8

MOLECULAR METHODS

Authors:
Søren M. Karst
Mads Albertsen
Rasmus H. Kirkegaard
Morten S. Dueholm
Per H. Nielsen

Reviewer:
Holger Daims

8.1 INTRODUCTION

Molecular methods can be applied in the wastewater field for a fast, reliable and cheap identification of relevant microorganisms. In some cases, it is also possible to link the identification to a function, but surprises are frequently encountered. A recent example is that certain nitrifiers, which were believed to have a simple and well described physiology, are now known to be much more diverse than hitherto known (Daims *et al.,* 2015). The function concerns both their physiology (heterotrophs, nitrifiers, fermenters etc.) as well as their morphology (filamentous or single cells), which is important for their overall effect in the wastewater systems. Through knowledge of the identity and function of the microorganisms, it may be possible to manipulate their presence to optimize the plant performance, e.g. ensure the presence of nitrifiers or removal of foam-forming filamentous species.

The molecular identification of microorganisms is usually based on the 16S rRNA gene. However, in some cases the identification is done using sequencing of functional genes instead, such as those encoding the ammonium monooxygenase enzyme (AMO) for ammonia oxidizers (Rotthauwe *et al.,* 1997; Okano *et al.,* 2004), as these provide higher phylogenetic resolution, which might be useful for fine-scale studies.

The most common methods applied in the wastewater field for identification have been real-time quantitative PCR (qPCR), clone library generation and fingerprinting techniques such as denaturing gradient gel electrophoresis (DGGE; Muyzer, 1999) or terminal restriction fragment length polymorphism (T-RFLP) (Marsh, 1999; Marzorati *et al.,* 2008). However, these fingerprinting methods are hardly used anymore as they are often more difficult to use and provide less information compared to their high-throughput sequencing-based counterparts and their continued use cannot be recommended.

High-throughput sequencing can be applied for metagenomics or metatranscriptomics, where all the DNA or expressed genes (mRNA) from a certain community is sequenced. We do not, however, regard these methods as relevant for most readers of this book, as they require significant skills in molecular biology and bioinformatics.

Instead, high-throughput amplicon sequencing is recommended for routine analyses of microbial communities and will be described in greater detail. The method provides a list of microbes and an estimate of their relative abundance. One of the first sequencing platforms to be used for high-throughput amplicon sequencing was the Roche 454 (often termed 'pyrosequencing'). It is, however, now outdated (ultimo 2016) and the Illumina platform is presently dominating the amplicon sequencing market. By using the Illumina platform it is possible to analyse hundreds of samples in a fast, easy and cheap way compared to prior techniques.

The identification of microorganisms is usually done by comparison of the unknown sequences to a known reference set with a defined taxonomy. In this chapter we recommend the MiDAS database (midasfieldguide.org), which is a curated database that specifically targets microorganisms in the wastewater treatment field. Canonical or putative names for most common genus-level taxa are included and can be used as a common vocabulary for all researchers in the field to refer to the same organisms. The MiDAS database also provides all the available functional information about the 150 most abundant microorganisms encountered in Danish wastewater treatment plants (WWTP) and probably also worldwide (McIlroy *et al.,* 2015).

In this chapter we will focus on the methods of choice today and the next few years in wastewater microbiology: DNA extraction, qPCR and amplicon sequencing.

8.2 EXTRACTION OF DNA

8.2.1 General considerations

An optimized and standardized protocol for DNA extraction is essential to any analysis of microbial composition using DNA sequencing. This is due to the simple fact that microbes differ enormously in their resistance to different lysing methods (Thomas *et al.*, 2012; Guillén-Navarro *et al.*, 2015). Hence, microbes with cell walls that are difficult to lyse will effectively seem less abundant if sub-optimal extraction protocols are used (Bollet *et al.*, 1991; Filippidou *et al.*, 2015). Furthermore, activated sludge samples contain various chemicals that render some techniques unsuccessful due to inhibition (Guo and Zhang, 2013). Thus, the method for DNA extraction needs to be robust in order to cope with the challenges presented by activated sludge. However, despite much research effort into different DNA extraction protocols, it seems unlikely that there will ever be a perfect protocol. The biases introduced in DNA extraction can only be minimised not circumvented (Guo and Zhang, 2013; Albertsen *et al.*, 2015). The aim of this section is to give a brief overview of the steps involved in DNA extraction and to make general recommendations when working with activated sludge. In addition, a protocol optimized for the use in activated sludge is presented, based on the protocol developed by Albertsen *et al.* (2015).

