the weighted UniFrac distance from each infant’s 1 week sample to 1 month sample and the distance to other infants’ 1 month samples, and similarly from 1 month to 3 months. Over the two time spans the median ratios were 0.94 and 0.86, respectively. These were confirmed to be significantly different from 1.0 by permutations of which column represented the infants’ own samples in the distance matrix (both $P < 0.0001$). When comparing 1 week samples to 3 months samples, the median ratio was 0.99 ($P = 0.14$), meaning that the distances to the infants’ own samples were not significantly lower than the distance to other infants’ samples.

Based on these distances, the development of each infant’s 1 week microbiota could not be linked directly to the microbiota at 3 months, but when including the 1 month microbiota there was a traceable development from 1 week microbiota to 3 months microbiota.

DISCUSSION

We here present the first extensive study of the early hypopharyngeal microbiota in 695 healthy infants at 1 week, 1 month, and 3 months of age.

We show that 69% of the hypopharyngeal microbiota at 3 months is from OTUs present already at our first time point (1 week). This result indicates that the very early colonization is important for the formation of the microbiota later in life. At the same time, the pattern of initial colonization by skin bacteria (Staphylococcus), followed by better-adapted colonizers (e.g. Streptococcus, Moraxella), is similar to how the gut microbiota is colonized in early life\(^2\). Taking into account the documented effect of delivery mode on bacteria present in the nasopharynx
directly after birth\(^{19}\), it could be hypothesized that such a difference could be traced for an extended period, as seen in the gut microbiota\(^{20}\).

We have presented five pneumotypes, which were defined using a robust methodology inspired by the approach used to identify gut enterotypes\(^{4}\), and characterized them with regard to composition, indicator OTUs, and development over time. We found that 9\% of the infants had the same PT at all three time points. The median abundance of the core microbiota in these infants was significantly higher than in infants which changed PT over time, and the difference in abundance of core microbiota between the two groups increased over time (Fig. 4, Supplementary Table 5).

![Figure 4: The core microbiota is more abundant in infants having one PT continuously. Percentage of reads from the core microbiota over time, separated by whether or not the sample belongs to an infant, which have one PT continuously. There is an increasing difference in core microbiota abundance between the two groups.](image)

The presence of pneumotypes and temporal non-random changes in these indicates that non-random drivers shape the hypopharyngeal microbiota. Whether these drivers are microbial, genetic, environmental, or immunological, they must be the focus of further studies. The gut microbiota in early life have been shown to affect development of the immune and life-style disorders such as obesity and diabetes\(^{21-23}\). The vast surface areas of the airways provide intimate contact between the hypopharyngeal microbiota and the immune system. Therefore, the early hypopharyngeal microbiota may possess similar abilities of immune modulation, e.g. leading to increased airway inflammation, which has been indicated using murine models\(^{16}\) and observed in children linking early life bacterial colonization to asthma\(^{14}\).
The weighted UniFrac distances suggest that the bacterial community present at one week might not be predictive of the microbiota at three months; however, we could follow the development when including the one month sample. This indicates that the time span between 1 week and 3 months may be too long when analyzing the initial colonization of the hypopharynx, and that it would be prudent to increase the sampling frequency in future studies.

In summary, our findings demonstrate that there is an overall time-dependent development of the hypopharyngeal microbiota, i.e. the decrease of *Staphylococcus* and increase of genera normally present in the airways, e.g. *Streptococcus, Moraxella*, and *Haemophilus*\textsuperscript{11,15,24}, leading to the formation of a microbiota community which can be separated into specific pneumotypes.

We reported several findings, which indicate that development and structural principles of the hypopharyngeal microbiota share similarities with the gut microbiota, and as such could be similarly important for health and development.

**METHODS**

**Ethics.** This study followed the principles of the Declaration of Helsinki, and was approved by the Ethics Committee for Copenhagen (The Danish National Committee on Health Research Ethics) (H-B-2008-093) and the Danish Data Protection Agency (2008-41-2599). Written informed consent was obtained from all participants.

**Study population.** Copenhagen Prospective Studies on Asthma in Childhood 2010 (COPSAC\textsubscript{2010}) is an ongoing Danish cohort study of 700 unselected children and their families followed prospectively from pregnancy week 24 in a protocol designed from the first COPSAC birth cohort (COPSAC\textsubscript{2000})\textsuperscript{25}. Exclusion criteria were gestational age below week 26, daily intake of more than 600 IU vitamin D during pregnancy, or having any endocrine, heart, or kidney disorders.

**Sample collection.** The microbiota of the airways was sampled at 1 week, 1 month and 3 months of age. Hypopharyngeal aspirates were collected with a soft suction catheter passed through the nose into the hypopharynx as previously described in detail.\textsuperscript{14} 1988 samples were collected and initially included (Supplementary Table 1). The aspirates were diluted in 1ml sterile 0.9% NaCl and transported to the microbiological laboratory at Statens Serum Institut, Copenhagen, Denmark. Here the samples were separated into 150 µl aliquots and stored at -80°C.
DNA extraction and 16s amplicon sequencing. Genomic DNA was extracted using the PowerMag® Soil DNA Isolation Kit optimized for epMotion® (MO-BIO Laboratories, Inc., Carlsberg, CA, US) using the epMotion® robotic platform model (Eppendorf) under manufacturer’s protocol. 150 µl was used from each sample. DNA concentrations were determined using the Quant-iT™ PicoGreen® quantification system (Life Technologies, CA, US). Extracted DNA was stored at -20°C.

The 16S rRNA gene amplification procedure was divided into two PCR steps. First, amplification of the hypervariable V4 region of the 16S rRNA gene, using the modified broad range primers 515F (5’-GTGCCAGCMGCGCGGTAA-3’) and 806R (5’-GGACTACHVGGGTWTCTAAT-3’), then sequencing primers and adaptors were added to the amplicon products in the second PCR step. The amplification products were then purified with Agencourt AMPure XP Beads (Beckman Coulter Genomics, MA, US), pooled equimolar, concentrated using the DNA Clean & Concentrator™-5 Kit (Zymo Research, Irvine, CA, US), and the concentration were then determined using the Quant-iT™ High-Sensitivity DNA Assay Kit (Life Technologies). Paired-end sequencing, of up to 192 samples, were performed on the Illumina MiSeq System (Illumina Inc., CA, US), including 1.0% PhiX as internal control. All reagents used were from the MiSeq Reagent Kits v2 (Illumina Inc., CA, US). All details have been included in Supplementary Methods 1.

