Benchmarking laboratory processes to characterise low-biomass respiratory microbiota

Raiza Hasrat
Wilhelmina Children's Hospital/University Medical Centre Utrecht
Jolanda Kool
National Institute for Public Health and the Environment
Wouter A.A. de Steenhuijsen Piters
Wilhelmina Children's Hospital/University Medical Centre Utrecht
Mei Ling J.N. Chu
Wilhelmina Children's Hospital/University Medical Centre Utrecht
Sjoerd Kuiling
National Institute for Public Health and the Environment
James Groot
National Institute for Public Health and the Environment
Elske van Logchem
National Institute for Public Health and the Environment
Susana Fuentes
National Institute for Public Health and the Environment
Eelco Franz
National Institute for Public Health and the Environment
Debby Bogaert
University of Edinburgh Centre for Inflammation Research, University of Edinburgh
Thijs Bosch (thijs.bosch@rivm.nl)
National Institute for Public Health and the Environment

Research Article

Keywords: low biomass, microbiota, MiSeq, 16S rRNA gene amplicon sequencing, DNA extraction, PCR bias, respiratory, negative controls, ZymoBIOMICS microbial community standard, AMPure XP

Posted Date: April 26th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-454841/v1

Abstract

The low biomass of respiratory samples makes it difficult to accurately characterise the microbial community composition. PCR conditions and contaminating microbial DNA can alter the biological profile. The objective of this study was to benchmark the currently available protocols to accurately analyse the microbial community of low biomass samples.

To study the effect of PCR conditions on the microbial community composition, we amplified the 16S rRNA gene of respiratory samples using various DNA input and different number of PCR cycles. Libraries were purified by gel electrophoresis or AMPure XP and sequenced by V2 and V3 MiSeq reagent kits by Illumina sequencing. The positive control was diluted in different solvents.

PCR conditions had no significant influence on the microbial community composition of low biomass samples. Purification methods and MiSeq reagent kits had only a modest impact on microbiota profiles, while profiles of positive controls were significantly influenced by type of dilution solvent. Microbiota profiles of low biomass samples can be accurately distinguished from DNA blanks.

Microbiota profiling of low biomass samples is stable under several PCR conditions, purification methods and MiSeq reagent kits. We recommend to use amplification with 30 PCR cycles. The amplicon pools can best be purified by two consecutive AMPure XP steps and sequenced by V3 MiSeq reagent kit. The benchmarked standardized workflow presented here ensures comparability of results within and between low biomass microbiome studies.

Introduction

The human microbiome consists of interacting networks of microorganisms, such as bacteria, archaea and fungi. The microbial community composition varies between individuals and body sites [1–3]. To date, the gut microbiota is the most well-studied niche, and has been shown to play a vital role in human health [4–8]. Evidence is accumulating that the microbiota in other niches, for example in the respiratory tract, similarly impacts human health [1,5,9–11]. The respiratory bacterial community is suggested to play an important role in protecting against acquisition and overgrowth of new pathogens, maturation and modulation of the immune system and increasing epithelial integrity thereby inhibiting bacterial translocation [5,12].

Complex microbial communities are more accurately characterised by culture-independent techniques. Especially next-generation sequencing techniques are commonly used for analysis of gut microbiota [1–3, 5,9–14]. In contrast to the gut microbiota, the respiratory tract is less densely colonized [15,16], which makes it more difficult to reliably characterise them. Particularly contaminating microbial DNA from the environment and from laboratory reagents can strongly skew bacterial profiles in low biomass materials [17–20]. Furthermore, differences in standard operating procedures including template concentration and

the number of PCR amplification cycles have shown to affect results significantly, making comparisons between different studies difficult [21–25].

Therefore, a consistent workflow including suitable controls should be applied to ensure reliable microbiota analyses of low biomass materials. Here we describe the optimization process for 16S rRNA gene MiSeq library preparation protocols [2,26], including testing the effects of template concentration, number of PCR cycles, library clean-up methods and MiSeq reagent kit chemistry on low biomass microbiota characterisation.

Methods

Study population/data collection

For optimization experiments, we used 218 random samples collected from the nasopharynx (n = 214), oropharynx (n = 2) and saliva (n = 2) from healthy individuals (Table 1) included in a Dutch cross-sectional population-wide study, named Pienter-3 [27]. All procedures performed were in accordance with the ethical standards of the institutional and/or national research committee. Ethical approval was granted by the national ethics committee in the Netherlands, METC Noord-Holland (METC number: M015–022). Written informed consent was obtained from all adult participants, and parents or legal guardians of minors included in the study [27]. Following collection, saliva samples were stored in a tube containing 50% glycerol, and nasopharyngeal (NP) and oropharyngeal (OP) swabs were stored in 1 ml of liquid Amies medium. Samples were directly frozen on dry-ice and stored at − 80°C until further processing [27]. We used the ZymoBIOMICS microbial community standard (Zymo mock; Zymo Research, Irvine, CA, USA) and the ZymoBIOMICS microbial community DNA standard (DNA mock; Zymo Research) as positive controls.

Table 1
Samples and statistical method per experiment. NP = Nasopharynx, OP = Oropharynx

Experiments	NP	OP	Saliva	Zymo mock	DNA mock	DNA blanks	NTC	Statistical method
Optimizing Zymo mock community profile				16				ANOVA-test, for global differences between groups, and Tukey's *post-hoc* test. Reference group: dilution in elution buffer. Lollipop plot
Effect of number of PCR cycles on microbial community composition		2	2					Relative abundance difference between 25, 30 and/or 35 cycles
Effect of DNA template concentration on microbial community composition		2	2					Relative abundance difference between 16, 125 and/or 1000 pg DNA input
Comparing DNA concentration of NP samples with DNA blanks	214					3		Difference of DNA concentration between NP amplified by 30 PCR cycles and DNA blanks amplified by 25, 30 and 35 cycles
Concordance gel-based and AMPure XP purification	214							Bray-Curtis dissimilarity, Pearson correlation coefficient and β-coefficient

Experiments	NP	OP	Saliva	Zymo mock	DNA mock	DNA blanks	NTC	Statistical method
Concordance V2 and V3 MiSeq reagent kits	214							Linear model, Bray-Curtis dissimilarity, Pearson correlation coefficient and β-coefficient
Comparing low-biomass samples with DNA blanks	140					8		Unsupervised hierarchical clustering based on Bray-Curtis dissimilarity

DNA extraction

DNA was extracted from NP swabs, OP swabs and saliva using an Agowa Mag DNA extraction kit (LGC genomics, Berlin, Germany) as previously described [26,28], with slight modifications shown to be more robust for low biomass DNA extractions [26]. In each isolation run, one 200 µl aliquot of 10^3 diluted Zymo mock was included as positive control, plus two negative controls containing lysis buffer only (referred to as DNA blanks). Samples were thawed on ice and robustly vortexed for 10 seconds. Per sample, 600 µl of lysis buffer with zirconium beads (diameter 0.1 mm, Biospec Products, Bartlesville, OK, USA) and 550 µl phenol (VWR International, Amsterdam, the Netherlands) was added in a conical 1.5 ml screw-cap Eppendorf tube. Samples were mechanically disrupted twice for 2 minutes at 3,500 oscillations/minute by bead beating (Mini-Beadbeater-24, Biospec Products) and transferred on ice for 2 minutes after each bead beating step. The tubes were centrifuged for 10 minutes at 4,500 x g. The clear aqueous phase was added to a 2 ml Eppendorf tube containing 1.3 ml binding buffer and 10 µl magnetic beads. After shaking for 30 minutes, the tubes were put in a magnetic separation rack. The supernatant was discarded, the magnetic beads were washed with wash buffer 1 and 2 and air-dried for 15 minutes at 55°C. DNA was eluted in either 35 µl or 50 µl elution buffer, depending on the starting material, by shaking for 15 minutes at 55°C. Supernatant was transferred to a 1.5 ml Eppendorf LoBind tube and stored at -20°C.

