Elina Häivälä

Troubleshooting the Workflow of a Next-Generation Sequencing Library Preparation

Metropolia University of Applied Sciences

Bachelor of Laboratory Services

Laboratory Sciences

Bachelor's Thesis

29 February 2024

Abstract

Author:	Elina Häivälä
Title:	Troubleshooting the Workflow of a Next-Generation Sequencing Library Preparation
Number of Pages:	48 pages + 2 appendices
Date:	29 February 2024
Degree:	Bachelor of Laboratory Services
Degree Programme:	Laboratory Sciences
Supervisors:	Tuomas Heini, Research Technician Anne Salonen, Docent, Principal Investigator Tiina Soininen, Senior Lecturer

Sequencing-based methods, especially bacterial 16S ribosomal RNA gene amplicon analysis are widely used when studying complex microbial communities. Preparing libraries for Illumina NGS platforms includes multiplying the targeted area of the gene with universal bacterial primers and attaching Illumina-compatible P5 and P7 tails to the amplicons.

This thesis project was done in a research group that has generated high-throughput microbiome data with NGS technology from over 10 000 samples using the 16S rRNA gene. The research group has lately experienced varying quality in their NGS results. The aim of this study was to identify and troubleshoot possible sources for the challenges focusing mainly on the workflow for preparing the NGS libraries.

During the thesis study a few issues in the workflow were identified. A dilution factor error on a formula counting DNA concentration was found. It has likely influenced the effectiveness of PCR and caused forming of primer-dimers. Inspection of the paramagnetic bead-based library purification protocol revealed low recovery percentages of DNA independent of the used conditions. The recovery percentage was slightly improved by changing to a higher bead to sample -ratio. In library quantification, the attachment rate of the Illumina-compatible P5 and P7 -tails in the prepared libraries was found to be low. Using two separate PCR reactions to prepare the libraries improved the attachment rate, but not to a satisfactory level. The reason behind the low attachment rate requires further work. Finally, the length of the NGS libraries was found out to be shorter than expected, indicating a possible problem with the PCR.

Keywords: next-generation sequencing, NGS library, polymerase chain reaction, 16S rRNA gene, microbiota analysis

Tiivistelmä

Tekijä:	Elina Häivälä
Otsikko:	Ongelmanlähteiden selvittäminen NGS-kirjastojen valmistukseen liittyvissä työvaiheissa
Sivumäärä:	48 sivua + 2 liitettä
Aika:	29.2.2024
Tutkinto:	Laboratorioanalyytikko
Tutkinto-ohjelma:	Laboratorioanalytiikka (AMK)
Ohjaajat:	Tutkimusteknikko Tuomas Heini Dosentti, vastuullinen tutkija Anne Salonen Lehtori Tiina Soininen

Sekvensointipohjaiset menetelmät ovat suosittuja monimutkaisten mikrobiyhteisöjen tutkimiseen. Yleisimmin käytössä oleva tekniikka on bakteerien 16S ribosomaalisen RNA-geenin monistustuotteiden analyysi. Kirjastojen valmistamiseen Illumina-alustoille kuuluu kohdealueen monistaminen universaalialukkeilla ja Illumina-yhteensopivien P5- ja P7-häntien kiinnittäminen amplikoneihin.

Opinnäytetyö tehtiin tutkimusryhmässä, jossa on analysoitu yli 10 000 mikrobistonäytettä NGS-tekniikan avulla hyödyntäen 16S rRNA -geeniä. Tutkimusryhmällä on viime aikoina ollut haasteita NGS-tulosten laadun vaihtelun kanssa. Opinnäytetyön tavoitteena oli selvittää mahdollisia ongelmanlähteitä keskittyen pääosin NGS-kirjastojen valmistukseen liittyviin työvaiheisiin.

Työn aikana havaittiin haasteita useissa työvaiheissa. Laboratorion DNA:n konsentraatiota laskevasta kaavasta löydettiin laimennoskerroinvirhe, joka on todennäköisesti vaikuttanut PCR:n tehokkuuteen ja alukkeiden pariutumiseen keskenään. Kirjastojen puhdistusprotokollaa testatessa havaittiin, että paramagneettisiin helmiin pohjautuvan puhdistuksen aikana DNA:ta häviää odotettua enemmän olosuhteista riippumatta. DNA:n saantoprosenttia saatiin hieman parannettua muuttamalla helmien ja näytteen välistä suhdetta korkeammaksi. Kirjastojen laatua tutkittaessa Illumina-yhteensopivien P5- ja P7-häntien kiinnittymisaste havaittiin heikoksi. Kirjastojen valmistaminen kahdella erillisellä PCR-reaktiolla paransi kiinnittymisastetta, mutta ei tyydyttävälle tasolle. Syy alhaiseen kiinnittymisasteeseen vaatii lisätutkimusta. Lisäksi kirjastojen koko havaittiin lyhyemmäksi kuin odotettiin, mikä saattaa johtua ongelmista PCR:n aikana.

Avainsanat: toisen sukupolven sekvensointi, NGS-kirjasto, polymeraasiketjureaktio, 16S rRNA -geeni, mikrobistoanalyysi

Contents

List of Abbreviations

bp:	Base pair.
DNA:	Deoxyribonucleic acid.
DMSO:	Dimethylsulfoxide.
dsDNA:	Double-stranded DNA.
EDTA:	Ethylenediaminetetraacetic acid.
gDNA:	Genomic DNA.
HCl:	Hydrochloric acid.
PCR:	Polymerase chain reaction.
RNA:	Ribonucleic acid.
rRNA:	Ribosomal RNA.
RT:	Room temperature.
SDS:	Sodium dodecyl sulfate.
ssDNA:	Single-stranded DNA.
Tris:	Tris(hydroxymethyl)aminomethane.
qPCR:	Quantitative polymerase chain reaction.
16S rRNA:	30S subunit’s RNA component of a procaryotic ribosome.

1 Introduction

The thesis project was done in the Microbes Inside -research group. The research group is part of the Human Microbiome Research program (HUMI) in the Faculty of Medicine, Research Programs Unit, University of Helsinki. The focus of the Microbes Inside -group is on characterizing the composition and function of the human intestinal and female reproductive tract microbiota populations and what kind of connection they have on human health and diseases. The research group manages its own Health and Early life Microbiota (HELMI) birth cohort study started in 2016, collecting information of early life gut microbiota development based on over 10 000 faecal samples collected from the cohort. (Microbes Inside n.d.)

The main research activity of the group is generating high-throughput microbiome data with next-generation sequencing (NGS) technology. The research group prepares NGS libraries from different sample materials in their own laboratory. Over 80 % of the libraries are made from human faecal samples. Other sample types include gynaecological, tissue and swap samples, mainly of human origin. The group prepares up to 5000 NGS libraries per year for their own research and for different collaborations. A functioning workflow to prepare the libraries is essential for the group.

For the past year the research group has had challenges that have been seen in the varying quality of NGS results. The aim of this study was to examine the laboratory's library preparation protocols and to investigate possible reasons behind the quality changes in NGS results focusing on the laboratory work. Bioinformatics and analysing of NGS results were not included in this study.

2 Theory

2.1 Studying Complex Microbial Communities with 16S rRNA Gene

Studying and understanding complex microbial communities can shed light on e. g. how the environment or human health is affected by the composition of and changes in microbial communities. Studying microbial communities with cultivation-based methods has been proven to be a difficult and unreliable method to capture the full diversity of bacteria. Finding suitable and comparable cultivation conditions for different groups or species is challenging. This has led to an increasing interest and popularity in sequencing-based methods. (Klindworth et al. 2012.)

The most popular cultivation-independent tool for studying microbial communities is using 16S ribosomal RNA (16S rRNA) gene amplicon analysis (Klindworth et al. 2012). The 16S rRNA gene is found in all bacteria and its structure is ideal for this type of research. The gene is approximately 1550 base pairs (bp) long and consists of both variable and conserved regions. The conserved regions are well preserved since the gene codes for important parts of the cell functions. The conserved regions provide an attachment site for universal primers for amplifying DNA with PCR making it possible to study most families and species of bacteria simultaneously. The variable areas provide enough variation for family or species level identification and statistically valid results. (Clarridge 2004.)

Millions of 16S rRNA gene sequences are available in sequence databases for comparisons. When studying completely new species the whole gene can be sequenced, but generally using a smaller part of it provides enough information for identification. (Clarridge 2004.) Choosing the right area and the right primers for the analysis is crucial. Some bacterial groups can be underrepresented, or certain species or whole groups can be unintentionally selected against with poor primer design. This can lead to misinterpretations of the composition of the microbial community. (Klindworth et al. 2012.)

2.2 Next Generation Sequencing

DNA sequencing methods have experienced a rapid development since they were first invented. First-generation sequencing methods were based on producing DNA fragments in which the last nucleotide of the fragment was known and using electrophoresis to separate the fragments by length. The order of the nucleic acids was found out fragment by fragment. Next-generation sequencing techniques, also called second-generation sequencing, are based on binding template DNA molecules to a surface. Attachment of individual nucleotides is detected by fluorometry during the synthesis of the DNA molecules complimentary to the templates. NGS techniques brought with them a genomics revolution, enabling a faster and more cost-effective way for acquiring sequencing data. (Heather and Chain 2016.) While it took 15 years to sequence the first complete human genome, dozens of human genomes can now be sequenced in a day with a single NGS machine. NGS techniques can be used in a variety of different methods in genomics, transcriptomics and epigenomics. (Illumina 2017.)

The current market leader in the NGS field is Illumina Inc. (San Diego, United States) who provides several different platforms for NGS. The most popular Illumina platforms have been HiSeq and MiSeq, the former being discontinued. The difference between these platforms is acquiring cost, cost per run, throughput of a run and the length of the DNA sequence read. (Heather and Chain 2016.) The Illumina NGS workflow consists of four steps: library preparation, cluster generation, sequencing, and data analysis (Illumina 2017).

Before DNA can be sequenced it needs to be made into sequencing libraries. Sequencing library is a pool of DNA fragments of similar size. The fragments contain adapter sequences compatible with the selected sequencing platform and indexes, also known as barcodes, for identification of individual samples. (Integrated DNA Technologies n.d.) Library preparation methods vary according to the aim of the study and whether a whole genome or a part of it is studied. Libraries are generally prepared by making amplicons of the region of interest,

and adding adapters to both ends of the amplicon with polymerase chain reaction (PCR). In Illumina platforms, the adapters, called P5 and P7 tails, are needed in the cluster generation phase. They allow the amplicons to bind to the flow cells in the NGS instrument. (Illumina 2017.)

In the cluster generation step the libraries are loaded into the flow cells of the Illumina NGS instrument. The surface of the flow cell is covered with oligonucleotides complimentary to the P5 and P7 adapters. The DNA templates bind to the flow cells and each template is amplified several times with PCR forming local clonal clusters. Following cluster generation the templates are sequenced. Sequencing is done by synthesis using reversible terminator nucleotides. The synthesis can continue only after the release of the fluorescent particle blocking the attachment site for the next nucleotide. This allows the sequencing to proceed in synchronous cycles. (Illumina 2017.)

In each cycle the newly attached nucleotides are detected by reading a signal from the exiting fluorescent particle. A four-channel image is produced, each channel representing one of the bases. In a paired-end (PE) sequencing both forward and reverse reads of the DNA template are performed and analysed as read pairs. The fourth step of the workflow is collecting and analysing the generated read data. (Illumina 2017.)

2.3 Indexed 16S rRNA Gene Amplicon Library Preparation for Illumina NGS Platforms

Preparing indexed 16S rRNA gene amplicon libraries for NGS in Illumina platforms involves amplifying the targeted part of the gene with universal primers and attaching P5 and P7 -adapters with unique indexes at both ends of the DNA amplicon (Figure 1). The indexes are short sequences that make it possible to pool several libraries into one NGS run and allocate each read back to a specific library in the data analysis step. (Illumina n.d.a.) Instead of using the same index in both ends of the amplicon, two different indexes, often referred to as i5 and i7, can be used. Each library is given a unique index pair

combination, which greatly reduces the number of indexes needed. For hundred samples the number of required indexes comes down from 100 individual indexes to 100 individual index pairs that can be reached using only 20 indexes. (Kozich et al. 2013.)