8.2.2 Sampling

It is of key importance that the sample is representative of the activated sludge in the process tank of the plant or in a lab-scale reactor. For large-scale systems it is recommended to sample a larger volume (1 L) from a well-mixed tank, then perform homogenisation and finally sub-sample 3×2 mL aliquots that can be readily frozen and stored for years at -20 °C until analysed. Ideally, biological replicates are stored to ensure that sample variance can be analysed and that extra biomass is available in case something goes wrong. It is important to minimise the time from sampling to freezing as the changed conditions outside the original environment might favour the growth of some species over others, rendering the sample unsuitable for comparative analysis (Guo and Zhang, 2013). Sampling should preferably take place quite often, e.g. every week and the samples stored frozen in a 'bio-bank' for later use. As the number of samples grows fast, it is important to label each of them clearly and keep a log with sample IDs, reactor ID, dates, related chemical measurements, and any additional information that might be relevant for a later microbial analysis.

8.2.3 DNA extraction

DNA extraction involves a few general steps that are modified and combined in different ways in a range of commercial kits depending on the target organisms, type of environment, and the purpose for the extracted DNA. The common steps are disruption and cell lysis, protein removal, chemical removal and DNA elution (Figure 8.1).

8.2.3.1 Cell lysis

Various methods have been developed to lyse the cells in order to release their DNA. Some methods use chemicals to burst the cells, some use enzymatic degradation of cell structures, and others use physical stress such as freeze-thaw cycles, or mechanical stress such as ultra sound or bead beating (Bollet *et al.*, 1991; Tsai and Olson, 1991;

Zhou *et al.*, 1996). Off-the-shelf kits have been developed that use combinations of these strategies optimised for different cell and sample types. The difficult nature of activated sludge presents a challenge for several of these approaches due to different kinds of inhibition (Tullis and Rubin, 1980). However, mechanical lysis has proven very robust and does not suffer from inhibition effects (Salonen *et al.*, 2010; Guo and Zhang, 2013; Albertsen *et al.*, 2015). The lysis of cells is often carried out in solutions with detergents and surfactants that support the disruption and removal of cell membrane constituents such as lipids by a subsequent centrifugation step.

If any parameters are modified in the cell lysis step, it is important to test the effects on yield and integrity of the extracted DNA. Increasing the intensity or duration of some steps will likely increase the yield until some level of saturation. However, the increased duration and intensity might shear DNA apart and fewer samples can be handled in a reasonable time (Bollet *et al.*, 1991; Bürgmann *et al.*, 2001).

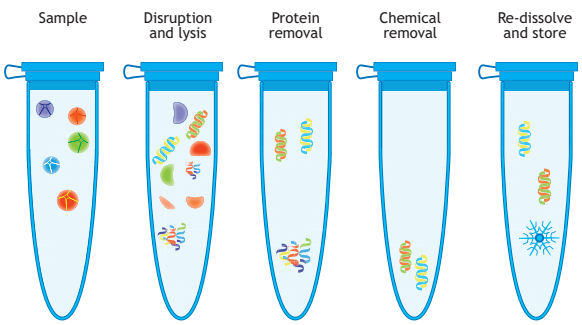

Figure 8.1 The main steps in DNA extraction.