ZymoBIOMICS microbial community standard

Zymo mock was stored by the manufacturer in DNA/RNA shield. To test the effect of dilution solvent on the generated Zymo mock profile, we prepared dilutions (10^1-10^3) in DNA/RNA shield, elution buffer (Qiagen, Hilden, Germany) and Milli-Q water, mimicking the DNA concentration of low biomass samples.

Bacterial DNA quantification

The bacterial load was quantified by quantitative PCR (StepOnePlus Real-Time PCR System, Thermo Fisher Scientific, the Netherlands) with universal primers and probe targeting the 16S rRNA gene,

containing forward primer 16S-F1 (5'-CGA AAG CGT GGG GAG CAA A -3'), reverse primer 16S-R1 (5'-GTT CGT ACT CCC CAG GCG G-3') and probe 16S-P1 (FAM- ATT AGA TAC CCT GGT AGT CCA -ZEN) (IDT, Leuven, Belgium) [15,26]. To optimize qPCR reproducibility and to allow comparisons of DNA concentrations reliably, we developed a standard curve by using a synthesized fragment of the 16S rRNA gene (gBlocks Gene Fragment, IDT, 5'-CGG TGC GAA AGC GTG GGG AGC AAA CAG GAT TAG ATA CCC TGG TAG TCC ACG CCG TAA ACG ATG TCT ACT AGC TGT TCG TGG TCT TGT ACT GTG AGT AGC GCA GCT AAC GCA CTA AGT AGA CCG CCT GGG GAG TAC GAA CGC AAG-3').

MiSeq library preparation and sequencing

The V4 region of the 16S rRNA gene was amplified by PCR using the 515F (5'- GTG CCA GCM GCC GCG GTA A-3') and 806R (5'-GGA CTA CHV GGG TWT CTA AT-3') primers including the Illumina adapters and sample specific barcodes [2,29,30]. Each 25 µl PCR reaction consisted of 0.5 µl Phusion Hot Start II High-Fidelity DNA Polymerase, 5 µl 5x Phusion HF Buffer (Thermo Fisher Scientific), 7 µl HPLC grade water (Instruchemie, Delfzijl, the Netherlands), 2.5 µl of 2mM dNTP mix (Roche, Mannheim, Germany), 5 µl of 5 µM barcoded primer 515F, 5 µl of 5 µM barcoded primer 806R and 5 µl template DNA. PCR reactions were executed by following the successive steps; 98°C for 30 s; 30 cycles at 98°C for 10 s, 55°C for 30 s and 72°C for 30 s and a final hold of 5 minutes at 72°C. Samples with a 16S rRNA gene DNA concentration of < 20 pg/µl (< 100 pg input DNA) were used undiluted, samples with a higher concentration were diluted in HPLC grade water accordingly. To study the effect of PCR conditions on the microbiota profile, 16, 125 and 1000 pg of template DNA from two OP and two saliva samples were amplified by 30 cycles. The input DNA of 125 pg was additionally, separately, amplified by 25 and 35 PCR cycles. DNA blanks, no template controls (NTC), Zymo mocks and DNA mocks were included in each PCR plate and sequenced alongside the samples. The amplicon fragment size was assessed using agarose gel electrophoresis and quantified by Quant-iT PicoGreen dsDNA Assay Kit (Thermo Fisher Scientific). Barcoded amplicons were pooled in equimolar ratios. To study the optimal purification method, we purified the pool with two different cleaning methods; 1. by running it on an agarose gel, extracting the DNA using GeneJET Gel Extraction and DNA Cleanup Micro Kit (Thermo Fisher Scientific) and subsequently purified by 0.9x AMPure XP magnetic beads (Beckman Coulter, the Netherlands), or 2. by two consecutive times purification using 0.9x AMPure XP. The library was prepared as recommended by Illumina and sequenced using the MiSeq reagent kit V2 or V3 (paired end, 500 bp) on an Illumina MiSeq instrument (Illumina Inc., San Diego, CA, US).

Data analysis

All sample libraries were simultaneously processed using an in-house bioinformatics pipeline [1,3,31]. First, we performed adaptive window trimming with a quality threshold of Q30, retaining those reads with a minimum length of 150 nucleotides (Sickle, version 1.33) [32]. Sequencing errors were reduced by an error correction algorithm (BayesHammer, SPAdes genome assembler toolkit, version 3.12.0) [33]. Paired-end sequenced reads were assembled into contigs using PANDAseq (version 2.10) and demultiplexed using

QIIME (version 1.9.1) [34,35]. Singleton sequences and chimeras were removed (UCHIME; implemented in the VSEARCH toolkit v2.0.3). VSEARCH abundance-based greedy clustering was performed to pick OTUs (operational taxonomic unit) with a 97% identity threshold [36]. OTUs were taxonomically annotated by the Naïve Bayesian RDP classifier using the SILVA 119 release reference database [37,38]. OTUs were assigned a rank number based on their abundance across the total dataset.

Analyses were performed in R version 4.0.2 within R studio version 1.4.623. OTU read counts were normalised using total sum scale resulting in relative abundances of OTUs. Microbiota profiles were visualized using stacked bar charts/boxplots. Lollipop plots were used to visualize the differences in relative abundance of each OTU between sequenced diluted Zymo mocks and the theoretical Zymo mock. To assess overall differences in microbial community composition between (pairs of) samples we used Bray-Curtis dissimilarity matrix, where zero indicates an identical composition between pairs. Non-metric multidimensional scaling (NMDS) based on the Bray-Curtis dissimilarity matrix was used to visualize differences in microbial profiles between low-biomass samples and DNA blanks [1]. We investigated the minimal DNA concentration for reliant microbiome analyses by comparing the microbial profiles of DNA blanks and low-density samples using an unsupervised hierarchical clustering approach based on the Bray-Curtis dissimilarity matrix, which was illustrated in a dendrogram. Silhouette and Calinski-Harabasz indices were used to determine the number of clusters [1]. To assess the impact of MiSeq reagent kits/purification methods, we determined the Pearson correlation of $\log_{10}+1$-transformed relative abundances of OTUs with > 0.1% abundance in at least 20 samples. To test for significant differences in Zymo mock composition with different dilution solvents we used an ANOVA-test with Tukey's *post hoc* test to determine statistical significance between specific groups. Linear models were used to calculate the statistical significance between the number of reads per sample sequenced by V2 and V3 reagents kit. A p-value < 0.05 was considered significant.

Results

DNA extraction

Zymo mock dilution optimization

To mimic the concentration of low-biomass samples, a dilution series (10^1-10^3x) was prepared of Zymo mock. Zymo mocks were diluted in DNA/RNA shield (n=6), elution buffer (n=5) and Milli-Q (n=5). Dilution in DNA/RNA shield resulted in a significantly different microbiota profile in comparison to elution buffer and Milli-Q across dilutions (Fig. 1a and b). We observed an overrepresentation of *Bacillus subtilis(11), Enterobacter (8), Escherichia coli (10)* and *Pseudomonas aeruginosa(15)* and underrepresentation of *Staphylococcus epidermidis (2), Lactobacillus fermentum (22)* and *Enterococcus faecium (29)* in Zymo mocks diluted in DNA/RNA shield compared to elution buffer (Fig. 1b). In contrast, when comparing milli-Q to elution buffer, only the *Lactobacillusfermentum(22)* relative abundance differed significantly (median 7.9% vs 10.0%, respectively, p-value < 0.001). The Zymo mock diluted in elution buffer most

closely resembled the theoretical Zymo mock bacterial composition (Fig. 1c). Therefore, for further experiments, elution buffer was chosen as preferred dilution solvent for the Zymo mocks.