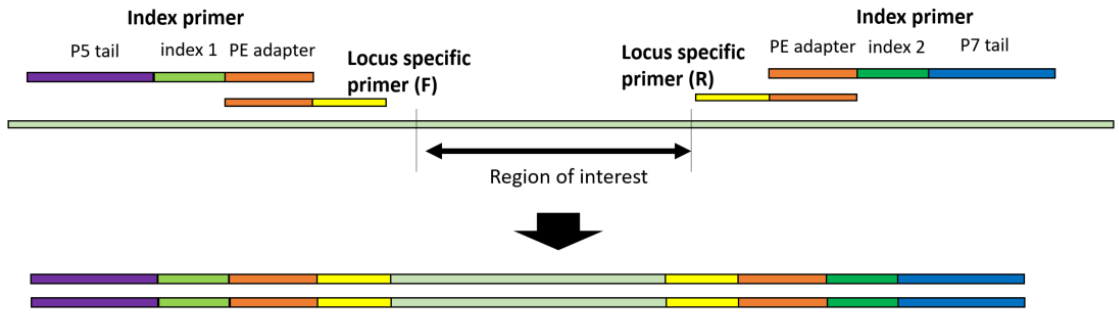

Figure 1. Dual-indexed 16S rRNA gene amplicon. Adapted from Raju et al. (2018).

Library preparation can be done in two PCR steps (two-step PCR) or in one PCR step (one-step PCR). In two-step PCR the first step is to restrict and amplify the targeted 16S region with locus specific primers including PE adapters. (Kozich et al. 2013.) The primers bind to both sides of the region of interest and initiate the amplification. The second step is another PCR using primers with the P5 and P7 Illumina compatible tails and indexes. The 16S rRNA gene amplicons from the first step are used as templates. (Raju et al. 2018.)

In the one-step method both primers are included in a single PCR run making it more cost effective, less time consuming and less prone to well-to-well contamination due to reduced processing steps. It may also reduce the amount of unwanted artifacts related to the high number of cycles performed during the two-step PCR method (Kozich et al. 2013). In the one-step method, the locus specific primers are consumed during the first cycles of amplification. The index primers will amplify the intermediate products in the following cycles creating the finished Illumina-compatible libraries. (Raju et al. 2018.)

16S rRNA gene amplicon libraries can be challenging in NGS runs. The conserved areas of the 16S rRNA gene make it possible to study many bacterial species simultaneously, but the low diversity in the beginning of the amplicons is problematic when sequencing with Illumina MiSeq platforms. The first cycles of a MiSeq run rely on the heterogeneity of the base composition of the targeted amplicons, which the conserved areas of the 16S rRNA gene lack. (Fadrosh et al. 2014.) The location of the clusters, phasing and colour matrix corrections are calculated during the first 25 cycles of a MiSeq sequencing. Calculations are done from the four-channel image and if one channel produces a high percentage of the image, the instrument will have problems identifying the location of the clusters and analysing the images when processing the data. (Illumina n.d.b.)

Homogeneity of the 16S rRNA gene amplicon libraries can be artificially reduced by adding a heterogenous control library to the library pool (Fadrosh et al. 2014). The PhiX Control v3 Library is the most used with percentages commonly varying between 5–10 %. The PhiX library is derived from a bacteriophage genome consisting of a balanced base composition of 45 % GC and 55 % AT. (Illumina n.d.c.) Using PhiX makes it possible to do successful runs with 16S rRNA gene amplicons but some of the sequence reads are lost to the non-targeted PhiX-template. Other methods to add heterogeneity have also been developed, e.g. adding a varying number of base pairs to the primer sequences which causes the samples to be sequenced out of phase and thus reduces the homogeneity of the bases in each cycle. (Raju et al. 2018; Fadrosh et al. 2014.)

2.4 Challenges Related to Library Preparation with PCR

PCR used in library preparation is a very sensitive technique both in the sense of being able to detect very small quantities of target DNA and being vulnerable to interferences. (Sidsted et al. 2019.) PCR protocols often need to be optimized for the samples of interest. PCR is sensitive to the quality of the template DNA, batch-related variability of reagents, problems related to the

thermal cycler, and PCR inhibitors. PCR inhibitors target the amplification process and can originate from the sample matrix, e.g. from the sample material or from the reagents used in extracting or storing DNA. These reagents include ethylenediaminetetraacetic acid (EDTA) and sodium dodecyl sulfate (SDS). (Monteiro et al. 1997.) Internal controls in PCR runs are important since they can reveal the presence of inhibitors, contaminants, and other problems in the reaction (Oikarinen et al. 2008).

Inhibitors present in the samples can lead to false negative results or insufficient yields of PCR products. DNA samples extracted from faecal samples and blood are one of the hardest sample materials to study as clinical samples often contain a varying number of inhibitors (Monteiro et al. 1997). There are several possible inhibitors present in clinical samples, including haemoglobin, complex polysaccharides, phenolic compounds, glycogen, fats, cellulose, constituents of bacterial cells, non-target nucleic acids and heavy metals (Oikarinen et al. 2008; Monteiro et al. 1997; Sidsted et al. 2019).

During PCR reactions, especially in case of low template concentration, primers can sometimes dimerize with other primers, resulting in untargeted fragments called primer-dimers. Primer-dimers may have the required sequence to bind to the NGS machine's flow cells, to form clusters and be sequenced. They are shorter than the target templates and make clusters more efficiently. If the proportion of primer-dimers is high, they can greatly decrease the number of reads from target templates and even stop the run completely. It is important to purify the libraries from primer-dimers and unpaired primers prior to sending them to be sequenced. (Illumina 2023.) Libraries and library pools can be purified using different methods: paramagnetic beads binding the DNA, solid-phase purification using DNA absorbing columns or reagents destroying ssDNA and single nucleotides. (Wang et al. 2021; Beckman Coulter Life Sciences n.d.)

The paramagnetic beads bind DNA molecules leaving contaminants and small DNA fragments in the solution. They are first mixed to the library, and then immobilized and bound out of the solution by a magnetic field. The beads are

magnetic only in the presence of an external magnetic field preventing them from clumping together while binding to DNA. The solution containing the contaminants and unwanted fragments is then discarded. The beads are next washed with ethanol to remove any remaining impurities. In the end DNA is eluted from the beads by removing the magnetic field, adding the desired eluting solution, and binding the beads again with magnets. (Wang et al. 2021; Beckman Coulter Life Sciences n.d.) The ratio of the beads to the sample volume can be used in size selection favouring certain DNA fragment lengths. The process is based on the chemical composition of the buffer solution the beads are in and how well different DNA fragment sizes stay attached to the beads in that specific chemical environment. (Wang et al. 2021; Hawkins 1998.)

3 Aims of This Study

The Microbes Inside -research group has been using the same protocol for library preparation for several years with both good results and occasional challenges. The laboratory's one-step indexed 16S rRNA gene amplicon library preparation protocol is based on a protocol made by a sequencing core facility. The protocol has been optimized for the laboratory and sample types. No apparent changes have been made in the laboratory before or during the latest challenges. The protocols or reagents used have not been changed recently. The personnel in the laboratory as well as the sequencing core facility have changed but not at the same time with the quality changes. All the personnel in the laboratory follow the same protocols.

The aim of this study was to focus on troubleshooting two phenomena:

1) **DNA concentrations of the indexed 16S rRNA gene amplicon libraries have declined.** The first aim of this study was to find out why the concentrations have declined. This was investigated by going through and testing protocols and faecal DNA extracts used in the library preparation workflow. The main focus was on the DNA concentration

measurement methods before the library preparation and library purification protocols after the libraries are prepared.

2) **The varying quality of the NGS results.** The second aim of this study was to investigate the quality of the libraries prepared in the laboratory. Poor quality libraries might compromise the NGS results. This was studied by testing two different methods, one-step and two-step index-PCR for library preparation and comparing the quality of the libraries visually from agarose gel and with library quantification analysis.

4 Methods

4.1 From Sample to NGS Data

The route from sample collection to NGS has several phases both inside and outside the research group's laboratory (Figure 2). Samples of human origin are collected by the study subjects themselves or by health care professionals. They are stored at -20 °C or -80 °C until arrival to the research group's laboratory, where the samples are stored at -80 °C.

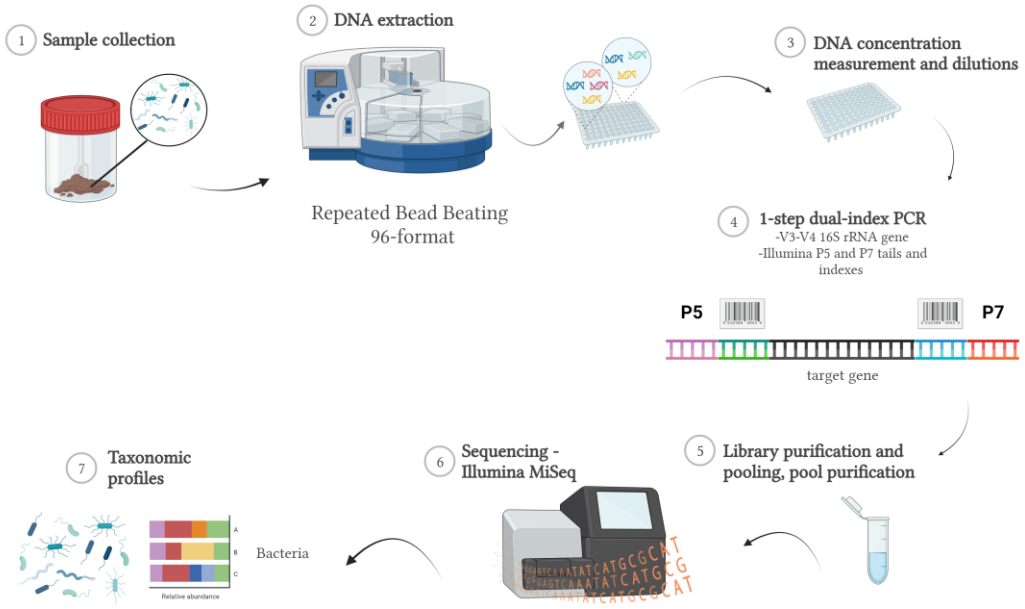

Figure 2. Route from sample to NGS data. Created with Biorender.com.

DNA is extracted from the samples by repeated bead beating method using a KingFisher Flex automated purification system (ThermoFisher Scientific, Massachusetts, United States) (Jokela et al. 2022) and stored in elution buffer (10 mM Tris-HCl, 0.5 mM EDTA, water, pH 8.5–9) at -20 °C. Concentration of the extracted DNA is measured after extraction and dilutions for later applications are prepared. Dilutions are stored at -20 °C.

Bacterial V3–V4 16S rRNA gene amplicon libraries for Illumina NGS are prepared from the extracted DNA. The PCR protocol used in the laboratory is modified from Illumina's protocol (Illumina n.d.a), using a dual-index strategy (Kozich et al. 2013) and one-step PCR (Raju et al. 2018) The primers targeting the V3–V4 region are based on Klindworth et al. (2012). More detailed information about the primers is provided in the Appendix 1. Target amplicon size is 600–640 bp (Raju et al. 2018). The libraries are purified after PCR using paramagnetic beads. The concentration of DNA is measured after purification and the libraries are pooled to a certain molarity based on the measurement. The library pool is then purified using the same magnetic beads used for library purification. NGS run is performed by a sequencing core facility and bioinformatics are handled in the research group.

4.2 Sample Information

All samples used for the experiments of this study were DNA extracted from human faecal samples. The extracted DNA has been stored in elution buffer at -20 °C. More detailed information about the samples is provided in Table 1. The samples derive from research projects approved by the ethical committee of The Hospital District of Helsinki and Uusimaa and Helsinki region hospital district (21/13/03/03/2014) and performed in accordance with the principles of the Helsinki Declaration. All participants signed an informed consent.

Table 1. Sample information.

