The use of 16S rRNA targeted next generation sequencing in diagnostics of polymicrobial invasive infections

Ruben Dyrhovden

Thesis for the degree of Philosophiae Doctor (PhD)
University of Bergen, Norway
2023

UNIVERSITY OF BERGEN

The use of 16S rRNA targeted next generation sequencing in diagnostics of polymicrobial invasive infections

Ruben Dyrhovden

Thesis for the degree of Philosophiae Doctor (PhD)
at the University of Bergen

Date of defense: 21.08.2023

Year: 2023

Title: The use of 16S rRNA targeted next generation sequencing in diagnostics of polymicrobial invasive infections

Name: Ruben Dyrhovden

Print: Skipnes Kommunikasjon / University of Bergen

Contents

Scientific environment

This work was done at the Department of Microbiology, Haukeland University Hospital Bergen, Norway.

Øyvind Kommedal (UiB / Haukeland University Hospital) was the main supervisor and Elling Ulvestad was co-supervisor. Scientific support was also provided by the staff at the Department of Microbiology.

The project was financed by the Western Norway Regional Health Authority (Helse-Vest) by a 3-year PhD fellowship.

This thesis is a part of the PhD program at the Department of Clinical Medicine (K2), Faculty of Medicine, University of Bergen.

Acknowledgements

The foundation for this work was laid in 2014 when I, as an intern at the Department of Surgery at Haukeland University Hospital, attended a lecture by Øyvind Kommedal on next generation sequencing. The inspiring lecture created an interest in and a desire to learn more about the field, and in turn led to me taking courage and visiting Øyvind Kommedal in his office a few weeks later. I did not imagine then that this first visit would lead to a doctoral thesis.

I would like to express my deepest gratitude to Øyvind Kommedal. Working with you has been a pleasure. Your great knowledge and understanding of the field, your selfless willingness to share this knowledge, and your unique ability to lift your gaze from details and identify new hypothesis-forming explanations and patterns, have been invaluable. Thank you for all the productive and less productive conversations, for your optimism, and for always taking time and being willing to answer all kind of questions.

Also, this endeavor would not have been possible without my co-supervisor Elling Ulvestad. You have the ability to widen the horizon and see things from different perspectives. I have enjoyed your reflections and constructive comments, as well as your ability to find and highlight the common thread of all the unfinished text I have sent you for review. Your research competence and thoroughness are admirable, and I am constantly amazed at how unfinished paragraphs and texts that I in pure resignation have sent to you and Øyvind have been turned into gold when they are returned.

I am also grateful to Randi Monsen Nygaard. You laid an important foundation for my work through your master's degree. Just as importantly, you patiently taught me, then a clumsy novice who had barely touched a pipette in my life, the necessary skills to master the laboratory work required for my thesis.

To my co-authors, Martin Rippin, Kjell Øvrebø, Robin Patel, and Magnus Vie Nordahl: Many thanks to all of you for the collaboration and help in different parts of my Ph.D. project!

A special thanks also to all my colleagues and the staff at the Department of Microbiology, both for all professional conversations, but not least for making time at work socially pleasant. I do not have a count of how many times we have solved both everyday and world problems during a simple lunch.

Many other people have worked with me on this project as well, and I am grateful and will express my deepest appreciation to all the clever and helpful people who have helped me along the way.

I would also like to thank the Western Norway Regional Health Authority for my 3-year PhD fellowship grant.

To my parents Kari and Sigmund. I am deeply grateful for the way you raised me. Thanks for always taking your time for me and my siblings and for always being supportive. You have given me values of life and a faith that I hope to pass on to my own children, and which I am sure has helped me also through this Ph.D. period. Thank you both for all the conversations, practical help, and support also during this, at times, busy period. Thanks also to my siblings, Mirjam, Joel and Sigve, and your families. It is always a pleasure to meet you and spend time together!

Finally, Gro. The love of my life and the mother of our three beloved children, Samuel (10), Rebekka Sofine (6) and Maria (4). Thank you for being my wife, friend, and support through all of life's joys and difficulties. I look forward to the continuation!

Abbreviations and definitions

ASV	Amplicon Sequence Variant
DNA	Deoxyribonucleic Acid
ssDNA	Single stranded DNA
dsDNA	Double stranded DNA
NGS	Next Generation Sequencing
OTU	Operational Taxonomic Unit
PCR	Polymerase Chain Reaction
RNA	Ribonucleic Acid
rpoB	Ribonucleic Acid Polymerase Beta Subunit
TNGS	Targeted Next Generation Sequencing
16S rRNA	16 Svedberg Ribosomal Ribonucleic Acid
Microbiota	The microorganisms of a particular site, habitat, or geological period.
Microbiome	The combined genetic material of the microorganisms in a particular environment.
AC	Acute cholangitis
NIBDS	Non-infectious bile duct stenosis

Abstract

As made evident from scientific investigations, 16S rRNA targeted next generation sequencing (TNGS) enables a more complete characterization of complex bacterial microbiotas than what can be obtained by standard culture-based microbiological techniques. Such data suggest that TNGS could be of use in the clinical laboratory as well, but so far, few studies have explored the adequacy of this method in the diagnostics of patients suffering from polymicrobial infections.

The main objective of this thesis was to investigate the use of 16S rRNA TNGS in microbiological diagnostics of polymicrobial invasive infections. As part of this endavour, we also wanted to evaluate the effect of supplementary *rpoB* gene TNGS on species-level resolution, explore the patterns of background contamination and suggest transparent approaches for management of DNA-contamination in post-sequencing processing and interpretation in a diagnostic setting.

Our results confirm the improved sensitivity of 16S rRNA TNGS as compared to traditional diagnostics for polymicrobial invasive infections. We also demonstrate the utility of *rpoB* sequencing, which provides more accurate species identifications for several clinically important genera. Upon exploring the unpredictable nature of background contamination in TNGS, we suggest and evaluate a method for managing the contamination, including rules and cutoffs for post-sequencing processing and interpretation to maximize accuracy of the results. The data also provide new insights into the pathogenesis of polymicrobial infections.

Our results thus demonstrate that methodological challenges inherent to TNGS can be overcome, and that TNGS may be useful for diagnostics of polymicrobial infections in individual patients.

Sammendrag

16S rRNA dypsekvensering muliggjør, som vist i flere vitenskapelige arbeider, en mer fullstendig karakterisering av komplekse bakterielle mikrobiota enn det som kan oppnås ved tradisjonelle, dyrkningsbaserte, mikrobiologiske teknikker. De vitenskapelige arbeidene indikerer også at 16S rRNA dypsekvensering kan være nyttig i kliniske mikrobiologisk diagnostikk, men så langt har få studier undersøkt egnetheten til denne metoden i diagnostikk av pasienter med polymikrobielle infeksjoner.

Hovedmålet med denne oppgaven var å undersøke bruken av 16S rRNA dypsekvensering i mikrobiologisk diagnostikk av polymikrobielle, invasive infeksjoner. Som en del av dette arbeidet ønsket vi også å evaluere effekten av supplerende dypsekvensering av *rpoB*-genet for å oppnå bedre oppløsning på artsnivå, utforske mønstrene ved DNA-kontaminasjon og foreslå transparente metoder for håndtering av DNA-kontaminasjon og tolkning av dypsekvensseringsdata i en diagnostisk setting.

Resultatene våre bekrefter den forbedrede sensitiviteten til 16S rRNA dypsekvensering sammenlignet med tradisjonell diagnostikk av polymikrobielle invasive infeksjoner. Vi demonstrerer også nytten av *rpoB*-gen dypsekvensering, som gir en mer presis artsidentifikasjon innen flere klinisk viktige bakterieslekter. Vi utforsker og beskriver den uforutsigbare naturen til DNA-kontaminasjon, og foreslår og evaluerer en metode for å håndtere denne kontaminasjonen og dermed bedre nøyaktigheten av dypsekvenseringsresultatene. Arbeidet gir også ny innsikt i patogenesen til polymikrobielle infeksjoner.

Resultatene våre viser at iboende metodiske utfordringer ved 16S rRNA dypsekvensering kan overvinnes, og at dypsekvensering kan være nyttig ved diagnostikk av polymikrobielle infeksjoner hos den enkelte pasient.

List of Publications

I. Dyrhovden R, Nygaard RM, Patel R, Ulvestad E, Kommedal O. “The bacterial aetiology of pleural empyema. A descriptive and comparative metagenomic study”. Clinical Microbiology and Infection 25.8 (2019): 981-986.

II. Dyrhovden R, Ovrebo KK, Nordahl MV, Nygaard RM, Ulvestad E, Kommedal O. “Bacteria and fungi in acute cholecystitis. A prospective study comparing next generation sequencing to culture”. Journal of Infection 80.1 (2020): 16-23.

III. Dyrhovden R, Rippin M, Ovrebo KK, Nygaard RM, Ulvestad E, Kommedal O. “Managing contamination and diverse bacterial loads in 16S rRNA deep sequencing of clinical samples - implications of the law of small numbers”. MBio 12.3 (2021): e00598-21.

1 Introduction

1.1 Background

Polymicrobial infections represent a particular challenge in culture-based diagnostic microbiology (1, 2). First, culture tends to facilitate growth of the subset of microbes that thrive on artificial media, thus outcompeting more demanding microbes. Second, anaerobic bacteria are difficult to keep alive during sample transportation, and some are not cultivable using the standard media and conditions provided in the routine laboratory. Third, culture-dependent diagnostics is generally limited by the fact that only viable microbes can be detected. The sensitivity of the method is therefore dramatically reduced for samples collected after the initiation of therapy. Such challenges has encouraged a quest for alternative diagnostic methods.

Universal amplification of the bacterial 16S rRNA gene directly from clinical samples followed by Sanger sequencing has been available as a culture-independent method in diagnostic bacteriology for more than 20 years. Despite initial high hopes, a relatively low sensitivity and a limited potential for resolving polymicrobial infections are hampering the usefulness of this approach.

The more recent development of targeted next generation sequencing (TNGS) has resolved many of the issues related to universal amplification of the 16S rRNA gene. Development of TNGS has been driven by microbiome and microbiota research, focusing on characterizations of the microbial flora in healthy individuals (Human Microbiome Project, http://www.hmpdacc.org/) as well as on how the gut and airway microbiota associate with various human diseases (3).

Several reviews have discussed the implementation of the method in microbiological routine diagnostics (4, 5) and there has been a gradual increase in TNGS-studies on infectious disease materials (1, 2, 6–10). Neverhteless, there are still surprisingly few studies exploring the use of TNGS in diagnostic microbiology (4).

The published research on the use of TNGS in clinical microbiology demonstrate the methods's diagnostic benefits as compared to traditional methods. In their study of human brain abscesses, Kommedal *et al*. found that culture and 16S Sanger sequencing using group-specific broad range PCR primers identified only 31% and 61% respectively of the bacteria identified by TNGS (1). The improved identification by TNGS also enabled discovery of three candidate key pathogens responsible for the establishment of primary polymicrobial brain abscesses. Bryan *et al.* used the method for diagnosing a young patient with an intra-abdominal infection of uncertain aetiology in which neither culture nor direct 16S rRNA Sanger sequencing allowed detection of the uropathogen *Actinotignum schaalii* (11)

Several challenges associated with TNGS have implications for the method's diagnostic sensitivity and specificity when used to investigate infectious agents. These challenges need to be properly addressed prior to implementing the method in diagnostic practice.

A major challenge – related to the method's high sensitivity – is the risk of bacterial DNA contamination during analysis, which reduces the method's diagnostic specificity and may lead to false positive results (12). This is particularly relevant for clinical samples with low microbial concentrations, where contaminating sequences may constitute most of the sequencing reads in the unfiltered sequencing results (13–15).

Another challenge – related to the method's capacity for resolution – is to obtain unambiguous species-level identification of detected microbes. Whereas identification to the genus or family level is often considered sufficient in microbiota research, species-level identification is normally required in clinical settings. Unfortunately, the 16S rRNA gene, by far the most dominant marker gene used for bacterial identification in TNGS, displays a too low inter-species variability to distinguish between several important infectious pathogens (16).

A third challenge - related to the method's clinical specificity (17) - is how to determine whether the detected sequence is causally related to the patient's current disease process. In accordance with Koch's postulates, identification of a microbe at the site of infection is not sufficient evidence to determine whether or not the microbe is associated with initiation or progression of the disease. The challenge related to clinical specificity has been highlighted by the use of sequencing technologies to explore the human microbiome, which demonstrates that body areas, including surfaces previously thought to be sterile such as the lower respiratory tract (18), can be the natural habitat for a large number of different microbes (19). Without knowledge of the normal microbial flora – or microbiota – at the site of infection, it is impossible to determine whether an identified microbe is causally related to the infection or not.

Finally, challenges associated with higher analysis costs, extensive workflows which often take many days to complete, and the need for specialized bioinformatics knowledge to process and interpret the sequencing results, may hamper the introduction og TNGS in clinical microbiology laboratories (5).

To conclude, there is a need for more studies exploring the use of TNGS for the diagnostics of polymicrobial infections. This need is emphasized by promising results from published studies, by potential benefits gained by better diagnostics of polymicrobial infections, and by methodological challenges inherent to traditional diagnostic practice. In this thesis, we have explored some of the above topics by performing TNGS on systematically collected samples from three different types of invasive polymicrobial infections: pleural empyema, acute cholecystitis and acute cholangitis. We focused on the performance of TNGS compared to traditional microbiological diagnostics, and on how to approach the two major challenges – precise species level identification and DNA contamination.

1.2 Polymicrobial infections

A polymicrobial infection is an infection caused by two or more microorganisms, sometimes displaying combinations of viruses, bacteria, fungi and parasites (20). Recent research has demonstrated that the frequency of polymicrobial infections is far higher than previously recognized, and that different species play different roles during the establishment and maintenance of polymicrobial infections (21, 22). Kommedal *et al*. suggested that certain bacteria are key pathogens for the establishment and development of polymicrobial brain abscesses (1), but wheter this result is representative for other types of polymicrobial invasive infections has so far not been ascertained.

1.3 Sequencing

Sequencing techniques identify bacteria by their DNA (23). These techniques are broadly categorized as either low-throughput sequencing (e.g. Sanger sequencing and pyrosequencing), high-throughput sequencing (e.g. next/second generation sequencing (NGS) or third-generation sequencing (long read sequencing). In theory, the techniques should represent a more sensitive alternative to culture for identification of fastidious, anaerobic, and non-viable bacteria.

1.3.1 Sanger-sequencing

Broad-range amplification of the bacterial 16S rRNA gene directly from clinical samples followed by Sanger amplicon sequencing, is widely adopted in diagnostic laboratories. Although mostly displaying a higher diagnostic sensitivity than culture, the technique has major limitations when it comes to diagnosing polymicrobial infections (24, 25). First, polymicrobial infections produce mixed Sanger chromatograms that may be impossible to interpret (24, 25). Second, the magnitude of signals from bacteria present in lower concentrations can be completely outcompeted by the signals from more dominant species, rendering them undetectable in the chromatograms. To reduce the effects of such factors, Kommedal *et al*. suggested to replace the single universal 16S rRNA PCR with a set of group-

specific broad-range PCRs, thus increasing the potential number of bacterial species to be identified in polybacterial samples from three to nine (25). Even though this modification represented a clear improvement of sensitivity, it remained insufficient for complex infections with a broad diversity of microbal species. The senstivity was also insufficient for detection of bacteria present at the lowest concentrations within each group.

1.3.2 Targeted next generation sequencing

The introduction of the pyrosequencing platform 454 Life Sciences (Branford, CT, USA) in 2005, initiated a revolution in DNA sequencing (23, 26). In the following years, several new sequencing platforms were developed and launched (e.g. Illumina and IonTorrent) (27). These techniques are capable of sequencing millions of small DNA fragments in parallel and are thus referred to as massive parallel sequencing or "next generation sequencing" (NGS).

A main NGS-application, targeted next generation sequencing (TNGS), uses PCR to amplify specific DNA sequences that are subsequently sequenced (5). Targeted next generation sequencing has revolutionized human microbiota research and is, together with whole genome shotgun sequencing, by now the major platform for research in descriptive microbiology (Human Microbiome Project, http://www.hmpdacc.org/).

In the present investigations we have used the Illumina MiSeq system, one of the major platforms for TNGS. The system utilizes paired-end sequencing and delivers sequences with a read length up to 2x300 base pairs (bp), thus enabling complete sequencing of amplicons up to around 500 basepairs. The process involves three sequential steps – libarary preparation, sequencing and data analysis.

1.3.2.1 Library preparation

Libarary preparation consists of two sequential PCRs (28). First, the targeted DNA-sequence is amplified by PCR (amplicon PCR). The primers used for this amplification have a dual function. In addition to being directed at the target DNA-sequence, the 5'end of the primers have an adapter sequence that is complementary to primers used for the subsequent PCR. Consequently, all amplicons have the same

adapter sequences at their ends. The next step is the index-PCR, where the amplified products from each sample are marked with uniqe sequences (indexes). This makes it possible to mix amplicons from multiple samples into a single library pool and still be able to separate the results from the individual samples during the data analyzes. The primers used for the index-PCR are directed towards the adapter-sequences, enabling indexing of all the amplified DNA from the amplicon-PCR. A dual indexing strategy is used, meaning that a uniqe combination of two different indexes (one at the 5'end and one at the 3'end) is added to each sample. The dual indexing strategy reduces the risk of index hopping/swithing, a phenomenon where reads are assigned to the wrong sample during sequencing, which is more likely to occur if only a single index is added to the DNA fragments. Finally, the 5'end of the primers used for the index-PCR contain adapters complementary to the oligonucleotides attached to the flow cell.

1.3.2.2 Illumina MiSeq Sequencing

Next, the dsDNA fragments from the library pool are denatured into ssDNA templates and loaded into a flow cell where the sequencing takes place. The ends of the ssDNA templates bind to complementary oligonucleotides attached to the inside surface of the flow cell where all ssDNA are multiplied by an isothermal DNA-polymerase in a process named bridge amplification. Each attached ssDNA is thereby transformed into a cluster of identical DNA-templates attached throughout the flow cell. The Illumina MiSeq system uses a sequencing by synthesis technology. The sequencing primers are targeted at the 3'end of the adapters. The sequences are then re-amplified with nucleotides (A, C, G, T) that are labeled with a distinct fluorophore in addition to a chemically inactivated 3'OH group. During sequencing, a single base is incorporated into the growing DNA chain per cycle. For each cycle the incorporated fluorophore is read before the fluorescent group is cleaved off and the 3'end is reactivated.

1.3.2.3 Post-sequence processing and data analysis

The post-sequencing bioinformatics processing consists of several steps with the goal of providing the most accurate taxonomic assignment possible for all sequencing

reads representing true biological signals. First, sample-separated raw sequencing files, most often in fastq format, are generated by demultiplexing of the MiSeq raw data based on the sample barcodes. Then three main steps follow (29): 1) Quality filtering and pre-processing of the sequencing raw data. 2) Either clustering of sequencing reads into operational taxonomic units (OTUs) or identification of exact amplicon sequencing variants (ASVs) after removal of error-containing sequences by denoising algorithms. 3) Taxonomic assignment of the OTUs/ASVs.

1.3.2.3.1 Quality filtering and pre-processing

This step includes demultiplexing of raw sequencing data, trimming of adapter sequences and eventual low-quality bases toward the sequence 3-end, filtering of short and low-quality reads and merging of paired end overlapping reads into a single, higher quality consensus read (29). Multiple software-systems which achieve different parts of these tasks are available (30–33). We used Adapterremoval (34), a software capable of performing all of the above-mentioned tasks.

1.3.2.3.2 OTU and ASV

The qualityfiltered and merged fastq-files contain a huge amount of sequencing reads representing the microbial taxa in the sample. The next step is therefore to group the sequences into clusters with the ideal intention that all sequences within a cluster is representative of a single species/the same species. Two main methods are used for this purpose; either clustering of the reads into *de novo* operational taxonomic units (*de novo* OTUs) based on a percent sequence similarity threshold, or the removal of erroneous sequences generated during PCR and sequencing followed by the identification of exact sequence variants (or ASVs) where only identical sequences are clustered together.

A third method for clustering is to map all the reads against a reference sequence database (closed-reference OTUs). In this approach any read representing a species that is lacking in the reference database is removed, and the sensitivity of the method is therefore limited by the content of the reference database.

In *de novo* OTU clustering, sequences within a specified sequence similarity are grouped together. Annotation is done by selection of a single sequence as a representative for all sequences in the cluster. Ideally this sequence should represent

the most common sequence type in the cluster. The most used sequencing similarity threshold for 16S rRNA sequencing within microbiota research is 97%. This threshold is based on the observation by Stackebrandt *et al*. that a 70% reassociation value by DNA-DNA hybridization, at that time the gold standard for species definition, corresponded to a 16S rRNA similarity of 97% or higher (35). Later it has been shown that a similarity threshold of 97% is too low and often leads to the inclusion of multiple species into the same OTU, and that a more conservative threshold of 99% is needed for OTUs to approximate the species concept (36, 37).
A challenge with the use of *de novo* OTU clustering is that in general, due to PCR and sequencing errors, the number of estimated OTUs will be higher than the real number of species (29). Sequences with an error rate above the chosen sequence similarity threshold will always introduce spurious OTUs (38). The primary sources of error are the error rate of the PCR polymerase, the formation of chimera during PCR amplification and, finally, errors introduced during sequencing, e.g. difficulties in accurate sequencing of stretches of DNA with the same base (homopolymers) (39). Approaches to handle these spurious OTUs include the application of various denoising algorithms and chimera removal tools following the *de novo* OTU clustering (38–40). Examples of denoising algorithms is the simple removal of low frequency OTUs or more advanced algorithms identifying sequencing reads as errors if they appear in low frequency together with a high frequency (dominant) highly similar OTU (38–40).
The ASV method differs from the *de novo* OTU clustering method in two major aspects. First, the construction of ASVs includes a *de novo* process where error sequences are removed from the data. This is done by the application of a denoising algorithm that is based on the expectation that biological true sequences are more likely to be repeatedly observed than error sequences (41). Second, following the denoising, the remaining reads are grouped based on 100% homology so that one group represents a single amplicon sequence variant (ASV). All ASVs may then be matched against a reference database, and ASVs that matches with the same species can be grouped together.

1.3.2.3.3 Taxonomic assignment of OTUs/ASVs

The final step in processing of sequencing data is to assign the OTUs/ASVs to taxonomic units. Taxonomic assignement is done by comparing the OTUs/ASVs to a reference database containing sequences of known and preferably well-described species. The keys to achive as the most accurate taxonomic assignment possible are I) an accurate and effective method for comparison of the query to the reference database, II) a high quality reference database and III) knowledge of the inherent limitations of the 16S rRNA when it comes to providing species level resolution among some closely related bacteria:

I) The sequencing-alignment-based method – in which derived sequences are compared directly to sequences in a database, is considered the gold standard method for sequencing comparison. The most common tool used for the sequencing-alignment-based method are different variants of the Basic Local Alignment Search Tool (BLAST) (42). This tool enables the search for similarity matches to a query sequence. The main disadvantage of BLAST-based approaches is that they require a lot of computing power, especially when working with large amounts of data. As a response, other taxonomy classifier softwares have been developed based on alignment-free algotritms (29). The most common are k-mer based algoritms which compare the frequency of k-mers between the query and the database sequences such as naïve Bayesian RDP classifier (43) and SPINGO (44). Because k-mer based algoritms rely on a proxy measurement of the sequence similarity between the query and the database sequence, it is inherently less accurate than the gold standard BLAST sequencing-alignment-based method (45). Gao and colleagues illustrate this in the publication of their sequencing alignment-based Bayesian based Lowest Common Ancestor algoritm (BLCA), which they found to significantly outperform k-mer based methods in accuracy of species-level classification (45). However, the higher accuracy achieved by the BLCA comes with the cost of a long computation time. The most used tools for taxonomic assignment in microbiome research are therefore softwares combining

alignment-free and alignment-based algoritms. An example of the latter is VSEARCH (46) which compares sequences in two phases; first by an initial filtering based on k-mers, followed by optimal global alignment of the query with the most promising candidates. The QIIME2 project has developed the q2-feature-classifier (47) that allows the researcher to choose between a novel machine-learning k-mer based taxonomy classifier and two alignment-based classifiers based on BLAST+, an improved version of the BLAST software (48) and VSEARCH (46). In their publication Bokulich *et al.* find that all three classifiers implemented in the q2-feature-classifier meet or exceed the species-level accuracy of other commonly used methods (47).

II) Choosing a a high quality database is crucial for correct taxonomic assignment of sequences. Reference databases can be divided into those that are curated and those that are not. Curated databases have undergone some sort of quality filtering, thus ensuring the correctness and quality of the reference sequences and their annotation. The largest uncurated sequencing database is Genbank (National Center for Biotechnology Information, NCBI), which is an annotated collection of all publicly available DNA sequences uploaded to either Genbank, DNA DataBank of Japan (DDBJ) or the European Nucleotide Archive (ENA). Genbank is unreliable – it contains thousands of identical sequences for some organisms, many references are misannotated, the annotation style is inconsistent (49), many of the sequences contain errors or represent chimeras (50) and a huge proportion of the uploaded 16S rRNA genes are from uncultured organisms (49). Several more curated databases have been developed for 16S rRNA gene microbiota analyses. The most popular include Greengenes (49), SILVA (51) and the Ribosomal Database Project (RDP) (52). Although these curated databases provide more reliable taxonomic classifications of sequences than the uncurated alternatives, they are still associateed with a significant level of uncertainty. First, the

databases are rarely updated and therefore will not include newly discovered species or taxonomic updates. Second, the error rates of the databases are still quite high. For example, Edgar (53) found an annotation error rate of approximately 10% in the RDP database and approximately 17% in the Greengenes and SILVA databases. Others have found that use of multiple curated databases lead to conflicting results, escpecially among less common genera (54, 55).

III) As elaborated in chapter 1.4.1, the inherent limitations of the 16S rRNA gene implies that a species level identification cannot be made for some bacterial families and genera based on the 16S rRNA gene alone. Avoidance of misclassified OTUs/ASVs thus require that taxonomic assignments of sequences are based on well-founded critariae for 16S rRNA sequence interpretation that includes not only specific cutoffs for homology with a high-quality reference sequence but also a cutoff for minimum sequence difference to alternative species (37, 45). If a taxonomic classifier software only reports the best match, independent of distance to the next alternative species, the user should have knowledge of which groups of species that cannot be reliably distinguished based on the 16S rRNA gene. If for example taxonomic assignment of a 16S rRNA sequence gives 100% match whith *Streptococcus mitis*, this should be altered to *Streptococcus mitis/oralis* group since the 16S rRNA gene cannot be used to distinguish species within this group.

1.3.2.3.4 DNA contamination

A crucial step in the post-sequence processing and data analyzes is to identify and filter contaminating DNA (12, 13, 56). Contaminating DNA can be defined as sequence reads from microbes that were not originally part of the sample. The source of contamination can be divided into background contamination and cross-contamination (13). Background contamination inludes DNA introduced during the sequencing processing, from sources like extracion reagents, plastic consumables and laboratory enviroment (13). Cross-contamination includes all transfer of DNA and

barcodes from neighboriging wells or tubes during PCR and sequencing processing as well as contamination occuring on the sequencing instrument either from barcode sequencing error, contamination from residual amplicons from past sequencing runs or index switching (13). Background contamination and cross-contamination are most prominent in low-biomass samples (12–15). Background contamination originating from extraction reagents appears to be the main source of DNA contamination and is therefore the type of contamination having the most significant impact on the sequencing results (12, 15).

There are multiple reports of how failure to properly identify and filter contaminating DNA may lead to misidentification of microbes or distinct microbial communities, which in turn can lead to the formation of new theories about the aetiology and pathogenesis of various diseases based on false grounds (13, 57). There is however no gold standard method for the management of contamination in deep sequencing studies. Whereas some researchers simply remove all sequences also found in the negative controls (2, 6, 11, 58), others utilize more advanced algoritms including pattern recognition (14, 59). The latter often include an extensive use of negative and positive controls. In microbiome research, the decontam tool is currently the most promising pattern recognition based method for filtering contamination (59). However, even decontam displays a rather poor specificity when used on samples with very low biomass (15) and like other pattern recognition methods they are dependent upon large batches of samples in order to identify significant patterns. These challenges makes transfer of the method from microbiome research to diagnostic settings challenging. In diagnostic laboratories, the focus on time to results and cost implies there may not always be room for sequencing large batches of samples and controls. Further, the focus is always on the individual patient and a high sensitivity and specificity is required to avoid false negative and false positive results.

1.3.2.3.5 RipSeq

The RipSeq NGS software (Pathogenomix, Santa Cruz, CA), the bioinformatic tool used for most of the post-sequencing processing in the three studies of this thesis, has been developed specifically for use in diagnostic microbiology and possesses some features that are advantageous in a diagnostic setting.

1.3.2.3.5.1 Clustering

The RipSeq NGS Software is OTU-based. The *de novo* clustering of OTUs is done by loading merged FASTQ files into the *RipSeq NGS Preprocessor,* a software installed on the local computer. The preprocessor provides several options for quality filtering of the FASTQ files before clustering. These include

1. checking and trimming for primers and eventually trimming of ends with no 3'end primer,
2. setting a lower sequence-length threshold for sequences to be included in the clustering,
3. setting a lower threshold for number of sequencing copies in a cluster to be included in the final results
4. setting a homology threshold for clustering.

1.3.2.3.5.2 Taxonomic assignment

The most prevalent sequence type from each of the *de novo* clustered OTUs is uploaded from the RipSeq NGS preprocessor to the RipSeq NGS web software for taxonomic assignment. The taxonomic assignment is done by a sequence alignment-based method using a BLAST variant against a curated or semi-curated database according to the user's preferences. The RipSeq *Pathogenomix Prime 16S* database that is recommended by the manufacturer for 16S analysis is a curated database which includes about 2500 manually curated references, all references from GenBank 16S RefSeq database, all type-strain references from GenBank, extracted 16S rRNA references from all GenBank complete genome references and all references in the Human Oral Microbiome database. The inclusion of the HOMD database also provides a biologically sound taxonomy for known but hitherto undescribed bacterial members of the human microbiota.

The vast majority of human pathogens have been included in the above mentioned databases either by type strains or by full genomes. In most circumstances the *Pathogenmix Prime 16S* database will therefore be broad enough to caracterize even the most complex clinical samples. If there are no matches to a species using RipSeq *Pathogenomix Prime 16S* database, queries against other reference databases

containing environmental sequences from uncultured bacteria should be performed. In the RipSeq web software the alignment results are listed hierarchically according to their sequence similarity with the query and every reference sequence and alignment can be analyzed and checked manually if needed. The software also flags the quality of each species assignment based on researcher-defined thresholds for % homology with best reference and % distance to the next alternative species.

1.3.2.3.5.3 Chimera check

The RipSeq web software includes the option to perform a chimera check following the annotation of the OTUs. The chimera check is based on the assumption that all chimeric OTUs are constructed of two or more species that are also identified from non-chimeric OTUs in the same sample. After an initial round of idenfication, all OTUs with a score below the genus level will be checked to see if they represent a construct of two or more OTUs that has been identified to the species level in the same sample. If this is the case, they will be flagged as potential chimeras and the program will also report the species that are involved in each one of the identified chimeras.

1.3.2.3.5.4 Management of contamination

The RipSeq web software includes an option to mark files as sample or negative control and to submit CT values of samples. This can be used for later marking of or automatic filtering of OTUs found in the negative controls or marking of or automatic filtering of OTUs below a CT-value based threshold (1).

1.4 Selection of marker genes for identification of bacteria

An ideal marker gene should enable the identification of all present microbes to the species level. Several characteristics are needed for such a marker gene (60); i) it must be found in all species within the kingdom; ii) it should function as a molecular chronometer useful for measuring phylogenetic relationships (61). Changes in the gene sequence must occur randomly in a clocklike manner and at a mutation rate high enough to provide discrimination to the species level for all species within the kingdom. At the same time the mutation rate must be slow enough to secure well

defined populations within a species over time (61); iii) the amplicon must be large enough, or contain enough functional domains, to provide adequate amounts of information and not be vulnerable to non-random changes in the sequence (61); iv) it must contain DNA-regions that are highly conserved and found in all species in the kingdom. These highly conserved regions must flank the variable areas to function as universal targets for the PCR amplification primers.

1.4.1 The 16S rRNA-gene

The 16 Svedberg(S) ribosomal ribonucleic acid (rRNA) gene, first described as an evolutionary marker by Woese et al. in 1977, fulfills most of the criteria for an ideal gene target for universal identification and classification of bacteria (62, 63). Since DNA sequencing gradually became more available in the 1990s and 2000s, the 16S rRNA gene has obtained an increasingly dominant role in bacterial species identification (16, 60, 64). The 16S rRNA gene is so far the only gene target that has been used for universal bacterial detection and identification in targeted next generation sequencing studies of clinical infectious material (1, 2, 6–8).

The 16S rRNA gene codes for a component of the prokaryotic 30S ribosomal small subunit, has a length of around 1500 base pairs, and contains nine “hypervariable regions” (V1-V9) flanked by short stretches that are highly conserved in most bacteria. The “hypervariable regions” are considered chronometers that can be used for inferring phylogenetic relationships and species assignments, while the conserved regions are ideal targets for universal primers. The hypervariable regions of different bacterial species exhibit different degrees of sequence-variability, and the regions can therefore vary in their ability to discriminate between different bacterial species (16). The Clinical and Laboratory Standard Institute (CLSI) has provided guidelines for 16S rRNA sequence interpretation (37). For the ~500 basepair long V1-V3 segment, ≥ 99% homology with a high-quality reference sequence combined with a minimum distance of >0.8% to the next alternative species is recommended for species level identification, and ≥97% homology with a high-quality reference for genus-level identification (37). Such cutoffs may be more conservative for regions with a lower variability and less conservative for regions with a higher variability (65).

Some bacteria display too low inter-species variation in their 16S rRNA genes to be unambiguously distinguished by 16S rRNA sequencing (16). Genera comprising species that can be difficult/impossible to distinguish at the species level include multiple genera within the *Enterobacteriaceae* family*, Staphylococcus, Streptococcus, Enterococcus* and *Mycobacteria* (16, 63).

1.4.2 The *rpoB* gene

The *rpoB* gene found in all bacteria, codes for one (the β subunit) of the five subunits building up the core enzyme of the RNA polymerase (66). *RpoB* gene sequence similarities correlate better with DNA–DNA hybridization values (DDH) than the 16S rRNA gene (67). The hypervariable regions of the *rpoB* gene have a higher mutation rate than the 16S rRNA gene, and provide a species resolution for many bacteria where 16S rRNA can only discriminate to the genus or family level, such as *Enterobacteriaceae* (68)*, Staphylococcus* (69)*, Streptococcus* (70) and *Enterococcus* (70). Like the 16S rRNA gene, the *rpoB* gene displays characteristics of a molecular chronometer and is suitable for phylogenetic analysis (66, 67, 71).

Unlike the 16S rRNA gene, the *rpoB* gene does not contain areas that are highly conserved throughout the eubacterial domain of the bacterial kingdom (66). Consequently, and in contrast to the 16S rRNA gene, it has not been possible to design universal primers that will amplify the *rpoB* gene from every bacterial species with a single universal PCR.

Randi M. Nygaard at the Department of Microbiology, Haukeland University Hospital, defended her master's thesis "Use of Massive Parallel Sequencing for Detection and Identification of Microbes in Bile in Patients with Acute Colecystitis and Acute Cholangitis" in 2017 (72). As a part of this work she developed a method for TNGS of a segment of the *rpoB* gene for selected groups of bacteria. Two different primer-pairs were designed, one targeting the *Enterobacteriaceae* (*rpoB*_Ent) and one targeting *Staphylococcus*, *Streptococcus* and *Enterococcus* (*rpoB*_ESS). She analyzed 20 clinical samples by TNGS of both 16S rRNA and *rpoB* gene amplicons in the same sequencing run, and found that *rpoB* gene sequencing improved the taxonomic classification for 14 bacterial identifications.

1.5 Pleural empyema - microbiology

Pleural empyema is associated with high morbitidy and a one-year mortality of 20%. (73). Microbial infiltration of the pleural cavity is an obligate part of the pathogenesis (73). Treatment includes antibiotics and adequate drainage of infected pleural fluid, and in twenty percent of patients also more extensive surgical interventions (73, 74). The view that pleural empyema is caused by bacteria that translocate from a bacterial pneumonic infiltrate (73), has recently been challenged. Many of the bacteria found in pleural empyema are not known to cause bacterial pneumonia, and many patients with pleural empyema have no signs of underlying pneumonia (73, 75, 76).

Pleural empyema comes in two "variants" – community-aquired (CA) and hospital-aquired (HA) which differ from each other bacteriologically. The MIST1 study, the so far largest randomised multicenter trial on pleural empyema including 454 patients, found that the *Streptococcus milleri* group (24%), *Pneumococci* (21%) and anaerobic bacteria (20%) were most common findings in CA pleural empyema, whereas *Staphylococcus aureus* (35%) *Enteribacteriaceae* (18%) and *Enterococcus* spp. (12%) were the most common findings in HA pleural empyema (77). Most other culture-based studies report a similar pattern (73, 78).

The literature on pleural empyema is largely reporting on culture-based studies. As already outlined, culture has clear limitations for the study of fastidious and anaerobic bacteria and polymicrobial infections. For example, in the already mentioned MIST1 study, 44% of the patients had a negative pleural fluid culture (77). To the best of our knowledge, only a single study had used TNGS on a large cohort of patients with pleural empyema prior to our study from 2018 (79). In their sequencing of 98 pleural fluids from the MIST2 study, Wrightson *et al.* found that 33% of the empyemas contained anaerobic bacteria (80) - a much higher prevalence than what was previously known from culture-based studies.

The poor performance of culture-based diagnostics and the promising results from Whightson *et al.*'s study underscores the need for better microbiological diagnostics of pleural empyemas.

1.6 Bile infections

Acute inflammatory/infectious diseases in bile organs affect either the gall bladder (acute cholecystitis), the bile duct channel (acute cholangitis) or both. Cholelithiasis is the most common cause of both diseases, accounting for 90-95% of acute cholecysitis and more than 50% of acute cholangitis (81). International evidence-based criteria for diagnosis and severity assessment of acute cholecystitis and acute cholangitis, as well as guidelines for antimicrobial treatment (Tokyo Guidelines 2007), were launched after an International Consensus Meeting held in Tokyo (82–85). These criteria and guidelines have subsequently been updated twice, most recently in 2018 (TG18) (86–88). Acute cholangitis is considered the more serious of the two diseases, but both can present as a continiuum from mild to severe with sepsis and organ dysfunction. The Tokyo Guidelines severity assessment criteria grades both diseases into mild, moderate and severe.

For acute cholangitis, biliary obstruction and bacterial growth in bile are obligate for the pathogenesis (81). This is in contrast to acute cholecystitis, which is primarily an acute inflammatory condition not necessarily involving bacteria (81). A positive bile culture has been detected in one third to almost two thirds of patients with acute cholecystitis (89, 90), and high age has been identified as a predisposing risk factor for bacteriobilia (90, 91). Bacterial infection in acute cholecystitis is considered a negative prognostic factor associated with more severe disase and local complications (90), and antibiotic treatment is therefore, as for acute cholangitis, recommended for all grades of severity (87).

Current knowledge of infecting bacteria in both acute cholecystitis and acute cholangitis, and consequently also the empiric antibiotic treatment guidelines, are based on bile culture studies only (87). In both conditions, the most frequently cultured bacteria are species within the family of *Enterobacteriaceae* and the *Enterococcus* genus (87). *Escherichia coli* is the most common isolate in larger bacteriological studies of both acute cholecystitis and acute cholangitis, followed by *Klebsiella* spp*., Enterobacter* spp., *Citrobacteri* spp. and *Pseudomonas aeruginosa* among the gram negative bacteria, and *Enterococcus* spp. and *Streptococcus* spp.

among the gram positive bacteria (85, 89–93). Anaerobic bacteria, predominantly *Clostridium* spp. and *Bacteroides* spp., have been reported in anywhere from 0 to 20% of cases (85, 89–93).

2 Aim of the thesis

The main objective of this thesis was to investigate the use of 16S rRNA TNGS in microbiological diagnostics of polymicrobial invasive infections. We also wanted to evaluate the usefulness of supplementary *rpoB* gene TNGS for species-level resolution within certain clinically important genera. During the work, we came to recognize the need for simple and transparent approaches for management of DNA-contamination in post-sequencing processing and interpretation that can be easily adopted by other diagnostic laboratories. We thus focused on the following aims:

1. To evaluate the utility of 16S rRNA TNGS in clinical microbiology.
2. To evaluate the usefulness of supplementary *rpoB* gene TNGS.
3. To investigate the use of TNGS to discover microbial patterns that may provide new insights into establishment, development and maintenance of polymicrobial infections.
4. To explore and manage DNA contamination in TNGS.

All three papers included in this thesis target the first and second aim. The two remaining aims are addressed in papers I and III, respectively.

3 Material and methods

This thesis is based on three scientific papers with differing study designs; paper I is a retrospective, descriptive and comparative study, paper II is a prospective, descriptive and comparative study and paper III is a combined methodological paper and a prospective, descriptive and comparative study.

All three studies include a comparison of microbiological diagnostics by culture and TNGS of systematically collected samples from invasive infections. Paper I is based on a retrospectively collected material of pleural empyemas. In paper II we study prospectively collected samples from the bile bladder. In paper III we investigate a prospectively collected material from the bile duct. In Paper I, we also compare the obtained results with results from a previous TNGS-based study of human brain abscesses (1) to explore and evolve a new hypothesis on the aetiology of pleural empyema. In Paper III we specifically address the challenge of microbial DNA contamination in TNGS and use TNGS on multiple negative and positive extraction controls and a commercial staggered bacterial mock community to both explore the pattern of DNA contamination and to evaluate a suggested approach for management of DNA contamination.

3.1 Patient inclusion and sample collection

In paper I the patients were included retrospectively. The laboratory information system was used to identify all culture-positive and/or 16S rRNA PCR-positive pleural fluid samples from patients ≥ 18 years of age during a two-year period, from January 2016 to December 2017. Pleural fluids from 11 patients with a low suspicion of infection, with negative bacterial cultures and a negative 16S rRNA gene PCR were included as a negative patient control group.

In paper II we prospectively included patients who were treated for acute cholecystitis with percutaneous or perioperative drainage of the gall bladder from July 2015 to April 2017 at Haukeland University Hospital. Acute cholecystitis was defined according to the Tokyo Guidelines 2013 (94) (TG13) for a definite diagnosis.

Perioperatively sampled bile samples from 16 patients with cholelithiasis and no signs of ongoing gallbladder inflammation undergoing cholecystectomy at Voss Hospital, Norway, were included as controls.

In paper III we prospectively collected bile samples from all patients undergoing endoscopic retrograde cholangiopancreatography (ERCP) at Haukeland University Hospital from July 2015 to April 2017. Patients diagnosed with either acute calculous cholangitis, defined according to the Tokyo Guideline 2013 (95) (TG13) criteria for a definite diagnosis, or non-infectious bile duct stone were included for further analysis.

3.2 Mock community

In paper III we used a staggered mock community from ZymoBIOMIC (ZymoBIOMICS Gut Microbiome Standard, catalog no. D6331; Zymo Research Corp., Irvine, CA, USA) consisting of 19 bacterial strains representing 15 bacterial species and two fungal species. The mock community was diluted with microbial DNA-free water (Qiagen) in seven rounds of a serial 10-fold dilution prior to DNA extraction. The dilutions were analyzed with a SYBR green real-time 16S rRNA PCR using the protocol described in section 3.3.3 to obtain a semi-quantitative measure of the bacterial load of each dilution. One dilution with high bacterial load (1:10) and two dilutions with low bacterial loads ($1:10^5$ and $1:10^6$) were selected for further analysis. Negative and positive extraction controls were included and followed all processing steps.

3.3 Sample processing

All clinical samples were analysed by TNGS and conventional culture. To obtain a semi-quantitative assessment of the amount of bacterial DNA present in each sample, we used a universal real-time 16S rRNA PCR. Culture, extraction of DNA and, for paper I, Sanger-based 16S rRNA gene PCR was done upon sample arrival to the Department of Microbiology, and in accordance with the laboratory's guidelines.

Deep-sequencing and, for paper II and III, Sanger-based 16S rRNA PCR, were done in batches at a later stage.

3.3.1 Culture procedures

Culture procedures for the pleural fluid samples (paper I) and the bile samples (paper II and III) are described in Table 1.

Table 1: Culture procedures for pleural fluid[a] and bile samples.

Material	Culture medium	Volume	Incubation atmosphere	Incubation temperature	Incubation time
Pleural fluid and bile	Blood agar	10 µl	CO2-enriched	35 °C	48 hours
	Fastidous anaerobic agar	10 µl	Anaerobe	35 °C	48 hours
	Fastidous anaerobic agar with kanamycin and vancomycin	10 µl	Anaerobe	35 °C	48 hours
	Brain heart infusion	Not specified	Normal	35 °C	48 hours
Bile	Lactose agar	10 µl	CO2-enriched	35 °C	24 hours
Pleural fluid	Chocolate blood agar	10 µl	CO2-enriched	35 °C	48 hours

[a]Some pleural fluids samples were also inoculated and cultured in blood culture bottles although this was not part of the routine diagnostics

3.3.2 DNA extraction

DNA extraction was performed as described previously (24, 72). Dependent upon viscosity, 200-800 µl of sample material was added to a bead-containing tube (SeptiFast Lysis kit, Roche) together with Bacterial Lysis Buffer (Roche) to reach a total volume of 600-1200 µl. The sample was then processed in a homogenizer (MagNA Lyser, Roche) for 2 x 45 seconds at speed 6500 rpm and centrifuged (16,000 x G, 5 minutes). Four hundred µl of the supernatant was transferred to a MagNa Pure Compact instrument (Roche, Mannheim, Germany) for automated nucleic acid extraction and purification using the "MagNA Pure Compact Nucleic Acid Isolation Kit I" (Roche) with the protocol "Bacteria_DNA_V3_2".

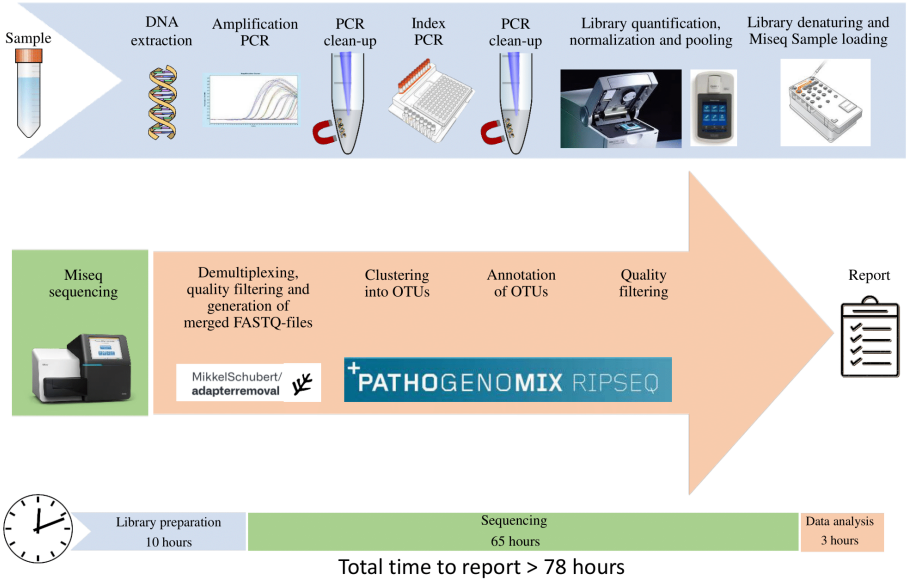
3.3.3 Sanger-based 16S rRNA gene PCR and sequencing

Sanger-based sequencing of the 16S rRNA gene was based on a previously described protocol (96) with a few modifications. The 5-end of the dual priming oligonucleotide (DPO) primers were optimized to eliminate a small tendency for primer-dimer formation (16S_DPO_Short-F: 5'-AGAGTTTGATCMTGGCTCAIIIIIAACGCT-3' (no LNA-bases) and 16S_DPO_Short-R 5'-CGGCTGCTGGCAIIIAITTRGC-3') and the annealing temperature in the PCR thermal profile was adjusted accordingly from 64 to 60°C. The PCR was run in a 25 µl reaction volume consisting of 23 µl mastermix and 2 µl of extracted DNA. The PCR-mastermix contained 1 µl of each primer, 12,5 µl of PCR-mastermix (SYBR Premix Ex Taq, TaKaRa, Shiga, Japan) and 8,5 µl of DNA-free water (Qiagen). The PCR thermal profile included an initial polymerase activation step of 10 s at 95°C, followed by 45 cycles of 10 s at 95°C (melt), 15 s at 60°C (annealing, DPO), and 20 s at 72°C (extension) (96). For interpretation of the Sanger electropherograms, the RipSeq mixed software (Pathogenomix, Santa Cruz, CA) was used for mixed DNA chromatograms (24), and the RipSeq single software (Pathogenomix, Santa Cruz, CA) for pure DNA chromatograms.

3.4 Targeted next generation sequencing

The same protocol for TNGS was used in all three papers. The workflow from sample arrival to report of results is showed in Figure 1.

Figure 1 Laboratory and post-sequencing workflow of targeted next generation sequencing[a]



[a] Illustrations in figure are collected from the Illumina 16S Metagenomic sequencing library preparation protocol (28) and the websites www.illumina.com, www.agilent.com and www.termofisher.com.

3.4.1 Choice of primers

Primers used for 16S rRNA and *rpoB* PCR are similar to those described by Nygaard (72) except for a few modifications made in the *rpoB*_ESS primers. All primers are listed in Table 2.

Table 2: Primers with adapter sequences. Target specific portions in capital letters and adapters in lower case letters.

Name	**Sequence[a]**	**Position[b]**
16S-F[c]	tcgtcggcagcgtcagatgtgtataagagacagCCTACGGGNGGCWGCAG	340-356
16S-R[c]	gtctcgtgggctcggagatgtgtataagagacagGACTACCAGGGTATCTAAKCC	784-803
*rpoB*_Ent-F	tcgtcggcagcgtcagatgtgtataagagacagGAAGGTCCRAAYATCGGTCT	1693-1712
*rpoB*_Ent-R	gtctcgtgggctcggagatgtgtataagagacagTGCATGTTCGCACCCAT	2041-2057
*rpoB*_ESS-F1	tcgtcggcagcgtcagatgtgtataagagacagGCRACAGCRTGTATYCCRTTC	1861-1881

*rpoB*_ESS-F2 paper I and II	tcgtcggcagcgtcagatgtgtataagagacagGCDACAGCATGTATTCCWTTC	1861-1881
*rpoB*_ESS-F2 paper III[d]	tcgtcggcagcgtcagatgtgtataagagacagGCDACMGCWTGTATYCCWTTC[d]	1861-1881
*rpoB*_ESS-R	gtctcgtgggctcggagatgtgtataagagacagGTTRTAMCCNTCCCAWGTCAT	2287-2307

[a] Nucleotide symbols and ambiguous base positions: A = Adenine, C = Cytosine, G = Guanine, T = Thymine, D = A/G/T, K = G/T, N = A/C/G/T, R = A/G, W = A/T, Y = C/T

[b] Positions for 16S based on *Escherichia coli* (GenBank accession J01859). Positions for *rpoB*_ESS based on *Staphylococcus aureus* (GenBank accession X64172). Positions for *rpoB*_Ent based on *Escherichia coli* (GenBank accession V00340).

[c] Abbreviations: F = forward primer. R = reverse primer.

[d] The *rpoB*_ESS F2 primer was modified during the work with paper III to correct three mismatches for *Enterococcus raffinosus*.

For the 16S rRNA PCR, the primers used were a modified version of those recommended in the Illumina protocol for the 16S library preparation (97), targeting the V3-V4 region. The modifications included replacement of the original “T” in position 3 from the 3-end of the reverse primer with a “K” (T/G) to avoid a mismatch with *Cutibacterium acnes* (24), and a change of the original H (A/C/T) and V (A/C/G) in position 7 and 8 from the 5-end to a C and A respectively. The latter changes do not reduce the primers’ ability to identify clinically relevant bacteria, but reduces the number of base pair combinations and thereby the risk of cross-reactivity with human DNA.

For the *rpoB* PCR, two different primer pairs were used (72). The *rpoB*_Ent primers targeting Enterobacteriaceae, and the *rpoB*_ESS primers targeting *Staphylococcus*, *Streptococcus* and *Enterococcus* species. For both the *rpoB*_Ent forward primer and the *rpoB*_ESS primers it was necessary to use degenerate primers to cover for all the intended microbes. For *rpoB*_ESS, two forward primers were used to limit the number of unnecessary base pair combinations.

3.4.2 Library preparation and sequencing

Targeted next generation sequencing for both 16S rRNA, *rpoB*_ESS and *rpoB*_Ent was performed using the MiSeq platform (Illumina, Redwood City, CA). Library

preparation and sequencing was done following a modified version of the Illumina protocol for the 16S library preparation (97) and the sequencing protocol described by Nygaard (72).

3.4.3 Amplicon PCR

PCR amplification of the sample templates (amplicon PCR) was processed in 96 well plates using the LightCycler 480 real-time PCR machine (Roche).
The DNA polymerase used was the TaKaRa-enzyme (SYBR Premix Ex Taq, TaKaRa, Shiga, Japan), and not the KAPA HiFi as described in the Illumina protocol. Reduced sensitivity of the *rpoB*_ESS PCR was one of the major challenges described by Nygaard (73), but this was improved using the TaKaRa-enzyme in the amplicon PCR. The use of a SYBR green real-time reaction also made it possible to perform a melting curve analysis to verify the presence of a PCR product with an expected melting point, eliminating the need for gel-based verification of the PCR product.
The content of the PCR mixture for the different target amplicons and the PCR temperature profile is described in Table 3. The PCRs had been optimized so that a single PCR thermal profile could be used for all four amplicons.

Table 3: PCR mixture and temperature profile for the amplicon PCR

Target gene	Primer name	Concentration and volume - primer	Volume - mastermix (µl)	Volume - H2O (µl)	Volume - template (µl)	Temperature profile
16S rRNA, V3-V4	16S-F	0,4 µM/ 1,0 µl	12,5	8,5	2,0	95 °C for 3 min (activation) 45 cycles of: - 95 °C for 20 s (melting) - 60 °C for 30 s (annealing) - 72 °C for 30 s (extension) Melting curve analysis: - 95 °C for 60 s - 40 °C for 2 min - 40-95 °C continous 40 °C for 30 s (cooling)
	16S-R	0,4 µM/ 1,0 µl				
*rpoB*_Ent *(*Targeting *Enterobacteriaceae*)	*rpoB*_Ent-F	0,4 µM/ 1,0 µl	12,5	8,5	2,0	
	*rpoB*_Ent-R	0,4 µM/ 1,0 µl				
*rpoB*_ESS (Targeting *Stapyholococcus, Streptococcus* and *Enterococcus)*	*rpoB*_ESS-F1	0,4 µM/ 1,0 µl	12,5	7,0	2,0	
	*rpoB*_ESS-F2	0,4 µM/ 1,0 µl				
	*rpoB*_ESS-R	0,6 µM/ 1,5 µl				

3.4.4 Index PCR and PCR clean-ups

Dual indexing of the amplicon PCR product and PCR clean-ups after both the amplicon PCR and the index PCR using *Agencourt AMPure XP* beads was performed as described in the Illumina protocol (97).

3.4.5 Library Quantification, Normalization, and Pooling

The DNA concentration in each sample (library) in nM was calculated using the formula:

$$\frac{(\text{DNA concentration in ng/µl})}{\left(660\frac{\text{g}}{\text{mol}}\times \text{average amplicon size}\right)} \times 10\text{^}6 = \text{concentration in nM}$$

We used a Qubit 3.0 Fluoremeter to measure the DNA concentration in each library (Fisher Scientific). The average amplicon size for the *rpoB*_Ent, *rpoB*_ESS and ITS2 libraries was found by analyzing each library on the Agilent 2100 Bioanalyzer, using a Bioanalyzer DNA 1000 chip that can measure the length of DNA strands between 25 and 1000 base pairs. For the 16S rRNA libraries, the average amplicon size was estimated to 630 base pairs based on the expected length of the 16S rRNA PCR product including primers and adapters. All libraries where then diluted to 4 nM using 10 mM Tris pH 8.5 and then pooled in a single tube using aliquot 5 µl of diluted DNA from each library, as described in the Illumina protocol (97).

3.4.6 Library Denaturing and MiSeq Sample Loading

Library denaturing and MiSeq sample loading was done as described in the Illumina protocol (97). A Phix control concentration of 5% was used in the final library. The loading concentration of the final library was 5 pM, which we experienced gave the best cluster density on the flow cell.

3.5 Post-sequencing processing

3.5.1 Demultiplexing and generation of merged FASTQ files

The MiSeq Reporter software was used for demultiplexing and generating FASTQ-files for each sample. We then used AdapterRemoval 2.2.2 (34) for trimming of

adapter sequences and low-quality bases and to merge the forward and reverse FASTQ-files of each sample by the following command:
AdapterRemoval—file1 <reads_1.fq> --file2 <reads_2.fq> --basename <mymergedfile> --threads 7 --trimns—trimqualities—minquality 20 --collapse—adapter-list <adapters>.txt—gzip

3.5.2 Clustering into Operational Taxonomic Units (OTUs)

The RipSeq NGS Software (Pathogenomix, Santa Cruz, CA) was used for downstream analysis. The merged FASTQ files were uploaded to the RipSeq NGS Preprocessor for further quality filtering and de novo clustering into Operational Taxonomic Units (OTUs). The following settings were used:

Primer check length: 12
Primer check max errors: 4
Trim ends with no 3'end primer: 10
Sequence length threshold: 250 (200 for *rpoB*_ESS and *rpoB*_Ent)
Copy number threshold: 10 (smallest acceptable OTU-size)
Max cluster variation: 1 % (i.e. 99% homology threshold for clustering)

3.5.3 OTU annotation

After clustering, a representativt sequence from each OTU were transferred to the RipSeq NGS online tool. BLAST searches against RipSeq curated databases were performed for all OTUs. The curated databases in RipSeq are regularly updated. For the last paper the following databases were used: The "*Pathogenomics Prime 16S*" for the 16S rRNA sequences and the "*Genbank Bacteria 1 – All bacterial targets, Valid Species and Pubmed*" and "*Pathogenomix rpoB_ESS*" / "*Pathogenomix rpoB_Ent*" for the *rpoB*_ESS and *rpoB*_Ent sequences. OTUs that did not yield a species or genus level using these databases were exported and individually analysed in GenBank using a standard BLAST search against the GenBank NCBI database in an attempt to obtain a better identification. The criteria used for taxonomy assignments are showed in Table 4.

Table 4: Criteria for unambiguous species assignments

Gene	Species	Species-group	Genus
16S[a]	≥99.3% homology with a high-quality reference, and minimum distance >0.7% to the next alternative species	≥99.3% homology with a high-quality reference, and minimum distance ≤0.7% to the next alternative species.	>97.0% homology with a high-quality reference
*rpoB*_Ent[b]	≥99.0% homology with a high-quality reference, and minimum distance >1.5% to the next alternative species	≥99.0% homology with a high-quality reference, and minimum distance ≤1.5% to the next alternative species	Not defined
*rpoB*_ESS[c]	≥97.0% homology with a high-quality reference, and minimum distance >2.0% to the next alternative species	≥97.0% homology with a high-quality reference, and minimum distance ≤2.0% to the next alternative species.	Not defined

[a] V3-V4 region of 16S rRNA-gene

[b] *rpoB*-gene sequence targeted at Enterobacteriaceae

[c] *rpoB*-gene sequence targeted at *Staphylococcus*, *Enterococcus* and *Streptococcus* species

3.5.4 Quality filtering in the RipSeq NGS online tool

OTUs consisting of human DNA sequences are reported as "no match" (0% similarity with reference sequences) when performing RipSeq NGS BLAST search, since the curated databases do not contain any human DNA references. To assure that such "no match" reports were not due to a lack of bacterial reference sequences in the curated databases, these OTUs were also analysed by performing BLAST search against the GenBank NCBI database. OTUs found to match with human DNA when performing the manually BLAST search against the GenBank NCBI database were removed.

Following the initial annotation of OTUs, a chimera-check was performed in all samples using the RipSeq NGS online tool.

If two or more OTUs mapped to the same reference sequence at a species or genus level they were manually merged. If a minor OTU only mapped to the genus level, but the BLAST search result in RipSeq showed that the best match was towards a

species that was also represented by a large OTU in the same sample, the two OTUs were also merged.
As a final quality filtering, OTUs represented by fewer than 50 reads were rejected. The use of such fixed lower cutoff for the number of representative sequences required to retain an OTU is recommended to reduce the risk of biased results related to sequencing noise and cross-contamination of samples (1, 6, 98–101).

3.6 Negative controls

For 16S rRNA sequencing, a negative extraction control containing 400 µL PCR-grade water and 400 µL lysis buffer was processed together with each sample. Before the sequencing procedure, all negative extraction controls were mixed into two or three pools, depending on the number of samples included in the sequencing run. A positive extraction control consisting of *Salmonella bongori* (paper I) or *Legionella pneumophila* (paper II and paper III) suspended in PCR-grade water was also included in each sequencing run.

3.7 Managing of background DNA contamination

The principles used for managing backgrund DNA contamination were similar in all three papers. Based on sequencing results from the pooled negative controls, a list of the most abundant contaminating bacteria was defined. These contaminants were used as indicators for the level of background DNA in clinical samples. Bacteria appearing in higher concentrations than any of the top background bacteria were accepted as valid identifications. Bacteria appearing in concentrations below the most abundant bacteria, but above a specified frequency threshold, was also accepted as valid (paper I and II) or likely valid (paper III) identifications. Based on the results in paper III from the in-depth characterization of background DNA contamination patterns and the evaluation of the above mentioned principles, a few changes were made compared to the criteria used in paper I and paper I: In paper I and II bacteria present in frequencies between 10% and 100% of the most abundant contaminant were accepted as valid identifications, if they were also absent from all the negative

controls. In paper III this range was altered to between 20% and 100%, and the status of these identifications were changed to “likely valid”. This was done to emphasize the gradually increased risk of including contaminants as true findings as the relative abundance of the bacteria compared to the top abundant contaminant decreases.

3.8 Literature

Search for literature used in this thesis ended 31st of December, 2021.

3.9 Ethics

All three studies were approved by the Regional Ethical Committee (REC) (2017/1095 – paper I; 2015/65 – paper II and III). For papers II and III, written informed consent was obtained from all participants. For the retrospective paper I, all participants received written information about the study and were given the opportunity to withdraw. The exemption from obtaining written informed consent in paper I was based on several conditions; 1) The project was assessed by REC to be of significant interest for the society, with a potential to alter the prevailing opinion regarding both pathogeneses, the microbes involved and antibiotic treatment of pleural empyema. 2) The REC considered that the personal welfare and integrity of the participants were safeguarded. This was a retrospective study on already collected biological material and study inclusion did not affect the patient or treatment in any way. 3) The REC recognized that it would be difficult to obtain consent from all participants, and that the strength of the study would be vulnerable to even a small number of non-responders because of the relatively limited number of participants expected to meet the inclusion criteria.

4 Paper summaries and results

4.1 Paper I

Introduction

Paper I is a retrospective, descriptive study. We used TNGS to describe the microbiological characteristics of 64 clinically well described pleural empyemas and to compare the results with those obtained by culture and 16S rRNA Sanger sequencing. Observed microbial parallels between pleural empyemas and brain abscesses were investigated aiming to further the understanding of pathophysiological mechanisms of pleural empyemas.

Methods

All available culture- and/or 16S rRNA gene PCR positive pleural fluids from a 2-year period were analyzed using TNGS of the 16S rRNA gene and, in selected cases, also the *rpoB* gene. Results from TNGS were compared with those obtained by culture and Sanger-based 16S rRNA gene sequencing. Clinical details were evaluated by medical records review. Comparative analysis with brain abscesses was performed using metagenomic data from a previous national Norwegian study (1). Pleural fluids from 11 patients with a low suspicion of infection, negative cultures and negative 16S rRNA gene PCR were included as a negative patient control group.

Results

64 patients (50 men, 14 women, mean age 62 years) were included. Among these, 43 (67%) had community acquired infections. Thirty-seven patients had a well-defined aetiology of their pleural empyemas, while 27 patients (24 men, 3 women), all of them with a community acquired infection, had an uncertain aetiology. Twenty-six out of these 27 empyemas contained either *Streptococcus intermedius*, *Fusobacterium nucleatum* or both. None of the 11 samples in the negative patient control group contained bacterial DNA beyond that found in the negative controls. Out of 385 bacterial detections made by TNGS, 38 (10%) were detected by culture

and 87 (22.5%) by Sanger-based 16S rRNA gene sequencing. Five detections were made exclusively by culture (one each *Streptococcus constellatus*, *Staphylococcus aureus*, *Klebsiella pneumoniae*, *Cutibacterium acnes* and *Staphylococcus epidermidis*, Table 5). The first three were made from complex polymicrobial infections inoculated and cultured in blood culture bottles. The last two were detected by TNGS, but since *C. acnes* and *S. epidermidis* were among the ten most abundant microbes in the negative controls, they were not considered valid detections according to the criteria applied. Sanger-based 16S rRNA gene sequencing did not detect any bacteria that were not also detected by TNGS.