Library preparation

Influence of PCR amplification cycles and bacterial density on the microbiota profile

We tested the effect of the number of PCR amplification cycles on the microbial community composition. To this end, 2 OP and 2 saliva samples with a DNA input of 125 pg were amplified using 25, 30 and 35 PCR cycles. We observed that a higher number of PCR cycles lead to minor increases in the relative abundance of high abundant OTUs. Both saliva samples had a higher relative abundance of *Neisseria (21)* (8.6/13.9%, 10.0/16.3% and 10.9/19.2% for 25, 30 and 35 cycles, respectively) with a higher number of PCR cycles (Fig. 2a). One OP sample had a higher relative abundance of *Prevotella melaninogenica (37)* (17.0%, 18.4% and 22.9% for 25, 30 and 35 cycles, respectively) and *Leptotrichia (74)* (16.8%, 17.3% and 22.6% for 25, 30 and 35 cycles, respectively) with more PCR cycles. A higher number of PCR cycles also resulted in an increased amplicon concentration in DNA blanks indicating increased amplification of contaminating DNA (Fig. 3). Given the increased risk of contamination bias when using 35 PCR cycles and higher low biomass sample dropout when using 25 PCR cycles (data not shown), we therefore recommend that the optimal number of PCR amplification cycles is 30.

To assess effects of DNA template input on microbial community profiles, we tested three different input DNA concentrations (16, 125 and 1000 pg) of 2 saliva and 2 OP samples. We noticed that increasing DNA concentrations modestly affect the relative abundance of high abundant OTUs (Fig. 2b). In 3 of the 4 samples, we observed a modest increase in the relative abundance of *Neisseria (21)* (9.4/8.7/14.5%, 10.8/10.0/16.3% and 11.1/11.0/17.4% for 16, 125 and 1000 pg, respectively). Another OP sample showed modest increased relative abundance of *Leptotrichia (74)* with increasing template input (16.8%, 17.3% and 17.9% for 16, 125 and 1000 pg, respectively). Despite minor differences, we propose to standardise to a 125 pg template input for MiSeq PCR in case of low-biomass samples, given that a 1000 pg yield is not always feasible.

Concordance between library clean-up methods

To further optimize our workflow, we studied the influence of the gel-based purification and the AMPure XP clean-up on the microbiota composition by purifying one amplicon pool containing 214 samples using both procedures (Table 1). Microbiota profiles of the same samples purified using different protocols were highly similar (paired Bray-Curtis dissimilarity median: 0.03; range: 0.0-0.06), indicating a high concordance between both clean-up methods (Fig. 4a). Furthermore, we compared the relative abundances of the top 8 OTUs per sample and observed a correlation and regression coefficient of ~1.0 for all OTU abundances observed by both methods (Fig. 4b), indicating a near perfect concordance for the most common OTUs, and thus negligible differences between the tested library clean-up methods. For the purpose of our study we chose to continue with the AMPure XP purification method as it is faster compared to gel-based purification.

MiSeq sequencing

Concordance between the V2 and V3 MiSeq reagent kits

To study the concordance between the V2 and V3 MiSeq reagent kits, we used the same set of samples as described when validating the library clean-up methods (Table 1). The mean number of reads per sample purified by AMPure XP was significantly different (p-value < 0.001), with 20,060 (range: 2,123-39,486) versus 36,981 reads (range: 3,781-72,469 reads) for the V2 and V3 kit, respectively (Fig. 5a). The overall microbial community composition only marginally differed between both sequencing methods, as indicated by a very high similarity between pairs (Bray-Curtis dissimilarity median: 0.05; range: 0.0-0.1) compared to unpaired samples (Bray-Curtis dissimilarity median: 0.8; range: 0.03-1.0) (Fig. 4a). Additionally, we compared the relative abundances of the top 8 OTUs and observed a correlation coefficient of ~1.0 for all those OTUs and a regression coefficient of ~1.0 for 7 of those OTUs (Fig. 5b), with *Streptococcus (7)* slightly underrepresented in the V2 kit (regression coefficient: 0.9). We conclude that given the high concordance between MiSeq reagent kits, we prefer to use the more recent V3 MiSeq kit, as it yields a higher number of reads per sample.

Microbiota profiles of low biomass samples compared to DNA isolation blanks

We tested to what extent the microbial community of samples with a range of low densities resembles that of DNA blanks. When comparing the microbiota profiles of 140 NP samples (range: 0.06-1.00 pg/µl) and 8 DNA blanks (0.02-0.07 pg/µl) (Table 1), we found that the blanks clustered separately from the NP samples (Fig. 6a). Using an unsupervised hierarchical clustering of both samples and blanks, we identified 8 different clusters, 7 clusters containing exclusively NP samples and one cluster containing DNA blanks and 2 NP samples (Fig. 6b). These 2 NP samples have a concentration lower than 0.10 pg/µl, while the other 2 NP samples with <0.10 pg/µl clustered with NP samples containing >0.10 pg/µl. Therefore, we advise to use a minimum concentration of 0.10 pg/µl to ensure a reliable microbiota analysis. Although, low biomass samples may still contain contaminating DNA, these samples are clearly distinguishable from DNA blanks and are more likely to still elicit sufficient reads after consecutive bioinformatic clean-up.

Discussion

To study high biomass fecal microbiota, Costea et al. recommended the use of a standardized protocol to ensure reproducility and comparibility among studies [39]. Here, we show the importance of a standardized DNA extraction and sequencing protocol for low biomass samples like respiratory materials as well. Noteworthy, the library clean-up methods (gel-based purification or AMPure XP), and the MiSeq reagent kits (V2 or V3 chemistry), resulted in modest to no effects on overall microbial community profiles.

We compared the labour-intensive gel-based size selection and a column-based clean-up method (AMPure XP), which can select for DNA size in a fast and effective manner [40–43]. A specific ratio of 0.9x AMPure XP lead to minimal loss of library DNA concentration and complete removal of primer dimers [44].

Comparison of microbial community compositions of samples processed using each of these two methods showed high similarity. Furthermore, the 8 most abundant OTUs of these samples showed a perfect concordance. Since the different cleaning procedures gave similar microbial composition, we propose to use AMPure XP for fast library clean-up.

The microbiota data obtained by sequencing using the V2 (2 x 150 bp) and V3 (2 x 300 bp) MiSeq reagent kits were highly similar [45]. We also compared the microbiota data of a 16S rRNA gene pool sequenced using both reagent kits (2 x 250 bp). We observed a very high concordance between the V2 and V3 kit; the observed modest underrepresentation of *Streptococcus* is likely a result of differences in number of freeze-thaw cycles of the library, rather than differences in kits used [46]. To understand the ecology of the respiratory microbiome, it is critical to study the whole microbiome including the low abundant bacteria [45], which underlines the importance of sufficient sequencing depth. Here, we noticed that the sequencing depth per sample almost doubled when sequenced using the V3 kit, thus being preferable above the V2 kit.

The inclusion of negative controls is important to accurately study the microbiota [17–20]. Contaminants can have a significant impact on the microbial data of low biomass samples [18]. Though not a primary research question in our study, we confirmed that samples with a concentration as low as 0.1 pg/µl can be consistenly amplified and show a microbiota composition that is distinguishable from the DNA extraction blanks. Discrimination between samples and blanks should be further improved by using bioinformatic tools [47,48] such as the *decontam* R-package, which allows for the identification and removal of contaminating OTUs, ideally based on a large number negative controls [47]. DNA extraction and no template controls will therefore not only help to identify limits within laboratory protocols, but also help to control for contaminating DNA in downstream analyses.

We demonstrated that the bacterial profile of Zymo mock, when diluted, can be influenced by the solvent used (DNA/RNA shield, MilliQ and elution buffer). Sample storage should therefore also be optimised for the positive controls. Dilution of Zymo mock in elution buffer closely resembled the bacterial profile of the theoretical mock.