Sample group and type	**DNA extracted**	**Details**
SG1 Adult faecal samples	2016	Set of eight samples. Used in: Comparing different concentration measurement methods.
SG2 Adult faecal samples	2020	Set of eight samples. Used in: Libraries for comparing one-step and two-step index-PCR. Libraries for testing the purification protocol.
SG3 Adult faecal samples	2023	Set of six samples. Used in: Libraries for the KAPA library quantification kit analysis.

The samples were stored either in +4 or -20 depending on the time between the experiments. Unnecessary freeze-thaw cycles were avoided.

4.3 Excluding Possible Problem Sources

Some possible problem sources, such as degradation or erroneous synthesis of primers, could be with fair confidence ruled out and were therefore not a part of this study. Since the one-step index-PCR protocol had already been optimized for use in the laboratory, it was not chosen as a priority for this study. During previous challenges some library and NGS run parameters were compared with the quality of the NGS results together with the sequencing core facility performing the NGS runs, but no clear factors had been identified.

Reagent batch effect was not considered a likely source for the problems. The reagents used in the laboratory are shared by other researchers, and they are not experiencing problems with their PCR or NGS results. They are using different sample materials and a two-step library preparation method, but problems related to the reagents would have come up despite these

differences. The reagent batches have also been changed and tested after the challenges were observed without notable differences in the results.

The PCR thermal cycler used by the laboratory has been tested by comparing quantitative PCR (qPCR) results from the same samples in another thermal cycler. No notable differences were observed.

A set of libraries were prepared following the current one-step protocol and added to the library pool of another researcher for an NGS run. This was done to see if the problem comes from the NGS run. The other researcher's samples from the same run performed better, which indicates that the problem is not likely originating from the NGS instrument or run settings.

4.4 DNA Concentration Measurements

The laboratory has three methods available for measuring DNA concentration: a NanoDrop ND-1000 Spectrophotometer (Marshall Scientific, Hampton, United States), a Qubit dsDNA HS assay and Qubit 4 Fluorometer (Invitrogen by Thermo Fisher Scientific, Massachusetts, United States), and a Quant-iT PicoGreen dsDNA assay (Invitrogen, Massachusetts, United States), hereafter referred to as NanoDrop, Qubit and Quant-iT, respectively. These three methods were compared by measuring the same eight genomic DNA (gDNA) samples extracted from adult faecal samples with each method. All samples used for the measurements belonged to SG1.

NanoDrop is a spectrophotometric method based on UV absorbance in certain wavelengths (260 and 280 nm). It can be used to measure the concentration of DNA and to evaluate the purity of a DNA sample. From the three compared methods, NanoDrop is the fastest since no sample preparation or standards are needed. (Desjardins and Concklin 2010.)

A blank sample (the elution buffer) was first measured. Its absorbance was deducted from the values measured from the samples by the NanoDrop's own

program. Two microlitres of sample was pipetted directly on the optical pedestal of the machine. The concentration and purity of the samples were calculated by the NanoDrop program.

Qubit and Quant-iT are both fluorometric methods based on fluorescence measurements of a fluorescent dye that attaches to double-stranded DNA (dsDNA) (ThermoFisher Scientific 2016). Quant-iT is used for measuring several samples at once on a 96-well plate. Qubit is used for a smaller set of samples, since each sample is measured individually with the Qubit 4 Fluorometer.

The laboratory's own protocol for Qubit measurements was followed. Two standards were first prepared by mixing 10 µl of the appropriate standard and 190 µl of Qubit working solution containing the fluorescent dye. The standards were measured first and the Qubit fluorometer's algorithm made a standard curve based on the results. The samples were prepared by pipetting 2 µl of sample to 198 µl of working solution. The volume of sample used needs to be entered in the fluorometer by the user for the program to automatically calculate the dilution factor. The sample concentrations were calculated by the fluorometer's software based on the measured fluorescence, standard curve, and dilution factor.

The laboratory's own protocol for Quant-iT measurement was followed. Five standard samples (between 0–1000 ng/ml of dsDNA) were prepared and measured simultaneously with the samples on a 96-well plate. Next, 2 µl of sample was mixed with 98 µl of 1xTE buffer and 100 µl of Picogreen fluorescent dye. The plate was read with a Hidex Sense microplate reader (Hidex Oy, Finland, Turku) to get the raw data. Deducting the blank (standard sample with 0 ng/ml of dsDNA), fitting the standard curve and calculating the sample concentrations based on the curve were done using the laboratory's own Excel template.

Due to scheduling reasons the measurements with Quant-iT were done a month earlier than the Qubit and NanoDrop measurements. The samples were stored at -20°C between the measurements and experienced one extra freeze-thaw cycle before they were measured with Qubit and NanoDrop.

4.5 Comparing One-Step and Two-Step Library Preparation

A diagram of the workflow of the one-step and two-step library preparations is provided in Appendix 2.

4.5.1 One-Step Library Preparation

One-step PCR to prepare indexed 16S rRNA gene amplicon libraries was done in 20 µl reactions with the composition of 1x Phusion Master Mix (ThermoFisher Scientific, Massachusetts, United States), 0.25 µM both forward and reverse TruSeq 16S primers (Merck KGaA, Darmstadt, Germany), 0.375 µM both P5 and P7 primers with indexes (Merck KGaA, Darmstadt, Germany), 5 ng template DNA, 3 % dimethylsulfoxide (DMSO) (ThermoFisher Scientific, Massachusetts, United States) and 3.4 µl water per well. More detailed information about the primers is provided in the Appendix 1.

Template DNA and index primers were pipetted separately before the master mix. The laboratory had recently moved to using dilutions of 5 ng/µl of template DNA instead of the 1 ng/µl mentioned in the protocol, since these currently performed better in the PCR. The 5 ng/µl dilutions were used for the experiments of this study, raising the amount of DNA template to 25 ng per reaction.

The thermal cycling conditions for one-step PCR are shown in Table 2. In all PCR reactions the same faecal DNA sample was always used as a positive control and water was used as a non-template control (blank sample). A C1000 Touch Thermal Cycler with a CFX96 Optical Reaction Module (Bio-Rad, California, United States) was used in all PCRs.

Table 2. Thermal cycling conditions for one-step index-PCR.

Temperature (°C)	Time	Cycles
98	60 s	1
98 64 72	10 s 30 s 30 s	27
72	10 min	1
4	forever	

Afterwards DNA concentrations were measured with Quant-iT. Several one-step libraries were prepared for the experiments of this study (Table 3).

Table 3. List of libraries used in this study prepared with one-step PCR method.

Used for	Used indexes	Number of libraries	Sample group
Comparing one-step and two-step PCR (unsuccessful runs)	SD501–SD508 & SD708	8	SG2
	SD501–SD508 & SD701–SD706	8	
Comparing one-step and two-step PCR	SD506 & SD706 same in each library	8	SG2
Library purification test 1	SD506 & SD706 same in each library	16	SG2
Library purification test 2	SD505 & SD705 same in each library	16	SG2
Library purification test 3	SD504 & SD704 same in each library	40	SG2
KAPA library quantification kit analysis	SD501–SD506 & SD705	6	SG3

4.5.2 Two-Step Library Preparation

Dual-indexed 16S rRNA gene amplicon libraries were prepared with two-step PCR according to the protocol used in the laboratory. The same primer sets which were used in the one-step PCR were used in the two-step method, but in two subsequent PCR reactions. Eight libraries were prepared from samples from SG2.

The reaction mixture was 1x Phusion Master Mix, 0.375 µM both forward and reverse Truseq 16S primers, 25 ng template DNA, 3 % DMSO and 5.4 µl water per well. Template DNA was pipetted separately before the master mix. The total reaction volume per well was 20 µl. Thermal cycling conditions are listed in Table 4 for both steps.

Table 4. Thermal cycling conditions for two-step index-PCR.

		First step (16S PCR)	**Second step (index-PCR)**
Temperature (°C)	**Time**	**Cycles**	**Cycles**
98	60 s	1	1
98 64 72	10 s 30 s 30 s	27	30
72	10 min	1	1
4	for ever		

After the first PCR, the products were run on agarose gel to confirm the presence of right sized products and that there was nothing in the blank. This was done to determine if the amplification was successful and that there was no

PCR contamination. The products were then purified using paramagnetic beads. Concentrations of the libraries were measured with Qubit and 5 ng/µl dilutions of DNA were prepared for the second step.

In the second step a dual-index PCR was done using a reaction mixture of 1x Phusion Master Mix, 0.5 µM both P5 and P7 primers with indexes (SD507 and SD707 for all samples in comparing the one-step and two-step PCRs and SD501–506 + SD708 for the library quality analysis), 4 ng of template DNA, 3 % µl DMSO and 1.4 µl water. Template DNA and primers were pipetted separately. The total reaction volume was 20 µl. Thermal cycling conditions are listed in Table 4 above. The indexed libraries were run on agarose gel to confirm if the PCR was successful.

4.5.3 Gel Electrophoresis and Library Purification

Gel electrophoresis was performed following a protocol used in the laboratory. Self-prepared 1.5 % agarose gel dyed with Midori Green Advance DNA Stain (Nippon Genetics Europe, Düren, Germany) was used. GeneRuler 1 kb DNA Ladder (ThermoFisher Scientific, Massachusetts, United States) was used as a size standard. The wells were filled with 3 µl of sample and 1 µl of 6x Gel DNA loading dye (ThermoFisher Scientific, Massachusetts, United States). The gels were run at 120 V for 45 minutes. Pictures of the gels were taken with Gel Doc XR+ Molecular Imager with Image Lab Software (Bio-Rad, California, United States).

Purification of the libraries was done with magnetic beads according to the library purification protocol used in the laboratory. Purification was performed on a 96-well plate with a compatible magnetic rack (Invitrogen by ThermoFisher Scientific, Massachusetts, United States). The protocol involved the following steps:

- Magnetic beads are added in each well in a 0.6:1 bead to sample volume ratio, vortexed and spun down.
- The mixture is incubated at room temperature (RT) for 5 min.

- The plate is placed on a magnetic rack and incubated for 2 min at RT.
- 150 µl of freshly made 80 % ethanol solution is added, incubated for 60 seconds, after which the ethanol is carefully removed. Repeated once.
- The beads are dried for 15 min at RT.
- The plate is removed from the magnetic rack and 30 µl of elution buffer is added. The plate is vortexed and spun down. Incubated for 5 min at RT.
- The plate is returned to the magnetic rack and incubated for 2 min.
- 20 µl of the purified product is transferred to a new plate.

The concentrations of purified libraries were measured with Quant-iT. The purified libraries were stored at -20 °C. The plates containing the rest of the purified products mixed with the beads were kept in the fridge for a couple of days.

4.5.4 Libraries for KAPA Library Quantification Kit

In addition to the fluorometric methods that quantify all dsDNA in the libraries, a KAPA library quantification kit (hereafter KAPA kit) for Illumina platforms was used. The KAPA kit analysis was performed by a sequencing core facility. The analysis is a qPCR based method for quantifying NGS libraries. The method uses primers that attach to the P5 and P7 tails. Only templates with correctly attached P5 and P7 tails will amplify during the PCR reaction and hence the method only quantifies sequencing-compatible full-length library fragments. The concentration of the templates with the tails can be calculated using a standard and compared with the concentration of dsDNA in the library. This gives the attachment rate of the P5 and P7 primers. (KAPABiosystems 2020.)

Two pools were prepared for the KAPA kit analysis. The first pool consisted of six libraries done with one-step index-PCR. The second pool consisted of six libraries done with two-step index-PCR. All samples belonged to SG3. Both methods and each individual library had different index pairs: one-step libraries SD501–506 + SD705 and two-step libraries SD501–506 + SD708. All libraries

were purified as described in Chapter 4.5.3. The one-step libraries were pooled to a 20 µM library pool and the two-step libraries into a 30 µM library pool.

4.6 Troubleshooting the Library Purification Protocol

4.6.1 Library Purification Test 1

In the first optimization test for library purification three factors were changed from the current protocol described in detail in Chapter 4.5.3. Only one factor was changed at a time. The four tested methods are listed in Table 5. Only the changes to the current protocol are included in the table (indicated with a green background), all other steps were done following the current protocol. Method 1 is the current purification protocol.