Bioinformatics analysis. Fastq-files demultiplexed by the MiSeq Controller Software were trimmed for amplification primers, diversity spacers, and sequencing adapters (biopieces29), mate-paired and quality filtered(usearch v7.0.109030, parameter: -maxee 0.5). UPARSE31 was used for OTU clustering as recommended, in particular removing singletons after dereplication. Chimera checking was performed with usearch against the gold database32 as recommended. Representative sequences were classified (Mothur v.1.25.033, wang() function at 0.8 confidence threshold). Qiime wrappers for PyNAST34, FastTree35, and filter_alignment.py36 were used to construct a phylogenetic tree. Alignments were built against the 2011 version of Greengenes37 (parameters: --allowed_gap_frac 0.999999 and --threshold 3.0).

The rarefaction curves (Supplementary Fig. 3) show the observed richness and the Shannon diversity as functions of count of sequences. Shannon diversity curves reach asymptotes with 1,000 sequences, based on this all samples with less than 2,000 sequences were excluded. 1,788 samples were included containing an average of 52,749 sequences per sample, representing 3,715 distinct OTUs. To avoid bias due to sampling depth, we removed the difference by
manuscript II

randomly subsampling the OTU table at even sequencing depth of 2,000 observations. All further post analyses were based on the even OTU table.

**Statistical analysis.** For data treatment and analysis we used the open source statistical program ‘R’\(^{38}\), predominantly the R-package “phyloseq”\(^{39}\). Wilcoxon rank sum test with continuity correction of difference in core OTUs abundance was achieved using function ‘wilcox.test’ (R-package ‘stats’). Pearson’s \(\chi^2\)-test of the significance in PT size variation between time-points was achieved function ‘chisq.test’ (R-package ‘stats’). The statistical significance of the difference in distances to each infant’s own sample and other infant’s samples were achieved using permutation, by randomly assigning which column contained the within sample.

**Characterization of pneumotypes.** Clustering analysis was performed using partitioning around medoids (PAM) clustering\(^{40}\). The Silhouette index, using both weighted UniFrac distances and Jensen-Shannon divergence, showed that 5 clusters were optimal (0.35 and 0.39, respectively; Supplementary Fig. 2)\(^{41}\); subsequent clustering was based on weighted UniFrac distances. NMDS ordination was performed using the function ‘metaMDS’ (R-package ‘vegan’\(^{42}\)) and the weighted UniFrac distances.

Indicator OTUs were identified using function ‘multipatt’ (func = “indVal.g”, R-package ‘indicspecies’\(^{18}\)).

**REFERENCES**

15. Teo, S. M. et al. The infant nasopharyngeal microbiome impacts severity of lower respiratory infection and risk of asthma development. Cell Host Microbe 17, 704–

**ACKNOWLEDGEMENTS**

We express our deepest gratitude to the children and families of the COPSAC2010 cohort study for all their support and commitment. We acknowledge and appreciate the unique efforts of the COPSAC research team. We thank Susanne Schjørring for critical logistic support, with handling and storing of samples at Statens Serum Institut, Denmark. Furthermore, we thank
Karin Pinholt Vestberg, Christina Brandt, and Jakob Russel (Section for microbiology, University of Copenhagen) for the help and support with DNA extraction, construction of the 16s rRNA gene amplicon libraries, and sequencing. We would like to thank the Lundbeck foundation for their support through grant no. R93-A8944.

AUTHOR CONTRIBUTIONS
M.S.M. performed DNA extraction, sequencing, and is the main author of this paper. A.D.B. performed the initial bioinformatics analysis. M.S.M. performed the microbiota analysis under supervision of M.R. and W.A.A. J.S. sampled the infants. J.S., J.T., J.W., and M.A.R. helped interpret the data. This project was conceived and designed by H.B., S.J.S., and K.A.K. All the authors have read and understood the manuscript.

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

SUPPLEMENTARY INFORMATION
Supplementary Figure 1. Relative abundance of hypopharyngeal microbiota taxa over time.
Interactive figure available at: http://goo.gl/z0yzLD
Krona plot showing the overall composition of the hypopharyngeal microbiota in all samples or at each time-point separately. This figure is interactive, and in the upper left corner the settings can be chosen; “Select dataset” to show data for one week, one month, 3 months, or all samples. “Max depth” sets to which taxonomic level the data is aggregated, 1: Phylum, 2: Class, 3: Order, 4: Family, 5: Genus, 6: OTU. “Collapse” simplifies the chart by collapsing "redundant" wedges that are entirely composed of another wedge. “Snapshot” creates a .svg image of the current display, and “Link” creates a link, including the current customized view.
Supplementary Figure 2. Silhouette index of PAM clusters using two metrics.

The silhouette index was calculated for 2 to 9 clusters using both Jensen-Shannon divergence and weighted UniFrac distances. 3 or 5 clusters were the best fit for our data, dependent on the metric, with 5 clusters giving a higher average silhouette index (0.735 and 0.738 for 3 and 5 clusters, respectively).
Supplementary Figure 3. Rarefaction curves showed a trend of an increasing diversity over time.

![Rarefaction curves](image_url)

Rarefaction curves were calculated for both observed richness and Shannon diversity, grouped by time-point and with bars indicating sd. The observed richness did not reach saturation before 15,000 reads, whereas the Shannon diversity reaches a maximum after 1,000 reads.

**Supplementary Table 1.** Overview of samples in this study and where in the workflow samples were removed. Complete infants indicates the number of infants from which all three samples are included.