8.2.3.2 Nuclease activity inhibition and protein removal

Microbial cells possess numerous enzymes that are specialised in degrading DNA (nucleases), and it is therefore essential that these are removed or inhibited as soon as possible after cell lysis. A common method to remove nuclease activity is by the addition of proteases, which are specialised in the degradation of proteins, including the nucleases. Afterwards the proteins are removed by increasing the salt concentration, which will cause them to precipitate. The precipitate can subsequently be removed by centrifugation, leaving the DNA in solution, and the proteins in the pellet (Miller *et al.*, 1999).

8.2.3.3 Purification

Along with the extracted DNA comes another type of nucleic acid, which is also found in the cell, the RNA. RNA molecules are often removed by the addition of the enzyme RNase that cleaves RNA specifically and leaves the DNA intact (Miller *et al.*, 1999). Following these steps, it is important to remove unwanted salts, detergents, proteins and other reagents used in the cell lysis process. This purification is usually carried out by precipitating the DNA with ethanol, as DNA is insoluble in this and can be pelleted by a centrifugation step (Bollet *et al.*, 1991). The DNA can be washed by replacing the supernatant with new ethanol. Alternatively, the DNA can be adsorbed to a matrix on a filter or a silica that allows washing steps and subsequent release by altering the salt concentration.

8.2.3.4 Elution and storage

After DNA has been isolated from the other cell constituents, and chemicals used in the extraction, the ethanol can be removed by evaporation and the DNA can be dissolved in DNase-free water or in a protective buffer solution such as Tris-EDTA (TE) buffer (Miller *et al.*, 1999). The purified DNA can be frozen and stored for years, or kept in the fridge for a few weeks.

8.2.4 Quantification and integrity

Depending on the downstream processing, it is important to quantify the DNA and check that it is not overly fragmented. The best quantification is provided using a fluorescence-based method that can distinguish between DNA and RNA, such as is implemented in the Qubit dsDNA kits. These kits can quantify DNA accurately at the very low concentrations needed for most sequencing-based methods (Singer *et al.*, 1997). However, the popular spectrophotometry-based nanodrop system can provide decent estimates for higher concentrations of DNA (> 20 ng μL^{-1}) when the DNA is very pure. The spectrum recorded by the nanodrop system also provides additional information about the purity of the DNA, as the ratio between the intensity at given wavelengths indicates the type of potential contaminants. This kind of information can be used to evaluate whether further purification steps are needed (Wilfinger *et al.*, 1997).

The size distribution of the DNA can be determined using classical gel electrophoresis (McMaster and Carmichael, 1977) or newer, more sensitive and typically

faster techniques implemented in dedicated and fully automated systems such as the 'bioanalyzer' and 'tape-station' instruments (Panaro *et al.*, 2000; Padmanaban *et al.*, 2013). Size determination relies on the fact that longer DNA fragments will move slower through a gel matrix when exposed to an electric current than smaller fragments. Comparing the travelled distance of your extracted DNA with the travelled distance of fragments of known sizes, it is possible to estimate the length distribution of your DNA fragments (McMaster and Carmichael, 1977). Accurate estimates of the length and concentration of DNA molecules are crucial for some types of molecular analysis.

8.2.5 Optimised DNA extraction from wastewater activated sludge

This protocol explains DNA extraction from activated sludge from WWTP. The protocol is based on the FastDNA spin kit for soil protocol (MP Biomedicals) with some modifications, mainly streamlining and longer bead beating, as published by Albertsen *et al.* (2015).

The key in DNA extraction is consistency and hence this protocol should be followed to the letter. If you choose to deviate from the protocol do it consistently, for all samples, throughout your experiment.

8.2.5.1 Materials

Materials needed for DNA extraction are:

- A FastDNA spin kit for soil (MP Biomedicals).
- A FastPrep-24 (MP Biomedicals).
- A microcentrifuge (preferably with a cooler).
- Spintubes (DNAse-free), 1.5 mL.
- Falcon tubes, 15 mL.
- Ice.
- Ethanol.
- Pipettes (range 1 µL to 1000 µL).
- DNAse-free tips (10 µL, 300 µL and 1000 µL).
- Nuclease-free H_2O (Qiagen).
- A permanent marker (freeze-resistant).
- A label printer (optional).
- PPE: lab coat, safety glasses, gloves.