Table 5. Different methods tested in Library purification test 1. Green background refers to deviations from the current protocol (method 1).

Method	**1**	**2**	**3**	**4**
Bead	Bead 1	Bead 1	Bead 1	Bead 2
Elution buffer T	RT	55 °C	RT	RT
Drying time	15 min	15 min	5 min	15 min

Two pools were used in the test. One pool was made from 2x8 libraries prepared with one-step index-PCR using exceptionally the same i5 and i7 index pair (SD506 and SD507) in each sample. This was done to avoid unexpected reactions between any unpaired index primers when pooling the unpurified libraries together, and to make the library pool as homogenous as possible. All samples belonged to SG2. All libraries were run on an agarose gel as described in Chapter 4.5.3 before pooling to check if the PCR was successful and to estimate the amount of unwanted fragments present. All libraries were then

combined to one pool, and equal volumes (17 µl) of the pool were distributed to four wells for each method (Figure 3, indexed library pool).

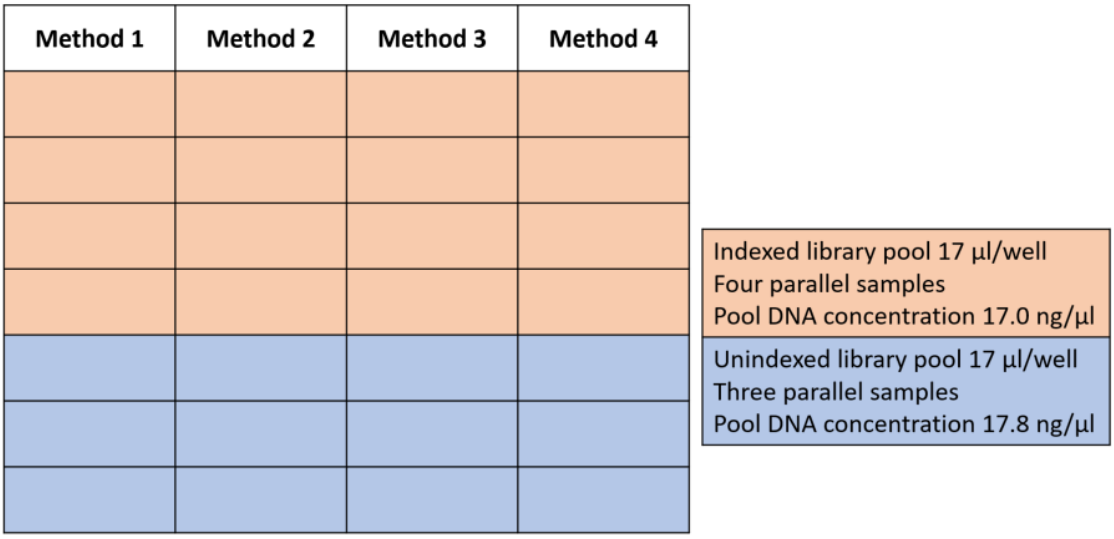

Figure 3. Parallel samples in Library purification test 1.

To reduce costs, another three parallel samples per method were taken from previously made libraries (2x8) prepared as described in Chapter 4.5.2 in the first step of the two-step library preparation. The libraries were also pooled together and 17 µl of the pool was distributed to three wells per method (Figure 3, unindexed library pool). Concentrations of the unpurified pools and all the purified parallel samples were measured with Quant-iT using the corrected dilution factor.

4.6.2 Library Purification Test 2

To exclude human-based error from the execution of the purification protocol, 2x8 libraries were made with one-step index-PCR, using the same index pair in each library (SD505 + SD705). All samples belonged to SG2. The libraries were run on agarose gel to check that the PCR was successful. The libraries were pooled together and divided to 2x8 parallel samples (16 µl/well). Two laboratory workers purified a set of eight parallel samples using the current protocol

described in Chapter 4.5.3. The concentration of the unpurified pool and the purified parallel samples were measured with Quant-iT.

4.6.3 Library Purification Test 3

Finally, to study the reason behind the loss of DNA during the library purification, 5x8 libraries were prepared with one-step index-PCR. All samples belonged to SG2. All libraries were prepared using the same index primer pair (SD505 + SD704) to make the library pool as homogenous as possible. The libraries were run on agarose gel to check that the PCR was successful and then pooled together. Four parallel samples were taken from the pool for each method. The volume of parallel samples was 20 µl in all other methods except Method 4, where 40 µl was used. Concentrations of the unpurified pool and the purified parallel samples were measured with Quant-iT.

The third purification test was performed with six different methods. The numbering of the methods was continued from the first purification test and method 1 follows the current purification protocol. The methods are listed in Table 6 and the deviations from method 1 are indicated with blue background.

Table 6. Different methods tested in Library purification test 3. Deviations from current protocol (method 1) are indicated with blue background.

Method	1 (current)	4	5	6	7	8
PCR product volume (µl)	20	40	20	20	20	20
Bead:Sample ratio (volume)	0.6:1	0.6:1	0.6:1	0.8:1	1.8:1	0.6:1
Elution volume (µl)	30	30	60	30	30	30
Eluted with	Elution buffer	Elution buffer	Elution buffer	Elution buffer	Elution buffer	TE buffer

Only one factor was changed in each method and only the changes are listed, all other steps were done following the current protocol.

The samples with the highest concentrations after purification from methods 1, 6, 7 and 8 were sent to a sequencing core facility to be analysed with an automated electrophoresis method called TapeStation (Agilent Technologies, California, United States) to check for the quality of the pooled libraries. TapeStation system allows for dsDNA concentration measurements as well as measuring the length and length distribution of DNA fragments (Agilent Technologies 2018).

5 Results and Discussion

5.1 DNA Concentration Measurements

NanoDrop, Qubit and Quant-iT for DNA concentration measurement were compared to get an estimate of how the results compare and if some methods should be preferred over the others. Results from these three different methods are shown in Table 7.

Qubit and Quant-iT are based on the same fluorometric principle detecting dsDNA. The difference between sample concentrations measured with Qubit and Quant-iT was always two-fold. Therefore, a systematic error was suspected. Both Qubit and Quant-iT measurements are susceptible for human error. The first possible source for error comes from preparing the standards and samples for measurements. With Quant-iT there is another possibility for error when the results are counted from the raw data. With Qubit the calculations are made by the algorithm but it is unclear what the calculations are based on.

Table 7. DNA concentration measurement results.

Sample	Qubit	Quant-iT		Quant-iT (corr.)		NanoDrop	
	DNA conc. (ng/µl)	DNA conc. (ng/µl)	Fold change from Qubit	DNA conc. (ng/µl)	Fold change from Qubit	DNA conc. (ng/µl)	Fold change from Qubit
S1	6.2	13.7	2.2	6.9	1.1	24.5	4.0
S2	48.6	104.5	2.2	52.2	1.1	174.8	3.6
S3	31.3	63.4	2.0	31.7	1.0	142.2	4.5
S4	6.0	12.8	2.1	6.4	1.1	15.9	2.6
S5	17.7	36.4	2.1	18.2	1.0	54.7	3.1
S6	12.7	27.7	2.2	13.9	1.1	60.7	4.8
S7	33.8	71.1	2.1	35.5	1.1	152.9	4.5
S8	29.3	60.1	2.1	30.0	1.0	109.9	3.8
Average fold change between Qubit and Quant-iT						2.1	
Average fold change between Qubit and Quant-iT with corrected dilution factor						1.1	
Average fold change between Qubit and NanoDrop						3.9	

The Excel template used when calculating concentrations from Quant-iT raw data was examined thoroughly. A mistake was noticed in the formula counting the original concentration of the samples from the dilutions prepared for the fluorometric measurement. The laboratory's Quant-iT protocol had been updated in June 2021 so that 2 µl of sample was always pipetted to 198 µl of other reagents. Previously 1 µl of sample and 199 µl of other reagents was used. The change was made to reduce variation caused by pipetting small volumes. The new dilution factor is 2:200 or 1:100 compared to the old protocol's 1:200. The change in the protocol was not transferred to the formula in the Excel template and the results were still multiplied by 200 to count the original concentrations of samples giving too high results.

This dilution factor error explained the systematic difference between the Qubit and Quant-iT results. When Quant-iT results were re-calculated with the 1:100 dilution factor (marked as Quant-iT (corr.) in the Table 7 above), the results were better in line with the values measured with Qubit. According to the corrected values Qubit gave slightly but systematically lower concentrations, on average 95 % of the values measured with Quant-iT. There was a statistically significant difference with the results of Qubit and Quant-iT at a 95 % confidence level (two-tailed paired Student's t-test, p-value 0.020). Both measurements should be repeated several times to eliminate stochastic effects and to get a more realistic evaluation of the actual difference. Further comparisons were not considered important in the scope of this study. The corrected dilution factor was used in all the DNA concentration measurements made during this study.

The dilution factor error found from the Excel template used for Quant-iT has likely been one reason behind the problems related to library preparation and a factor causing low library concentrations. In the laboratory, the dilutions for PCR are usually based on the Quant-iT measurements. Using the twice too high concentrations when preparing dilutions for the one-step PCR have resulted in dilutions closer to 0.5 ng/µl instead of the intended 1 ng/µl. A low DNA template concentration is susceptible to pipetting errors and reduced representativeness of the low abundant taxa within the microbial communities and might have been too low for the PCR to function properly. The dilution factor error would also explain why using the dilutions thought to be 5 ng/µl in recent PCRs had improved the PCR results. A low amount of template is also a factor causing excessive primer and adapter dimers in the sequencing libraries (Illumina 2023). Based on several agarose gels of previous libraries made in the laboratory, this has been a problem in the laboratory, leading to decreased quantity and quality of the NGS output.

After noticing the dilution factor error, it became problematic to analyse and compare data with older index-PCRs prepared in the laboratory, because it was unclear when the correct dilution factor had been used and when not. The

personnel in the laboratory have changed several times while the same protocols have been used. Based on some of the original raw data and calculations found, some of them had been correcting the dilution factor according to volume of sample being used and some not. For these reasons comparisons with older PCR results were not made during this study.

Results from NanoDrop stood out from both Qubit and corrected Quant-iT results. NanoDrop gave on average 370 % higher results than Qubit. There was a statistically significant difference with the results between NanoDrop and Qubit at a 95 % confidence level (two-tailed paired Student's t-test, p-value 0.0042). There was a linear correlation between the Qubit and NanoDrop results as seen in Figure 4.

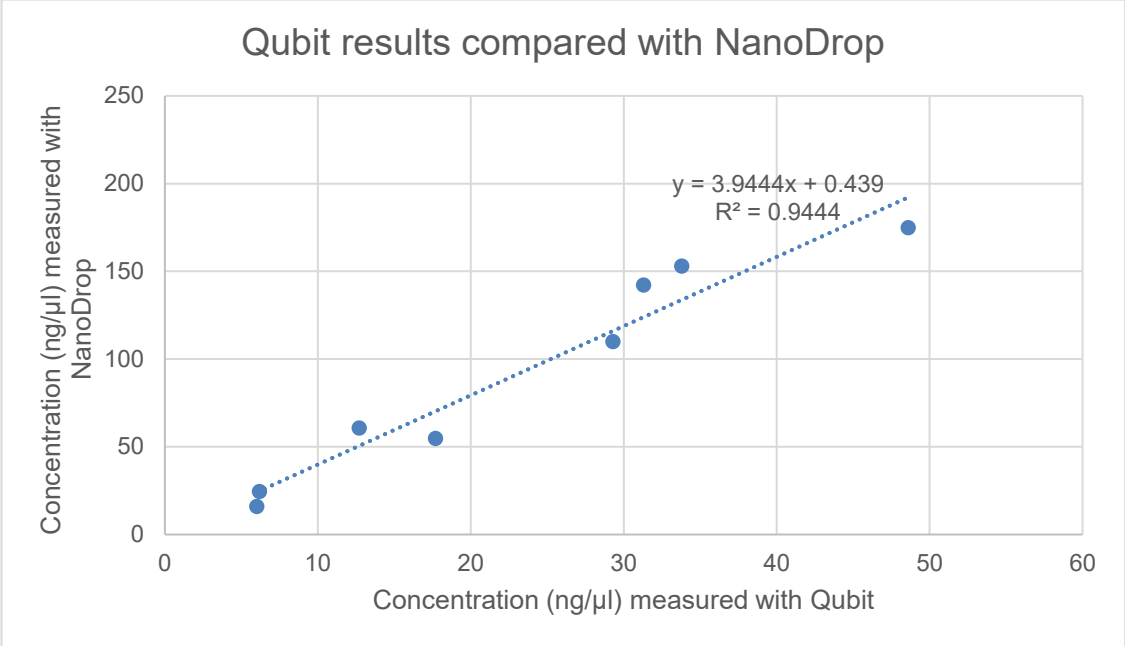

Figure 4. Comparison of DNA concentration measured with Qubit and NanoDrop.