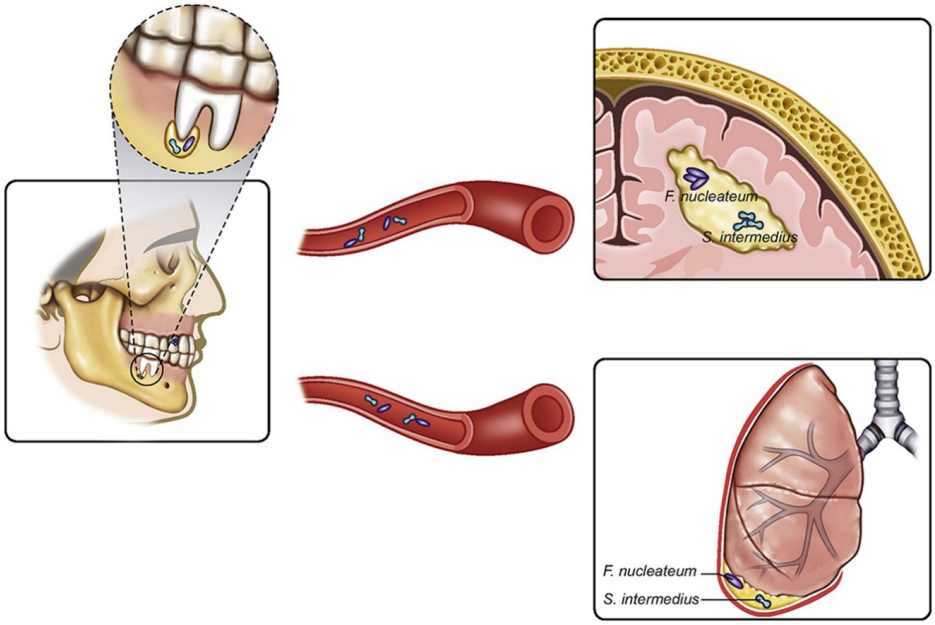
Supplementary sequencing of the *rpoB* gene allowed for species level identification of 25 more bacteria.

Unifrac analysis of the 64 included empyemas revealed that most of the 27 samples with uncertain aetiology clustered in neighboring branches, indicating significant similarities in microbial patterns. Venn diagram analysis comparing species found in the 27 empyemas of poorly described aetiology and 25 brain abscesses with assumed oral/sinus origin demonstrated a significant microbial overlap with 19 (65.5%) species present in both infection types. The most frequent common species were *Fusobacterium nucleatum*, *Streptococcus intermedius*, *Parvimonas micra*, and *Eubacterium brachy*, all present in more than 30% of samples from both infections.

Conclusion

Targeted next generation sequencing led to a 10-fold increase in bacterial detections as compared to standard culture. The increased sensitivity led to the discovery that a major subgroup of pleural empyemas is caused by a limited set of bacteria not normally involved in pneumonia, and that such empyemas have a similar microbial profile to oral/sinus-derived brain abscesses. The two distinct patient groups also share several clinical risk factors. We therefore suggest that these pleural empyemas, like brain abscesses, result from haematogenous seeding of bacteria from an oral focus (Figure 2) rather than by micro-aspiration trough the respiratory tract as currently postulated.

Figure 2[a] – Illustration of suggested shared aetiology of primary empyema and oral derived brain abscesses



[a] Reprinted from paper I (79)

4.2 Paper II

Introduction

Paper II is a prospective single-center study. It presents the results of TNGS performed on bile samples from 36 patients with moderate or severe acute cholecystitis. Sequencing results are compared with those obtained by culturing.

Methods

Bile samples were collected from patients undergoing percutaneous or perioperative drainage of the gall bladder. All samples were analyzed using both culture and TNGS of the bacterial 16S rRNA, the fungal ITS2-segment and in selected samples the bacterial *rpoB* gene. Clinical details were evaluated by medical records review. Bile

samples taken at cholecystectomy from 16 patients with cholelithiasis and no signs of ongoing gallbladder inflammation were included as a patient control group.

Results

Bile from 31 (86%) of the 36 patients contained bacteria (29) and/or fungi (5) as determined by TNGS. Only 40 (38%) of the 106 microbial detections made by TNGS were detected by culture as well. In none of the 15 polymicrobial samples did culture detect all present microbes. Two bacterial detections, a *Klebsiella pneumonia* and a *Staphylococcus epidermidis*, were made by culture only (Table 5). The *rpoB* gene sequencing allowed for species level identification of 21 more bacteria. Bacteria detected by TNGS that were frequently missed by culture included oral *Streptococci*, anaerobic bacteria, *Enterococci* and *Enterobacteriaceae* other than *Klebsiella* spp. and *Escherichia coli*. Culture had a particular low recovery rate for anaerobe bacteria. TNGS detected 24 anaerobic bacteria from 10 samples whereof only two (8%) were also identified by culture. In the patient control group, only three (19%) out of the 16 controls had detectable microbes in bile with *Streptococcus parasanguinis*, *Bifidobacterium animalis* and *Haemophilus parainfluenzae* identified from one patient each.

Conclusion

The article demonstrates that culture-based methods alone are insufficient for microbiological diagnostics of moderate and severe acute cholecystitis, leaving more than 60% of the microbes undetected. The clinical consequences of not detecting or treating all these bacteria should be further addressed in future studies as should consequences for empiric treatment recommendations.

4.3 Paper III

Introduction

Paper III is primarily a methodological article although we also describe the microbial composition of bile samples from patients with acute cholangitis. The paper

investigates the patterns of microbial contamination in 16S rRNA TNGS and their implications for post-sequencing filtering of results. The impact of sequencing depth and the inherent sensitivity limitations that remain in TNGS is demonstrated. An approach for managing DNA contamination in clinical diagnostics using 16S rRNA TNGS is suggested. The approach is evaluated by TNGS of a diluted staggered mock community and of prospectively collected bile samples from 41 patients with acute cholangitis or non-infectious bile duct stenosis.

Methods and Results

(I) The patterns of DNA contamination in 16S rRNA TNGS were studied by sequencing two negative and one positive extraction control. Each of the three controls was split into five replicates before the amplicon PCR (hereafter named “PCR replicates”). One PCR replicate from each of the three controls was further split into five replicates before sequencing (hereafter named “sequencing replicates”). This was done to isolate the impact of the PCR amplification of the sample template (amplicon PCR) from the impact of the following index PCR and sequencing procedure. All PCR and sequencing replicates were then indexed and sequenced in the same run.

Sequencing results showed that in all replicates a few species dominated, and that these dominating species were the only species consistently found across all replicates. The PCR replicates displayed a high diversity among the low abundant background contaminants, while the sequencing replicates had a very low diversity. We used the results from this part of the study to formulate criteria for filtration of sequencing data from clinical samples (Box 1), which we further evaluated on a staggered mock community and a collection of bile samples.

Box 1: Suggested criteria for filtering of DNA contaminants in clinical samples.

(i) Any bacterium appearing with a higher abundance than the top five abundant contaminants, as determined by the sequencing of negative and positive extraction controls, is accepted as a valid identification, even if it occurs as a low abundance species in the controls.

(ii) Bacteria present in frequencies between 20% and 100% of the most abundant contaminant are accepted as likely valid identifications, but only if they are also absent from all the negative controls.

(iii) Bacteria present in frequencies below 20% of the most abundant contaminant are always rejected.

(iv) In samples where none of the top five abundant contaminants are detected, all identifications are accepted as valid

(II) The suggested approach was tested out by sequencing a diluted staggered mock community where we aimed (i) to assess the actual abundance of the contaminants detected in our negative controls and to determine at what level the observed high variability in PCR replicates occur and (ii) to assess the sensitivity and specificity of our suggested criteria for filtering bacterial contaminants and to compare the strategy to other common methods for contaminant filtering.

By using the calculated concentration of mock community species as a reference, we observed that the most dominating contaminants appeared at concentrations of approximately 10 16S rRNA copies per 2 µl template, corresponding to about 500 bacterial cells per ml in the original sample. The less abundant contaminants appeared in concentrations close to or less than one single 16S rRNA copy per 2 µl PCR template, approaching the lower limit of detection in the PCR. This corresponds to an initial concentration of up to 100 bacterial cells per ml sample. We found that filtering using our suggested criteria had

a sensitivity and specificity for the identification of mock community bacteria in the tested samples of 83% and 97% respectively, giving an overall test accuracy of 93%. Among the other methods tested, decontam (59) with use of the function “isNotContaminant” showed the highest sensitivity (100%) and specificity (77%) for the identification of mock community bacteria, giving a test accuracy of 86%.

(III) Finally, the effectiveness of the approach for analysis of polymicrobial infectious samples was demonstrated by sequencing 41 bile samples from patients with acute cholangitis (AC, n=15) or non-infectious bile duct stenosis (NIBDS, n=26). The 41 samples were sequenced in two replicates with different sequencing depths to also demonstrate the importance of sequencing depth in TNGS. In selected samples *rpoB* genes were also sequenced.

The same patterns of contamination as for the analyses of extraction controls and the staggered mock community were observed. The results showed discrepancies in microbial findings between the two replicates for 22 (53.7%) of the 41 samples. Ninety-four bacterial identifications were made in only one of the two replicates. Out of these, 90 (96%) were found in the sample with the highest sequencing depth. The *rpoB* gene sequencing allowed for species level identification of 24 more bacteria. Compared to culture, sequencing found a much higher species richness in most samples. In the acute cholangitis group, 84 microbial identifications were made by sequencing whereof only 26 (30%) were cultured. With regards to the bacterial findings, some differences between samples from AC and NIBDS were observed. All samples from patients with AC were both culture and sequencing positive, while only 21 out of 26 NIBDS were sequencing positive, and only 17 out of the 26 were culture positive. In general the AC samples had a lower CT-value (mean CT-value 19,8, range 12.5–27.9) than the NIBDS samples (mean CT value 25,6, range 12,2-33,4), indicating a higher microbial mass in the AC samples. Many of the identified bacterial species were

found in both AC and NIBDS samples, but we observed that known bile pathogens, such as *Escherichia coli*, *Klebsiella spp* and *Enterococcus* spp, constituted a larger part of microbial detections in AC samples compared to the NIBDS samples.

Conclusion

Based on the findings in this study we hypothesized that the major contributor to the variation found between TNGS replicates with a low microbial mass, both in our papers and others (79, 102, 103), is the random inclusion of low-abundant contaminant microbial DNA during pipetting of the PCR-template. The low-abundant contaminants are under the law of small numbers, meaning that a random sample is not likely to reflect the actual population.

We further suggested and tested criteria for filtering contaminants, according to which the most abundant background contaminant species defined the background level of contamination. We showed that below this level, due to the law of small numbers, discrimination between background contamintants and true findings rapidly becomes highly uncertain, firmly defining a sensitivity limit for current deep sequencing approaches. We further demonstrated that the most abundant contaminant DNA can serve as a marker of sequencing depth. Adequate sequencing depth can only be claimed when the analysis also picks up some background contamination.

4.4 Sensitivity of targeted next generation sequencing versus culture

A comparison of culture and TNGS results from all three infectious materials combined (pleural empyemas, acute cholecystitis and acute cholangitis samples) is provided in Figure 3. Culture detected only 101 (18%) out of the 562 bacterial detectections made by TNGS (Figure 3A). In contrast, only 13 bacterial detections were made by culture but not by TNGS (Table 5).

As expected, the sensitivity of culture depended upon both the bacteria's aerotolerance and on the family to which the detected bacteria belonged. The sensitivity of culture was lowest for anaerobic bacteria where only 5% (13/247) of

bacterial detectctions made by TNGS were reproduced by culture (Figure 3D). The overall sensitivity of culturing for aerobic gram-negative and gram-positive bacteria were 32% (54/169) and 27% (70/260) respectively. Major differences in the sensitivity of culture for bacterial families within each of these groups were observed, spanning from close to or above 50% for bacteria belonging to well-known families of pathogens like *Enterobacteriaceae*, *Pseudomonaceae* and *Enterococcaceae* to zero or close to zero for other families like *Pasteurellaceae*, *Campylobacteraceae* and *Actinomycateae* (Figure 3B and 2C). The sensitivity of culture for monomicrobial compared to polymicrobial samples, as determined by TNGS, is shown in Figure 4. In only one out of 60 polymicrobial samples did culture detect all bacteria detected by TNGS.

Figure 3

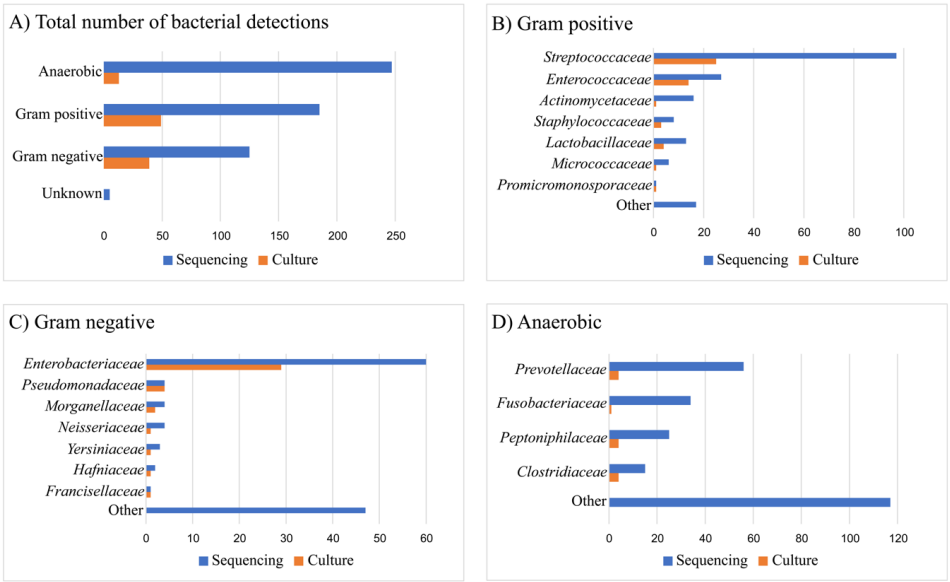


Figure 3: Total number of bacterial detections by TNGS, and the number of these detections that were also made by culture, in all samples from pleural empyemas, acute cholecystitis and acute cholangitis combined. The x-axis shows the absolute number of bacterial detections in all samples combined **A)** All bacterial detections are sorted into anaerobic, gram positive and gram negative bacteria. Five bacterial sequences did not match to any known species and are named as unknown. **B)** All detected gram positive bacteria sorted at family level. The group "other" includes 5 bacterial families that were detected by sequencing only: *Aerococcaceae* (1), *Bacillaceae* (1), *Bacillales incertae sedis* (9), *Carnobacteriaceae* (5) and *Corynebacteriaceae* (1). **C)** All detected gram negative bacteria sorted at family level. The group "other" includes 4 bacterial families that were detected by

sequencing only: *Campylobacteraceae* (21), *Moraxellaceae* (1), *Mycoplasmataceae* (7) and *Pasteurellaceae* (18). **D)** All detected anaerobic bacteria sorted at family level. The group "other" includes 24 bacterial families that were detected by sequencing only: *Atopobiaceae* (4), *Bacteriodaceae* (9), *Bacteroidetes* (F-1) (1), *Barnesiellaceae* (1), *Bifidobacteriaceae* (8), *Clostridiales* (F-1) (1), *Coriobacteriaceae* (1), *Desulfovibrionaceae* (1), *Eggerthellaceae* (3), *Erysipelotrichidae* (2), *Eubacteriaceae* (12), *Eubacteriales incertae sedis* (1), *Lachnospiraceae* (9), *Leptotrichiaceae* (2), *Peptostreptococcaceae* (12), *Porphyromonadaceae* (4), *Propionibacteriaceae* (2), *Rikenellaceae* (1), *Ruminococcaceae* (3), *Saccharimonadaceae* (2), *Selenomonadaceae* (3), *Tannerellaceae* (5), *Treponemataceae* (3) and *Veillonellaceae* (27)

Figure 4

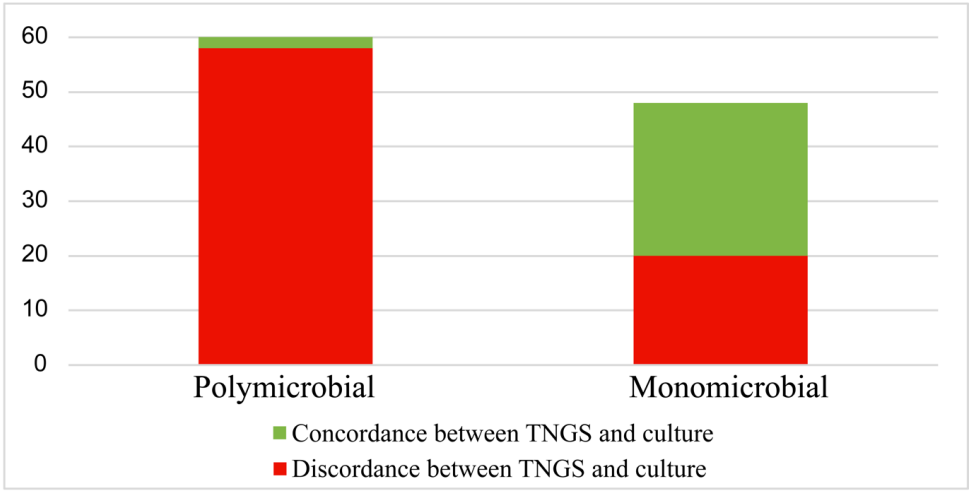


Figure 4: Comparison of diagnostic efficacy of TNGS and culturing in polymicrobial (n=60) and monomicrobial (n=48) samples when TNGS was used as gold standard. In the polymicrobial samples culture succeeded in detecting all bacteria that were detected by TNGS in only two (2 %) out of 60 samples. The corresponding number for the monomicrobial samples was 28 (58 %) out of 48 samples.

5 Discussion

5.1 Methodological considerations

5.1.1 Study design

Papers I to III report results from single-center studies performed at Haukeland University Hospital/Helse Bergen and may therefore not be generalisable to other populations. However, since many of the pathogens detected in our studies are reported as typical pathogens for the specific types of infections in other parts of the world, the generalisation problem probably applies more to the sequencing results than to the culturing results. The lack of deep-sequencing studies of the infections, and the fact that microbiomes of humans vary geographically and across populations (104), implies that our findings need to be confirmed by similar studies in other populations/countries to assure that the results are generalisable.

5.1.1.1 Paper I

The patients in paper I were included retrospectively. In general, retrospective studies allow researchers to establish associations and formulate hypotheses but are not suitable for confirmation of causal relations. In order to establish cause-effect relations between exposure and outcome, like this paper's proposed causal hypothesis of hematogenous spread of oral pathogens from a dental focus to pleural infections, larger, prospective studies are needed.

The study was biased by the availability of material. As stated in the paper, 18 culture-positive pleural fluids were unavailable for inclusion because DNA had not been extracted and stored.

The study may also have been vulnerable for information bias. The collection of information from the medical record was done by the first author, who was aware of the culture and sequencing results. This implies that he might have tended to confirm his own hypotheses when there was information in the journal open to interpretation. Since the diagnosis of pneumonia sometimes can be difficult to establish, the evaluation as to whether pneumonia was the cause of the pleural empyema or

whether the empyema was of uncertain aetiology was considered the assessment most vulnerable to such information bias. To reduce the risk of this bias we tried to standardize the interpretation of collected data by defining specific and objective criteria for bacterial pneumonia.

Another bias is missing information on study variables in the medical records. The retrospective design implies that there is no systematic collection of relevant data at inclusion of patients. When analysing associations between risk factors and type of pleural empyema the lack of information may have biased the results in both directions. An example is the lack of information about dental status, which was available for only six patients.

5.1.1.2 Paper II

Although the prospective design is a strength of paper II as compared to paper I, there is still a room for selection bias. Since inclusion took place at the time of surgical intervention, we do not know whether there were patients with moderate or severe acute cholecystitis that were treated conservatively and never underwent surgical interventions. As commented in the paper, we observed a higher mean age in our group of patients then what is reported in historic studies on moderate and severe cholecystitis. If younger patients with moderate disease were less likely to be referred to surgical interventions than older patients with moderate disease, this would be a systematic error explaining the observed age difference and making our results less generalisable also for moderate and severe acute cholecystitis.

5.1.1.3 Paper III

Patients in paper III were also included prospectively at the time of intervention. We included samples from all patients undergoing ERCP at the Department of Surgery, and then excluded all but those patients that either filled the criteria for acute calculous cholangitis or noninfectious bile duct stone. Since almost all patients with acute calculous cholangitis undergo ERCP as part of the treatment, we believe the risk of selection bias in this study to be low.

5.1.2 Quality of samples

The collection of representative sample material is crucial when performing next generation sequencing studies, as any contamination during sampling can disturb the results. In paper I and II, samples were collected either percutaneously (paper I and II) or perioperative during cholecystectomy (paper II). In both cases, the risk of contamination from the surroundings were low, and we believe the results to be representative. In paper III, samples were more prone to contamination. Bile-fluid was collected by aspiration from the bile duct through a sterile sphincterotome catheter as soon as the position of the catheter in the bile duct was verified. Although the sphincterotome is hidden in the endoscope until the major duodenal papilla is located, it is not possible to completely avoid exposure to the duodenal microbial flora when the tip of the sphincterotome is extended from the endoscope to enter the bile duct through the papilla. However, it is likely that the relative concentration of contamination will remain low due to the relatively low microbial density in the upper gastrointestinal tract combined with a large sample volume of 2-5 ml bile. Many of the bacterial findings both in the acute cholangitis group and the non-infectious bile duct stenosis group were low abundant bacteria known to colonize the upper gastrointestinal tract (Supplementary Figure S1, paper III). It is possible that these low abundant bacteria represented contamination during sampling. We also found a high rate of bacterial precense in NIBDS samples, but, as explained in chapter 4.3, the microbial mass in many of these samples were relatively low as determined by their 16S rRNA PCR CT-value. The reported rate of bacterial precense in NIBDS samples should therefore be interpreted with caution, as it be may be explained by contamination during sampling.

5.1.3 Choice of clustering method

We chose to use *de novo* OTU clustering when analysing our sequencing results, since this approach makes it easier to manually assess the quality of the data-analysis. The use of Amplicon Sequencing Variants (ASV) has, as discussed in the introduction, emerged as an alternative to *de novo* OTU-clustering for analyses of TNGS data. Some argue that the method should replace *de novo* OTU clustering (41), while others finds that both methods have their pros and cons (29, 41, 105, 106). The

following section will provide a brief discussion around our choice of an OTU-based method.

The major arguments put forward in support of the ASV method are that (i) it gives a more accurate description of alpha diversity (the species diversity in a sample), since erroneous sequences that could have created spurious OTUs are removed, (ii) results are reproducible and can be compared across different studies since every ASV represents a biological reality, and finally (iii) the ASV method provides a more precise identification at a species level since even single-nucleotide differences between sequencing reads are identified (29, 41).

There is no doubt that erroneous sequences can disturb the results and that the ASV-method offers a novel approach to reduce the impact of these errors. However, the aggressive filtering of all error sequences by the ASV-method have several caveats. The validity of the ASV-method is strongly affected by the quality of the data. Removal of all reads containing PCR-errors may lead to depletion of a large amount of useful information. In contrast, in the *de novo* OTU approach error sequences within the chosen similarity threshold will tend to merge with more prevalent correct sequences, and consequently data are not lost by later error filtering. In our studies we controlled for more aberrant error sequences by setting a lower threshold for the number of sequencing copies in a cluster and by merging clusters assigned to the same reference taxon. We also considered genus-level identifications from small OTUs that matched best with a species represented by an OTU of higher abundance to be outliers from that larger species-level OTU. We believe that the use of these methods combined minimized the impact of sequencing errors on the diversity-measures of our results.

We do agree that it is an advantage of the ASV method that one can directly compare the ASVs across studies. However, in clinical microbiology it is more important to compare the species or genera that these sequences represent, than to compare the specific sequence of an OTU or ASV from different samples or studies. In a clinical setting, we want to know whether there was a *Fusobacterium nucleatum* present in the sample or not rather than the exact sequence variant of a part of the 16S rRNA gene of that *Fusobacterium nucleatum.* We will also emphazise that in studies on

clinical materials, in contrast to environmental microbiome studies, it is uncommon to find OTUs representing species not having been given a valid or provisional name. With proper filtering, most OTUs will therefore be assigned to a known species. Finally, the claim that a more precise species identification is possible by the ASV method because it identifies singe-nucleotide differences (59) is not in concordance with the recommended criteria for unambiguous species-level assignment that requires a minimum distance of >0.8% to the next species (37). In other words, a single-nucleotide difference in the 16S rRNA gene is not sufficient to reliably distinguish between two closely related species. This constitutes one of the major problems of 16S rRNA TNGS and represents an inherent limitation that remains independent of the chosen clustering method. Provided that a biologically meaningful level of homology is used for clustering (e.g. 99% or higher), we thus believe that in a clinical microbiology setting, choosing the OTU method does not impair the results as compared to what can be achieved by the ASV method.

5.1.4 The comparison of different contamination filtering methods – paper III.

In paper III we used different dilutions of a staggered mock community to compare our suggested criteria for filtering contaminants with other commonly used methods, including the decontam R-package. For decontam, we found a sensitivity of 100% for the identification of a true bacterium as valid. A similar high sensitivity for decontam has been demonstrated both in the decontam introduction paper and in other papers examinating the capabilities of the package, when used on dilution-series of a sample with a known bacterial composition (14, 15, 59). However, we will argue that the use of such serially diluted samples to demonstrate the efficacy of filtering methods like decontam produces results that are not generalizable to a clinical setting. There is a risk of over-estimating both the obtainable sensitivity and specificity of the method as compared to what can be expected when used on clinical samples.

The prevalence-based method of decontam is grounded on the expectation that the prevalence of contaminants will be higher in extraction controls than in true samples due to the absence of competing DNA in the sequencing process, and further that non-contaminants will be more prevalent in true samples than in extraction controls

(59). The latter expectation will tend to get self-fulfilled when analyzing a single mock community in multiple dilutions, since all samples will contain the same bacteria. However, this will normally not be the case in diagnostic microbiology where multiple samples from different patients and types of infections are analyzed in the same run. Regarding the first expectation, this will not hold true for the most dominant contaminants as these will also appear in all samples provided an adequate sequencing depth.

In our paper we did not evaluate the frequency-based decontam method since we had not included a quantitative PCR to quantify the DNA concentration of our samples and because this method is not recommended for low biomass samples by the authors of the introduction paper (59). However, we will argue that the limitations of using a single serially diluted sample to evaluate the performance of the decontam package also applies when using the frequency-based method, most obviously by giving a higher sensitivity than what will be achievable in a clinical material. The frequency-based method is based on the expectation that the frequency of contaminant DNA varies inversely with the total sample bacterial DNA concentration, while the frequency of non-contaminant DNA does not (59). Just like for the prevalence-based method, this expectation will tend to be self-fulfilling when analyzing a single mock community in multiple dilutions, thus explaining why the authors of the introduction paper of decontam found that no true *Salmonella bongori* identifications (sensitivity 100%) were classified as contaminants when testing a dilution-series of that bacterium (59). Testing of a dilution-series does not take into account the real world situation where low abundant true OTUs may only randomly appear because they are under the law of small numbers or may be present in only some samples dependent upon sequencing depth.

In contrasts to decontam, our suggested approach for contaminant filtering is not dependent upon the sequencing of multiple samples and/or extraction controls to calculate what is likely to represent contaminant or true sequences. We believe this to be one of the strengths of our method. Although the principle of decontam might be reasonable in microbiota research with large batches of the same sample material from a naturally occurring and presumably relatively stable microbiota, this is not

something that can be expected in clinical microbiology. In the routine diagnostic, samples from different infection types with unpredictable and variable bacterial compositions will be run together and many pathogenic bacteria will be present in only a single sample, maybe even at a low concentration.

5.2 *rpoB* gene sequencing as a supplement to the 16S rRNA gene sequencing to improve species differentiation

The rationale for including TNGS of selected parts of the *rpoB* gene in our protocol was the acknowledgement that differentiation between closely related species by the 16S rRNA gene alone is not always achievable. Targeted next generation sequencing of the *rpoB* gene proved useful for increasing the taxonomic resolution in all three study materials. The number of species that could be identified at a higher taxonomic level by use of partial *rpoB* gene as compared to partial 16S rRNA gene sequencing (V3–V4) were 25 for the pleural empyema samples (Paper I, Supplementary Table S3), 21 for the bile bladder samples (paper II, Table 3) and 24 for the bile duct samples (Paper III, Supplementary Table S7). It is important to remember that we only targeted a few selected genera with this approach.

We chose to predefine the 16S rRNA TNGS as the gold standard for detection of bacteria. This implied that if we detected a bacterium by the *rpoB* gene TNGS alone, it was not considered a valid identification. The reason for using the 16S rRNA TNGS as a gold standard was mainly the lack of experience of using *rpoB* gene sequencing to characterize the bacterial metagenome of a sample. We did observe that this conservative approach in some cases probably reduced the sensitivity of TNGS. The sensitivity of TNGS in polymicrobial samples is dependent on the sequencing depth, as will be discussed in more detail in chapter 5.4.1.2. In samples containing a mixture of bacteria where only some of them are covered by the *rpoB* gene PCRs, the universal 16S rRNA gene will amplify more PCR targets than the more specific *rpoB* PCRs. In such samples, a lower sequencing depth will be needed to cover all targets of the *rpoB* gene PCR by TNGS than by 16S rRNA TNGS. An example of this is found for sample number 1 and 9 in Table 5. In these samples

culture identified a *K. pneumoniae* and a *S. aureus* that were also identified by the semi-selective *rpoB*_Ent and *rpoB*_ESS TNGS respectively. However, due to the presence of a range of other bacteria at much higher concentrations they remained undetected by the universal 16S TNGS.

Another interesting observation resulting from the inclusion of the *rpoB* gene was the demonstration of the inability of 16S rRNA TNGS to distinguish between many closely related species because of the low inter-species variation in the 16S rRNA gene. For example, in sample 38 in paper III, the 16S rRNA TNGS identified an OTU cluster of 4723 reads mapping towards the *Streptococcus mitis/oralis* group. However, with *rpoB*_ESS TNGS, we identified five distinct OTUs that all mapped to different species within the *S. mitis/oralis* group, including one OTU that mapped as *Streptoccous infantis* (cluster 1), one as *Streptococcus infantis* (cluster 2), one as *Streptococcus* sp. (300_SSPC 371), one as *Streptococcus oralis* and one as *Streptococcus mitis*.

5.3 Comparison of targeted next generation sequencing to traditional microbial diagnostics and its utility in clinical microbiology

Several studies and reviews have been published that discuss and argue for the implementation of NGS, including TNGS, as a diagnostic tool in routine diagnostics (2, 4, 10, 107, 108). The main argument for the implementation of TNGS is the increased sensitivity of the method. The findings in our three studies support this argument. We found that TNGS was highly superior to culture (paper I-III) for the detection of microbes in both pleural empyemas, acute cholecystitis and acute cholangitis (Figure 3). In paper I it was also highly superior to Sanger-based 16S rRNA gene PCR/sequencing. The higher sensitivity for TNGS as compared to culture was clearly most pronounced for the polymicrobial infections (Figure 4). This corresponds well with other studies (1, 2, 6, 10, 108).
The role of TNGS in two different fields of clinical microbiology will be discussed in the following sections: (1) Use of TNGS for routine microbial diagnostics of

polymicrobial invasive infections and (2) use of TNGS in clinical microbiology research on polymicrobial invasive infections.

5.3.1 The utility of TNGS in routine diagnostics

A main aim of the thesis has been to evaluate the usefulness of TNGS compared to culture. For this endeavour, it is not enough to demonstrate an increased accuracy compared to current diagnostic tools (109). In addition, the test's impact on i) workflow, ii) clinical costs, iii) clinical decision making and iiii) patient outcomes should be evaluated (109). The following sections will provide a brief discussion of TNGS's impact on these four categories.

5.3.1.1 Workflow

The workflow of TNGS is more complicated and, in most cases, more time-consuming when compared to other diagnostic tools like routine culture, universal 16SrRNA Sanger sequencing and the emerging rapid PCR-based syndromic panels (110). Targeted next generation sequencing requires specialized technicians and microbiologists both to run the analysis and interpret the results, and even though we have managed to simplify and shorten parts of the library preparation protocol, the laboratory turn-around-time from sample arrival to final results is >78 hours (Figure 1).

5.3.1.2 Costs

Despite a remarkable reduction in cost per sample since the early development of TNGS, the cost of a MiSeq run remain high compared to other available diagnostics. Illumina advertises with a cost of $18 USD per sample (www.illumina.com) with use of the maximum capacity of the MiSeq of 96 samples per sequencing run. However, this is not feasible in most laboratories owing to scarcity of samples. In addition, 96 samples per run will result in an insufficient sequencing depth for reliable characterization of the sample metagenome. At the Department of Microbiology at HUS we include 24-30 samples per run with an estimated cost per sample of 800-1000 NOK (about $80-100 USD)

5.3.1.3 Clinical decision making and patient outcome

Clinical decision making includes factors such as selection of antibiotics and changes in hospital procedures, while patient outcome focuses on patient mortality and morbidity (109). Were TNGS to have a positive impact on such parameters, it should provide additional information relevant for steering the patient's disease progress and treatment. Expressed in other words, TNGS may have a positive impact in situations where it can be used to identify clinically relevant pathogens that otherwise would remain undetected and where this information can be used to guide treatment in a way that favourably improves patient outcome.

Papers I-III clearly demonstrates the improved sensitivity of TNGS for polymicrobial infections compared to culture, thus implying that culture alone should be considered insufficient for guiding of antibiotic treatment of such infections. Although some may argue that what grows by culture reflects the most important pathogenic bacteria in the polymicrobial infection, this argument rests on insufficient knowledge obtained from only one technique – culturing. As argued in paper II, the clinical relevance of individual bacteria in complex infections should not be considered based on relative abundance or method of detection. Rather, such inference should be based on in-depth ecological knowledge of each type of infection, including microbial dynamics over time, microbial aggregate formation, metabolic interdependencies and synergisms.

With regards to the relevance of TNGS for clinical decision making and patient outcome, the most important task will therefore be to define the clinical relevance of microbes identified by TNGS. The complete microbial characterization of infectious diseases as provided in our three papers is a necessary, but only a first step, towards answering questions related to decision making and patient outcome. As argued in paper II, prevalence studies, experimental studies and larger clinical studies are needed to increase our knowledge of the pathogenic role of individual bacterial species.

Although not firmly investigated in our studies, we believe that many of the bacteria detected by TNGS might have a pathogenic role. First, it can be argued that species detected in normally sterile fluids or tissues are more likely to represent pathogens

(111, 112). Second, it can be argued that species found repeatedly in samples from a certain type of infection are more likely to represent pathogens (111). Both assumptions are fulfilled for many of the bacterial species detected in pleural infections. The belief that pleural fluid is sterile unless there is an ongoing infection (73) was confirmed by our inclusion of a control group in which no bacteria were identified. Additionally, many of the species, such as *S. intermedius*, *F. nucleatum, P. micra* and *Eubacterium brachy,* were found in a high proportion of the patient samples, increasing the probability that these species are clinically relevant.

For samples where it is more questionable whether the sterility assumtion holds, like samples from the biliary bladder and bile duct, the clinical relevance of the TNGS findings become more difficult to interpret. The inclusion of a control group may be helpful also in such cases. In the study of acute cholecystitis (paper II) we found a clear association between presence of bacteria in the bile bladder and the diagnosis of acute cholecystitis; The rate of samples with microbes identified by TNGS in the 36 patients with infection compared to the 16 patients in the control group was 86% vs 18% (paper II). Also, the three bacteria identified in the bile of the control group patients were considered low-grade pathogens, in contrast to the well-known pathogens constituting most species found in the group of acute cholecystitis patients (Table 5, paper II). Such observations strengthen the likelihood that the identified bacteria have a role in the disease pathogenesis. However, considerations of clinical relevance are complicated by the fact that the pathogenesis of acute cholecystitis does not necessarily involve a bacterial infection. Illustrative of this is the finding of no bacteria in 14% of bile samples from patients with acute cholecystitis.

Evaluation of clinical relevance is even more complicated for the acute cholangitis findings in paper III. The belief that bacterial invasion is necessary for the pathogenesis of acute cholangitis, was supported by our study in which all the acute cholangitis samples contained bacteria. However, and in contrast to pleural empyemas and acute cholecystitis, the frequency of bacterial detections in the control group was high. Eigthy-one percent of the patients in the group of non-infectious bile duct stenosis had microbes detected in their bile, including well-known pathogens from the *Enterobacteriaceae* family and the *Enterococcus* genus. The risk of

contamination during sampling, as discussed in chapter 5.1.2, represents a possible explanation for the high prevalence of bacteria in samples from non-infectious patients. Furthermore, it is reasonable to suggest that the biliary duct of patients suffering from bile duct stenosis are more easily colonized with bacteria due to inhibition of bile flow. Whether the high rate of microbial presence in non-infectious patients and the high rate of low abundant bacteria in both infectious and non-infectious materials represent contamination during sampling, colonization, asymptomatic carriage, or pathogens (for those found in acute cholangitis samples only) of the bile duct cannot be determined based on our study. The high rate of bacteria in non-infectious clinical samples and the possibility of contamination during sampling thus reduces the clinical value of TNGS in the routine diagnostics of acute cholangitis.

The conciderations above are important for the evaluation of clinical relevance, but they do not address to which extent the results from TNGS should guide clinical decision making, for example the administration of antibiotics. Results from TNGS only provide indirect information about antimicrobial susceptibility and resistance, for example by the clinician's knowledge of the bacterium's innate resistance patterns. In that sense it has a limitation compared to both culture, which can provide phenotypic resistance profiles, and to shotgun metagenomics aiming at providing genotypic resistance profiles.

To summarize, the three papers illustrate the importance of carefully selecting type of infection as well as sampling method when considering the use of TNGS in routine diagnostics. The type of infections where TNGS will be most likely to provide clinically relevant information are infections occurring in normally sterile areas where samples can be collected with a low risk of contamination. In contrast to culture-based methods, where procedures for interpretation of findings have been established and honed through multiple clinical trials and decades of experience, much research remains to be done before the clinical utility of TNGS in routine diagnostics can be firmly established (17). Future studies should aim at assessing not only the sensitivity of NGS methods, but also the clinical significance of the findings and the impact of the findings on patient outcome.

5.3.2 The utility of TNGS in clinical microbiology research of polymicrobial invasive infections

While TNGS and other NGS methods are still struggling to find their place in clinical diagnostics of polymicrobial infections, their role in expanding our understanding of polymicrobial invasive infections is easier to identify. The discovery of new pathogens may lead to an altered or completely new understanding of the aetiology and pathogenesis of a disease, and eventually to better prevention, diagnostics, and treatment. Examples of this includes the discovery of *Helicobacter pylori* (113) as the causative agent of peptic ulcer and gastritis and the discovery of *Legionella pneumophila* as the causative agent of Legionnaires' disease (114). Both examples refer to monomicrobial infections in which identification of a single pathogen was sufficient to establish a possible aetiology and pathogenesis. In contrast, polymicrobial infections consist of multiple bacteria that may acting together through synergisms and microbial aggregate formations to establish and maintain the infection, or by the presence of key pathogens that are obligate for establishing the disease. A full understanding of the aetiology and pathogenesis of polymicrobial infections requires knowledge of the microbiota of the infection. This can only be achieved by deep sequencing studies, and not by culture alone. For example, the low recovery rate of anaerobic bacteria and several families of aerobic and facultative anaerobic bacteria by culture compared to TNGS (Figure 3) has almost certainly led to an underestimation of the importance of these bacteria in such infections.

Paper I illustrates how TNGS can serve as a tool for the formation of novel hypotheses of disease aetiology. The potential existence of key pathogens (e.g. *F. nucleatum* and *S. intermedius*) would not have been discovered without the more complete and consistant microbial descriptions of pleural empyema obtained by TNGS. *Fusobacterum nucleatum* was for example found in 17 pleural empyema by TNGS, but successfully cultured from only one of them. Further, the observed similarity between oral derived brain abscesses and primary empyema also required the sensitivity of TNGS to be revealed.

For acute cholecystitis and acute cholangitis, we did not identify any potential key pathogens, nor did we suggest any new hypothesis for the aetiology or pathogenesis

of these diseases based on the improved microbial characterization of the infections. However, these studies were relatively small, and the investigation of larger sample collections might still contribute to discover or understand hitherto unknown microbial connections. Our results could therefore serve as a useful first step for bettering our understanding and treatment of the diseases.

5.4 Limit of detection and management of contamination in TNGS

For a PCR, representing the first step in a TNGS analysis, the limit of detection (LoD) - sometimes termed the analytical sensitivity of the PCR - is the lowest concentration of a target gene that produces at least 95% positive results in replicate PCR runs. However, the final analytical sensitivity and specificity of a TNGS analysis is also dependent upon the ability to distinguish between sequences representing true target genes and contamination. In our three papers we observed 11 samples where a total of 13 bacteria were identified by culture but not by TNGS. An overview of these cases is provided in Table 5. We believe that most of these 11 cases can be explained by the sensitivity and specificity limitations of TNGS. In the following sections the different factors important for the LoD and contamination management in TNGS will be discussed and the 11 cases in Table 5 will be used as illustrative examples.

Table 5: Overview of the 13 species that were identified by culture but not with targeted next generation sequencing (TNGS)

#	Clinical sample, sample study number	CT value[a]	Species identified only by culture	Additional information	Probable causes of negative TNGS
1	Acute cholecysitits, 9	12,4	*Klebsiella pneumoniae*	Identified both by *rpoB_Ent* and 16S TNGS but with only 16 reads and thus below the fixed OTU-size cutoff	Inadequate sequencing depth (analytical sensitivity).
2	Acute cholecysitits, 27	26,1	*Staphylococcus epidermidis*	*rpoB*_ESS also negative	Growth represents contamination in culture lab / Concentration of *S. epidermidis* below the absolute lower LoD of TNGS (analytical sensitivity).

3	Acute cholangitis, 4	12,9	*Granulicatella adiacens*		Inadequate sequencing depth (analytical sensitivity)/ Concentration of *G. adiacens* below the absolute lower LoD of TNGS (analytical sensitivity)
4	Non-infectious bile duct stenosis, 27	33,3	*Staphylococcus warneri*		Growth represents contamination in culture lab / Concentration of *S. warneri* below the absolute lower LoD of TNGS (analytical sensitivity).
5	Non-infectious bile duct stenosis, 19	12,2	*Enterococcus faecalis*	Identified by TNGS but with only 12 reads in the replicate with highest sequencing depth	Inadequate sequencing depth (analytical sensitivity).
6	Non-infectious bile duct stenosis, 38	22,4	*Staphylococcus epidermidi*, *Corynebacterium pseudodiphtericum*	*rpoB*_ESS also negative	Growth represents contamination in culture lab / Both bacteria present in concentrations below the absolute LoD of TNGS
7	Non-infectious bile duct stenosis, 39	17,0	*Aeromonas* species	Growth of two morphologically different Aeromonas species by culture, while only a single OTU assigned to the Aeromonas veronii group were identified by TNGS.	16S gene sequencing not able to distinguish between different Aeromonas species. All will cluster in one group / Only one species but with morphologically heterogeneous appearance
8	Pleural empyema, 17	30,4	*Streptococcus constellatus*	*rpoB*_ESS also negative	Concentration of *S. constellatus* below the absolute lower LoD of TNGS (analytical sensitivity).
9	Pleural empyema, 18	12,2	*Staphylococcus aureus, Klebsiella pneumoniae*	*S. aureus* also identified by *rpoB*_ESS. *rpoB*_Ent negative.	Inadequate sequencing depth (analytical sensitivity) / Concentration of bacteria below the absolute lower LoD of TNGS (analytical sensitivity).
10	Pleural empyema, 48	19,6	*Cutibacterium acnes*	Both detected by TNGS, but since *C. acnes* and *S. epidermidis* were among the ten most abundant microbes in the negative controls, they were not considered valid identifications according to the criteria applied. Both bacteria were also identified by Sanger 16S rRNA sequencing. *S. epidermidis* was additionally identified by *rpoB*_ESS.	Contaminant filtering principle
11	Pleural empyema, 24	19,3	*Staphylococcus epidermidis*		

[a] Cycle threshold value of 16S rRNA amplicon PCR. A low CT-value indicates a high amount of target DNA.

5.4.1 Limit of detection

The factors determining the LoD can be divided into factors that are sample and sequencing run dependent and factors that are independent of the features of the specific run or samples analyzed. This is well illustrated in the experiments performed in paper III. The run and sample dependent factors include the sequencing depth of the sequencing run, the microbial mass of the target sample and the microbial diversity of the target sample. In sequencing runs with an inadequate sequence coverage, this will define the LoD for that sequencing run. However, if the sequencing coverage is adequate and therefore not the limiting factor, the LoD will be defined by the target gene concentration needed for the gene target to be included in the PCR-template. This is a factor that is always present and thus run and sample independent. In the following section we will term the latter the absolute limit of detection of TNGS, to distinguish it from the LoD determined by the sequencing coverage.

5.4.1.1 Absolute limit of detection

In paper III the absolute limit of detection was defined as the value where the DNA input of a given species in a sample approached one copy of the target gene per PCR. Our results confirmed that for bacteria approaching this concentration it will be random whether they are included in the pieppted PCR template or not. By theoretical calculations we found that the target gene of a bacterium is getting close to one copy per PCR when the concentration is around 100 CFU/ml in the original sample (Supplementary Table S3, paper III).

The theoretical absolute limit of detection for our sequencing protocol of around 100 CFU/ml, was supported by our replicate sequencing of mock community dilutions. A more accurate determination of the absolute LoD is of little value as it will vary somewhat dependent upon e.g. the number of target gene copies in the specific target microbe, the extraction method and the exact volume of PCR-template used. Some

sequencing protocols use 5 µl, instead of the 2 µl PCR-template used in our protocol, thus theoretically slightly increasing the absolute LoD. The LoD of our sequencing processing is in concordance with findings in other studies, like Culbreath et al. who found an absolute LoD of 10-100 CFU/ml (108).
Three of the 13 species found exclusively by culture (Sample #8 and #9 in Table 5) grew in blood culture bottles only. These three species may very well represent cases where the absolute LoD of TNGS explaines the false negative sequencing results. Culturing in blood culture bottles is recommended as a routine diagnostics (115) of pleural empyema because of the increased sensitivity compared to standard agar culture (116). In a blood culture bottle up to 10 ml of sample volume is used, and the LoD for living bacteria has been shown to be close to or below 1 CFU/ml, which is far better than the LoD obtainable by TNGS (117).

5.4.1.2 Sequencing coverage

Sequencing coverage, a term used both in whole-genome shotgun sequencing and TNGS, is defined as the fraction of the metagenome that is represented in the metagenomic dataset (118). For 16S rRNA TNGS, coverage can more specifically be defined as the fraction of the total collection of 16S rRNA genes in the sample that is represented among the sequencing reads. An insufficient sequencing coverage will reduce the limit of detection and may cause a false low alpha diversity (119). It is therefore important to assess whether a sufficient sequencing coverage has been achieved when performing TNGS analyses. Surprisingly few studies address the issue of sequencing coverage in TNGS (119). In paper III we suggest using the confirmed presence of background contaminant sequences among the sequencing reads as a marker for sufficient sequencing depth, thus serving as a marker for an adequate sequencing coverage.

The two major factors influencing the sequencing coverage are the complexity of the sample microbial community and the sequencing depth (119). It is obvious, that for complex samples containing numerous species with a wide variation in relative concentrations, a higher sequencing depth is needed to cover all present species as compared to a sample containing a single or a few evenly distributed species. This is

well demonstrated in paper III by our sequencing of samples in two replicates with varying sequencing depths. Increased sequencing depth led to better sequencing coverage and thus to the identification of more species in many of the samples (Figure 8, Paper III). Insufficient sequencing depth is the most likely explanation for at least three out of the 13 species identified by culture and not by TNGS (*K. pneumonia* in #1, *E. faecalis* in #5 and *S. aureus* in #9, Table 5). As shown in Table 5, these three species were all either identified by 16SrRNA TNGS but below the fixed OTU size cutoff threshold of 50 reads, or by *rpoB* TNGS, or both, thus indicating that DNA from these species were indeed present in the samples, but in low abundancies.

There are no general consensus recommendations on sequencing depth for polymicrobial infections. The Illumina MiSeq Protocol for 16S TNGS recommends a sequencing depth of > 100.000 reads for full characterization of the bacterial composition of a sample, but others have found that a sequencing depth of 1.500.000 reads was necessary to reach the absolute LoD (108). Our result demonstrates that the main factor determining the adequacy of sequencing depth is the bacterial load of the sample. We found that even a sequencing depth of several hundred thousand reads was insufficient to obtain an adequate coverage in samples with a high bacterial load. Again, the above mentioned three species found by culture only are good examples. All three samples had a high bacterial load (CT-values 12,2-12,4) and despite a relatively high sequencing depth (158.165 reads in #1, 291.151 reads in #5 and 68.485 reads in #9, Table 5) we did not identify any contaminant reads in any of the three samples. In contrast, in samples with a low bacterial load, (e.g. replicate 1, sample 23, 28 and 35, Supplementary Table S6, Paper III) a sequencing depth even below 50.000 reads was sufficient to sequence deep into the background contamination and reaching the absolute LoD of TNGS.

5.4.2 Background contamination in TNGS

As described in the introduction, background contamination is one of the main obstacles in providing accurate and trustworthy sequencing results in TNGS. As demonstrated by the work in this thesis, understanding the patterns of contamination

in sequencing results are crucial for proper handling. The concepts in paper III, where we suggest a specific method for the management of contamination in diagnostic laboratories, were conceived while working on the first two papers, thus explaining why background contamination was managed slightly differently in the three papers.

The most important principles are:

1. Know your contamination. The level and composition of background contamination will, as we underscore in paper III, vary between laboratories, between extraction kits and PCR reagents, and even between batches of the same extraction kits and PCR reagents. This implies that the specific method for managing background contamination in 16S rRNA TNGS should be elaborated in each laboratorium.
2. The level of background contamination is relatively stable and will always appear when sequencing at deep levels. However, its relative abundance will increase as the microbial mass of the sample gets lower (12, 14, 15). A key to handle background contamination properly is to identify at what level contaminating OTUs start to appear, knowing that OTUs appearing below that level may represent contaminants. In our material we were able to accurately identify at what level contamination started to appear since the major contaminant OTUs were consistent, making it possible to establish an individual, robust cutoff for valid identifications in each sample. Whether there is a similar pattern of highly consistent top contaminants should be part of the initial investigation of every laboratorium setting up their own TNGS assay.
3. The microbial mass of the analyzed sample should be assessed together with the sequencing depth as this will determine the relative level of background contamination in the sequencing results. A high microbial mass implies that sequencing coverage may be a limiting factor, and thus that contamination may not appear at all if an adequate sequencing depth is not achieved. In contrast, background contamination will be a major challenge in samples with a low microbial mass and/or adequate sequencing depth.

Evaluation of the microbial mass is particularly important if a bacterial species present in relative high abundance in the sample is also one of the dominant background contaminants. As discussed in paper III, information about the microbial mass and sequencing depth may enable the determination of whether that specific bacterium is a true finding or represents contamination (1). Two of the species identified by culture only from paper I (#10 and #11 in Table 5) serve as examples of such a situation. Both species, *S. epidermidis* and *C. acnes*, were detected by TNGS as well as by culture, but were removed from the TNGS results as invalid identifications as they were also among the top abundant contaminant bacteria in the extraction controls. However, the CT-values in both samples were low (19,6 and 19,3 respectively), indicating a high microbial mass. By applying the method suggested by Kommedal *et al.* (2014) to calculate the sample-specific cutoff, we determined that the cutoff from where contamination may start to appear was 461 reads for sample #10 containing *C. acnes*, and 284 reads for sample #11 containing *S. epidermidis*. Both *C. acnes* (# of reads 51505) and *S. epidermidis* (# of reads 16084) had an abundance way above that cutoff and could consequently have been considered as true findings.

4. The above example underscores a final principle: It is important not to be locked into a single method or algorithm when dealing with background contamination. In studies aiming at validating and evaluating a specific algorithm for managing contamination, as in our paper III, it is important to strictly follow the algorithm tested. If not, the pros and cons of the method will not be discovered and made visible. However, in a routine diagnostic setting, it may be more beneficial to combine the information acquired by our suggested approach with other methods. One approach, as we discuss in paper III, could be to combine our contamination management method with the use of an expert review of the results. The use of an expert review to remove biologically unexpected finding is suggested as a supplementary method for managing contaminants in several articles (12, 56) and have also been used as the main method in a few articles (9, 10). However, an expert opinion should

not be based on knowledge of clinical relevance only. It needs to be combined with an in depth understanding of the general patterns of contamination in TNGS, as well as the benefits and limitations of the chosen approach for contaminant filtering.

6 Conclusions

In this thesis, we have confirmed the improved sensitivity of TNGS as compared to traditional diagnostics for the microbiological characterization of pleural empyema, acute cholecystitis and acute cholangitis, thus further underpinning the inadequacy of culture for polymicrobial infections. We have also established that the inclusion of *rpoB* sequencing can provide more accurate species identification within certain genera. We have explored the unpredictable nature of background contamination in TNGS and highlighted challenges related to filtering of sequencing results. We have suggested and evaluated a method for managing laboratory contamination, including rules and cutoffs for post-sequencing processing and interpretation to maximize the accuracy of the results, that is suitable for diagnostic laboratories. Finally, we have shown how knowledge obtained by TNGS can lead to new hypotheses for the aetiology and pathogenesis of an infection.

7 Future research

Most studies of TNGS on clinical materials have focused on comparing the sensitivity of the method to conventional culture. Future research should also aim at performing clinical studies to evaluate the clinical utility of the increased sensitivity obtained by TNGS. There is also a need for more studies aiming at characterizing the microbiome of different invasive polymicrobial infections to increase our knowledge and possibly also our understanding of their pathogenesis and aetiology. The attempts at complete microbial characterizations of pleural empyema, acute cholangitis and acute cholecystitis as provided in our three studies needs to be confirmed by other studies. Such studies should also address the methodological weaknesses of our studies as outlined in chapter 5.1.

Developments within deep sequencing technology is progressing rapidly. In particular, long read sequencing as provided by Oxford Nanopore Technologies represents a promising tool for both metagenomic and targeted amplicon sequencing (120). The short turn-around time combined with a format that allows for analyzing individual samples at a relatively reasonable cost resolves two of the major obstacles related to current second-generation platforms. We will emphasize that the use of novel technologies such as nanopore sequencing for characterization of polymicrobial infections needs to be validated and compared to the current gold standard; either TNGS or shotgun metagenomic sequencing. This places high demands on laboratories that will carry out this type of studies, as they must have sufficient knowledge and experience across the methods.

For pleural empyema, the identification of potential key pathogens should be further elaborated. Basic microbiological research should aim at figuring out why these microbes seem to be so important. The 1990s research on oral microbial communities, and in particular the role of *F. nucleatum*, could serve as a model for such studies (121–123). Another research question of high interest is the degree to which targeted antibiotic treatment towards these few key pathogens is sufficient for antibiotic management of primary pleural empyema.

Source of data

1. Kommedal O, Wilhelmsen MT, Skrede S, Meisal R, Jakovljev A, Gaustad P, Hermansen NO, Vik-Mo E, Solheim O, Ambur OH, Saebo O, Hostmaelingen CT, Helland C. 2014. Massive parallel sequencing provides new perspectives on bacterial brain abscesses. J Clin Microbiol 52:1990–7.

2. Cummings LA, Kurosawa K, Hoogestraat DR, SenGupta DJ, Candra F, Doyle M, Thielges S, Land TA, Rosenthal CA, Hoffman NG, Salipante SJ, Cookson BT. 2016. Clinical Next Generation Sequencing Outperforms Standard Microbiological Culture for Characterizing Polymicrobial Samples. Clin Chem 62:1465–1473.

3. Zheng D, Liwinski T, Elinav E. 2020. Interaction between microbiota and immunity in health and disease. Cell Res 30:492–506.

4. Boers SA, Jansen R, Hays JP. 2019. Understanding and overcoming the pitfalls and biases of next-generation sequencing (NGS) methods for use in the routine clinical microbiological diagnostic laboratory. Eur J Clin Microbiol 38:1059–1070.

5. Zhang Y, Hu A, Andini N, Yang S. 2019. A “culture” shift: Application of molecular techniques for diagnosing polymicrobial infections. Biotechnol Adv 37:476–490.

6. Salipante SJ, Sengupta DJ, Rosenthal C, Costa G, Spangler J, Sims EH, Jacobs MA, Miller SI, Hoogestraat DR, Cookson BT, McCoy C, Matsen FA, Shendure J, Lee CC, Harkins TT, Hoffman NG. 2013. Rapid 16S rRNA next-generation sequencing of polymicrobial clinical samples for diagnosis of complex bacterial infections. Plos One 8:e65226.

7. Kozlov A, Bean L, Hill EV, Zhao L, Li E, Wang GP. 2018. Molecular Identification of Bacteria in Intra-abdominal Abscesses Using Deep Sequencing. Open Forum Infect Dis 5:ofy025.

8. Tarabichi M, Shohat N, Goswami K, Alvand A, Silibovsky R, Belden K, Parvizi J. 2018. Diagnosis of Periprosthetic Joint Infection: The Potential of Next-Generation Sequencing. J Bone Jt Surg 100:147–154.

9. Fida M, Wolf MJ, Hamdi A, Vijayvargiya P, Garrigos ZE, Khalil S, Greenwood-Quaintance KE, Thoendel MJ, Patel R. 2021. Detection of Pathogenic Bacteria from Septic Patients Using 16S rRNA gene Targeted Metagenomic Sequencing. Clin Infect Dis ciab349-.

10. Flurin L, Wolf MJ, Greenwood-Quaintance KE, Sanchez-Sotelo J, Patel R. 2021. Targeted next generation sequencing for elbow periprosthetic joint infection diagnosis. Diagn Micr Infec Dis 101:115448.

11. Bryan A, Kirkpatrick LM, Manaloor JJ, Salipante SJ. 2017. 16S rRNA deep sequencing identifies Actinotignum schaalii as the major component of a polymicrobial intra-abdominal infection and implicates a urinary source. Jmm Case Reports 4:e005091.

12. Salter SJ, Cox MJ, Turek EM, Calus ST, Cookson WO, Moffatt MF, Turner P, Parkhill J, Loman NJ, Walker AW. 2014. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. Bmc Biol 12:87.

13. Eisenhofer R, Minich JJ, Marotz C, Cooper A, Knight R, Weyrich LS. 2019. Contamination in Low Microbial Biomass Microbiome Studies: Issues and Recommendations. Trends Microbiol 27:105–117.

14. Karstens L, Asquith M, Davin S, Fair D, Gregory WT, Wolfe AJ, Braun J, McWeeney S. 2019. Controlling for Contaminants in Low-Biomass 16S rRNA Gene Sequencing Experiments. Msystems 4.

15. Drengenes C, Wiker HG, Kalananthan T, Nordeide E, Eagan TML, Nielsen R. 2019. Laboratory contamination in airway microbiome studies. Bmc Microbiol 19:187.

16. Chakravorty S, Helb D, Burday M, Connell N, Alland D. 2007. A detailed analysis of 16S ribosomal RNA gene segments for the diagnosis of pathogenic bacteria. J Microbiol Meth 69:330–9.

17. Miller S, Chiu C, Rodino KG, Miller MB. 2020. Point-Counterpoint: Should We Be Performing Metagenomic Next-Generation Sequencing for Infectious Disease Diagnosis in the Clinical Laboratory? J Clin Microbiol 58.

18. Dickson RP, Erb-Downward JR, Martinez FJ, Huffnagle GB. 2015. The Microbiome and the Respiratory Tract. Annu Rev Physiol 78:1–24.

19. Gilbert J, Blaser MJ, Caporaso JG, Jansson J, Lynch SV, Knight R. 2018. Current understanding of the human microbiome. Nat Med 24:392–400.

20. Brogden KA, Guthmiller JM, Taylor CE. 2005. Human polymicrobial infections. Lancet 365:253–5.

21. Peters BM, Jabra-Rizk MA, O'May GA, Costerton JW, Shirtliff ME. 2012. Polymicrobial Interactions: Impact on Pathogenesis and Human Disease. Clin Microbiol Rev 25:193–213.

22. Tay WH, Chong KK, Kline KA. 2016. Polymicrobial-Host Interactions during Infection. J Mol Biol 428:3355–71.

23. Besser J, Carleton HA, Gerner-Smidt P, Lindsey RL, Trees E. 2018. Next-generation sequencing technologies and their application to the study and control of bacterial infections. Clin Microbiol Infec 24:335–341.

24. Kommedal O, Kvello K, Skjastad R, Langeland N, Wiker HG. 2009. Direct 16S rRNA gene sequencing from clinical specimens, with special focus on polybacterial samples and interpretation of mixed DNA chromatograms. J Clin Microbiol 47:3562–8.

25. Kommedal O, Lekang K, Langeland N, Wiker HG. 2011. Characterization of polybacterial clinical samples using a set of group-specific broad-range primers targeting the

16S rRNA gene followed by DNA sequencing and RipSeq analysis. J Med Microbiol 60:927–36.

26. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z, Dewell SB, Du L, Fierro JM, Gomes XV, Godwin BC, He W, Helgesen S, Ho CH, Irzyk GP, Jando SC, Alenquer ML, Jarvie TP, Jirage KB, Kim JB, Knight JR, Lanza JR, Leamon JH, Lefkowitz SM, Lei M, Li J, Lohman KL, Lu H, Makhijani VB, McDade KE, McKenna MP, Myers EW, Nickerson E, Nobile JR, Plant R, Puc BP, Ronan MT, Roth GT, Sarkis GJ, Simons JF, Simpson JW, Srinivasan M, Tartaro KR, Tomasz A, Vogt KA, Volkmer GA, Wang SH, Wang Y, Weiner MP, Yu P, Begley RF, Rothberg JM. 2005. Genome sequencing in microfabricated high-density picolitre reactors. Nature 437:376–80.

27. Buermans HP, Dunnen JT den. 2014. Next generation sequencing technology: Advances and applications. Biochim Biophys Acta 1842:1932–1941.

28. Illumina. 2013. 16S Metagenomic Sequencing Library Preparation : Preparing 16S Ribosomal RNA Gene Amplicons for the Illumina MiSeq System. Illumina Inc.

29. Perez-Cobas AE, Gomez-Valero L, Buchrieser C. 2020. Metagenomic approaches in microbial ecology: an update on whole-genome and marker gene sequencing analyses. Microb Genom 6.

30. Gordon A. 2010. FASTX-Toolkit. http://hannonlab.cshl.edu/fastx_toolkit/.

31. Zhang J, Kobert K, Flouri T, Stamatakis A. 2014. PEAR: a fast and accurate Illumina Paired-End reAd mergeR. Bioinformatics 30:614–20.

32. Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30:2114–20.

33. Aronesty E. 2013. Comparison of Sequencing Utility Programs. Open Bioinform J 7:1–8.

34. Schubert M, Lindgreen S, Orlando L. 2016. AdapterRemoval v2: rapid adapter trimming, identification, and read merging. Bmc Res Notes 9:88.

35. Stackebrandt E, Goebel BM. 1994. A Place for DNA-DNA Reassociation and 16s Ribosomal-Rna Sequence-Analysis in the Present Species Definition in Bacteriology. Int J Syst Bacteriol 44:846–849.

36. Edgar RC. 2018. Updating the 97% identity threshold for 16S ribosomal RNA OTUs. Bioinformatics 34:2371–2375.

37. Petti CA, Institute. Clinical and Laboratory Standards. 2008. Interpretive criteria for identification of bacteria and fungi by DNA target sequencing : approved guideline. Clinical and Laboratory Standards Institute, Wayne, PA.

38. Edgar RC, Flyvbjerg H. 2015. Error filtering, pair assembly and error correction for next-generation sequencing reads. Bioinformatics 31:3476–82.

39. Schloss PD, Gevers D, Westcott SL. 2011. Reducing the effects of PCR amplification and sequencing artifacts on 16S rRNA-based studies. Plos One 6:e27310.

40. Porter TM, Hajibabaei M. 2018. Scaling up: A guide to high-throughput genomic approaches for biodiversity analysis. Mol Ecol 27:313–338.

41. Callahan BJ, McMurdie PJ, Holmes SP. 2017. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. Isme J 11:2639–2643.

42. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol 215:403–410.

43. Wang Q, Garrity GM, Tiedje JM, Cole JR. 2007. Naïve Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy ▿ †. Appl Environ Microb 73:5261–5267.

44. Allard G, Ryan FJ, Jeffery IB, Claesson MJ. 2015. SPINGO: a rapid species-classifier for microbial amplicon sequences. Bmc Bioinformatics 16:324.

45. Gao X, Lin H, Revanna K, Dong Q. 2017. A Bayesian taxonomic classification method for 16S rRNA gene sequences with improved species-level accuracy. Bmc Bioinformatics 18:247.

46. Rognes T, Flouri T, Nichols B, Quince C, Mahé F. 2016. VSEARCH: a versatile open source tool for metagenomics. Peerj 4:e2584.

47. Bokulich NA, Kaehler BD, Rideout JR, Dillon M, Bolyen E, Knight R, Huttley GA, Caporaso JG. 2018. Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2's q2-feature-classifier plugin. Microbiome 6:90.

48. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. Bmc Bioinformatics 10:421.

49. DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, Huber T, Dalevi D, Hu P, Andersen GL. 2006. Greengenes, a Chimera-Checked 16S rRNA Gene Database and Workbench Compatible with ARB. Appl Environ Microb 72:5069–5072.