Several studies described the effect of PCR conditions on the microbial composition. A higher number of PCR cycles can lead to an increased concentration of contaminating DNA, point mutation artifacts and chimera formation [18,21,22,24,25]. An input of 16 pg is feasible for most of the NP samples used in this study, though more samples would have to be diluted, resulting in a higher contamination risk [18]. We demonstrated that template DNA concentration and PCR cycles resulted in minor differences in the microbiota profile. Eventually, 30 amplification cycles with a DNA input of 125 pg resulted in sufficient amplicon concentrations for MiSeq sequencing and low background contamination.

This study has several strengths. We improved the laboratory processes by optimizing protocols in the workflow e.g. clean-up methods and PCR conditions. This resulted in an optimized MiSeq protocol where we even can analyse low-biomass samples. We used diluted positive controls to mimic the concentration

of low biomass samples and studied the influence of dilution solvents on the bacterial profile of these controls. To adjust for contaminating DNA, we included appropriate negative controls, which are extremely important when studying low biomass samples. We compared the MiSeq reagents kits with the same MiSeq settings. Our study also has some limitations. Despite the advantages of the Zymo mock as a positive control, it only contained few respiratory bacteria and had a low diversity. Preferably, we would like to use a mock which mimics the microbiota composition of NP samples, has a more diverse profile and consists of different ratios of bacteria. Furthermore, we did not include a sufficient number of Zymo mocks to test whether different PCR conditions and different MiSeq reagent kits have an influence on their bacterial profile.

Conclusion

In this study, we demonstrated the reliability of our DNA extraction and 16S rRNA gene MiSeq library preparation protocol for low biomass samples. Template concentration and number of PCR cycles had a modest influence on the microbiota profiles, while the PCR purification method and MiSeq sequencing kit had no significant effects on the microbial composition. Therefore, we propose to use samples with a DNA concentration of 0.1–20 pg/µl which can be amplified with 30 PCR cycles. After pooling, the library can be purified by two consecutive 0.9x AMPure XP purification steps and sequenced with the V3 MiSeq reagent kit. We confirmed that even extremely low density samples can be distinguished from DNA blanks. We present a benchmarked standardized workflow that when consistently and more widely used ensures comparability of results within and between studies. In addition, the workflow could be useful to study the microbiota of other low bacterial colonization density samples, e.g. lung, skin, blood, but also environmental samples in a standardized way.

Declarations

Data availability

Sequence data that support the findings of this study have been deposited in the National Center for Biotechnology Information Sequence Read Archive database with BioProject ID PRJNA718293.

Acknowledgements

The serosurveys in the Netherlands (PIENTER-3) and in Caribbean Netherlands (HSCN) are conducted by the National Institute for Public Health and the Environment (RIVM), in close collaboration with the local Public Health Services (GGD) and Statistics Netherlands (CBS). We would like to thank all volunteers who participated in this study. This work (salaries R.H., W.A.A.d.S.P.) was also supported by The Netherlands Organisation for Scientific research (NWO-VIDI; grant 91715359).

Author contributions statement

S.F., E.F., T.B. and D.B. conceived and designed the experiments. M.L.J.N.C., J.G., E.v.L., S.K and J.K. were responsible for the execution and quality control of the laboratory work. R.H., W.A.A.d.S.P., D.B. and T.B. analysed the data. R.H., D.B. and T.B. wrote the paper. All authors significantly contributed to interpretation of the results, critically revised the manuscript for important intellectual content and approved the final manuscript.

Competing interests

No authors report financial disclosures. None of the authors report competing interests.

References

1 Bosch, A. *et al.* Maturation of the Infant Respiratory Microbiota, Environmental Drivers, and Health Consequences. A Prospective Cohort Study. *Am J Respir Crit Care Med***196**, 1582-1590, doi:10.1164/rccm.201703-0554OC (2017).

2 Bosch, A. *et al.* Development of Upper Respiratory Tract Microbiota in Infancy is Affected by Mode of Delivery. *EBioMedicine***9**, 336-345, doi:10.1016/j.ebiom.2016.05.031 (2016).

3 Reyman, M. *et al.* Impact of delivery mode-associated gut microbiota dynamics on health in the first year of life. *Nat Commun***10**, 4997, doi:10.1038/s41467-019-13014-7 (2019).

4 Cho, I. & Blaser, M. J. The human microbiome: at the interface of health and disease. *Nat Rev Genet***13**, 260-270, doi:10.1038/nrg3182 (2012).

5 de Steenhuijsen Piters, W. A., Sanders, E. A. & Bogaert, D. The role of the local microbial ecosystem in respiratory health and disease. *Philos Trans R Soc Lond B Biol Sci***370**, doi:10.1098/rstb.2014.0294 (2015).

6 Human Microbiome Project, C. Structure, function and diversity of the healthy human microbiome. *Nature***486**, 207-214, doi:10.1038/nature11234 (2012).

7 Kamada, N., Chen, G. Y., Inohara, N. & Nunez, G. Control of pathogens and pathobionts by the gut microbiota. *Nat Immunol***14**, 685-690, doi:10.1038/ni.2608 (2013).

8 O'Hara, A. M. & Shanahan, F. The gut flora as a forgotten organ. *EMBO Rep***7**, 688-693, doi:10.1038/sj.embor.7400731 (2006).

9 Biesbroek, G. *et al.* Early respiratory microbiota composition determines bacterial succession patterns and respiratory health in children. *Am J Respir Crit Care Med***190**, 1283-1292, doi:10.1164/rccm.201407-1240OC (2014).

10 de Steenhuijsen Piters, W. A. A., Binkowska, J. & Bogaert, D. Early Life Microbiota and Respiratory Tract Infections. *Cell Host Microbe***28**, 223-232, doi:10.1016/j.chom.2020.07.004 (2020).

11 Man, W. H. *et al.* Respiratory Microbiota Predicts Clinical Disease Course of Acute Otorrhea in Children With Tympanostomy Tubes. *Pediatr Infect Dis J***38**, e116-e125, doi:10.1097/inf.0000000000002215 (2019).

12 Man, W. H., de Steenhuijsen Piters, W. A. & Bogaert, D. The microbiota of the respiratory tract: gatekeeper to respiratory health. *Nat Rev Microbiol***15**, 259-270, doi:10.1038/nrmicro.2017.14 (2017).

13 Biesbroek, G. *et al.* The impact of breastfeeding on nasopharyngeal microbial communities in infants. *Am J Respir Crit Care Med***190**, 298-308, doi:10.1164/rccm.201401-0073OC (2014).

14 de Steenhuijsen Piters, W. A. *et al.* Dysbiosis of upper respiratory tract microbiota in elderly pneumonia patients. *Isme j***10**, 97-108, doi:10.1038/ismej.2015.99 (2016).

15 Bogaert, D. *et al.* Variability and diversity of nasopharyngeal microbiota in children: a metagenomic analysis. *PLoS One***6**, e17035, doi:10.1371/journal.pone.0017035 (2011).

16 Salonen, A. *et al.* Comparative analysis of fecal DNA extraction methods with phylogenetic microarray: effective recovery of bacterial and archaeal DNA using mechanical cell lysis. *J Microbiol Methods***81**, 127-134, doi:10.1016/j.mimet.2010.02.007 (2010).

17 Ducarmon, Q. R., Hornung, B. V. H., Geelen, A. R., Kuijper, E. J. & Zwittink, R. D. Toward Standards in Clinical Microbiota Studies: Comparison of Three DNA Extraction Methods and Two Bioinformatic Pipelines. *mSystems***5**, doi:10.1128/mSystems.00547-19 (2020).

18 Salter, S. J. *et al.* Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biol***12**, 87, doi:10.1186/s12915-014-0087-z (2014).

19 Eisenhofer, R. *et al.* Contamination in Low Microbial Biomass Microbiome Studies: Issues and Recommendations. *Trends Microbiol***27**, 105-117, doi:10.1016/j.tim.2018.11.003 (2019).