NanoDrop is a spectrophotometric method, which measures all absorbance from the sample, not only dsDNA. The measured samples probably contain impurities which absorb in the same wavelengths as DNA. The distraction comes from the sample material itself because the elution buffer was used as a blank and its possible effect on the absorbance is already deducted from the

results. The A260/A280-ratio of the samples was on average 1.9 (range: 1.5–2.1) while a ratio of 1.8 is considered pure for nucleic acids, indicating a fairly good quality DNA. (Desjardins and Concklin 2010.)

Based on this experiment using NanoDrop is not recommended when measuring DNA concentration at least from samples that are extracted from complex matrixes potentially interfering with the absorbance measurement until the reason behind the anomaly compared to fluorometric results is clarified.

5.2 Index-PCR and Comparing One-Step and Two-Step Library Preparation

The first attempts to prepare libraries with the one-step index-PCR protocol in the context of this study were not successful. Visual inspection of the libraries and a positive control on agarose gel showed very weak bands from the target amplicon (~640 bp) and stronger bands from smaller products (< 250 bp) (Figure 5).

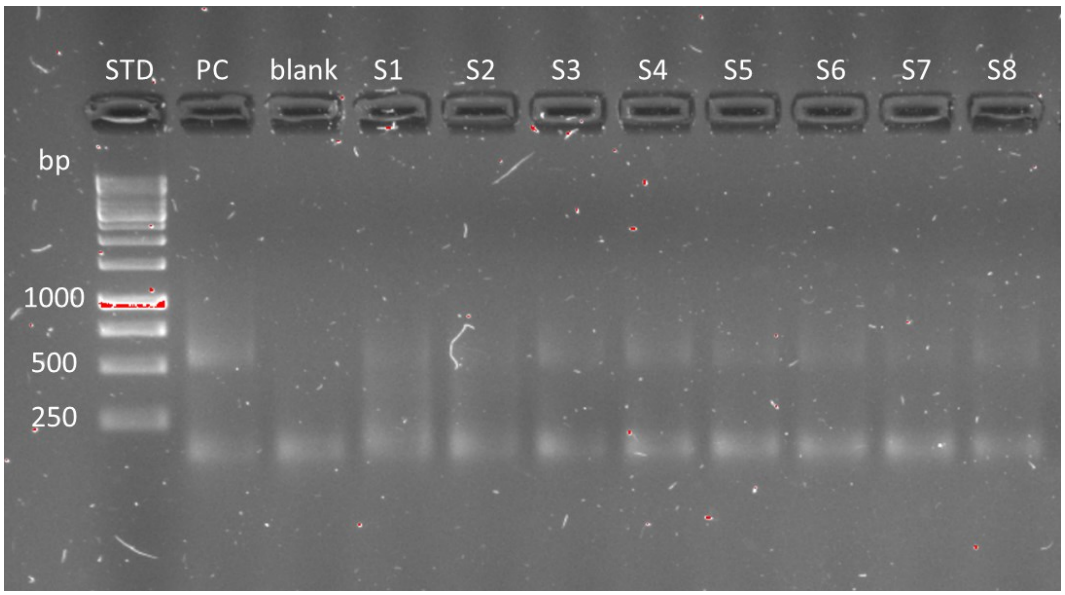

Figure 5. Unsuccessful one-step index-PCR run on agarose gel. S1–S8 = libraries 1 to 8, pc=positive control.

The dsDNA concentrations measured from blank samples were high, around 15 ng/µl. The blank samples had no visible bands from the target amplicon, but the bands from smaller products were present. Based on the gel electrophoresis results the DNA concentration measured from the blank sample originated from smaller molecules and not from a PCR contamination.

The smaller products contained dsDNA since they were also picked up by the fluorometric measurements specific for dsDNA. They were most likely primers that had dimerized with themselves or other primers. Another option is that they originate from the dsDNA template which has been broken into small fragments.

This observation provided a problem for analysis because it became obvious that DNA concentration could not be used directly to evaluate and compare the success of the one-step and two-step PCR protocols as the concentration of the target amplicon could not be separated from the unspecific PCR products. Even after removing the smaller fragments with purification, DNA concentration-based evaluation of the libraries after the purification was not considered reliable after the results of the library purification tests. A decision was made to visually evaluate the success of the PCRs and the quality of the libraries from agarose gels. Numerical comparisons based on concentrations between the one-step and two-step protocols were abandoned.

Based on the excessive amount of possible primer-dimers seen on the agarose gel a problem with the PCR preparation was considered likely. A study by Chou et al. (1992) done with Taq polymerase showed that keeping the mixed reagents at room temperature before PCR even for a few minutes led to considerable mispriming and in an increase of primer-dimers. The unwanted reactions started happening even before the thermal cycler and the quantity of primer artifacts correlated inversely with the quantity of the target product.

In the experiments of this study Phusion High-Fidelity DNA polymerase was used instead of Taq. To test the effect observed in Chou et al. (1992), the reactions were kept as cold as possible during PCR preparation. This improved

the quality of the libraries substantially, yielding stronger bands from the target amplicon on gel and reducing the amount of small fragments (Figure 6). The DNA concentration of the blank sample was 0.8 ng/µl and showed only a very weak band from small fragments. Based on these observations this practice was adopted to every PCR preparation, greatly reducing the amount of non-targeted fragments.

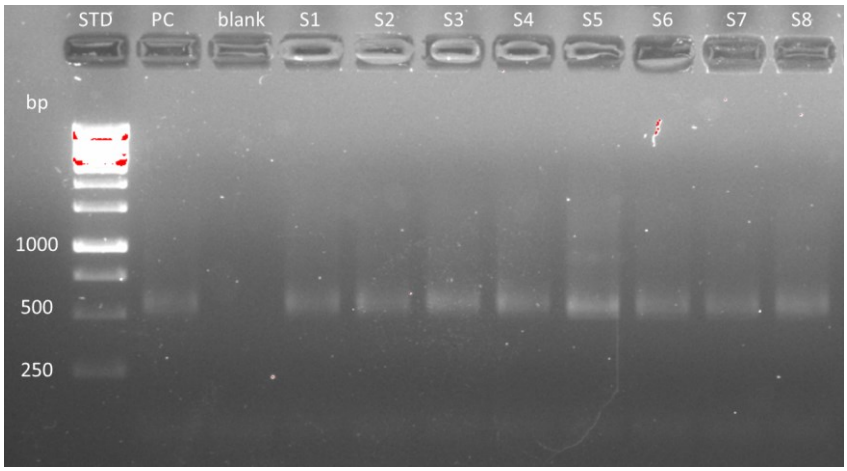

Figure 6. One-step index-PCR libraries on agarose gel. S1–S8 = libraries 1 to 8, pc=positive control.

The first step of the two-step protocol, 16S rRNA gene amplicon PCR, was considered successful based on the visual inspection of the libraries on agarose gel (Figure 7).

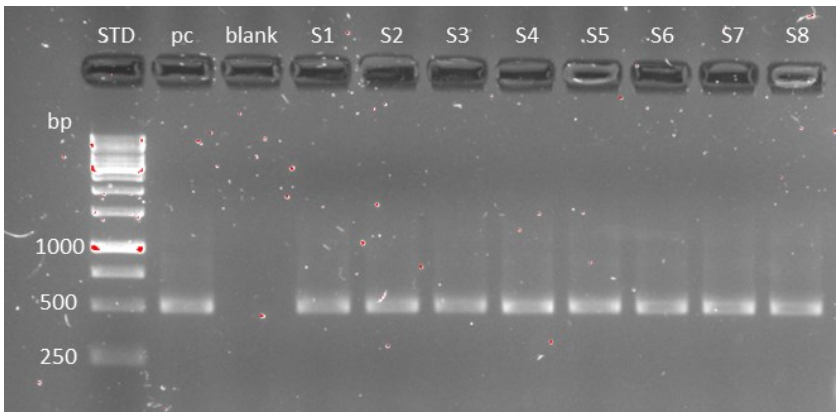

Figure 7. 16S rRNA gene PCR libraries after the first PCR of the two-step protocol on agarose gel. S1–S8 = libraries 1 to 8, pc=positive control.

There were no clear bands from untargeted fragments and the blank sample had no bands. The libraries were purified and diluted and used as a template for the second step.

The indexed libraries from the second step, after two PCR reactions, showed right sized products on the gel, but also bands from longer untargeted fragments (Figure 8). Target amplicon bands from samples 4 and 8 were weaker than the others. This might result from a problem in the PCR or from a pipetting error when pipetting the samples to the gel. There were no clearly visible bands from the shorter untargeted fragments in the samples or the positive control, only in the blank, indicating the presence of short fragments in the scarcity of template DNA. There is also a weak band in the blank corresponding to the targeted amplicon size. After the second step all samples have gone through over 50 PCR cycles and even small amounts of contaminants have been amplified.

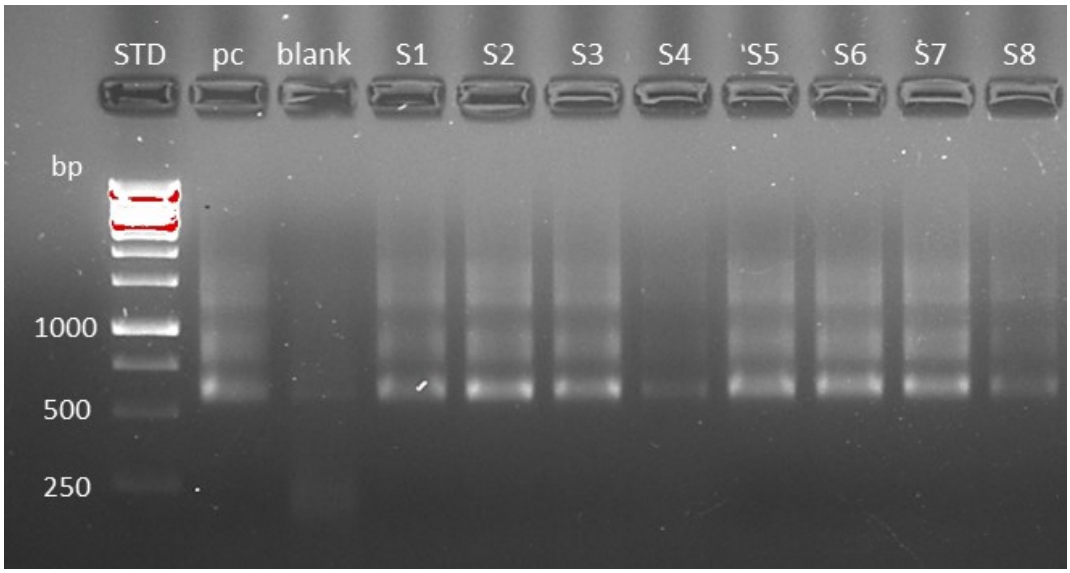

Figure 8. Index-PCR libraries prepared with two-step protocol on agarose gel. S1–S8 = libraries 1 to 8, pc=positive control.