8.2.5.2 DNA Extraction

The total time needed for DNA extraction from a sample is approximately 4 h.

1. Sample input
   a. Target volume: 500 µL.
   b. Target Total Solids (TS): 2 mg.
   ▲ Critical step Never spin the sample down to increase concentration!
2. Prepare and mark tubes for the whole workflow (per sample):
   a. 1 × Lysing Matrix E tube (from the kit).
   b. 1 × SPIN™ filter (from the kit).
   c. 1 × catch tube (from the kit).
   d. 3 × 1.5 mL DNAse-free tubes.
   e. 1 × 15 mL Falcon tube.
3. Thaw the sample aliquot at room temperature and store on ice until used.
4. Add 480 µL Sodium Phosphate Buffer: PBS (pH 8.0) and 120 µL MT Buffer to each of the Lysing Matrix E tubes.
5. Add 250 µL PPS (Protein Precipitation Solution) to one 1.5 mL spintube for each sample.
6. Re-suspend the binding matrix and add 1.0 mL to each of the 15 mL Falcon tubes.

- **Bead-beating**

1. Mix the sample before use e.g. by vortexing.
2. Transfer a sample volume equal to 2 mg TS[1] to a Lysing Matrix E tube and add PBS so the total added volume is 500 µL[2].
3. Perform bead-beating in the FastPrep-24 instrument
   a. Time: 4 × 40 s.
   b. Speed: 6 m/s.
   c. Adaptor: Custom.
   d. ▲ Critical step Remember to load the bead-beating tubes in a balanced loading pattern. A balance tube may be required.
   e. ▲ Critical step Between each 40-seconds interval, the samples should be kept on ice for 2 min to cool down.

- **Protein precipitation and binding of DNA to matrix**

1. Spin down the samples at > 10,000 × g for 10 min, preferably at 4 °C.
2. After centrifugation, transfer the supernatants to 1.5 mL spintubes with PPS and then shake the tubes 10 times by hand.
3. ▲ Critical step Keep the tubes on ice until all the samples have been processed.
4. Centrifuge the tubes at 14,000 × g for 5 min to pellet the precipitate.

[1] 1-4 mg of TS is usually acceptable.

[2] Use a pipette tip with a wide orifice so that large granules are also selected.

5. Transfer the supernatant to the 15 mL tube with the binding matrix suspension.
6. Invert by hand for 2 min to allow binding of DNA to the matrix.
7. Place the tube in a rack for 3-5 min (or until the liquid appears clear) to allow settling of the silica matrix.
8. Remove and discard up to 2 × 750 µL of supernatant, being careful to avoid the settled binding matrix.
9. Re-suspend the binding matrix in the remaining amount of supernatant.

- **DNA washing and elution**

1. Transfer approximately 750 µL of the mixture to a SPIN™ filter and then centrifuge at 14,000 × g for 1 min.[3]
2. Empty the catch tube.
3. ▲ Critical step Ensure that ethanol has been added to the concentrated SEWS-M.
4. Add 500 µL prepared SEWS-M and gently re-suspend the pellet using the force of the liquid from the pipette tip - or by stirring with a pipette tip.
5. Centrifuge at 14,000 × g for 1 min.
6. Empty the catch tube and use it again.
7. Centrifuge at 14,000 × g for 2 min to 'dry' the matrix of residual wash solution.
8. Discard the catch tube and replace it with a new tube.[4]
9. Allow the SPIN™ filter to dry for 5 min at room temperature with an open lid.
10. Gently re-suspend the Binding Matrix (above the SPIN filter) in 60 µL of DES. Use a pipette tip to stir the matrix until it becomes liquid. Make sure not to disrupt the filter.
11. Centrifuge at 14,000 × g for 1 min to bring the eluted DNA into the clean catch tube. Discard the SPIN filter.
12. Label the tube appropriately either with a printed label or with a freeze-resistant marker.
13. ■ Pause Point Store DNA at -20 °C for short-term storage and -80 °C for long-term storage.