50. Ashelford KE, Chuzhanova NA, Fry JC, Jones AJ, Weightman AJ. 2005. At Least 1 in 20 16S rRNA Sequence Records Currently Held in Public Repositories Is Estimated To Contain Substantial Anomalies. Appl Environ Microb 71:7724–7736.

51. Pruesse E, Quast C, Knittel K, Fuchs BM, Ludwig W, Peplies J, Glöckner FO. 2007. SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. Nucleic Acids Res 35:7188–7196.

52. Cole JR, Wang Q, Fish JA, Chai B, McGarrell DM, Sun Y, Brown CT, Porras-Alfaro A, Kuske CR, Tiedje JM. 2014. Ribosomal Database Project: data and tools for high throughput rRNA analysis. Nucleic Acids Res 42:D633–D642.

53. Edgar R. 2018. Taxonomy annotation and guide tree errors in 16S rRNA databases. Peerj 6:e5030.

54. Sierra MA, Li Q, Pushalkar S, Paul B, Sandoval TA, Kamer AR, Corby P, Guo Y, Ruff RR, Alekseyenko AV, Li X, Saxena D. 2020. The Influences of Bioinformatics Tools and Reference Databases in Analyzing the Human Oral Microbial Community. Genes-basel 11:878.

55. Balvočiūtė M, Huson DH. 2017. SILVA, RDP, Greengenes, NCBI and OTT — how do these taxonomies compare? Bmc Genomics 18:114.

56. Goffau MC de, Lager S, Salter SJ, Wagner J, Kronbichler A, Charnock-Jones DS, Peacock SJ, Smith GCS, Parkhill J. 2018. Recognizing the reagent microbiome. Nat Microbiol 3:851–853.

57. Theis KR, Romero R, Winters AD, Greenberg JM, Gomez-Lopez N, Alhousseini A, Bieda J, Maymon E, Pacora P, Fettweis JM, Buck GA, Jefferson KK, F. 3rd Strauss J, Erez O, Hassan SS. 2019. Does the human placenta delivered at term have a microbiota? Results of cultivation, quantitative real-time PCR, 16S rRNA gene sequencing, and metagenomics. Am J Obstet Gynecol 220:267 e1-267 e39.

58. Abayasekara LM, Perera J, Chandrasekharan V, Gnanam VS, Udunuwara NA, Liyanage DS, Bulathsinhala NE, Adikary S, Aluthmuhandiram JVS, Thanaseelan CS, Tharmakulasingam DP, Karunakaran T, Ilango J. 2017. Detection of bacterial pathogens from clinical specimens using conventional microbial culture and 16S metagenomics: a comparative study. Bmc Infect Dis 17:631.

59. Davis NM, Proctor DM, Holmes SP, Relman DA, Callahan BJ. 2018. Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data. Microbiome 6:226.

60. Petti CA. 2007. Detection and identification of microorganisms by gene amplification and sequencing. Clin Infect Dis 44:1108–14.

61. Woese CR. 1987. Bacterial evolution. Microbiol Rev 51:221–71.

62. Woese CR, Fox GE. 1977. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proc Natl Acad Sci U S A 74:5088–90.

63. Janda JM, Abbott SL. 2007. 16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils, and pitfalls. J Clin Microbiol 45:2761–4.

64. E. 3rd Clarridge J. 2004. Impact of 16S rRNA gene sequence analysis for identification of bacteria on clinical microbiology and infectious diseases. Clin Microbiol Rev 17:840–62, table of contents.

65. Villmones HC, Haug ES, Ulvestad E, Grude N, Stenstad T, Halland A, Kommedal O. 2018. Species Level Description of the Human Ileal Bacterial Microbiota. Sci Rep-uk 8:4736.

66. Adekambi T, Drancourt M, Raoult D. 2009. The rpoB gene as a tool for clinical microbiologists. Trends Microbiol 17:37–45.

67. Adekambi T, Shinnick TM, Raoult D, Drancourt M. 2008. Complete rpoB gene sequencing as a suitable supplement to DNA-DNA hybridization for bacterial species and genus delineation. Int J Syst Evol Micr 58:1807–14.

68. Drancourt M, Bollet C, Carta A, Rousselier P. 2001. Phylogenetic analyses of Klebsiella species delineate Klebsiella and Raoultella gen. nov., with description of Raoultella ornithinolytica comb. nov., Raoultella terrigena comb. nov. and Raoultella planticola comb. nov. Int J Syst Evol Micr 51:925–932.

69. Drancourt M, Raoult D. 2002. rpoB gene sequence-based identification of Staphylococcus species. J Clin Microbiol 40:1333–8.

70. Drancourt M, Roux V, Fournier PE, Raoult D. 2004. rpoB gene sequence-based identification of aerobic Gram-positive cocci of the genera Streptococcus, Enterococcus, Gemella, Abiotrophia, and Granulicatella. J Clin Microbiol 42:497–504.

71. Mollet C, Drancourt M, Raoult D. 1997. rpoB sequence analysis as a novel basis for bacterial identification. Mol Microbiol 26:1005–11.

72. Nygaard RM. 2017. Bruk av massiv parallell sekvensering for påvisning og identifikasjon av mikrober i galle hos pasienter med akutt kolangitt. The University of Bergen.

73. Corcoran JP, Wrightson JM, Belcher E, DeCamp MM, Feller-Kopman D, Rahman NM. 2015. Pleural infection: past, present, and future directions. Lancet Respir Medicine 3:563–77.

74. Shen KR, Bribriesco A, Crabtree T, Denlinger C, Eby J, Eiken P, Jones DR, Keshavjee S, Maldonado F, Paul S, Kozower B. 2017. The American Association for Thoracic Surgery consensus guidelines for the management of empyema. J Thorac Cardiovasc Surg 153:e129–e146.

75. Tobin CL, Lee YC. 2012. Pleural infection: what we need to know but don't. Curr Opin Pulm Med 18:321–5.

76. Lisboa T, Waterer GW, Lee YC. 2011. Pleural infection: changing bacteriology and its implications. Respirology 16:598–603.

77. Maskell NA, Batt S, Hedley EL, Davies CW, Gillespie SH, Davies RJ. 2006. The bacteriology of pleural infection by genetic and standard methods and its mortality significance. Am J Resp Crit Care 174:817–23.

78. Hassan M, Cargill T, Harriss E, Asciak R, Mercer RM, Bedawi EO, McCracken DJ, Psallidas I, Corcoran JP, Rahman NM. 2019. The microbiology of pleural infection in adults: a systematic review. Eur Respir J 54:1900542.

79. Dyrhovden R, Nygaard RM, Patel R, Ulvestad E, Kommedal O. 2019. The bacterial aetiology of pleural empyema. A descriptive and comparative metagenomic study. Clin Microbiol Infec 25:981–986.

80. Wrightson J, Wray J, Street T, Chapman S, Crook D, Rahman N. 2014. S114 Previously Unrecognised Oral Anaerobes In Pleural Infection. Thorax 69:A61–A62.

81. Kimura Y, Takada T, Kawarada Y, Nimura Y, Hirata K, Sekimoto M, Yoshida M, Mayumi T, Wada K, Miura F, Yasuda H, Yamashita Y, Nagino M, Hirota M, Tanaka A, Tsuyuguchi T, Strasberg SM, Gadacz TR. 2007. Definitions, pathophysiology, and epidemiology of acute cholangitis and cholecystitis: Tokyo Guidelines. J Hepato-biliary-pan 14:15–26.

82. Hirota M, Takada T, Kawarada Y, Nimura Y, Miura F, Hirata K, Mayumi T, Yoshida M, Strasberg S, Pitt H, Gadacz TR, Santibanes E, Gouma DJ, Solomkin JS, Belghiti J, Neuhaus H, Büchler MW, Fan S, Ker C, Padbury RT, Liau K, Hilvano SC, Belli G, Windsor JA, Dervenis C. 2007. Diagnostic criteria and severity assessment of acute cholecystitis: Tokyo Guidelines. J Hepato Biliary Pancreat Surg 14:78–82.

83. Yoshida M, Takada T, Kawarada Y, Tanaka A, Nimura Y, Gomi H, Hirota M, Miura F, Wada K, Mayumi T, Solomkin JS, Strasberg S, Pitt HA, Belghiti J, Santibanes E de, Fan ST, Chen MF, Belli G, Hilvano SC, Kim SW, Ker CG. 2007. Antimicrobial therapy for acute cholecystitis: Tokyo Guidelines. J Hepato-biliary-pan 14:83–90.

84. Wada K, Takada T, Kawarada Y, Nimura Y, Miura F, Yoshida M, Mayumi T, Strasberg S, Pitt HA, Gadacz TR, Buchler MW, Belghiti J, Santibanes E de, Gouma DJ, Neuhaus H, Dervenis C, Fan ST, Chen MF, Ker CG, Bornman PC, Hilvano SC, Kim SW, Liau KH, Kim MH. 2007. Diagnostic criteria and severity assessment of acute cholangitis: Tokyo Guidelines. J Hepato-biliary-pan 14:52–8.

85. Tanaka A, Takada T, Kawarada Y, Nimura Y, Yoshida M, Miura F, Hirota M, Wada K, Mayumi T, Gomi H, Solomkin JS, Strasberg SM, Pitt HA, Belghiti J, Santibanes E de, Padbury R, Chen MF, Belli G, Ker CG, Hilvano SC, Fan ST, Liau KH. 2007. Antimicrobial therapy for acute cholangitis: Tokyo Guidelines. J Hepato-biliary-pan 14:59–67.

86. Kiriyama S, Kozaka K, Takada T, Strasberg SM, Pitt HA, Gabata T, Hata J, Liau KH, Miura F, Horiguchi A, Liu KH, Su CH, Wada K, Jagannath P, Itoi T, Gouma DJ, Mori Y, Mukai S, Gimenez ME, Huang WS, Kim MH, Okamoto K, Belli G, Dervenis C, Chan ACW, Lau WY, Endo I, Gomi H, Yoshida M, Mayumi T, Baron TH, Santibanes E de, Teoh AYB, Hwang TL, Ker CG, Chen MF, Han HS, Yoon YS, Choi IS, Yoon DS, Higuchi R, Kitano S, Inomata M, Deziel DJ, Jonas E, Hirata K, Sumiyama Y, Inui K, Yamamoto M. 2018. Tokyo Guidelines 2018: diagnostic criteria and severity grading of acute cholangitis (with videos). J Hepato-bil-pan Sci 25:17–30.

87. Gomi H, Solomkin JS, Schlossberg D, Okamoto K, Takada T, Strasberg SM, Ukai T, Endo I, Iwashita Y, Hibi T, Pitt HA, Matsunaga N, Takamori Y, Umezawa A, Asai K, Suzuki K, Han H, Hwang T, Mori Y, Yoon Y, Huang WS, Belli G, Dervenis C, Yokoe M, Kiriyama S, Itoi T, Jagannath P, Garden OJ, Miura F, Santibañes E, Shikata S, Noguchi Y, Wada K, Honda G, Supe AN, Yoshida M, Mayumi T, Gouma DJ, Deziel DJ, Liau K, Chen M, Liu K, Su C, Chan ACW, Yoon D, Choi I, Jonas E, Chen X, Fan ST, Ker C, Giménez

ME, Kitano S, Inomata M, Mukai S, Higuchi R, Hirata K, Inui K, Sumiyama Y, Yamamoto M. 2018. Tokyo Guidelines 2018: antimicrobial therapy for acute cholangitis and cholecystitis. J Hepato-bil-pan Sci 25:3–16.

88. Yokoe M, Hata J, Takada T, Strasberg SM, Asbun HJ, Wakabayashi G, Kozaka K, Endo I, Deziel DJ, Miura F, Okamoto K, Hwang T, Huang WS, Ker C, Chen M, Han H, Yoon Y, Choi I, Yoon D, Noguchi Y, Shikata S, Ukai T, Higuchi R, Gabata T, Mori Y, Iwashita Y, Hibi T, Jagannath P, Jonas E, Liau K, Dervenis C, Gouma DJ, Cherqui D, Belli G, Garden OJ, Giménez ME, Santibañes E, Suzuki K, Umezawa A, Supe AN, Pitt HA, Singh H, Chan ACW, Lau WY, Teoh AYB, Honda G, Sugioka A, Asai K, Gomi H, Itoi T, Kiriyama S, Yoshida M, Mayumi T, Matsumura N, Tokumura H, Kitano S, Hirata K, Inui K, Sumiyama Y, Yamamoto M. 2018. Tokyo Guidelines 2018: diagnostic criteria and severity grading of acute cholecystitis (with videos). J Hepato-bil-pan Sci 25:41–54.

89. Csendes A, Burdiles P, Maluenda F, Diaz JC, Csendes P, Mitru N. 1996. Simultaneous bacteriologic assessment of bile from gallbladder and common bile duct in control subjects and patients with gallstones and common duct stones. Arch Surg 131:389–94.

90. Asai K, Watanabe M, Kusachi S, Tanaka H, Matsukiyo H, Osawa A, Saito T, Kodama H, Enomoto T, Nakamura Y, Okamoto Y, Saida Y, Nagao J. 2012. Bacteriological analysis of bile in acute cholecystitis according to the Tokyo guidelines. J Hepato-bil-pan Sci 19:476–486.

91. Galili O, S. JrE, Matter I, Madi H, Brodsky A, Galis I, S. SrE. 2008. The effect of bactibilia on the course and outcome of laparoscopic cholecystectomy. Eur J Clin Microbiol 27:797–803.

92. Gomi H, Takada T, Hwang TL, Akazawa K, Mori R, Endo I, Miura F, Kiriyama S, Matsunaga N, Itoi T, Yokoe M, Chen MF, Jan YY, Ker CG, Wang HP, Wada K, Yamaue H, Miyazaki M, Yamamoto M. 2017. Updated comprehensive epidemiology, microbiology, and outcomes among patients with acute cholangitis. J Hepato-bil-pan Sci 24:310–318.

93. Nitzan O, Brodsky Y, Edelstein H, Hershko D, Saliba W, Keness Y, Peretz A, Chazan B. 2017. Microbiologic Data in Acute Cholecystitis: Ten Years' Experience from Bile Cultures Obtained during Percutaneous Cholecystostomy. Surg Infect 18:345–349.

94. Yokoe M, Takada T, Strasberg SM, Solomkin JS, Mayumi T, Gomi H, Pitt HA, Gouma DJ, Garden OJ, Buchler MW, Kiriyama S, Kimura Y, Tsuyuguchi T, Itoi T, Yoshida M, Miura F, Yamashita Y, Okamoto K, Gabata T, Hata J, Higuchi R, Windsor JA, Bornman PC, Fan ST, Singh H, Santibanes E de, Kusachi S, Murata A, Chen XP, Jagannath P, Lee S, Padbury R, Chen MF, Revision CTG. 2012. New diagnostic criteria and severity assessment of acute cholecystitis in revised Tokyo Guidelines. J Hepato-bil-pan Sci 19:578–85.

95. Kiriyama S, Takada T, Strasberg SM, Solomkin JS, Mayumi T, Pitt HA, Gouma DJ, Garden OJ, Buchler MW, Yokoe M, Kimura Y, Tsuyuguchi T, Itoi T, Yoshida M, Miura F, Yamashita Y, Okamoto K, Gabata T, Hata J, Higuchi R, Windsor JA, Bornman PC, Fan ST, Singh H, Santibanes E de, Gomi H, Kusachi S, Murata A, Chen XP, Jagannath P, Lee S, Padbury R, Chen MF, Dervenis C, Chan AC, Supe AN, Liau KH, Kim MH, Kim SW, Revision CTG. 2013. TG13 guidelines for diagnosis and severity grading of acute cholangitis (with videos). J Hepato-bil-pan Sci 20:24–34.

96. Kommedal O, Simmon K, Karaca D, Langeland N, Wiker HG. 2012. Dual priming oligonucleotides for broad-range amplification of the bacterial 16S rRNA gene directly from human clinical specimens. J Clin Microbiol 50:1289–94.

97. Illumina. 16S Metagenomic Sequencing Library Preparation : Preparing 16S Ribosomal RNA Gene Amplicons for the Illumina MiSeq System. 2013.

98. Stebner A, Ensser A, Geissdorfer W, Bozhkov Y, Lang R. 2020. Molecular diagnosis of polymicrobial brain abscesses with 16S-rDNA-based next-generation sequencing. Clin Microbiol Infec 27:76–82.

99. Tremblay J, Yergeau E. 2019. Systematic processing of ribosomal RNA gene amplicon sequencing data. Gigascience 8.

100. Franzen O, Hu J, Bao X, Itzkowitz SH, Peter I, Bashir A. 2015. Improved OTU-picking using long-read 16S rRNA gene amplicon sequencing and generic hierarchical clustering. Microbiome 3:43.

101. Bokulich NA, Subramanian S, Faith JJ, Gevers D, Gordon JI, Knight R, Mills DA, Caporaso JG. 2013. Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing. Nat Methods 10:57–9.

102. Erb-Downward JR, Falkowski NR, D'Souza JC, McCloskey LM, McDonald RA, Brown CA, Shedden K, Dickson RP, Freeman CM, Stringer KA, Foxman B, Huffnagle GB, Curtis JL, Adar SD. 2020. Critical Relevance of Stochastic Effects on Low-Bacterial-Biomass 16S rRNA Gene Analysis. Mbio 11:e00258-20.

103. Dyrhovden R, Ovrebo KK, Nordahl MV, Nygaard RM, Ulvestad E, Kommedal O. 2019. Bacteria and fungi in acute cholecystitis. A prospective study comparing next generation sequencing to culture. J Infection 80:16–23.

104. Gupta VK, Paul S, Dutta C. 2017. Geography, Ethnicity or Subsistence-Specific Variations in Human Microbiome Composition and Diversity. Front Microbiol 8:1162.

105. Glassman SI, Martiny JBH. 2018. Broadscale Ecological Patterns Are Robust to Use of Exact Sequence Variants versus Operational Taxonomic Units. mSphere 3.

106. Johnson JS, Spakowicz DJ, Hong BY, Petersen LM, Demkowicz P, Chen L, Leopold SR, Hanson BM, Agresta HO, Gerstein M, Sodergren E, Weinstock GM. 2019. Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. Nat Commun 10:5029.

107. Rizal NSM, Neoh H, Ramli R, Periyasamy P@ RAK, Hanafiah A, Samat MNA, Tan TL, Wong KK, Nathan S, Chieng S, Saw SH, Khor BY. 2020. Advantages and Limitations of 16S rRNA Next-Generation Sequencing for Pathogen Identification in the Diagnostic Microbiology Laboratory: Perspectives from a Middle-Income Country. Diagnostics 10:816.

108. Culbreath K, Melanson S, Gale J, Baker J, Li F, Saebo O, Kommedal O, Contreras D, Garner OB, Yang S. 2019. Validation and Retrospective Clinical Evaluation of a

Quantitative 16S rRNA Gene Metagenomic Sequencing Assay for Bacterial Pathogen Detection in Body Fluids. J Mol Diagnostics 21:913–923.

109. Miller MB, Atrzadeh F, Burnham CA, Cavalieri S, Dunn J, Jones S, Mathews C, McNult P, Meduri J, Newhouse C, Newton D, Oberholzer M, Osiecki J, Pedersen D, Sweeney N, Whitfield N, Campos J, Clinical ASM, Microbiology CPH, the ASMCC. 2019. Clinical Utility of Advanced Microbiology Testing Tools. J Clin Microbiol 57.

110. Ramanan P, Bryson AL, Binnicker MJ, Pritt BS, Patel R. 2017. Syndromic Panel-Based Testing in Clinical Microbiology. Clin Microbiol Rev 31.

111. Schlaberg R, Simmon KE, Fisher MA. 2012. A Systematic Approach for Discovering Novel, Clinically Relevant Bacteria. Emerg Infect Dis 18:422–430.

112. Simner PJ, Miller S, Carroll KC. 2018. Understanding the Promises and Hurdles of Metagenomic Next-Generation Sequencing as a Diagnostic Tool for Infectious Diseases. Clin Infect Dis Official Publ Infect Dis Soc Am 66:778–788.

113. Marshall B, Warren JR. 1984. UNIDENTIFIED CURVED BACILLI IN THE STOMACH OF PATIENTS WITH GASTRITIS AND PEPTIC ULCERATION. Lancet 323:1311–1315.

114. McDade JE, Shepard CC, Fraser DW, Tsai TR, Redus MA, Dowdle WR. 1977. Legionnaires' Disease — Isolation of a Bacterium and Demonstration of Its Role in Other Respiratory Disease. New Engl J Medicine 297:1197–1203.

115. Miller JM, Binnicker MJ, Campbell S, Carroll KC, Chapin KC, Gilligan PH, Gonzalez MD, Jerris RC, Kehl SC, Patel R, Pritt BS, Richter SS, Robinson-Dunn B, Schwartzman JD, Snyder JW, S. 3rd Telford, Theel ES, B. JrT R, Weinstein MP, Yao JD. 2018. A Guide to Utilization of the Microbiology Laboratory for Diagnosis of Infectious Diseases: 2018 Update by the Infectious Diseases Society of America and the American Society for Microbiology. Clin Infect Dis 67:e1–e94.

116. Menzies SM, Rahman NM, Wrightson JM, Davies HE, Shorten R, Gillespie SH, Davies CWH, Maskell NA, Jeffrey AA, Lee YCG, Davies RJO. 2011. Blood culture bottle culture of pleural fluid in pleural infection. Thorax 66:658.

117. Lancaster DP, Friedman DF, Chiotos K, Sullivan KV. 2015. Blood Volume Required for Detection of Low Levels and Ultralow Levels of Organisms Responsible for Neonatal Bacteremia by Use of Bactec Peds Plus/F, Plus Aerobic/F Medium, and the BD Bactec FX System: an In Vitro Study. J Clin Microbiol 53:3609–3613.

118. Rodriguez-R LM, Konstantinidis KT. 2014. Estimating coverage in metagenomic data sets and why it matters. Isme J 8:2349–2351.

119. Ma Z (Sam). 2020. Estimating the Optimum Coverage and Quality of Amplicon Sequencing With Taylor's Power Law Extensions. Frontiers Bioeng Biotechnology 8:372.

120. Kerkhof LJ. 2021. Is Oxford Nanopore sequencing ready for analyzing complex microbiomes? Fems Microbiol Ecol 97:fiab001.

121. Bradshaw DJ, Marsh PD, Watson GK, Allison C. 1998. Role of Fusobacterium nucleatum and Coaggregation in Anaerobe Survival in Planktonic and Biofilm Oral Microbial Communities during Aeration. Infect Immun 66:4729–4732.

122. Bradshaw DJ, Marsh PD, Watson GK, Allison C. 1997. Oral anaerobes cannot survive oxygen stress without interacting with faculTative/aerobic species as a microbial commmunity. Lett Appl Microbiol 25:385–387.

123. Kolenbrander PE, London J. 1993. Adhere today, here tomorrow: oral bacterial adherence. J Bacteriol 175:3247–3252.

Appendices

Paper I, including all supplementary documents

Paper II, including all supplementary documents

Paper III, including all supplementary documents except Supplementary Table S2, S4, S5 and S6. These tables are OTU tables with a format that is not compatible with a printed version of the thesis. However, these supplementary tables were primary published to ensure full transparency, but they are not necessary for the understanding of the rest of the article.

I

Clinical Microbiology and Infection 25 (2019) 981–986

Contents lists available at ScienceDirect

Clinical Microbiology and Infection

journal homepage: www.clinicalmicrobiologyandinfection.com

Original article

The bacterial aetiology of pleural empyema. A descriptive and comparative metagenomic study

R. Dyrhovden [1, *], R.M. Nygaard [1], R. Patel [2, 3], E. Ulvestad [1, 4], Ø. Kommedal [1]

[1] Department of Microbiology, Haukeland University Hospital, Bergen, Norway
[2] Division of Clinical Microbiology, Department of Laboratory Medicine and Pathology, Mayo Clinic, Rochester MN, USA
[3] Division of Infectious Diseases, Department of Medicine, Mayo Clinic, Rochester MN, USA
[4] Department of Clinical Science, University of Bergen, Bergen, Norway

ARTICLE INFO

Article history:
Received 28 August 2018
Received in revised form
15 November 2018
Accepted 30 November 2018
Available online 21 December 2018

Editor: S. J. Cutler

Keywords:
16S rRNA
Aetiology
Brain abscess
Metagenomics
Pleural empyema
RpoB

ABSTRACT

Objectives: The view of pleural empyema as a complication of bacterial pneumonia is changing because many patients lack evidence of underlying pneumonia. To further our understanding of pathophysiological mechanisms, we conducted in-depth microbiological characterization of empyemas in clinically well-characterized patients and investigated observed microbial parallels between pleural empyemas and brain abscesses.

Methods: Culture-positive and/or 16S rRNA gene PCR-positive pleural fluids were analysed using massive parallel sequencing of the 16S rRNA and *rpoB* genes. Clinical details were evaluated by medical record review. Comparative analysis with brain abscesses was performed using metagenomic data from a national Norwegian study.

Results: Sixty-four individuals with empyema were included. Thirty-seven had a well-defined microbial aetiology, while 27, all of whom had community-acquired infections, did not. In the latter subset, *Fusobacterium nucleatum* and/or *Streptococcus intermedius* was detected in 26 patients, of which 18 had additional facultative and/or anaerobic species in various combinations. For this group, there was 65.5% species overlap with brain abscesses; predisposing factors included dental infection, minor chest trauma, chronic obstructive pulmonary disease, drug abuse, alcoholism and diabetes mellitus. Altogether, massive parallel sequencing yielded 385 bacterial detections, whereas culture detected 38 (10%) and 16S rRNA gene PCR/Sanger-based sequencing detected 87 (23%).

Conclusions: A subgroup of pleural empyema appears to be caused by a set of bacteria not normally considered to be involved in pneumonia. Such empyemas appear to have a similar microbial profile to oral/sinus-derived brain abscesses, supporting spread from the oral cavity, potentially haematogenously. We suggest reserving the term 'primary empyema' for these infections. **R. Dyrhovden, Clin Microbiol Infect 2019;25:981**

Introduction

Empyema, defined as the presence of bacteria or pus in the pleural cavity, is a serious infection with high morbidity and mortality (15%–20%) and increasing incidence [1]. Predisposing conditions include bacterial pneumonia, surgery, trauma, oesophageal perforation, thoracentesis, subdiaphragmatic infection, spontaneous pneumothorax and septicaemia [2]. Traditionally, the predominant cause has been assumed to be bacterial pneumonia in which bacteria breach the visceral pleura to establish an infected parapneumonic effusion [1,2]. This concept has recently been challenged. Differences between the typical bacteriology of pneumonia and empyema have been highlighted, with many patients with empyema having no evidence of an underlying pneumonia [1,3,4]. Such observations suggest that pleural empyema and pneumonia should, in some cases, be considered separate conditions [1,3–5], and—furthermore—that the mechanisms and sources of bacterial invasion of the pleural cavity are poorly understood [1,4,5].

* Corresponding author. R. Dyrhovden, Department of Microbiology, Haukeland University Hospital, Jonas Lies vei 65, 5021 Bergen, Norway.
E-mail address: ruben.dyrhovden@helse-bergen.no (R. Dyrhovden).

https://doi.org/10.1016/j.cmi.2018.11.030

In 2014, we conducted a nationwide Norwegian study using targeted 16S rRNA gene-based metagenomics on brain abscesses [6]. Three bacterial species (*Streptococcus intermedius*, *Fusobacterium nucleatum* and *Aggregatibacter aphrophilus*) were either sole pathogens or a dominant part of abscesses with an assumed oral or sinus origin. This led us to hypothesize that these three microbes may be involved in initial establishment of these infections, with the additional species detected representing later colonizers.

In our routine clinical practice, we more recently observed that many pleural empyema samples have a microbiome similar to that of brain abscesses. This suggested a new hypothesis that a small group of highly specialized bacteria in the oral cavity may, under appropriate conditions, spread to and establish purulent infections in highly oxygenated organs, including the brain and lung.

The aim of the present investigation was to conduct a thorough bacterial characterization of pleural empyemas from clinically well-characterized patients to further our understanding of pathophysiological mechanisms. This endeavour included a comparison between pleural empyemas and brain abscesses.

Materials and methods

We conducted a retrospective study at the Department of Microbiology, Haukeland University Hospital, Bergen, Norway. The study was approved by the regional ethical committee (2017/1095).

Study definition of pleural empyema

Pleural empyema was defined as the presence of bacteria in pleural fluid by Gram stain, culture or 16S rRNA gene PCR [1]. Since previous studies have not included detection of bacterial DNA in their definitions, all patients were also evaluated according to the traditional criteria for pleural empyema, defined as a positive Gram stain or bacterial culture of pleural fluid *or* macroscopic purulent pleural fluid *or* pleural pH < 7.2 combined with clinical evidence of infection [1,3,7].

Clinical samples

The Department covers a population of ~500 000 people and also receives occasional samples from other regional hospitals. Upon clinical suspicion of infection, the laboratory routinely performs partial amplification and Sanger-based sequencing of a portion of the bacterial 16S rRNA genes directly from pleural fluid (see Supplementary material, Document S1). For the present investigation, we searched the laboratory information system for culture-positive and/or 16S rRNA gene PCR-positive pleural fluid samples obtained from patients >17 years of age, admitted from January 2016 through to December 2017. Remnant extracted DNA from these samples was collected from a Biobank for targeted metagenomic analysis. Samples from 64 unique patients were eligible for inclusion. Eighteen patients with culture-positive empyemas could not be included because 16S rRNA gene PCR had not been performed and no extracted DNA was available (see Supplementary material, Table S1). Pleural fluids from 11 patients with a low suspicion of infection, negative cultures and negative 16S rRNA gene PCR were included as a negative patient control group.

Clinical definition of pneumonia

Pneumonia was defined by fulfilment of the following criteria: (i) at least two of the following: fever, cough, sputum production and/or chest pain; and (ii) radiographic evidence of pneumonia, as interpreted by a radiologist [7].

Massive parallel sequencing of partial 16S rRNA and rpoB genes

Massive parallel sequencing was performed using the MiSeq system (Illumina, Redwood City, CA, USA). Some clinically important bacterial families and genera display too few variations in the 16S rRNA gene to reliably identify them to the species level [8]. Therefore, for selected groups of bacteria, we supplemented the analysis with massive parallel sequencing of parts of *rpoB* [9], which is more discriminatory than the 16S rRNA gene.

A detailed protocol for sequencing of all targets, including primers, description of negative controls, management of background DNA and sequence data analysis can be found in the Supplementary material (Document S1). For novel undescribed species we used the provisional HMT-taxonomy [10].

Statistical analysis

Statistical analysis was performed using SPSS 25 (IBM Corp). Differences between subgroups were analysed with Pearson's chi-squared test. Microbial β-diversity analysis for the empyemas, including unweighted and weighted UniFrac metrics [11,12], was performed in QIIME2. Comparison between a subset of the data from this study and previous data from the Norwegian brain abscess study [6], was undertaken using UniFrac and Principal Coordinate Analysis in QIIME2. EMPeror [13] was used to visualize Principal Coordinate Analysis plots. Venn-diagram analysis was performed using the web tool http://bioinformatics.psb.ugent.be/webtools/Venn/. For the Venn-diagram analysis, we included species that were found in at least two cases in one of the two groups.

Results

Description of study population

The mean age of the 64 patients was 62 years (median 68, range 18–93). Fifty (78%) patients were male. Twenty-one (32%) had no history of smoking, while 24 (38%) were current smokers and 19 (30%) were former smokers. The most frequent chronic diseases were hypertension (15 patients, 23%) and chronic obstructive pulmonary disease (11 patients, 17%). Nineteen patients (30%) had no chronic disease. Sixty (94%) patients fulfilled the traditional criteria [1,3,7] for pleural empyema. The four patients who did not (Patients 5, 28, 47 and 62) included two postoperative infections, one *Francisella tularensis* infection and one infection of poorly described aetiology.

Technical sequencing data

After removal of short reads (<250 base pairs), small clusters (<20 reads) and chimeras, the mean number of valid reads was 75 426 per sample (range 19 892 to 311 160, median 65 286).

Microbiological findings and correlations with clinical data

Out of 385 bacterial detections made by massive parallel sequencing, culture detected only 38 (10%) and Sanger-based 16S rRNA gene sequencing detected 87 (22.5%) (see Supplementary material, Table S2). Thirty-nine (61%) of the 16S rRNA gene PCR-positive samples were culture-negative. None of the 11 samples

in the negative patient control group contained bacterial DNA beyond that found in the negative controls.

By 16S rRNA and *rpoB* gene massive parallel sequencing, bacteria from 183 species were identified, of which 157 could be assigned to the species level, 13 to a species group level, and 8 to a genus level. Five Operational taxonomic units yielded no significant match with published sequences. *rpoB* sequencing was performed for 44 samples (69%) and provided identification at a higher taxonomic level than 16S rRNA gene sequencing for 25 species (see Supplementary material, Table S3).

Thirty-seven patients had a well-defined aetiology underlying their pleural fluid infection (Table 1). Among these were 19 patients with monomicrobial infections caused by *Streptococcus pneumoniae* (8), *Staphylococcus aureus* (5), *Pseudomonas aeruginosa* (2), *Streptococcus pyogenes* (1), *Escherichia coli* (1), *Klebsiella pneumoniae* (1) or *Francisella tularensis* (1). *Streptococcus pneumoniae* was identified exclusively in monomicrobial samples, including seven cases of community-acquired pneumonia (CAP) and one case of postoperative empyema after lung resection. All identifications of *Staphylococcus aureus* and *Streptococcus pyogenes* were also made in monomicrobial samples, except for one patient with CAP who had a mixed infection with the two species. Empyemas resulting from surgical complications or spontaneous rupture of the oesophagus presented the highest microbial diversity. An overview of clinical data and identified microbes in each of these 37 individuals is provided in the Supplementary material (Table S4).

For the remaining 27 patients, the aetiology was less obvious. The Supplementary material (Table S5) summarizes the clinical history, computed tomography findings and identified bacteria in these individuals. All 27 infections were community-acquired. Several potential predisposing factors were identified. Six patients reported a minor blunt chest trauma before the onset of symptoms (such as blunt violence or fall accidents) and five of six patients in whom dental status was addressed had poor dental health. Other lung-related pathology among patients included chronic obstructive pulmonary disease (4) and computed tomography-diagnosed lung abscess (3). Only 12 out of 27 fulfilled the diagnostic criteria for CAP and all patients were negative for CAP-associated bacteria. Other possible predisposing comorbidities were drug abuse (4), alcoholism (3) and diabetes mellitus (3). Table 2 provides a comparison of demographics and clinical features between the 27 patients with empyema with poorly described aetiology versus the ten patients with classic post-pneumonia empyema.

In both the weighted and unweighted UniFrac analyses, most of the 27 samples with uncertain aetiology clustered in neighbouring branches, indicating significant similarities in microbial patterns (Fig. 1). Though they contained many of the species found in oesophageal ruptures and postoperative infections, they were distinguished from these by having lower microbial diversity and higher inter-sample consistency, both quantitatively and qualitatively. In total, they harboured 54 different species, of which 33 were identified in single patients. Among the 21 species found in multiple samples, *Streptococcus intermedius* and *Fusobacterium nucleatum* stood out both as being the most frequent bacteria detected on a patient-level and as being among the most abundant. Twenty-six (96%) of the 27 samples contained either *Streptococcus intermedius* ($n = 9$) or *F. nucleatum* ($n = 10$), or a combination of the two ($n = 7$). *Streptococcus intermedius* was the only detected species in eight samples, whereas *F. nucleatum* was detected alone in two. The single sample without either *Streptococcus intermedius* or *F. nucleatum* was a monomicrobial infection caused by *A. aphrophilus* in a patient with pulmonary sarcoidosis.

The microbial patterns observed in the empyema cases of uncertain aetiology had similarities with oral-type bacterial brain abscesses; Fig. 2 is a Venn diagram analysis for both infection types showing 19 (65.5%) shared species. The most frequent common species were *F. nucleatum*, *Streptococcus intermedius*, *Parvimonas micra* and *Eubacterium brachy*, all present in more than 30% of samples from both infection types. The most notable difference was observed for *A. aphrophilus*, found in 40% of brain abscesses but only in a single empyema (3.7%). In a comparative analysis of the two specimen types, Principal Coordinates Analysis plots based on UniFrac metrics showed no clear clustering into separate groups (Fig. 3).

Table 1
Patients with well-defined aetiologies of empyema

Postoperative[a]	12
Community-acquired pneumonia[b]	10
Metastatic cancer affecting the lung	4
Sepsis	4
Spontaneous rupture of the oesophagus	2
Hospital-acquired pneumonia	2
Lemierre syndrome	1
Francisella tularensis pneumonia	1
Post-traumatic	1
Sum	37

[a] Includes five cases occurring after upper gastrointestinal tract surgery, four cases after heart and lung surgery, and three cases after abdominal surgery.
[b] Fulfilment of diagnostic criteria for community-acquired pneumonia (CAP) and identification of bacteria known to cause CAP (*Streptococcus pneumoniae* (7), *Streptococcus pyogenes* (1), *Pseudomonas aeruginosa* (1), *Staphylococcus aureus* and *S. pyogenes* (1)).

Table 2
Demographic and clinical characteristics of empyemas with poorly described aetiology (PDE) versus those occurring after classic community-acquired pneumonia[a] (CAP)

	PDE	CAP	p-value[b]
Number of patients	27	10	
Male, *n*	24	5	0.01
Mean age, years (SD)	56 (18)	65 (18)	0.17
Smoker, *n*	11	4	0.97
Fever, *n*	21	8	0.88
Chest pain, *n*	25	5	0.003
Dyspnoea, *n*	24	9	0.92
Cough, *n*	18	8	0.43
Purulent sputum, *n*	8	3	0.98
Mean CRP[c] (SD)	230 (107)	313 (83)	0.03
Mean leucocytes[d] (SD)	19.3 (8.2)	21.0 (11.8)	0.60
Purulent pleural fluid, *n*	19	9	0.21
Mean pH of pleural fluid (SD)	7.0 (0.6)	7.1 (0.3)	0.69
Mean glucose[e] (SD)	2.2 (2.8)	1.0 (2.2)	0.41
Intensive care, *n*[f]	2	2	0.27
Death, *n*[g]	0	1	0.10

[a] Fulfilment of diagnostic criteria for CAP and identification of bacteria known to cause CAP: *Streptococcus pneumoniae*, *Staphylococcus aureus*, *Pseudomonas aeruginosa* and *Streptococcus pyogenes*.
[b] Pearson's chi-squared test for categorical variables. Students *t*-test for continuous variables.
[c] C-reactive protein (mg/L) at admission.
[d] Leucocytes (10^9/L) at admission.
[e] Mean glucose level (mmol/L) in pleural fluid.
[f] Admitted to intensive care unit during the hospital stay.
[g] Death during the hospital stay.

Discussion

In this study, we provide a clinical and microbiological characterization of 64 patients with empyema. To the best of our knowledge, this is the first investigation in which such a large

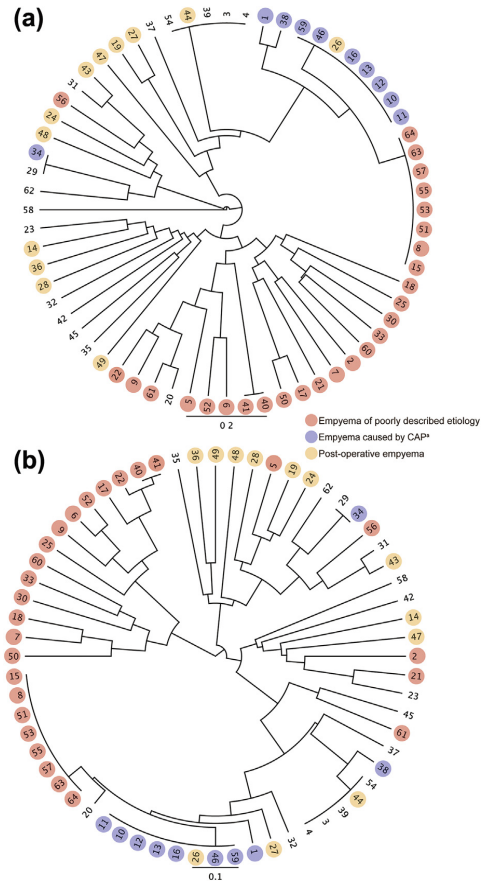


Fig. 1. UniFrac-analysis of empyemas. Phylogenetic tree of unweighted (a) and weighted (b) UniFrac-analysis of all included pleural empyemas. Each number represents one sample. CAP, fulfilment of diagnostic criteria for community-acquired pneumonia (CAP) and identification of bacteria known to cause CAP: *Streptococcus pneumoniae*, *Staphylococcus aureus*, *Pseudomonas aeruginosa* and *Streptococcus pyogenes*.

sample of pleural empyemas has been characterized using a targeted metagenomics analysis.

We identified a subgroup of 27 patients with community-acquired infections with unclear explanation as to how the bacteria had reached the pleural cavity. All empyemas harbored oral bacteria not normally associated with pneumonia, and shared distinct microbial patterns overlapping with those previously described for oral/sinus-type brain abscesses. Samples from 26 of 27 patients contained either *Streptococcus intermedius* or *F. nucleatum* or a combination of the two, and these were also, together with *A. aphrophilus*, the only bacteria detected in monomicrobial infections. We therefore hypothesize that *F. nucleatum* and *Streptococcus intermedius* are possible key pathogens for establishing these empyemas, much like previously suggested for bacterial brain abscesses [6]. *Aggregatibacter aphrophilus*, found in a single monomicrobial empyema, does not seem to be very common in pleural empyemas, though it was found in 40% of brain abscesses.

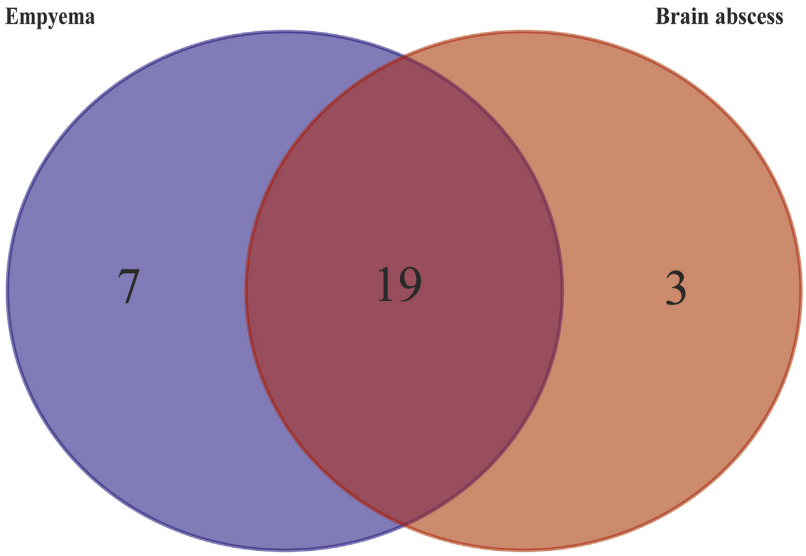
Fusobacterium nucleatum was found in all empyemas harbouring strict anaerobic bacteria, except for one containing *Fusobacterium gonidiaformans*. Although traditionally thought of as a strict anaerobe, *F. nucleatum* is a moderate anaerobe with capacity for oxygen adaption [14,15]. In periodontitis research, it is considered a key organism in the transition between early facultative colonizers and later obligate anaerobes [16].

In patients with pneumonia, pathogens may overcome pulmonary defence mechanisms and reach the pleural cavity by direct transpleural spread from the respiratory alveoli. The capacity of *Streptococcus pneumoniae* to do this is well-described [1], but for *Streptococcus intermedius* and *F. nucleatum*, we have found no evidence of a similar ability. For anaerobic bacteria this mode of dissemination would be halted by the high oxygen tension in the respiratory tract. When examining the clinical characteristics for the 27 patients with empyemas of uncertain pathogenesis, only 12 had possible pneumonia. However, since diagnostic criteria for CAP are based on clinical features with low diagnostic specificity and sensitivity [7,17] and display considerable overlap with the most common signs and symptoms of pleural empyema, we believe the actual number of pneumonias may have been even lower than reported [7].

For brain abscesses, an important route of infection is dissemination of bacteria via the haematogenous route from primary sinusitis or an odontogenic focus [18–20]. Bacteria are thought to reach the brain parenchyma by transit trough valveless emissary and diploic veins that drain through the skull bone and into the venous system of the brain [19]. This mode of transmission is supported by the observation that abscesses formed by oral pathogens are the dominant type of abscess in all regions of the brain, not just in the frontal lobe [6]. The lung parenchyma and the visceral pleura are therefore natural sites along the same route of infection. Infected venous blood follows the venous draining system to the right ventricle of the heart and is pumped into the pulmonary arteries, ending up in the capillary network of alveoli and parts of the visceral pleura. Indeed, several studies and case reports describe the simultaneous occurrence of brain abscess and lung abscess or pleural empyema [21–24], both in general and for *F. nucleatum* and *Streptococcus intermedius*, specifically.

We therefore suggest, based on the findings in this study, that facultative and anaerobic oral bacteria, able to spread via deoxygenated venous blood to establish purulent infections in brain tissue, are also capable of reaching and establishing pyogenic infections in the lung parenchyma or pleural cavity, and that right-sided haematogenous spread is a plausible route of infection in oral-type bacterial pleural empyema.

We found a remarkably high involvement of males in oral-type bacterial empyema that significantly differed from the even gender distribution of classic post-pneumonia empyema (Table 2). The same uneven gender distribution has been observed in other studies on pleural empyema, and specifically among those caused by the *Streptococcus anginosus* group [25–27]. Odontogenic infections have been identified as a potential risk factor for pleural empyema [27] and one explanation of the male predominance might be a higher frequency of serious odontogenic infections in men [28]. Except for dental caries, we found the most common comorbidities to be hypertension, chronic obstructive pulmonary disease, injection drug use, alcoholism and diabetes mellitus (see Supplementary material, Table S5). This correlates with findings from previous studies [27,29]. Male gender, odontogenic infections and injection drug use are also definite risk factors for brain abscesses [19,20]. Another finding that might represent a hitherto unappreciated predisposing factor, is that six patients had a history of a minor blunt chest trauma just before symptom onset. In brain



Species found only in empyema:
Catonella morbi (3)
Gemella bergeriae (2)
Peptostreptococcaceae sp. HMT 369(2)
Treponema maltophilum (2)
Dialister pneumosintes (2)
Mycoplasma salivarium (2)
Prevotella buccae (2)

Species found in both infections:
Fusobacterium nucleatum (17/18)
Streptococcus intermedius (16/19)
Parvimonas micra (11/17)
Eubacterium brachy (8/9)
Campylobacter rectus (5/4)
Prevotella oris (4/3)
Streptococcus constellatus (4/2)
Porphyromonas endodontalis (4/1)
Actinomyces meyeri (3/12)
Prevotella pleuritidis (3/3)
Campylobacter gracilis (2/7)
Eikenella corrodens (2/6)
Prevotella sp. HMT 317 (2/3)
Alloprevotella tannerae (2/1)
Aggregatibacter aphrophilus (1/10)
Eubacterium yurii (1/2)
Filifactor alocis (1/2)
Gemella morbillorum (1/2)
Tannerella forsythia (1/2)

Species found only in brain abscess:
Actinomyces georgiae (2)
Anaeroglobus geminatus (2)
Capnocytophaga sp. HMT 338 (2)

Fig. 2. Venn diagram analysis comparing species found in 27 empyemas of poorly described aetiology and 25 brain abscesses with assumed oral/sinus origin. The numbers in parenthesis are the total number of species found in each group. In the middle column, the first number in parenthesis represents empyemas and the second brain abscesses.

abscesses one of the main predisposing pathogenic factors is the presence of ischaemic or devitalized brain tissue occurring after incidents such as trauma or cerebrovascular accidents [20]. Blunt chest trauma may cause damage to the visceral pleura or adjacent lung parenchyma, leading to the formation of a poorer oxygenated *locus minores resistentiae* (e.g. a haematoma or atelectasis) facilitating colonization by blood-borne oral microbes.

The retrospective design is a weakness of our study; 18 culture-positive samples were lost due to lack of sample availability (see Supplementary material, Table S1). Information about dental status was available from only six patients, and information about preceding minor trauma was not systematically collected. Another limitation is that this is a case study from one particular area of the world. The oral microbiome varies between geographical areas [30] and this may influence the bacterial composition of pleural empyema as well. The findings and hypotheses proposed in this article should clearly be challenged in future, prospective studies.

We have shown that a large subgroup of community-acquired pleural empyemas is caused by a limited set of oral bacteria not normally involved in pneumonia. We provide microbiological, anatomical and epidemiological arguments to support that these pleural empyemas and oral/sinus-derived brain abscesses might be two sides of the same coin sharing microbial composition, haematogenous routes of infection and risk factors. We suggest that the term 'primary empyema' should be reserved for this type of infection to distinguish it from pleural empyema secondary to other lung conditions, including classic post-pneumonia empyema and post-operative empyema. Our finding also suggest that traditional culture-based methods and even Sanger-based 16S rRNA gene PCR/sequencing may be insufficient in characterizing the microbial spectrum of primary empyemas.

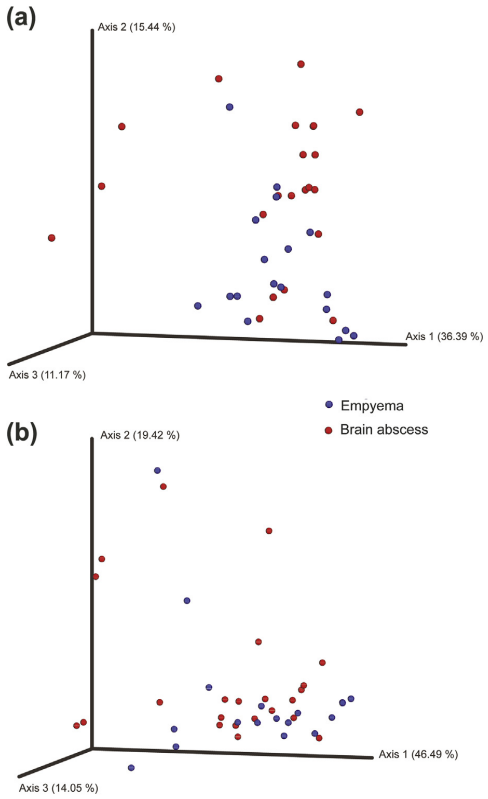


Fig. 3. Comparative UniFrac-analysis of empyemas and brain abscesses. Principal Coordinates Analysis of unweighted (a) and weighted (b) UniFrac-analysis of the 27 empyemas of poorly described aetiology and 25 oral/sinus-derived brain abscesses. The figure shows no clear clustering into separate groups. Each dot represents one sample.

Funding

This work was supported by the Western Norway Regional Health Authority's research funding (grant number 912206) and by the Department of Clinical Microbiology, Haukeland University Hospital.

Transparency declaration

Dr. Patel reports grants from CD Diagnostics, BioFire, Curetis, Merck, Contrafect, Hutchison Biofilm Medical Solutions, Accelerate Diagnostics, Allergan, and The Medicines Company. Dr. Patel is or has been a consultant to Curetis, Specific Technologies, Selux Dx, GenMark Diagnostics, PathoQuest, Heraeus Medical, and Qvella; monies are paid to Mayo Clinic. In addition, Dr. Patel has a patent on *Bordetella pertussis/parapertussis* PCR issued, a patent on a device/method for sonication with royalties paid by Samsung to Mayo Clinic, and a patent on an anti-biofilm substance issued. Dr. Patel receives travel reimbursement from ASM and IDSA and an editor's stipend from ASM and IDSA, and honoraria from the NBME, Up-to-Date and the Infectious Diseases Board Review Course.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cmi.2018.11.030.

References

[1] Corcoran JP, Wrightson JM, Belcher E, DeCamp MM, Feller-Kopman D, Rahman NM. Pleural infection: past, present, and future directions. Lancet Respir Med 2015;3:563–77.
[2] Bryant RE, Salmon CJ. Pleural empyema. Clin Infect Dis 1996;22:747–62. quiz 63–4.
[3] Maskell NA, Batt S, Hedley EL, Davies CW, Gillespie SH, Davies RJ. The bacteriology of pleural infection by genetic and standard methods and its mortality significance. Am J Respir Crit Care Med 2006;174:817–23.
[4] Lisboa T, Waterer GW, Lee YC. Pleural infection: changing bacteriology and its implications. Respirology 2011;16:598–603.
[5] Tobin CL, Lee YC. Pleural infection: what we need to know but don't. Curr Opin Pulm Med 2012;18:321–5.
[6] Kommedal O, Wilhelmsen MT, Skrede S, Meisal R, Jakovljev A, Gaustad P, et al. Massive parallel sequencing provides new perspectives on bacterial brain abscesses. J Clin Microbiol 2014;52:1990–7.
[7] Ahmed RA, Marrie TJ, Huang JQ. Thoracic empyema in patients with community-acquired pneumonia. Am J Med 2006;119:877–83.
[8] Janda JM, Abbott SL. 16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils, and pitfalls. J Clin Microbiol 2007;45:2761–4.
[9] Mollet C, Drancourt M, Raoult D. rpoB sequence analysis as a novel basis for bacterial identification. Mol Microbiol 1997;26:1005–11.
[10] Dewhirst FE, Chen T, Izard J, Paster BJ, Tanner AC, Yu WH, et al. The human oral microbiome. J Bacteriol 2010;192:5002–17.
[11] Lozupone C, Knight R. UniFrac: a new phylogenetic method for comparing microbial communities. Appl Environ Microbiol 2005;71:8228–35.
[12] Lozupone CA, Knight R. Global patterns in bacterial diversity. Proc Natl Acad Sci U S A 2007;104:11436–40.
[13] Vazquez-Baeza Y, Pirrung M, Gonzalez A, Knight R. EMPeror: a tool for visualizing high-throughput microbial community data. Gigascience 2013;2:16.
[14] Silva VL, Diniz CG, Cara DC, Santos SG, Nicoli JR, Carvalho MA, et al. Enhanced pathogenicity of *Fusobacterium nucleatum* adapted to oxidative stress. Microb Pathog 2005;39:131–8.
[15] Gursoy UK, Pollanen M, Kononen E, Uitto VJ. Biofilm formation enhances the oxygen tolerance and invasiveness of *Fusobacterium nucleatum* in an oral mucosa culture model. J Periodontol 2010;81:1084–91.
[16] Kolenbrander PE, London J. Adhere today, here tomorrow: oral bacterial adherence. J Bacteriol 1993;175:3247–52.
[17] Niederman MS. Guidelines for the management of community-acquired pneumonia. Current recommendations and antibiotic selection issues. Med Clin North Am 2001;85:1493–509.
[18] Brook I. Microbiology and treatment of brain abscess. J Clin Neurosci 2017;38:8–12.
[19] Mathisen GE, Johnson JP. Brain abscess. Clin Infect Dis 1997;25:763–79. quiz 80–1.
[20] Brook I. Brain abscess and other focal pyogenic infections of the central nervous system. In: Cohen JP, William G, Opal SM, editors. Infectious diseases. 4th ed. Amsterdam: Elsevier Ltd.; 2017. 198–207.e1.
[21] Takayanagi N, Kagiyama N, Ishiguro T, Tokunaga D, Sugita Y. Etiology and outcome of community-acquired lung abscess. Respiration 2010;80:98–105.
[22] Russell W, Taylor W, Ray G, Gravil J, Davidson S. Not just a 'simple stroke'. Acute Med 2011;10:35–7.
[23] Albrecht P, Stettner M, Husseini L, Macht S, Jander S, Mackenzie C, et al. An emboligenic pulmonary abscess leading to ischemic stroke and secondary brain abscess. BMC Neurol 2012;12:133.
[24] Trabue C, Pearman R, Doering T. Pyogenic brain and lung abscesses due to *Streptococcus intermedius*. J Gen Intern Med 2014;29:407.
[25] Shinzato T, Uema H, Inadome J, Shimoji K, Kusano N, Fukuhara H, et al. Bacteriological and clinical studies in 23 cases of thoracic empyema—the role of oral streptococci and anaerobes. Nihon Kyobu Shikkan Gakkai Zasshi 1993;31:486–91.
[26] Jerng JS, Hsueh PR, Teng LJ, Lee LN, Yang PC, Luh KT. Empyema thoracis and lung abscess caused by viridans streptococci. Am J Respir Crit Care Med 1997;156:1508–14.
[27] Kobashi Y, Mouri K, Yagi S, Obase Y, Oka M. Clinical analysis of cases of empyema due to *Streptococcus milleri* group. Jpn J Infect Dis 2008;61:484–6.
[28] Cachovan G, Phark JH, Schon G, Pohlenz P, Platzer U. Odontogenic infections: an 8-year epidemiologic analysis in a dental emergency outpatient care unit. Acta Odontol Scand 2013;71:518–24.
[29] Nielsen J, Meyer CN, Rosenlund S. Outcome and clinical characteristics in pleural empyema: a retrospective study. Scand J Infect Dis 2011;43:430–5.
[30] Gupta VK, Paul S, Dutta C. Geography, ethnicity or subsistence-specific variations in human microbiome composition and diversity. Front Microbiol 2017;8:1162.

SUPPLEMENTARY MATERIAL

Supplementary Document S1

Sample preparation

DNA-extraction included mechanical disruption of bacterial cells using the SeptiFast Lysis kit and a MagNA Lyser apparatus followed by DNA extraction and purification on a MagNA Pure compact automated extractor (Roche, Mannheim, Germany), as described previously [1].

Routine Sanger-based 16S rRNA gene PCR/sequencing directly from clinical samples

Sanger-based 16S rRNA gene PCR/sequencing had been performed on all study samples as part of routine clinical practice using a modified version of a previously described protocol [2]. The modifications included 5-end improvements of the primers (16S_DPO_Short-F: 5'-AGAGTTTGATCMTGGCTCAIIIIIAACGCT-3' (no LNA-bases) and 16S_DPO_Short-R 5'-CGGCTGCTGGCAIIIAITTRGC-3') and a concomitant reduction of annealing temperature from 64 to 60°C. Mixed electropherograms were interpreted using RipSeq mixed software (Pathogenomix, Santa Cruz, CA) [1].

Massive parallel sequencing of partial 16S rRNA and *rpoB* genes

For both the 16S rRNA and *rpoB* genes, the Illumina protocol for 16S library preparation [3] was used, with a few modifications applied to the first round PCR (amplicon PCR, pages 6-7 in the original protocol). 16S rRNA gene primers are listed at the end of Document S1. They bind to the same area as the original primers in the Illumina protocol, targeting the 16S V3 and V4 regions, but were modified to better suit the human microbial spectrum. For example, the original "T" in position 3 from the 3-end of the reverse primer was replaced with a "K" (T/G) to avoid a mismatch with *Cutibacterium acnes*.

For *rpoB* analysis, two novel broad-range primer pairs were designed one targeting clinically important species of *Enterobacteriaceae* (RpoB_Ent), and the other targeting enterococci, staphylococci and streptococci *(*RpoB_ESS*)*. The primers are listed at the end of Document S1. Massive parallel sequencing of the *rpoB* targets was performed only on samples where analysis of the 16S rRNA gene revealed bacteria that could not be accurately identified using this approach. Except for the primers, the protocol for *rpoB* sequencing was identical to the modified protocol used for the 16S rRNA gene. The TaKaRa-enzyme used in the first PCR was necessary to obtain efficient amplification with the *rpoB* primers.

For PCR, we used a LightCycler 480 real-time PCR machine (Roche). The PCR mixture consisted of 12.5 µl SYBR Premix Ex Taq (TaKaRa, city, Japan), 8.5 µl PCR-grade water, 1 µl of each primer (from a 10 µM solution, giving a final concentration of 0.4 µM in the PCR) and 2 µl template. For the RpoB_ESS solution we used 1 µl each of the forward primers and 1.5 µl of the reverse primer and reduced the amount of water correspondingly to 7 µl. The PCR thermal profile included an initial polymerase activation step of 30 s at 95°C followed by 45 cycles of 20 s at 95°C (melting), 30 s at 60°C (annealing), and 30 s at 72°C (extension). After completion, PCR products were spun out of the SmartCycler reaction tubes and used directly in downstream steps. The SYBR-green real-time reaction with melting-curve analysis eliminated the need for gel-based verification of the PCR-product.

Negative controls

For each clinical sample, a negative extraction control consisting of lysis buffer and PCR-grade water was processed in parallel. Before sequencing, negative extraction controls were mixed into three pools. Each pool of negative controls was sequenced in duplicate. A positive extraction control consisting of *Salmonella bongori* suspended in PCR-grade water was also included and sequenced in duplicate.

Sequence data analysis

After Illumina-sequencing, barcode separated FASTQ-files were processed individually using RipSeq NGS software [4] (Pathogenomix, Santa Cruz, CA). For all targets, reads shorter than 250 base pairs were removed before *de novo* clustering into operational taxonomic units (OTUs) using a similarity threshold of 99%. OTUs containing less than 50 sequences were rejected. For unambiguous 16S rRNA gene-based species-level identification, we used a cutoff of ≥99.3% homology with a high-quality reference sequence combined with a minimum distance of >0.8% to the next alternative species. OTUs obtaining species-level homology but with an insufficient distance to the next species were assigned to a species-group or listed as a slashed result. Homology between 97.0 and 99.3% qualified for genus-level identification.

The two *rpoB* PCRs target separate parts of the *rpoB* gene with different levels of difference between species. Based on observed clustering patterns and intra-species variation among GenBank references, we adopted the following rules: For RpoB_ESS we used a cut-off of ≥97.0% homology with a high-quality reference combined with a minimum distance to next species of >2.0% for species-level identification. OTUs with a similarity ≥97.0% but with a distance to next species of ≤2.0% were assigned to a group of species (i.e., group-level identification). The RpoB_Ent amplicon displayed smaller inter-species distances and lower intra-species variation. In addition, the taxonomy for important groups of *Enterobacteriaceae* is still evolving and, as such, it was challenging to identify up-to-date high-quality references. For this target, we therefore applied more stringent rules, requiring ≥99% homology with a high-quality reference combined with a minimum distance of >1.5% to the next alternative species for a specie-level assignment. OTUs with a similarity >99.0% but with a distance to next species of ≤1.5% was assigned to a group of species (i.e., group-level identification)

Background DNA

Based on sequencing results from the pooled negative controls, we defined a list of the ten most abundant contaminating bacteria. Since samples had to be divided into two groups and sequenced in separate runs to have enough space on the sequencing chip, we got two lists of contaminating bacteria. Starting with the most abundant, the top ten contaminating bacteria of the first run were *Cutibacterium acnes, Aquabacterium citratiphilum, Ralstonia pickettii, Staphylococcus capitis / Staphylococcus caprae / Staphylococcus epidermidis, Pseudomonas fluorescens, Phenylobacterium koreense, Hydrotalea flava, Pseudomonas extremorientalis / Pseudomonas fluorescens / Pseudomonas poae,* Unknown bacteria 6 and *Aquabacterium* spp. The top ten contaminating bacteria of the second run were *C. acnes, A. citratiphilum, Paracoccus chinensis / Paracoccus marinus, P. koreense, R. pickettii, S. capitis / S. caprae,* Unknown bacteria 7, *H. flava, Staphylococcus saccharolyticus* and *Afipia broomeae / Afipia felis.*

For the dominant contaminants, there was high consistency across all negative controls. The top-ten contaminants were used as indicators for the level of background DNA in clinical samples. Bacteria appearing in higher concentrations than any of the top ten background bacteria were accepted as valid identifications. Bacteria present in concentrations between 10 and 100% of the most abundant background bacterium identified in a sample were also accepted as valid identifications, if they were absent from the negative controls. Bacteria present in concentrations below 10% of the most abundant background species in a sample were rejected as invalid.

Primers with adapter sequences. Sequences of the target specific portions in capital letters

Name	Sequence	Position[a]
16S-F[b]	tcgtcggcagcgtcagatgtgtataagagacagCCTACGGGNGGCWGCAG	340-356
16S-R[b]	gtctcgtgggctcggagatgtgtataagagacagGACTACCAGGGTATCTAAKCC	784-803
RpoB_Ent-F	tcgtcggcagcgtcagatgtgtataagagacagGAAGGTCCRAAYATCGGTCT	1693-1712
RpoB_Ent-R	gtctcgtgggctcggagatgtgtataagagacagTGCATGTTCGCACCCAT	2041-2057
RpoB_ESS-F1	tcgtcggcagcgtcagatgtgtataagagacagGCRACAGCRTGTATYCCRTTC	1861-1881
RpoB_ESS-F2	tcgtcggcagcgtcagatgtgtataagagacagGCDACAGCATGTATTCCWTTC	1861-1881
RpoB_ESS-R	gtctcgtgggctcggagatgtgtataagagacagGTTRTAMCCNTCCCAWGTCAT	2287-2307

[a] Positions for 16S based on *Escherichia coli* (GenBank accession J01859). Positions for RpoB_ESS based on *Staphylococcus aureus* [*rpoB* coding sequence (CDS); GenBank accession X64172]. Positions for RpoB_Ent based on *Escherichia coli* [*rpoB* coding sequence (CDS); GenBank accession V00340].

[b] Abbreviations: F = forward primer. R = reverse primer.

References

[1] Kommedal O, Kvello K, Skjastad R, Langeland N, Wiker HG. Direct 16S rRNA gene sequencing from clinical specimens, with special focus on polybacterial samples and interpretation of mixed DNA chromatograms. J Clin Microbiol. 2009;47(11):3562-8.

[2] Kommedal O, Simmon K, Karaca D, Langeland N, Wiker HG. Dual priming oligonucleotides for broad-range amplification of the bacterial 16S rRNA gene directly from human clinical specimens. J Clin Microbiol. 2012;50(4):1289-94.