The longer unwanted fragments are not removed in the current library purification method, which targets only small DNA fragments. Small fragments with P5 and P7 tails are problematic in NGS runs, because they attach to the flow cell surface more effectively than the longer target amplicons. Longer

untargeted fragments are not as effective in attaching to the flow cell, but any untargeted products with Illumina-compatible tails can still affect the results of the whole NGS run by competing with the target amplicons and in the worst scenario cause the run to abort (Illumina 2024). The more sequencing capacity the untargeted products take, the less space there is available for the target amplicons.

Based on visual inspection the two-step index-PCR seemed to help with the problem of small untargeted PCR fragments but keeping the reagents very cold during the PCR preparation had already reduced the same problem with the one-step method. The second step brought a new problem with the long untargeted fragments. The PCR program could be optimised to remove the problem, but performing the two-step protocol takes considerably more time, reagents, and consumables than the one-step protocol and is therefore not the preferred method. Based on the results, switching to the two-step library preparation method would not result in a major improvement in the library quality to make the switch from one-step method profitable. Optimising the two-step protocol might be reconsidered if the one-step method continues to be unreliable.

Based on the excessive amount of possible primer-dimers, low library concentrations noticed in the laboratory and quality issues with the NGS results, it was suspected that the index primers containing the P5 and P7 tails were not attaching properly. According to the KAPA kit analysis the mean percentage of correctly attached P5 and P7 tails was only 11.40 % for the one-step library pool and 20.20 % for the two-step library pool. Conclusions should not be drawn based on individual tests, but in this experiment the two-step method had functioned better.

The most significant finding here is the very low percentage of amplicons with properly attached tails indicating a problem in the library preparation. Another researcher outside the research group using the same two-step protocol and same reagents reported an attachment rate of nearly 100 %. Two libraries from

the one-step library pool were also analysed separately to investigate variation between individual libraries. The percentages of correctly attached tails were 8.22 % and 5.30 %. It seems that there was some variation in the attachment rates of the P5 and P7 -tails between the different libraries, but not some libraries performing exceptionally better than the others.

The low attachment rate of the tails undoubtedly affects the quality of the NGS runs as only a small fraction of the library pool gets sequenced. No data would be received from 80–90 % of the library pools tested here, since they would not attach to the flow cells. The reason behind the low attachment rate is not known, and it is uncertain how it affects the data. It would be important to know whether this can cause a taxonomic bias or if the phenomenon is random. If the tails are attaching better to specific DNA templates in a sample, this may lead to these templates being overrepresented. Variation in the attachment rate between different libraries might lead to differences in the number of reads received from some of the libraries.

5.3 Troubleshooting the Library Purification Protocol

In all library purification tests the libraries were inspected visually on agarose gel before the pools for the purification tests were made. In each inspection there was only a very weak band from the small untargeted fragments and the DNA concentration of the blank samples were on average 1.1 ng/µl. There was always a strong band from a product of the targeted size. Based on these observations it was assumed that dsDNA concentration of the pool came mainly from the target amplicon and a smaller concentration after the purification process meant a loss of the target amplicon.

5.3.1 Library Purification Test 1

The laboratory had changed to beads from a new manufacturer before this study. The performance of both bead brands was compared with methods 1 and 4. Bead 1 is the currently used brand. In the purification protocol provided

by the manufacturer of bead 1 it was mentioned that warming up the elution buffer to 55 °C might improve the yield. This was chosen as one of the factors to be tested in method 2. Normally elution buffer is at room temperature. Both bead manufacturers warned against long drying times after the ethanol wash to prevent the beads from drying out completely. If the beads became too dry, some of the DNA would stick to the beads and would not eluate in the final steps of the process. A shorter drying time was tested in method 3. The results from the first purification test are listed in Table 8.

Table 8. Results from Library purification test 1.

Method	**1 (current)**	**2**	**3**	**4**
Bead	Bead 1	Bead 1	Bead 1	Bead 2
Elution buffer	RT	55 °C	RT	RT
Drying time	15 min	15 min	5 min	15 min
Results				
Indexed libraries				
Conc. before purification (ng/µl)	17			
Average conc. after (ng/µl)	7.9	5.8	7.3	6.9
Std. deviation	0.53	1.37	1.40	0.73
Recovery%	46.3	33.9	42.9	40.4
Unindexed libraries				
Conc. before purification (ng/µl)	17.8			
Average conc. after (ng/µl)	4.7	3.8	5.8	4.2
Std. deviation	2.10	0.23	0.74	2.58
Recovery%	26.3	21.4	32.7	23.4

Library concentrations before purification had not been measured in the laboratory before. Therefore, there was no prior numeric data available of the usual recovery percentages. The recovery percentages in this test were considered low, especially with the unindexed libraries. Beckman Coulter Life Sciences has published an application note where they show results of recovery percentages from different bead brands. Almost all the beads have a recovery percentage of over 70 % and Beckman Coulter Life Sciences recommend a number between 80–95 % to be optimal for downstream processes. (Beckman Coulter Life Sciences 2020.) These numbers should be looked at with some caution since the application note is an advertisement for Beckman Coulter Life Science's own product, but it might give some indication of an expected recovery rate.

There was also considerable variation between the wells within the same method, as seen in Table 9 especially for methods 1 and 4, using the unindexed rRNA gene amplicon libraries as an example.

Table 9. The DNA concentration of the unindexed 16S rRNA gene amplicon libraries in individual wells E to G in Library purification test 1.

	Method1	**Method2**	**Method3**	**Method4**
Well ID on 96 well plate (number+letter)	**1**	**2**	**3**	**4**
E	4.7 ng/µl	3.8 ng/µl	6.6 ng/µl	1.9 ng/µl
F	2.5 ng/µl	3.6 ng/µl	5.6 ng/µl	7.0 ng/µl
G	6.8 ng/µl	4.0 ng/µl	5.2 ng/µl	3.7 ng/µl

Since all wells had parallel samples from the same library pool, they all had the same volume and concentration before the purification. Variation between the wells of the same method were caused by something unrelated to the samples. The most likely causes are related to pipetting and how homogenous the magnetic bead solution is. Since the library volumes are small, even small

changes in the amount of magnetic beads in the well can change the attachment rate of DNA. Accidentally disturbing the beads with a pipette tip can also cause differences in the recovery percentage from individual wells. Some amount of variation is to be expected with the bead purification method done by hand due to pipetting and the uneven distribution of magnetic force to the beads (Klose 2016:8–9.) The differences might also come from the magnets in the magnetic rack. The rack used in this experiment has magnets attaching to four wells each except the ones on the edges, which attach to two wells. A malfunctioning magnet would therefore affect groups of 2 or 4 samples similarly. No such patterns were noted in this test or in the following tests.

The experiment design had several problems. Using both indexed and unindexed 16S rRNA gene amplicon libraries was not a good choice. It was thought that the libraries would behave similarly during the purification providing more parallel samples for each method without having to make new libraries just for the test. There was a statistically significant difference with the indexed and unindexed libraries at a 95 % confidence level when the recovery percentages (two-tailed paired Student’s t-test, p-value 0.0067). The difference between the performance of the libraries during the purification might originate from the size of the product or differences related to the number of PCR cycles. The bead to sample -ratio was likely slightly more favourable for the longer indexed libraries. The observed level of variation between the wells would likely lead to misinterpretations of the differences between the methods.

The results from Library purification test 1 were not analysed further and the test should be repeated with more parallel samples to get comparable results. The results were used to understand that there is a notable loss of DNA during the purification process and variation between wells. Based on these results the rest of the purification tests were done using only indexed 16S rRNA gene amplicon libraries and more parallel samples per method to account for the variation. After Library purification test 1 the focus switched from optimizing the purification protocol to studying where the loss of DNA during the purification originates from. This was considered important to understand, since the library

concentrations are measured in the laboratory after purification and the challenge of low library concentrations could partly be caused by the purification protocol.

5.3.2 Library Purification Test 2

After the first library purification test, a human error while performing the protocol was suspected. This was tested in purification test 2. The recovery percentages from the eight parallel samples purified with the current protocol by two different laboratory workers are shown in Figure 9.

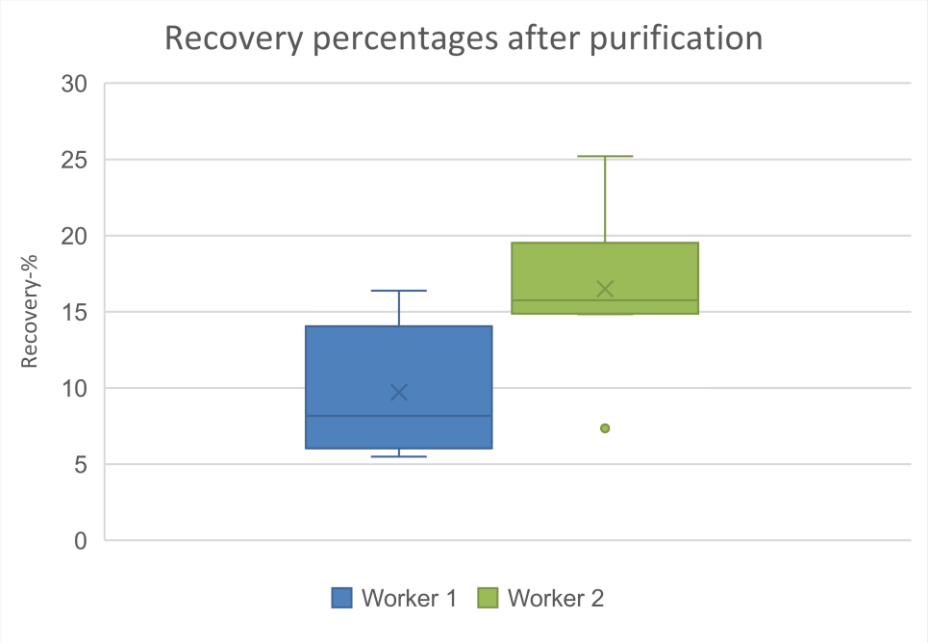

Figure 9. Recovery percentages after purification in Library purification test 2.

There was a statistically significant difference between the recovery percentages of the laboratory workers at a 95 % confidence level (two-tailed homoscedastic Student's t-test, p-value 0.012, variance inspected with F-test, p-value 0.66). There is a possible outlier (7.4 %) in the results of worker 2, which was included in the statistical analysis. Overall, the recovery percent of worker 2 was slightly better.

The observed variation between the workers here might come from different ways of performing the same protocol. One worker shakes the ethanol from the wells after the washes by inverting the plate over sink and the other uses a multichannel pipette. Different pipettes are also used: one worker uses manual multichannel pipettes and the other electrical. The protocol does not specify these details and the laboratory personnel use the methods they prefer. Based on these results, unifying the methods should be considered.

The most important finding from the comparison between the workers is that the low recovery percentage seen in Purification test 1 is not caused by a single worker but is a general phenomenon and troubleshooting the low recovery rates should be continued. Although there was a difference between the results from the workers, both recovery percentages were still considered so low that the difference was not meaningful considering the challenges related to the preparation of the libraries. The average recovery percentage in this test was even lower than in the previous test, but this might be due to coincidence because of the low number of parallel samples used in the first test leading to unreliable results.

5.3.3 Library Purification Test 3

Library purification test 3 focused on finding the cause behind the loss of DNA during the purification. During the first and second purification tests it was noticed that the surface of the sample did not reach the point where the magnet physically touches the side of the well when the plate is first placed on the magnetic rack or when the DNA is eluted from the beads after the ethanol washes. Most beads were still drawn out from the liquid by the magnets but by visual inspection some were left in. Whether this had any effect on the recovery percentage was tested both in the beginning of the process by using a higher volume of the PCR product and in the end by using a higher elution buffer volume (methods 4 and 5).

A bead to sample volume ratio of 0.8:1 had previously been used in the laboratory, but it had been changed to a 0.6:1 ratio. The latter had been chosen from a plot made by the manufacturer of bead 2 showing the best selection performance for the target amplicon size. The previous ratio and a clearly higher ratio given by the manufacturer of bead 1 were tested in methods 6 and 7. To make sure there was not a problem with the elution buffer used, TE buffer was tested in Method 8. Both buffers were 10 mM Tris-HCl and pH 8, but elution buffer is 0.5 mM EDTA and TE buffer 1.0 mM EDTA.