20 Douglas, C. A. *et al.* DNA extraction approaches substantially influence the assessment of the human breast milk microbiome. *Sci Rep***10**, 123, doi:10.1038/s41598-019-55568-y (2020).

21 Wu, J. Y. *et al.* Effects of polymerase, template dilution and cycle number on PCR based 16 S rRNA diversity analysis using the deep sequencing method. *BMC Microbiol***10**, 255, doi:10.1186/1471-2180-10-255 (2010).

22 Polz, M. F. & Cavanaugh, C. M. Bias in template-to-product ratios in multitemplate PCR. *Appl Environ Microbiol***64**, 3724-3730 (1998).

23 Haas, B. J. *et al.* Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. *Genome Res***21**, 494-504, doi:10.1101/gr.112730.110 (2011).

24 de Muinck, E. J., Trosvik, P., Gilfillan, G. D., Hov, J. R. & Sundaram, A. Y. M. A novel ultra high-throughput 16S rRNA gene amplicon sequencing library preparation method for the Illumina HiSeq platform. *Microbiome***5**, 68, doi:10.1186/s40168-017-0279-1 (2017).

25 Kennedy, K., Hall, M. W., Lynch, M. D., Moreno-Hagelsieb, G. & Neufeld, J. D. Evaluating bias of illumina-based bacterial 16S rRNA gene profiles. *Appl Environ Microbiol***80**, 5717-5722, doi:10.1128/AEM.01451-14 (2014).

26 Biesbroek, G. *et al.* Deep sequencing analyses of low density microbial communities: working at the boundary of accurate microbiota detection. *PLoS One***7**, e32942, doi:10.1371/journal.pone.0032942 (2012).

27 Verberk, J. D. M. *et al.* Third national biobank for population-based seroprevalence studies in the Netherlands, including the Caribbean Netherlands. *BMC Infect Dis***19**, 470, doi:10.1186/s12879-019-4019-y (2019).

28 Wyllie, A. L. *et al.* Streptococcus pneumoniae in saliva of Dutch primary school children. *PLoS One***9**, e102045, doi:10.1371/journal.pone.0102045 (2014).

29 Kozich, J. J., Westcott, S. L., Baxter, N. T., Highlander, S. K. & Schloss, P. D. Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform. *Appl Environ Microbiol***79**, 5112-5120, doi:10.1128/AEM.01043-13 (2013).

30 Caporaso, J. G. *et al.* Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proc Natl Acad Sci U S A***108 Suppl 1**, 4516-4522, doi:10.1073/pnas.1000080107 (2011).

31 Reyman, M., van Houten, M. A., Arp, K., Sanders, E. A. M. & Bogaert, D. Rectal swabs are a reliable proxy for faecal samples in infant gut microbiota research based on 16S-rRNA sequencing. *Sci Rep***9**, 16072, doi:10.1038/s41598-019-52549-z (2019).

32 Joshi, N. A. & Fass, J. N. *Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files (Version 1.33)*, 2011).

33 Nikolenko, S. I., Korobeynikov, A. I. & Alekseyev, M. A. BayesHammer: Bayesian clustering for error correction in single-cell sequencing. *BMC Genomics***14 Suppl 1**, S7, doi:10.1186/1471-2164-14-S1-S7 (2013).

34 Masella, A. P., Bartram, A. K., Truszkowski, J. M., Brown, D. G. & Neufeld, J. D. PANDAseq: paired-end assembler for illumina sequences. *BMC Bioinformatics***13**, 31, doi:10.1186/1471-2105-13-31 (2012).

35 Caporaso, J. G. *et al.* QIIME allows analysis of high-throughput community sequencing data. *Nat Methods***7**, 335-336, doi:10.1038/nmeth.f.303 (2010).

36 Rognes, T., Flouri, T., Nichols, B., Quince, C. & Mahe, F. VSEARCH: a versatile open source tool for metagenomics. *PeerJ***4**, e2584, doi:10.7717/peerj.2584 (2016).

37 Wang, Q., Garrity, G. M., Tiedje, J. M. & Cole, J. R. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl Environ Microbiol***73**, 5261-5267, doi:10.1128/AEM.00062-07 (2007).

38 Quast, C. *et al.* The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res***41**, D590-596, doi:10.1093/nar/gks1219 (2013).

39 Costea, P. I. *et al.* Towards standards for human fecal sample processing in metagenomic studies. *Nat Biotechnol***35**, 1069-1076, doi:10.1038/nbt.3960 (2017).

40 Borgstrom, E., Lundin, S. & Lundeberg, J. Large scale library generation for high throughput sequencing. *PLoS One***6**, e19119, doi:10.1371/journal.pone.0019119 (2011).

41 Hawkins, T. L., O'Connor-Morin, T., Roy, A. & Santillan, C. DNA purification and isolation using a solid-phase. *Nucleic Acids Res***22**, 4543-4544, doi:10.1093/nar/22.21.4543 (1994).

42 Westen, A. A., van der Gaag, K. J., de Knijff, P. & Sijen, T. Improved analysis of long STR amplicons from degraded single source and mixed DNA. *Int J Legal Med***127**, 741-747, doi:10.1007/s00414-012-0816-1 (2013).

43 DeAngelis, M. M., Wang, D. G. & Hawkins, T. L. Solid-phase reversible immobilization for the isolation of PCR products. *Nucleic Acids Res***23**, 4742-4743, doi:10.1093/nar/23.22.4742 (1995).

44 McElhoe, J. A. *et al.* Development and assessment of an optimized next-generation DNA sequencing approach for the mtgenome using the Illumina MiSeq. *Forensic Sci Int Genet***13**, 20-29, doi:10.1016/j.fsigen.2014.05.007 (2014).

45 Ranjan, R., Rani, A., Metwally, A., McGee, H. S. & Perkins, D. L. Analysis of the microbiome: Advantages of whole genome shotgun versus 16S amplicon sequencing. *Biochem Biophys Res Commun***469**, 967-977, doi:10.1016/j.bbrc.2015.12.083 (2016).

46 Shao, W., Khin, S. & Kopp, W. C. Characterization of effect of repeated freeze and thaw cycles on stability of genomic DNA using pulsed field gel electrophoresis. *Biopreserv Biobank***10**, 4-11, doi:10.1089/bio.2011.0016 (2012).

47 Davis, N. M., Proctor, D. M., Holmes, S. P., Relman, D. A. & Callahan, B. J. Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data. *Microbiome***6**, 226, doi:10.1186/s40168-018-0605-2 (2018).

48 Proctor, D. M. *et al.* A spatial gradient of bacterial diversity in the human oral cavity shaped by salivary flow. *Nat Commun***9**, 681, doi:10.1038/s41467-018-02900-1 (2018).