[3] Illumina. 16S Metagenomic Sequencing Library Preparation : Preparing 16S Ribosomal RNA Gene Amplicons for the Illumina MiSeq System. 2013. https://support.illumina.com/downloads/16s_metagenomic_sequencing_library_preparation.html.

[4] Kommedal O, Wilhelmsen MT, Skrede S, Meisal R, Jakovljev A, Gaustad P, et al. Massive parallel sequencing provides new perspectives on bacterial brain abscesses. J Clin Microbiol. 2014;52(6):1990-7.

Supplementary Table S1: Samples with growth of bacteria, not included in current study because of the lack of availability of residual specimen.

Patient	Microbes detected by growth	Diagnosis according to medical record
1	*Enterococcus faecalis*	Endocarditis and empyema
2	*Pseudomonas aeruginosa*	Pseudomonas empyema
3	*Enterococcus faecium*	Postoperative empyema; thoracotomy
4	*Staphylococcus aureus*	Pancreatic cancer with lung metastasis; empyema.
5	*E. faecalis,* coagulase negative *Staphylococcus* species*, Prevotella bivia, Prevotella disiens*	Endocarditis, operated, postoperative empyema.
6	*Streptococcus intermedius*	Empyema
7	*S. intermedius*	Empyema
8	*Prevotella vulgaris, E. faecium*	Acute pancreatitis; empyema
9	*Streptococcus mitis/oralis*	Pneumonia and empyema
10	*S. intermedius*	Pneumonia and empyema
11	*S. aureus*	Pneumonia and empyema
12	*S. intermedius*	Empyema
13	*Escherichia coli, E. faecalis*	Postoperative empyema; abdominal surgery
14	*S. aureus*	Liver failure; sepsis; empyema.
15	*E. faecalis, E. faecium*	Postoperative empyema; abdominal surgery.
16	*E. coli*	Pneumonia and empyema
17	*Streptococcus anginosus*	Gastric cancer with lung metastasis and empyema
18	*Streptococcus salivarius, Streptococcus parasanguinis, Haemophilus parainfluenzae, Rothia mucilangilosa, Prevotella melaninogenica*	Esophageal rupture

Supplementary Table S2: Comparison between parallel sequencing, Sanger-sequencing and culture for all patients

Pleural empyema with a poorly described etiology (n = 27)			
ID	**16S rRNA and *rpoB* gene sequencing - sorted by decreasing abundance**	**16S rRNA Sanger sequencing (V1-V3)**	**Culture**
02	*Prevotella oris* *Campylobacter gracilis* *Streptococcus intermediusa*[a] *Eikenella corrodens* *Fusobacterium nucleatum* Unknown bacterium 1 *Eikenella* sp. (MDA2346-4) *Alloprevotella tannerae* *Mycoplasma salivarium* *Eubacterium brachy* *Tannerella forsythia* *Parvimonas micra*	*P. oris* *C. gracilis* *S. intermedius*	Negative
05	*Escherichia coli*[a] *S. intermedius*[a] *F. nucleatum* *Klebsiella michiganensis*[a] *Klebsiella variicola*[a] *Parvimonas micra* *Clostridium perfringens*	*Escherichia/Cronobacter/Citrobacter* sp. *S. intermedius*	Negative
06	*F. nucleatum* *S. intermedius*[a] *E. brachy* *Actinomyces meyeri*	*F. nucleatum* *S. intermedius*	*S. intermedius*
07	*Fusobacterium gonidiaformans*[b] *P. oris* *Leptotrichia amnionii* *P. micra* *Streptococcus anginosus*[a] *Gemella asaccharolytica* *Gemella bergeriae* *E. corrodens* *Alloprevotella* sp. HMT 308	*F. gonidiaformans* *P. oris*	*P. micra* *S. anginosus* *E. corrodens*
08	*S. intermedius*[a]	*S. intermedius*	*S. intermedius*
09	*F. nucleatum* *S. intermedius*[a] *E. brachy* *Eubacterium yurii* *A. meyeri*	*F. nucleatum* *S. intermedius*	Negative
15	*S. intermedius*[a]	*S. intermedius*	Negative
17	*F. nucleatum* *P. micra* *Prevotella pleuritidis*	Negative	*S. constellatus*[c]
18	*P. micra* *F. nucleatum* *Sneathia* sp. *Prevotella denticola* *P. oris* *Dialister* sp. *A. tannerae* *Prevotella baroniae* *Prevotella buccae* *Colibacter massiliensis* *Dialister pneumosintes* *Porphyromonas asaccharolytica/uenonis* *Streptococcus mitis* group *Campylobacter rectus/showae*	CTC[d]	*P. denticola* *P. baroniae* *P. buccae* *S. constellatus* *Actinomyces* sp. *K. pneumoniae*[c] *S. aureus*[c]

	Dialister invisus *Catonella morbi* *Peptostreptococcaceae* (XI)(G-4) sp. HMT 369 *Streptococcus constellatus* *G. bergeriae* *Actinomyces funkei* *Alloprevotella rava*		
21	*S. constellatus* *Prevotella* sp. (HMT 314) *F. nucleatum* *P. buccae* *Peptostreptococcus stomatis*	*S. constellatus* *Prevotella sp.* *F. nucleatum*	Negative
22	*F. nucleatum* *S. intermedius*[a]	*F. nucleatum*	Negative
25	*F. nucleatum* *Prevotella conseptionensis* *E. brachy* *Porphyromonas endodontalis* *E. yurii* *C. morbi* *C. rectus/showae*	*Fusobacterium sp.* *P. endodontalis*	Negative
30	*P. micra* *E. brachy* *P. endodontalis* *F. nucleatum* *Treponema maltophilum* *Mycoplasma faucium* *C. rectus/showae* *Peptostreptococcaceae* (XI)(G-4) sp. (HMT 369)	*P. micra* *E. brachy*	Negative
33	*Mycoplasma salivarium* *P. micra* *F. nucleatum* *C. rectus/showae* *P. oris* *P. endodontalis* *Prevotella nigrescens* *E. brachy* *D. pneumosintes* *S. constellatus* *Prevotella conceptionensis* *Treponema lecithinolyticum* *Eubacterium saphenum* *Catonella* sp. (oral clone FL073) *G. morbillorum* *Mogibacterium timidum*	CTC[d]	Negative
40	*F. nucleatum*	*F. nucleatum*	Negative
41	*F. nucleatum*	*F. nucleatum*	Negative
50	*P. pleuritidis* *F. nucleatum* *P. micra* *E. brachy*	*P. pleuritidis* *F. nucleatum* *P. micra*	*F. nucleatum* *P. micra*
51	*S. intermedius*[a]	*S. intermedius*	*S. intermedius*
52	*F. nucleatum* *S. intermedius*[a] *A. meyeri* *C. gracilis*	*F. nucleatum* *S. intermedius*	Negative
53	*S. intermedius*[a]	*S. intermedius*	*S. intermedius*
55	*S. intermedius*[a]	*S. intermedius*	Negative
56	*Aggregatibacter aphrophilus*	*A. aphrophilus*	Negative
57	*S. intermedius*[a]	*S. intermedius*	*S. intermedius*
60	*C. rectus/showae* *F. nucleatum* *P. pleuritidis* *P. endodontalis* *Eubacterium nodatum* *S. constellatus*	CTC[d]	Negative

	T. maltophilum *P. micra* *E. brachy* *A. meyeri*		
61	*S. intermedius*[a] *F. nucleatum* *Filifactor alocis* *P. micra*	*S. intermedius* *F. nucleatum*	*S. intermedius*
63'	*S. intermedius*[a]	*S. intermedius*	Negative
64	*S. intermedius*[a]	*S. intermedius*	Negative
Post-operative infections (n=12)			
ID	**16S rRNA and *rpoB* gene sequencing - sorted by decreasing abundance**	**16S rRNA Sanger sequencing (V1-V3)**	
14	*Prevotella oris* *Bifidobacterium dentium* *Lactobacillus fermentum* *Veillonella parvula* group[e] *Rothia dentocariosa* *Prevotella sp.* *Streptococcus intermedius*[a] *Streptococcus oralis*[a] *Alloscardovia omnicolens* *Streptococcus mitis*[a] *Enterococcus faecalis* *Haemophilus parainfluenzae* *Mycoplasma salivarium* *Rothia mucilaginosa* *Streptococcus parasanguinis*[a] *Campylobacter curvus* *Porphyromonas* sp. (HMT 279)	*P. oris* *V. parvula* group	Negative
19	*Raoultella terrigena*[a] *Streptococcus* sp.[f] *Haemophilus haemolyticus/influenzae*	*Enterobacter/Raoultella* sp. *Streptococcus cristatus* *H. haemolyticus*	Negative
24	*Serratia marcescens/nematodiphila/urealyticum*[a] *(Staphylococcus epidermidis)*[g]	*S. marcescens* *S. epidermidis*	*S. marcescens* *S. epidermidis*
26	*Streptococcus pneumoniae*[a]	*S. mitis/oralis* group	Negative
27	*S. parasanguinis* *H. parainfluenzae* *Streptococcus cristatus*	CTC[a]	Negative
28	*Enterobacter aerogenes*[a] *Anaeroglobus geminatus* *Campylobacter gracilis* *Megasphaera micronuciformis* *Prevotella* sp. *Veillonella* sp. *Parvimonas micra* *M. salivarium* *Campylobacter* sp. *Prevotella denticola* *Porphyromonas* sp. (HMT 279) *P. oris* *S. mitis*[a] *Prevotella* sp. (HMT 314) *Dialister pneumosintes* *Selenomonas artemidis* *Streptococcus anginosus*[a] *Prevotella* sp. (HMT 313) *Alloprevotella tannerae* *Slackia exigua* *Veillonella atypica* *Dialister invisus* *Prevotella baroniae* *Selenomonas* sp. *C. curvus* *Selenomonas noxia* *S. oralis*[a]	*E. aerogenes/Raoultella planticola* *A. geminatus*	Negative

36	*M. salivarium* *S. parasanguinis*[a] *H. parainfluenzae* *H. haemolyticus/influenzae* *S. mitis*[a] *Klebsiella variicola*[a] *Gemella haemolysans/sanguinis* *Prevotella veroralis* *Morganella morganii* *Pseudomonas aeruginosa* *Fusobacterium nucleatum* *D. pneumosintes* *P. oris* *Granulicatella adiacens* *Prevotella pallens* *Oribacterium sinus* *Streptococcus gordonii*[a] *H. haemolyticus* *S. intermedius*[a] *Campylobacter concisus* *Prevotella salivae* *Atopobium rimae* *Stomatobaculum longum* *Prevotella melaninogenica* *D. invisus* *Lachnoanaerobaculum umaense* *Alloprevotella* sp. (HMT 914) Unknown bacterium 2 *Actinomyces odontolyticus* *S. oralis*[a] *Streptococcus infantis*[a] *S. cristatus*[a]	*Haemophilus* sp. *(Pasteurella pneumotropica*[h]*)*	*P. aeruginosa*
43	*Klebsiella pneumoniae/quasipneumoniae*[a]	*K. pneumoniae/quasipneumoniae*	Negative
44	*Staphylococcus aureus*[a]	*S. aureus*	Negative
47	*Bacteroides fragilis* *Enterococcus faecium*[a] *Bacteroides xylanisolvens*	*B. fragilis* *E. faecium/hirae*	Negative
48	*(Cutibacterium acnes)*[g] *Escherichia coli*[a]	*C. acnes* *E. coli*	*C. acnes*

49	*F. nucleatum* *Enterococcus avium*[a] *Citrobacter amalonaticus*[a] *Klebsiella michiganensis*[a] *Clostridium bolteae/clostridioforme* *Alistipes onderdonkii* *K. pneumoniae/quasipneumoniae*[a] *Bacteroides stercoris* *Eggerthella lenta* *Bacteroides uniformis* *Bacteroides dorei* *Parabacteroides goldsteinii* *Barnesiella sp.* *E. coli*[a] *E. faecalis* *Parabacteroides distasonis* *Bacteroides caccae* *Bacteroides ovatus* *Veillonella parvula/dispar* *Parabacteroides distasonis* *Enterobacter cloacae/hormaechei*[a] *Ruminococcus gnavus* *Clostridium subterminale* *C. gracilis* *Clostridium glycolicum* Unknown bacterium 4 Unknown bacterium 5 *H. parainfluenzae* *Clostridium perfringens* *Citrobacter/Kluyvera* speices *B. xylanisolvens* *Clostridium nexile*	Negative	*K. pneumoniae*
Community acquired pneumonia with typical pneumonia-associated bacteria (n=10)			
ID	**16S rRNA and *rpoB* gene sequencing - sorted by decreasing abundance**	**16S rRNA Sanger sequencing (V1-V3)**	
1	*Streptococcus pyogenes*	*S. pyogenes*	*S. pyogenes*
10	*S. pneumoniae*[a]	*S. mitis/oralis* group	*S. pneumoniae*
11	*S. pneumoniae*[a]	*S. mitis/oralis* group	Negative
12	*S. pneumoniae*[a]	*S. mitis/oralis* group	*S. pneumoniae*
13	*S. pneumoniae*[a]	*S. mitis/oralis* group	*S. pneumoniae*
16	*S. pneumoniae*[a]	*S. mitis/oralis* group	Negative
34	*P. aeruginosa*	*P. aeruginosa*	*P. aeruginosa*
38	*S. aureus*[a] *S. pyogenes*	*S. aureus* *S. pyogenes*	Negative
46	*S. pneumoniae*[a]	*S. mitis/oralis* group	Negative
59	*S. pneumoniae*[a]	*S. mitis/oralis* group	Negative
Metastatic cancer affecting the lung (n=4)			
ID	**16S rRNA and *rpoB* gene sequencing - sorted by decreasing abundance**	**16S rRNA Sanger sequencing (V1-V3)**	
20	*S. intermedius*[a] *P. micra* *Fusobacterium periodonticum*	*S. intermedius* *P. micra*	Negative

35	*Finegoldia magna* *H. haemolyticus/influenzae* *Peptostreptococcus stomatis* *P. oris* *Peptoniphilus* sp. *Atopobium parvulum* *Prevotella* sp. *P. micra* *Prevotella* sp. (HMT 315) *Peptoniphilus lacrimalis* *Tannerella forsythia* *Prevotella* sp. (HMT 475) *P. baroniae* *Olsenella uli* *Gemella morbillorum* *Eikenella corrodens* *Streptococcus constellatus*	*F. magna* *H. influenzae*	*F. magna* *Peptoniphilus sp.* *Prevotella sp.*
39	*S. aureus*[a]	*S. aureus*	Negative
42	*P. denticola* *Streptococcus* sp.[f] *A. parvulum* *A. odontolyticus* *S. oralis*[a]	*P. denticola*	Negative
Sepsis (n=4)			
ID	**16S rRNA and *rpoB* gene sequencing - sorted by decreasing abundance**	**16S rRNA Sanger sequencing (V1-V3)**	
31	*E. coli**	*E. coli*	Negative
37	*Lactobacillus casei/rhamnosus/paracesei* *E. faecium*[a]	*L. rhamnosus* *E. faecium*	Negative
54	*S. aureus*[a]	*S. aureus*	*S. aureus*
58	*Prevotella timonensis* *Anaerococcus obesiensis* *Bacteroidetes* (G-7) sp. (HMT 911) *Anaerococcus lactolyticus* *Porphyromonas uenonis* *Prevotella disiens* *Anaerococcus murdochii/degenerii* *Peptoniphilus massiliensis* *Prevotella bergensis* *P. lacrimalis* *F. magna* *Peptoniphilus* sp. *Bacteroidales* sp. (vaginal isolate KA00251) *Peptoniphilus coxii* *Peptostreptococcus* sp. *Lachnospiraceae* sp. (vaginal isolate KA00044)	CTC[d]	Negative
Spontaneous rupture of esophagus (n=2)			
ID	**16S rRNA and *rpoB* gene sequencing - sorted by decreasing abundance**	**16S rRNA Sanger sequencing (V1-V3)**	
23	*Streptococcus vestibularis*[a] *Prevotella* sp. (HMT 306) *S. mitis*[a] *P. melaninogenica* *Prevotella* sp. (HMT 313) *Streptococcus salivarius*[a] *G. haemolysans/sanguinis* *S. parasanguinis*[a] *Prevotella histicola* *H. parainfluenzae* *R. mucilaginosa* *V. atypica* *Veillonella* sp. (HMT 780) *Alloprevotella* sp. (HMT 308) *Veillonella* sp. *S. oralis*[a] *Aggregatibacter* sp. (HMT 458)	*S. salivarius* group *S. mitis/oralis* group	*S. vestibularis* *Streptococcus mitis/oralis* group *Streptococcus salivarius* group *S. oralis* *R. mucilaginosa*

32	*S. salivarius*[a] *S. mitis/pneumoniae*[a] *G. sanguinis* *Streptococcus* sp. (most similar to *S. oralis*) *S. infantis* *Mycoplasma faucium* *Ruminococcaceae* [G-2] sp. (HMT 085) TM7 (G-1) sp. (HMT 352) *S. mitis*[a] *Clostridiales (F-1)(G-1)* sp. (HMT 093) *S. vestibularis*[a] *S. oralis*[a] *S. gordonii* *S. cristatus*[a] *S. parasanguinis*[a] *Prevotella* sp. (HMT 306) *Streptococcus termophilus* TM7(G-6) sp. (HMT 870) *G. adiacens* *S. timonensis* *L. rhamnosus/casei/paracasei* *Streptococcus* sp. (HMT 056) *Granulicatella elegans* *Ruminococcaceae* [G-1] sp. (HMT 075) *Stomatobaculum* sp. (HMT 097) *C. concisus* *F. nucleatum* *F. periodonticum* *O. sinus* *Bifidobacterium animalis*	*S. salivarius* group *S. mitis/oralis* group	Negative
Hospital acquired pneumonia (n=2)			
ID	**16S rRNA and *rpoB* gene sequencing - sorted by decreasing abundance**	**16S rRNA Sanger sequencing (V1-V3)**	
4	*S. aureus*[a]	*S. aureus*	Negative
29	*P. aeruginosa*	*P. aeruginosa*	*P. aeruginosa*
Lemierre syndrome (n=1)			
ID	**16S rRNA and *rpoB* gene sequencing - sorted by decreasing abundance**	**16S rRNA Sanger sequencing (V1-V3)**	
45	*P. stomatis* *Fusobacterium necrophorum* *S. anginosus* *P. baroniae* *P. oris* *G. morbillorum* *P. micra* *S. constellatus* *Prevotella intermedia* *D. pneumosintes* *Solobacterium moorei* *Filifactor alocis* *Coriobacteriaceae* sp. S9 PR-11 *Bulleidia extructa* Unknown bacterium 3	*P. stomatis* *F. necrophorum* *S. anginosus*	Negative
***Francicella tularensis* pneumonia (n=1)**			
ID	**16S rRNA and *rpoB* gene sequencing - sorted by decreasing abundance**	**16S rRNA Sanger sequencing (V1-V3)**	
62	*Francicella tularensis*	*F. tularensis*	*F. tularensis*
Trauma (n=1)			
ID	**16S rRNA and *rpoB* gene sequencing - sorted by decreasing abundance**	**16S rRNA Sanger sequencing (V1-V3)**	
3	*S. aureus*[a]	*S. aureus*	*S. aureus*

[a] *rpoB* sequencing provided identification at a higher taxonomic level than 16S rRNA gene sequencing.

[b] Not distinguishable from the horse pathogen *Fusobacterium equinum.*

[c] Growth in blood culture. Not detected by sequencing.

[d] Abbreviations: CTC = Mixed chromatogram too complex to allow for interpretation.

[e] *Veillonella parvula/dispar/tobetsuensis/dentocariosa.*

[f] 95.8% match with *Streptoccus cristatus.*

[g] Not a valid identification according to our criteria. *S. epidermidis* and *C. acnes* were also among the ten most abundant microbes in the negative controls.

[h] Considered a false positive from the theoretical deconvolution of a mixed chromatogram using RipSeq mixed software.

Supplementary Table S3: Species identified at a higher taxonomic level with use of partial *rpoB* compared to partial 16S rRNA gene sequencing (V3-V4)

	16S rRNA gene sequencing results	*rpoB* gene sequencing results
1	*Enterococcus avium/raffinosus*	*Enterococcus avium*
2	*Enterococcus durans/faecium/hirae*	*Enterococcus faecium*
3	*Staphylococcus aureus/croceolyticus/petrasii/simiae*	*Staphylococcus aureus*
4	*Streptococcus intermedius/anginosus*	*Streptococcus anginosus*
5	*S. intermedius/anginosus*	*Streptococcus intermedius*
6	*Streptococcus mitis/oralis* group	*Streptococcus cristatus*
7	*S. mitis/oralis* group	*Streptococcus infantis*
8	*S. mitis/oralis* group	*Streptococcus mitis*
9	*S. mitis/oralis* group	*Streptococcus mitis/pneumoniae*[a]
10	*S. mitis/oralis* group	*Streptococcus oralis*
11	*S. mitis/oralis* group	*Streptococcus pneumoniae*
12	*S. mitis/oralis* group	*Streptococcus timonensis*
13	*Streptococcus salivarius* group	*Streptococcus salivarius*
14	*S. salivarius* group	*Streptococcus thermophilus*
15	*S. salivarius* group	*Streptococcus vestibularis*
16	*Streptococcus sanguinis* group	*Streptococcus parasanguinis*
17	*Citrobacter/Enterobacter* species	*Citrobacter amalonaticus*
18	*Raoultella* species/*Klebsiella aerogenes*	*Enterobacter aerogenes*
19	*Enterobacter asburiae/cloacae/hormaechei*	*Enterobacter cloacae/hormaechei*
20	*Escherichia albertii/coli/fergusonii/Shigella* species	*Escherichia coli/Shigella* species
21	*Enterobacter cancerogenus/Klebsiella michiganensis/oxytoca*	*Klebsiella michiganensis*
22	*Klebsiella pneumoniae/variicola/quasipneumoniae*	*Klebsiella variicola*
24	*K. pneumoniae/variicola/quasipneumoniae*	*Klebsiella pneumoniae/quasipneumoniae*
23	*Raoultella* species/*Klebsiella aerogenes*	*Raoultella terrigena*
25	*Serratia marcescens/nematodiphila/Cronobacter dublinensis/Escherichia coli*	*Serratia marcescens /nematodiphila/urealyticum*

[a] Discrimination between S. mitis and S. pneumoniae was possible for all but a single OTU, presumably due to lack of relevant mitis-reference in GenBank

Supplementary Table S4: Samples and results overview for pleural empyema with a well-defined etiology (n=37).

Post-operative infections (n=12)			
ID Sex/Age	**Clinical history**	**16S rRNA and *rpoB* gene sequencing - sorted by decreasing abundance**	**Relevant comorbidities**
14 M/72	Rupture of esophagus after surgical treatment of gastric cancer.	*Prevotella oris* *Bifidobacterium dentium* *Lactobacillus fermentum* *Veillonella parvula* group[a] *Rothia dentocariosa* *Prevotella sp.* *Streptococcus intermedius*[b] *Streptococcus oralis*[b] *Alloscardovia omnicolens* *Streptococcus mitis*[b] *Enterococcus faecalis* *Haemophilus parainfluenzae* *Mycoplasma salivarium* *Rothia mucilaginosa* *Streptococcus parasanguinis*[b] *Campylobacter curvus* *Porphyromonas* sp. (HMT 279)	None
19 F/82	Postoperative infection after cholecystectomy. Biloma and pleural empyema.	*Raoultella terrigena*[b] *Streptococcus* sp.[c] *Haemophilus haemolyticus/influenzae*	Hypertension
24 M/71	Postoperative infection after pulmonary resection.	*Serratia marcescens/nematodiphila/urealyticum*[b] *(Staphylococcus epidermidis)*[e]	PVD[d] Renal failure Hypertension
26 M/74	Postoperative infection after pulmonary resection.	*Streptococcus pneumoniae*[b]	CHD[d] COPD[d] Renal failure
27 M/48	Anastomotic leakage after esophageal surgery	*S. parasanguinis* *H. parainfluenzae* *Streptococcus cristatus*	CHF[d]
28 F/64	Esophageal rupture after surgical treatment of hiatal hernia.	*Enterobacter aerogenes*[b] *Anaeroglobus geminatus* *Campylobacter gracilis* *Megasphaera micronuciformis* *Prevotella* sp. *Veillonella* sp. *Parvimonas micra* *M. salivarium* *Campylobacter* sp. *Prevotella denticola* *Porphyromonas* sp. (HMT 279) *P. oris* *S. mitis*[b] *Prevotella* sp. (HMT 314) *Dialister pneumosintes* *Selenomonas artemidis*	None

		Streptococcus anginosus[b] *Prevotella* sp. (HMT 313) *Alloprevotella tannerae* *Slackia exigua* *Veillonella atypica* *Dialister invisus* *Prevotella baroniae* *Selenomonas* sp. *C. curvus* *Selenomonas noxia* *S. oralis*[b]	
36 M/65	Esophageal rupture after endoscopic removal of esophageal tumor.	*M. salivarium* *S. parasanguinis*[b] *H. parainfluenzae* *H. haemolyticus/influenzae* *S. mitis*[b] *Klebsiella variicola*[b] *Gemella haemolysans/sanguinis* *Prevotella veroralis* *Morganella morganii* *Pseudomonas aeruginosa* *Fusobacterium nucleatum* *D. pneumosintes* *P. oris* *Granulicatella adiacens* *Prevotella pallens* *Oribacterium sinus* *Streptococcus gordonii*[b] *H. haemolyticus* *S. intermedius*[b] *Campylobacter concisus* *Prevotella salivae* *Atopobium rimae* *Stomatobaculum longum* *Prevotella melaninogenica* *D. invisus* *Lachnoanaerobaculum umaense* *Alloprevotella* sp. (HMT 914) Unknown bacterium 2 *Actinomyces odontolyticus* *S. oralis*[b] *Streptococcus infantis*[b] *S. cristatus*[b]	None
43 F/72	Perforation of gastric bowel after surgical treatment of diaphragmatic hernia.	*Klebsiella pneumoniae/quasipneumoniae*[b]	COPD Hypertension
44 M/73	Postoperative infection after bilobectomy.	*Staphylococcus aureus*[b]	None
47 M/58	Postoperative infection after mitral valve surgery	*Bacteroides fragilis* *Enterococcus faecium*[b] *Bacteroides xylanisolvens*	None
48 M/73	Postoperative infection after liver surgery for metastatic colon cancer.	*(Cutibacterium acnes)*[e] *Escherichia coli*[b]	None

49 M/73	Biliary stricture relieved endoscopically. Postoperative pleural empyema.	*F. nucleatum* *Enterococcus avium*[b] *Citrobacter amalonaticus*[b] *Klebsiella michiganensis*[b] *Clostridium bolteae/clostridioforme* *Alistipes onderdonkii* *K. pneumoniae/quasipneumoniae*[b] *Bacteroides stercoris* *Eggerthella lenta* *Bacteroides uniformis* *Bacteroides dorei* *Parabacteroides goldsteinii* *Barnesiella sp.* *E. coli*[b] *E. faecalis* *Parabacteroides distasonis* *Bacteroides caccae* *Bacteroides ovatus* *Veillonella parvula/dispar* *Parabacteroides distasonis* *Enterobacter cloacae/hormaechei*[b] *Ruminococcus gnavus* *Clostridium subterminale* *C. gracilis* *Clostridium glycolicum* Unknown bacterium 4 Unknown bacterium 5 *H. parainfluenzae* *Clostridium perfringens* *Citrobacter/Kluyvera* speices *B. xylanisolvens* *Clostridium nexile*	None
Community acquired pneumonia with typical pneumonia-associated bacteria (n=10)			
ID Sex/Age	**Clinical history**	**16S rRNA and *rpoB* gene sequencing - sorted by decreasing abundance**	**Relevant comorbidities**
1 F/69	Dry cough for 3 weeks. Gradually increasing retrosternal chest pain. No fever. 5 kg weight loss. Admission CT[a] showed pleural effusion and adjacent pneumonia.	*Streptococcus pyogenes*	None
10 M/61	Cough, chest pain, fever and chills for 1 week. Admission CT[a] showed pleural effusion and adjacent pneumonia.	*S. pneumoniae*[b]	None
11 F/70	Cough and left sided chest pain for 1-2 weeks prior to hospitalization. Admission CT[a] showed pleural effusion and adjacent pneumonia.	*S. pneumoniae*[b]	None
12 F/87	Fever, cough, dyspnea and impaired general condition for 3 days prior to hospital admission. Chest x-ray showed consolidations and pleural fluid. Worsening condition despite adequate treatment. Death 9 days after hospitalization.	*S. pneumoniae*[b]	CHF[d] RPD[d]

13 M/30	Cough and impaired general condition for 3 weeks. At hospital admission respiratory failure. Intubated. Admission CT[a] showed pleural effusion and adjacent pneumonia.	*S. pneumoniae*[b]	Asthma Injection drug use
16 F/71	Persistent infection and impaired general condition after completing 3 weeks of antibiotics for CAP. Hospitalized and diagnosed with empyema.	*S. pneumoniae*[b]	Asthma Hypertension
34 M/88	1-2 weeks fever and dyspnea. Hospitalized, diagnosed with and treated for CAP. Exacerbation of symptoms 10 days after starting treatment. CT[a]-diagnosed lung abscess and adjacent pleural empyema.	*P. aeruginosa*	COPD Renal failure Hypertension
38 M/43	A few weeks of retrosternal chest pain. Fever, purulent cough and dyspnea in the week before hospitalization; hospitalized with septic shock and multiorgan failure. Intubated. CT[a]-diagnosed bilateral pneumonia and pleural empyema.	*S. aureus*[b] *S. pyogenes*	None
46 M/57	10 days purulent cough, fever, dyspnea and chest pain prior to hospitalization. Admission CT[a]-diagnosed left-sided pleural empyema and adjacent consolidation in lung parenchyma.	*S. pneumoniae*[b]	DM[d]
59 F/75	Flu-like symptoms for 6 days prior to hospitalization. Treated for CAP when hospitalized. 4 days after hospitalization CT[a] showed large bilateral pneumonia and left-sided empyema.	*S. pneumoniae*[b]	COPD[d] PVD[d]
Metastatic cancer affecting the lung (n=4)			
ID Sex/Age	**Clinical history**	**16S rRNA and *rpoB* gene sequencing - sorted by decreasing abundance**	**Relevant comorbidities**
20 M/80	Metastatic NSCLC[a] with chronic persistent pleural fluid in the right hemithorax which developed into a pleural empyema.	*S. intermedius*[b] *P. micra* *Fusobacterium periodonticum*	CHD[d] Asthma NSCLC
35 M/59	Metastatic lung cancer. After surgery for cerebral metastasis, exacerbation of respiratory symptoms and infection symptoms. CT[a] of thorax showed lung abscesses and pleural empyema.	*Finegoldia magna* *H. haemolyticus/influenzae* *Peptostreptococcus stomatis* *P. oris* *Peptoniphilus* sp. *Atopobium parvulum* *Prevotella* sp. *P. micra* *Prevotella* sp. (HMT 315) *Peptoniphilus lacrimalis* *Tannerella forsythia*	COPD[d] Alcoholism Metastatic lung cancer Hypertension

		Prevotella sp. (HMT 475) *P. baroniae* *Olsenella uli* *Gemella morbillorum* *Eikenella corrodens* *Streptococcus constellatus*	
39 M/50	Metastatic thymoma carcinoma and local growth of tumor into mediastinum and pleura. Chronic persistent pleural fluid which developed into pleural empyema.	*S. aureus*[b]	Thymoma carcinoma
42 M/79	Metastatic NSCLC[a]. Pleural fluid adjacent to lung tumor, developing into empyema.	*P. denticola* *Streptococcus* sp.[c] *A. parvulum* *A. odontolyticus* *S. oralis*[b]	NSCLC[d] DM[d]
Sepsis (n=4)			
ID Sex/Age	**Clinical history**	**16S rRNA and *rpoB* gene sequencing - sorted by decreasing abundance**	**Relevant comorbidities**
31 M/74	Persistent pleural fluid caused by chronic lymphatic leukemia. Hospitalized with *E. coli* sepsis followed by pleural empyema 3 months prior to current event. Readmitted to hospital several times because of persistent symptoms of infection. Current sample taken after follow-up CT[a]-showed persistent right empyema.	*E. coli*[b]	COPD[d] Chronic lymphatic leukemia
37 F/74	Prolonged course of acute pancreatitis. Persistent pleural fluid. 2 months after hospitalization sepsis with *Enterococcus faecium* developing into empyema.	*Lactobacillus casei/rhamnosus/paracesei* *E. faecium*[b]	Hypertension
54 M/74	14 days of impaired general condition. Fever 5 days before hospitalization. Sepsis and multiorgan failure when hospitalized; blood culture positive for *S. aureus*. Probable focus of infection determined to be right leg wound. During hospital stay diagnosed with both empyema and spondylodiscitis.	*S. aureus*[b]	Hypertension
58 F/47	Injection of drugs in the femoral groin leading to infected venous thrombus. Spread of bacteria to the lungs, developing into lung abscesses and adjacent empyema.	*Prevotella timonensis* *Anaerococcus obesiensis* *Bacteroidetes* (G-7) sp. (HMT 911) *Anaerococcus lactolyticus* *Porphyromonas uenonis* *Prevotella disiens* *Anaerococcus murdochii/degenerii* *Peptoniphilus massiliensis* *Prevotella bergensis* *P. lacrimalis* *F. magna*	Injection drug use

		Peptoniphilus sp. *Bacteroidales* sp. (vaginal isolate KA00251) *Peptoniphilus coxii* *Peptostreptococcus* sp. *Lachnospiraceae* sp. (vaginal isolate KA00044)	
Spontaneous rupture of esophagus (n=2)			
ID Sex/Age	**Clinical history**	**16S rRNA and *rpoB* gene sequencing - sorted by decreasing abundance**	**Relevant comorbidities**
23 M/93	Acute illness with hematemesis. Admission CT[a] chest showed pneumothorax and right pleural effusion. Death 2 days after hospital admission.	*Streptococcus vestibularis*[b] *Prevotella* sp. (HMT 306) *S. mitis*[b] *P. melaninogenica* *Prevotella* sp. (HMT 313) *Streptococcus salivarius*[b] *G. haemolysans/sanguinis* *S. parasanguinis*[b] *Prevotella histicola* *H. parainfluenzae* *R. mucilaginosa* *V. atypica* *Veillonella* sp. (HMT 780) *Alloprevotella* sp. (HMT 308) *Veillonella* sp. *S. oralis*[b] *Aggregatibacter* sp. (HMT 458)	CHD[d] Dementia
32 M/58	Acute illness with abdominal pain. Endoscopy showed ulceration and perforation of distal esophagus. Diagnosed with empyema 4 days after hospital admission.	*S. salivarius*[b] *S. mitis/pneumoniae*[b] *G. sanguinis* *Streptococcus* sp. (most similar to *S. oralis*) *S. infantis* *Mycoplasma faucium* *Ruminococcaceae* [G-2] sp. (HMT 085) TM7 (G-1) sp. (HMT 352) *S. mitis*[b] *Clostridiales (F-1)(G-1)* sp. (HMT 093) *S. vestibularis*[b] *S. oralis*[b] *S. gordonii* *S. cristatus*[b] *S. parasanguinis*[b] *Prevotella* sp. (HMT 306) *Streptococcus termophilus* TM7(G-6) sp. (HMT 870) *G. adiacens* *S. timonensis* *L. rhamnosus/casei/paracasei* *Streptococcus* sp. (HMT 056) *Granulicatella elegans* *Ruminococcaceae* [G-1] sp. (HMT 075) *Stomatobaculum* sp. (HMT 097) *C. concisus* *F. nucleatum*	GURS[d] Obesity Hypertension

		F. periodonticum *O. sinus* *Bifidobacterium animalis*	
Hospital acquired pneumonia (n=2)			
ID **Sex/Age**	**Clinical history**	**16S rRNA and *rpoB* gene sequencing - sorted by decreasing abundance**	**Relevant comorbidities**
4 M/69	Hospitalized for CAP. Recurrence of symptoms 13 days after discharge. Readmitted to hospital. X-ray and ultrasound of chest showed pleural fluid which appeared to be purulent when drained.	*S. aureus*[b]	COPD[d] CHD[d] PVD[d]
29 M/61	Admitted to hospital with wound infection and sepsis 14 days after open fracture of humerus while on vacation in Asia. Intubated 4 days after hospitalization. Developed ventilator assisted pneumonia with *P. aeruginosa.* Empyema developed a few days later.	*P. aeruginosa*	Asthma
Lemierre syndrome (n=1)			
ID **Sex/Age**	**Clinical history**	**16S rRNA and *rpoB* gene sequencing - sorted by decreasing abundance**	**Relevant comorbidities**
45 M/38	Lemierre syndrome. Neck abscesses, mediastinitis and pleural empyema starting from dental focus.	*P. stomatis* *Fusobacterium necrophorum* *S. anginosus* *P. baroniae* *P. oris* *G. morbillorum* *P. micra* *S. constellatus* *Prevotella intermedia* *D. pneumosintes* *Solobacterium moorei* *Filifactor alocis* *Coriobacteriaceae* sp. S9 PR-11 *Bulleidia extructa* Unknown bacterium 3	Dental caries/ periodontitis
***Francicella tularensis* pneumonia (n=1)**			
ID **Sex/Age**	**Clinical history**	**16S rRNA and *rpoB* gene sequencing - sorted by decreasing abundance**	**Relevant comorbidities**
62 M/51	Low-grade fever for >6 months. On hospital admission diagnosed with atypical pneumonia and adjacent pleural fluid. Died 11 days after hospitalization.	*Francisella tularensis*	Alcoholism Chronic pancreatitis
Trauma (n=1)			
ID **Sex/Age**	**Clinical history**	**16S rRNA and *rpoB* gene sequencing - sorted by decreasing abundance**	**Relevant comorbidities**

3 F/73	Fell on his right chest 3 weeks prior to hospitalization. Clinical diagnosis of rib fracture. 1 week before hospitalization developed gradually increasing dyspnea and impaired general condition. CT[a] of thorax at hospitalization showed empyema and adjacent rib fracture.	*S. aureus*[b]	None

[a] *Veillonella parvula/dispar/tobetsuensis/dentocariosa.*

[b] *rpoB* sequencing provided identification at a higher taxonomic level than 16S rRNA gene sequencing.

[c] 95.8% match with *Streptococcus cristatus.*

[d] Abbreviations: CT = Computer tomography. CAP = Community acquired pneumonia. COPD = Chronic obstructive pulmonary disease. DM = Diabetes mellitus. CHF = Congestive heart failure. CHD = Coronary heart disease. PVD = Peripheral vascular disease. NSCLC = Non-small cell lung cancer. RPD = Restrictive pulmonary disease. GURS = Gastroesophageal reflux syndrome. CTC = Mixed chromatogram too complex to allow for interpretation.

[e] Not a valid identification according to our criteria. *S. epidermidis* and *C. acnes* was also among the ten most abundant microbes in the negative controls.

Supplementary Table S5: Sample and results overview for pleural empyema of poorly described etiology (n=27)

ID Sex/ Age	Clinical history	Computer tomography (CT)	16S rRNA and *rpoB* gene sequencing sorted by decreasing abundance	Relevant comorbidities/ risk factors
02 M/51	Possible CAP[a]. Cough and intermittent fever for 1 month prior to hospitalization. No effect of pre-hospital penicillin and first-generation cephalosporin.	Empyema and pulmonary consolidation consistent with atelectasis.	*Prevotella oris* *Campylobacter gracilis* *Streptococcus intermedius*[b] *Eikenella corrodens* *Fusobacterium nucleatum* Unknown bacterium 1 *Eikenella* sp. (MDA2346-4) *Alloprevotella tannerae* *Mycoplasma salivarium* *Eubacterium brachy* *Tannerella forsythia* *Parvimonas micra*	None
05 M/70	Dry cough and impaired general condition for 1 week prior to admission. Acute exacerbation on day of admission.	Large empyema with no apparent infection in the lung parenchyma.	*Escherichia coli*[b] *S. intermedius*[b] *F. nucleatum* *Klebsiella michiganensis*[b] *Klebsiella variicola*[b] *Parvimonas micra* *Clostridium perfringens*	Hypertension
06 F/69	Hemoptysis, night sweats and weight loss 3 months prior to admission. Acute exacerbation with respiratory failure 3 days after percutaneous biopsy of possible lung tumor/abscess.	First CT: Possible abscess/ tumor in lung parenchyma. Second CT: Large empyema. Findings on first CT determined to be a lung abscess.	*F. nucleatum* *S. intermedius*[b] *E. brachy* *Actinomyces meyeri*	Injection drug use COPD[a]
07 M/66	Possible CAP[a]. Dyspnea and cough from 12 days prior to hospitalization.	Large empyema and adjacent pulmonary consolidation consistent with pneumonia.	*Fusobacterium gonidiaformans*[c] *P. oris* *Leptotrichia amnionii* *P. micra* *Streptococcus anginosus*[b] *Gemella asaccharolytica* *Gemella bergeriae* *E. corrodens* *Alloprevotella* sp. HMT 308	None
08 M/34	Possible CAP[a]. Fell from bicycle and hit right hemithorax 3 weeks prior to hospital admission. Increasing pain in right hemithorax. Developed fever and impaired general condition prior to admission	Large right empyema and small consolidations in apes of right lung.	*S. intermedius*[b]	Injection drug use Minor blunt trauma
09 M/43	Treated for dental abscess 4 weeks prior to admission. 2 weeks later developed cough, fever and pain in left chest. After	First CT: Consolidation compatible with tumor or abscess in left lung Second CT: rupture of lung abscess	*F. nucleatum* *S. intermedius*[b] *E. brachy* *Eubacterium yurii* *A. meyeri*	Dental abscess

	admission initially treated with ampicillin while awaiting biopsy. 10 days later acute exacerbation, SIRS and respiratory failure.	causing large empyema.		
15 M/81	Possible CAP[a]. Cough, pain in left chest and increasing dyspnea 2 weeks prior to admission. on admission, SIRS and multi-organ failure.	Left empyema. Right-sided pulmonary consolidations consistent with edema and possible parenchymal infection.	*S. intermedius*[b]	None
17 M/43	Dyspnea and acute pain in right chest after heavy lifting 8 days prior. Persistent dull pain followed by acute exacerbation two days before admission. While in hospital also diagnosed with periodontitis and root canal infection.	Large right empyema	*F. nucleatum* *P. micra* *Prevotella pleuritidis*	Asthma Periodontitis/ root canal infection
18 F/41	Possible CAP[a]. Exposed to blunt violence to head and chest 2 weeks prior to hospitalization. Persistent pain in left chest. Cough. At admission impaired general condition and respiratory failure. Noted poor dental health.	Large left empyema. Some ground glass opacification in lung parenchyma on right side.	*P. micra* *F. nucleatum* *Sneathia* sp. *Prevotella denticola* *P. oris* *Dialister* sp. *A. tannerae* *Prevotella baroniae* *Prevotella buccae* *Colibacter massiliensis* *Dialister pneumosintes* *Porphyromonas asaccharolytica/uenonis* *Streptococcus mitis* group *Campylobacter rectus/showae* *Dialister invisus* *Catonella morbi* *Peptostreptococcaceae* (XI) (G-4) sp. HMT 369 *Streptococcus constellatus* *G. bergeriae* *Actinomyces funkei* *Alloprevotella rava*	Injection drug use Poor dental health Minor blunt trauma
21 M/78	Possible CAP[a]. Fall accident and fracture of left arm 3 months previously. Impaired general condition, weight loss and dyspnea up to admission. Acute exacerbation day before admission.	Right empyema and right pulmonary consolidations.	*S. constellatus* *Prevotella* sp. (HMT 314) *F. nucleatum* *P. buccae* *Peptostreptococcus stomatis*	COPD[a] Alcoholism DM[a] Minor blunt trauma
22 M/18	Possible CAP[a]. Fell off a sledge 1-2 weeks prior to	Left empyema and right-sided	*F. nucleatum* *S. intermedius*[b]	Primary ciliary dyskinesia PDR[a]

	hospital admission. Hit his left chest. Impaired general condition and pain in left hemithorax 4 days prior to admission.	pulmonary consolidations.		Minor blunt trauma
25 M/81	Possible CAP[a]. Cough and symptoms of upper respiratory tract infection the last month prior to hospital admission. Acute exacerbation with pain in left chest.	Left empyema and pulmonary consolidations consistent with atelectasis and/or pneumonia.	*F. nucleatum* *Prevotella conseptionensis* *E. brachy* *Porphyromonas endodontalis* *E. yurii* *C. morbi* *C. rectus/showae*	CHF[a] CHD[a] Hypertension
30 M/57	Possible CAP[a]. Purulent cough 1 week prior to hospital admission.	Left empyema. In right lung small (12 mm diameter) area of consolidation.	*P. micra* *E. brachy* *P. endodontalis* *F. nucleatum* *Treponema maltophilum* *Mycoplasma faucium* *C. rectus/showae* *Peptostreptococcaceae* (XI)(G-4) sp. HMT 369	Asthma Hypertension
33 M/49	Possible CAP[a]. Completed treatment for presumed CAP[a]. 8 days before admission. Admitted due to persistent pain in left chest.	Empyema and adjacent pulmonary consolidations consistent with infection.	*Mycoplasma salivarium* *P. micra* *F. nucleatum* *C. rectus/showae* *P. oris* *P. endodontalis* *Prevotella nigrescens* *E. brachy* *D. pneumosintes* *S. constellatus* *Prevotella conceptionensis* *Treponema lecithinolyticum* *Eubacterium saphenum* *Catonella* sp. (oral clone FL073) *G. morbillorum* *Mogibacterium timidum*	Alcoholism
40 M/48	Possible CAP[a]. Pain in right chest 2-3 weeks prior to admission. Acute exacerbation with increasing pain, fever and cough on day of admission.	Large empyema. Ground glass opacification in both lungs.	*F. nucleatum*	Poor dental health Hypertension
41 M/71	Impaired general condition 2 months. Weight loss, dyspnea and pain in right chest. Exacerbation few days before admission; high fever and purulent cough.	Empyema and possible malignant tumor (later diagnosed as adenocarcinoma).	*F. nucleatum*	None
50 M/34	For 6 weeks pain in left chest. Acute exacerbation several days before admission; Impaired general condition, high fever and purulent	Left empyema. Pulmonary consolidations consistent with atelectasis adjacent to empyema.	*P. pleuritidis* *F. nucleatum* *P. micra* *E. brachy*	Dental abscess

	cough. Dental root abscess diagnosed while in hospital.			
51 M/75	Increasing dyspnea and chest pain 2 weeks prior to hospital admission.	Large right empyema with adjacent atelectasis.	*S. intermedius*[b]	None
52 M/44	Pain in right chest, cough and fever 12 days before hospital admission.	Large right empyema with adjacent atelectasis.	*F. nucleatum* *S. intermedius*[b] *A. meyeri* *C. gracilis*	None
53 M/33	Generalized body pain 4 weeks prior to hospital admission. Purulent cough, fever/chills for 2-3 weeks.	Large left empyema with adjacent atelectasis.	*S. intermedius*[b]	None
55 M/73	Pain in right chest, dyspnea, fever and hemoptysis 1 week prior to hospital admission.	Large right empyema with right-sided atelectasis.	*S. intermedius*[b]	COPD[a] DM[a] Hypertension
56 M/70	Acute left sided flank pain and fever the day before admission.	Large left empyema and left lung abscesses.	*Aggregatibacter aphrophilus*	Pulmonary sarcoidosis
57 M/46	Upper respiratory infection and cough for a few days, followed by acute pain in the right chest 3 days before hospital admission.	Large right empyema with adjacent atelectasis.	*S. intermedius*[b]	None
60 M/49	Possible CAP[a]. Frequent falls. 3 days before admission fell down stairs at home, following which he developed pain in the right chest and increasing dyspnea.	Large right empyema with adjacent atelectasis and pulmonary consolidations. Ground glass opacification in left lung.	*C. rectus/showae* *F. nucleatum* *P. pleuritidis* *P. endodontalis* *Eubacterium nodatum* *S. constellatus* *T. maltophilum* *P. micra* *E. brachy* *A. meyeri*	Alcoholism DM[a] Epilepsy Minor blunt trauma
61 F/44	Impaired general condition for 10 days. 2 days before admission fever, dyspnea and left chest pain.	Large left empyema.	*S. intermedius*[b] *F. nucleatum* *Filifactor alocis* *P. micra*	Hypertension
63 M/51	Two recent fall accidents. Pain in right chest after this. Acute exacerbation the night before admission with dyspnea and increasing chest pain.	Right empyema.	*S. intermedius*[b]	COPD[a] Injection drug use Minor blunt trauma
64 M/85	Acute chest pain and dyspnea. Admitted to hospital and drained large	Chest X-ray showed pleural fluid but no pulmonary consolidation.	*S. intermedius*[b]	None

	amount of pleural fluid the same day symptoms started.			

[a]Abbreviations: CAP = Community acquired pneumonia. COPD = Chronic obstructive pulmonary disease. DM = Diabetes mellitus. PDR = Psychomotor development retardation. CHF = Congestive heart failure. CHD = Coronary heart disease. CTC = Mixed chromatogram too complex to allow for interpretation.

[b] *rpoB* sequencing provided identification at a higher taxonomic level than 16S rRNA gene sequencing.

[c] Not distinguishable from the horse pathogen *Fusobacterium equinum.*

II

Journal of Infection 80 (2020) 16–23

Contents lists available at ScienceDirect

Journal of Infection

journal homepage: www.elsevier.com/locate/jinf

Bacteria and fungi in acute cholecystitis. A prospective study comparing next generation sequencing to culture

Ruben Dyrhovden[a,*], Kjell Kåre Øvrebø[b], Magnus Vie Nordahl[c], Randi M. Nygaard[a], Elling Ulvestad[a,d], Øyvind Kommedal[a]

[a] Department of Microbiology, Haukeland University Hospital, Jonas Lies vei 65, 5021 Bergen, Norway
[b] Department of Surgery, Haukeland University Hospital, Bergen, Norway
[c] Department of Surgery, Voss Hospital, Voss, Norway
[d] Department of Clinical Science, University of Bergen, Bergen, Norway

ARTICLE INFO

Article history:
Accepted 27 September 2019
Available online 2 October 2019

Keywords:
Next generation sequencing
16S
rpoB
ITS
Acute cholecystitis
Bile
Massive parallel sequencing

SUMMARY

Objectives: Guidelines for antibiotic treatment of acute cholecystitis are based on studies using culture techniques for microbial identification. Microbial culture has well described limitations and more comprehensive data on the microbial spectrum may support adjustments of these recommendations. We used next generation sequencing to conduct a thorough microbiological characterization of bile-samples from patients with moderate and severe acute cholecystitis.

Methods: We prospectively included patients with moderate and severe acute cholecystitis, undergoing percutaneous or perioperative drainage of the gall bladder. Bile samples were analyzed using both culture and deep sequencing of bacterial 16S rRNA and *rpoB* genes and the fungal ITS2-segment. Clinical details were evaluated by medical record review.

Results: Thirty-six patients with moderate and severe acute cholecystitis were included. Bile from 31 (86%) of these contained bacteria (29) and/or fungi (5) as determined by sequencing. Culture identified only 40 (38%) of the 106 microbes identified by sequencing. In none of the 15 polymicrobial samples did culture detect all present microbes. Frequently identified bacteria often missed by culture included oral streptococci, anaerobic bacteria, enterococci and Enterobacteriaceae other than *Klebsiella* spp. and *Escherichia coli*.

Conclusions: Culture techniques display decreased sensitivity for the microbial diagnostics of acute cholecystitis leaving possible pathogens undetected.

Introduction

Acute cholecystitis is defined as an acute inflammation of the gall bladder. It is one of the most common inpatient diagnoses at surgical departments[1,2] and in more than 90% of patients it arises as complications of cholelithiasis (calculous cholecystitis).[1,2] Bacterial growth in bile is reported in 20% to 70% of patients.[3–8] Bacterial infection is believed to represent a secondary complication and not the initiating event of the disease.[2] Infection is considered an important negative prognostic factor, and antibiotics are included in treatment recommendations for all grades of severity.[4,7,9–11]

Empiric treatment with piperacillin/tazobactam or a cephalosporin +/− metronidazole is recommended for moderate and severe acute cholecystitis irrespective of whether there is growth by culture.[9–11] The microbiological studies constituting the basis for choosing these antibiotic regimens were all performed with conventional culture techniques.[10] For other purulent infections, recent comparisons of microbial detection by culture versus culture-free identification of microbial DNA by next generation sequencing (NGS) have demonstrated that conventional culture detects only a fraction of the bacteria being present.[12,13] The lower sensitivity is particularly pronounced for samples containing anaerobic bacteria and for samples collected after the initiation of antimicrobial therapy.

Incomplete data on the microbial spectrum associated with acute cholecystitis may lead to sub-optimal antibiotic treatment, thus worsening patient outcome. A study from Israel found that

* Corresponding author.
E-mail address: ruben.dyrhovden@helse-bergen.no (R. Dyrhovden).

https://doi.org/10.1016/j.jinf.2019.09.015

discordant antibiotic therapy for acute cholecystitis, in most cases because of a non-susceptible *Enterobacter* spp. or *Enterococcus* spp., resulted in a relative risk for in-hospital death of 6.28 compared to patients who received concordant therapy.[7]

The aim of the present investigation was to use NGS to conduct a thorough microbiological characterization of bile-samples from clinically well-characterized patients with acute cholecystitis. We further sought to compare the results from culture-free NGS with results obtained by conventional microbiological culture and discuss discrepancies from a diagnostic and clinical perspective.

Materials and methods

We conducted a prospective, single-center study at Haukeland University Hospital, Bergen, Norway. The study was approved by the regional ethical committee (2015/65). Written informed consent was obtained from all participants.

Patients

From July 2015 to April 2017, we collected bile samples from 36 patients who underwent treatment with percutaneous (34) or perioperative (2) drainage for acute cholecystitis, defined according to the Tokyo Guideline 2013 (TG13) criteria for a definite diagnosis.[14] Clinical details were evaluated by medical record review. Although debated,[15,16] at Haukeland University Hospital acute mild cholecystitis is treated with observation and/or antibiotics, sometimes followed by delayed cholecystectomy 2–4 months later. For moderate and severe disease percutaneous drainage is the treatment of choice. Consequently, only patients with moderate or severe disease were available for inclusion, and percutaneous drainage was the dominating sampling method. As a patient control group, we included bile samples taken at cholecystectomy from 16 patients with cholelithiasis and no signs of ongoing gallbladder inflammation, operated at Voss Hospital, Voss, Norway.

Sample material, routine diagnostics and DNA-extraction

Bile fluid was aseptically collected during surgery or percutaneous drainage and injected into a sterile tube. All samples were cultured according to the laboratory's guidelines; 10 µl sample material was spread on plates of blood agar, lactose agar, and fastidious anaerobic agar with and without kanamycin and vancomycin. An aliquot of bile was inoculated into brain heart infusion (BHI) as an enrichment procedure. Blood agars and BHIs were incubated in a CO2-enriched atmosphere for 48 h. Lactose agar was incubated for 24 h. Anaerobe agars were incubated in an anaerobe atmosphere for 48 h. Isolates were identified by MALDI-TOF MS Bruker Microflex (Bruker Biotyper, Bremen, Germany.)

DNA was extracted from each sample using a volume of 400 µl bile as described previously.[17] The eluate was stored at −80 °C for later NGS analysis.

Massive parallel sequencing of 16S rRNA, ITS2 and rpoB *genes*

Sequencing of partial bacterial 16S rRNA and the fungal ITS2-segment were performed from all samples. Sequencing of partial *rpoB*-genes were done whenever 16S rRNA sequencing revealed bacteria from the Enterobacteriaceae family or from the *Enterococcus, Streptococcus* or *Staphylococcus* genera that could be identified at a higher taxonomic level by the selected *rpoB*-gene segments.[13] Amplification and sequencing of 16S rRNA- and *rpoB*-genes was performed as described previously using the Illumina MiSeq system (Illumina, Redwood City, CA).[13] For the fungal ITS2-segment we used the primers recommended by Khot et al.[18] and otherwise followed the protocol as described for 16S rRNA.[13] All primers are listed in Supplementary Table S1.

Negative controls

Each clinical sample was processed together with a parallel negative extraction control consisting of lysis buffer and PCR-grade water. Before sequencing, the negative extraction controls were mixed into three pools. A positive extraction control consisting of *Legionella pneumophila* suspended in PCR-grade water was also included and sequenced in the same run.

Sequence data analysis

After Illumina-sequencing, barcode separated FASTQ-files were processed using the RipSeq NGS software[12] (Pathogenomix, Santa Cruz, CA) where sequences were *de novo* clustered into operational taxonomic units (OTUs) using a similarity threshold of 99%. OTUs containing less than 50 sequences were rejected.[13] Criteria for sequence interpretations are provided in Table 1.

Background DNA

Management of background contaminant bacterial DNA was done as described previously.[13] There was a high consistency across all negative and positive extraction controls for the dominant contaminant bacterial species.

Background contaminant fungal DNA showed a higher variation across negative and positive extraction controls. For management of background fungal DNA, we defined a list of the ten most abundant contaminating fungi based on results from negative and positive extraction controls. Additionally, the laboratory keeps a list of

Table 1
Criteria for sequence interpretations.

Gene	Species	Species-group	Genus
16S[a]	≥99.3% homology with a high-quality reference, and minimum distance >0.7% to the next alternative species.[13]	≥99.3% homology with a high-quality reference, and minimum distance ≤0.7% to the next alternative species.	>97.0% homology with a high-quality reference
rpoB_Ent[b]	≥99.0% homology with a high-quality reference, and minimum distance >1.5% to the next alternative species.[13]	≥99.0% homology with a high-quality reference, and minimum distance ≤1.5% to the next alternative species	Not applicable
rpoB_ESS[c]	≥97.0% homology with a high-quality reference, and minimum distance >2.0% to the next alternative species.[13]	≥97.0% homology with a high-quality reference, and minimum distance ≤2.0% to the next alternative species.	Not applicable
ITS-2	≥99.0% homology with a high-quality reference, and minimum distance >2.0% to the next alternative species	≥99.0% homology with a high-quality reference, and minimum distance ≤2.0% to the next alternative species	Not applicable

[a] V3-V4 region of 16S rRNA-gene.
[b] rpoB-gene sequence targeted at Enterobacteriaceae.
[c] rpoB-gene sequence targeted at Staphylococcus, Enterococcus and Streptococcus species.

common contaminant fungi based on previous sequencing of negative and positive extraction controls. Fungi appearing in higher concentrations than any of these contaminants were accepted as valid identifications.

Statistical analysis

Statistical analyses were performed using SPSS 25 (IBM Corp). Clinical and microbial differences between subgroups were analyzed with Pearson's chi squared test for categorical data. For continuous data the Studentś *t*-test was used for normal distributed variables and Mann-Whitney U test for skewed variables.

Results

Clinical description of patients

Thirty-six patients – 19 (53%) males and 17 (47%) females – were included. The mean age was 70 years (median 72, range 37–94). Clinical and demographic characteristics together with main microbiological findings are presented in Table 2. Patients were categorized as having either moderate (24) or severe (12) acute cholecystitis according to the TG18/TG13 severity assessment criteria[19] (Supplementary Table S2). Compared to the moderate disease group, patients in the severe disease group were older, scored higher on Charlson's comorbidity index[20] and had higher prevalence of *Streptococcus* spp. and Enterobacteriaceae other than *Klebsiella* spp. and *E. coli*. Antibiotic treatment had been initiated for all patients except one prior to sample collection. Piperacillin/Tazobactam was the most frequently administered antibiotic, being part of or the only antimicrobial treatment for 28 patients. Eleven patients were diagnosed with local complications including marked local inflammation and/or perforated cholecystitis. The microbial findings by both NGS and culture from these patients can be found in Supplementary Table S3. Individual clinical characteristics, microbial findings and antibiotic treatment are provided in Supplementary Table S4. One patient died during hospital stay (Patient number 26, Supplementary Table S4). This patient had no detectable microbe in bile, neither by culture nor by sequencing.

Characteristics of the patient-control group are detailed in Table 2. Only three (19%) out of the 16 controls had detectable microbes in bile; *Streptococcus parasanguinis, Bifidobacterium animalis* and *Haemophilus parainfluenzae* were identified in one sample each.

Technical sequencing data

For the 16S rRNA amplicon the mean number of accepted reads per sample was 145,155 (range 28,592–404,981, median 115,519) after removal of short reads (<250 base pairs), small clusters (<50 reads) and chimeras. For the ITS2 amplicon the corresponding number was 23,024 (range 5150–47,568, median 15,514).

Microbial findings

Thirty-one samples (86%) contained bacteria (29) and/or fungi (5) as determined by sequencing (Table 2). Among these, five

Table 2
Characteristic, sequencing and culture results for all patients.

Patient group	
Demographic and clinical characteristics	
Number of patients	36
Male	19 (53%)
Mean age (SD; median; min–max), years	70 (16; 72; 37–94)
Community-acquired	30 (83%)
Mean CCI[a] (SD; median; min–max)	1,7 (1,9; 1,0; 0–8)
Gall bladder stone	30 (83%)
Bile duct stone	10 (28%)
Concomitant acute cholangitis	8 (22%)
Ongoing antibiotic therapy	35 (97%)
Severity grade:	
Moderate	24 (67%)
Severe	12 (33%)
Sequencing and culture results	
Samples with detected microbes by sequencing	31 (86%)
Samples with growth in bile culture	26 (72%)
Samples with detected bacteria by sequencing	29 (81%)
Samples with detected fungi by sequencing	5 (14%)
Polymicrobial samples by sequencing	15 (42%)
Major groups of bacteria detected by sequencing:	
Samples with *Klebsiella* spp.	11
Samples with *E. coli*	10
Samples with Enterobacteriaceae other than *Klebsiella* spp. and *Escherichia coli*	7
Samples with *Enterococcus* spp.	7
Samples with *Streptococcus* spp.	13
Samples with anaerobic bacteria	10
Patient control-group	
Number of patients	16
Male	6 (38%)
Mean age (SD; median; min–max), years	53 (18; 55; 20–79)
Mean CCI[a] (SD; median; min–max)	0,25 (0,5; 0; 0–1)
Samples with detected microbes by sequencing and/or culture	3
Species detected	
Bifidobacterium animalis (detected by)	1 (sequencing and culture)
Streptococcus parasanguinis (detected by)	1 (sequencing and culture)
Haemophilus parainfluenzae (detected by)	1 (sequencing)

[a] CCI = Charlsons comorbidity index.

Table 3
Species identified at a higher taxonomic level with use of partial *rpoB*-gene compared to partial 16S rRNA gene sequencing (V3–V4).

	16S rRNA gene sequencing results	*rpoB*-gene sequencing results
1	*Citrobacter werkmanii/Citrobacter freundii/Citrobacter braakii/Citrobacter pasteurii/Kluyvera ascorbata*	*Citrobacter* sp.
2	*Klebsiella michiganensis/Enterobacter ludwigii/Enterobacter asburiae/Enterobcater cloacae/Enterobacter kobei/Citrobacter freundii/Salmonella enterica*	*Enterobacter asburiae/Enterobacter cloacae/Enterobacter kobei*
3	*Enterobacter asburiae/Enterobacter cloacae/Enterobacter hormaechei/Klebsiella michiganensis/Klebsiella oxytoca/Klebsiella pneumoniae/Klebsiella quasipneumoniae*	*Enterobacter cloacae/Enterobacter hormaechei*
4	*Enterococcus gallinarum /Enterococcus casseliflavus*	*Enterococcus casseliflavus*
5	*Enterococcus durans/Enterococcus faecium/Enterococcus hirae*	*Enterococcus faecium*
6	*Escherichia coli/Escherichia albertii/Escherichia fergusonii/Shigella species*	*Escherichia coli/Shigella* sp.
7	*Hafnia alvei/Hafnia paralvei/Ewingella americana*	*Hafnia alvei*
8	*Klebsiella michiganensis/Klebsiella oxytoca/Enterobacter asburiae/Enterobacter cloacae/Enterobacter hormaechei*	*Klebsiella michiganensis*
9	*Klebsiella michiganensis/Klebsiella oxytoca/Enterobacter asburiae/Enterobacter hormaechei/Enterobacter cloacae/Salmonella enterica*	*Klebsiella oxytoca*
10	*Klebsiella aerogenes/Enterobacter asburiae/E. cancerogenes/Enterobacter cloacae/Enterobacter hormaechei/Enterobacter ludwigii/Enterobacter xiangfangensis/Klebsiella pneumoniae/Klebsiella oxytoca/Klebsiella michiganensis/Klebsiella variicola*	*Klebsiella pneumoniae/Klebsiella quasipneumoniae*
11	*Klebsiella pneumoniae/Klebsiella variicola*	*Klebsiella variicola*
12	*Proteus hauseri/Proteus penneri/Proteus vulgaris*	*Proteus vulgaris*
13	*Salmonella enterica/Enterobacter cloacae/Enterobacter kobei/Enterobacter ludwigii/Citrobacter amalonaticus/Klebsiella michiganensis*	*Salmonella enterica*
14	*Streptococcus anginosus/Streptococcus intermedius*	*Streptococcus anginosus*
15	*Streptococcus gordonii/Streptococcus cristatus*	*Streptococcus gordonii*
16	*Streptococcus mitis/oralis group*	*Streptococcus mitis*
17	*Streptococcus mitis/oralis group*	*Streptococcus oralis*
18	*Streptococcus sanguinis group*	*Streptococcus parasanguinis*
19	*Streptococcus salivarius group*	*Streptococcus salivarius*
20	*Streptococcus sanguinis group*	*Streptococcus sanguinis*
21	*Streptococcus salivarius group*	*Streptococcus thermophilus*

samples were culture negative. From the 106 microbial detections made by sequencing (100 bacteria and 6 fungi), only 40 were cultured (38%). The 100 bacteria detected by sequencing represented 53 different species of which 38 were identified to the species level, 14 to a species group level, and 1 to the genus level. The *rpoB* gene improved identification for 21 species (Table 3). Two bacterial identifications were made by culture alone, one *Klebsiella pneumoniae* and one *Staphylococcus epidermidis*. A detailed comparison of identifications made by sequencing versus culture is provided in Table 4. Table 5 provides an overview of the bacterial genera found in each patient and the proportion of samples containing each genus. In patients with polymicrobial infections culture failed to detect one or more microbes in all 15 samples (Supplementary Table S4). For the monomicrobial infections, there was a higher concordance (81%) between culture and sequencing. Only three of the 16 monomicrobial samples were culture negative (Supplementary Table S4).