The recovery percentages from Library purification test 3 are shown in Figure 10.

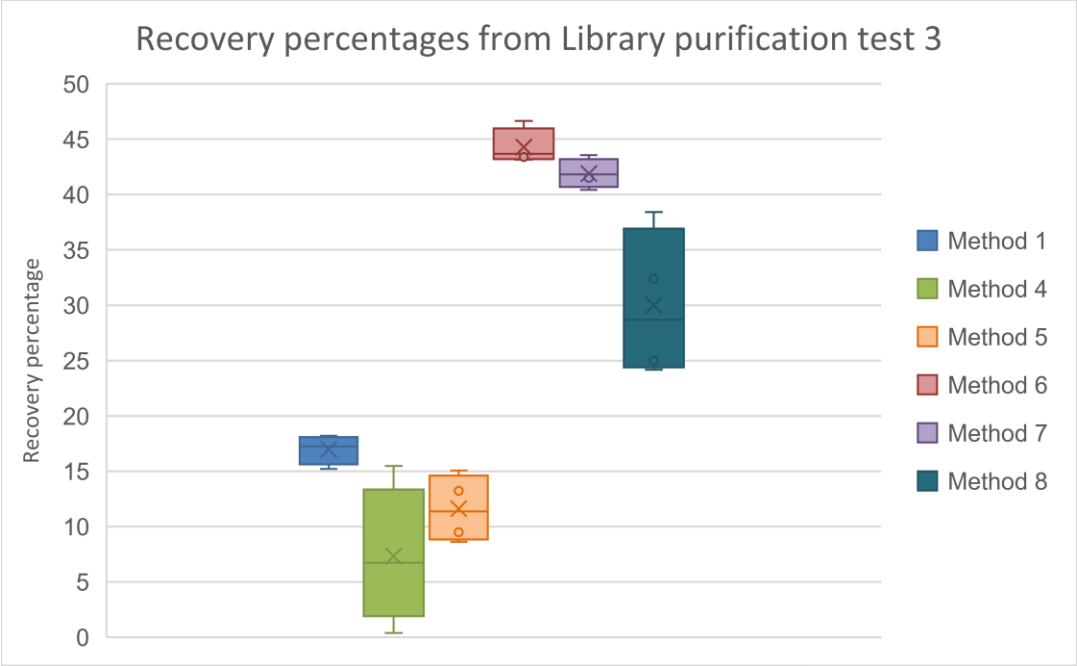

Figure 10. Recovery percentages from Library purification test 3.

Increasing the volume of the PCR product to 40 µl (method 4), increasing the elution volume to 60 µl (method 5) and using TE-buffer as the elution buffer (method 8) seemed to result in greater variability in the recovery percentages between the parallel samples than in the other methods. Raising the volume of the PCR product while keeping the bead to sample -ratio the same increases the number of beads in the solution. If there are too many beads, the outmost

beads will experience weaker magnetic force than the ones closest to the magnet when bound to the magnet. The outmost beads are more likely to detach from the magnet during the washing steps and the DNA bound to them is lost. This is especially true for the type of magnetic rack used in the laboratory, where the beads bind to one spot on the well wall. This might explain the greater variation in the recovery rates between parallel samples with method 4. (Klose 2016:8–9.) It is unclear where the increase in variation between wells comes from when using higher elution volume (method 5) or TE-buffer (method 8). Besides causing more variation, increasing the volume of the PCR product or the elution buffer did not help in achieving a higher recovery percentage. The phenomenon of the liquid surface not reaching the point where the magnet touches the well is not likely to influence the recovery percentage.

Using TE-buffer to elute the DNA from the beads increased the recovery percentage. DNA is eluted from the beads by lowering the salt concentration. The only difference between the TE-buffer and the elution buffer is a higher EDTA concentration. Water can also be used for eluting, so the DNA should not need a stronger buffer to elute from the beads (Hawkins 1998; Wang et al. 2021). Higher concentrations of EDTA may also be problematic in downstream applications. Water was not tested here but it would provide an interesting comparison with the current elution buffer.

The best recovery percentages were achieved with methods 6 and 7, where the bead to sample -ratio was higher than in the current protocol (0.8:1 and 1.8:1 respectively). There was a statistically significant difference in the 95 % confidence level between the recovery percentage of the current method and method 6 that gave the best results (two-tailed homoscedastic Student’s t-test, p-value 0.00000020; variance inspected with F-test, p-value 0.74). The manufacturer of bead 1 was contacted and the results from Library purification test 3 and the purification protocol were provided to them. They confirmed that the purification workflow is correct but recommended that a 1:1 bead to sample -ratio should be used with an amplicon of this size. The manufacturer’s protocol mentions only one bead to sample -ratio (1.8:1) and they ask to contact them

for a tailored ratio according to what is needed. The 1:1-ratio should be tested next.

Beckman Coulter Life Sciences has an application note for their own bead brand where they show figures from tests made with beads from different manufacturers. The recovery percentages of some bead brands seem to strongly depend on the bead to sample -ratio. The most favoured fragment size with each ratio is also a changing factor between the different brands. (Beckman Coulter Life Sciences 2020.) The manufacturer of bead 1 does not give an easily available expected recovery rate for DNA when using their products nor did they provide one when contacted. One of the problems with the current purification protocol might be that the bead to sample -ratio was copied from the instructions of the manufacturer of the previously used beads but the performance of the beads of different brands may not be equal in all circumstances. However, it should be noted that the comparisons made by Beckman Coulter Life Sciences are a part of an advertisement for their own product, where they also warn against changing to other brands (Beckman Coulter Life Sciences 2020).

The laboratory had previously noticed that a higher bead to sample -ratio had left unwanted smaller fragments in the library pool even after purification. TapeStation analysis did not show big differences in the purification efficiency between the methods, but the unpurified pool did not contain a lot of small fragments to begin with (Figure 11).

However, the analysis revealed that most of the fragments in the libraries are shorter than expected. The final indexed 16S rRNA gene amplicon libraries should be approximately 640 bp long (Illumina n.d.a) but the highest peaks from the library pools are at approximately 500 bp (Figure 11), indicating that most of the fragments in the pools are of that size. There is some variation in the length of the V3–V4 region between different bacteria that originates mostly from the V3 region. The variation was found to be less than 50 bp and does not therefore explain completely the difference noted here. (Vargas-Albores et al. 2017.) The

length of the P5 and P7 tails missing from some of the amplicons is not enough to explain the difference either.

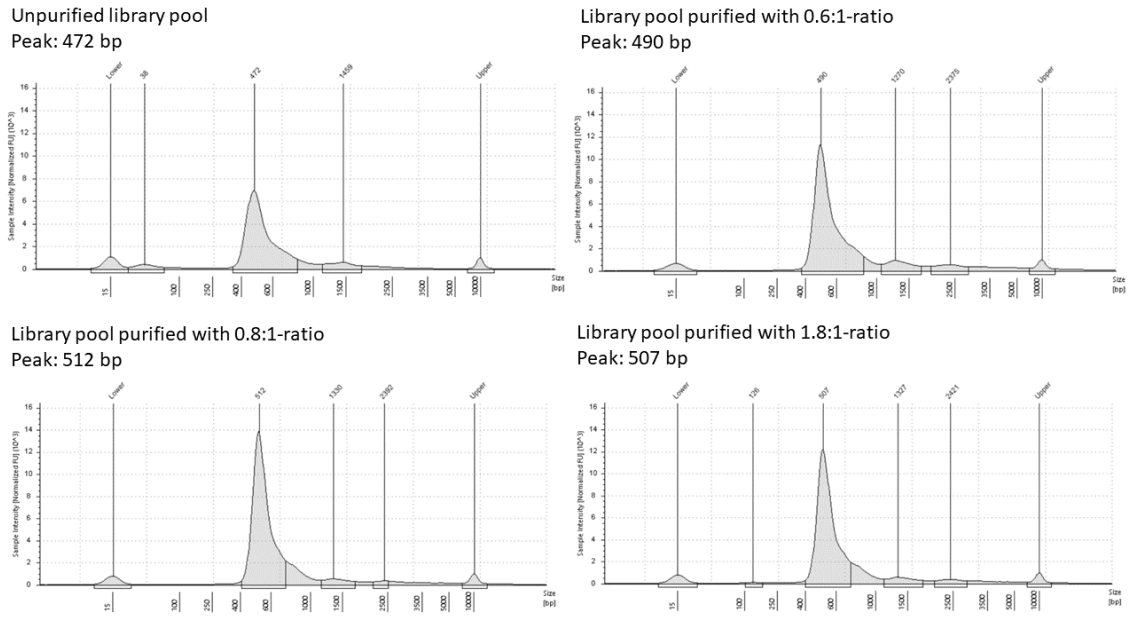

Figure 11. Results from the TapeStation analysis from libraries purified with different bead to sample -ratios.

PCR conditions can lead to unspecific priming and the formation of unspecific products. This could also explain the difference in the observed and the expected fragment size. (Bio-Rad n.d.) At least one set of the V3–V4 primer pairs has been observed to produce more unspecific amplification artefacts than primer pairs targeting other regions of 16S rRNA gene (Claesson et al. 2010). The PCR reaction and the formation of possible unspecific products should be studied further.

In addition to testing the 1:1 bead to sample -ratio the manufacturer recommended, further tests could be made to find out in which step of the purification the DNA is lost. The wells are normally eluted with 30 µl of elution buffer, but only 20 µl is transferred to a new plate. Three wells from the plate containing the rest of the elution buffer and the paramagnetic beads (stored at +4 °C overnight) were eluted again with the currently used elution buffer. The concentrations measured were the same as after the first elution after

accounting for the dilution factor. All the DNA came from the 10 µl of elution buffer left in the wells, not from the beads. Either DNA is permanently attached to the beads, or it is lost in one of the washing steps where liquid is discarded. This could be studied by measuring DNA concentration from the liquids after the beads are first drawn out from the solution to see if the beads are able to bind all the DNA, and from the ethanol discarded during the wash steps to see if the DNA does not stay attached to the beads.

6 Conclusions

Several parts of the NGS library preparation workflow used in the laboratory of the Microbes Inside -research group were experimentally studied during the thesis project. No single explaining factor for the varying quality of the NGS results and the low library concentrations was found. None of the findings made during this study completely explain why these challenges have arisen during the past year. The only factor dating close to that time period is the dilution factor error resulting in too low template concentrations in PCR. The lack of available templates in the library preparation PCR may have caused the low DNA concentrations in the libraries, and the excessive formation of primer-dimers observed in the laboratory. The poor recovery percentage of DNA during library purification poses further challenges, since the purification step is a compromise between an efficient removal of short interfering fragments and loss of the targeted PCR product.

The protocol for making indexed 16S rRNA gene amplicon libraries had previously been optimized for a reaction containing 1 ng/µl DNA. Increasing the amount of template to correct for the then unnoticed dilution factor error had helped to lower the amount of unwanted fragments, but the reaction should be optimized again for the higher concentration of template. However, it should be noted that raising the amount of DNA template extracted from faecal samples can result in problems with PCR inhibitors that are present in high concentrations in faecal samples (Monteiro et al. 1997). The protocol should be tested again using accurate dilutions to see if this improves the PCR results. If it

does not, re-optimising the PCR conditions should be considered to improve the concentrations of the libraries.

The most important observation made during the thesis project was the low attachment rate of the index primers during library preparation. More libraries should be analysed with the KAPA kit analysis to verify the results seen in the experiment of this study. It is important to explore this finding further since no data is received from an NGS run from amplicons without the Illumina compatible P5 and P7 -tails. The data loss can cause a bias in the results. The most likely cause for both the missing tails and the short amplicon size noticed in the TapeStation analysis is the PCR used in library preparation. The quality changes in the NGS results might originate from the anomalies noted here. The quality of the extracted DNA and the PCR conditions should be further examined.