Figures

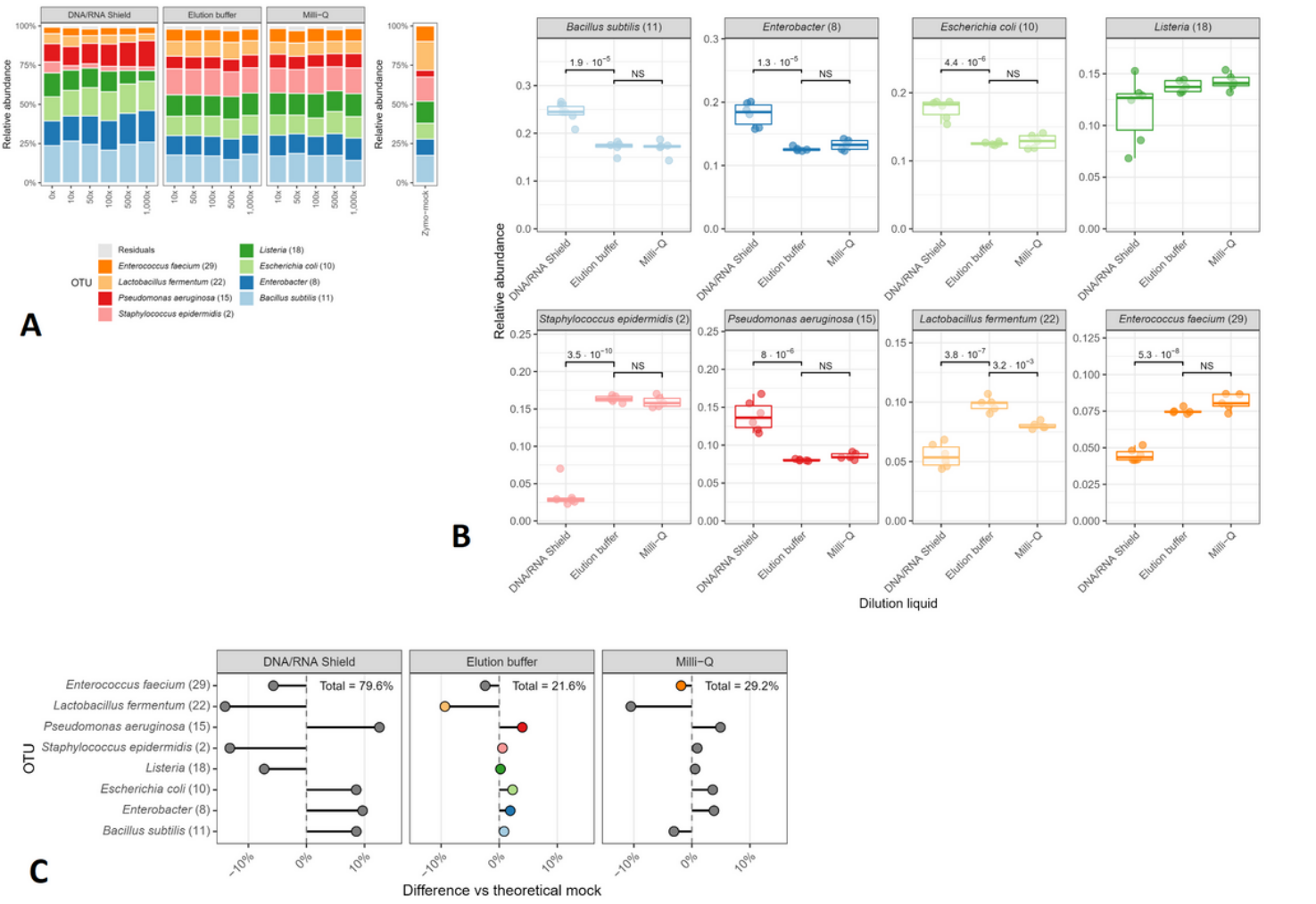

Figure 1

Bacterial composition of Zymo mock diluted in DNA/RNA shield (n=6; undiluted and 101-103x diluted), elution buffer (n=5; 101-103x diluted) and Milli-Q (n=5; 101-103x diluted). a. Stacked bar charts show the relative abundance of the top 8 operational taxonomic units (OTUs) in (un)diluted Zymo mock stratified by dilution solvent and the theoretical undiluted Zymo mock composition. The diluted Zymo mocks are annotated in the bioinformatic pipeline and have different annotations than the OTUs of the theoretical Zymo mock. Based on inspection of community profiles we found that the OTU annotated as Salmonella enterica refers to Enterobacter, Listeria monocytogenes as Listeria, Staphylococcus aureus as Staphylococcus epidermidis and Enterococcus faecalis as Enterococcus faecium. b. Boxplots show the relative abundance of each OTU in the dilution solvents. Boxplot depicts the 25th and 75th percentiles by lower and upper hinges, respectively, the median is depicted by a horizontal line in the box. The measurements that fall within 1.5 times the interquartile range are shown by whiskers. Statistical significance in relative abundance between dilution solvents were assessed by ANOVA test. A Tukey's post hoc test was used to determine statistical significance between elution buffer and DNA/RNA shield

or Milli-Q. c. Lollipop plot shows the differences in relative abundance of each OTU between the 103x diluted Zymo mocks and the theoretical Zymo mock. A strong positive value indicates a higher relative abundance of this OTU found than expected based on the theoretical mock and a strong negative value means that there is less of this OTU observed in the diluted Zymo mock. Coloured points indicate the lowest difference compared to the theoretical Zymo mock for that specific OTU across the dilution solvents. The percentage demonstrates the total cumulative absolute difference in relative abundance across the 8 OTUs compared to the theoretical mock.

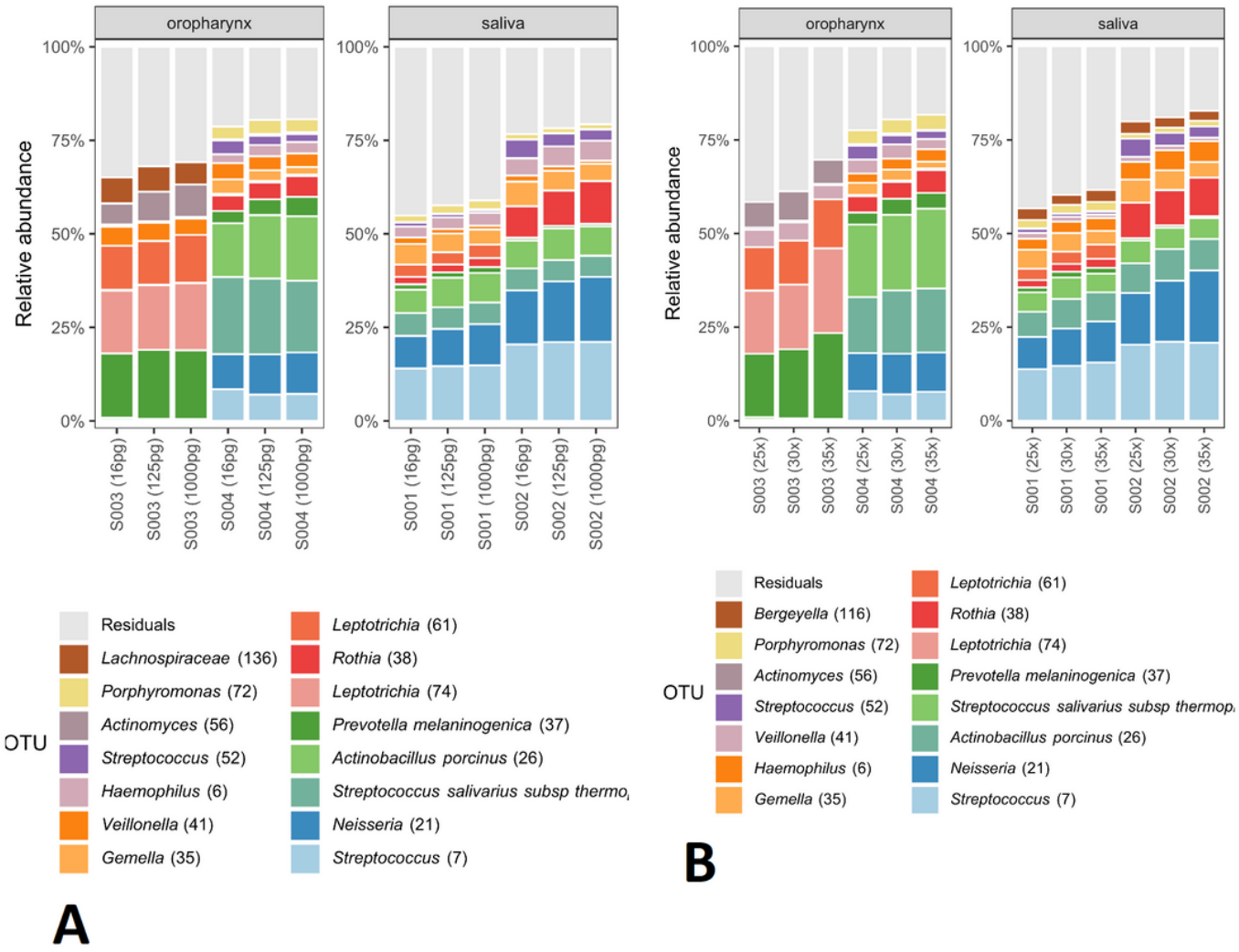

Figure 2

Microbiota composition profiles of 2 oropharynx (S003 and S004) and 2 saliva samples (S001 and S002). Stacked bar charts of the relative abundance of the top 15 OTUs and the residuals are shown. a. Microbiota profile of OP and saliva samples amplified by 25, 30 or 35 PCR cycles. b. Microbiota composition of OP and saliva samples with a DNA template input of 16, 125 or 1000 pg.