Six fungi were identified by sequencing (Table 4) whereof one, a *Candida albicans*, was also cultured. Two samples were monomicrobial containing *C. albicans*; one severe postoperative acalculous cholecystitis after pancreatic cancer surgery who also had *C. albicans* in blood culture, and one community-acquired calculous cholecystitis of moderate severity. The other identified fungi, *Saccharomyces cerevisiae* (2), *C. albicans* (1) and *Candida humilis* (1) were part of poly-microbial infections (Supplementary Table S4). Only the patient with severe postoperative acalculous cholecystitis received antifungal treatment.

Blood culture samples were collected from 24 patients whereof five had a detectable bacteremia (Supplementary Table S4). Antibiograms of all bacteria cultured from bile or in blood culture are provided in Supplementary Table S5.

Discussion

To the best of our knowledge, this is the first study that uses NGS for microbial characterization of bile samples from patients with acute cholecystitis, with the exception of a small study on six patients.[21] This is also the first study to describe the bacteriology of severe acute cholecystitis according to the TG18/TG13 severity grading.[5,8]

Although bactobilia is considered a negative prognostic factor in acute cholecystitis, there is, with the exception of the aforementioned Israeli study,[7] little evidence on the clinical importance of the individual bacterial species. In many of the polymicrobial samples in our study, the relative abundance of the identified bacteria varied widely (Table 5). Some might dismiss the clinical relevance of low abundance species in complex infections, in particular if found be sequencing only. However, several of the bacteria identified were anaerobic, fastidious, slow growing and/or antibiotics-affected, and their failure to survive and grow in the laboratory does not mean that they are eradicated from the infection site nor that they are of lower clinical relevance. We would also like to point out that abundant growth does not necessarily reflect in-vivo dominance but might as well reflect a microbe's ability to thrive and compete during transportation and cultivation. We have frequently observed, also in this study, that bacteria with abundant growth constitute only minor parts of the population as determined by sequencing or that a dominant microbe as determined by sequencing fails to grow. In our opinion, the clinical relevance of individual bacteria in complex infections should not be considered based on relative quantifications or by method of detection. Rather, such inference should be based on in-depth ecological knowledge of each type of infection, including microbial dynamics over time, microbial aggregate formation, metabolic interdependencies and synergisms.[22,23] Complete microbial characterizations as provided in this study represent the first step in obtaining such knowledge but needs to be followed up by both experimental studies and larger clinical studies.

Except for *Klebsiella* spp. and *E. coli*, we found that 50% of species in the Enterobacteriaceae family, including species from the genera *Citrobacter, Enterobacter, Proteus, Hafnia, Salmonella, Serratia, Morganella and Raoultella* remained undiscovered by culture

Table 4
Identified bacteria and fungi from bile samples by sequencing compared to conventional culture.

	Total number of identifications by sequencing (% of all microbial detections)	Growth by culture
Total identifications	106	40
Gram negative[a]	41 (39%)	23
Klebsiella	11 (10%)	9
pneumoniae/quasipneumoniae	3	3
Michiganensis[c]	3	2
Oxytoca[c]	3	2
Variicola[c]	2	2
Escherichia coli[b,c]	10 (9%)	9
Campylobacter	4 (4%)	0
Concisus	1	0
Concisus/mucosalis	1	0
Curvus	1	0
Rectus/showae	1	0
Citrobacter	3 (3%)	1
Species[c]	2	1
Amalonaticus/farmeri	1	0
Haemophilus parainfluenzae	3 (3%)	0
Enterobacter	2 (2%)	1
Asburiae/cloacae/kobei[c]	1	1
Cloacae/hormaechei[c]	1	0
Morganella morganii	2 (2%)	1
Hafnia alvei[c]	1 (1%)	0
Proteus vulgaris	1 (1%)	1
Pseudomonas aeruginosa/otidis	1 (1%)	1
Raoultella ornithinolytica/planticola	1 (1%)	0
Salmonella enterica[c]	1 (1%)	0
Serratia marcescens	1 (1%)	0
Gram positive[a]	35 (33%)	14
Streptococcus	15 (14%)	6
Anginosus[c]	3	1
Salivarius[c]	3	2
Sanguinis[c]	2	1
Gordonii	1	1
Massiliensis	1	1
Mitis[c]	1	0
Mutans	1	0
Oralis[c]	1	0
Parasanguinis[c]	1	0
Termophilus[c]	1	0
Enterococcus	11 (10%)	5
Faecalis	4	2
Faecium[c]	4	2
Avium/raffinosus[c]	2	1
Casseliflavus[c]	1	0
Lactobacillus casei/paracasei/rhamnosus	4 (4%)	3
Actinomyces	5 (5%)	0
Gerencseriae	1	0
Naeslundii/oris	1	0
Naeslundii/oris/johnsonii	1	0
sp. (oral taxon 848)	1	0
Turicensis	1	0
Anaerobic	24 (23%)	2
Clostridium perfringens	5 (5%)	2
Fusobacterium nucleatum	5 (5%)	0
Bifdobacterium	4 (4%)	0
Animalis	2	0
Dentium	1	0
Longum	1	0
Veillonella	3 (3%)	0
Dispar/parvula	2	0
Parvula/tobetsuensis/dentocariosa	1	0
Intestinibacter bartletti	3 (3%)	0
Slackia exigua	1 (1%)	0
Dialister invisius	1 (1%)	0
Bilophila wadsworthia	1 (1%)	0
Propionibacterium acidifaciens	1 (1%)	0
Fungus	6 (6%)	1
Candida	4 (4%)	1
Albicans	3	1
Humilis	1	0
Saccharomyces cerevisiae	2 (2%)	0

[a] One *K. pneumoniae* and one *S. epidermidis* detected exclusively by culture is not included in table.
[b] Not distinguishable from *Shigella* spp.
[c] *rpoB* sequencing provided identification at a higher taxonomic level than 16S rRNA gene sequencing.

Table 5
Heatmap of all bacterial genus identified in each patient. Only samples containing bacteria are included in the table.

Bacterial ID	Number of samples containing the bacterium (%)[a]	1	2	4	7	8	9	10	12	15	16	17	18	19	20	21	22	23	24	25	27	28	29	30	31	32	33	34	35	36
Streptococcus spp.	13 (45%)	5,6	100		0,6			100	21,5	0,2	100						59,6	0,1						88,5	100				84,2	0,5
Klebsiella spp.	11 (38%)	2,3			46,4	100	42,0		0,2	91,0		100		11,6				6,2		32,2										75,4
Escherichia coli	10 (34%)	90,3		100	53,0				20,8				100								100		100			5,6	51,3	100		
Enterococcus spp.	7 (24%)	1,0					37,7		0,1					88,3				11,6		41,0							46,3			
Clostridium perfringens	5 (17%)														100			0,8								68,5	2,3			12,9
Fusobacterium nucleatum	5 (17%)	0,1							42,8								7,0	47,4												11,2
Campylobacter spp.	4 (14%)									0,1							0,5	0,4		10,0										
Lactobacillus spp.	4 (14%)									0,3							7,4		100							23,8				
Actinomyces spp.	3 (10%)	0,3								2,2							5,1													
Bifidobacterium spp.	3 (10%)									0,5							0,6	0,2												
Citrobacter spp.	3 (10%)	0,1					1,8											0,3												
Haemophilus parainfluenzae	3 (10%)																0,3					100							15,8	
Intestinibacter bartlettii	3 (10%)													0,1				0,1								2,1				
Veillonella spp.	3 (10%)	0,3															2,8	7,1												
Enterobacter spp.	2 (7%)									3,3						100														
Morganella morganii	2 (7%)																	0,6		9,6										
Bilophila wadsworthia	1 (3%)								14,6																					
Dialister invisus	1 (3%)																	0,1												
Hafnia alvei	1 (3%)						0,1																							
Propionibacterium acidifaciens	1 (3%)																4,0													
Proteus vulgaris	1 (3%)																	25,1												
Pseudomonas spp.	1 (3%)						18,4																							
Raoultella spp.	1 (3%)									2,3																				
Salmonella enterica	1 (3%)																							11,5						
Serratia marcescens	1 (3%)																			7,3										
Slackia exigua	1 (3%)																12,8													

[a] Percentage in parenthesis represents the proportion of samples, out of all bacterial sequencing-positive samples, containing the bacterium.

(Table 4). These bacteria are generally considered clinically relevant and there is evidence to support their role in the pathogenesis of acute cholecystitis.[24] For *Enterobacter* spp. there is also a possible association with a poorer patient outcome.[7] The capability of acquiring or inducing antibiotic resistance, and a high frequency of multi-resistant clones among some Enterobacteriaceae, increases the likely clinical benefit of identifying these bacteria.[25]

Only five out of eleven enterococci were found by conventional culture (Table 4 and Supplementary Table S4). The clinical significance of enterococci in acute cholecystitis and in intra-abdominal infections in general remains uncertain. Most empiric guidelines for antibiotic treatment of acute cholecystitis do not include specific enterococcal coverage,[9,11] except for the Tokyo Guidelines' recommendation of adding vancomycin for severe cholecystitis.[10] However, in complicated acute cholecystitis and/or severely ill patients it is recommended to use microbiology culture results to guide antimicrobial treatment.[10,11,26] This implies that if the enterococci found only by sequencing in our cohort had also been found by culture, it might have led to an adjustment of antibiotic treatment. As mentioned, failure to culture microbes does not mean that they have been eradicated from the infection site. Future studies addressing the relevance of enterococci should therefore not rely on culture-based diagnostics alone but also include molecular approaches like sequencing or PCR.

Anaerobic bacteria may be sub-optimally covered by monotherapy with a third-generation cephalosporin whereas Piperacillin/tazobactam provides good coverage of anaerobic bacteria. In this study, NGS detected 24 anaerobic bacteria from 10 samples whereof only two (8%) were also detected by culture (Table 4). The two most common anaerobe species were *Clostridium perfringens* and *Fusobacterium nucleatum. Clostridium perfringens* is known for its pathogenicity and its ability to cause emphysematous cholecystitis. *Fusobacterium nucleatum* has to the best of our knowledge not previously been reported in acute cholecystitis but is considered an important anaerobe pathogen in both odontogenic infections, pleural empyemas and brain abscesses.[12,13]

In healthy individuals, the bile is considered to be sterile,[27–29] but gallstone disease might lead to bacterial colonization. Culture-based studies report bacteria in between 9% and 54% of patients with gallstone disease without infection.[6,28,29] Two NGS-based studies addressing this issue report conflicting results. One study found a very high rate of colonization (100%) and suggest the existence of a bile core microbiome comprising 208 Operational Taxonomic Units (OTUs)/species.[30] Another study found the rate of colonization to be 13% with a mean bacterial diversity of 5 OTUs per sample.[31] Both studies fail to explain how they addressed the problem of contaminant background DNA, chimera formation and sequencing noise. These are fundamental challenges in microbiome studies and will significantly inflate microbial diversity if not considered properly.[32,33] In our patient control group only three (19%) bile samples were colonized, each with a single bacterial species (Table 2), providing little support for the existence of a bile microbiome.

Some limitations to this study should be noted. It is a single center investigation with a relatively low number of patients, and the general validity of our results therefore needs confirmation by other studies. The patients in our cohort were also of higher mean age than in historic studies on moderate and severe cholecystitis which may in part explain the higher rate of bactobilia observed.[4,5,29] Due to the severity of the disease, antibiotic treatment had been initiated for most patients prior to sample collection. Although bacterial DNA is very stable in undrained purulent infections this might still have impacted the observed relative abundancies of species in the polymicrobial infections.

We have shown that culture-based methods alone are insufficient in the microbiological diagnostics of moderate and severe acute cholecystitis, leaving more than 60% of the microbes undetected. The clinical consequences of not detecting or treating all these bacteria should be further addressed in future studies as should eventual consequences for empiric treatment recommendations. Yet, clinicians should be aware of the risk of leaving clinical important bacteria untreated if antimicrobial treatment is customized based on culture results only. For anaerobic bacteria, the low recovery rate may imply that anaerobic coverage should be considered regardless of a negative anaerobic culture. This and other studies emphasize the need for rapid and reliable culture-independent microbial detection and susceptibility testing in diagnostic microbiology.

Funding

This work was supported by the Western Norway Regional Health Authority's research funding [grant number 912206].

Declaration of Competing Interest

O.K. contributed to the development of the RipSeq software and is a minor shareholder of Pathogenomix Inc.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.jinf.2019.09.015.

References

1. Kimura Y, Takada T, Kawarada Y, Nimura Y, Hirata K, Sekimoto M, et al. Definitions, pathophysiology, and epidemiology of acute cholangitis and cholecystitis: Tokyo guidelines. *J Hepatobiliary Pancreat Surg* 2007;**14**(1):15–26.
2. Indar AA, Beckingham IJ. Acute cholecystitis. *BMJ* 2002;**325**(7365):639–43.
3. Csendes A, Burdiles P, Maluenda F, Diaz JC, Csendes P, Mitru N. Simultaneous bacteriologic assessment of bile from gallbladder and common bile duct in control subjects and patients with gallstones and common duct stones. *Arch Surg* 1996;**131**(4):389–94.
4. Yun SP, Seo HI. Clinical aspects of bile culture in patients undergoing laparoscopic cholecystectomy. *Medicine* 2018;**97**(26):e11234.
5. Asai K, Watanabe M, Kusachi S, Tanaka H, Matsukiyo H, Osawa A, et al. Bacteriological analysis of bile in acute cholecystitis according to the Tokyo guidelines. *J Hepatobiliary Pancreat Sci* 2012;**19**(4):476–86.
6. Darkahi B, Sandblom G, Liljeholm H, Videhult P, Melhus A, Rasmussen IC. Biliary microflora in patients undergoing cholecystectomy. *Surg Infect* 2014;**15**(3):262–5.
7. Nitzan O, Brodsky Y, Edelstein H, Hershko D, Saliba W, Keness Y, et al. Microbiologic data in acute cholecystitis: ten years' experience from bile cultures obtained during percutaneous cholecystostomy. *Surg Infect* 2017;**18**(3):345–9.
8. Cueto-Ramos R, Hernandez-Guedea M, Perez-Rodriguez E, Reyna-Sepulveda F, Munoz-Maldonado G. Incidence of bacteria from cultures of bile and gallbladder wall of laparoscopic cholecystectomy patients in the University Hospital "Dr. Jose Eleuterio Gonzalez". *Cir Cir* 2017;**85**(6):515–21.
9. Bellows C. *Cholecystitis. BMJ best practice*. BMJ Publishing Group; 2018. 2018 [updated March 29, 2018. Available from: https://bestpractice.bmj.com/topics/en-us/78 .
10. Gomi H, Solomkin JS, Schlossberg D, Okamoto K, Takada T, Strasberg SM, et al. Tokyo guidelines 2018: antimicrobial therapy for acute cholangitis and cholecystitis. *J Hepatobiliary Pancreat Sci* 2018;**25**(1):3–16.
11. Ansaloni L, Pisano M, Coccolini F, Peitzmann AB, Fingerhut A, Catena F, et al. 2016 WSES guidelines on acute calculous cholecystitis. *World J Emerg Surg* 2016;**11**:25.
12. Kommedal O, Wilhelmsen MT, Skrede S, Meisal R, Jakovljev A, Gaustad P, et al. Massive parallel sequencing provides new perspectives on bacterial brain abscesses. *J Clin Microbiol* 2014;**52**(6):1990–7.
13. Dyrhovden R, Nygaard RM, Patel R, Ulvestad E, Kommedal O. The bacterial aetiology of pleural empyema. A descriptive and comparative metagenomic study. *Clin Microbiol Infect*. 2019;**25**(8):981–6.
14. Yokoe M, Takada T, Strasberg SM, Solomkin JS, Mayumi T, Gomi H, et al. New diagnostic criteria and severity assessment of acute cholecystitis in revised Tokyo guidelines. *J Hepatobiliary Pancreat Sci* 2012;**19**(5):578–85.
15. Loozen CS, Blessing MM, van Ramshorst B, van Santvoort HC, Boerma D. The optimal treatment of patients with mild and moderate acute cholecystitis: time for a revision of the Tokyo guidelines. *Surg Endosc* 2017;**31**(10):3858–63.

16. Loozen CS, van Santvoort HC, van Duijvendijk P, Besselink MG, Gouma DJ, Nieuwenhuijzen GA, et al. Laparoscopic cholecystectomy versus percutaneous catheter drainage for acute cholecystitis in high risk patients (CHOCOLATE): multicentre randomised clinical trial. *BMJ* 2018;**363**:k3965.
17. Kommedal O, Kvello K, Skjastad R, Langeland N, Wiker HG. Direct 16S rRNA gene sequencing from clinical specimens, with special focus on polybacterial samples and interpretation of mixed DNA chromatograms. *J Clin Microbiol* 2009;**47**(11):3562–8.
18. Khot PD, Ko DL, Fredricks DN. Sequencing and analysis of fungal rRNA operons for development of broad-range fungal PCR assays. *Appl Environ Microbiol* 2009;**75**(6):1559–65.
19. Kiriyama S, Kozaka K, Takada T, Strasberg SM, Pitt HA, Gabata T, et al. Tokyo guidelines 2018: diagnostic criteria and severity grading of acute cholangitis (with videos). *J Hepatobiliary Pancreat Sci* 2018;**25**(1):17–30.
20. Charlson ME, Pompei P, Ales KL, MacKenzie CR. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. *J Chronic Dis* 1987;**40**(5):373–83.
21. Kujiraoka M, Kuroda M, Asai K, Sekizuka T, Kato K, Watanabe M, et al. Comprehensive diagnosis of bacterial infection associated with acute cholecystitis using metagenomic approach. *Front Microbiol* 2017;**8**:685.
22. Bradshaw DJM, P D, Watson GK, Allison C. Oral anaerobes cannot survive oxygen stress without interacting with facultative/aerobic species as a microbial community. *Lett Appl Microbiol* 1997;**25**:385–7.
23. Zengler K, Zaramela LS. The social network of microorganisms – how auxotrophies shape complex communities. *Nat Rev Microbiol* 2018;**16**(6):383–90.
24. Liu J, Yan Q, Luo F, Shang D, Wu D, Zhang H, et al. Acute cholecystitis associated with infection of Enterobacteriaceae from gut microbiota. *Clin Microbiol Infect* 2015;**21**(9):851.e1–851.e9.
25. Iredell J, Brown J, Tagg K. Antibiotic resistance in Enterobacteriaceae: mechanisms and clinical implications. *BMJ* 2016;**352**:h6420.
26. Solomkin JS, Mazuski JE, Bradley JS, Rodvold KA, Goldstein EJ, Baron EJ, et al. Diagnosis and management of complicated intra-abdominal infection in adults and children: guidelines by the surgical infection society and the infectious diseases society of America. *Surg Infect* 2010;**11**(1):79–109.
27. Ikeda T, Yanaga K, Kusne S, Fung J, Higashi H, Starzl TE. Sterility of bile in multiple-organ donors. *Transplantation* 1990;**49**(3):653.
28. Abeysuriya V, Deen KI, Wijesuriya T, Salgado SS. Microbiology of gallbladder bile in uncomplicated symptomatic cholelithiasis. *Hepatobiliary Pancreat Dis Int* 2008;**7**(6):633–7.
29. Csendes A, Becerra M, Burdiles P, Demian I, Bancalari K, Csendes P. Bacteriological studies of bile from the gallbladder in patients with carcinoma of the gallbladder, cholelithiasis, common bile duct stones and no gallstones disease. *Eur J Surg* 1994;**160**(6–7):363–7.
30. Wu T, Zhang Z, Liu B, Hou D, Liang Y, Zhang J, et al. Gut microbiota dysbiosis and bacterial community assembly associated with cholesterol gallstones in large-scale study. *BMC Genom* 2013;**14**:669.
31. Tsuchiya Y, Loza E, Villa-Gomez G, Trujillo CC, Baez S, Asai T, et al. Metagenomics of microbial communities in gallbladder bile from patients with gallbladder cancer or cholelithiasis. *Asian Pac J Cancer Prev* 2018;**19**(4):961–7.
32. Glassing A, Dowd SE, Galandiuk S, Davis B, Chiodini RJ. Inherent bacterial DNA contamination of extraction and sequencing reagents may affect interpretation of microbiota in low bacterial biomass samples. *Gut Pathog* 2016;**8**:24.
33. Auer L, Mariadassou M, O'Donohue M, Klopp C, Hernandez-Raquet G. Analysis of large 16S rRNA Illumina data sets: impact of singleton read filtering on microbial community description. *Mol Ecol Resour* 2017;**17**(6):e122–ee32.

Supplementary Table S1: Primers with adapter sequences. Sequences of the target specific portions in capital letters

Name	Sequence	Position[a]
16S-F[b]	tcgtcggcagcgtcagatgtgtataagagacagCCTACGGGNGGCWGCAG	340-356
16S-R[b]	gtctcgtgggctcggagatgtgtataagagacagGACTACCAGGGTATCTAAKCC	784-803
ITS2-F	tcgtcggcagcgtcagatgtgtataagagacagGTGAATCATCGARTCTTTGAA	NA[c]
ITS2-R	gtctcgtgggctcggagatgtgtataagagacagTATGCTTAAGTTCAGCGGGTA	NA[c]
RpoB_Ent-F	tcgtcggcagcgtcagatgtgtataagagacagGAAGGTCCRAAYATCGGTCT	1693-1712
RpoB_Ent-R	gtctcgtgggctcggagatgtgtataagagacagTGCATGTTCGCACCCAT	2041-2057
RpoB_ESS-F1	tcgtcggcagcgtcagatgtgtataagagacagGCRACAGCRTGTATYCCRTTC	1861-1881
RpoB_ESS-F2	tcgtcggcagcgtcagatgtgtataagagacagGCDACAGCATGTATTCCWTTC	1861-1881
RpoB_ESS-R	gtctcgtgggctcggagatgtgtataagagacagGTTRTAMCCNTCCCAWGTCAT	2287-2307

[a] Positions for 16S based on *Escherichia coli* (GenBank accession J01859). Positions for RpoB_ESS based on *Staphylococcus aureus* [*rpoB* coding sequence (CDS); GenBank accession X64172]. Positions for RpoB_Ent based on *Escherichia coli* [*rpoB* coding sequence (CDS); GenBank accession V00340].

[b] Abbreviations: F = forward primer. R = reverse primer.

[c] Not applicable. Both F and R primers are flanking the ITS2-segment and is located at 5.8S and 28S respectively.

Supplementary Table S2: Characteristics of patients according to the Tokyo Guidelines severity grading

	Grade II Moderate	Grade III Severe	p[a]
Demographic and clinical characteristics			
Number of patients	24	12	
Male	11	8	0,24
Mean age (SD; median; min-max), years	66 (17; 68; 37-94)	79 (13; 84; 51-92)	0,03
Community-acquired	20	10	1,0
CCI[b] (SD; median; min-max)	1,1 (1,3; 1,0; 0-5)	3,0 (2,3; 2,5; 0-8)	0,01
Gall bladder stone	21	9	0,34
Bile duct stone	8	2	0,29
Concomitant acute cholangitis	5	3	0,78
Growth in blood culture (of tested)	2 (15)	4 (9)	0,09
Ongoing antibiotic therapy	24	11	
In-hospital death	1	0	
Severity grading parameters:			
Cardiovascular dysfunction	0	3	
Neurological dysfunction	0	4	
Respiratory dysfunction	0	10	
Renal dysfunction	0	2	
Hepatic dysfunction	0	3	
WBC > 18 10^9/L	14	6	
Palpable tender mass	3	2	
Duration > 72 hours	24	10	
Marked local inflammation	4	2	
Sequencing and culture results			
Detected microbes by sequencing	19	12	0,09
Growth in bile culture	16	10	
Detected bacteria by sequencing	18	11	0,23
Detected fungi by sequencing	2	3	0,73
Polymicrobial sample by sequencing	8	7	0,15
Major groups of bacteria detected by Sequencing:			
Enterobacteriaceae other than *Klebsiella* spp. and *Escherichia coli*	2	5	0,02
Klebsiella spp.	5	6	0,07
Escherichia coli	9	1	0,07
Enterococcus spp.	4	3	0,55
Streptococcus spp.	5	8	0,01
Anaerobic	6	4	0,6

[a] Pearson's chi-squared test for categorical variables. Students t-test for continuous, normal distributed variables. Mann-Whitney U-test for continuous, skewed variables.

[b] Charlsons comorbidity index

Supplementary Table S3: Microbial findings by NGS and culture in patients with local complications

ID	Local complications	16S rRNA, *rpoB* gene and ITS sequencing - sorted by decreasing abundance[a]	Culture
4	Biliary peritonitis	*Escherichia coli* [b]	*E. coli* [ag]
8	Pericholecystic abscess	*Klebsiella pneumoniae / Klebsiella quasipneumoniae*[b]	*K. pneumoniae* [c, ag]
12	Pericholecystic abscess	*Fusobacterium nucleatum* *Streptococcus massiliensis* *E. coli* [b] *Bilophila wadsworthia* *Klebsiella oxytoca* [b] *Enterococcus faecalis*	*E. coli*[ag] *S. massiliensis* [mg]
15	Perforated cholecystitis	*Klebsiella michiganensis* [b] *Enterobacter cloacae / Enterobacter hormaechei* [b] *Raoultella ornithinolytica / Raoultella planticola* *Actinomyces naeslundii / Actinomyces oris* *Bifidobacterium animalis* *Lactobacillus rhamnosus / Lactobacillus casei / Lactobacillus paracasei* *Streptococcus mutans* *Campylobacter concisus / Campylobacter mucosalis*	*K. oxytoca* [c, ag]
17	Perforated cholecystitis	*K.pneumoniae / K. quasipneumoniae*[b]	*K. pneumoniae* [c, ag]
20	Gangrenous cholecystitis	*Clostridium perfringens* *Saccharomyces cerevisiae*	*C. perfringens*
21	Biliary peritonitis	*Enterobacter asburiae / E. cloacae / Enterobacter kobei* [b]	*Enterobacter cloacae complex* [mg]
23	Gangrenous cholecystitis	*F. nucleatum* *Proteus vulgaris* [b] *Veillonella parvula / Veillonella dispar* *K. pneumoniae / K. quasipneumoniae* [b]	*K. pneumoniae*[c, ag] *P. vulgaris* [ag] *E. faecium* [ag]

		Enterococcus faecium [b] *Enterococcus avium / Enterococcus raffinosus* [b] *C. perfringens* *Morganella morganii* *Campylobacter curvus* *Citrobacter sp.* [b] *B. animalis* *Dialister invisus* *I. bartlettii* *Streptococcus salivarius* [b] *S. cerevisiae* *Candida humilis*	
27	Perforated cholecystitis	*E. coli* [b]	*E. coli* [mg] *S. epidermidis* [e, mg]
32	Perforated cholecystitis	*C. perfringens* *L. rhamnosus / L. casei / L. paracasei* *E. coli* [b] *I. bartlettii*	*L. rhamnosus* [sg]
35	Perforated cholecystitis	*Streptococcus anginosus*[b] *Haemophilus parainfluenzae*	Negative

[a] It was not possible to compute the relative abundance of fungi compared to bacteria since they were identified in different PCR reactions. Fungi are therefore listed at the end of each list of microbes.

[b] *rpoB* sequencing provided identification at a higher taxonomic level than 16S rRNA gene sequencing.

[c] MALDI/TOF spectra database does not contain *K. michiganensis* nor *K. quasipneumoniae*. These two species will consequently most probably be reported as a *K. pneumoniae* when analyzed with MALDI-TOF.

[e] *S. epidermidis* most likely represents contamination as it was found by culture only

[ag] abundant growth, [mg] medium abundant growth, [sg] sparse growth

Supplementary Table S4: Clinical characteristics of all patients and comparison between parallel sequencing and culture for all patients

ID Sex/ Age	**Hospital (HA) / community (CA) acquired infection, severity grading, calculous/acalculous, bile duct stone present, marked local inflammation, etiology, other relevant concomitant diseases, sampling method.**	**16S rRNA, *rpoB* gene and ITS sequencing - sorted by decreasing abundance[a]**	**Culture**	**Growth in blood culture**	**Antibiotic treatment before sampling. Ongoing (O) treatment and treatment last 14 days (L)**
1 F/37	CA, grade 2, acalculous, no bile duct stone, no marked local inflammation, inflammation in common bile duct probable cause of cholecystitis, concomitant acute cholangitis, PTHC	*Escherichia coli* [b] *Streptococcus oralis* [b] *Klebsiella michiganensis* [b] *Enterococcus casseliflavus* [b] *Streptococcus thermophilus* [b] *Actinomyces turicensis* *Veillonella parvula / Veillonella tobetsuensis / Veillonella dentocariosa* *Enterococcus faecium* [b] *Fusobacterium nucleatum* *Citrobacter amalonaticus / Citrobacter farmeri* *Candida albicans*	*E. coli* [mg]	Negative	O = iv PIT
2 F/92	CA, grade 3, calculous, no bile duct stone, no marked local inflammation, cholelithiasis probable cause of cholecystitis, PTHC.	*Streptococcus salivarius* [b]	*S. salivarius* [ag]	Negative	O = iv CTA and MET.
3 M/69	HA, grade 3, acalculous, no bile duct stone, no marked local inflammation, postoperative cholecystitis 10 days after cancer surgery for pancreatic adenocarcinoma, PTHC.	*Candida albicans*	Negative	*C. albicans*	Peroperative treatment with MET and dox. 7 days before sampling
4 F/63	HA, grade 2, calculous, no bile duct stone, perforated cholecystitis with biliary peritonitis, cholelithiasis probable cause of cholecystitis, concomitant biliary acute pancreatitis and chronic cholecystitis, PTHC.	*E. coli* [b]	*E. coli* [ag]	*Negative*	O = iv CTA and MET.

5 M/71	HA, grade 2, calculous, bile duct stone present, no marked local inflammation, cholelithiasis probable cause of cholecystitis, already hospitalized for a spinal cord injury when cholecystitis occurs, concomitant acute cholangitis, PTHC.	Negative	Negative	Negative	O = iv PIT, L = po TRS
6 M/53	CA, grade 2, calculous, bile duct stone present, no marked local inflammation, cholelithiasis probable cause of cholecystitis, PTHC.	Negative	Negative	Not taken	O = iv PIT
7 M/92	CA, grade 3, calculous, no bile duct stone, no marked local inflammation, cholelithiasis probable cause of infection, PTHC.	*E. coli* [b] *Klebsiella oxytoca* [b] *Streptococcus parasanguinis* [b]	*E. coli* [ag] *K. oxytoca / Raoultella ornithinolytica* [ag]	Negative	O = iv PIT and MET.
8 F/77	CA, grade 2, calculous, no bile duct stone, pericholecystic abscess, cholelithiasis probable cause of cholecystitis, concomitant chronic cholecystitis, PTHC.	*Klebsiella pneumoniae / Klebsiella quasipneumoniae*[b]	*K. pneumoniae* [c, ag]	Negative	O = iv PIT, L = po MET and TRS
9 F/51	HA, grade 3, calculous, no bile duct stone, no marked local inflammation, common bile duct stent probable cause of cholecystitis, concomitant acute cholangitis, PTHC.	*K. oxytoca* *Enterococcus avium / Enterococcus raffinosus* [b] *Pseudomonas aeruginosa* *Citrobacter sp.* [b] *E. faecium* [b] *Hafnia alvei* [b]	*Citrobacter species* [ag] *K. oxytoca* [ag] *K. pneumoniae* [ag] *P. aeruginosa* [ag] *E. raffinosus* [ag]	Negative	O = iv PIT
10 F/50	CA, grade 2, calculous, no bile duct stone, no marked local inflammation, cholelithiasis probable cause of cholecystitis, PTHC, perioperative.	*Streptococcus mitis* [b]	Negative	Not taken	O = iv PIT L = po MET and TRS
11 F/57	CA, grade 2, calculous, no bile duct stone, no marked local inflammation, cholelithiasis probable cause of cholecystitis, PTHC.	*C. albicans*	*C. albicans*[ag]	Negative	O = iv PIT and po MET and TRS
12 M/50	CA, grade 2, acalculous, no bile duct stone, pericholecystic abscess, unknown etiology, concomitant chronic pancreatitis, PTHC.	*F. nucleatum* *Streptococcus massiliensis* *E. coli* [b] *Bilophila wadsworthia* *K. oxytoca* [b] *Enterococcus faecalis*	*E. coli*[ag] *S. massiliensis* [mg]	Negative	O = iv PIT

13 F/67	CA, grade 2, calculous, no bile duct stone, no marked local inflammation, cholelithiasis probable cause of cholecystitis, PTHC.	Negative	Negative	Negative	O = iv PIT, MET, gent and penc.
14 F/92	HA, grade 2, calculous, no bile duct stone, no marked local inflammation, cholelithiasis probable cause of cholecystitis, recently hospitalized for surgical treatment of femoral neck fracture, PTHC.	Negative	Negative	Not taken	O = iv PIT
15 M/85	CA, grade 3, acalculous, no bile duct stone, perforated cholecystitis, unknown etiology, PTHC.	*K. michiganensis* [b] *Enterobacter cloacae / Enterobacter hormaechei* [b] *R. ornithinolytica / Raoultella planticola* *Actinomyces naeslundii / Actinomyces oris* *Bifidobacterium animalis* *Lactobacillus rhamnosus / Lactobacillus casei / Lactobacillus paracasei* *Streptococcus mutans* *Campylobacter concisus / Campylobacter mucosalis*	*K. oxytoca* [c, ag]	Not taken	O = iv PIT L = po MET and TRS.
16 F/82	CA, grade 3, calculous, no bile duct stone, no marked local inflammation, cholelithiasis probable cause of cholecystitis, PTHC.	*S. salivarius* [b]	*S. salivarius* [sg]	Not taken	O = iv PIT and po MET and TRS.
17 M/53	CA, grade 2, calculous, no bile duct stone, perforated cholecystitis, cholelithiasis probable cause of cholecystitis, PTHC.	*K. pneumoniae / K. quasipneumoniae* [b]	*K. pneumoniae* [c, ag]	*K. pneumoniae*	O = iv PIT L = po MET and TRS
18 M/51	CA, grade 2, calculous, bile duct stone present, no marked local inflammation, cholelithiasis probable cause of cholecystitis, PTHC.	*E. coli* [b]	*E. coli* [ag]	Negative	O = iv PIT
19 M/88	CA, grade 3, calculous, bile duct stone present, no marked local inflammation, cholelithiasis probable cause of cholecystitis, concomitant acute cholangitis and acute pancreatitis, PTHC.	*E. faecalis* *K. variicola* [b] *Intestinibacter bartlettii*	*E. faecalis* [ag] *K. pneumoniae* [d, sg]	*K. pneumoniae*[d]	O = iv PIT and IMI.

20 M/68	CA, grade 2, calculous, no bile duct stone, gangrenous cholecystitis, cholelithiasis probable cause of cholecystitis, perioperative.	*Clostridium perfringens* *Saccharomyces cerevisiae*	*C. perfringens* [ag]	Not taken	O = iv CLI and CTA.
21 M87	CA, grade 3, calculous, no bile duct stone, perforated cholecystitis with biliary peritonitis, cholelithiasis probable cause of cholecystitis, concomitant chronic cholecystitis, PTHC.	*E. asburiae / E. cloacae / E. Kobei* [b]	*Enterobacter cloacae complex* [mg]	Negative	O = iv PIT and po MET and TRS.
22 M/87	CA, grade 2, calculous, no bile duct stone, no marked local inflammation, cholelithiasis probable cause of cholecystitis, PTHC.	*Streptococcus anginosus* [b] *Slackia exigua* *L. rhamnosus / L. casei / L. paracasei* *F. nucleatum* *A. naeslundii / A. oris / Actinomyces johnsonii* *Propionibacterium acidifaciens* *Streptococcus sanguinis* [b] *Veillonella dispar / V. parvula* *Campylobacter rectus / Campylobacter showae* *Bifidobacterium longum* *Actinomyces gerencseriae* *Haemophilus parainfluenzae* *Bifidobacterium dentium* *Actinomyces sp. (oral taxon 848)*	*Streptococcus anginosus* [ag] *Lactobacillus paracasei* [na]	*S. anginosus*	O = iv PIT
23 M/72	CA, grade 3, acalculous, no bile duct stone, gangrenous cholecystitis, unknown etiology, PTHC.	*F. nucleatum* *Proteus vulgaris* [b] *V. parvula / V. dispar* *K. pneumoniae / K. quasipneumoniae* [b] *E. faecium* [b] *E. avium / E. raffinosus* [b] *C. perfringens* *Morganella morganii* *Campylobacter curvus* *Citrobacter sp.* [b] *B. animalis* *Dialister invisus* *I. bartlettii* *S. salivarius* [b] *S. cerevisiae* *Candida humilis*	*K. pneumoniae* [c, ag] *P. vulgaris* [ag] *E. faecium* [ag]	*K. pneumoniae* [c] *E. faecium*	O = iv PIT
24 F/41	CA, grade 2, calculous, no bile duct stone, no marked local inflammation, cholelithiasis probable cause of cholecystitis, PTHC.	*L. rhamnosus / L. casei / L. paracasei*	*L. rhamnosus* [ag]	Negative	O = iv PIT

25 M/76	CA, grade 2, calculous, no bile duct stone, no marked local inflammation, cholelithiasis probable cause of cholecystitis, PTHC.	*K. michiganensis* [b] *E. faecium* [b] *C. concisus* *M. morganii* *E. faecalis* [b] *Serratia marcescens*	*K. oxytoca* [c, ag] *M. morganii* [ag] *E. faecium* [ag]	Negative	O = iv PIT
26 M/85	CA, grade 2, acalculous, no bile duct stone, no marked local inflammation, unknown etiology, concomitant acute cholangitis, PTHC.	Negative	Negative	Negative	O = iv PIT
27 F/94	CA, grade 2, calculous, bile duct stone present, perforated cholecystitis, cholelithiasis probable cause of cholecystitis, PTHC.	*E. coli* [b]	*E. coli* [mg] *S. epidermidis* [e, mg]	Negative	O = iv CTA L = po CIP
28 F/53	CA, grade 2, calculous, no bile duct stone, no marked local inflammation, cholelithiasis probable cause of cholecystitis.	*H. parainfluenzae*	Negative	Not taken	O = iv PIT and po MET and TRS
29 F/48	CA, grade 2, calculous, no bile duct stone, no marked local inflammation, cholelithiasis probable cause of cholecystitis, PTHC.	*E. coli* [b]	*E. coli* [sg]	Not taken	O = po MET, TRS and CIP.
30 M/60	CA, grade 3, calculous, no bile duct stone, no marked local inflammation, cholelithiasis probable cause of cholecystitis.	*S. anginosus* [b] *Salmonella enterica* [b]	Negative	Negative	O = iv PIT L = po MET, TRS and AMO.
31 M/77	CA, grade 3, calculous, no bile duct stone, no marked local inflammation, cholelithiasis probable cause of cholecystitis, PTHC.	*Streptococcus gordonii* [b]	*S. gordonii* [sg]	Not taken	O = iv PIT
32 F/87	CA, grade 2, calculous, bile duct stone present, perforated cholecystitis, cholelithiasis probable cause of cholecystitis, concomitant acute cholangitis, PTHC.	*C. perfringens* *L. rhamnosus / L. casei / L. paracasei* *E. coli* [b] *I. bartlettii*	*L. rhamnosus* [sg]	Not taken	O = iv PIT
33 M/89	CA, grade 2, calculous, bile duct stone present, no marked local inflammation, cholelithiasis probable cause of cholecystitis, PTHC.	*E. coli* [b] *E. faecalis* *C. perfringens*	*E. coli* [ag] *E. faecalis* [ag]	Not taken	O = iv PIT L = MET and TRS

34	HA, grade 2, calculous, bile duct stone present, no marked local inflammation, cholelithiasis probable cause of cholecystitis. Hospitalized over long time because of a spinal cord injury, PTHC.	*E. coli* [b]	*E. coli* [ag]	Negative	O = iv CTA and MET
35 F/72	CA, grade 2, calculous, bile duct stone present, perforated cholecystitis, cholelithiasis probable cause of cholecystitis, concomitant acute cholangitis, PTHC.	*S. anginosus*[b] *H. parainfluenzae*	Negative	Not taken	O = iv MET and CIP
36 F/89	CA, grade 3, calculous, bile duct stone present, no marked local inflammation, cholelithiasis probable cause of cholecystitis, concomitant acute cholangitis, PTHC.	*K. variicola*[b] *C. perfringens* *F. nucleatum* *S. sanguinis*[b]	*K. pneumoniae*[d, ag] *C. perfringens* [ag]	*C. perfringens* *K. pneumoniae*[d]	O = iv PIT L = MET and TRS

[a] It was not possible to compute the relative abundance of fungi compared to bacteria since they were identified in different PCR reactions. Fungi are therefore listed at the end of each list of microbes.

[b] *rpoB* sequencing provided identification at a higher taxonomic level than 16S rRNA gene sequencing.

[c] MALDI/TOF spectra database do not contain *K. michiganensis* or *K. quasipneumoniae.* These two species will most likely be reported as a *K. pneumoniae* when analyzed with MALDI-TOF.

[d] MALDI/TOF spectra of *K. pneumoniae* isolate were reanalyzed with an updated database with *Klebsiella variicola* included. The isolate was re-classified as *Klebsiella variicola* according to MALDI/TOF results.

[e] *S. epidermidis* most likely represents contamination as it was found by culture only

Abbreviations: AMO amoxicillin, CTA cefotaxime, CIP ciprofloxacin, CLI clindamycin, IMI imipenem, MET metronidazol, PIT piperacillin/tazobactam, TRS trimetoprim/sulfamethoxazole, iv intravenous, po per oral.

[ag] abundant growth, [mg] medium abundant growth, [sg] sparse growth

Supplementary Table S5: Antibiogram of bacteria identified by bile or blood culture

ID	Bile culture	Blood culture	Antibiogram
1	*Escherichia coli*	Negative	AMP S(23), PIT S(28), CUR S(27), CTA S(32), CTZ S(27), GEN S(22), CIP S(29), TRS S(30), MER S(36)
2	*Streptococcus salivarius*	Negative	BEN S(0,125), CLI S(0,125)
3	Negative	*Candida albicans*	AMB S(0,125), FLC S(0,064), AFG S(0,016)
4	*E. coli*	*Negative*	AMP S(19), PIT S(25), CUR S(22), CTA S(27), CTZ S(24), GEN S(21), CIP S(34), TRS S(28), MER S(33)
5	Negative	Negative	
6	Negative	Negative	
7	*E. coli*	Negative	Not available
	Klebsiella oxytoca / Raoultella ornithinolytica		Not available
8	*Klebsiella pneumoniae*	Negative	**AMP R(6)**, PIT S(22), CUR S(19), CTA S(25), CTZ S(25), GEN S(20), CIP S(23), TRS S(16), MER S(29)
9	*Citrobacter species*	Negative	**AMP R(18)**, PIT S(28), CUR S(25), CTA S(28), CTZ S(25), GEN S(21), CIP S(37), TRS R(6), MER S(37)
	K. oxytoca		**AMP R(6)**, PIT I(19), CUR R(11), CTA S(21), CTZ S(23), GEN S(18), CIP S(31), TRS R(6), MER S(34)
	K. pneumoniae		**AMP R(6)**, PIT R(14), CUR S(22), CTA S(27), CTZ S(22), GEN S(17), CIP S(29), TRS R(6), MER S(30)
	Pseudomonas aeruginosa		PIT S(4), CTZ S(2), AZT I(4), MER S(0,25), TOB S(0,5), CIP R(2)
	Enterococcus raffinosus		AMP R(6), VAN S, LIN S(27)
10	Negative	Not taken	
11	*C. albicans*	Negative	Not available
12	*E. coli*	Negative	AMP S(17), PIT S(24), CUR S(23), CTA S(28), CTZ S(26), GEN S(20), CIP S(31), TRS S(28), MER S(31)
	Streptococcus massiliensis		Not available
13	Negative	Negative	
14	Negative	Not taken	
15	*K. oxytoca*	Not taken	**AMP R(6)**, PIT R(23), CUR S(20), CTA S(27), CTZ S(27), GEN S(20), CIP S(34), TRS S(28), MER S(30)
16	*S. salivarius*	Not taken	BEN S(0,125), CLI S(0,032)
17	*K. pneumoniae*		**AMP R(6)**, PIT S(21), CUR R(14), CTA S(26), CTZ S(26), GEN S(21), CIP S(29), TRS S(26), MER S(31)
		K. pneumoniae	**AMP R(10)**, PIT S(26), CUR S(27), CTA S(32), CTZ S(28), GEN S(25), CIP S(31), TRS S(29), MER S(31)
18	*E. coli*	Negative	AMP S(18), PIT S(25), CUR S(22), CTA (S(28), CTZ S(27), GEN S(18), CIP S(33), TRS S(20), MER S(35)

19	*K. pneumoniae*[c]		**AMP R(16)**, PIT S(27), CUR S(26), CTA S(29), CTZ S(26), GEN S(19), CIP S(38), TRS S(26), MER S(31)
	Enterococcus faecalis		AMP S(14), VAN S, LIN S(23)
		K. pneumoniae[c]	**AMP R(16)**, PIT S(26), CUR S(27) CTA S(31), CTZ S(28), GEN S(19), CIP S(34), TRS S(28), MER S(30)
20	*Clostridium perfringens*	Not taken	BEN S(0,064), PIT S(0,064), CLI S(1), MET S(4)
21	*Enterobacter cloacae complex*	Negative	**AMP R(6)**, PIT I(18), CUR R(10), CTA R(16), CTZ R(18), GEN S(22), CIP S(32), TRS S(28), MER S(29)
22	*Lactobacillus paracasei*		Not available
	Streptococcus anginosus		Not available
	Streptococcous sanguinis		BEN S(0,016), CLI S(0,016)
		S. anginosus	BEN S(0,032), CUR S(0,064), CTA S(0,064), CLI S(0,032)
23	*K. pneumoniae*		**AMP R(9)**, PIT S(22), CUR S(27), CTA S(28), CTZ S(30), GEN S(18), CIP S(29), TRS S(24), MER S(27)
	Proteus vulgaris		**AMP R(24)**, PIT S(31), CUR S(27), CTA S(31), CTZ S(30), GEN S(23), CIP S(37), TRS S(28), MER S(30)
	Enterococcus faecium		AMP S(18), VAN S, LIN S(25)
		K. pneumoniae	**AMP R**, PIT S(23), CUR S(24), CTA S(26), CTZ S(25), GEN S(22), CIP S(30), TRS S(21), MER S(26)
		E. faecium	AMP S(18), IMI S(21), GEN S(23), TIG S(24) VAN S, LIN S(24)
24	*L. rhamnosus*	Negative	Not available
25	*K. oxytoca*	Negative	**AMP R(9)**, PIT S(25), CUR S(26), CTA S(31), CTZ S(30), GEN S(20), CIP S(33), TRS S(25), MER S(33)
	Morganella morganii		**AMP R(6)**, PIT S(29), CUR S(22), CTA S(33), CTZ S(31), GEN S(22), CIP S(38), TRS S(27), MER S(32)
	E. faecium		AMP R(06) VAN S, LIN S(29)
26	Negative	Negative	
27	*E. coli*	Negative	**AMP R(6)**, PIT S(23), CUR S(21), CTA S(29), CTZ S(26), GEN S(19), CIP R(9), TRS R(6), MER S(35)
	Staphylococcus epidermidis[d]		Not available
28	Negative	Not taken	
29	*E. coli*	Not taken	AMP S(17), PIT S(24), CUR S(21), CTA S(27), CTZ S(25), GEN S(19), CIP S(31), TRS S(29), MER S(31)
30	Negative	Negative	
31	*Streptococcus gordonii*	Not taken	BEN S(0,008), CLI S(0,032)
32	*L. rhamnosus*	Not taken	Not available
33	*E. coli*	Not taken	AMP S(20), PIT S(26), CUR S(23), CTA S(30), CTZ S(27), GEN S(21), CIP S(37), TRS S(33), MER S(35)
	E. faecalis		AMP S(17), VAN S, LIN S(23)

34	*E. Coli*	Negative	AMP S(18), PIT S(22), CUR S(21), CTA S(27), CTZ S(25), GEN S(19), CIP S(30), TRS S(26), MER S(31)
35	Negative	Not taken	
36	*K. pneumoniae*[c]		**AMP R(6)**, PIT S(22), CUR S(24), CTA S(28), CTZ S(26), GEN S(20), CIP S(29), TRS S(25), MER S(30)
	C. perfringens		BEN S(0,064), PIT S(0,064), CLI S(0,5), MET S(0,5)
		K. pneumoniae[c]	**AMP R(8)**, PIT S(22), CUR S(24), CTA S(29), CTZ S(25), GEN S(20), CIP S(29), TRS S(24), MER S(32)
		C. perfringens	BEN S(0,064), PIT S(0,064), CLI S(0,5), MET S(0,5), MER S(0,016)

Abbreviations: AFG anidulafungin, AMB amphotericin B, AMO amoxicillin, AMP ampicillin, AZT aztreonam, BEN benzylpenicillin, CTA cefotaxime, CTZ ceftazidime, CUR cefuroxime, CIP ciprofloxacin, CLI clindamycin, FLC fluconazol, GEN gentamicin, IMI imipenem, LIN linezolid, MER meropenem, MET metronidazol, PIT piperacillin/tazobactam, TOB tobramycin, TRS trimetoprim/sulfamethoxazole, VAN vancomycin

III

AMERICAN SOCIETY FOR MICROBIOLOGY | mBio®

RESEARCH ARTICLE

Managing Contamination and Diverse Bacterial Loads in 16S rRNA Deep Sequencing of Clinical Samples: Implications of the Law of Small Numbers

Ruben Dyrhovden,[a] Martin Rippin,[b*] Kjell Kåre Øvrebø,[c,d] Randi M. Nygaard,[a] Elling Ulvestad,[a,d] Øyvind Kommedal[a,d]

[a]Department of Microbiology, Haukeland University Hospital, Bergen, Norway
[b]Section for Bioinformatics, Haukeland University Hospital, Bergen, Norway
[c]Department of Surgery, Haukeland University Hospital, Bergen, Norway
[d]Department of Clinical Science, University of Bergen, Bergen, Norway

ABSTRACT In this article, we investigate patterns of microbial DNA contamination in targeted 16S rRNA amplicon sequencing (16S deep sequencing) and demonstrate how this can be used to filter background bacterial DNA in diagnostic microbiology. We also investigate the importance of sequencing depth. We first determined the patterns of contamination by performing repeat 16S deep sequencing of negative and positive extraction controls. This process identified a few bacterial species dominating across all replicates but also a high intersample variability among low abundance contaminant species in replicates split before PCR amplification. Replicates split after PCR amplification yielded almost identical sequencing results. On the basis of these observations, we suggest using the abundance of the most dominant contaminant species to define a threshold in each clinical sample from where identifications with lower abundances possibly represent contamination. We evaluated this approach by sequencing of a diluted, staggered mock community and of bile samples from 41 patients with acute cholangitis and noninfectious bile duct stenosis. All clinical samples were sequenced twice using different sequencing depths. We were able to demonstrate the following: (i) The high intersample variability between sequencing replicates is caused by events occurring before or during the PCR amplification step. (ii) Knowledge about the most dominant contaminant species can be used to establish sample-specific cutoffs for reliable identifications. (iii) Below the level of the most abundant contaminant, it rapidly becomes very demanding to reliably discriminate between background and true findings. (iv) Adequate sequencing depth can be claimed only when the analysis also picks up background contamination.

IMPORTANCE There has been a gradual increase in 16S deep sequencing studies on infectious disease materials. Management of bacterial DNA contamination is a major challenge in such diagnostics, particularly in low biomass samples. Reporting a contaminant species as a relevant pathogen may cause unnecessary antibiotic treatment or even falsely classify a noninfectious condition as a bacterial infection. Yet, there are few studies on how to filter contamination in clinical microbiology. Here, we demonstrate that sequencing of extraction controls will not reveal the full spectrum of contaminants that could occur in the associated clinical samples. Only the most abundant contaminant species were consistently detected, and we present how this can be used to set sample specific thresholds for reliable identifications. We believe this work can facilitate the implementation of 16S deep sequencing in diagnostic laboratories. The new data we provide on the patterns of microbial DNA contamination is also important for microbiome research.

Citation Dyrhovden R, Rippin M, Øvrebø KK, Nygaard RM, Ulvestad E, Kommedal Ø. 2021. Managing contamination and diverse bacterial loads in 16S rRNA deep sequencing of clinical samples: implications of the law of small numbers. mBio 12:e00598-21. https://doi.org/10.1128/mBio.00598-21.

Editor Julian Parkhill, Department of Veterinary Medicine

Address correspondence to Ruben Dyrhovden, ruben.dyrhovden@helse-bergen.no.

* Present address: Martin Rippin, Department of Immunology, Genetics and Pathology, Rudbecklaboratoriet, Uppsala University, Uppsala, Sweden.

Received 4 March 2021
Accepted 22 April 2021
Published 8 June 2021

KEYWORDS 16S rRNA, acute cholangitis, contamination, NGS, targeted amplicon sequencing, *rpoB*

Microbial DNA contamination from extraction kits and other PCR and sequencing reagents (background bacterial DNA) is a major challenge in 16S rRNA amplicon sequencing (16S deep sequencing) of polymicrobial infections (1, 2). Several studies have demonstrated the risk for erroneously interpreting contaminating DNA as bacteria originating from the sample (1, 3–5). In clinical microbiology, reporting a contaminant species as a relevant pathogen may cause unnecessary antibiotic treatment or even falsely classify a noninfectious condition as a bacterial infection. Unfortunately, many of the studies on infectious disease materials do not address background contamination (6–8), and among those that do, the approaches vary. The most used method is to sequence extraction controls along with the samples and remove those bacteria from the sample reports which were also found in the controls (9–12). However, the sensitivity of this method is reduced if bacteria truly present in clinical samples are also present in the negative controls (2). Further, the specificity of this approach relies on the assumption that sequencing of the negative controls provides an exhaustive identification of contaminants.

Within microbiota research, a range of methods have been developed to diminish the problem of background contamination (1, 2, 4, 13, 14), but many of these approaches are not easily transferable to diagnostic laboratories. Despite the common aim of describing bacterial flora, clinical microbiologists and microbiota researchers have partly divergent challenges and goals. In microbiota research, typically large sets of the same sample type are analyzed in multiple batches over a limited period. Combined with extensive use of negative and positive controls, and even multiple sequencing techniques (14), this allows labs to use pattern recognition and statistical calculations to filter their data sets (4). Although they make considerable effort to ensure the overall quality of a data set, there is less focus on the individual sample, and identifications are usually limited to the genus level or above. In clinical microbiology, there is a broad spectrum of sample types with highly divergent bacterial concentrations and compositions, and background contamination will vary over time with different batches of reagents and consumables. The focus is always the individual patient and species level identification is normally required. Finally, time to results and cost are crucial matters, limiting the room for extensive assessments of background contamination.

Accurate filtering of background contamination is more critical in weakly positive samples, where it constitutes a larger portion of the total bacterial DNA (1, 2). Sequencing depth is another essential factor, in particular for strongly positive, polymicrobial samples where the use of too few reads may result in failure to detect low abundance species. In clinical microbiology, especially in samples from normally sterile body sites, the detection of a bacterium at any concentration is *a priori* considered potentially relevant. A sample from a polymicrobial infection must be considered a snapshot of a potentially dynamic process, and species present at low abundances in the sample can flourish at the site of infection at a later stage, especially if antibiotic treatment is directed only against the dominant flora. Also, the relative microbial abundances in a sample cannot be assumed to be representative of the entire site of infection. For example, the abundance of a given species in pus aspirated from the necrotic, anaerobic center of an abscess is not necessarily representative of the abundance of the same species in the more oxygenated periphery on the transition to intact tissue. Despite these issues, there has been little attention to the relationship between sequencing depth and sensitivity.

In this study, we aim to describe and evaluate simple and transparent approaches for dealing with contamination in 16S deep sequencing in clinical microbiology. We base our suggestions on the observation that the presence of a few dominant contaminant species is highly consistent across all controls, while in the same controls the presence of less dominant contaminant species seems to vary (15, 16). We use these

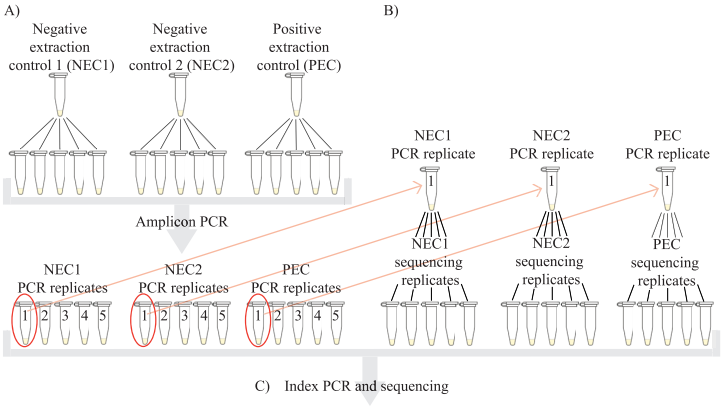


FIG 1 Illustration of workflow for PCR amplification and sequencing of the three extraction controls. (A) All three samples were split into five replicates before 16S rRNA amplicon PCR, resulting in five PCR replicates from each original sample after the PCR. (B) From each of the three groups of PCR replicates, one of the five replicates was then split into five new replicates. (C) Index PCR and sequencing were than performed for both PCR replicates and sequencing replicates on the same sequencing run. One PEC sequencing replicate was lost due to technicalities, leaving 15 PCR replicates and 14 sequencing replicates eligible for postsequencing analysis.

most abundant contaminant species and their abundances in the corresponding clinical samples to set sample-specific cutoffs for the number of reads needed to reliably classify a species as a noncontaminant. We perform repeat sequencing of a set of extraction controls, both before and after the 16S rRNA PCR amplification step, to underpin our approach and to demonstrate sensitivity limitations that remain even in deep sequencing. We further test the approach on a diluted, standardized staggered mock community and on prospectively collected bile samples from patients with acute cholangitis or noninfectious bile duct stenosis. To demonstrate the importance of sequencing depth, all patient samples were sequenced twice with different sequencing depths in each replicate.

RESULTS

Experiment 1. Repeat sequencing of extraction controls. (i) Experimental design. We first sought to understand the mechanisms behind the observed phenomenon that the presence of a few dominant contaminant species is highly consistent across all controls, while the presence of less dominant contaminant species seems to vary (15, 16). To investigate this, we analyzed a set of extraction controls in a separate sequencing run (Fig. 1). Three different samples were analyzed, two negative extraction controls (NEC1 and NEC2), consisting of PCR-grade water and lysis buffer, and one weakly positive extraction control (PEC) containing *Legionella pneumophila*. To isolate the impact of the PCR amplification of the sample template (amplicon PCR) from the impact of the following index PCR and sequencing procedure, each of the three controls was split into five replicates before the amplicon PCR (hereafter named "PCR replicates"). One PCR replicate from each of the three controls was further split into five replicates before sequencing (hereafter named "sequencing replicates"). All PCR and sequencing replicates were then indexed and sequenced in the same run. We used the results from this part of the study to formulate criteria for filtration of sequencing data from clinical samples, which we further evaluated on a staggered mock community and a collection of bile samples.

(ii) Results. The Venn diagrams in Fig. 2 illustrate the higher diversity between PCR replicates compared to the sequencing replicates. Among the five PCR replicates of one sample, most species were present in only a single replicate. Only four (NEC1 and

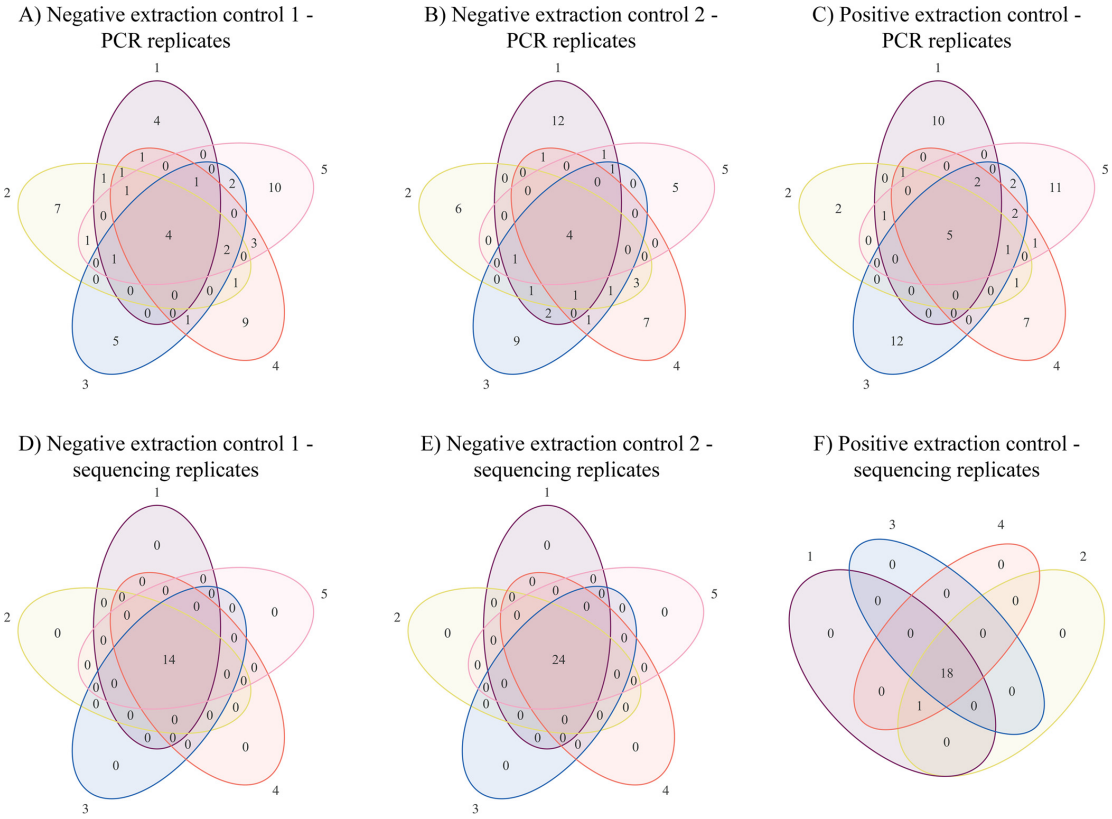
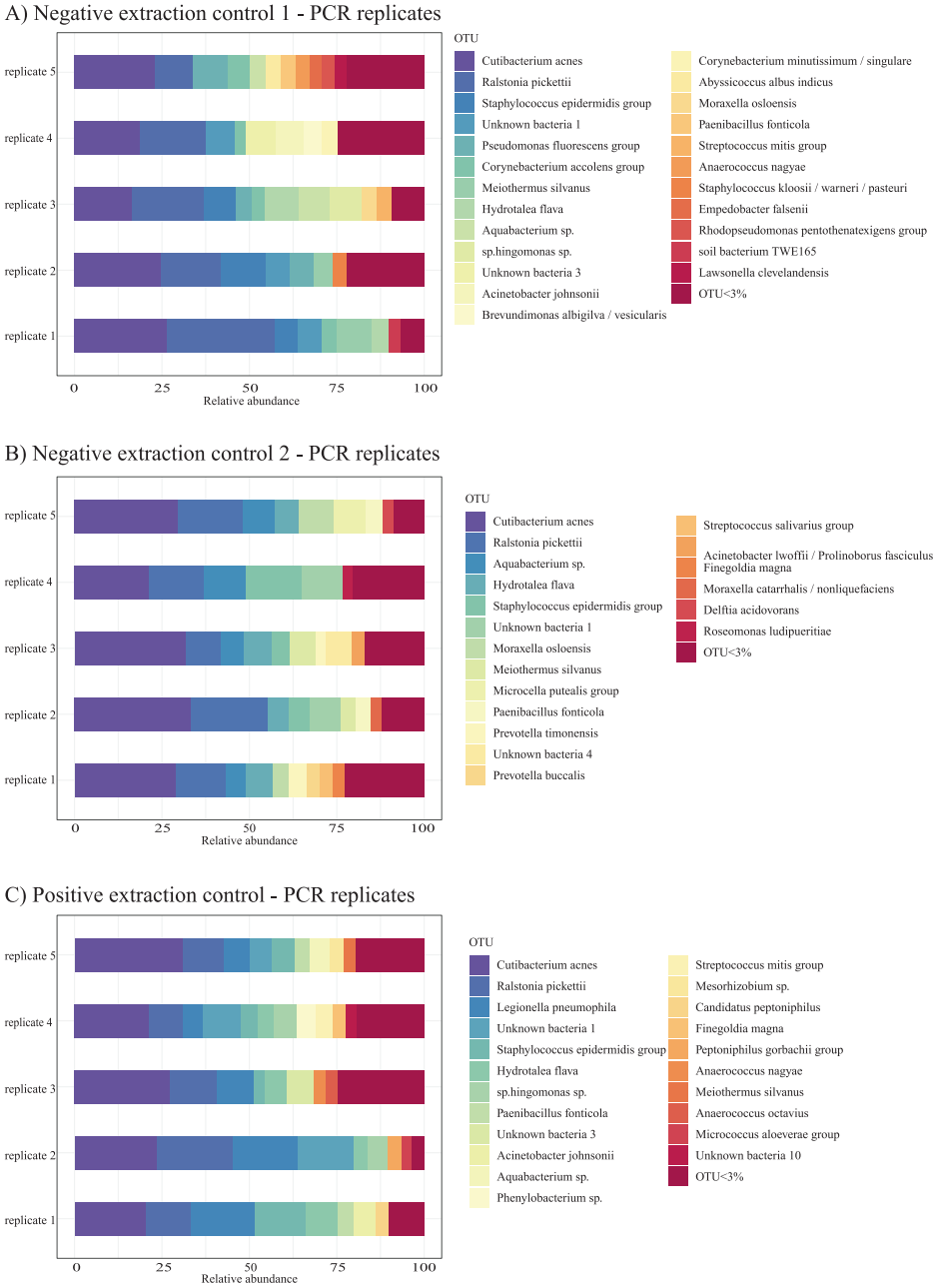


FIG 2 Venn diagram of bacterial identifications in PCR replicates (A to C) and sequencing replicates (D to F), showing a high diversity between the PCR replicates originating from the same sample, in contrast to the high similarity between the sequencing replicates originating from the same PCR. (A) Fifty-five different bacteria were found in all five NEC1 PCR replicates combined. Only 4 (7%) identifications were shared between all five replicates, while 35 (64%) were found in only 1 of the 5 replicates. (B) Fifty-six different bacteria were found in all NEC2 PCR replicates in total. Only 4 (7%) identifications were detected in all five replicates, and 39 (70%) identifications were found in only a single replicate. (C) Fifty-eight different bacteria were found in all PEC PCR replicates in total. Five (9%) bacterial identifications were shared across all five replicates, including *Legionella pneumophila*, which was the positive control. Forty-two (72%) identifications were detected in only a single replicate. (D and E) Bacterial identifications were identical in all sequencing replicates of both NEC1 and NEC2. (F) Bacterial identifications were identical in all sequencing replicates of PEC except for one *Phenylobacterium* species which was identified in only three out of the four replicates.