<table>
<thead>
<tr>
<th>Step</th>
<th>Samples</th>
<th>Reason removed</th>
<th>Samples removed</th>
<th>Complete infants</th>
</tr>
</thead>
<tbody>
<tr>
<td>Samples received</td>
<td>1988</td>
<td></td>
<td></td>
<td>592</td>
</tr>
<tr>
<td>DNA extraction</td>
<td>1972</td>
<td>Contamination</td>
<td>16</td>
<td>583</td>
</tr>
<tr>
<td>Sequencing</td>
<td>1947</td>
<td>No sequences from sample</td>
<td>26</td>
<td>567</td>
</tr>
<tr>
<td>Removal of unwanted samples</td>
<td>1946</td>
<td>Duplicate 3 month sample</td>
<td>1</td>
<td>567</td>
</tr>
<tr>
<td></td>
<td>1944</td>
<td>Unusual high alpha-diversity</td>
<td>2</td>
<td>565</td>
</tr>
<tr>
<td></td>
<td>1788</td>
<td>Less than 2000 reads</td>
<td>156</td>
<td>438</td>
</tr>
</tbody>
</table>
Supplementary Table 2. Most common OTUs in the core microbiota, shown as the percentage of infants having this OTU as part of their core microbiota.

<table>
<thead>
<tr>
<th>OTU</th>
<th>Infants (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Streptococcus_OTU4</td>
<td>96.1</td>
</tr>
<tr>
<td>Staphylococcus_OTU1</td>
<td>83.6</td>
</tr>
<tr>
<td>Moraxella_OTU6</td>
<td>62.3</td>
</tr>
<tr>
<td>Gemella_OTU12</td>
<td>59.4</td>
</tr>
<tr>
<td>Haemophilus_OTU11</td>
<td>32.6</td>
</tr>
<tr>
<td>Corynebacterium_OTU15</td>
<td>28.8</td>
</tr>
<tr>
<td>Staphylococcus_OTU3295</td>
<td>25.3</td>
</tr>
<tr>
<td>Streptococcus_OTU14</td>
<td>20.5</td>
</tr>
<tr>
<td>Dolosigranulum_OTU23</td>
<td>18.5</td>
</tr>
<tr>
<td>Veillonella_OTU29</td>
<td>16.9</td>
</tr>
</tbody>
</table>

Supplementary Table 3. Indicator values for the most significant indicator OTUs for each pneumotype. A is the Positive predictive power of OTU as indicator for the given PT (Specificity). B is the sensitivity of the OTU as indicator for the given PT (Fidelity). Stat is the statistical value of OTU as indicator for the given PT.

<table>
<thead>
<tr>
<th>PT</th>
<th>OTU</th>
<th>All time-points</th>
<th>1 week</th>
<th>1 month</th>
<th>3 months</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>A</td>
<td>B</td>
<td>stat</td>
<td>A</td>
</tr>
<tr>
<td>PT I</td>
<td>Staphylococcus_OTU1</td>
<td>0.7</td>
<td>1</td>
<td>0.9</td>
<td>0.6</td>
</tr>
<tr>
<td></td>
<td>Staphylococcus_OTU3295</td>
<td>0.7</td>
<td>1</td>
<td>0.8</td>
<td>0.6</td>
</tr>
<tr>
<td></td>
<td>Staphylococcus_OTU5181</td>
<td>0.7</td>
<td>0.8</td>
<td>0.8</td>
<td>0.6</td>
</tr>
<tr>
<td></td>
<td>Staphylococcus_OTU3678</td>
<td>0.7</td>
<td>0.6</td>
<td>0.7</td>
<td>0.6</td>
</tr>
<tr>
<td></td>
<td>Staphylococcus_OTU765</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PT II</td>
<td>Streptococcus_OTU4</td>
<td>0.5</td>
<td>1</td>
<td>0.7</td>
<td>0.6</td>
</tr>
<tr>
<td></td>
<td>Streptococcus_OTU14</td>
<td>0.6</td>
<td>0.7</td>
<td>0.6</td>
<td>0.5</td>
</tr>
<tr>
<td></td>
<td>Veillonella_OTU16</td>
<td>0.6</td>
<td>0.7</td>
<td>0.6</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Gemella_OTU12</td>
<td></td>
<td>0.5</td>
<td>0.9</td>
<td>0.7</td>
</tr>
<tr>
<td>PT III</td>
<td>Moraxella_OTU6</td>
<td>0.7</td>
<td>1</td>
<td>0.8</td>
<td>0.7</td>
</tr>
<tr>
<td>PT IV</td>
<td>Corynebacterium_OTU15</td>
<td>0.9</td>
<td>0.9</td>
<td>0.9</td>
<td>0.9</td>
</tr>
<tr>
<td></td>
<td>Corynebacterium_OTU37</td>
<td>0.8</td>
<td>0.7</td>
<td>0.8</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Dolosigranulum_OTU23</td>
<td>0.7</td>
<td>0.7</td>
<td>0.7</td>
<td></td>
</tr>
<tr>
<td>PT V</td>
<td>No OTU found</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Supplementary Table 4. Distribution of samples between the five pneumotypes at each time-point.

<table>
<thead>
<tr>
<th>Pneumotype</th>
<th>1 week (544)</th>
<th>1 month (621)</th>
<th>3 months (623)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PT I</td>
<td>297 (55%)</td>
<td>137 (22%)</td>
<td>54 (9%)</td>
</tr>
<tr>
<td>PT II</td>
<td>87 (16%)</td>
<td>211 (34%)</td>
<td>196 (31%)</td>
</tr>
<tr>
<td>PT III</td>
<td>51 (9%)</td>
<td>89 (14%)</td>
<td>165 (26%)</td>
</tr>
<tr>
<td>PT IV</td>
<td>24 (4%)</td>
<td>53 (9%)</td>
<td>30 (5%)</td>
</tr>
<tr>
<td>PT V</td>
<td>85 (16%)</td>
<td>131 (21%)</td>
<td>178 (29%)</td>
</tr>
</tbody>
</table>

Supplementary Table 5. Number of infants presenting the same PT continuously. Observed infants is how many infants we observed in this study. Random infants is how many infants would have had this pneumotype continuously, if the infants were randomly distributed between the pneumotypes. Ratio is shows how many times more infants were observed to have the pneumotype continuously than what would be expected by random distribution.