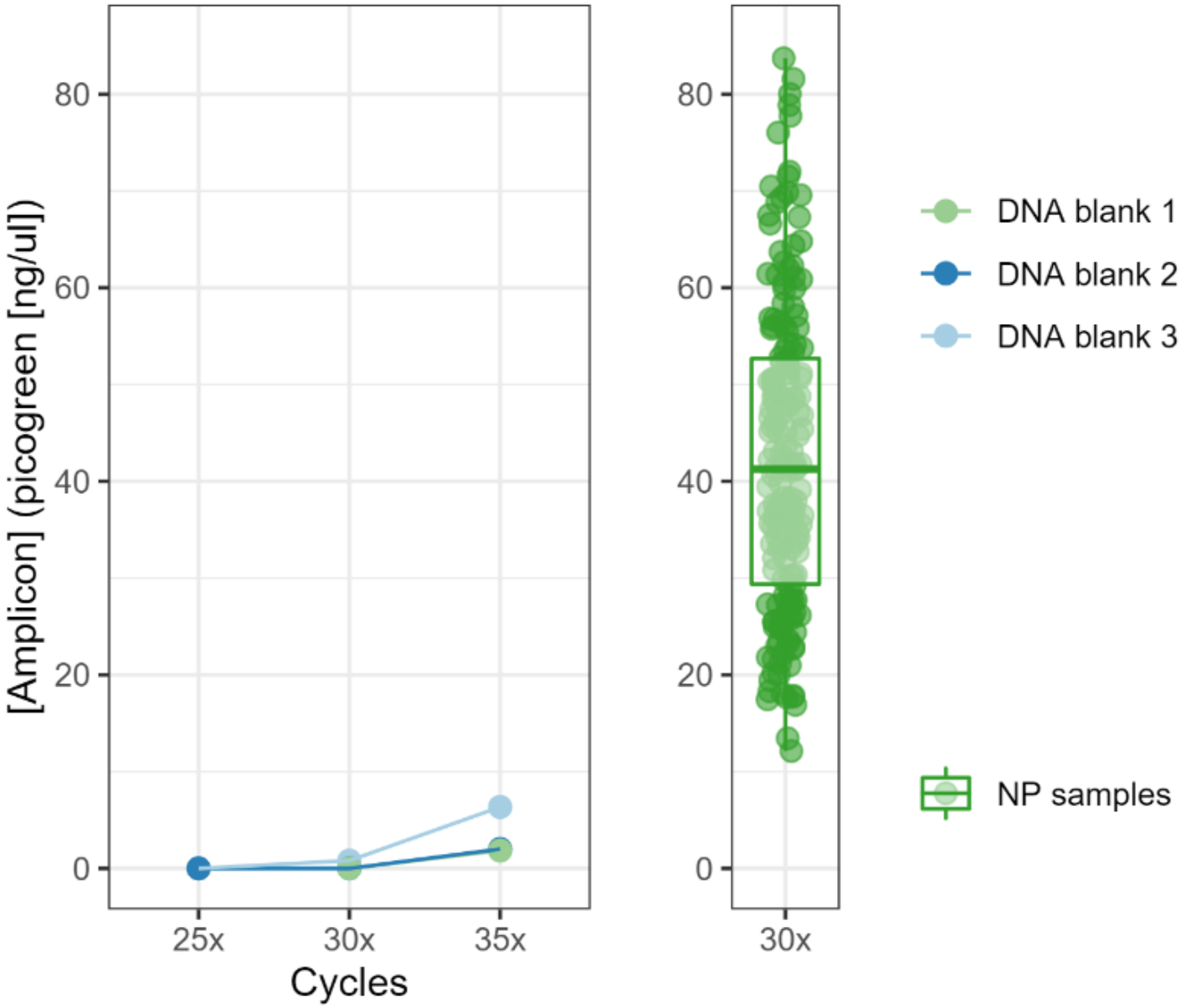

Figure 3

DNA concentration of 3 DNA blanks and 214 NP samples after PCR amplification. DNA concentration was estimated by picogreen quantification. The concentration of DNA blanks is visualized in the left graph shown for 25, 30 and 35 amplification cycles. The DNA concentration of the NP samples amplified by 30 PCR cycles are shown in the boxplot. NP samples had a median concentration of 41.3 ng/µl (range: 12.1-83.7). Boxplot depicts the 25th and 75th percentiles by lower and upper hinges, respectively, the median is depicted by a horizontal line in the box. The measurements that fall within 1.5 times the interquartile range are shown by whiskers. Each green dot is an individual NP sample.