NEC2 PCR replicates) or five (PEC PCR replicates) species were found in all five replicates. In contrast, there were no differences between sequencing replicates originating from the same PCR except for a single species missing in one replicate. This bacterium, a *Phenylobacterium* sp., was also present in the latter replicate but by less than 50 reads and consequently below our cutoff for a valid identification.

A few species dominated all replicates, whereas the majority of contaminants appeared at relatively low abundances. The most abundant bacterium in each replicate was represented by 19 to 33% of the total number of valid reads (Fig. 3). In all replicates, *Ralstonia pickettii* and *Cutibacterium acnes* were the two most dominant species. These were also the only bacteria present in all replicates from all groups.

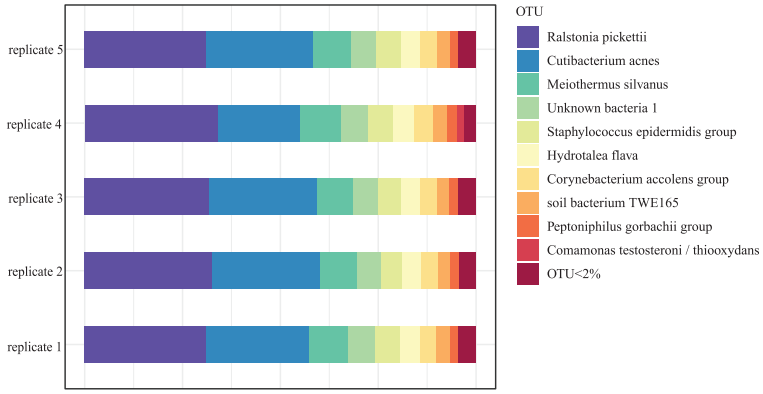
In Fig. 4, we have defined a "frequency threshold rate" (FTR) as a percentage of the most dominant contaminant bacteria in each replicate measured in the number of sequencing reads. For example, if the most abundant contaminant bacterium is present in 10,000 reads, then the 20% FTR is 2,000 reads. Figure 4 shows a steep decrease in the number of bacterial identifications with an abundance above the FTR as the rate increases from 0% to 50%, from >55 bacteria to ≤ 5 bacteria in each group of



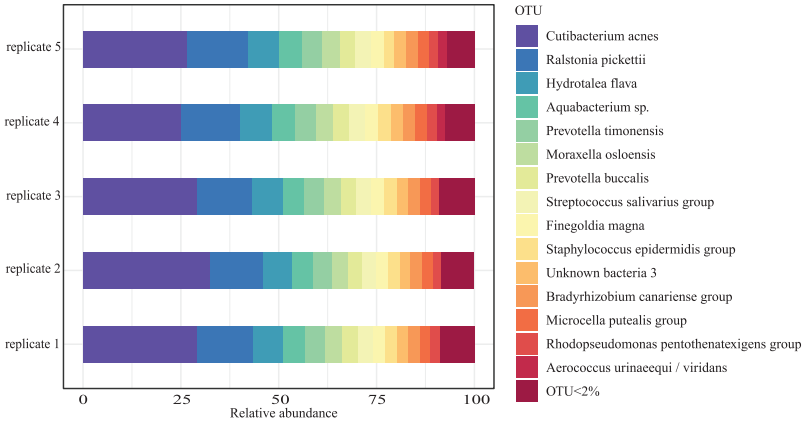
AQ:fig

FIG 3 Bacterial composition in PCR (A to C) and sequencing (D to F) replicates of the extraction controls. A few species dominated in all replicates. (A to C) The replicates from each sample show a high degree of variability. Only two species were present in every PCR replicate from all three samples: *Ralstonia picketti* and *Cutibacterium acnes*. These two species were also the most abundant species in all PCR replicates. (D to F) Bacterial identifications were identical in all sequencing replicates of NEC1 and NEC2, while one *Phenylobacterium* species was identified in only three out of the four replicates of PEC. In those three replicates containing *Phenylobacterium*, it appeared with the lowest number of reads (105, 51, and 90) and relative abundance (0.03%, 0.02%, and 0.04% of total number of reads, respectively) of all bacterial identifications.

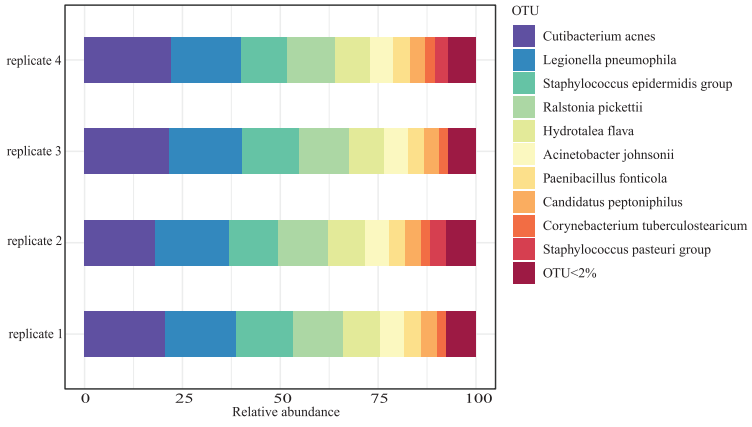
D) Negative extraction control 1 - Sequencing replicates



E) Negative extraction control 2 - Sequencing replicates



F) Positive extraction control - Sequencing replicates



AQ:fig

FIG 3 (Continued)

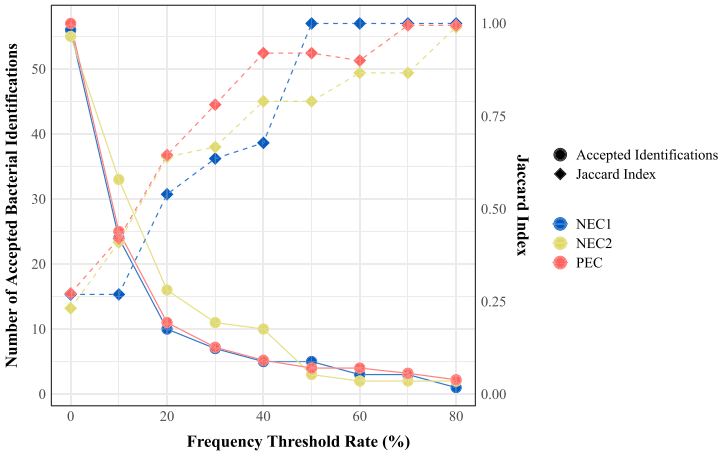


FIG 4 Graph showing the correlation between a chosen frequency threshold rate and the resulting number of accepted bacterial identifications and similarity between PCR replicates. The *x* axis shows the frequency threshold rate (FTR) calculated as a percentage of the most dominant contaminant bacteria in each replicate measured in the number of sequencing reads. The left *y* axis shows the total number of accepted bacterial species for all five PCR replicates for each control when only bacteria represented by more reads than the chosen FTR were accepted. The right *y* axis shows the mean sample to sample Jaccard index of the five PCR replicates when only bacteria represented by more reads than the chosen FTR cutoff were accepted.

replicates. The plot also shows the correlation between the FTR and the similarity between the PCR replicates. The chance for the same bacterial species to be found in all PCR replicates increases as the relative abundance of that bacterium increases. For NEC1, bacteria with abundances above 50% of the dominant contaminant were present in all replicates. The corresponding thresholds for NEC2 and PEC were 80% and 70%, respectively.

On the basis of these findings, we suggest the following criteria for filtering bacterial contaminants in clinical samples. (i) Any bacterium appearing with a higher abundance than the top five abundant contaminants, as determined by the sequencing of negative and positive extraction controls, is accepted as a valid identification, even if it occurs as a low abundance species in the controls. (ii) Bacteria present in frequencies between 20% and 100% of the most abundant contaminant are accepted as likely valid identifications, but only if they are also absent from all the negative controls. (iii) Bacteria present in frequencies below 20% of the most abundant contaminant are always rejected as invalid. (iv) In samples where none of the top five abundant contaminants are detected, all identifications are accepted as valid.

Detailed data from these experiments, including technical sequencing results and sample diversity measures, is provided in Table S1 of the supplemental material. Operational taxonomy unit (OTU) lists for all extraction control replicates can be found in Table S2.

Experiment 2. Sequencing of a staggered mock community. (i) Experimental design. Our next experiment included sequencing of a staggered mock community together with negative and positive extraction controls. The aims of this experiment were twofold: (i) to assess the actual abundance of the contaminants detected in our negative controls and to determine at what level the observed high variability in PCR replicates occur and (ii) to assess the sensitivity and specificity of our suggested criteria for filtering bacterial contaminants and to compare it to other common methods for contaminant filtering.

We performed deep sequencing of three different dilutions of the staggered mock community: a 1:10 dilution, representing a high bacterial load sample (16S PCR

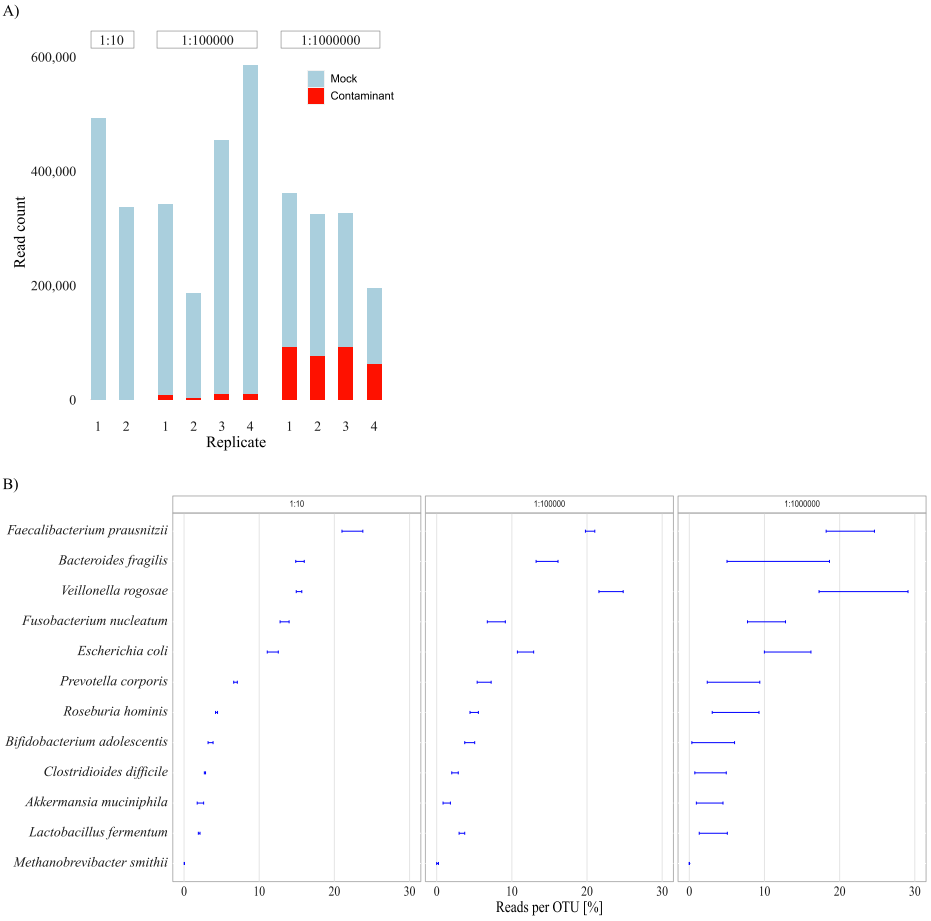


FIG 5 Analysis of mock community dilutions. (A) Number of reads per sample and distribution of reads from mock community and DNA contaminants. The absolute and relative amount of reads from DNA contaminants increases with the subsequent dilutions. (B) Identified mock microbes in each of the three dilutions investigated, and the variation (range) in relative abundance of each identified bacteria between the different PCR replicates within each dilution. The species identified in the most diluted sample showed a higher variation in relative abundance between PCR replicates.

threshold cycle [C_T] value of 11.2), and a $1{:}10^5$ and a $1{:}10^6$ dilution, representing low bacterial load samples (16S PCR C_T values of 27.3 and 31.7, respectively). The theoretical composition of bacterial cells and the estimated 16S rRNA copy counts in each of these dilutions is presented in Table S3. The 1:10 dilution was split into two PCR replicates (1:10_1 and 1:10_2) and the $1{:}10^5$ dilution and $1{:}10^6$ were split into four PCR replicates each ($1{:}10^5$_1 to $1{:}10^5$_4 and $1{:}10^6$_1 to $1{:}10^6$_4) before the PCR amplification step. A negative and a positive extraction control were split into five PCR replicates each before the PCR amplification step and sequenced together with the mock community samples.

(ii) Results. *(a) Mock community result variability*. The total number of accepted reads from the mock community samples after quality filtering was 3,606,622. Each sample had between 186,375 and 586,295 reads, and the relative proportion of DNA contaminants increased with subsequent dilutions (Fig. 5A). No contaminant microbes were identified in the high bacterial load sample (Fig. 5A). In the replicates from the $1{:}10^5$ dilution, contamination constituted 2 to 3% of the total number of reads,

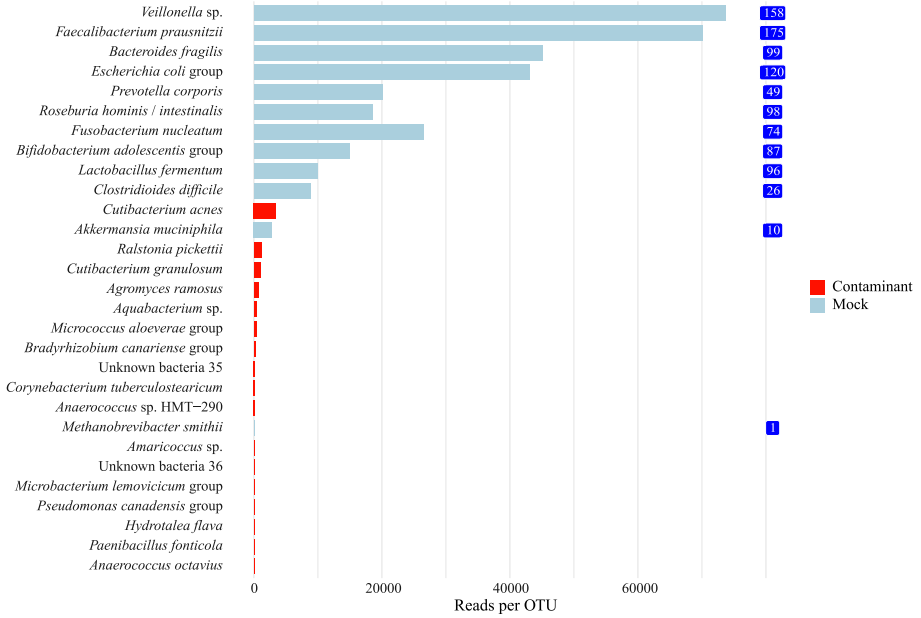
increasing to 24 to 32% in the 1:10^6 dilution replicates. Figure 5B shows the variation in relative abundance between the PCR replicates for each identified mock community bacterium and how the variability increases in the higher dilutions. Twelve of the 15 bacterial species present in the mock community were identified in each of the 1:10 diluted replicates, representing 100% of the identified OTUs in both samples. The three mock community species that were not detected were those with the lowest abundances, ranging from 0.00009% to 0.0065% of the total microbial content of the mock community (Table S3). In the 1:10^5 dilution, the same 12 bacterial species were identified in two out of four PCR replicates, while the least abundant of these species, *Methanobrevibacter smithii*, remained undetectable in two PCR replicates. In the 1:10^6 dilutions, the 11 most abundant mock community species were identified in all four PCR replicates. An OTU table for all mock community PCR replicates and extraction controls is provided in Table S4.

(b) Assessment of the abundance of laboratory contamination. The absolute abundances of all OTUs found in the first replicate of the 1:10^5 and 1:10^6 dilutions are presented in Fig. 6. Using the calculated concentration of mock community species as a reference, we see that the most dominating contaminants appeared at concentrations around 10 16S copies per 2 μl template, corresponding to about 500 cells per ml in the original sample (Table S3). The less abundant contaminants appeared in concentrations close to or less than a single 16S copy per 2 μl PCR template, approaching the lower limit of detection in the PCR. This corresponded to an initial concentration of up to 100 bacterial cells per ml sample (Table S3).

(c) Composition of the five negative and five positive extraction control replicates. The mean number of species identified in each of the extraction controls were 18 (range, 12 to 23) with *C. acnes* and *R. picketti* being the only species consistently detected in all negative and positive extraction control replicates. As in experiment 1, these two species were the most dominant contaminants in all replicates, and we observed the same high diversity between PCR replicates originating from the same extraction control. Eighty-three different species were found in the 10 replicates combined (Table S4). The mean Jaccard distance was 0.80 (range, 0.65 to 0.87) for the negative extraction control replicates and 0.76 (range, 0.65 to 0.81) for the positive extraction control replicates. Forty out of 58 species from the negative extraction controls were found in a single replicate only. The corresponding number for the positive extraction control replicates was 35 out of 52.

(d) Filtering contaminants based on our suggested criteria versus other common methods for contaminant filtering. For the first replicate of the 1:10^5 and 1:10^6 mock community dilutions, five different methods for removing contaminants were evaluated: (i) our suggested criteria, (ii) removing all OTUs found in one preselected negative and one preselected positive extraction control replicate, (iii) removing all OTUs found in all five negative extraction control PCR replicates and all five positive extraction control PCR replicates, and (iv and v) use of Decontam prevalence-based contaminant identification, including both the *isContaminant* and the *isNotContaminant* function which are both recommended for low biomass samples (4). Results are presented in Fig. 7. Filtering using our suggested criteria had a sensitivity and specificity for the identification of mock community bacteria in the two dilutions combined of 83% and 97% with an overall test accuracy of 93%. One out of 39 contaminants were wrongly classified as a mock community microbe, and four mock community microbes were wrongly classified as contamination (*M. smithii* in the 1:10^5 dilution, and *Bifidobacterium adolescentis*, *Clostridioides difficile*, and *Akkermansia muciniphila* in the 1:10^6 dilution). Filtering using a single preselected negative and positive extraction control gave a sensitivity of 100%, a specificity of 39%, and a test accuracy of 61%. Filtering using all 10 extraction controls had a sensitivity of 100%, a specificity of 64%, and a test accuracy of 77%. Filtering using Decontam *isContaminant* function had a sensitivity of 100%, a specificity of 39%, and a test accuracy of 61%. Filtering using Decontam *isNotContaminant* function increased specificity to 77%, giving a test accuracy of 86%.

A) Mock dilution 1:100000, replicate 1



B) Mock dilution 1:1000000, replicate 1

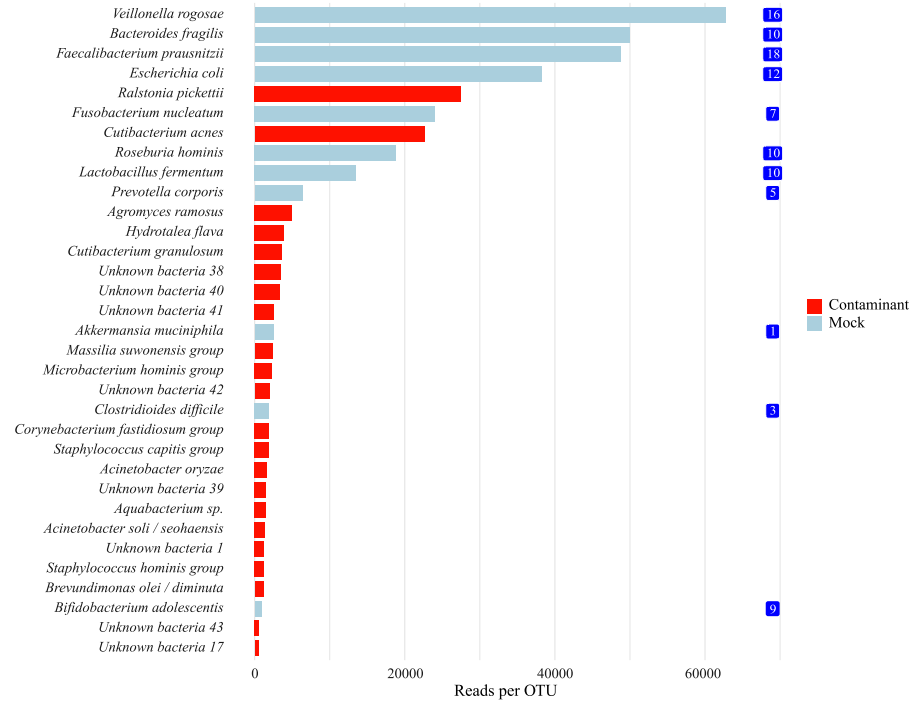


FIG 6 The abundance of all species found in the first replicate of the 1:10^5 and 1:10^6 dilutions. The theoretical numbers of 16S copies of each mock community species in 2 μl PCR template are shown as white numbers on blue rectangles. The most

(Continued on next page)

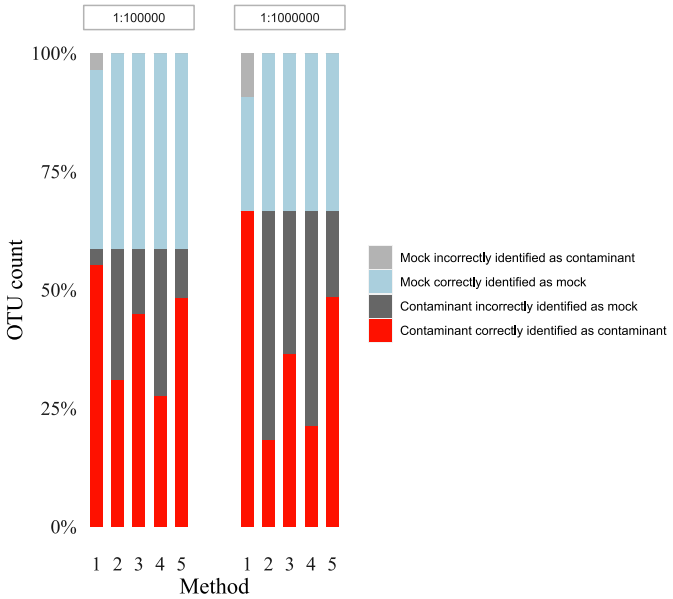


FIG 7 Comparison of five different methods for filtering DNA contaminants for a $1{:}10^5$ and $1{:}10^6$ dilution replicate of the mock community. Method 1 is our suggested method. Method 2 is filter all OTUs found in one NEC and PEC. Method 3 is filter OTUs found in all 10 extraction controls. Method 4 is Decontam prevalence based isContamintant function. Method 5 is Decontam prevalence-based *isNotContaminant* function.

Sequencing of bile samples from patients with acute cholangitis and bile duct stenosis. Forty-one patients with either acute cholangitis ($n = 15$) or noninfectious bile duct stenosis caused by bile duct stones ($n = 26$) were analyzed. Patient characteristics together with culture and sequencing results are summarized in Table 1. Bacterial loads were categorized as high in 15 samples, moderate in 8, and low in 18 (Table 2). Each sample was split into two replicates before 16S rRNA sequencing and were sequenced using different sequencing depths (16S rRNA replicate 1 and 16S rRNA replicate 2). Table 3 gives an overview of technical sequencing results for the two sets of replicates. Five sequencing runs were performed to include all samples.

(a) Identifying and filtering bacterial contaminants in the 16S rRNA replicates. The combined number of extraction controls analyzed in all clinical sequencing runs were 18. An OTU table for all these is provided as Table S5. The top five abundant species in each extraction control in each of the sequencing runs were identified. If any of these were found in a clinical sample, the most abundant of them defined a level from where contamination could be expected to occur in that sample and were used to filter contaminants as described for experiment 1. Based on this, OTUs were categorized as either valid, likely valid, or contaminant. One or more of the most abundant contaminant species from the controls were identified in 22 and 24 of the 41 samples in the two 16S PCR replicate runs, respectively (Table 2). As shown in Table 2, detection of contaminant bacteria was inversely correlated with the bacterial load of the samples.

FIG 6 Legend (Continued)
dominating contaminants were found in the same concentration as mock microbes with a theoretical concentration of approximately 10 16S copies per 2 μl PCR template. This corresponds to approximately 100 cells per ml in the original sample, or about 500 16S copies per 100 μl extracted DNA. The less abundant contaminants appear in the same concentration as mock microbes with a theoretical concentration close to or less than only a single 16S copy per 2 μl PCR template. This corresponds to an initial concentration of less than 100 bacterial cells per ml sample.

TABLE 1 General characteristics, culture, and sequencing results of all included patients

Characteristic, culture, or sequencing result	No. (%) of patients or indicated value for characteristic or result	
	Acute cholangitis	Noninfectious bile duct stenosis
No. of patients	15	26
General characteristics		
Male	9 (60)	7 (27)
Mean age, yrs	73	54
SD; median; range	10; 71; 58–95	17; 51; 20–83
Previous biliary interventions	5 (33)	10 (39)
ERCP with papillotomy	4 (27)	3 (12)
ERCP without papillotomy	0	1 (4)
Choledocus stent still in place	1 (7)	0
Choledocus stent removed	0	1 (4)
Cholecystectomy	0	7 (27)
Ongoing antibiotic therapy at time of sampling	14 (93)	2 (8)
Concomitant acute pancreatitis	1 (7)	0
Concomitant acute cholecystitis	3 (20)	0
Culture and sequencing results		
C_T value[a] for sample, mean	19.8	25,6
SD; median; range	5,1; 19.2; 12.5–27.9	7.5; 28.4; 12.2–33.4
C_T value[a] for NEC[b], mean	32.7	33.4
SD; median; range	1.3; 32.9; 29.7–34.6	1.1; 33.7; 31.3–35.5
Growth in blood culture (of tested)	4 (7)	
Samples with detected bacteria by sequencing[c]	15 (100)	21 (81)
Samples with growth of bacteria in bile culture	14 (93)	17 (65)
Polybacterial samples by culture	8 (53)	9 (35
Polybacterial samples by sequencing[c]	14 (93)	15 (58%)
Mean species richness by sequencing[c]	5,7	8,1
SD; median; range	3.1; 6.0; 1–13	12.8; 3.0; 0–59
Mean species richness by culture	2.1	1.7
SD; median; range	1.4; 2,0; 1–6	1.8; 1,0; 0–5

[a]Cycle threshold of SYBR green real-time 16S rRNA PCR.
[b]Negative extraction control.
[c]All valid and likely valid identifications included, irrespective of whether they were identified in only one or in both of the two 16S rRNA replicates.

(b) Comparison of 16S rRNA PCR replicates from the clinical samples. The total number of accepted identifications for all samples in 16S replicate 1 was 209 (173 valid and 36 likely valid). The corresponding number for replicate 2 was 295 (239 valid and 56 likely valid). The mean species richness was significantly higher in replicate 2 (Table 3). Figure S1 in the supplemental material shows a prevalence bar chart per sample for 16S rRNA replicate 2, categorized according to our filtering criteria. Verifications by other methods (culture, corresponding 16S rRNA replicate, or *rpoB* sequencing) are also indicated in the figure. An OTU table for all samples in both replicates is provided in Table S6.

Discrepancies between the two replicates were observed for 22 (53.7%) of the 41 samples (Fig. 8). Ninety-four bacterial identifications, from now on called singletons,

TABLE 2 Samples where contaminant bacteria were identified, categorized by the bacterial load of the sample

Bacterial load of sample	No. of samples where contaminant bacteria were identified	
	Replicate 1	Replicate 2
High ($n = 15$)	0	0
Moderate ($n = 8$)	4	6
Low ($n = 18$)	18	18
Total ($n = 41$)	22	24

TABLE 3 Overview of the two 16S rRNA sequencing replicates of the 41 bile samples

Characteristic(s)	16S rRNA replicate 1	16S rRNA replicate 2	*P* value[c]
Valid reads,[a] mean (median)	67,608 (50,999)	229,904 (198,331)	<0.001
Range	16,179–221,713	52,068–583,052	
Accepted reads when identified contaminants excluded, mean (median)	53,289 (33,192)	185,522 (166,919)	<0.001
Range	0–221,713	0–583,052	
Total no. of bacterial identifications[b]	208	291	
Mean no. of bacterial identifications per sample[b] (median)	5.1 (3.0)	7.2 (4.0)	<0.001
Range	0–26	0–59	

[a]Accepted reads per sample after quality filtering.
[b]After exclusion of identified contaminants.
[c]Student's *t* test for continuous, normal distributed variables. Mann-Whitney U-test for continuous, skewed variables.

were made in only one replicate (Fig. 8 and Table S6). As expected, most singleton findings were found in the replicate with the highest sequencing depth (Fig. 8), and all singleton findings were among bacteria with either a low relative abundance or from samples with a low bacterial load (Table S6).

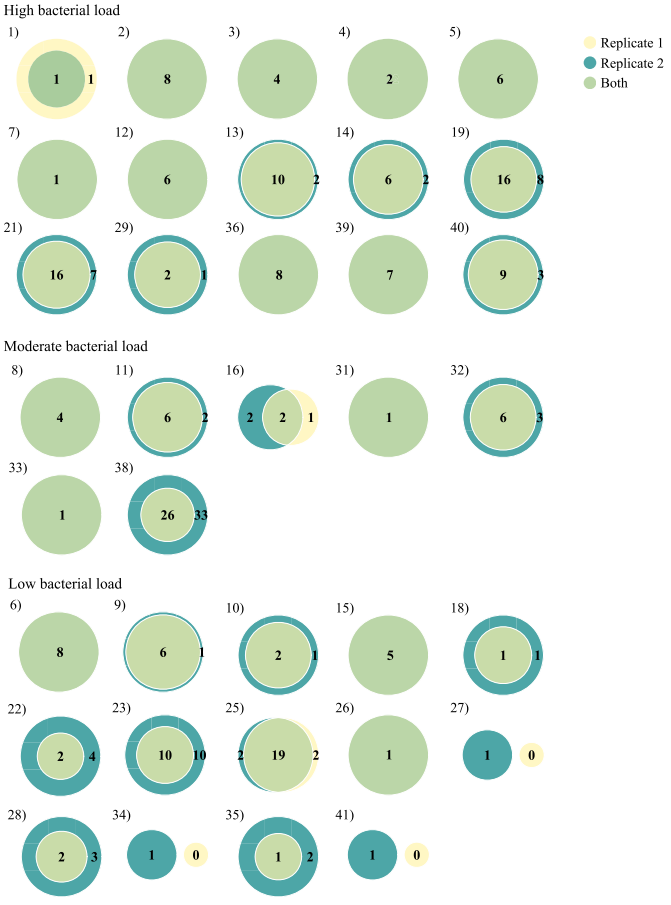


FIG 8 Venn diagram for each clinical sample, comparing the bacterial findings in the two sequencing replicates. There were discrepancies in bacterial findings between the two replicates for 22 (53.7%) of the 41 samples. Ninety-four bacterial identifications were made in only one of the two replicates. Out of these, 90 (96%) were found in the sample with the highest sequencing depth.

(c) Microbial findings in bile samples. When summarizing the microbial findings, we included all valid and likely valid identifications from both replicates. A summary of all bacterial findings, grouped at the genus level, is presented in Table 4.

All samples from patients with acute cholangitis contained bacteria as determined by both culture and sequencing (Table 1 and Fig. S1). For patients with noninfectious bile duct stenosis, 21 out of 26 (81%) samples contained bacteria as determined by sequencing. Among these, four samples were culture negative.

Compared to culture, sequencing found a much higher species richness in most samples (Tables 1 and 4). In the acute cholangitis group, 84 microbial identifications were made by sequencing whereof only 26 (30%) were cultured (Table 4). One identification, *Granulicatella adiacens*, was made solely by culture. In the group of noninfectious bile duct stone patients, 215 identifications were made by sequencing, whereof only 40 (19%) were cultured. Four unique identifications, one *Staphylococcus epidermidis*, one *Staphylococcus warneri*, one *Corynebacterium pseudodiphtericum*, and one *Enterococcus faecalis* were made by culture only.

DISCUSSION

In this study, we investigate patterns of microbial contamination in targeted amplicon sequencing and their implications for postsequencing filtering of results. We demonstrate how the most dominant contaminant species can be used to establish sample-specific cutoffs for reliable identifications. We also show how sample bacterial load and sequencing depth affect sequencing results.

Sequencing of negative controls does not reveal all contaminants. Most current approaches for identifying and filtering contaminant bacteria rely on the assumption that sequencing of appropriate extraction controls will reveal the full spectrum of background contaminants that could possibly occur in the associated clinical samples (1–4). Our results contradict this assumption. We found that less than 10% of the contaminant species were detectable in all five replicates from the same negative control when split before the PCR amplification step (Fig. 2).

Recently, Erb-Downward et al. described the same PCR replicate variability (17). They suggested that the phenomenon occurred because of sequencing errors, possibly due to underloading of the flow cell and very low cluster densities. On the basis of data from both pre- and post-PCR replicates, we provide an alternative hypothesis, that the major contributor to the variation between pre-PCR replicates is the random inclusion of low abundance contaminant microbial DNA during pipetting of the PCR template. While a few contaminants, having a relatively higher concentration, will always be part of the PCR template, the majority of contaminants will be present at such low concentrations that they will only occasionally be included. They are under the law of small numbers (18), where a random sample is not likely to reflect the population from which it is drawn, and the similarity between different samples is low. This would explain why we robustly detect the most abundant contaminant taxa across all samples and extraction controls in a sequencing run, whereas the presence and identity of less abundant background contaminants vary from sample to sample (Fig. 3; see also Tables S1, S2, and S4 in the supplemental material). Further, the negative-control replicates that were split after the 16S PCR, i.e., after massive amplification of any low abundance target and therefore with expected low intersample variability, showed a very high homogeneity (Fig. 2 and Tables S1 and S2). The latter finding contradicts the hypothesis by Erb-Downward et al. (17). It is essential to acknowledge the difference between pre- and postamplification replicates, and only the latter is useful for addressing the reproducibility of the sequencing technology itself.

Lower limit of detection. When the DNA input of a given species in a sample is getting close to one copy per PCR, it will, like the low abundance contaminants, be under the law of small numbers. This thus constitutes a lower limit of detection for 16S deep sequencing as a method. Our sequencing of the mock community illustrates this point. From the $1{:}10^5$ dilution, we identified *M. smithii* in only 2 out of four PCR

TABLE 4 Identified bacteria by sequencing compared to conventional culture

Condition, parameter, and bacterial species[a]	Identifications by 16S rRNA sequencing		Growth by culture	
	Total no.	% of all microbial detections	No.	% of identified by 16S rRNA sequencing
Acute cholangitis (15 patients)				
Total no. of identifications	84		26	31
Gram negative	**27**	**32**	**9**	**33**
Klebsiella spp.	8	9.5	2	25
Escherichia coli	7	8.3	6	86
Campylobacter spp.	3	3.6		
Enterobacter spp.	2	2.4		
Haemophilus parainfluenzae	2	2.4		
Aggregatibacter spp.	2	2.4		
Hafnia alvei	1	1.2	1	100
Moraxella osloensis	1	1.2		
Serratia odorifera	1	1.2		
Gram positive	**42**	**50**	**15**	**35**
Enterococcus spp.	11	13.1	9	82
Streptococcus spp.	10	11.9	3	30
Lactobacillus spp.	6	7.1	1	17
Actinomyces spp.	4	4.8		
Granulicatella adiacens	2	2.4		
Rothia mucilaginosa	2	2.4		
Staphylococcus spp.	2	2.4	1	50
Abiotrophia defectiva	1	1.2		
Bacillus halodurans	1	1.2		
Cellulosimicrobium sp.	1	1.2	1	100
Corynebacterium provencense	1	1.2		
Kocuria sp.	1	1.2		
Anaerobic	**15**	**18**	**2**	**14**
Fusobacterium spp.	4	4.8		
Veillonella spp.	4	4.8		
Clostridium perfringens	4	3.6	2	67
Bifidobacterium dentium	1	1.2		
Cutibacterium avidum	1	1.2		
Finegoldia magna	1	1.2		
Intestinibacter bartlettii	1	1.2		
Noninfectious bile duct stenosis (26 patients)				
Total no. of identifications	215	40	19	
Gram negative	**44**	**21**	**16**	**36**
Escherichia spp.	7	3.3	6	86
Klebsiella spp.	7	3.3	5	71
Haemophilus spp.	6	2.8	1	17
Campylobacter spp.	3	1.4		
Enterobacter spp.	3	1.4		
Neisseria spp.	3	1.4		
Citrobacter/Cronobacter	2	0.9		
Pseudomonas aeruginosa	2	0.9		
Aeromonas sp.	1	0.5	2	200
Bergeyella sp. (HMT-322)	1	0.5		
Capnocytophaga gingivalis	1	0.5		
Citrobacter amalonaticus	1	0.5	1	100
Hafnia alvei	1	0.5		
Hymenobacter sp.	1	0.5		
Kluyvera ascorbata	1	0.5		
Pluralibacter gergoviae	1	0.5		
Proteus sp.	1	0.5		
Pseudolabrys sp.	1	0.5		
Serratia marcescens	1	0.5	1	100
Gram positive	**75**	**35**	**21**	**28**
Streptococcus spp.	35	16.3	8	23
Actinomyces spp.	13	6.0	2	15
Enterococcus spp.	7	3.3	6	86
Granulicatella adiacens	4	1.9		

(Continued on next page)

TABLE 4 (Continued)

Condition, parameter, and bacterial species[a]	Identifications by 16S rRNA sequencing		Growth by culture	
	Total no.	% of all microbial detections	No.	% of identified by 16S rRNA sequencing
Rothia mucilaginosa	4	1.9	2	50
Saccharibacteria (TM7) spp.	4	1.9		
Staphylococcus spp.	3	1.4	2	67
Gemella spp.	2	0.9		
Corynebacterium sp.	1	0.5		
Kocuria palustris	1	0.5		
Leuconostoc lactis	1	0.5	1	100
Anaerobic	**96**	**45**	**3**	**3**
Veillonella spp.	18	8.4		
Prevotella spp.	14	6.5	1	7
Fusobacterium spp.	7	3.3		
Oribacterium spp.	6	2.8		
Leptotrichia spp.	5	2.3		
Clostridium spp.	5	2.3	1	20
Bifidobacterium spp.	4	1.9		
Stomatobaculum longum	3	1.4		
Peptostreptococcus spp.	3	1.4		
Atopobium parvulum	2	0.9		
Bacteroides spp.	2	0.9	1	50
Lachnoanaerobaculum spp.	2	0.9		
Megasphaera micronuciformis	2	0.9		
Alloprevotella tannerae	1	0.5		
Alloscardovia omnicolens	1	0.5		
Anaerococcus vaginalis	1	0.5		
Bilophila wadsworthia	1	0.5		
Catabacter hongkongensis	1	0.5		
Catonella morbi	1	0.5		
Colibacter massiliensis	1	0.5		
Cryptobacterium curtum	1	0.5		
Dialister pneumosintes	1	0.5		
Eggerthella lenta	1	0.5		
Eubacterium sulci	1	0.5		
Finegoldia magna	1	0.5		
Fretibacterium fastidiosum	1	0.5		
Lachnospiraceae (G-2) sp.	1	0.5		
Mogibacterium sp.	1	0.5		
Parasutterella excrementihominis	1	0.5		
Parvimonas micra	1	0.5		
Porphyromonas pasteri	1	0.5		
Selenomonas sp.	1	0.5		
Slackia exigua	1	0.5		
Solobacterium moorei	1	0.5		
Veillonellaceae [G-1] sp.	1	0.5		

[a]Microbes that could be identified only to a species group or genus level are listed at the genus level. For microbes that could be identified to the species level, and where there were no other species identified within the same genus, the species name is listed.

replicates. The theoretical abundance of *M. smithii* in the 1:10^5 dilution was 67 cells per ml, corresponding to a little less than one copy in 2 μl PCR template (Table S3).

Comparison of filtering methods. The major strength of our method for filtering contaminants is its high specificity, found to be 97% when evaluated on mock community dilutions (Fig. 7). As expected, the use of a single negative and positive extraction control had a very low specificity (39%) (method 2; Fig. 7). The law of small numbers implies that increasing the number of extraction controls should provide a more complete description of the background contaminants, and including OTUs found in all 10 extraction controls (method 3; Fig. 7) did result in filtering of more true contaminants. However, many of the low abundance contaminants were still not flagged, and the specificity of this method remained low (64%).

Our findings might explain why promising postanalytic methods for removing contaminants, like the R-package Decontam, still display reduced specificity in low

biomass/highly diluted samples (4, 5, 13). Decontam filtering of contaminant taxa had a specificity of 39% and 77%, respectively, in our mock community experiments (Fig. 7). The prevalence-based method in Decontam, which is recommended for low biomass samples (4), relies on the assumption that contaminating taxa are likely to have a higher prevalence in control samples than in true samples. Our results indicate that this assumption may be true only for the more abundant contaminant bacteria. Low abundance contaminant taxa that appear randomly in the negative extraction controls might not be recognized as contaminants.

The sensitivity of our suggested filtering method on the diluted mock communities was 83%. The mock community microbes wrongly classified as contaminants were those present in concentrations close to the absolute lower limit of detection, having a theoretical copy number ranging from <1 to 9 copies per 2 μl PCR template. Thus, this delineates the lower limit of detection for our filtering method. Many of the bile samples from the noninfectious patients also had low bacterial loads. They contained bacteria known to be part of the human oral microbiota, possibly reflecting contamination of the sampling catheter during the endoscopic retrograde cholangiopancreatography (ERCP) procedure. In some of these samples (e.g., samples 23 and 28 [see Fig. S1 in the supplemental material]), due to the low bacterial loads, many human oral bacteria were categorized as background contaminants by our filtering approach.

A major concern when subtracting all findings in the negative controls (1, 2, 13) is the situation where a species truly present in the sample is also found in the bacterial background. Our method allows for correct classification of these as relevant if they are represented by more reads than the most abundant contaminants. For the specific situation where the infection is caused by a species that is also among the dominant contaminants, one must look at alternative approaches. It is possible to calculate a sample-specific cutoff for differentiating between true and contaminant bacteria by using a combination of sequencing depth (number of reads) and the C_T values of the sample and the corresponding negative control in the 16S rRNA PCR (ΔC_T) (19). Although specific, this approach has a lower sensitivity.

Another suggested approach for contaminant filtering is to have an expert review of the samples and remove taxa that are considered biologically unexpected (1, 20). This method will however fall short if contaminant species are also biologically plausible, like many of the species identified in our extraction controls (e.g. *Anaerococcus* sp., *Actinomyces* sp., *Corynebacterium* sp., *Cutibacterium acnes*, *Staphylococcus* sp., *Finegoldia magna*, *Haemophilus* sp., *Pseudomonas* sp., *Prevotella* sp., *Streptococcus* sp., and *Moraxella* sp.) (Tables S2 and S4). However, combining our suggested filtering method with expert removal of biologically unexpected findings could possibly further increase the accuracy of results. In such a setting, clinically plausible findings below the cutoff of a valid identification could also be reported, but with more caution and as part of a broader clinical assessment.

Using the most abundant contaminant to establish a cutoff for likely valid identifications represents a dynamic approach taking into account both sequencing depth and the relative level of contamination in each individual sample. This is in contrast to some approaches based on a fixed cutoff, either a specified read count or a specified proportion of the total number of sequencing reads in each sample (21, 22). Such approaches will not be expedient for filtering samples with diverse bacterial loads or with dissimilar sequencing depths.

Setting the lower cutoff for acceptable bacterial identifications. We removed any species represented by less than 20% of the reads of the most abundant contaminant. This was a pragmatic cutoff, based on the observation that, with our reagents, inclusion of random background contaminants seemed to increase exponentially below this threshold (Fig. 4). However, as seen in Fig. 4, contaminants could occasionally occur at abundances up to 80% of the dominant background bacteria. We must therefore assume that some of the bacteria defined as "likely valid" in our clinical samples could represent contaminants. A likely example of this is the soil bacterium

Hymenobacter sp., found in sample 41 with an abundance of 33% compared to the most abundant contaminant (Fig. S1).

The relative number of reads representing a given bacterium in a sample will fluctuate somewhat from sequencing run to sequencing run. Such variations will be more pronounced among low abundance bacteria since they, like the background contaminants, are more affected by random differences in the number of target DNAs pipetted for the amplification PCR. This is illustrated by the repeat sequencing of mock community dilutions, where the interreplicate variations in relative species abundances were higher in the most diluted samples (Fig. 5B). Low abundance species will therefore be vulnerable for accidentally falling below the cutoff in some runs, explaining why altogether 27 bacteria were validly detected in only one of the 16S rRNA replicates among the "low bacterial load" bile samples (Fig. 8).

The relationship between bacterial load, sequencing depth, and diagnostic sensitivity. Background contamination is described as mainly constituting a challenge in low biomass samples, and many studies report the inverse relationship between the bacterial load of a sample and the relative abundance of contaminating DNA (1–3, 5). We will argue that the absence of contaminant species in data from a high biomass sample is actually an indication of inadequate sequencing depth. If you are not seeing any contaminants, there may remain undiscovered species with lower abundances that you could have detected using a higher number of reads (as in our sequencing of the 1:10 dilution of the mock community). This is well exemplified by our repeated 16S rRNA sequencing of clinical samples, where all except 1 out of 62 singleton findings were made in the replicate with the highest sequencing depth (Fig. 8). All these extra identifications were also, as expected, among the low abundance species in their samples with relative abundances of $<$1% of the total number of accepted bacterial reads (Fig. S1). Sample 38 (moderate bacterial load, C_T value of 22.5) represented the most extreme example (Fig. S1). For this sample, the number of accepted reads increased from 17,966 reads in the first replicate to 188,744 in the second. With this increase, we were able to identify 32 additional species and, as an indication of sufficient depth, small amounts of contamination (113 reads/0.001% with *Ralstonia picketti*). The high number of reads needed for robust description of polymicrobial clinical infections is emphasized by our data. For samples with moderate to high bacterial loads, even a sequencing depth of hundreds of thousands reads was frequently insufficient to start seeing contaminant bacteria (Table 2 and Fig. S1).

Cross-contamination. Another possible source of contamination in target amplicon sequencing is cross-contamination between samples (2). The level of cross-contamination is difficult to determine with certainty. To minimize the risk of sequencing noise and cross-contamination disturbing our results, we rejected all OTU clusters containing less than 50 sequences. This is a similar or even more strict criterion than other studies have used (9, 19, 22–24).

Limitations. We believe the general principles outlined in the study will be transferable to other clinical labs. However, background contamination will vary between labs, between extraction kits and PCR reagents, and even between batches of the same extraction kits and PCR reagents (2). Every lab should analyze and monitor the pattern of contamination in their own sequencing results if adopting our approach for filtering of contaminants and adjust their filtering cutoffs according to their findings. Adjustments could include, e.g., the number of "top contaminants" or the "frequency threshold rate."

Conclusion. In this study, we demonstrate the limitations of simply using microbial identifications in negative controls as the basis for filtering background bacterial contamination. The main concern regarding this strategy until now has been that the negative controls may contain bacteria that are also truly present in the clinical samples or that the negative controls may be contaminated with DNA from the clinical samples during the sequencing process (1, 2, 13) and that true findings therefore will be discarded as contaminants. We demonstrate that due to the law of small numbers, the

risk of accepting contaminants as true findings should be of equal concern using this strategy.

We suggest using the most abundant background contaminant species to define a level in each sample from where identifications might represent contamination. Below this level, again due to the law of small numbers, it rapidly becomes very demanding to discriminate between background and true findings. The most abundant contaminant DNA can also serve to evaluate sequencing depth. Adequate sequencing depth can be claimed only when the analysis also picks up background contamination.

MATERIALS AND METHODS

Inclusion of patients and collection of bile samples. This was a prospective, single-center study performed at Haukeland University Hospital, Bergen, Norway. The study was approved by the regional ethical committee (2015/65). Written informed consent was obtained from all participants.

From July 2015 to April 2017, bile samples were collected from all patients undergoing endoscopic retrograde cholangiopancreatography (ERCP). Patients diagnosed with either acute calculous cholangitis, defined according to the Tokyo Guidelines 2013 (25) (TG13) criteria for a definite diagnosis or noninfectious bile duct stone were included for further analysis.

Bile samples were immediately placed in sterile sample glass and sent to the laboratory for analysis after sampling. Upon arrival to the laboratory, DNA was extracted directly from 400 μl of bile as described previously (15, 16, 26). The eluate was stored at −80°C for later deep sequencing analysis. All samples were also routinely cultured according to our previously described laboratory guidelines (16).

Endoscopic retrograde cholangiography and pancreatography procedure. The intestine was rinsed with a solution of water and Minifom before procedure. ERCP was performed with the patient in the supine position. The patient was sedated with midazolam and pethidine, and if needed supplemented with buscopan for bowel relaxation. A side-viewing, sterilized, endoscope (Evis Exera III Duodenovideoscope, Olympus TJF – Q190V, Olympus) was used. Wire guided selective bile duct cannulation was performed with use of a guidewire (Dreamwire 0.035 in., 260 cm; Boston Scientific, Costa Rica) passed through a sterile sphincterotome catheter (Jagtome RX 44; Boston Scientific, Costa Rica). The position in the bile duct was confirmed by X-ray to identify the position of the catheter and guide wire before aspiration of approximately 2 to 5 ml bile. If there was any concern about the location of the guidewire, the sphincterotome was gently advanced over the guidewire, and a small amount of contrast was injected to delineate the anatomy. If there were any difficulties with cannulation of the ampulla of Vater, normal saline was injected to dilate the bile duct. Normal saline injections was also used to flush the bile ducts if bile aspiration attempts yielded little or no fluid in return on the catheter.

Mock community dilution. A staggered mock community from ZymoBIOMIC were used (ZymoBIOMICS Gut Microbiome Standard, catalog no. D6331; Zymo Research Corp., Irvine, CA, USA). This mock community consists of 19 bacterial strains representing 15 bacterial species (*Faecalibacterium prausnitzii*, *Veillonella rogosae*, *Roseburia hominis*, *Bacteroides fragilis*, *Prevotella corporis*, *Bifidobacterium adolescentis*, *Fusobacterium nucleatum*, *Lactobacillus fermentum*, *Clostridioides difficile*, *Akkermansia muciniphila*, *Methanobrevibacter smithii*, *Salmonella enterica*, *Enterococcus faecalis*, *Clostridium perfringens*, and *Escherichia coli* strains JM109, B-3008, B-2207, B-766, and B-1109) and two fungal species (*Saccharomyces cerevisiae* and *Candida albicans*). The mock community was diluted with microbial DNA-free water (Qiagen) in seven rounds of a serial 10-fold dilution prior to DNA extraction. The dilutions were analyzed with a SYBR green real-time 16S rRNA PCR using a previously described protocol (15) to obtain a semiquantitative measure of the bacterial load of each dilution. A dilution with high bacterial load (1:10) and two different dilutions with low bacterial load ($1:10^5$ and $1:10^6$) were selected for further analysis. Negative and positive extraction controls were included and followed all processing steps.

Gene targets. In all bile samples, mock community samples, and extraction control samples, the 16S rRNA gene V3-V4 region was sequenced (see Table S8 in the supplemental material). For selected bile samples, a part of the *rpoB* gene were also sequenced in a separate sequencing run to obtain a higher taxonomic resolution for *Enterobacteriaceae*, *Enterococcus*, *Streptococcus*, and *Staphylococcus* species identified by the 16S rRNA sequencing (16). Species identified at a higher taxonomic level with partial *rpoB* gene sequencing compared to partial 16S rRNA gene sequencing (V3-V4) are listed in Table S7. All primers used were the same as described previously (16), except for a modification of one of the two forward *RpoB*_ESS primers to obtain better coverage of *Enterococcus raffinosus* (Table S8). All primers are listed in Table S8.

Sequencing procedure. The Illumina Miseq system (Illumina, Redwood City, CA) was used for sequencing. The sequencing protocol was a modified version of the of the Illumina 16S Metagenomic Library Preparation protocol (27) as described previously (15, 16). Briefly, the sequencing workflow included the following stages. The target genes were amplified in an amplicon PCR using the same temperature profile for all targets. An overview of the PCR mixture for the different gene targets and the temperature profile of the amplicon PCR is provided in Table S8. After PCR cleanup of the amplicon PCR product with use of AMPure XP beads, the next step was attachment of dual indices and Illumina sequencing adapters in an index PCR. The index PCR product underwent a similar cleanup, followed by a fluorometric quantification of the DNA content of each sample using Qubit 3.0 fluorometer (Fisher Scientific) and the QubitR dsDNA (double-stranded DNA) HS (high-sensitivity) assay kit (0.2 to 100 ng). Samples were then diluted using 10 mM Tris (pH 8.5) to reach a final concentration of 4 nM, before they

were pooled together into a final library pool that was denatured, diluted, and mixed with a Phix control before loaded on the Miseq system as described in the Illumina protocol (27).

For the 16S rRNA amplicon sequencing of bile samples, each sample was split into two replicates (16S rRNA replicate 1 and 16S rRNA replicate 2) after DNA extraction and then processed in different PCR amplification and sequencing runs. The second replicate from each sample was sequenced with fewer samples per sequencing run to obtain a higher sequencing depth.

Assessing the bacterial load in the bile samples. A semiquantitative measure of bacterial load in each sample was calculated using the C_T value from the SYBR green real-time 16S rRNA PCR, following the same protocol as for the mock community experiment. According to their C_T value, samples were categorized as having either high bacterial load (C_T values ranging from lowest to 19), moderate bacterial load (C_T values ranging from 20 to 24) or low bacterial load (C_T values ranging from 25 to highest).

Extraction controls. Each sample was processed together with a parallel negative extraction control consisting of lysis buffer and PCR-grade water. For the bile samples, all negative extraction controls were mixed into two or three pools before sequencing, depending on the number of samples included in the sequencing run. In addition, a weakly positive extraction control consisting of *Legionella pneumophila* suspended in PCR-grade water was included.

Postsequencing processing. The Miseq Reporter software was used for removing primers, demultiplexing, and generating FASTQ files for each sample. AdapterRemoval 2.2.2 (28) was used for trimming adapter sequences and low-quality bases and to merge the forward and reverse FASTQ files of each sample, using the following command: AdapterRemoval –file1 <reads_1.fq> --file2 <reads_2.fq> --basename <mymergedfile> --threads 7 –trimns –trimqualities –minquality 20 –collapse – adapter-list <adapters>.txt –gzip.

Downstream analysis was then performed using the RipSeq next-generation sequencing (NGS) software (Pathogenomix, Santa Cruz, CA) (15, 16) (*de novo* clustering into operational taxonomic units [OTUs] using a 99% similarity threshold). A chimera check was performed with the RipSeq online tool.

Taxonomic assignment. OTUs were assigned using the RipSeq online BLAST search against the RipSeqs curated database "Pathogenomics Prime 16S" (16S), "*Pathogenomix rpoB_ESS*," "*Pathogenomix rpoB_Ent*," and "*GenBank Bacteria 1 – All bacterial targets*, *Valid Species and Pubmed*" (*rpoB*). OTUs that did not match a reference sequence using these RipSeq curated databases were manually assigned by performing a BLAST search against the GenBank NCBI database and the Human Oral Microbiome Database (www.homd.org). OTUs mapping to the same reference species were merged.

Criteria used for taxonomy assignments for both 16S rRNA and *rpoB* gene were the same as described previously (16) (for 16S rRNA species-level identification, ≥99.3% homology with a high-quality reference, and minimum distance >0.7% to the next alternative species). OTUs obtaining species-level homology but with an insufficient distance to the next species were assigned to a species group or listed as a slashed result. OTUs that did not assign to any known species were indicated as "Unknown bacteria #." A full list of all species groups and of the best BLAST search match in GenBank NCBI database for all OTUs termed as "Unknown bacteria #" is found in Table S9.

Secondary filtration of sequencing results. A lower cutoff for the number of representative sequences required to retain an OTU is recommended as a secondary filtration to diminish problems related to sequencing noise and cross-contamination of samples (9, 19, 22–24, 29). We rejected OTUs represented by fewer than 50 reads. Further filtering of bacterial background DNA from the sequencing results is outlined in Results.

Statistical analysis. Statistical analyses were performed using SPSS 25 (IBM Corp.) and the R programming language (30). Clinical and microbial characteristics of categorical and continuous data were analyzed with Pearson's chi-squared test and Student's *t* test, respectively. Mann-Whitney U-test was used for continuous, skewed variables. Figures illustrating microbial distribution were produced using the R-packages "VennDiagram" (31) version 1.6.0 and "ggplot2" (32) version 3.2.1. Diversity analyses were performed using the R-package "phyloseq" (33) version 1.30.0. Rarefaction of data used in diversity measures was performed using the phyloseq package in R with the following arguments: rarefy_even_depth (Otu_table, sample.size = min(sample_sums(Otu_table)), rngseed = TRUE, replace = TRUE, verbose = TRUE).

Data availability. The source data from experiment 1 and experiment 2 have been deposited in the European Nucleotide Archive (ENA) at EMBL-EBI under accession number PRJEB44556 (https://www.ebi.ac.uk/ena/browser/view/PRJEB44556).

Other source data of this study are available from the corresponding author upon request. Not all patient data are publicly available due to restrictions from the Regional Ethical Committee.

SUPPLEMENTAL MATERIAL

Supplemental material is available online only.

FIG S1, PDF file, 0.5 MB.

TABLE S1, PDF file, 0.2 MB.

TABLE S2, XLSX file, 0.02 MB.

TABLE S3, PDF file, 0.2 MB.

TABLE S4, XLSX file, 0.02 MB.

TABLE S5, XLSX file, 0.02 MB.

TABLE S6, XLSX file, 0.1 MB.

TABLE S7, PDF file, 0.1 MB.

TABLE S8, PDF file, 0.2 MB.
TABLE S9, PDF file, 0.2 MB.

ACKNOWLEDGMENTS

We thank the staff at the Department of Surgery at Haukeland University Hospital for their invaluable help in collecting the bile samples.

This work was supported by the Western Norway Regional Health Authority's research funding (grant 912206).

O.K. contributed to the development of the RipSeq software and is a minor shareholder of Pathogenomix Inc. All other authors had no conflicts of interest.

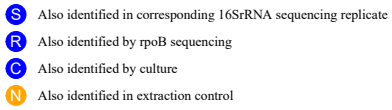
REFERENCES

1. Salter SJ, Cox MJ, Turek EM, Calus ST, Cookson WO, Moffatt MF, Turner P, Parkhill J, Loman NJ, Walker AW. 2014. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. BMC Biol 12:87. https://doi.org/10.1186/s12915-014-0087-z.
2. Eisenhofer R, Minich JJ, Marotz C, Cooper A, Knight R, Weyrich LS. 2019. Contamination in low microbial biomass microbiome studies: issues and recommendations. Trends Microbiol 27:105–117. https://doi.org/10.1016/j.tim.2018.11.003.
3. Boers SA, Jansen R, Hays JP. 2019. Understanding and overcoming the pitfalls and biases of next-generation sequencing (NGS) methods for use in the routine clinical microbiological diagnostic laboratory. Eur J Clin Microbiol Infect Dis 38:1059–1070. https://doi.org/10.1007/s10096-019-03520-3.
4. Davis NM, Proctor DM, Holmes SP, Relman DA, Callahan BJ. 2018. Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data. Microbiome 6:226. https://doi.org/10.1186/s40168-018-0605-2.
5. Drengenes C, Wiker HG, Kalananthan T, Nordeide E, Eagan TML, Nielsen R. 2019. Laboratory contamination in airway microbiome studies. BMC Microbiol 19:187. https://doi.org/10.1186/s12866-019-1560-1.
6. Kozlov A, Bean L, Hill EV, Zhao L, Li E, Wang GP. 2018. Molecular identification of bacteria in intra-abdominal abscesses using deep sequencing. Open Forum Infect Dis 5:ofy025. https://doi.org/10.1093/ofid/ofy025.
7. Tarabichi M, Shohat N, Goswami K, Alvand A, Silibovsky R, Belden K, Parvizi J. 2018. Diagnosis of periprosthetic joint infection: the potential of next-generation sequencing. J Bone Joint Surg Am 100:147–154. https://doi.org/10.2106/JBJS.17.00434.
8. Liu H, Guo M, Jiang Y, Cao Y, Qian Q, He X, Huang K, Zhang J, Xu W. 2019. Diagnosing and tracing the pathogens of infantile infectious diarrhea by amplicon sequencing. Gut Pathog 11:12. https://doi.org/10.1186/s13099-019-0292-y.
9. Salipante SJ, Sengupta DJ, Rosenthal C, Costa G, Spangler J, Sims EH, Jacobs MA, Miller SI, Hoogestraat DR, Cookson BT, McCoy C, Matsen FA, Shendure J, Lee CC, Harkins TT, Hoffman NG. 2013. Rapid 16S rRNA next-generation sequencing of polymicrobial clinical samples for diagnosis of complex bacterial infections. PLoS One 8:e65226. https://doi.org/10.1371/journal.pone.0065226.
10. Cummings LA, Kurosawa K, Hoogestraat DR, SenGupta DJ, Candra F, Doyle M, Thielges S, Land TA, Rosenthal CA, Hoffman NG, Salipante SJ, Cookson BT. 2016. Clinical next generation sequencing outperforms standard microbiological culture for characterizing polymicrobial samples. Clin Chem 62:1465–1473. https://doi.org/10.1373/clinchem.2016.258806.
11. Bryan A, Kirkpatrick LM, Manaloor JJ, Salipante SJ. 2017. 16S rRNA deep sequencing identifies Actinotignum schaalii as the major component of a polymicrobial intra-abdominal infection and implicates a urinary source. JMM Case Rep 4:e005091. https://doi.org/10.1099/jmmcr.0.005091.
12. Abayasekara LM, Perera J, Chandrasekharan V, Gnanam VS, Udunuwara NA, Liyanage DS, Bulathsinhala NE, Adikary S, Aluthmuhandiram JVS, Thanaseelan CS, Tharmakulasingam DP, Karunakaran T, Ilango J. 2017. Detection of bacterial pathogens from clinical specimens using conventional microbial culture and 16S metagenomics: a comparative study. BMC Infect Dis 17:631. https://doi.org/10.1186/s12879-017-2727-8.
13. Karstens L, Asquith M, Davin S, Fair D, Gregory WT, Wolfe AJ, Braun J, McWeeney S. 2019. Controlling for contaminants in low-biomass 16S rRNA gene sequencing experiments. mSystems 4:e00290-19. https://doi.org/10.1128/mSystems.00290-19.
14. Theis KR, Romero R, Winters AD, Greenberg JM, Gomez-Lopez N, Alhousseini A, Bieda J, Maymon E, Pacora P, Fettweis JM, Buck GA, Jefferson KK, Strauss JF, III, Erez O, Hassan SS. 2019. Does the human placenta delivered at term have a microbiota? Results of cultivation, quantitative real-time PCR, 16S rRNA gene sequencing, and metagenomics. Am J Obstet Gynecol 220:267.e1–267.e39. https://doi.org/10.1016/j.ajog.2018.10.018.
15. Dyrhovden R, Nygaard RM, Patel R, Ulvestad E, Kommedal O. 2019. The bacterial aetiology of pleural empyema. A descriptive and comparative metagenomic study. Clin Microbiol Infect 25:981–986. https://doi.org/10.1016/j.cmi.2018.11.030.
16. Dyrhovden R, Ovrebo KK, Nordahl MV, Nygaard RM, Ulvestad E, Kommedal O. 2019. Bacteria and fungi in acute cholecystitis. A prospective study comparing next generation sequencing to culture. J Infect 80:16−23. https://doi.org/10.1016/j.jinf.2019.09.015.
17. Erb-Downward JR, Falkowski NR, D'Souza JC, McCloskey LM, McDonald RA, Brown CA, Shedden K, Dickson RP, Freeman CM, Stringer KA, Foxman B, Huffnagle GB, Curtis JL, Adar SD. 2020. Critical relevance of stochastic effects on low-bacterial-biomass 16S rRNA gene analysis. mBio 11:e00258-20. https://doi.org/10.1128/mBio.00258-20.
18. Tversky A, Kahneman D. 1971. Belief in law of small numbers. Psychological Bull 76:105–110. https://doi.org/10.1037/h0031322.
19. Kommedal Ø, Wilhelmsen MT, Skrede S, Meisal R, Jakovljev A, Gaustad P, Hermansen NO, Vik-Mo E, Solheim O, Ambur OH, Sæbø Ø, Høstmælingen CT, Helland C. 2014. Massive parallel sequencing provides new perspectives on bacterial brain abscesses. J Clin Microbiol 52:1990–1997. https://doi.org/10.1128/JCM.00346-14.
20. de Goffau MC, Lager S, Salter SJ, Wagner J, Kronbichler A, Charnock-Jones DS, Peacock SJ, Smith GCS, Parkhill J. 2018. Recognizing the reagent microbiome. Nat Microbiol 3:851–853. https://doi.org/10.1038/s41564-018-0202-y.
21. Minich JJ, Zhu Q, Janssen S, Hendrickson R, Amir A, Vetter R, Hyde J, Doty MM, Stillwell K, Benardini J, Kim JH, Allen EE, Venkateswaran K, Knight R. 2018. KatharoSeq enables high-throughput microbiome analysis from low-biomass samples. mSystems 3:e00218-17. https://doi.org/10.1128/mSystems.00218-17.
22. Stebner A, Ensser A, Geissdorfer W, Bozhkov Y, Lang R. 2020. Molecular diagnosis of polymicrobial brain abscesses with 16S-rDNA-based next-generation sequencing. Clin Microbiol Infect 27:76−82. https://doi.org/10.1016/j.cmi.2020.03.028.
23. Franzen O, Hu J, Bao X, Itzkowitz SH, Peter I, Bashir A. 2015. Improved OTU-picking using long-read 16S rRNA gene amplicon sequencing and generic hierarchical clustering. Microbiome 3:43. https://doi.org/10.1186/s40168-015-0105-6.
24. Tremblay J, Yergeau E. 2019. Systematic processing of ribosomal RNA gene amplicon sequencing data. Gigascience 8:giz146. https://doi.org/10.1093/gigascience/giz146.
25. Kiriyama S, Takada T, Strasberg SM, Solomkin JS, Mayumi T, Pitt HA, Gouma DJ, Garden OJ, Buchler MW, Yokoe M, Kimura Y, Tsuyuguchi T, Itoi T, Yoshida M, Miura F, Yamashita Y, Okamoto K, Gabata T, Hata J, Higuchi R, Windsor JA, Bornman PC, Fan ST, Singh H, de Santibanes E, Gomi H, Kusachi S, Murata A, Chen XP, Jagannath P, Lee S, Padbury R, Chen MF, Dervenis C, Chan AC, Supe AN, Liau KH, Kim MH, Kim SW, Tokyo Guidelines Revision Committee. 2013. TG13 guidelines for diagnosis and severity grading of acute cholangitis (with videos). J Hepatobiliary Pancreat Sci 20:24–34. https://doi.org/10.1007/s00534-012-0561-3.