<table>
<thead>
<tr>
<th>PT</th>
<th>Observed infants (%)</th>
<th>Random infants (%)</th>
<th>Ratio (Observed/Random)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PT I</td>
<td>10 (4.1%)</td>
<td>4.4 (1.82%)</td>
<td>2.25</td>
</tr>
<tr>
<td>PT II</td>
<td>14 (19.2%)</td>
<td>7.4 (10.1%)</td>
<td>1.9</td>
</tr>
<tr>
<td>PT III</td>
<td>4 (9.3%)</td>
<td>1.8 (4.17%)</td>
<td>2.23</td>
</tr>
<tr>
<td>PT IV</td>
<td>0</td>
<td>0</td>
<td>-</td>
</tr>
<tr>
<td>PT V</td>
<td>13 (20.6%)</td>
<td>4.0 (6.3%)</td>
<td>3.27</td>
</tr>
</tbody>
</table>
Sequencing vs. Culturing:

Is sequencing ready to replace culturing in a clinical setting?

Martin S. Mortensen¹, Susanne Schjørring², Asker D. Brejnrod¹, Waleed Abu Al-Soud¹, Jakob Stokholm³, Hans Bisgaard³, Karen A. Krogfelt², Søren J. Sørensen¹
¹ Section of Microbiology, Department of Biology, University of Copenhagen, 2100 Copenhagen, Denmark
² Microbiology and Infection Control, Statens Serum Institut, Artillerivej 5, 2300 Copenhagen S, Denmark
³ Copenhagen Prospective Studies on Asthma in Childhood, Faculty of Health Sciences, University of Copenhagen, Copenhagen University Hospital Gentofte, Denmark
* Corresponding author: S.J.S. (sjs@bio.ku.dk)

ABSTRACT

Background: The correlation of microbiota for our health has been investigated since the discovery of bacteria. For a long time such studies have been limited by the inability to culture all bacteria. Recent developments in culture-independent methods and especially with development of high throughput sequencing (HTS) have transformed the way microbiota is analyzed. While being widely used in microbiota studies, HTS of 16S rRNA has not yet replaced culturing for clinical and diagnostic purposes. The aim of this study was to compare the diagnostic efficiency of culturing compared to sequencing.

Methods: Based on samples from the COPSAC₂₀₁₀ cohort, following 700 mother-infant pairs, we have compared culturing and HTS of 16S rRNA of 3552 samples. Both de-novo OTU picking and closed reference OTU picking (based on culturable species) have been performed for analysis of the sequencing data.

Results: Using de-novo OTU picking of sequencing data bacterial phyla cultured from fecal samples represented ≤50% of the reads classified. In hypopharyngeal samples, more than half the reads belonged to genera also identified by culturing.
Of the 140 cultured bacteria, which had a type strain in the RDP database there were 103 unique sequences when comparing the V4 region. Grouping the cultured bacteria to match the 103 unique bacteria, between 1 and 11 bacterial species were cultured from each sample (mean < 3 species/sample). Between 1 and 32 species/sample were identified by sequencing (mean > 10 species/sample). Vaginal samples had the lowest richness by sequencing, but the highest richness when cultured, and less than 50% of culturable bacteria, identified by culturing, were identified by HTS as well. In fecal samples, the percentage of cultured species found by sequencing decreases as the infants get older, whereas the percentage of sequenced bacteria found by culturing increases.

**Conclusion:** We have shown that HTS is more sensitive than culturing as well as able to identify bacteria that cannot be cultured, and using closed reference OTU picking we were able to separate most of the cultured species using only the V4 region of 16S rRNA. Based on these findings we believe that sequencing will be useful for bacterial identification as soon as a standardized protocol and pipeline are developed.

**Keywords:** Molecular diagnostic techniques, Bacteria, Microbiology, Culturing, 16S rRNA, HTS

**BACKGROUND**

Knowledge about the importance of bacteria for human health and disease have evolved together with the methods used to identify and investigate bacteria. From the time of Pasteur, the goal of most health related microbiology have been to isolate, identify and describe bacteria. For this purpose, culturing is a very suited method and it is still used to identify bacteria related to acute infections. After the hygiene hypothesis was first proposed (Strachan, 1989), the study of correlations between bacteria and disease has been expanded to also include chronic diseases such as obesity (Turnbaugh et al., 2009; Tilg & Kaser, 2011; Turnbaugh et al., 2006), diabetes (Qin et al., 2012; Larsen et al., 2010; Sanz et al., 2015), asthma, and allergies (Bisgaard et al., 2011; Gollwitzer et al., 2014; Ege et al., 2011). Over the same period, new DNA-based diagnostic techniques, such as PCR, DNA finger printing and high throughput sequencing (HTS) have been developed and become increasingly rapid, sensitive and cost efficient. This have meant that the methods used for large scale studies of human populations have also changed, and the development is particularly noticeable when comparing techniques used by a single research group. One such example is the COPSAC group, managing the COPSAC2000 (Bisgaard, 2004)
MANUSCRIPT III

and COPSAC2010 (Bisgaard et al., 2013) cohorts. This group have published data using classic culturing (Stokholm et al., 2012), DGGE (Bisgaard et al., 2011) and now HTS of 16S rRNA (Mortensen et al., 2016a, 2016b).

Culturing enables classification of bacteria to species and strain level, and in a clinical setting, the right choice of media and incubation settings will allow for selective growth of relevant bacteria. When studying a bacterial community or a sample of unknown composition, there are several disadvantages with using culturing; Only bacteria able to grow under the selected conditions can be identified, culturing is work intensive, there is a possibility of misidentification (Rhoads et al., 2012), and the results will only be qualitative (presence/absence).

For HTS of 16S rRNA, DNA is first extracted, a specific region of 16S rRNA is amplified, sequenced, and then identification of generated sequences is based on similarity to reference 16S rRNA sequences available in public databases, such as the Ribosomal Database Project (RDP) (Cole et al., 2014). The advantages of high throughput 16S rRNA sequencing are; the method does not rely on the ability to culture the bacteria in a sample, the relative abundance of all bacteria in the sample can be determined, and the method allow for parallel sequencing of hundreds of samples at the same time. As with culturing there are disadvantages with high throughput 16S rRNA sequencing; the primers used for amplification will introduce a bias, as they bind to regions which are not 100 % conserved across all bacteria, and bacteria can only be identified to genus level due to high similarity between 16S rRNA from closely related species.

Efforts to compare the results from culturing and sequencing have mostly been focused on identification of pure cultures. Woo et al. (2008) have made a very thorough review of the usefulness of 16S rRNA sequencing in 2008, where they listed diseases for which the bacterial cause cannot be cultured, e.g. Whipple’s disease, culture negative endocarditis.