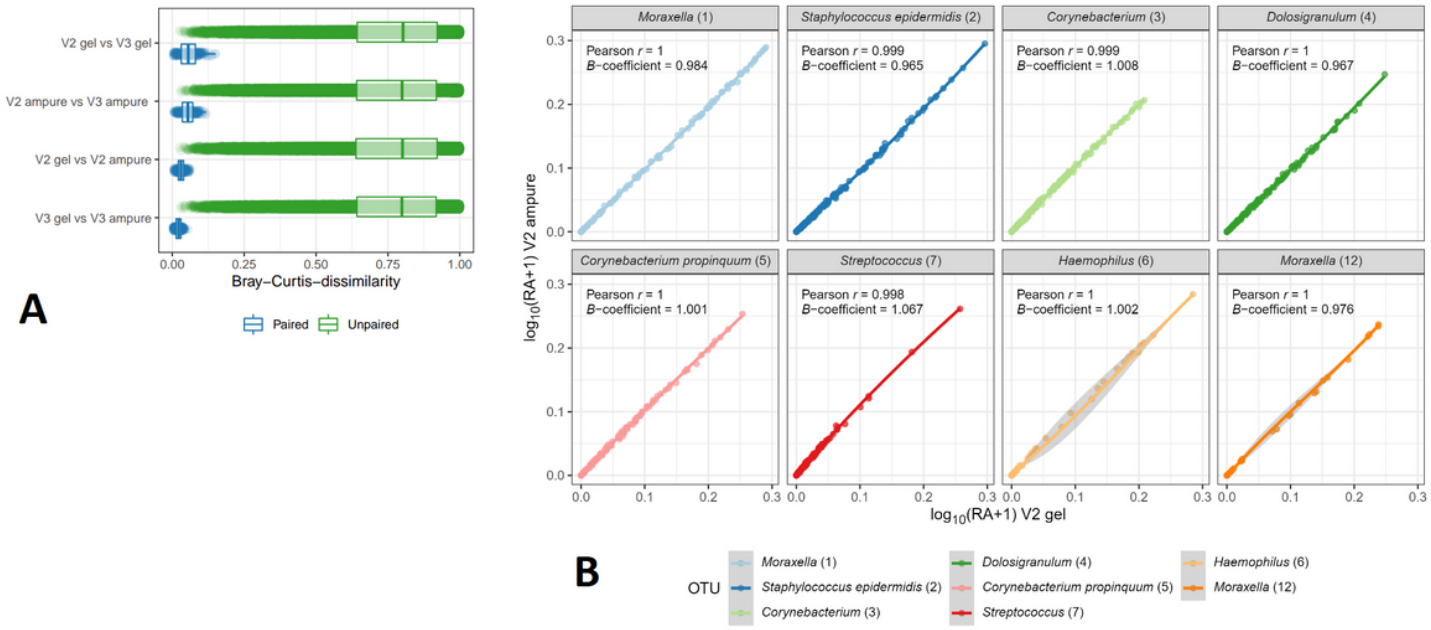

Figure 4

Similarity of a library (n=214) purified by gel or AMPure XP. a. To quantify differences in the overall microbial community composition between (pairs of) samples, Bray-Curtis dissimilarity was used, where zero indicates an identical composition between pairs. Unpaired dissimilarity was determined by calculating the dissimilarity of a given sample to all other (unpaired) samples in the other group, whereas paired dissimilarity refers to the dissimilarity between pairs of samples in both groups. Boxplot depicts the 25th and 75th percentiles by lower and upper hinges, respectively, the median is depicted by a horizontal line in the box. The measurements that fall within 1.5 times the interquartile range are shown by whiskers. b. Correlation plots visualizes log10+1-transformed relative abundances of the top 8 OTUs of a pool sequenced using the V2 reagent kit, comparing gel-based and AMPure XP purification methods. For each OTU, the Pearson correlation coefficient and regression coefficient (slope) was calculated. Both the correlation coefficient and the slope show a value close to 1.0, indicating a perfect correlation between purification methods for these OTUs.

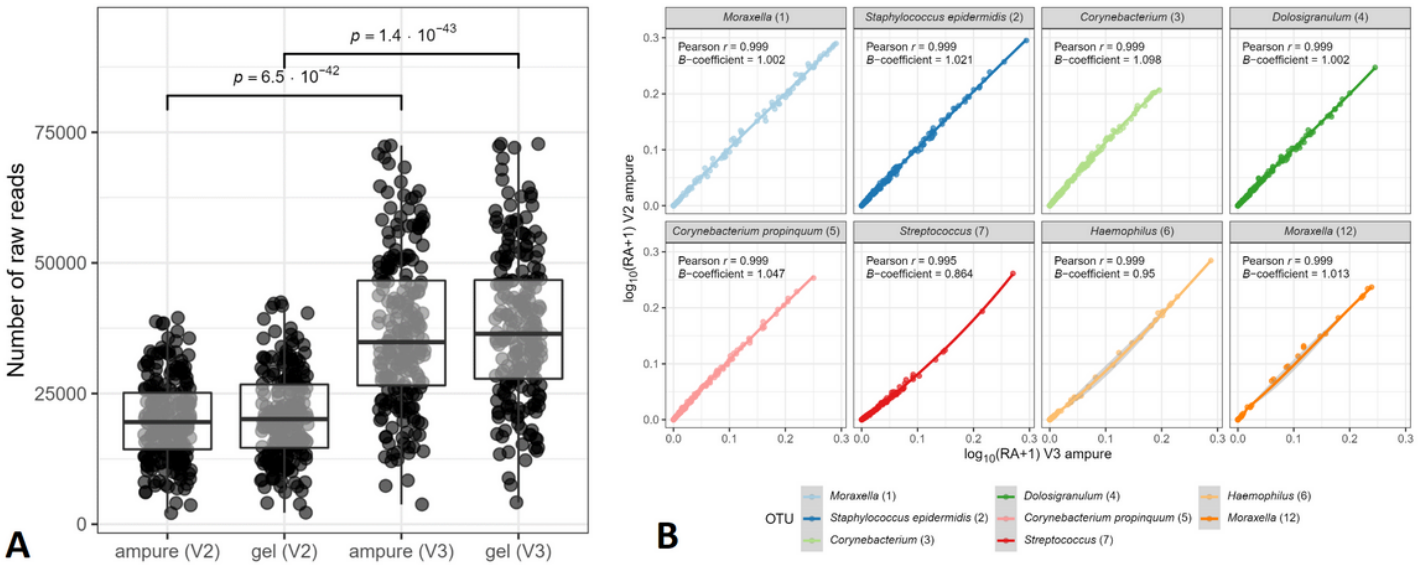

Figure 5

Similarity of a library (n=214) sequenced by MiSeq V2 and V3 kits. a. Number of reads of samples sequenced by V2 and V3 kit stratified by purification method. Statistical significance in number of reads between V2 and V3 kit was calculated by a linear model. b. Correlation plots visualize log10+1-transformed relative abundances of the top 8 OTUs of a pool purified by AMPure XP, comparing the V2 or V3 reagent kit. For each OTU, the Pearson correlation coefficient and regression coefficient was calculated. Both the correlation coefficient and almost all slopes show a value close to 1.0, indicating a near perfect correlation between purification methods for these OTUs.

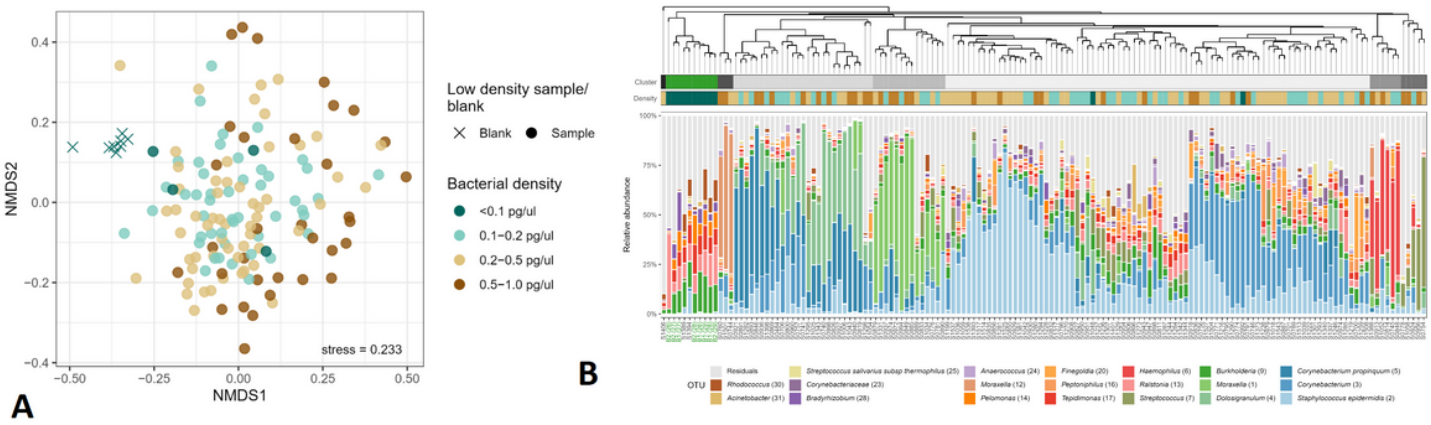

Figure 6

Microbiota profiles of nasopharyngeal (NP) samples (n= 140) and DNA blanks (n=8). a. Two-dimensional nonmetric multidimensional scaling (NMDS) plot, based on the Bray-Curtis dissimilarity matrix, visualizes the microbial community composition (each point) of the NP samples and DNA blanks (cross). Samples are stratified by DNA concentration based on 16S rRNA gene qPCR quantification (concentration pg/µl:

<0.1; 0.1-0.2; 0.2-0.5; 0.5-1.0). The stress value illustrates how well the high-dimensional data are captured in the two-dimensional space; a value of around 0.2 indicates an acceptable representation. b. Dendrogram visualizes an average linkage hierarchical clustering of NP samples and DNA blanks based on the Bray-Curtis dissimilarity index. The length of the branches represents the similarities between samples and DNA blanks. Stacked bar charts show the relative abundance of the top 20 OTUs and the residuals. Based on the clustering indices, 8 clusters were identified, 7 clusters solely consisted of NP samples (grey) and one separate cluster contained all DNA blanks and 2 NP samples (in green). Bacterial density colours correspond with colours used in the NMDS plot.