26. Kommedal O, Kvello K, Skjastad R, Langeland N, Wiker HG. 2009. Direct 16S rRNA gene sequencing from clinical specimens, with special focus on polybacterial samples and interpretation of mixed DNA chromatograms. J Clin Microbiol 47:3562–3568. https://doi.org/10.1128/JCM.00973-09.
27. Illumina. 2013. 16S metagenomic sequencing library preparation: preparing 16S ribosomal RNA gene amplicons for the Illumina MiSeq system. Illumina, San Diego, CA. https://support.illumina.com/downloads/16s_metagenomic_sequencing_library_preparation.html.
28. Schubert M, Lindgreen S, Orlando L. 2016. AdapterRemoval v2: rapid adapter trimming, identification, and read merging. BMC Res Notes 9:88. https://doi.org/10.1186/s13104-016-1900-2.
29. Bokulich NA, Subramanian S, Faith JJ, Gevers D, Gordon JI, Knight R, Mills DA, Caporaso JG. 2013. Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing. Nat Methods 10:57–59. https://doi.org/10.1038/nmeth.2276.
30. R Core Team. 2019. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.
31. Chen H. 2018. VennDiagram: generate high-resolution Venn and Euler plots. https://CRAN.R-project.org/package=VennDiagram.
32. Hadley W. 2016. ggplot2: elegant graphics for data analysis. Springer-Verlag, New York, NY.
33. McMurdie PJ, Holmes S. 2013. phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. PLoS One 8:e61217. https://doi.org/10.1371/journal.pone.0061217.

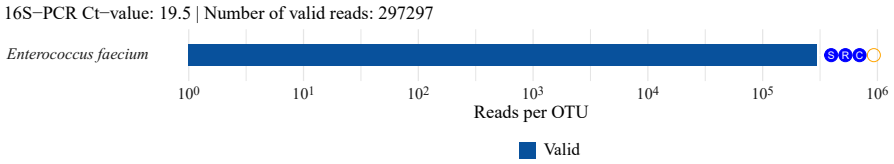
Supplementary Figure S1:

Barchart illustrating the sequencing result of each bile sample in 16S rRNA replicate 2. Number of reads per OTU are given on a log-transformed scale. Bars for species with an abundance above the top abundant contaminant is categorized as "Valid" and colored dark blue. Bars for species with an abundance between 20 to 100% of the top abundant contaminant is categorized as "Likely valid" and colored light blue. Bars with an abundance below 20% of the top abundant contaminant is categorized as "Contaminant" and colored red.

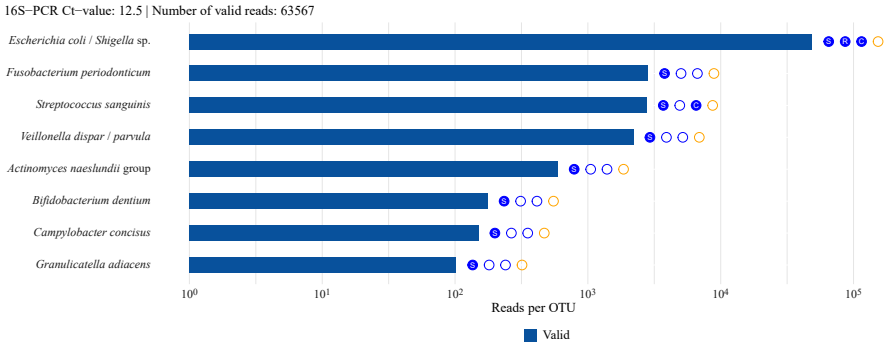


Sample 01-15: Bile samples from patients with acute cholangitis.

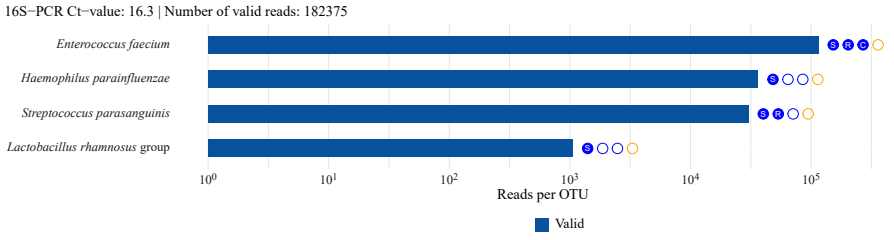
Sample 01



Sample 02

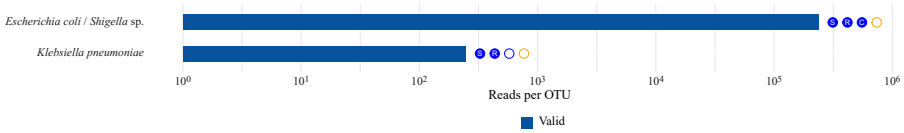


Sample 03



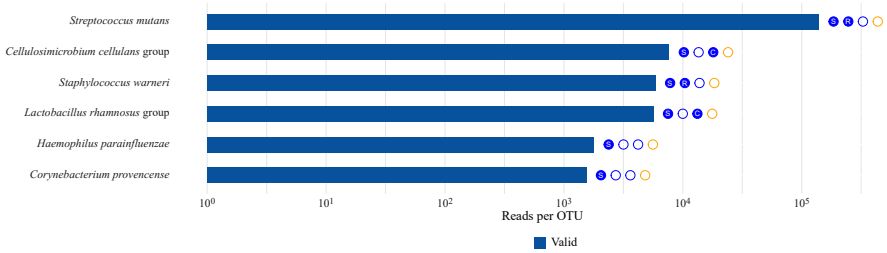
Sample 04

16S−PCR Ct−value: 12.9 | Number of valid reads: 235807



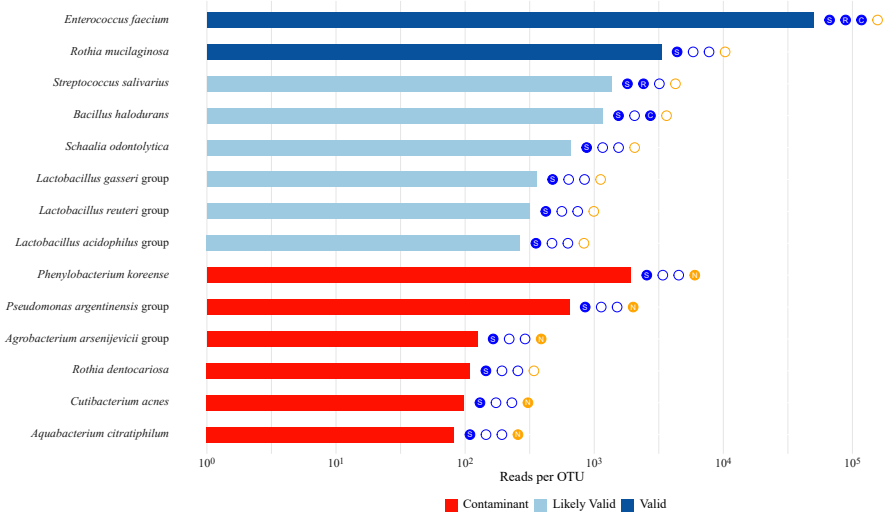
Sample 05

16S−PCR Ct−value: 17.6 | Number of valid reads: 161737



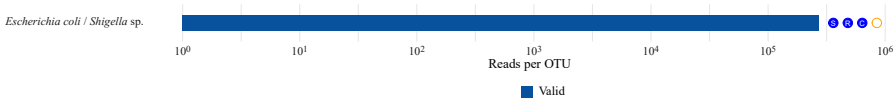
Sample 06

16S−PCR Ct−value: 27.9 | Number of valid reads: 60399



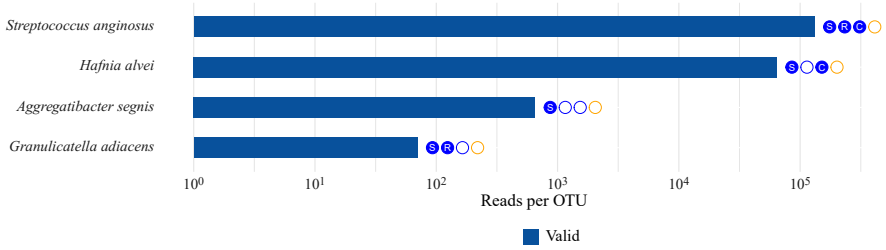
Sample 07

16S−PCR Ct−value: 15.1 | Number of valid reads: 271147



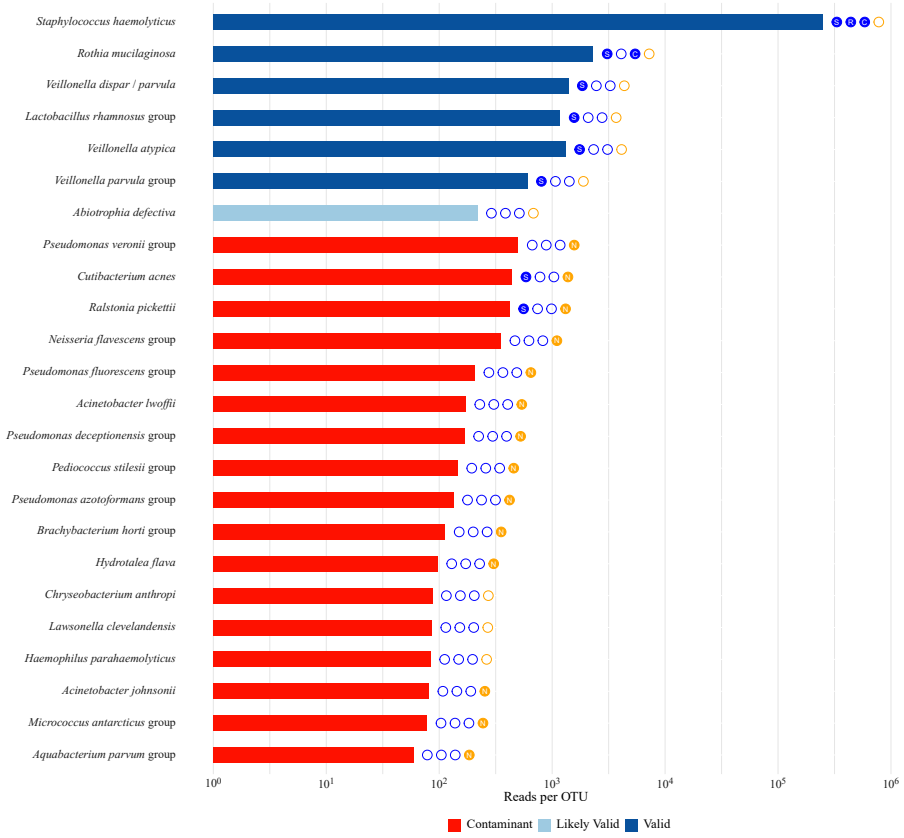
Sample 08

16S−PCR Ct−value: 20.5 | Number of valid reads: 196768



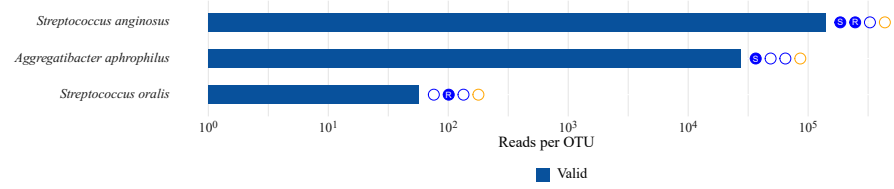
Sample 09

16S−PCR Ct−value: 26.7 | Number of valid reads: 259551



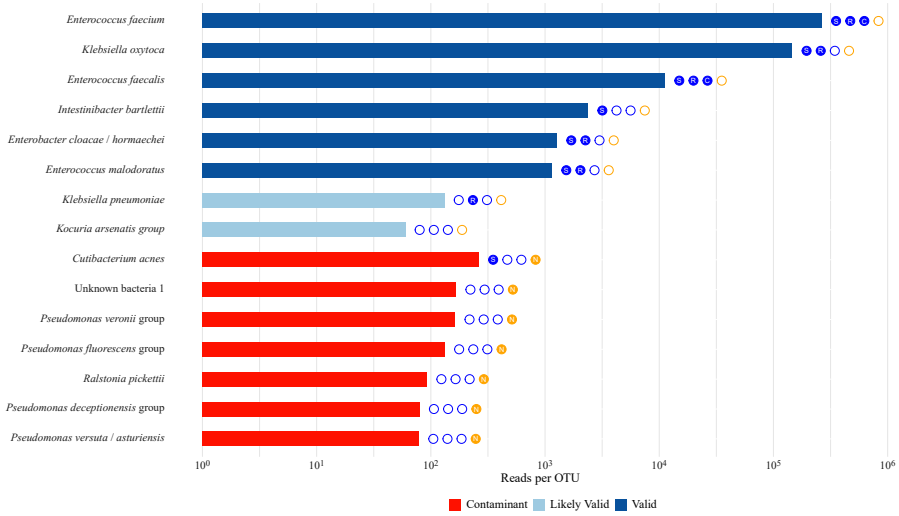
Sample 10

16S−PCR Ct−value: 22.6 | Number of valid reads: 166919



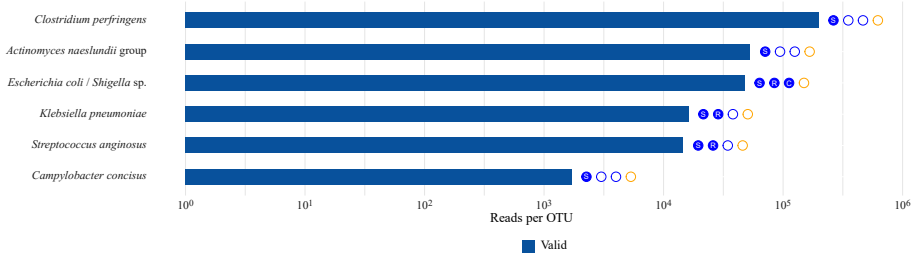
Sample 11

16S−PCR Ct−value: 24.4 | Number of valid reads: 429752



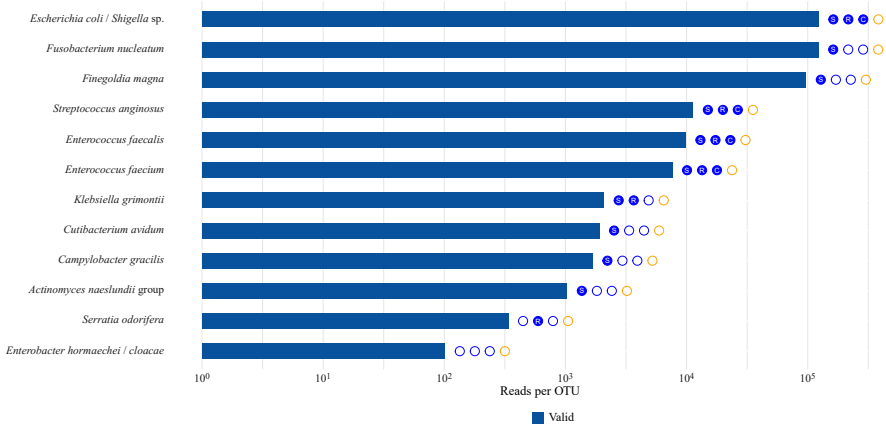
Sample 12

16S−PCR Ct−value: 18.5 | Number of valid reads: 332011

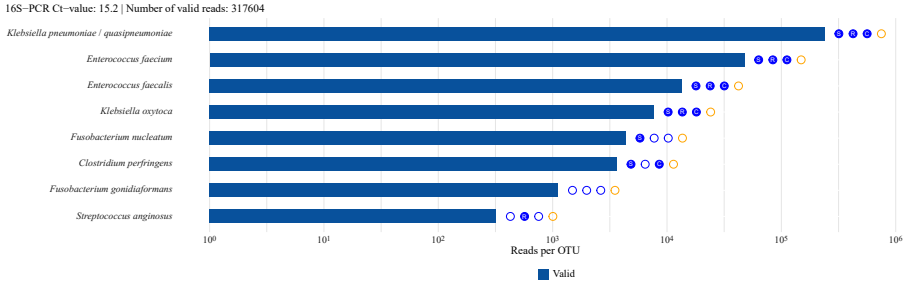


Sample 13

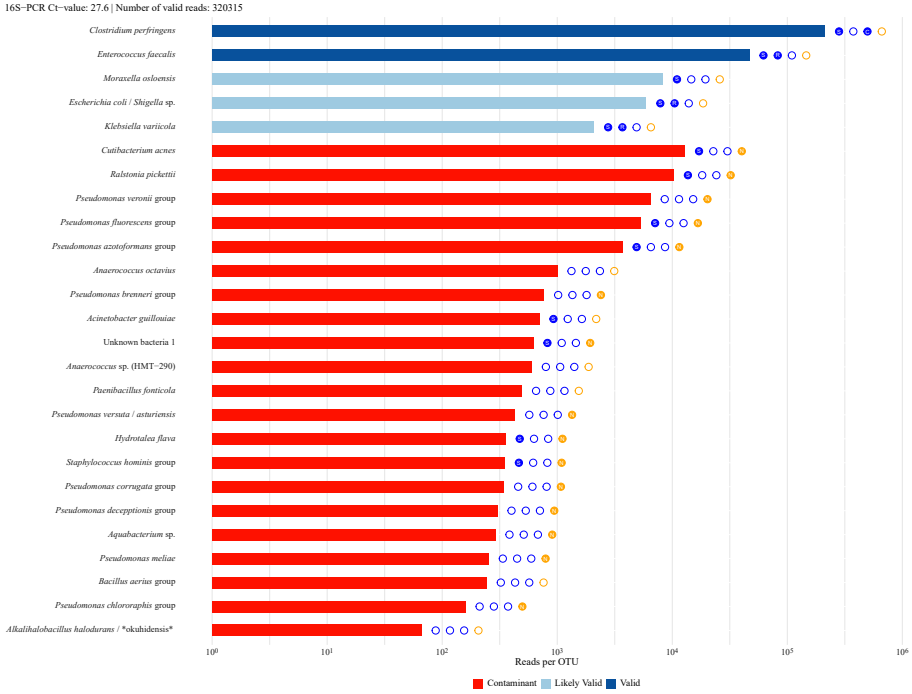
16S−PCR Ct−value: 19.2 | Number of valid reads: 376639



Sample 14



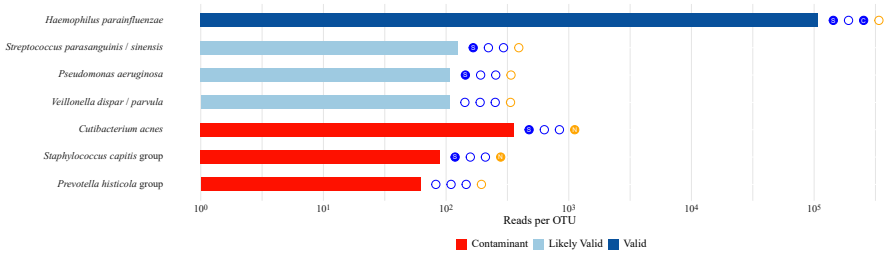
Sample 15



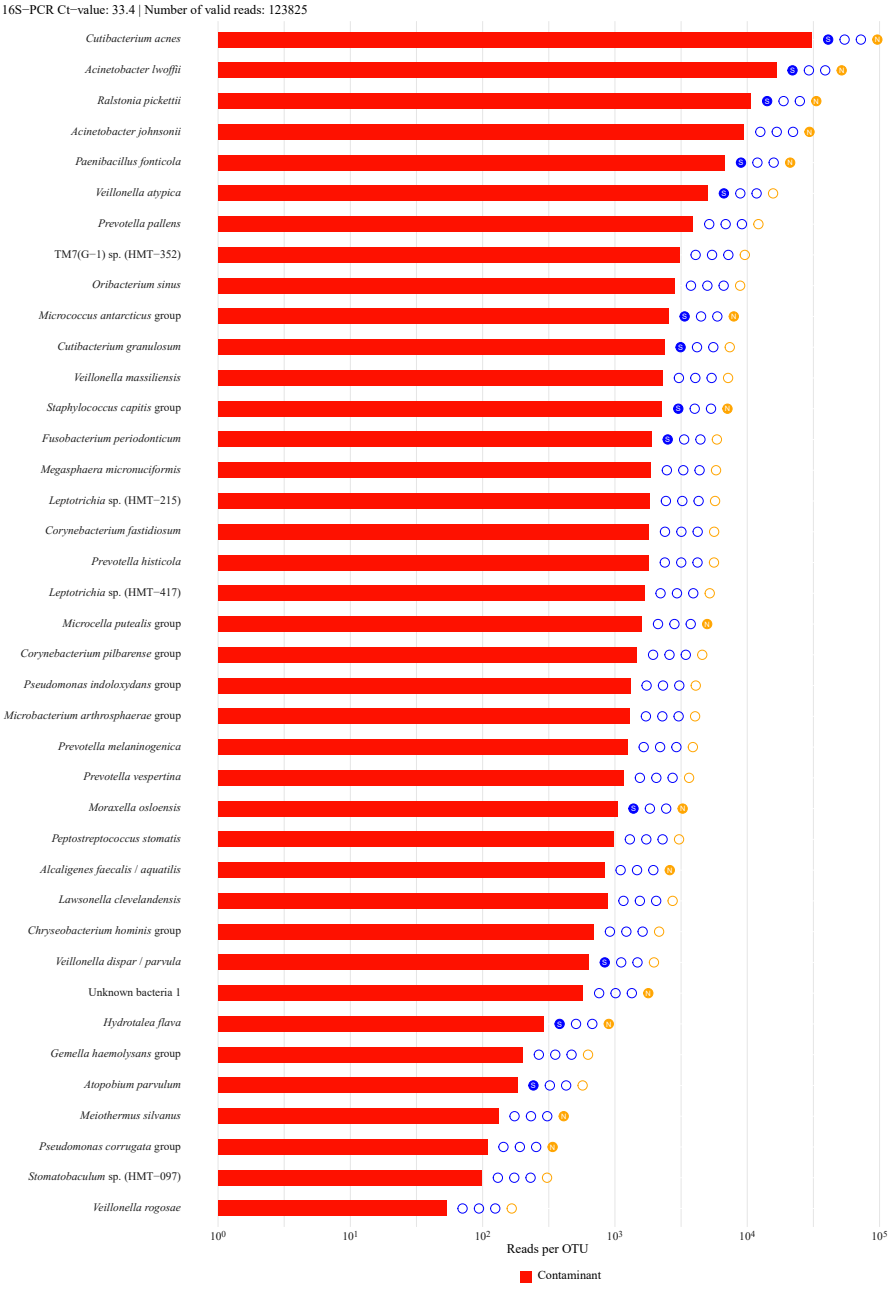
<u>Sample 16-41: Biles samples from patiens with non-infectious bile duct stenosis caused by bile duct stone.</u>

Sample 16

16S–PCR Ct–value:22.2 | Number of valid reads: 108617

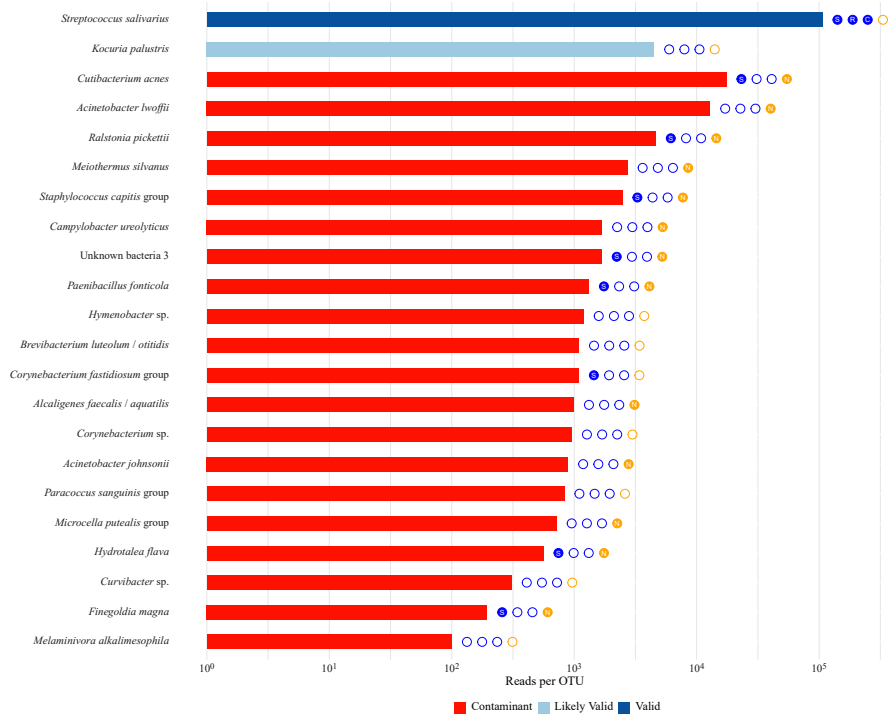


Sample 17


16S−PCR Ct−value: 33.4 | Number of valid reads: 123825
Cutibacterium acnes
Acinetobacter lwoffii
Ralstonia pickettii
Acinetobacter johnsonii
Paenibacillus fonticola
Veillonella atypica
Prevotella pallens
TM7(G−1) sp. (HMT−352)
Oribacterium sinus
Micrococcus antarcticus group
Cutibacterium granulosum
Veillonella massiliensis
Staphylococcus capitis group
Fusobacterium periodonticum
Megasphaera micronuciformis
Leptotrichia sp. (HMT−215)
Corynebacterium fastidiosum
Prevotella histicola
Leptotrichia sp. (HMT−417)
Microcella putealis group
Corynebacterium pilbarense group
Pseudomonas indoloxydans group
Microbacterium arthrosphaerae group
Prevotella melaninogenica
Prevotella vespertina
Moraxella osloensis
Peptostreptococcus stomatis
Alcaligenes faecalis / aquatilis
Lawsonella clevelandensis
Chryseobacterium hominis group
Veillonella dispar / parvula
Unknown bacteria 1
Hydrotalea flava
Gemella haemolysans group
Atopobium parvulum
Meiothermus silvanus
Pseudomonas corrugata group
Stomatobaculum sp. (HMT−097)
Veillonella rogosae
10^0
10^1
10^2
10^3
10^4
10^5
Reads per OTU
Contaminant

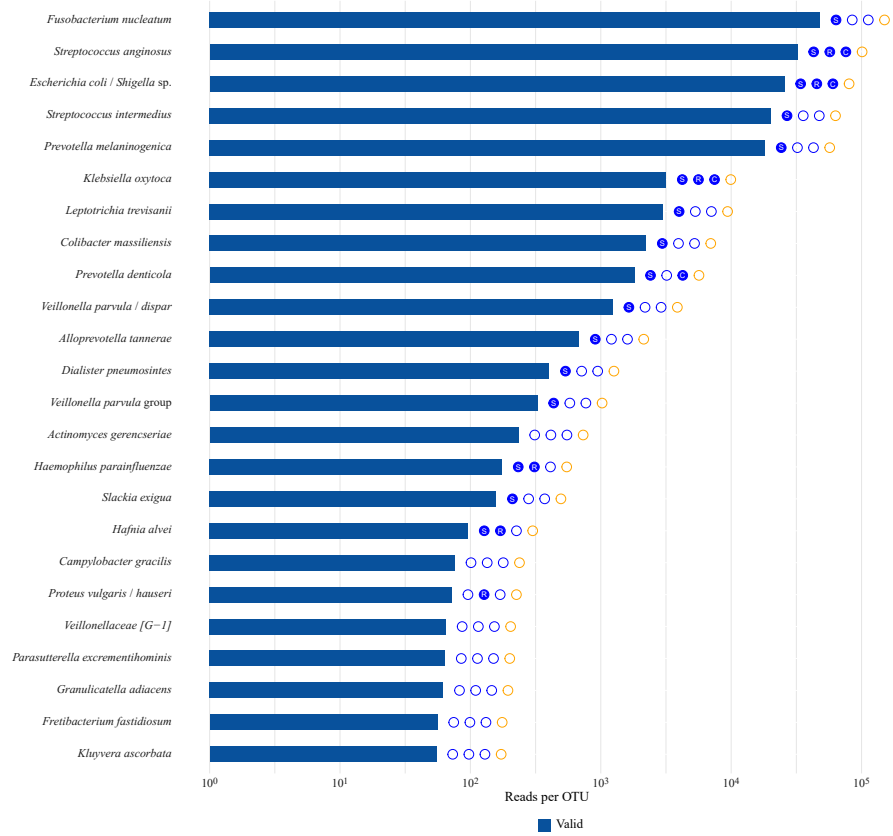
Sample 18

16S−PCR Ct−value: 30.6 | Number of valid reads: 164345

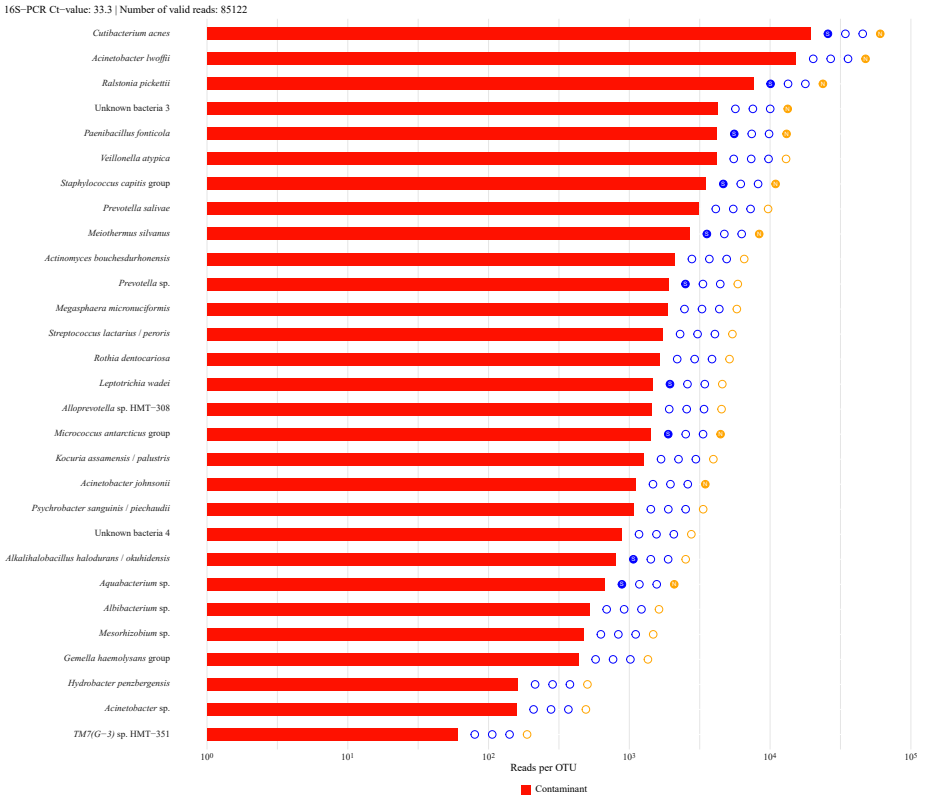


Sample 19

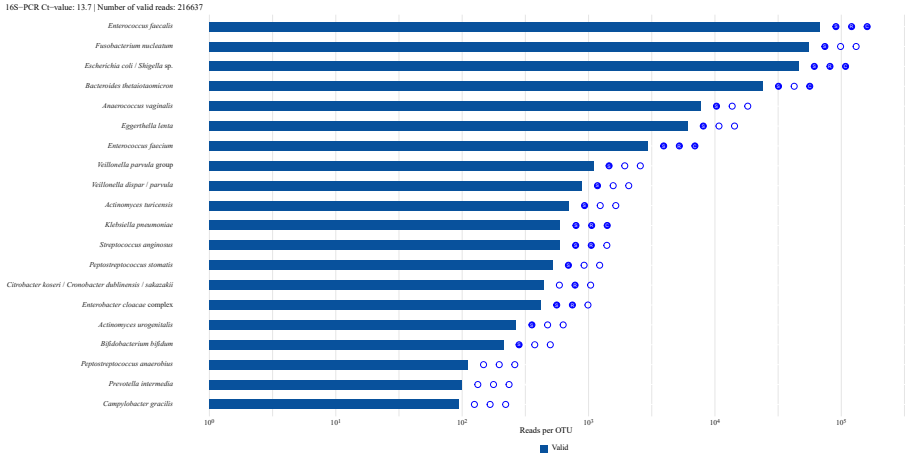
16S−PCR Ct−value: 12.2 | Number of valid reads: 158615



Sample 20

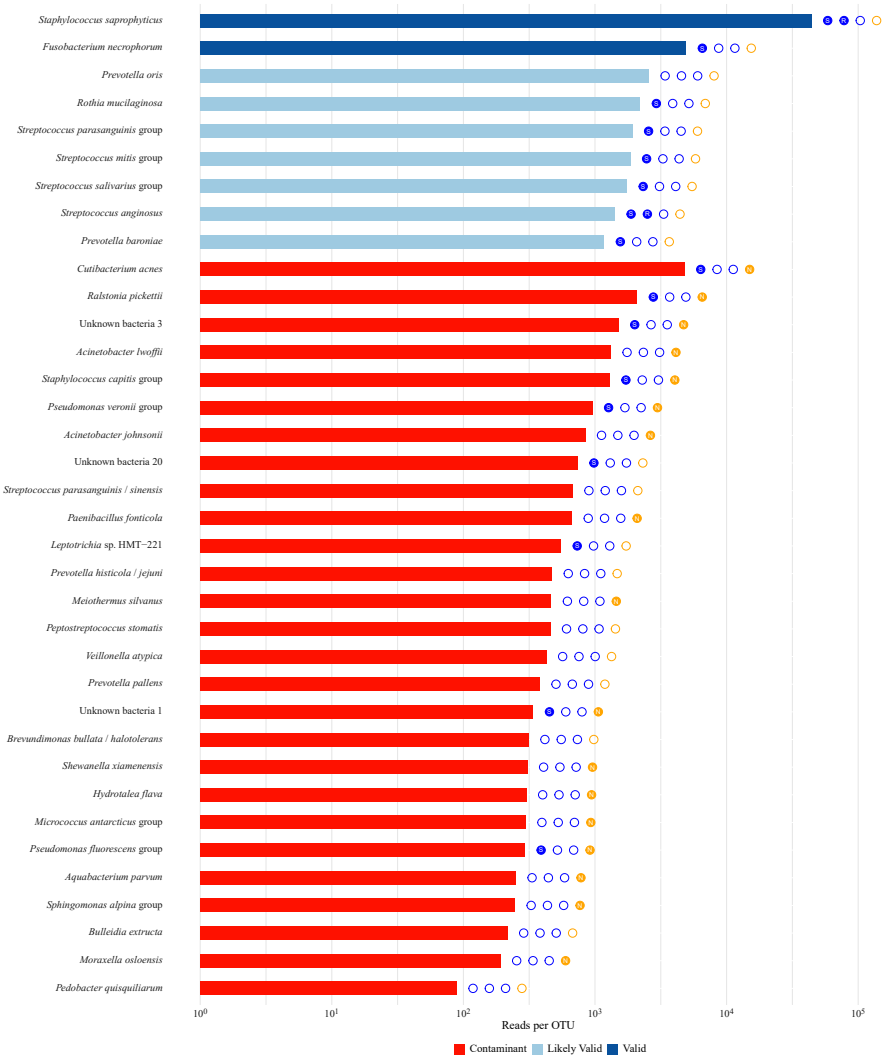


Sample 21



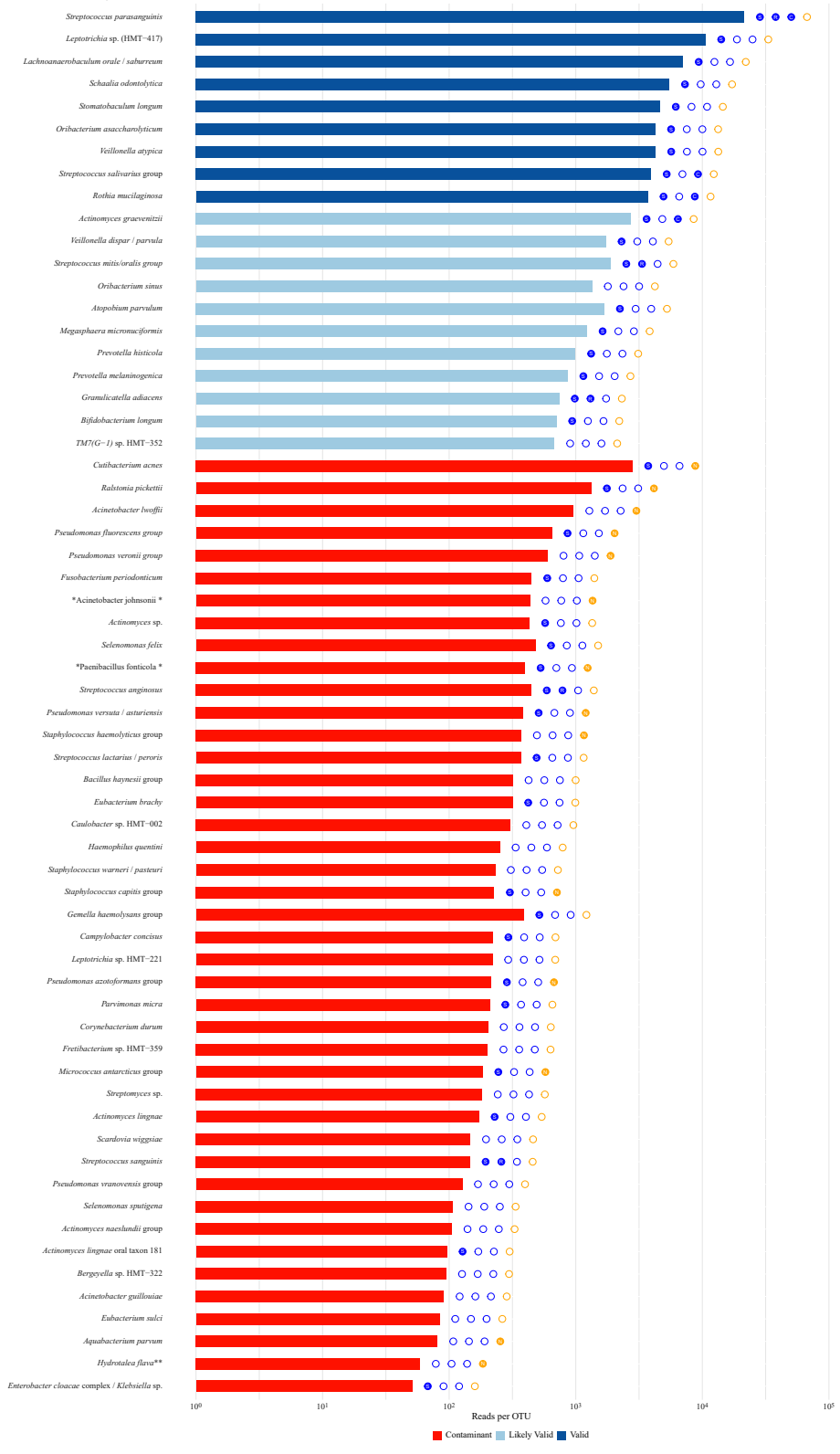
Sample 22

16S−PCR Ct−value: 30.9 | Number of valid reads: 76201

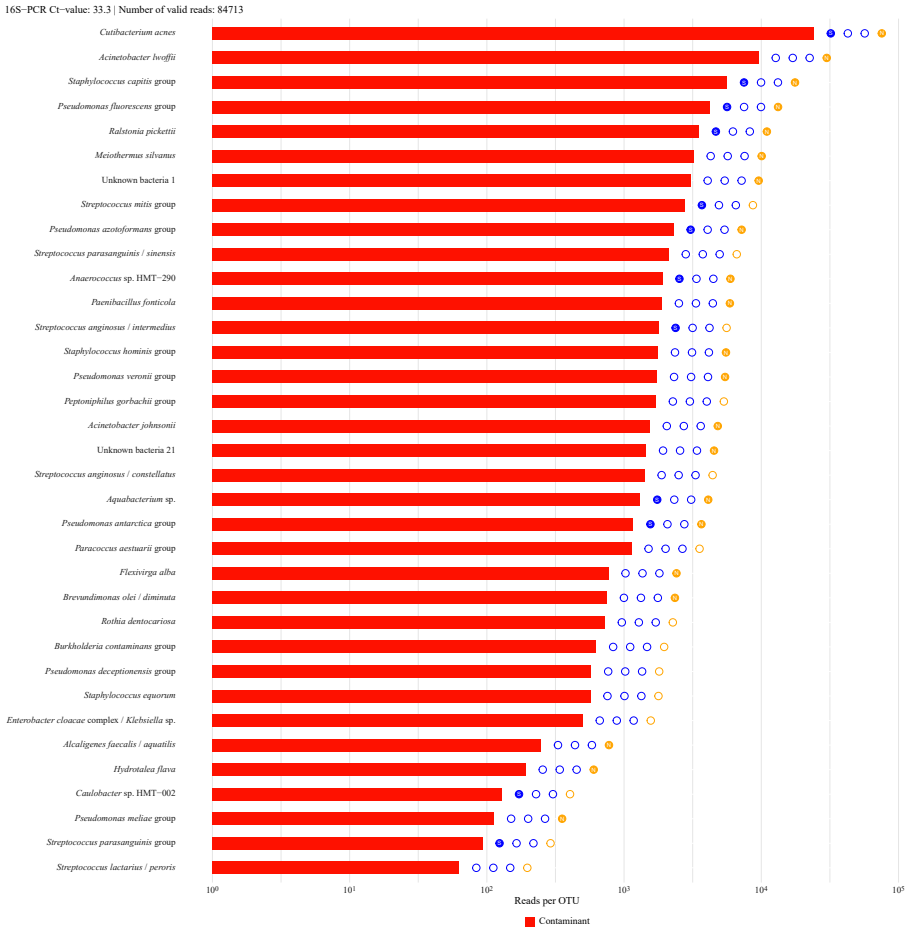


Sample 23

16S−PCR Ct−value: 29.9 | Number of valid reads: 97083

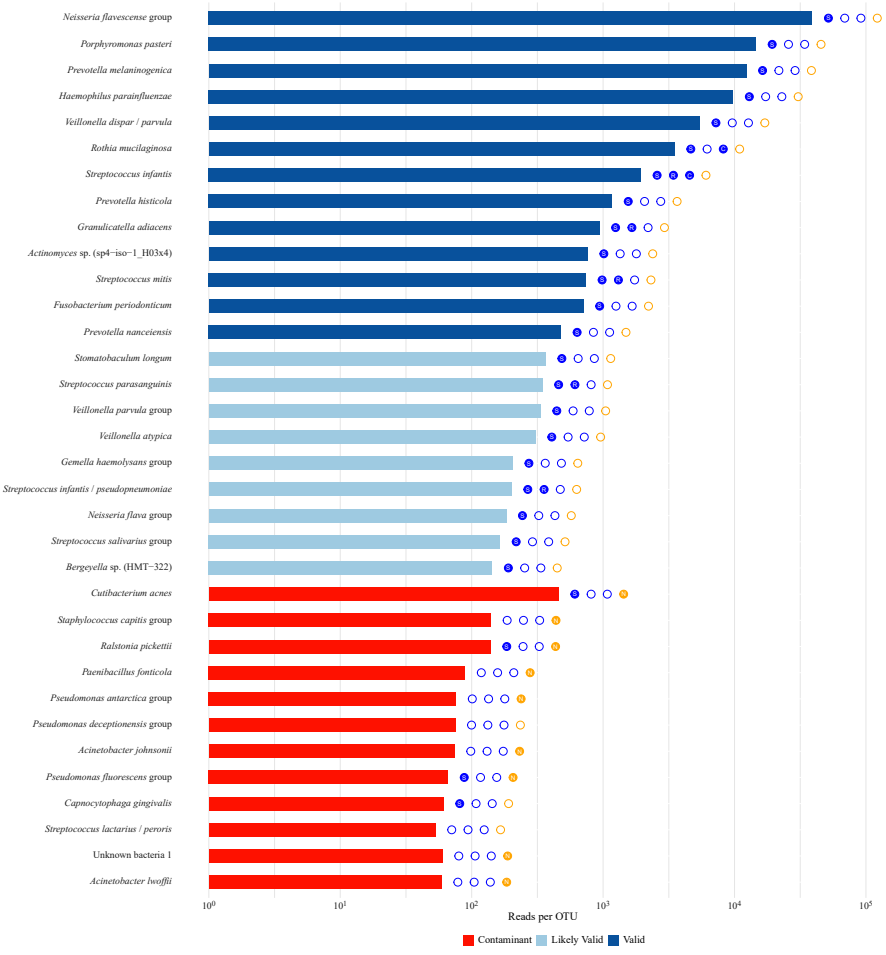


Sample 24


16S−PCR Ct−value: 33.3 | Number of valid reads: 84713
Cutibacterium acnes
Acinetobacter lwoffii
Staphylococcus capitis group
Pseudomonas fluorescens group
Ralstonia pickettii
Meiothermus silvanus
Unknown bacteria 1
Streptococcus mitis group
Pseudomonas azotoformans group
Streptococcus parasanguinis / sinensis
Anaerococcus sp. HMT−290
Paenibacillus fonticola
Streptococcus anginosus / intermedius
Staphylococcus hominis group
Pseudomonas veronii group
Peptoniphilus gorbachii group
Acinetobacter johnsonii
Unknown bacteria 21
Streptococcus anginosus / constellatus
Aquabacterium sp.
Pseudomonas antarctica group
Paracoccus aestuarii group
Flexivirga alba
Brevundimonas olei / diminuta
Rothia dentocariosa
Burkholderia contaminans group
Pseudomonas deceptionensis group
Staphylococcus equorum
Enterobacter cloacae complex / Klebsiella sp.
Alcaligenes faecalis / aquatilis
Hydrotalea flava
Caulobacter sp. HMT−002
Pseudomonas meliae group
Streptococcus parasanguinis group
Streptococcus lactarius / peroris
10⁰
10¹
10²
10³
10⁴
10⁵
Reads per OTU
Contaminant

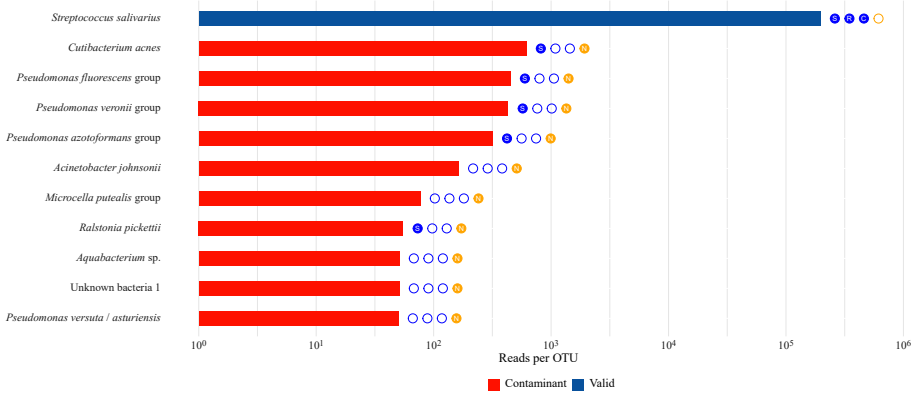
Sample 25

16S−PCR Ct−value: 26.8 | Number of valid reads: 94801



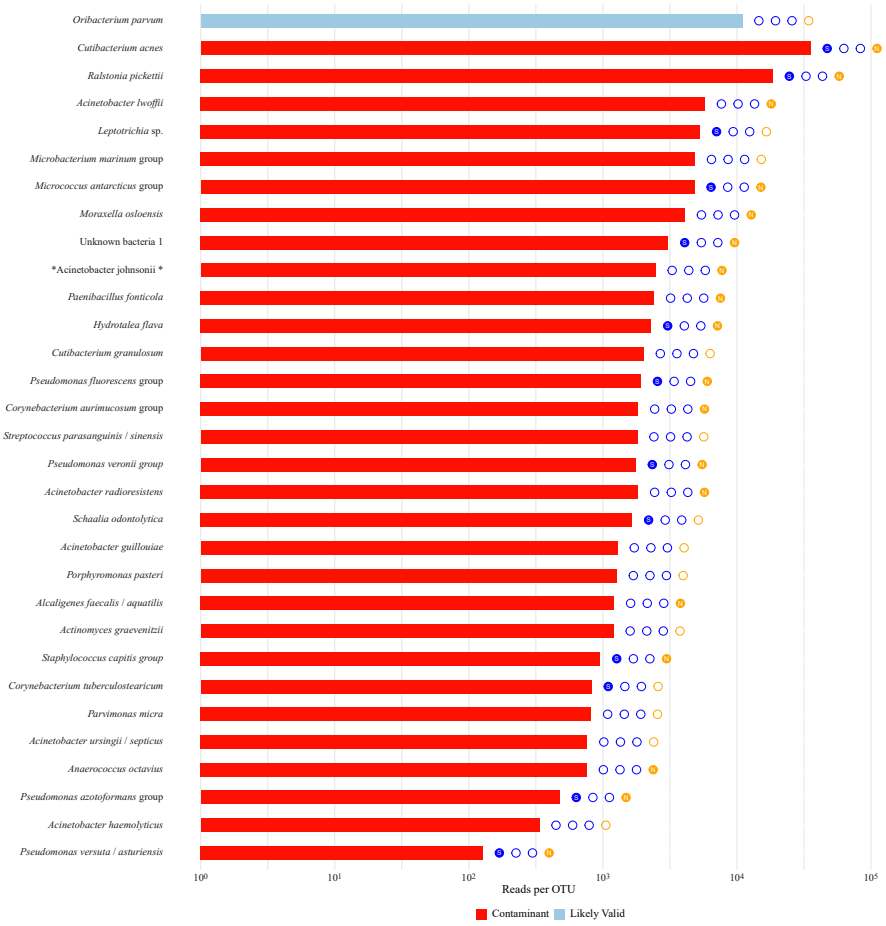
Sample 26

16S−PCR Ct−value: 25.8 | Number of valid reads: 198331



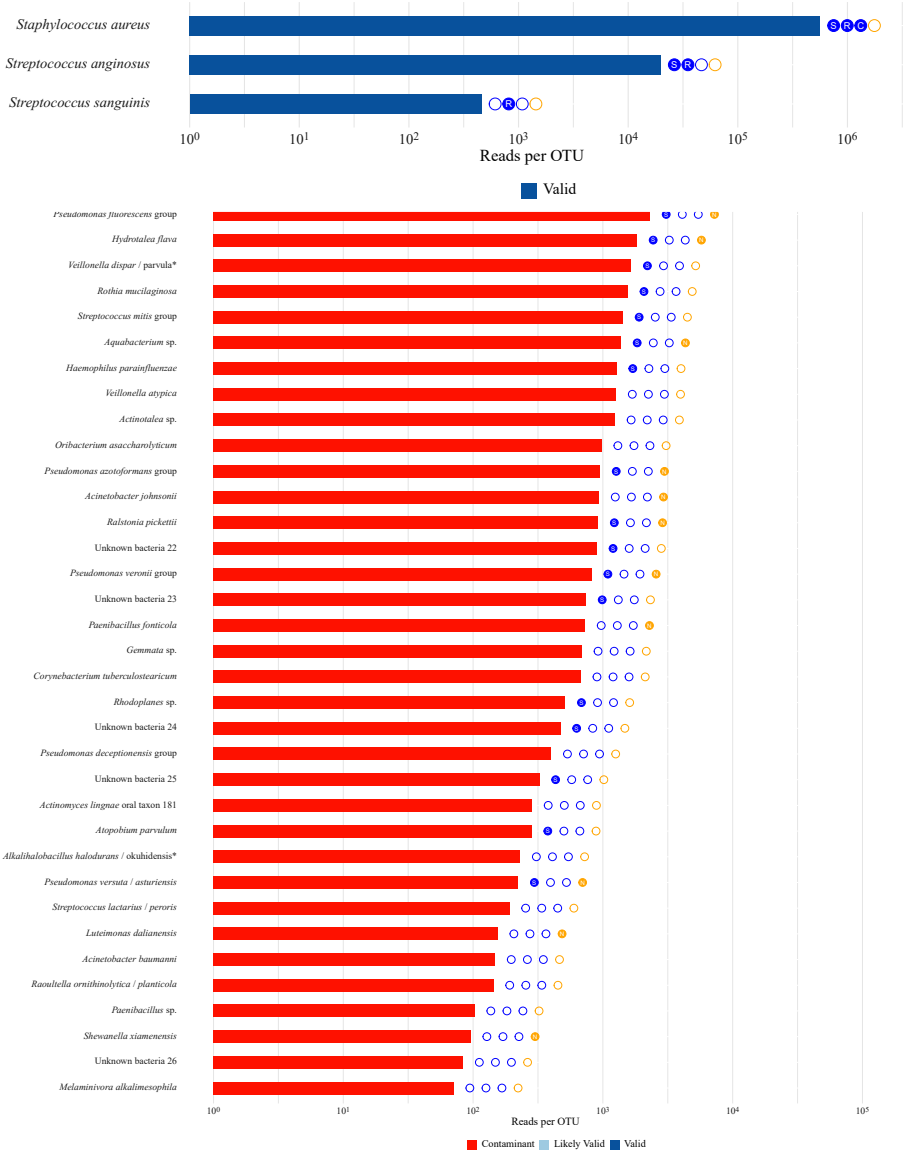
Sample 27

16S−PCR Ct−value: 33.3 | Number of valid reads: 122790



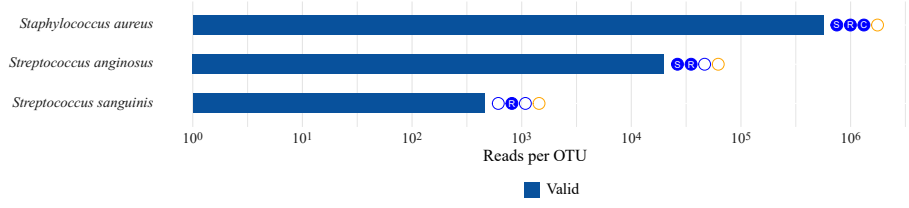
Sample 28

16S−PCR Ct−value: 16.3 | Number of valid reads: 583052

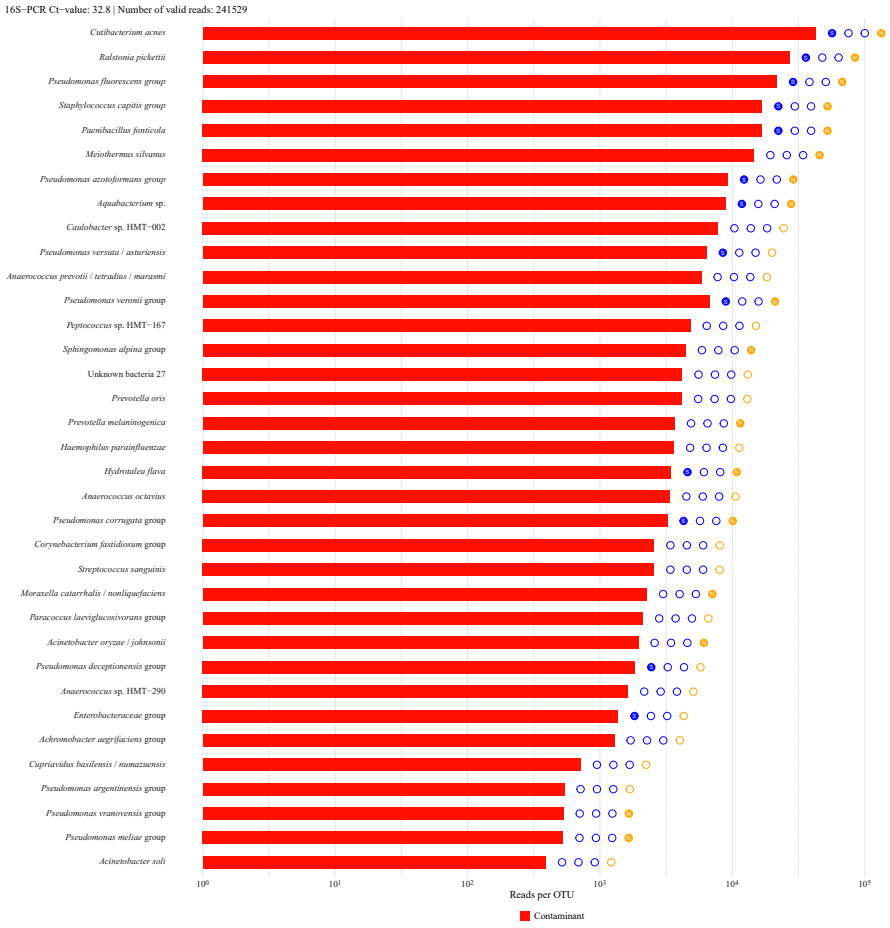


Sample 29

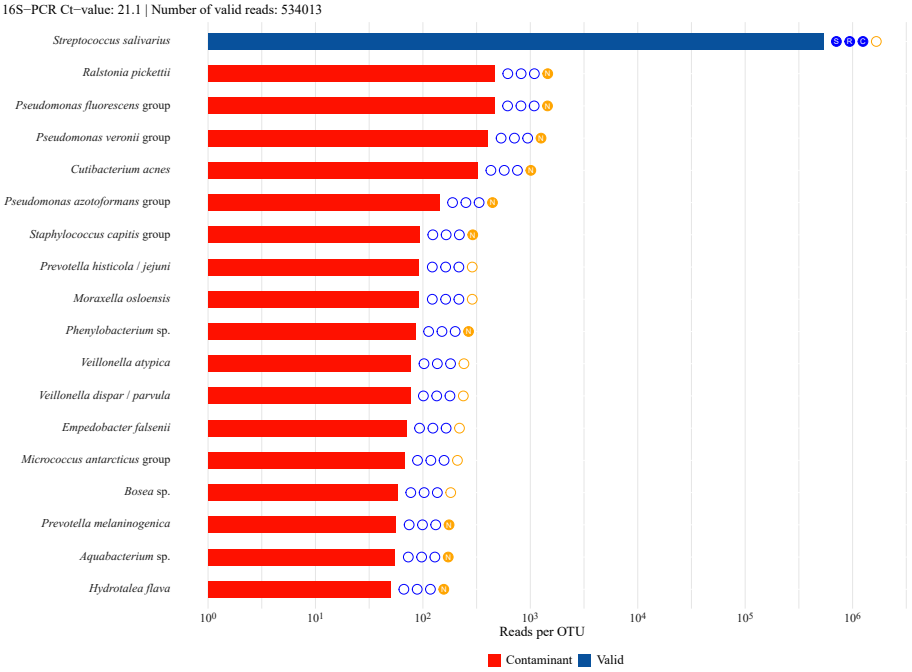
16S−PCR Ct−value: 16.3 | Number of valid reads: 583052



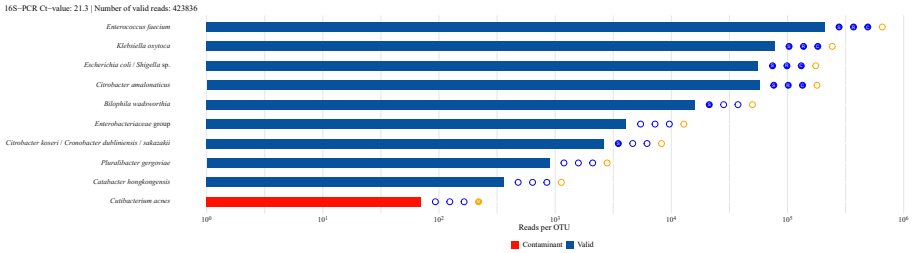
Sample 30


16S−PCR Ct−value: 32.8 | Number of valid reads: 241529
Cutibacterium acnes
Ralstonia pickettii
Pseudomonas fluorescens group
Staphylococcus capitis group
Paenibacillus fonticola
Meiothermus silvanus
Pseudomonas azotoformans group
Aquabacterium sp.
Caulobacter sp. HMT−002
Pseudomonas versuta / asturiensis
Anaerococcus prevotii / tetradius / marasmi
Pseudomonas veronii group
Peptococcus sp. HMT−167
Sphingomonas alpina group
Unknown bacteria 27
Prevotella oris
Prevotella melaninogenica
Haemophilus parainfluenzae
Hydrotalea flava
Anaerococcus octavius
Pseudomonas corrugata group
Corynebacterium fastidiosum group
Streptococcus sanguinis
Moraxella catarrhalis / nonliquefaciens
Paracoccus laeviglucosivorans group
Acinetobacter oryzae / johnsonii
Pseudomonas deceptionensis group
Anaerococcus sp. HMT−290
Enterobacteraceae group
Achromobacter aegrifaciens group
Cupriavidus basilensis / numazuensis
Pseudomonas argentinensis group
Pseudomonas vranovensis group
Pseudomonas meliae group
Acinetobacter soli
10⁰
10¹
10²
10³
10⁴
10⁵
Reads per OTU
Contaminant