Comparison between the methods when looking at bacteria communities have been shown by only a few studies, mostly looking at the bioburden in chronic wounds (Rhoads et al., 2012) or on a very small number of subjects (Westergren et al., 2003; Dickson et al., 2014).

As part of the COPSAC2010 cohort, we have collected fecal samples, vaginal swabs and hypopharyngeal aspirations. Initially the microbiota of all samples were characterized using traditional culturing techniques, and now we have performed HTS of 16S rRNA gene, using
Illumina bridge amplification technology, to provide a more detailed microbiota characterization. Here we compare 3552 samples, which have been both cultured and sequenced. In addition to de-novo OTU picking, we analyzed the sequencing data using closed reference OTU picking (based on culturable species) to achieve to species level identification, which is more relevant for clinical settings.

RESULTS AND DISCUSSION

Characterization of samples by culturing. In this project we cultivated bacteria from vaginal (692), hypopharyngeal (1914), and fecal samples (1935). From the 4541 samples, we isolated 9,983 pure cultures and identified 179 unique bacteria. The fecal samples contained on average 2.7 bacterial species, most commonly \textit{Escherichia coli} (57 \%), \textit{Staphylococcus epidermidis} (29 \%), and \textit{Enterococcus faecalis} (23 \%). The hypopharyngeal samples contained on average 2.4 bacterial species, most commonly \textit{Staphylococcus aureus} (48 \%), \textit{S. epidermidis} (42 \%), and \textit{Corynebacterium spp.} (28 \%). The vaginal samples contained on average 3.4 bacterial species with \textit{S. epidermidis} (61 \%), \textit{Corynebacterium spp.} (42 \%), and \textit{Lactobacillus spp.} (40 \%) being most commonly isolated from the samples. In a prior study (Stokholm et al., 2012), describing the vaginal samples from week 36 and how having household pets correlates with vaginal colonization for specific bacteria. The most commonly identified bacteria are shown in table 1 (full list in supplementary table 1)

<table>
<thead>
<tr>
<th>Top Five</th>
<th>Vaginal</th>
<th>Hypopharyngeal</th>
<th>Fecal</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>\textit{S. epidermidis}</td>
<td>61%</td>
<td>\textit{S. aureus}</td>
</tr>
<tr>
<td>2</td>
<td>\textit{Corynebacterium sp}</td>
<td>42%</td>
<td>\textit{S. epidermidis}</td>
</tr>
<tr>
<td>3</td>
<td>\textit{Lactobacillus sp}</td>
<td>40%</td>
<td>\textit{Corynebacterium sp}</td>
</tr>
<tr>
<td>4</td>
<td>\textit{E. faecalis}</td>
<td>27%</td>
<td>\textit{M. catarrhalis}</td>
</tr>
<tr>
<td>5</td>
<td>\textit{S. haemolyticus}</td>
<td>25%</td>
<td>\textit{S. mitis/oralis}</td>
</tr>
</tbody>
</table>

Characterization of samples by high throughput sequencing. The richness of the fecal samples was on average 36 ± 13 genera per sample and the most dominant genera were \textit{Bacteroides} (31 \%), family \textit{Enterobacteriaceae} (21 \%), \textit{Bifidobacterium} (13 \%), \textit{Clostridium} (4 \%), and \textit{Veillonella} (4 \%). The vaginal samples had an average richness of 24 ± 12 genera per sample and were dominated by the genera \textit{Lactobacillus} (79 \%), \textit{Gardnerella} (8.7 \%), \textit{Enterobacteriaceae} family (1.5 \%), and \textit{Bifidobacterium} (1.5 \%) (Mortensen et al., 2016a).
Streptococcus (26 %), Staphylococcus (26 %), Moraxella (16 %), and Haemophilus (5 %) dominated the hypopharyngeal samples (average richness 25 ± 9 genera per sample) (Mortensen et al., 2016b).

We classified the OTUs to genus level, except for a few families, e.g. *Enterobacteriaceae*, where members of different genera and even the genera were not distinguishable. Several genera had multiple OTUs associated to them, but if any of these represented a unique species we could not tell. To investigate how well the culturing and HTS data correlated we analyzed what proportion of the sample abundance, from sequencing, that represented bacteria cultured in the specific sample. This was done at all taxonomic levels from Phylum to genus, and can be seen in figure 1. There are large differences between the sample types; bacteria cultured from fecal samples are less abundant. For the vaginal samples, at order, family and genus level, the standard deviations are very large, and there is a big difference in mean abundance between order and family level.

![Figure 1: Abundance of sequence reads, classified using de-novo OTU clustering, matching bacteria identified, by culturing, in each sample.](image)

**Closed reference clustering.** To improve resolution of the sequencing data and allow for comparison at species level we performed closed reference OTU picking at a 100% identity. We compiled a reference database of type strains, from the RDP database, for the species identified by culturing. We disregarded isolates not identified to species level (Supplementary table 2). Of
the 187 unique bacteria cultured, 24 were only identified to genus level, these bacteria represented 12.5% of the identified bacteria in the samples, with the majority being *Corynebacterium* (7.5%) and *Lactobacillus* (2.7%). Of the remaining species 14 did not have a type strain in the RDP database (1.7%), and two are not covered by our primers (0.07%) (Supplementary table 3). In addition, we pooled the bacteria, which had identical sequences, giving 103 unique bacterial species and groups to compare. Out of our 204 million high quality reads, 34.1 % matched to our reference database, An average of 19,962 sequences per sample, and represented 86 of 103 unique sequences in the database.

Hypopharyngeal samples had the highest abundance of matching sequences (55.2 ± 20.3 %), followed by fecal samples (20.7 ± 22.7 %), while a very low percentage of the vaginal samples matched sequences in our reference database (4.0 ± 13.4 %). The low abundance of matching sequences in vaginal samples is most likely because we have no *Lactobacillus* species in our reference database as these *Lactobacillus* sequences were only identified to genus level by culturing. The sample richness, total and separated by sample type and time-point, for both culturing and closed reference sequencing, can be seen in table 2.

Bacteria found by culturing and sequencing (5,374) represented 74.0 % of all cultured bacteria, but only account for 13.9 % of sequenced bacteria. Bacteria, which were dominant in sequencing, were more likely to have been cultured (62.1 %) when compared to major (29.6 %) and minor bacteria (7.4 %) (Table 3, Figure 2).