It would be beneficial to have numerical data available from different steps of functioning workflows and different steps of a protocol. When challenges are noticed, it is easier to spot any changes and evaluate whether the anomaly is related to the current challenges or not. Previous information also helps in experiment design. When previous data is not available, pre-tests are needed to evaluate what kind of results are to be expected and plan the experiment accordingly. Unprocessed raw data from measurements was important during this study in understanding the background of the dilution factor error. Keeping raw data and calculations based on it for later inspection may also help to understand reasons behind challenges.

The personnel in a laboratory can change and ideally information should not be lost when changes happen. The protocols of the laboratory should be taught to new personnel explaining the meaning of each step or formula so errors can be spotted more easily. Insufficient time and the capacity to internalize new information might prevent detailed familiarization of protocols when new personnel start. Therefore, it is important that everything is written in the protocols in detail. Differences in methods not detailed in the protocols may

affect the results, as seen in comparing the results from two different workers in Library purification test 2. The library purification protocol might need to be written in more detail if human based variation would like to be removed from the results. When updating protocols, it is important that the older versions are removed from use. A good practice is to indicate what was changed and when. Since the reason behind the change might not be evident later, it is good practice to indicate why the changes were made to prevent loss of information.

Based on the results of this study, the next steps in the laboratory would be to verify the anomalies in the indexed 16S rRNA gene amplicon libraries noticed in the KAPA library quantification kit analysis and the TapeStation analysis. If the results are verified, both the low library concentrations and varying quality of the NGS results might originate from the PCR to make the libraries. The purification protocol should be tested with the bead to sample -ratio recommended by the manufacturer and the step where the DNA is lost should be detected.

References

Agilent Technologies (2018), ‘Performance characteristics of the D1000 and High Sensitivity D100 ScreenTape assays for the 4150 TapeStation System’, technical overview. https://www.agilent.com/en-us/library/technicaloverviews?N=130

Beckman Coulter Life Sciences (2020), ‘Impact of clean-up kits of DNA sequencing quality and efficiency’, accessed 17 November 2023. https://media.beckman.com/-/media/pdf-assets/application-notes/genomics-application-note-ampure-xp-ngs-cleanup.pdf

Beckman Coulter Life Sciences (n.d.), ‘Manual or automated DNA size selection’, accessed 19 November 2023. https://media.beckman.com/-/media/pdf-assets/data-sheets/genomics-data-sheet-spriselect-reagent-2023-03.pdf

Bio-Rad (n.d.), ‘PCR Troubleshooting’, accessed 2 February 2024. https://www.bio-rad.com/en-fi/applications-technologies/pcr-troubleshooting?ID=LUSO3HC4S#gel2

Chou Quin, Russel Marion, Birch David, Raymond Jonathan and Bloch Will (1992), ‘Prevention of pre-PCR mis-priming and primer dimers improves low-copy-number amplifications’, *Nucleic Acids Research,* 120(7). doi: 10.1093/nar/20.7.1717

Claesson Marcus, Wang Qiong, O’Sullivan Orla, Greene-Diniz Rachel, Cole James, Ross Paul and O’Toole Paul (2010), ‘Comparison of two next-generation sequencing technologies for resolving highly complex microbiota composition using tandem variable 16S rRNA gene regions’, *Nucleic Acids Research,* 38(22). doi: 10.1093/nar/gkq873

Clarridge Jill III (2004), ‘Impact of 16S rRNA gene sequence analysis for identification of bacteria on clinical microbiology and infectious diseases’, *Clinical Microbiology Reviews,* 17(4). doi: 10.1128/CMR.17.4.840-862.2004

Desjardins Philippe and Conklin Deborah (2010), ‘NanoDrop microvolume quantitation of nucleic acids’, *Journal of Visualized Experiment,s* 45. doi: 10.3791/2565

Fadrosh Douglas, Ma Bing, Gajer Pawel, Sengamalay Naomi, Ott Sandra, Brotman Rebecca and Ravel Jacques (2014), ‘An improved dual-indexing approach for multiplexed 16S rRNA gene sequencing on the Illumina MiSeq platform’, *Microbiome,* 2. doi: 10.1186/2049-2618-2-6

Hawkins Trevor (1998), ‘DNA purification and isolation using magnetic particles’, United States Patent 5705628.

Heather James and Chain Benjamin (2016), ‘The sequence of sequencers: The history of sequencing DNA’, *Genomics,* 107(1). doi: 10.1016/j.ygeno.2015.11.003

Illumina (2024), ‘How short inserts affect sequencing performance’, accessed 13 February 2024. https://knowledge.illumina.com/library-preparation/general/library-preparation-general-reference_material-list/000003874

Illumina (2023) ‘Adapter dimers causes, effects, and how to remove them’, Illumina Knowledge article #1911, accessed 5 November 2023. https://knowledge.illumina.com/library-preparation/general/library-preparation-general-troubleshooting-list/000001911

Illumina (2017), ‘An introduction to next-generation sequencing technology’, accessed 14 February 2024. https://www.illumina.com/science/technology/next-generation-sequencing.html

Illumina (n.d.a), ‘What is the PhiX Control v3 Library and what is its function in Illumina Next Generation Sequencing’, accessed 5 November 2023. https://knowledge.illumina.com/library-preparation/general/library-preparation-general-reference_material-list/000001545

Illumina (n.d.b), ‘What is nucleotide diversity and why is it important?’, accessed 5 November 2023. https://knowledge.illumina.com/instrumentation/general/instrumentation-general-reference_material-list/000001543

Illumina (n.d.c), ‘16S Metagenomic sequencing library preparation’, accessed 19 November 2023. https://support.illumina.com/documents/documentation/chemistry_documentation/16s/16s-metagenomic-library-prep-guide-15044223-b.pdf

Integrated DNA Technologies (n.d.), ‘Sample library preparation for next generation sequencing’, accessed 13 February 2024. https://www.idtdna.com/pages/technology/next-generation-sequencing/library-preparation

Jokela Roosa, Korpela Katri, Jian Ching, Dikareva Evgenia, Nikkonen Anne, Saisto Terhi, Skogberg Kirsi, de Vos Willem, Kolho Kaija-Leena and Salonen Anne (2022), ‘Quantitative insights into effects of intrapartum antibiotics and birth mode on infant gut microbiota in relation to well-being during the first year of life’, *Gut Microbes,* 14(1). doi: 10.1080/19490976.2022.2095775

KAPABiosystems (2020), ‘KAPA Library Quantification Kit’, technical data sheet, accessed 23 January 2024. https://rochesequencingstore.com/wp-content/uploads/2017/10/KAPA-Library-Quantification-Kit.pdf

Klindworth Ann, Pruesse Elmar, Schweer Timmy, Peplies Jörg, Quast Christian, Horn Matthias and Glöckner Frank (2021), ‘Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies’, *Nucleic Acids Research*, 14(1). doi: 10.1093/nar/gks808

Klose Alanna (2016), ‘The basic guide for magnetic DNA purification,’ Sepmag Systems. https://www.sepmag.eu/free-guide-magnetic-dna-purification

Kozich James, Westcott Sarah, Baxter Nielson and Schloss Patrick (2013), ‘Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform’, *Applied and Environmental Microbiology*, 79(17). doi: 10.1128/AEM.01043-13

Microbes Inside (n.d.), accessed 5 November 2023. https://www.helsinki.fi/en/researchgroups/microbes-inside

Monteiro Lurdes, Bonnemaison Dominique, Vekris Antoine, Petry Klaus G., Bonnet Jacques, Vidal Rui, Cabrita José and Mégraud Francis (1997), ‘Complex polysaccharides as PCR inhibitors in feces: Heliobacter pylori model’, *Journal of Clinical Microbiology*, 35(4). doi: 10.1128/jcm.35.4.995-998.1997

Oikarinen Sami, Tauriainen Sisko, Viskari Hanna, Simell Olli, Knip Mikael, Virtanen Suvi and Hyöty Heikki (2008), ‘PCR inhibition in stool samples in relation to age of infants’, *Journal of Clinical Virology*, 44. doi: 10.1016/j.jcv.2008.12.017

Raju Sajan, Lagström Sonja, Ellonen Pekka, de Vos Willem, Eriksson Johan, Weiderpass Elisabete and Rounge Trine (2018), ‘Reproducibility and repeatability of six high-throughput 16S rDNA sequencing protocols for microbiota profiling’, *Journal of Microbiological Methods*, 147. doi: 10.1016/j.mimet.2018.03.003

Sidstedt Maja, Rådström Peter and Hedman Johannes (2020), ‘PCR inhibition in qPCR, dPCR and MPS – mechanisms and solutions’, *Analytical and Bioanalytical Chemistry*, 412(9). doi: 10.1007/s00216-020-02490-2

ThermoFisher Scientific (2016), ‘Comparison of Quant-iT and Qubit DNA quantitation assays for accuracy and precision’, accessed 5 November 2023.

https://assets.thermofisher.com/TFS-Assets/LSG/Application-Notes/comparison-quantit-qubit-dna-quantitation-app-note.pdf

Vargas-Albores Francisco, Ortiz-Suárez Luis, Villalpando-Canchola Enrique and Martínez-Porchas Marcel (2017), 'Size-variable zone in V3 region of 16S rRNA', *RNA Biology* 14(11). doi: 10.1080/15476286.2017.1317912

Wang Xi, Zhao Ling, Wu Xiaoxing, Luo Huaxiu, Wu Di, Zhang Meng, Zhang Jing, Pakvasa Mikhail, Wagstaff William, He Fang, Mao Yukun, Zhang Yontao, Niu Changchun, Wu Meng, Zhao Xia, Wang Hao, Huang Linjuan, Shi Deyao, Liu Qing, Ni Na, Fu Kai, Hynes Kelly, Strelzow Jason, El Dafrawy Mostafa, He Tong-Chuan, Qi Hongbo and Zeng Zongyue (2021), 'Development of a simplified and inexpensive RNA depletion method for plasmid DNA purification using size selection magnetic beads (SSMBs)', *Genes & Diseases*, 8(3). doi: 0.1016/j.gendis.2020.04.013

Appendix 1. Primer Sequences

All primer sequences used in the experiments of this study.

Primer	Sequence 5’ – 3’
16S Truseq V3 Forward	ACACTCTTTCCCTACACGACGCTCTTCCGATCTCCTACGGGNGGCWGCAG
16S Truseq V4 Reverse	AGACGTGTGCTCTTCCGATCTGACTACHVGGGTATCTAATCC
SD501	AATGATACGGCGACCACCGAGATCTACACAAGCAGCAACACTCTTTCCCTACACGACGCTCTTCCGATC*T
SD502	AATGATACGGCGACCACCGAGATCTACACACGCGTGAACACTCTTTCCCTACACGACGCTCTTCCGATC*T
SD503	AATGATACGGCGACCACCGAGATCTACACCGATCTACACACTCTTTCCCTACACGACGCTCTTCCGATC*T
SD504	AATGATACGGCGACCACCGAGATCTACACTGCGTCACACACTCTTTCCCTACACGACGCTCTTCCGATC*T
SD505	AATGATACGGCGACCACCGAGATCTACACGTCTAGTGACACTCTTTCCCTACACGACGCTCTTCCGATC*T
SD506	AATGATACGGCGACCACCGAGATCTACACCTAGTATGACACTCTTTCCCTACACGACGCTCTTCCGATC*T
SD507	AATGATACGGCGACCACCGAGATCTACACGATAGCGTACACTCTTTCCCTACACGACGCTCTTCCGATC*T
SD508	AATGATACGGCGACCACCGAGATCTACACTCTACACTACACTCTTTCCCTACACGACGCTCTTCCGATC*T
SD704	CAAGCAGAAGACGGCATACGAGATCACGATAGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC*T
SD705	CAAGCAGAAGACGGCATACGAGATCGTATCGCGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC*T
SD706	CAAGCAGAAGACGGCATACGAGATCTGCGACTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC*T

SD707	CAAGCAGAAGACGGCATACGAGATGCTGTAACGTGAC TGGAGTTCAGACGTGTGCTCTTCCGATC*T

Appendix 2. The Workflows of One-Step and Two-Step Library Preparation Protocols.

The workflows of the one-step and the two-step library preparation protocols.