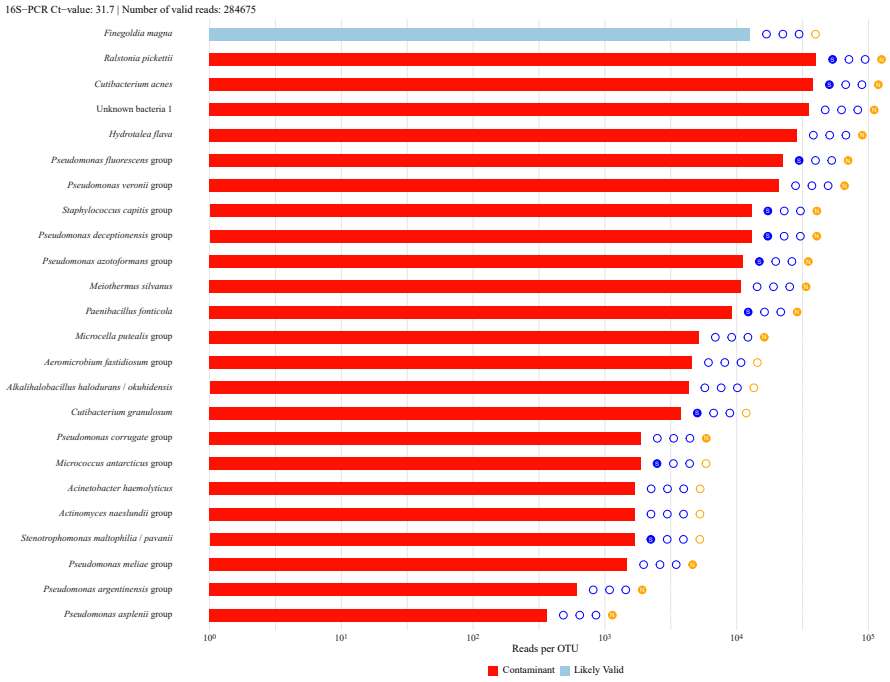
Sample 31



Sample 32

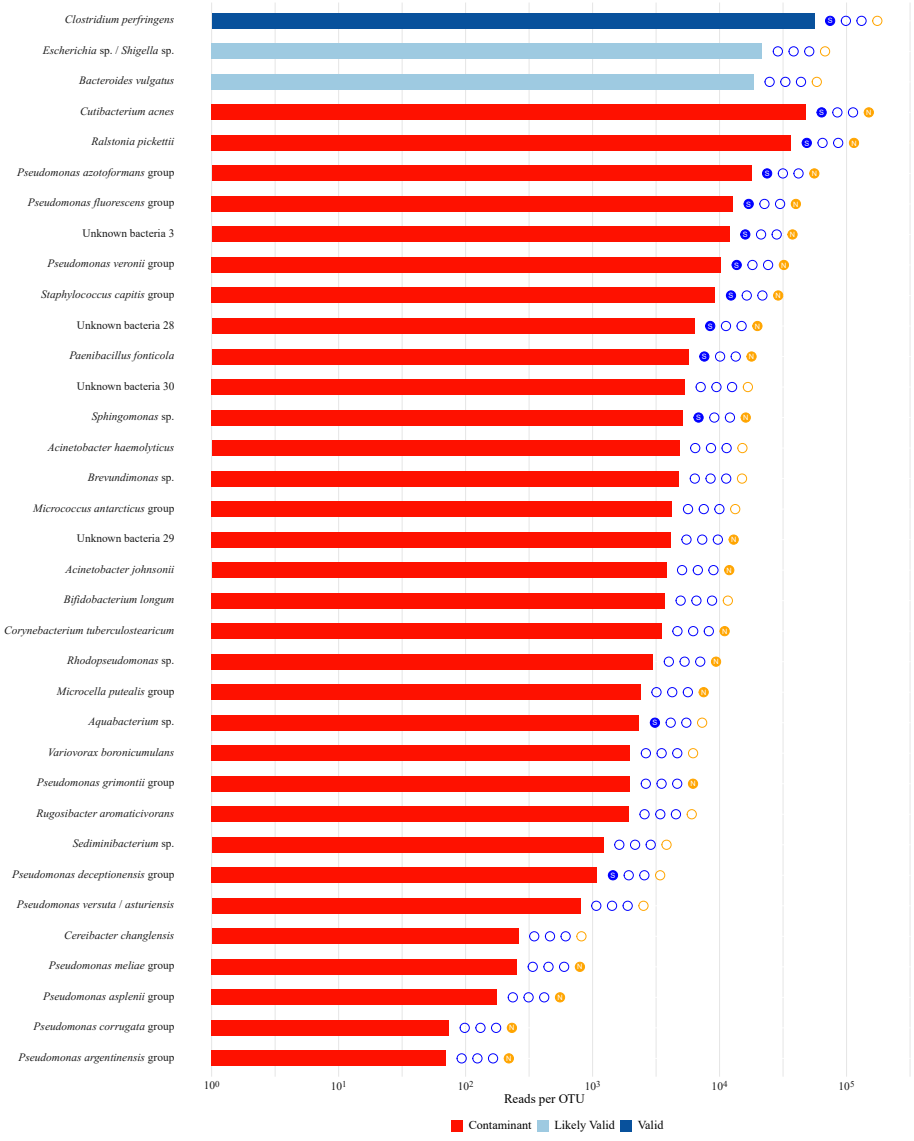


Sample 34


16S−PCR Ct−value: 31.7 | Number of valid reads: 284675
Finegoldia magna
Ralstonia pickettii
Cutibacterium acnes
Unknown bacteria 1
Hydrotalea flava
Pseudomonas fluorescens group
Pseudomonas veronii group
Staphylococcus capitis group
Pseudomonas deceptionensis group
Pseudomonas azotoformans group
Meiothermus silvanus
Paenibacillus fonticola
Microcella putealis group
Aeromicrobium fastidiosum group
Alkalihalobacillus halodurans / okuhidensis
Cutibacterium granulosum
Pseudomonas corrugate group
Micrococcus antarcticus group
Acinetobacter haemolyticus
Actinomyces naeslundii group
Stenotrophomonas maltophilia / pavanii
Pseudomonas meliae group
Pseudomonas argentinensis group
Pseudomonas asplenii group
10⁰
10¹
10²
10³
10⁴
10⁵
Reads per OTU
Contaminant
Likely Valid

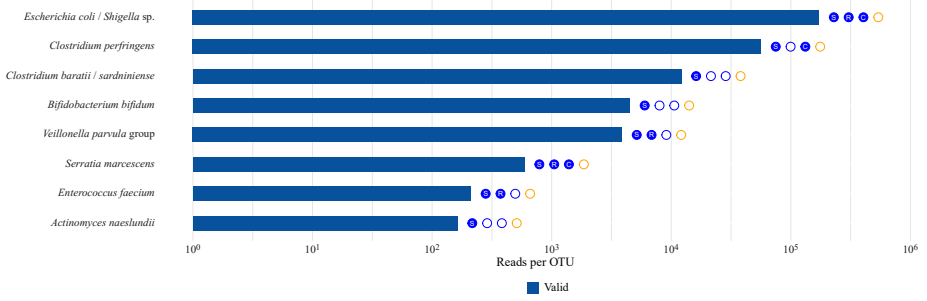
Sample 35

16S−PCR Ct−value: 33 | Number of valid reads: 311709



Sample 36

16S−PCR Ct−value: 12.9 | Number of valid reads: 249750

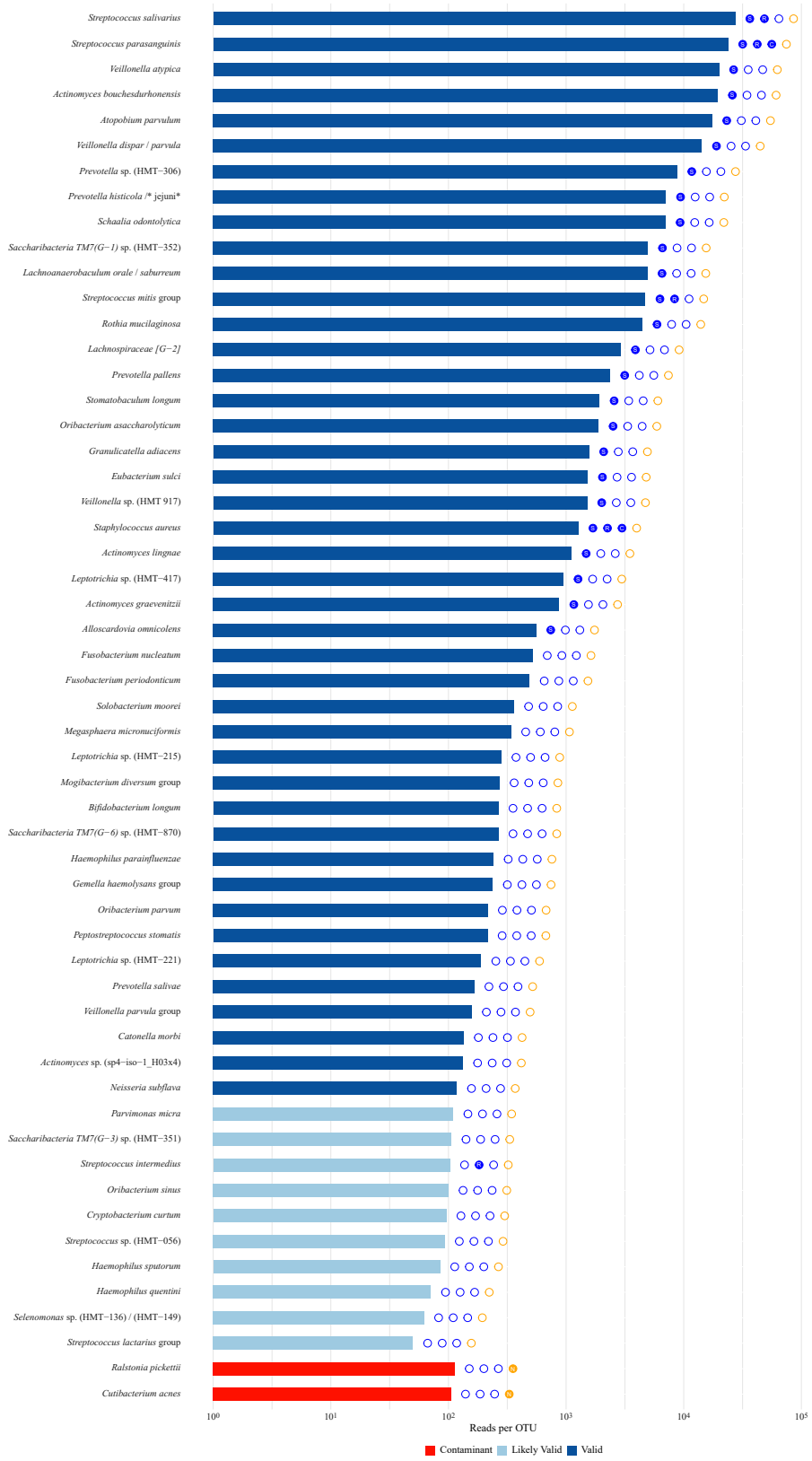


Sample 37

16S−PCR Ct−value: 33.4 | Number of valid reads: 289331

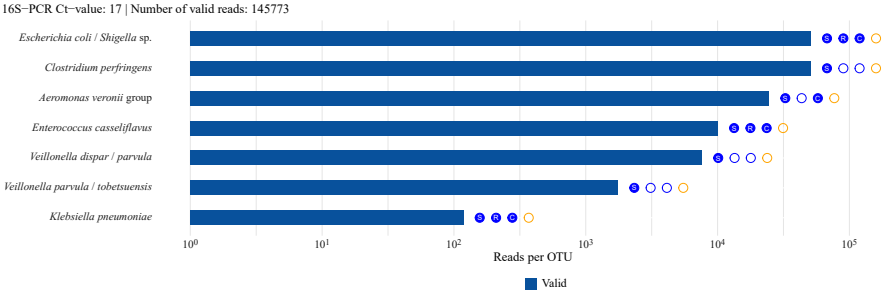
Sample 38

16S−PCR Ct−value: 22.4 | Number of valid reads: 188962



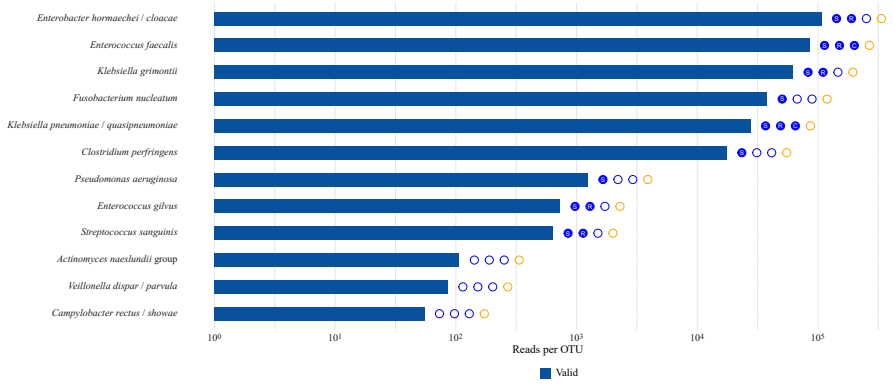
Sample 39

16S−PCR Ct−value: 17 | Number of valid reads: 145773



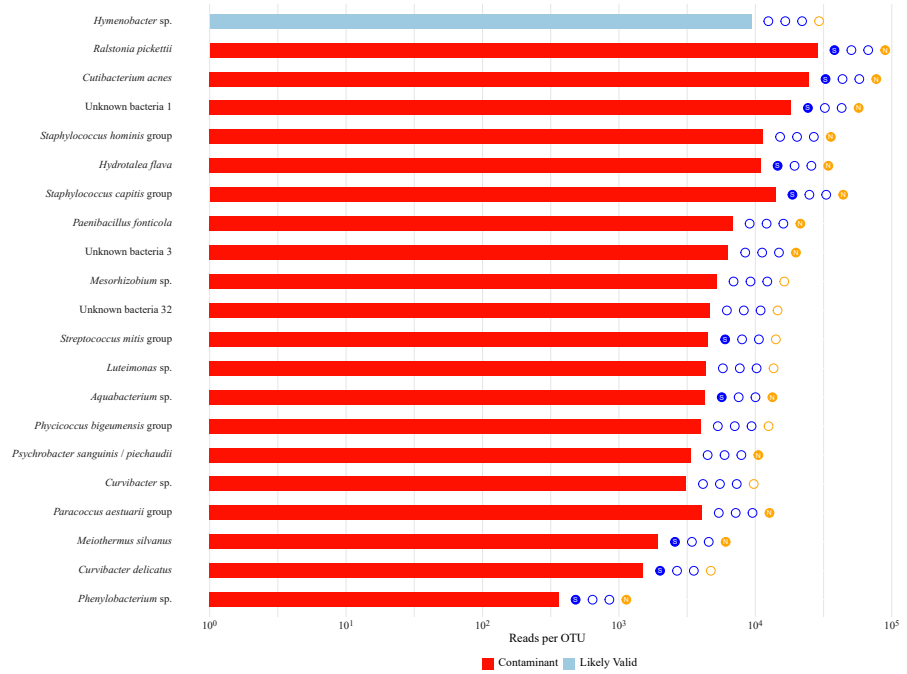
Sample 40

16S−PCR Ct−value: 14.4 | Number of valid reads: 342012



Sample 41

16S−PCR Ct−value: 32.3 | Number of valid reads: 173034



Supplementary Table S1

A) Characteristics and beta-diversity measures of the extraction control replicates in Experiment 1.

	Negative extraction control 1		
	PCR replicates	Sequencing replicates	p-value[a]
Number of samples	5	5	
Accepted reads per sample[b], mean	278,466	322,314	
min-max	227,539-319871	303,225-336,425	
Bacterial identifications per sample (ranged min to max)	14, 16, 19, 24, 25	14, 14, 14, 14, 14	
Jaccard distance[c], mean	0.73	0.00	<0,001
min-max	0.68-0.78		
Bray Curtis dissimilarity[c], mean	0.49	0.04	<0,001
min-max	0.37-0.57	0.01-0.07	

	Negative extraction control 2		
	PCR replicates	Sequencing replicates	p-value[a]
Number of samples	5	5	
Accepted reads per sample[b], mean	290,391	299,105	
min-max	206,617-352,522	227,323-357,574	
Bacterial identifications per sample (ranged min to max)	12, 17, 18, 21, 24	24, 24, 24, 24, 24	
Jaccard distance[c], mean	0.77	0.00	<0,001
min-max	0.65-0.85		
Bray Curtis dissimilarity[c], mean	0.44	0.04	<0,001
min-max	0.36-0.53	0.01-0.07	

	Positive extraction control		
	PCR replicates	Sequencing replicates	p-value[a]
Number of samples	5	4	
Accepted reads per sample[b], mean	313,545	309,860	
min-max	222,116-395,935	256,164-395,935	
Bacterial identifications per sample (ranged min to max)	11, 19, 20, 24, 25	18, 19, 19, 19	
Jaccard distance[c], mean	0.73	0.03	<0,001
min-max	0.65-0.81	0.00-0.05	
Bray Curtis dissimilarity[c], mean	0.44	0.04	<0,001
min-max	0.37-0.51	0.02-0.06	

[a] Students t-test for continuous, normal distributed variables. Mann-Whitney U-test for continuous, skewed variables.

[b] Number of reads per sample after removal of short reads (< 250 base pairs), small clusters (< 50 reads), human reads and chimeras.

[c] Data rarified as described in materials and method

B) Number of accepted reads and alpha-diversity measures for all extraction control replicates in Experiment 1.

		Total number of reads[a]	Accepted reads[b] (n)	Bacterial identifications (n)	Shannon Index[a]	InvSimpson index[a]	Fisher's alpha index[c]
Negative extraction control 1							
PCR replicates							
	1	583348	319871	14	1,98	5,23	1,15
	2	547193	271946	19	2,42	7,93	1,60
	3	470255	227539	16	2,41	8,87	1,33
	4	547230	262352	24	2,64	9,87	2,07
	5	637553	310622	25	2,73	10,60	2,16
Sequencing replicates							
	1	583348	319871	14	1,98	5,23	1,12
	2	573792	324402	14	1,93	4,88	1,12
	3	538117	303646	14	1,95	5,01	1,12
	4	552688	303225	14	2,01	5,30	1,12
	5	651891	360425	14	1,96	5,11	1,12
Negative extraction control 2							
PCR replicates							
	1	686972	338240	24	2,55	7,81	2,09
	2	633095	331274	17	2,12	5,52	1,43
	3	432850	206617	21	2,40	7,00	1,80
	4	444629	223301	18	2,31	7,77	1,52
	5	655334	352522	12	2,11	6,31	0,98
Sequencing replicates							
	1	686972	338240	24	2,55	7,83	2,07
	2	601795	305697	24	2,48	6,91	2,07
	3	553463	266693	24	2,55	7,80	2,07
	4	722975	357574	24	2,62	9,04	2,07
	5	474732	227323	24	2,58	8,41	2,07

Positive extraction control							
PCR replicates							
	1	797249	395935	19	2,28	7,67	1,61
	2	573511	285593	11	1,94	5,83	0,88
	3	737625	352059	24	2,57	8,23	2,07
	4	648329	312020	20	2,69	10,96	1,70
	5	461088	222116	25	2,52	7,42	2,17
Sequencing replicates							
	1	797249	395935	19	2,28	7,66	1,58
	2	530500	256164	19	2,36	8,35	1,58
	3	555651	275015	18	2,25	7,40	1,49
	4	629303	312325	19	2,33	7,85	1,58

[a] Total number of reads: Number of reads per sample before post-sequencing data processing

[b] Accepted reads: Number of reads per sample for after removal of short reads (< 250 base pairs), small clusters (< 50 reads), human reads and chimeras.

[c] Data rarified as described in materials and methods.

C) Sample-to-sample beta diversity measures for all extraction control replicates in Experiment 1.

Negative extraction control 1 - PCR replicates

Jaccard_distance

	replicate 1	replicate 2	replicate 3	replicate 4
replicate 2	0,68			
replicate 3	0,75	0,75		
replicate 4	0,73	0,74	0,75	
replicate 5	0,78	0,74	0,68	0,71

Bray_curtis_dissimilarity

	replicate 1	replicate 2	replicate 3	replicate 4
replicate 2	0,37			
replicate 3	0,48	0,48		
replicate 4	0,49	0,47	0,54	
replicate 5	0,57	0,48	0,51	0,52

Negative extraction control 1 - Sequencing replicates

Jaccard_distance

	replicate 1	replicate 2	replicate 3	replicate 4
replicate 2	0,00			
replicate 3	0,00	0,00		
replicate 4	0,00	0,00	0,00	
replicate 5	0,00	0,00	0,00	0,00

Bray_curtis_distance

	replicate 1	replicate 2	replicate 3	replicate 4
replicate 2	0,03			
replicate 3	0,02	0,01		
replicate 4	0,06	0,07	0,06	
replicate 5	0,01	0,02	0,01	0,07

Negative extraction control 2 - PCR replicates

Jaccard_distance

	replicate 1	replicate 2	replicate 3	replicate 4
replicate 2	0,79			
replicate 3	0,71	0,73		
replicate 4	0,83	0,65	0,78	
replicate 5	0,76	0,79	0,78	0,85

Bray_curtis_distance

	replicate 1	replicate 2	replicate 3	replicate 4
replicate 2	0,45			
replicate 3	0,38	0,38		
replicate 4	0,53	0,41	0,52	
replicate 5	0,36	0,40	0,42	0,51

Negative extraction control 2 - Sequencing replicates

Jaccard_distance

	replicate 1	replicate 2	replicate 3	replicate 4
replicate 2	0,00			
replicate 3	0,00	0,00		
replicate 4	0,00	0,00	0,00	
replicate 5	0,00	0,00	0,00	0,00

Bray_curtis_distance

	replicate 1	replicate 2	replicate 3	replicate 4
replicate 2	0,04			
replicate 3	0,01	0,03		
replicate 4	0,04	0,07	0,04	
replicate 5	0,03	0,06	0,04	0,03

Positive extraction control - PCR replicates

Jaccard_distance

	replicate 1	replicate 2	replicate 3	replicate 4
replicate 2	0,70			
replicate 3	0,81	0,79		
replicate 4	0,74	0,65	0,71	
replicate 5	0,78	0,76	0,68	0,68

Bray_curtis_distance

	replicate 1	replicate 2	replicate 3	replicate 4
replicate 2	0,43			
replicate 3	0,45	0,45		
replicate 4	0,51	0,40	0,46	
replicate 5	0,46	0,47	0,37	0,40

Positive extraction control - Sequencing replicates

Jaccard_distance

	replicate 1	replicate 2	replicate 3
replicate 2	0,00		
replicate 3	0,05	0,05	
replicate 4	0,00	0,00	0,05

Bray_curtis_distance

	replicate 1	replicate 2	replicate 3
replicate 2	0,05		
replicate 3	0,02	0,06	
replicate 4	0,05	0,04	0,04

[a] Data rarified as described in materials and methods.

Supplementary Table S3: Theoretical microbial composition and theoretical abundance of each bacteria for different mock community dilutions used in Experiment 2

A) Undiluted mock community. Data from producer of the staggered mock community.

	Theoretical composition (%)			Theoretical abundance of each microbe
Species	**Genomic DNA**	**16S Only**	**Cell Number**	**Cell (n) per ml**
Faecalibacterium prausnitzii	14	17,63	14,82	583908000
Veillonella rogosae	14	15,87	20,01	788394000
Roseburia hominis	14	9,89	12,47	491318000
Bacteroides fragilis	14	9,94	8,36	329384000
Prevotella corporis	6	4,98	6,28	247432000
Bifidobacterium adolescentis	6	8,78	8,86	349084000
Fusobacterium nucleatum	6	7,49	7,56	297864000
Lactobacillus fermentum	6	9,63	9,71	382574000
Clostridioides difficile	1,5	2,62	1,10	43340000
Akkermansia muciniphila	1,5	0,97	1,62	63828000
Methanobrevibacter smithii	0,1	0,066	0,17	6698000
Salmonella enterica	0,01	0,009	0,01	256100
Enterococcus faecalis	0,001	0,0009	0,00	43340
Clostridium perfringens	0,0001	0,0002	0,00	3546
Escherichia coli	14	12,12	8,73	343962000
Saccharomyces cerevisiae	1,4	N/A	0,16	6304000
Candida albicans	1,5	N/A	0,16	6304000
Sum	100		100	3940696986

B) 1:10 dilution of mock community

	Theoretical abundance of each bacteria			
Species	**Cell (n) per ml**	**Cell (n) input in DNA extraction:**	**16S copies in 100 µl extraction eluate**	**16S copies in 2 µl extraction eluate**
Faecalibacterium prausnitzii	58390800	14597700	87586200	1751724
Veillonella rogosae	78839400	19709850	78839400	1576788
Roseburia hominis	49131800	12282950	49131800	982636
Bacteroides fragilis	32938400	8234600	49407600	988152
Prevotella corporis	24743200	6185800	24743200	494864
Bifidobacterium adolescentis	34908400	8727100	43635500	872710
Fusobacterium nucleatum	29786400	7446600	37233000	744660
Lactobacillus fermentum	38257400	9564350	47821750	956435
Clostridioides difficile	4334000	1083500	13002000	260040
Akkermansia muciniphila	6382800	1595700	4787100	95742
Methanobrevibacter smithii	669800	167450	334900	6698
Salmonella enterica	25610	6403	44818	896
Enterococcus faecalis	4334	1084	4334	87
Clostridium perfringens	355	89	887	18

Escherichia coli	34396200	8599050	60193350	1203867
Sum	392808899	98202225	496765838	9935317

C) 1:10^5 dilution of mock community

	Theoretical abundance of each bacteria			
Species	**Cell (n) per ml**	**Cell input in DNA extration:**	**16S copies in 100 µl extraction eulate**	**16S copies in 2 µl extraction eluate**
Faecalibacterium prausnitzii	5839	1460	8759	175
Veillonella rogosae	7884	1971	7884	158
Roseburia hominis	4913	1228	4913	98
Bacteroides fragilis	3294	823	4941	99
Prevotella corporis	2474	619	2474	49
Bifidobacterium adolescentis	3491	873	4364	87
Fusobacterium nucleatum	2979	745	3723	74
Lactobacillus fermentum	3826	956	4782	96
Clostridioides difficile	433	108	1300	26
Akkermansia muciniphila	638	160	479	10
Methanobrevibacter smithii	67	17	33	0,7
Salmonella enterica	2,6	0,6	4,5	0,1
Enterococcus faecalis	0	0	0	0,0
Clostridium perfringens	0	0	0	0,0
Escherichia coli	3440	860	6019	120
Sum	39281	9820	49677	994

D) 1:10^6 dilution of mock community

	Theoretical abundance of each bacteria			
Species	**Cell (n) per ml**	**Cell input in DNA extration:**	**16S copies in 100 µl extraction eulate**	**16S copies in 2 µl extraction eluate**
Faecalibacterium prausnitzii	584	146	876	17,5
Veillonella rogosae	788	197	788	15,8
Roseburia hominis	491	123	491	9,8
Bacteroides fragilis	329	82	494	9,9
Prevotella corporis	247	62	247	4,9
Bifidobacterium adolescentis	349	87	436	8,7
Fusobacterium nucleatum	298	74	372	7,4
Lactobacillus fermentum	383	96	478	9,6
Clostridioides difficile	43	11	130	2,6
Akkermansia muciniphila	64	16	48	1,0
Methanobrevibacter smithii	7	2	3	0,1
Salmonella enterica	0	0	0	0,0
Enterococcus faecalis	0	0	0	0,0
Clostridium perfringens	0	0	0	0,0

Escherichia coli	344	86	602	12,0
Sum	3928	982	4968	99

Supplementary Table S7: Species identified at a higher taxonomic level with use of partial rpoB -gene compared to partial 16S rRNA gene sequencing (V3–V4).

	16S rRNA gene sequencing results	**rpoB- gene sequencing results**
1	*Citrobacter amalonaticus / Citrobacter farmeri*	*Citrobacter amalonaticus*
2	*Enterobacter hormaechei / Enterobacter cloacae / Klebsiella grimontii / Klebsiella michiganensis / Klebsiella oxytoca*	*Enterobacter cloacae / Enterobacter hormaechei*
3	*Enterobacter kobei / Enterobacter ludwigii / Enterobacter cloacae / Salmonella enterica*	*Enterobacter cloacae* complex
4	*Enterococcus avium / Enterococcus devriesei / Enterococcus gilvus / Enterococcus pseudoavium / Enterococcus viikkiensis / Enterococcus canintestini / Enterococcus gilvus / Enterococcus malodoratus / Enterococcus raffinosus / Enterococcus xianfangensis / Enterococcus casseliflavus / Enterococcus dispar / Enterococcus hermanniensis*	*Enterococcus gilvus*
5	*Enterococcus casseliflavus / Enterococcus canintestini / Enterococcus gallinarum / Enterococcus saigonensis / Enterococcus devriesei / Enterococcus dispar / Enterococcus gilvus / Enterococcus viikkiensis*	*Enterococcus casseliflavus*
6	*Enterococcus durans / Enterococcus faecium / Enterococcus hirae / Enterococcus sanguinicola / Enterococcus casseliflavus / Enterococcus mundtii / Enterococcus ratti / Enterococcus villorum*	*Enterococcus faecium*
7	*Escherichia albertii / Escherichia coli / Escherichia fergusonii / Escherichia marmotae / Shigella* sp.	*Escherichia coli / Shigella* sp.
8	*Klebsiella grimontii / Klebsiella oxytoca / Salmonella enterica*	*Klebsiella grimontii*
9	*Klebsiella michiganensis / Klebsiella oxytoca / Enterobacter asburiae / Enterobacter hormaechei / Enterobacter cloacae*	*Klebsiella oxytoca*
10	*Klebsiella oxytoca / Salmonella enterica*	*Klebsiella oxytoca*
11	*Klebsiella pneumoniae / Klebsiella quasipneumoniae / Enterobacter asburiae / Enterobacter bugandensis / Enterobacter cancerogenus / Enterobacter cloacae / Enterobacter hormaechei*	*Klebsiella pneumoniae*
12	*Klebsiella pneumoniae / Klebsiella variicola*	*Klebsiella pneumoniae*
13	*Klebsiella pneumoniae / Klebsiella variicola*	*Klebsiella variicola*
14	*Proteus cibarius / Proteus hauseri / Proteus terrae / Proteus vulgaris*	*Proteus vulgaris / Proteus hauseri*

15	*Serratia odorifera / Yersinia enterocolitica / Yersinia rohdei*	*Serratia odorifera*
16	*Staphylococcus aureus / Staphylococcus croceolyticus / Staphylococcus petrasii*	*Staphylococcus aureus*
17	*Staphylococcus haemolyticus / Staphylococcus croceolyticus / Staphylococcus petrasii / Staphylococcus epidermidis / Staphylococcus capitis / Staphylococcus caprae / Staphylococcus hominis*	*Staphylococcus haemolyticus*
18	*Staphylococcus saprophyticus / Staphylococcus xylosus / Staphylococcus gallinarum*	*Staphylococcus saprophyticus*
19	*Staphylococcus warneri / Staphylococcus pasteuri*	*Staphylococcus warneri*
20	*Streptococcus anginosus / Streptococcus intermedius*	*Streptococcus intermedius*
21	*Streptococcus anginosus / Streptococcus intermedius*	*Streptococcus anginosus*
22	*Streptococcus mitis/oralis group*	*Streptococcus infantis*
23	*Streptococcus mitis/oralis group*	*Streptococcus mitis*
24	*Streptococcus mitis/oralis group*	*Streptococcus oralis*

Supplementary Table S8:

A) Primers with adapter sequences (1). Sequences of the target specific portions in capital letters.

Name	Sequence	Position[a]
16S-F[b]	tcgtcggcagcgtcagatgtgtataagagacagCCTACGGGNGGCWGCAG	340-356
16S-R[b]	gtctcgtgggctcggagatgtgtataagagacagGACTACCAGGGTATCTAAKCC	784-803
rpoB_Ent-F	tcgtcggcagcgtcagatgtgtataagagacagGAAGGTCCRAAYATCGGTCT	1693-1712
rpoB_Ent-R	gtctcgtgggctcggagatgtgtataagagacagTGCATGTTCGCACCCAT	2041-2057
rpoB_ESS-F1	tcgtcggcagcgtcagatgtgtataagagacagGCRACAGCRTGTATYCCRTTC	1861-1881
rpoB_ESS-F2[c]	tcgtcggcagcgtcagatgtgtataagagacagGCDACMGCWTGTATYCCWTTC	1861-1881
rpoB_ESS-R	gtctcgtgggctcggagatgtgtataagagacagGTTRTAMCCNTCCCAWGTCAT	2287-2307

[a] Positions for 16S based on *Escherichia coli* (GenBank accession J01859). Positions for RpoB_ESS based on *Staphylococcus aureus* [*rpoB* coding sequence (CDS); GenBank accession X64172]. Positions for RpoB_Ent based on *Escherichia coli* [*rpoB* coding sequence (CDS); GenBank accession V00340].

[b] Abbreviations: F = forward primer. R = reverse primer.

[c] The modifications in rpoB_ESS, which were made for better coverage of *Enterococcus raffinosus,* are marked with red.

1. Dyrhovden R, Ovrebo KK, Nordahl MV, Nygaard RM, Ulvestad E, Kommedal O. 2019. Bacteria and fungi in acute cholecystitis. A prospective study comparing next generation sequencing to culture. J Infect doi:10.1016/j.jinf.2019.09.015.

B) PCR mixture for the different gene targets and the temperature profile of the amplicon PCR

Target gene	Primer name	Concentration and volume - primer	Volume - enzyme (µl)	Volume - H2O (µl)	Volume - template (µl)	Temperature profile (all targets)
16S rRNA, V3-V4	16S-F	0,4 µM/ 1,0 µl	12,5	8,5	2,0	95 °C for 3 min (activation) 45 cycles of: - 95 °C for 60 s (melting) - 60 °C for 30 s (annealing) - 72 °C for 30 s (annealing) Melting curve analysis: - 95 °C for 60 s - 40 °C for 2 min - 95 °C continous 40 °C for 30 s (cooling)
	16S-R	0,4 µM/ 1,0 µl				
*rpoB*_Ent *(*targeting Enterobacteriaceae)	*rpoB*_Ent-F	0,4 µM/ 1,0 µl	12,5	8,5	2,0	
	*rpoB*_Ent-R	0,4 µM/ 1,0 µl				
*rpoB*_ESS (targeting *Stapyhlococcus, Streptococcus* and *Enterococcus)*	*rpoB*_ESS-F1	0,4 µM/ 1,0 µl	12,5	8,0	2,0	
	*rpoB*_ESS-F2	0,4 µM/ 1,0 µl				
	*rpoB*_ESS-R	0,6 µM/ 1,5 µl				

Supplementary Table S9: Species included in species groups (S9A), and best BLAST search match for bacteria termed as "Unknown bacteria #" (S9B).

A)

Name of species group	Species included
Achromobacter aegrifaciens group	Achromobacter aegrifaciens / Achromobacter insuavis / Achromobacter marplatensis / Achromobacter deleyi / Achromobacter spanius / Achromobacter kerstersii / Achromobacter pestifer / Achromobacter piechaudii / Achromobacter xylosoxidans
Acidovorax temperans group	Acidovorax temperans / Acidovorax delafieldii / Acidovorax facilis / Acidovorax radicis
Actinomyces naeslundii group	Actinomyces naeslundii / Actinomyces johnsonii / Actinomyces oralis / Actinomyces oris
Actinomyces oris group	Actinomyces oris / Actinomyces viscosus / Actinomyces sp. / Actinomyces johnsonii / Actinomyces oralis / Actinomyces naeslundii / Actinomyces bowdenii
Aeromicrobium fastidiosum group	Aeromicrobium fastidiosum / Aeromicrobium ginsengisoli / Aeromicrobium erythreum / Aeromicrobium panaciterrae / Aeromicrobium choanae / Aeromicrobium tamlense
Aeromonas veronii group	Aeromonas veronii / Aeromonas allosaccharophila / Aeromonas australiensis / Aeromonas lacus / Aeromonas rivipollensis
Agrococcus citreus group	Agrococcus citreus / Agrococcus jenensis / Agrococcus baldri / Agrococcus carbonis
Agrobacterium arsenijevicii group	Agrobacterium arsenijevicii / Agrobacterium deltaense / Agrobacterium tumefaciens / Beijerinckia fluminensis / Rhizobium nepotum / Agrobacterium rubi / Neorhizobium alkalisoli / Neorhizobium huautlense / Rhizobium loessense / Rhizobium skierniewicense
Aquabacterium parvum group	Aquabacterium parvum / Imtechium assamiensis / Aquabacterium commune
Bacillus aerius group	Bacillus aerius / Bacillus aerophilus / Bacillus altitudinis / Bacillus stratosphericus / Bacillus xiamenensis / Bacillus australimaris / Bacillus pumilus / Bacillus safensis / Bacillus zhangzhouensis / Peribacillus acanthi
Bacillus haynesii group	Bacillus haynesii / Bacillus licheniformis / Bacillus paralicheniformis / Bacillus piscis / Bacillus sonorensis / Bacillus subtilis / Bacillus glycinifermentans / Bacillus oryzaecorticis / Bacillus swezeyi
Blastococcus litoris group	Blastococcus litoris / Blastococcus colisei / Blastococcus deserti / Blastococcus jejuensis
Brachybacterium horti group	Brachybacterium horti / Brachybacterium nesterenkovii / Brachybacterium rhamnosum / Brachybacterium endophyticum / Brachybacterium zhongshanense / Brachybacterium sacelli / Brachybacterium timonense
Bradyrhizobium canariense group	Bradyrhizobium sp. / Afipia sp. / Pseudomonas carboxydohydrogena / Nitrobacter sp. / Oligotropha carboxidovorans

Burkholderia contaminans group	Burkholderia contaminans / Burkholderia lata / Burkholderia paludis / Burkholderia cenocepacia / Burkholderia cepacia / Burkholderia latens / Burkholderia territorii / Burkholderia ubonensis / Burkholderia vietnamiensis / Burkholderia arboris / Burkholderia metallica / Burkholderia puraquae / Burkholderia seminalis / Burkholderia stabilis
Cellulosimicrobium cellulans group	Cellulosimicrobium cellulans / Cellulosimicrobium funkei / Cellulosimicrobium terreum / Cellulosimicrobium aquatile / Luteimicrobium subarcticum
Chryseobacterium hominis group	Chryseobacterium hominis / Chryseobacterium arachidiradicis / Chryseobacterium bovis
Corynebacterium accolens group	Corynebacterium accolens / Corynebacterium fastidiosum / Corynebacterium macginleyi / Corynebacterium segmentosum / Corynebacterium tuberculostearicum
Corynebacterium amycolatum group	Corynebacterium amycolatum / Corynebacterium lactis / Corynebacterium freneyi / Corynebacterium xerosis
Corynebacterium lipophiloflavum group	Corynebacterium lipophiloflavum / Corynebacterium sanguinis / Corynebacterium senegalense / Corynebacterium mycetoides
Corynebacterium aurimucosum group	Corynebacterium aurimucosum / Corynebacterium minutissimum / Corynebacterium singulare / Corynebacterium phoceense
Corynebacterium fastidiosum group	Corynebacterium fastidiosum / Corynebacterium accolens / Corynebacterium macginleyi / Corynebacterium segmentosum
Corynebacterium gottingense group	Corynebacterium gottingense / Corynebacterium hadale / Corynebacterium imitans / Corynebacterium godavarianum
Corynebacterium mucifaciens group	Corynebacterium mucifaciens / Corynebacterium fournierii / Corynebacterium ihumii / Corynebacterium ureicelerivorans / Corynebacterium pilbarense
Conyzicola nivalis group	Conyzicola nivalis / Cryobacterium tepidiphilum / Cryobacterium mesophilum / Cryobacterium arcticum / Cryobacterium zongtaii / Klugiella xanthotipulae / Leifsonia kafniensis / Leifsonia psychrotolerans / Pseudolysinimonas kribbensis / Chryseoglobus frigidaquae / Glaciihabitans arcticus / Homoserinibacter gongjuensis / Leifsonia poae / Protaetiibacter intestinalis
Corynebacterium pilbarense group	Corynebacterium pilbarense / Corynebacterium mucifaciens / Corynebacterium coyleae / Corynebacterium afermentans / Corynebacterium uhumii
Corynebacterium vitaeruminis group	Corynebacterium vitaeruminis / Corynebacterium ulcerans / Corynebacterium pseudotuberculosis
Deinococcus grandis group:	Deinococcus grandis / Deinococcus soli / Deinococcus daejeonensis / Deinococcus radiotolerans
Dermacoccus nishinomiyaensis group	Dermacoccus nishinomiyaensis / Dermacoccus abyssi / Dermacoccus barathri / Dermacoccus profundi
Dietzia kunjamensis group	Dietzia kunjamensis / Dietzia maris / Dietzia schimae / Dietzia alimentaria
Enterobacteriaceae group	Enterobacter cloacae complex / Salmonella enterica / Leclercia adecarboxylata / Morganella morganii / Pantoea agglomerans

Enterococcus casseliflavus group	Enterococcus casseliflavus / Enterococcus canintestini / Enterococcus gallinarum / Enterococcus saigonensis / Enterococcus casseliflavus / Enterococcus devriesei / Enterococcus dispar / Enterococcus gilvus / Enterococcus pseudoavium / Enterococcus viikiensis
Escherichia albertii group	Escherichia albertii / Escherichia coli / Pseudescherichia vulneris / Escherichia fergusonii / Escherichia marmotae / Shigella sp.
Friedmanniella okinawensis group	Friedmanniella okinawensis / Friedmanniella sagamiharensis / Friedmanniella spumicola
Gemella haemolysans group	Gemella haemolysans / Gemella taiwanensis / Gemella parahaemolysans / Gemella sanguinis
Gordonia namibiensis group	Gordonia namibiensis / Gordonia paraffinivorans / Gordonia rubripertincta / Gordonia westfalica / Gordonia hankookensis / Gordonia hongkongensis / Gordonia hydrophobica / Gordonia lacunae / Gordonia terrae / Gordonia alkanivorans / Gordonia amicalis / Gordonia insulae / Gordonia neofelifaecis / Gordonia spumicola
Kluyvera ascorbata group	Citrobacter europaeus / Enterobacter soli / Kluyvera ascorbata / Kluyvera cryocrescens / Raoultella terrigena / Klebsiella aerogenes
Kocuria flava group	Kocuria flava / Kocuria turfanensis / Kocuria oceani / Kocuria sediminis
Kocuria arsenatis group	Kocuria arsenatis / Kocuria rhizophila / Kocuria tytonicola / Kocuria atrinae / Kocuria carniphila / Kocuria gwangalliensis / Kocuria salsicia / Kocuria varians
Klenkia terrae group	Klenkia terrae / Klenkia brasiliensis / Klenkia taihuensis / Klenkia marina / Klenkia soli
Kribbella endophytica group	Kribbella endophytica / Kribbella flavida / Kribbella italica / Kribbella amoyensis / Kribbella alba / Kribbella karoonensis / Kribbella pittospori / Kribbella swartbergensis
Kribbella hippodromi group	Kribbella hippodromi / Kribbella jejuensis / Kribbella karoonensis / Kribbella podocarpi / Kribbella shirazensis / Kribbella solani / Kribbella soli / Kribbella swartbergensis / Kribbella aluminosa / Kribbella pittospori / Kribbella sindirgiensis
Lactobacillus acidophilus group	Lactobacillus acidophilus / Lactobacillus crispatus / Lactobacillus gallinarum / Lactobacillus helveticus / Lactobacillus kitasatonis
Lactobacillus gasseri group	Lactobacillus gasseri / Lactobacillus paragasseri / Lactobacillus johnsonii / Lactobacillus taiwanensis
Lactobacillus reuteri group	Lactobacillus reuteri / Lactobacillus vaginalis / Lactobacillus antri / Lactobacillus frumenti / Lactobacillus oris
Lactobacillus rhamnosus group	Lactobacillus rhamnosus / Lactobacillus casei / Lactobacillus paracasei / Lactobacillus zeae
Leifsonia aquatica group	Leifsonia aquatica / Leifsonia naganoensis / Leifsonia shinshuensis / Leifsonia xyli / Leifsonia soli / Leifsonia lichenia
Massilia aurea group	Massilia violaceinigra / Massilia aurea / Massilia brevitalea
Massilia suwonensis group	Massilia suwonensis / Massilia niabensis / Massilia haematophila
Massilia mucilaginosa group	Massilia mucilaginosa / Massilia eurypsychrophila / Massilia frigida / Massilia niabensis
Massilia frigida group	Massilia frigida / Massilia mucilaginosa / Massilia rubra / Massilia violaceinigra / Massilia aquatica

Microbacterium arthrosphaerae group	Microbacterium arthrosphaerae / Microbacterium murale / Microbacterium shaanxiense / Microbacterium invictum / Microbacterium lacus / Microbacterium profundi
Microbacterium marinum group	Microbacterium marinum / Microbacterium maritypicum / Microbacterium liquefaciens / Microbacterium oxydans / Microbacterium foliorum / Microbacterium hydrocarbonoxydans / Microbacterium luteolum / Microbacterium phyllosphaerae / Microbacterium saperdae
Microbacterium lemovicicum group	Microbacterium lemovicicum / Microbacterium binotii / Microbacterium diaminobutyricum / Microbacterium endophyticum / Microbacterium neimengense / Microbacterium sediminicola
Microbacterium hominis group	Microbacterium hominis / Microbacterium laevaniformans / Microbacterium pyrexiae / Microbacterium flavum / Microbacterium paraoxydans / Microbacterium proteolyticum / Microbacterium aerolatum / Microbacterium assamensis / Microbacterium flavescens / Microbacterium foliorum / Microbacterium ginsengiterrae / Microbacterium marinum / Microbacterium oleivorans / Microbacterium radiodurans / Microbacterium resistens / Microbacterium testaceum
Microbacterium sediminis group	Microbacterium sediminis / Microbacterium petrolearium / Microbacterium halimionae / Microbacterium hatanonis / Microbacterium sediminicola / Microbacterium binotii / Microbacterium telephonicum
Microcella putealis group	Microcella putealis / Microcella alkaliphila / Labedella endophytica / Labedella gwakjiensis
Micrococcus antarcticus group	Micrococcus antarcticus / Micrococcus endophyticus / Micrococcus luteus / Micrococcus yunnanensis / Micrococcus aloeverae / Micrococcus cohnii / Micrococcus flavus
Modestobacter marinus group	Modestobacter marinus / Modestobacter muralis / Modestobacter versicolor / Modestobacter caceresii
Mogibacterium diversum group	Mogibacterium diversum / Mogibacterium neglectum / Mogibacterium pumilum / Mogibacterium vescum
Neisseria flava group	Neisseria flava / Neisseria macacae / Neisseria mucosa / Neisseria sicca
Neisseria flavescens group	Neisseria flavescens / Neisseria perflava / Neisseria subflava
Nocardia coeliaca group	Nocardia coeliaca / Rhodococcus degradans / Rhodococcus erythropolis / Rhodococcus qingshengii
Paracoccus aestuarii group	Paracoccus aestuarii / Paracoccus beibuensis / Paracoccus hibisci / Paracoccus marinus / Paracoccus pueri / Paracoccus aeridis / Paracoccus rhizosphaerae / Paracoccus tibetensis / Paracoccus zhejiangensis / Paracoccus alimentarius / Paracoccus isoporae / Paracoccus siganidrum
Paracoccus aestuarii group	Paracoccus aestuarii / Paracoccus beibuensis / Paracoccus hibisci / Paracoccus marinus / Paracoccus pueri / Paracoccus aeridis / Paracoccus rhizosphaerae / Paracoccus tibetensis / Paracoccus zhejiangensis / Paracoccus alimentarius / Paracoccus isoporae / Paracoccus siganidrum
Paracoccus aestuarii group	Paracoccus aestuarii / Paracoccus hibisci / Paracoccus marinus / Paracoccus rhizosphaerae / Paracoccus tibetensis / Paracoccus alimentarius / Paracoccus isoporae / Paracoccus beibuensis / / Paracoccus zhejiangensis / Paracoccus siganidrum

Paracoccus simplex group	Paracoccus simplex / Paracoccus aminovorans / Paracoccus caeni / Paracoccus chinensis / Paracoccus fontiphilus / Paracoccus huijuniae / Paracoccus subflavus / Paracoccus niistensis / Paracoccus aerius / Paracoccus angustae / Paracoccus communis / Paracoccus contaminans / Paracoccus denitrificans / Paracoccus halophilus / Paracoccus sanguinis / Paracoccus speluncae / Paracoccus tibetensis
Paracoccus laeviglucosivorans group	Paracoccus laeviglucosivorans / Paracoccus yeei / Paracoccus carotinifaciens / Paracoccus hibiscisoli / Paracoccus marcusii
Paracoccus sanguinis group	Paracoccus sanguinis / Paracoccus panacisoli / Paracoccus aminovorans / Paracoccus caeni / Paracoccus chinensis / Paracoccus fontiphilus / Paracoccus huijuniae / Paracoccus subflavus / Paracoccus angustae / Paracoccus communis / Paracoccus contaminans / Paracoccus halophilus / Paracoccus simplex / Paracoccus sphaerophysae
Pediococcus stilesii group	Pediococcus acidilactici / Pediococcus stilesii / Pediococcus claussenii
Peptoniphilus gorbachii group	Peptoniphilus gorbachii / Peptoniphilus lacydonensis / Peptoniphilus grossensis / Peptoniphilus harei / Peptoniphilus timonensis / Peptoniphilus phoceensis
Peptoniphilus grossensis group	Peptoniphilus grossensis / Peptoniphilus gorbachii / Peptoniphilus lacydonensis / Peptoniphilus harei / Peptoniphilus timonensis
Phycicoccus bigeumensis group	Phycicoccus bigeumensis / Phycicoccus ginsenosidimutans / Phycicoccus aerophilus / Phycicoccus soli / Phycicoccus dokdonensis
Prevotella histicola group	Prevotella histicola / Prevotella veroralis / Prevotella jejuni
Pseudomonas aylmerense group	Pseudomonas aylmerense / Pseudomonas palleroniana / Pseudomonas tolaasii / Pseudomonas costantinii / Pseudomonas lurida
Pseudomonas aeruginosa group	Pseudomonas aeruginosa / Pseudomonas guezennei / Pseudomonas guangdongensis / Pseudomonas otitidis / Pseudomonas resinovorans / Pseudomonas indica
Pseudomonas antarctica group	Pseudomonas antarctica / Pseudomonas extremorientalis / Pseudomonas fluorescens / Pseudomonas kairouanensis / Pseudomonas kitaguniensis / Pseudomonas meridiana / Pseudomonas poae / Pseudomonas simiae / Pseudomonas trivialis / Pseudomonas extremaustralis / Pseudomonas marginalis / Pseudomonas cerasi / Pseudomonas nabeulensis / Pseudomonas veronii
Pseudomonas argentinensis group	Pseudomonas argentinensis / Pseudomonas cremoricolorata / Pseudomonas fulva / Pseudomonas parafulva / Pseudomonas punonensis / Pseudomonas straminea / Pseudomonas koreensis
Pseudomonas asplenii group	Pseudomonas asplenii / Pseudomonas brassicacearum / Pseudomonas fuscovaginae / Pseudomonas asturiensis / Pseudomonas fluorescens / Pseudomonas synxantha / Pseudomonas versuta / Pseudomonas agarici / Pseudomonas caspiana / Pseudomonas deceptionensis / Pseudomonas fragi / Pseudomonas frederiksbergensis / Pseudomonas putida / Pseudomonas thivervalensis / Pseudomonas vranovensis
Pseudomonas azotoformans group	Pseudomonas azotoformans / Pseudomonas lactis / Pseudomonas paralactis / Pseudomonas synxantha / Pseudomonas libanensis / Pseudomonas mucidolens / Pseudomonas fluorescens / Pseudomonas gessardii
Pseudomonas brenneri group	Pseudomonas brenneri / Pseudomonas fluorescens / Pseudomonas proteolytica

Pseudomonas canadensis group	Pseudomonas canadensis / Pseudomonas fluorescens / Pseudomonas salomonii / Pseudomonas corrugata
Pseudomonas chlororaphis group	Pseudomonas chlororaphis / Pseudomonas fluorescens / Pseudomonas glycinae / Pseudomonas kribbensis / Pseudomonas entomophila / Pseudomonas guariconensis / Pseudomonas mosselii / Pseudomonas sichuanensis / Pseudomonas soli
Pseudomonas corrugata group	Pseudomonas corrugata / Pseudomonas canadensis / Pseudomonas fluorescens / Pseudomonas salomonii
Pseudomonas deceptionensis group	Pseudomonas deceptionensis / Pseudomonas fragi / Pseudomonas lundensis / Pseudomonas psychrophila / Pseudomonas weihenstephanensis
Pseudomonas chloritidismutans group	Pseudomonas chloritidismutans / Pseudomonas knackmussii / Pseudomonas stutzeri / Pseudomonas zhaodongensis / Pseudomonas kunmingensis
Pseudomonas fluorescens group	Pseudomonas fluorescens / Pseudomonas glycinae / Pseudomonas kribbensis / Pseudomonas granadensis / Pseudomonas koreensis / Pseudomonas turukhanskensis
Pseudomonas grimontii group	Pseudomonas grimontii / Pseudomonas marginalis / Pseudomonas rhodesiae
Pseudomonas indoloxydans group	Pseudomonas indoloxydans / Pseudomonas oleovorans / Serratia plymuthica / Pseudomonas sediminis
Pseudomonas cannabina group	Pseudomonas cannabina / Pseudomonas syringae / Pseudomonas cerasi / Pseudomonas congelans / Pseudomonas ficuserectae
Pseudomonas meliae group	Pseudomonas meliae / Pseudomonas savastanoi / Pseudomonas tremae / Pseudomonas cerasi / Pseudomonas chlororaphis / Pseudomonas congelans / Pseudomonas ficuserectae / Pseudomonas protegens / Pseudomonas syringae / Pseudomonas lini / Pseudomonas caricapapayae
Pseudomonas veronii group	Pseudomonas veronii / Pseudomonas extremaustralis / Pseudomonas fluorescens / Pseudomonas marginalis / Pseudomonas antarctica / Pseudomonas extremorientalis / Pseudomonas kairouanensis / Pseudomonas kitaguniensis / Pseudomonas meridiana / Pseudomonas nabeulensis / Pseudomonas poae / Pseudomonas simiae / Pseudomonas trivialis
Pseudomonas vranovensis group	Pseudomonas vranovensis / Pseudomonas alkylphenolica / Pseudomonas asplenii / Pseudomonas fuscovaginae / Pseudomonas hutmensis
Rhodopseudomonas pentothenatexigens group	Rhodopseudomonas pentothenatexigens / Rhodopseudomonas thermotolerans / Rhodopseudomonas faecalis / Rhodopseudomonas palustris
Sphingomonas alpina group	Sphingomonas alpina / Sphingomonas echinoides / Sphingomonas oligophenolica / Sphingomonas asaccharolytica / Sphingomonas insulae / Sphingomonas kyungheensis / Sphingomonas mali / Sphingomonas mucosissima / Sphingomonas panacis / Sphingomonas populi / Sphingomonas pruni / Sphingomonas aquatilis / Sphingomonas dokdonensis / Sphingomonas jeddahensis / Sphingomonas melonis
Staphylococcus aureus group	Staphylococcus argenteus / Staphylococcus aureus / Staphylococcus schweitzeri / Staphylococcus simiae / Staphylococcus haemolyticus / Staphylococcus petrasii

Staphylococcus caeli group	Staphylococcus caeli / Staphylococcus pseudoxylosus / Staphylococcus saprophyticus / Staphylococcus edaphicus / Staphylococcus xylosus / Staphylococcus arlettae / Staphylococcus gallinarum
Staphylococcus capitis group	Staphylococcus capitis / Staphylococcus caprae / Staphylococcus epidermidis / Staphylococcus saccharolyticus / Staphylococcus cohnii / Staphylococcus haemolyticus / Staphylococcus hominis
Staphylococcus epidermidis group	Staphylococcus capitis / Staphylococcus caprae / Staphylococcus epidermidis / Staphylococcus saccharolyticus / Staphylococcus cohnii / Staphylococcus haemolyticus / Staphylococcus hominis
Staphylococcus haemolyticus group	Staphylococcus haemolyticus group / Staphylococcus petrasii / Staphylococcus hominis / Staphylococcus argenteus / Staphylococcus aureus / Staphylococcus devriesei / Staphylococcus lugdunensis / Staphylococcus schweitzeri / Staphylococcus simiae
Staphylococcus hominis group	Staphylococcus hominis / Staphylococcus haemolyticus / Staphylococcus lugdunensis / Staphylococcus petrasii / Staphylococcus capitis / Staphylococcus caprae / Staphylococcus epidermidis / Staphylococcus pasteuri
Staphylococcus hominis group	Staphylococcus hominis / Staphylococcus haemolyticus / Staphylococcus lugdunensis / Staphylococcus petrasii / Staphylococcus capitis / Staphylococcus caprae / Staphylococcus epidermidis / Staphylococcus pasteurii
Streptococcus mitis group	Streptococcus cristatus / Streptococcus gordonii / Streptococcus gwangjuense / Streptococcus infantis / Streptococcus mitis / Streptococcus oralis / Streptococcus periodonticum / Streptococcus pneumoniae / Streptococcus timonensis / Streptococcus pseudopneumoniae / Streptococcus sanguinis / Streptococcus chosunense
Streptococcus lactarius group	Streptococcus lactarius / peroris / parasanguinis
Streptococcus parasanguinis group	Streptococcus parasanguinis / Streptococcus australis / Streptococcus cristatus / Streptococcus rubneri
Streptococcus salivarius group	Streptococcus salivarius / Streptococcus vestibularis / Streptococcus thermophilus
Variovorax gossypii group	Variovorax gossypii / Variovorax guangxiensis / Variovorax paradoxus
Veillonella parvula group	Veillonella parvula / Veillonela dentocariosa / Veillonella tobetsuensis / Veillonella rodentium / Veillonella rogosae

B)

Unknown bacteria #	**Best match in Genbank BLAST search**
Unknown bacteria 1	100% match with uncultured *Vampirovibrio* sp. clone Z3AcetBAC91 16S ribosomal RNA gene, accession number KX350762.1
Unknown bacteria 2	100% match with uncultured Syntrophobacteraceae bacterium clone 327 16S ribosomal RNA gene, accession number KX366231.1
Unknown bacteria 3	99.8% match with Tuber borchii symbiont b-17BO 16S ribosomal RNA gene, accession number AF070444.1

Unknown bacteria 4	100% match with uncultured bacterium clone 3500 16S ribosomal RNA gene, accession number MF082879.1
Unknown bacteria 5	100% match with uncultured bacterium gene for 16S rRNA, clone: 11Aug11-129, accession number LC336118.1
Unknown bacteria 6	99,8% match with uncultured delta proteobacterium clone 4M1_F12 16S ribosomal RNA gene, accession number EU052019.1
Unknown bacteria 7	99,3% match with uncultured bacterium partial 16S rRNA gene, isolate BACT_OTU_235, accession number LT842698.1
Unknown bacteria 8	100% match with uncultured bacterium clone KTB502 16S ribosomal RNA gene, accession number MG388843.1
Unknown bacteria 9	98,6% match with uncultured prokaryote gene for 16S ribosomal RNA, OTU:NOR1708, accession number LC248638.1
Unknown bacteria 10	100% match with Neisseriaceae [G-1] [G-1] bacterium HMT 174, accession number FM873692.1
Unknown bacteria 11	100% match with uncultured bacterium RNA for 16S rRNA, partial sequence, clone: B0423R003_K06
Unknown bacteria 12	100% match with uncultured bacterium partial 16S rRNA gene, OTU05254, accession number LT009308.1
Unknown bacteria 13	97.8% match with uncultured bacterium clone 12-2B-102 16S ribosomal RNA gene, accession number KM221296.1
Unknown bacteria 14	98.8 % match with uncultured bacterium clone lp233 16S ribosomal RNA gene, accession number KC331451.1
Unknown bacteria 15	99.1% match with uncultured bacterium clone 2290 16S ribosomal RNA gene, accession number MF081669.1
Unknown bacteria 16	99.3% match with uncultured bacterium clone OTU424_L_1_A_2109484 16S ribosomal RNA gene, accession number MG858260.1
Unknown bacteria 17	100% match with Desulfatiglans anilini strain WB91 16S ribosomal RNA gene, accession number MH196469.1
Unknown bacteria 18	99.5% match with uncultured bacterium clone S5 16S ribosomal RNA gene, accession number JX133359.1
Unknown bacteria 19	100% match with uncultured bacterium clone OTU1014 16S ribosomal RNA gene, accession number MF689080.1
Unknown bacteria 20	100% match with Myxobacterium AT3-01 gene for 16S rRNA, accession number AB246772.1
Unknown bacteria 21	100% match with uncultured bacterium clone ncd1960f07c1 16S rRNA, accession number JF171142.1
Unknown bacteria 22	98.8% match with unidentified bacterium clone M2_Bulk_T7s_9 16S rRNA, accession number EF605681.1
Unknown bacteria 23	98.5% match with Uncultured bacterium clone Otu2063, accession number MW084188.1
Unknown bacteria 24	99,0% match with Bacterium Kaz2, accession number AB491166.1
Unknown bacteria 25	99.8% match with uncultured bacterium clone KD_68, accession number HQ911196.1
Unknown bacteria 26	100% match with uncultured bacterium partial 16S rRNA gene, clone FG34B-43, accession number FR846901.1
Unknown bacteria 27	98.8 % match with uncultured bacterium clone OTU1241_Y_10_A_1708905, accession number MG859032.1
Unknown bacteria 28	100% match with uncultured bacterium clone SupSIB047, accession number MW128096.1
Unknown bacteria 30	92,6% match with uncultured organism clone KBTEX_233 genomic sequence, accession number MN079308.1
Unknown bacteria 31	100% match with uncultured alpha proteobacterium partial 16S rRNA gene, isolate BACT_OTU_1497, accession number LT842726.1

Unknown bacteria 32	100% match with uncultured bacterium partial 16S rRNA gene, isolate BACT_OTU_2798, accession number LT844046.1
Unknown bacteria 33	99,8% match with uncultured bacterium clone OTU276, accession number MG928809.1
Unknown bacteria 34	94,6% match with uncultured Spirosoma sp. clone OTU627 16S ribosomal RNA gene, accession number MW144078.1
Unknown bacteria 35	100% match with uncultured Planctomyces sp. clone 507, accession number MF042869.1
Unknown bacteria 36	99,5% match with uncultured bacterium isolate DGGE gel band 03_M3 clone 05, accession number JX986140.1
Unknown bacteria 37	100% match with uncultured bacterium clone ncd2030b12c1, accession number JF175604.1
Unknown bacteria 38	99,3% match with uncultured Chitinophagaceae bacterium clone CNY_00868, accession number JQ400929.1
Unknown bacteria 39	99,8% match with uncultured bacterium clone KTTB22, accession number MG392744.1
Unknown bacteria 40	98,0% match with uncultured bacterium clone OTU7082, accession number KT790215.1
Unknown bacteria 41	93,9% match with uncultured bacterium clone 16S(V3+V4)-356 16S ribosomal RNA gene, accession number MH096412.1
Unknown bacteria 42	100% match with uncultured bacterium clone 7A_10-029, accession number KY190683.1
Unknown bacteria 43	100% match with Uncultured Gemmatimonas sp. clone CNY_00734, accession number JQ400828.1
Unknown bacteria 44	99,3% match with uncultured Kofleriaceae bacterium clone 414, accession number KX366318.1
Unknown bacteria 45	99,8% match with uncultured bacterium clone OTU8133, accession number KT791146.1
Unknown bacteria 46	98,8% match with uncultured bacterium clone OTU_6667, accession number MH530315.1
Unknown bacteria 47	100% match with uncultured bacterium clone OTU1315_Control_T1.3100, accession number MF950297.1
Unknown bacteria 48	100% match with uncultured bacterium clone KTTB699, accession number MG393421.1
Unknown bacteria 49	100% match with uncultured actinobacterium partial 16S rRNA gene, clone UMAB-cl-58, accession number FN811242.1
Unknown bacteria 50	100% match with uncultured bacterium clone 3-200, accession number KC554166.1
Unknown bacteria 51	99,8% match with uncultured bacterium clone KTB2725, accession number MG391066.1
Unknown bacteria 52	99,3% match with uncultured bacterium partial 16S rRNA gene, isolate BI-2-66, accession number HG970682.1
Unknown bacteria 53	98,6% match with uncultured Bacteroidetes bacterium partial 16S rRNA gene, clone 169-1, accession number AJ871244.1
Unknown bacteria 54	100% match with uncultured bacterium clone MS-9-5 16S ribosomal RNA gene, accession number KX284677.1
Unknown bacteria 55	100% match with uncultured bacterium clone OTU39 16S ribosomal RNA gene, accession number MW143612.1
Unknown bacteria 56	99,3% match with uncultured bacterium clone 1016 16S ribosomal RNA gene, accession number MG716571.1
Unknown bacteria 57	99,5% match with Candidatus Saccharibacteria bacterium isolate NC_groundwater_1927_Pr3_S-0.2um_48_6 chromosome, accession number CP066696.1
Unknown bacteria 58	99,3% match with uncultured bacterium clone ELA_111314_OTU_7272 16S ribosomal RNA gene, accession number KY522290.1

Unknown bacteria 59	100% match with uncultured bacterium clone OTU333_L_2_A_979036 16S ribosomal RNA gene, accession number MG858172.1
Unknown bacteria 60	100% match with uncultured bacterium clone denovo16199 16S ribosomal RNA gene, accession number MG910600.1
Unknown bacteria 61	98,8% match with uncultured bacterium partial 16S rRNA gene, isolate M1-261, accession number HE653888.1
Unknown bacteria 62	100% match with uncultured bacterium clone OTU7317 16S ribosomal RNA gene, accession number KT790439.1
Unknown bacteria 63	97,6 % match with uncultured bacterium clone ELA_111314_OTU_6424 16S ribosomal RNA gene, accession number KY521460.1

Graphic design: Communication Division, UiB / Print: Skipnes Kommunikasjon AS

uib.no

ISBN: 9788230851180 (print)
9788230841365 (PDF)