<table>
<thead>
<tr>
<th>Type</th>
<th>Time</th>
<th>Mean</th>
<th>Min</th>
<th>Max</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fecal</td>
<td>All</td>
<td>2.06</td>
<td>1</td>
<td>7</td>
<td>0.91</td>
</tr>
<tr>
<td></td>
<td>One week</td>
<td>2.01</td>
<td>1</td>
<td>5</td>
<td>0.84</td>
</tr>
<tr>
<td></td>
<td>One month</td>
<td>1.96</td>
<td>1</td>
<td>5</td>
<td>0.87</td>
</tr>
<tr>
<td></td>
<td>One year</td>
<td>2.16</td>
<td>1</td>
<td>7</td>
<td>0.97</td>
</tr>
<tr>
<td>Hypopharyngeal</td>
<td>All</td>
<td>2.00</td>
<td>1</td>
<td>6</td>
<td>0.97</td>
</tr>
<tr>
<td></td>
<td>One week</td>
<td>1.79</td>
<td>1</td>
<td>6</td>
<td>0.91</td>
</tr>
<tr>
<td></td>
<td>One month</td>
<td>2.01</td>
<td>1</td>
<td>6</td>
<td>0.94</td>
</tr>
<tr>
<td></td>
<td>Three months</td>
<td>2.17</td>
<td>1</td>
<td>6</td>
<td>1.00</td>
</tr>
<tr>
<td>Vaginal</td>
<td>All</td>
<td>2.15</td>
<td>1</td>
<td>7</td>
<td>1.03</td>
</tr>
<tr>
<td></td>
<td>Week 36</td>
<td>2.14</td>
<td>1</td>
<td>7</td>
<td>1.04</td>
</tr>
<tr>
<td></td>
<td>Birth</td>
<td>2.18</td>
<td>1</td>
<td>4</td>
<td>0.98</td>
</tr>
<tr>
<td>Fecal</td>
<td>All</td>
<td>10.84</td>
<td>2</td>
<td>32</td>
<td>4.18</td>
</tr>
<tr>
<td></td>
<td>One week</td>
<td>13.77</td>
<td>5</td>
<td>32</td>
<td>3.72</td>
</tr>
<tr>
<td></td>
<td>One month</td>
<td>11.06</td>
<td>3</td>
<td>22</td>
<td>3.44</td>
</tr>
<tr>
<td></td>
<td>One year</td>
<td>8.39</td>
<td>2</td>
<td>21</td>
<td>3.35</td>
</tr>
<tr>
<td>Hypopharyngeal</td>
<td>All</td>
<td>12.84</td>
<td>3</td>
<td>27</td>
<td>3.18</td>
</tr>
<tr>
<td></td>
<td>One week</td>
<td>12.82</td>
<td>5</td>
<td>21</td>
<td>2.83</td>
</tr>
<tr>
<td></td>
<td>One month</td>
<td>13.21</td>
<td>5</td>
<td>24</td>
<td>3.05</td>
</tr>
<tr>
<td></td>
<td>Three months</td>
<td>12.46</td>
<td>3</td>
<td>27</td>
<td>3.55</td>
</tr>
<tr>
<td>Vaginal</td>
<td>All</td>
<td>5.61</td>
<td>1</td>
<td>20</td>
<td>3.64</td>
</tr>
<tr>
<td></td>
<td>Week 36</td>
<td>5.05</td>
<td>1</td>
<td>18</td>
<td>3.12</td>
</tr>
<tr>
<td></td>
<td>Birth</td>
<td>11.92</td>
<td>5</td>
<td>20</td>
<td>3.17</td>
</tr>
</tbody>
</table>

Table 2: Richness of samples by closed reference OTU picking and culturing. Shown for all samples and split by sample type and time-point.
Figure 2: Average of sequenced bacteria also identified by culturing. Separated by sample type and how dominant the bacteria were. Bars indicate the three groups separately and the line shows the average percentage for each sample type.

Table 3: Bacteria identified by culturing, sequencing or both. The counts represent the number of bacteria identified per type of sample. + calculated as both/(both+cultured), * calculated as sequenced/(both+sequenced), $ show the percentage the bacteria identified by both methods from each class.

<table>
<thead>
<tr>
<th>Area</th>
<th>Time</th>
<th>Importance</th>
<th>cultured</th>
<th>Closed reference sequencing</th>
<th>both</th>
<th>cul found by seq+</th>
<th>seq found by cul*</th>
<th>Percent of positive$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Feces</td>
<td>One Week</td>
<td>All</td>
<td>151</td>
<td>4974</td>
<td>673</td>
<td>81.7%</td>
<td>11.9%</td>
<td>30.3%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Dominant</td>
<td>123</td>
<td>204</td>
<td>62.4%</td>
<td>30.9%</td>
<td>29.6%</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Major</td>
<td>445</td>
<td>199</td>
<td>30.9%</td>
<td>29.6%</td>
<td>29.6%</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Minor</td>
<td>4406</td>
<td>270</td>
<td>5.8%</td>
<td>40.1%</td>
<td>30.9%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>One Month</td>
<td>All</td>
<td>159</td>
<td>3161</td>
<td>488</td>
<td>75.4%</td>
<td>13.4%</td>
<td>37.1%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Dominant</td>
<td>55</td>
<td>181</td>
<td>76.7%</td>
<td>32.7%</td>
<td>18.2%</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Major</td>
<td>183</td>
<td>89</td>
<td>32.7%</td>
<td>18.2%</td>
<td>18.2%</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Minor</td>
<td>2923</td>
<td>218</td>
<td>6.9%</td>
<td>44.7%</td>
<td>32.7%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>One Year</td>
<td>All</td>
<td>305</td>
<td>3542</td>
<td>819</td>
<td>72.9%</td>
<td>18.8%</td>
<td>69.2%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Dominant</td>
<td>18</td>
<td>68</td>
<td>79.1%</td>
<td>8.3%</td>
<td>8.3%</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Major</td>
<td>137</td>
<td>184</td>
<td>57.3%</td>
<td>22.5%</td>
<td>22.5%</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Minor</td>
<td>3387</td>
<td>567</td>
<td>14.3%</td>
<td>69.2%</td>
<td>14.3%</td>
<td></td>
</tr>
<tr>
<td>Hypopharyngeal</td>
<td>One Week</td>
<td>All</td>
<td>187</td>
<td>5913</td>
<td>743</td>
<td>79.9%</td>
<td>11.2%</td>
<td>60.3%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Dominant</td>
<td>256</td>
<td>448</td>
<td>63.6%</td>
<td>60.3%</td>
<td>60.3%</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Major</td>
<td>642</td>
<td>136</td>
<td>17.5%</td>
<td>18.3%</td>
<td>18.3%</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Minor</td>
<td>5015</td>
<td>159</td>
<td>1.1%</td>
<td>21.4%</td>
<td>21.4%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>One Month</td>
<td>All</td>
<td>220</td>
<td>6812</td>
<td>966</td>
<td>81.5%</td>
<td>12.4%</td>
<td>41.1%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Dominant</td>
<td>469</td>
<td>348</td>
<td>42.6%</td>
<td>36.0%</td>
<td>36.0%</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Major</td>
<td>716</td>
<td>237</td>
<td>24.9%</td>
<td>24.5%</td>
<td>24.5%</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Minor</td>
<td>5627</td>
<td>381</td>
<td>6.3%</td>
<td>39.4%</td>
<td>6.3%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Three Months</td>
<td>All</td>
<td>246</td>
<td>6122</td>
<td>994</td>
<td>80.2%</td>
<td>14.0%</td>
<td>30.9%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Dominant</td>
<td>312</td>
<td>438</td>
<td>58.4%</td>
<td>44.1%</td>
<td>44.1%</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Major</td>
<td>729</td>
<td>249</td>
<td>25.5%</td>
<td>25.1%</td>
<td>25.1%</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Minor</td>
<td>5081</td>
<td>307</td>
<td>5.7%</td>
<td>30.9%</td>
<td>5.7%</td>
<td></td>
</tr>
<tr>
<td>Vaginal</td>
<td>Week 36</td>
<td>All</td>
<td>639</td>
<td>2277</td>
<td>568</td>
<td>47.1%</td>
<td>20.0%</td>
<td>91.7%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Dominant</td>
<td>14</td>
<td>22</td>
<td>61.1%</td>
<td>3.9%</td>
<td>3.9%</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Major</td>
<td>29</td>
<td>25</td>
<td>46.3%</td>
<td>4.4%</td>
<td>4.4%</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Minor</td>
<td>2234</td>
<td>521</td>
<td>18.9%</td>
<td>91.7%</td>
<td>18.9%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Week Birth</td>
<td>All</td>
<td>43</td>
<td>530</td>
<td>66</td>
<td>60.6%</td>
<td>11.1%</td>
<td>12.1%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Dominant</td>
<td>13</td>
<td>8</td>
<td>38.1%</td>
<td>12.1%</td>
<td>12.1%</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Major</td>
<td>34</td>
<td>14</td>
<td>29.2%</td>
<td>21.2%</td>
<td>21.2%</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Minor</td>
<td>483</td>
<td>44</td>
<td>8.3%</td>
<td>66.7%</td>
<td>66.7%</td>
<td></td>
</tr>
</tbody>
</table>
As shown by Rhoads et al. (2012), sequencing is the more sensitive method, identifying more bacteria per sample than culturing, despite only including sequences identical to the reference database sequences. While sequencing is more sensitive the resolution is not sufficient for all genera. In a clinical setting the difference between S. aureus and S. epidermidis is very relevant, but when sequencing the V4 region of 16S rRNA the two species are 100% identical. For Enterobacteriaceae the problem is more pronounced, as many of the genera cannot be separated.

**Sequencing an additional region of 16S rRNA.** The Illumina MiSeq platform has been updated to enable longer sequencing reads, this means that an extra variable region can be sequenced (Fadrosh et al., 2014). Using two variable regions, V3 and V4, the species in our reference database represented 127 unique sequences, compared to 103 when using only V4 (Supplementary table 4). This increased resolution was clear when looking at the species that, using only V4, had been grouped as S. aureus group and Enterobacteriaceae group A (figure 3). The species from Enterobacteriaceae all can be separated using an extra variable region. From the S. aureus group three species still had identical sequences (S. epidermidis, S. capatis, and S. caprae), but notably S. aureus was not identical to any other species.

![Figure 3](image-url)

The resolution of 16S rRNA for clinical relevant species was investigated by Woo et al. (Woo et al., 2007), based on their findings that sequencing only the 16S rRNA will not provide sufficient information to discriminate between all bacterial species. The number of sequence reads per sample on the Illumina MiSeq system is more than what is needed for identification in a clinical setting, so it would be possible to additionally sequence other household genes, e.g. HSP60 to identify *Staphylococcus* species (Kwok et al., 1999). With a well curated closed reference database containing sequences for clinical relevant bacteria in addition to sequencing of 16S rRNA, it would be easy to automatically get an output showing an overall picture of microbial community and accurate identification relevant bacterial species.
CONCLUSIONS

Comparison between HTS data and culture-dependent identification is very difficult when using de-novo OTU clustering for treatment of the sequencing data. One clear result from such analysis is that a large part of the bacterial composition is from bacteria that cannot be cultured under aerobic conditions. Using closed reference OTU picking, is closer to a use case scenario for diagnostic purposes, and using such methods we have shown that HTS is more sensitive than culturing, and identify five times more bacteria from fecal and hypopharyngeal samples, and twice as many from vaginal samples. When using HTS, sequencing just one variable region of 16S rRNA is not sufficient to separate the bacterial species, identified by culturing, to species level. We have shown that including an additional variable region increases the resolution sufficiently for almost all bacteria identified in this study and if necessary, increased resolution can be achieved by sequencing other household genes than just 16S rRNA.

These findings show that sequencing, as a technology, is ready for diagnostic purposes, but development of a standardized and automated pipeline is needed for sequencing to replace culture dependent methods.

METHODS

Ethics. The study follows the principles of the Declaration of Helsinki and was approved by the Ethics Committee for Copenhagen (The Danish National Committee on Health Research Ethics) (H-B- 2008-093) and the Danish Data Protection Agency (2008-41- 2599). Written informed consent was obtained from all participants. The study is reported in accordance with the STROBE guidelines(von Elm et al., 2007).

Study population. The novel Copenhagen Prospective Study on Asthma in Childhood 2010 (COPSAC2010) is an ongoing Danish cohort study of 743 unselected pregnant women and their children followed prospectively from pregnancy week 24 in a protocol largely similar to the first COPSAC birth cohort (COPSAC2000)(Bisgaard, 2004; Bisgaard et al., 2006, 2007). Recruitment lasted during 2009–10. Exclusion criteria were chronic cardiac, endocrinological, nephrological or lung disease other than asthma.

Sample collection. Vaginal swaps were collected at 36 weeks of pregnancy and at birth. Hypopharyngeal aspirates were collected at 1 week, 1 month, and 3 months after birth. Fecal samples were collected at 1 week, 1 month, and 12 months.(Bisgaard et al., 2013) All samples
were transported to Statens Serum Institut (Copenhagen, Denmark) where they were cultured and stored within 24 hours of sampling.

**Culturing.** Bacterial samples were cultured with standard methods on non-selective and selective media (SSI Diagnostica, Hillerød, Denmark). One set of blood agar plates and chocolate agar plates (both supplemented with with 5% horse blood) were incubated aerobically at 37°C for 18–20 hours. Another set of blood agar and chocolate agar plates were incubated under microaerophilic conditions (5% CO₂, 3% H₂, 5% O₂ and 87% N₂) at 37°C for 48 hours. Additionally for vaginal samples, one HBT (Human Blood Tween) agar plate was used for selection of Gardnerella vaginalis incubated at microaerophilic conditions at 37°C for 48 hours. Fecal samples were also cultured on an anaerobic plate, under anaerobic conditions at 37°C for 72 hours. Subsequently, microbial identification was performed according to growth on selective media, characteristics of colonies, and cellular morphology. Bacteria cultured anaerobically were not identified further. All bacteria identifications were confirmed biochemically by automated identification system VITEK 2 (BioMérieux, France). Isolates were preserved at -80°C for future identification. No quantification was performed.

**DNA extraction and 16S amplicon sequencing.** Genomic DNA was extracted using the PowerMag® Soil DNA Isolation Kit optimized for epMotion® (MO-BIO Laboratories, Inc., Carlsberg, CA, US) using the epMotion® robotic platform model (Eppendorf) according to manufacturer’s protocol. 150 µl was used from each sample. DNA concentrations were determined using the Quant-iT™ PicoGreen® quantification system (Life Technologies, CA, US). Extracted DNA was stored at -20°C.

The 16S rRNA gene amplification procedure was divided into two PCR steps. First, amplification of the hypervariable V4 region of the 16S rRNA gene, using the modified broad range primers 515F (5’-GTGCCAGCMGCCGCGGTAA-3’) and 806R (5’-GGACTACHVGGGTWTCTAAT-3’),26–28 then sequencing primers and adaptors were added to the amplicon products in the second PCR step. The amplification products were then purified with Agencourt AMPure XP Beads (Beckman Coulter Genomics, MA, US), pooled equimolar, concentrated using the DNA Clean & Concentrator™-5 Kit (Zymo Research, Irvine, CA, US), and the concentration were then determined using the Quant-iT™ High-Sensitivity DNA Assay Kit (Life Technologies). Paired-end sequencing, of up to 192 samples, were performed on the Illumina MiSeq System (Illumina Inc., CA, US), including 1.0% PhiX as internal control. All
Reagents used were from the MiSeq Reagent Kits v2 (Illumina Inc., CA, US). All details have been included in Supplementary Methods 1.

Fastq-files demultiplexed by the MiSeq Controller Software were trimmed for amplification primers, diversity spacers, and sequencing adapters (biopieces), mate-paired and quality filtered (usearch v7.0.1090, parameter: -maxee 0.5).

**De-novo OTU clustering.** Samples with less than 2000 reads were excluded from the analysis. This means that 1,301 fecal, 1,790 hypopharyngeal, and 722 vaginal samples were included to describe the microbiota, whereas just 3,552 samples, which had also been cultured successfully, were included for the comparisons. UPARSE was used for OTU clustering as recommended, in particular removing singletons after dereplication. Chimera checking was performed with usearch against the gold database as recommended. Representative sequences were classified (Mothur v.1.25.0, wang() function at 0.8 confidence threshold). Qiime wrappers for PyNAST, FastTree, and filter_alignment.py were used to construct a phylogenetic tree. Alignments were built against the 2011 version of Greengenes (parameters: --allowed_gap_frac 0.999999 and --threshold 3.0).

The rarefaction curves (Supplementary Fig. 3) show the observed richness and the Shannon diversity as functions of count of sequences. Shannon diversity curves reach asymptotes with 1,000 sequences, based on this all samples with less than 2,000 sequences were excluded. 3,813 samples were included containing an average of 56,979 sequences per sample (217,260,802 total), representing 511 distinct genera.

**Closed reference OTU picking.** Based on the cultured species, a reference database was created from the matching type strains in the RDP database (Cole et al., 2014). The database was rewritten into a Qiime-friendly format using a custom script. In silico PCR was performed using Mothur 1.33.2 with two mismatches allowed (pdiff = 2), selecting only species that could be amplified. Finally, reads were matched at 100% using QIIMEs pick_closed_reference_otus.py (parameter: enable_rev_strand_match = True). Some of the bacteria had more than one type strain, but with the amplified region of all strains included, we had 101 unique sequences (supplementary table 1).
The abundances of the bacteria were calculated as the percent of all reads, including those, which did not match to the reference database. Based on their abundances the bacteria were classified as dominant bacteria (abundance > 10%), major (1-10%), or minor (<1%).

REFERENCES


**SUPPLEMENTARY MATERIAL**

Supplementary tables are available online at: [http://goo.gl/bx6Q1E](http://goo.gl/bx6Q1E)

**Supplementary table 1**: Full list of cultured bacteria and the percentage of samples they were isolated from.

**Supplementary table 2**: Typestrains used for the reference database, including the sequence of their V4 region, and new naming for species with identical sequences.

**Supplementary table 3**: Cultured species not included in the reference database, their counts, the percentage of counts they represent from their genus and all samples.

**Supplementary table 4**: Bacteria identified by culturing and their naming when group with identical sequences from their V4 region and V3-V4 regions.