8.3 REAL-TIME QUANTITATIVE PCR (qPCR)

8.3.1 General considerations

Although high-throughput techniques such as metagenome and amplicon sequencing (see Section 8.4) have revolutionized the way we interrogate microbial communities, real-time quantitative PCR (qPCR) still remains the most sensitive technique for quantification of specific DNA species. Also, under optimal conditions it allows the detection of a single target sequence within the analysed sample, although such conditions are rarely achieved for environmental samples due to PCR inhibitory substances. Furthermore, qPCR may be used to convert the relative abundance data obtained from amplicon sequencing into absolute quantities, although this is rarely required.

In wastewater treatment systems, qPCR can be used to estimate the overall bacterial abundance (Horz *et al.*, 2005) or to quantify bacteria belonging to specific taxonomic groups (Matsuda *et al.*, 2007) using primers that target conserved or variable regions of the rRNA (16S or 23S) genes, respectively. It may also be used to estimate the abundance of bacteria belonging to specific functional groups, such as nitrifiers or polyphosphate-accumulating bacteria, using primers that target key functional genes (Ge *et al.*, 2015). qPCR is also a useful tool for determining the fate of individual bioaugmentation strains. This can be done using primers that target unique genomic regions (Dueholm *et al.*, 2015). qPCR may furthermore be used to track the spread of antibiotic resistance genes (Volkmann *et al.*, 2004) and infectious viruses (Kitajima *et al.*, 2014). By combining qPCR with reverse transcription (RT-qPCR), the activity (transcription) of specific genes can be quantified (Nolan *et al.*, 2006), but this will not be covered here. qPCR is a refinement of the classical PCR (see Section 8.4.3) (Saiki *et al.*, 1985) in which the PCR products are detected after each PCR cycle (Figure 8.2). The technique relies on the fact that in a typical PCR, the target sequence is amplified approximately two-fold for each PCR cycle until one or more reagents become limiting (Kubista *et al.*, 2006). The PCR cycle, for which a detectable product appears, is known as the quantification cycle (C_q), and it relates to the abundance of the target sequence in the original sample (Brzoska and Hassan, 2014). The most widely used qPCR technologies are based on fluorescence reporters that allow the PCR products to be detected in real-time on thermal cyclers equipped with fluorescence detectors (Brzoska and Hassan, 2014). Two different chemistries are widely used for the detection, each with its benefits and drawbacks (Table 8.1). The first relies on the nucleic acid stain SYBR Green I, which becomes highly fluorescent upon intercalation with double-stranded DNA (dsDNA) (Figure 8.3A) (Zipper *et al.*, 2004). The second relies on single-stranded DNA (ssDNA) hydrolysis probes that contain a fluorophore

[3] If you have more sample than 750 µL, you should repeat this step.

[4] The new tube is the tube that the sample is to be stored in so be sure to label it properly.

with an intrinsically strong fluorescence at the 5' end and a quencher molecule at the 3' end (Figure 8.3B) (Holland *et al.*, 1991). The intact probes do not fluoresce as the close proximity of the fluorophore and the quencher results in energy transfer from the fluorophore to the quencher by fluorescence resonance energy transfer (FRET) (Holland *et al*., 1991). During the annealing step the probes bind to a DNA segment between the sequencing primers and are subsequently hydrolysed during the elongation by the 5'-3' exonuclease activity of DNA polymerase. This liberates the fluorophore from the quencher, leading to a fluorescence signal (Holland *et al*., 1991). Other less common techniques have also been reviewed (Kubista *et al*., 2006).

Table 8.1 Comparison of SYBR Green I and hydrolysis probe-based detection.

	SYBR Green I	Hydrolysis probe
Specificity	Detects all amplified double-stranded DNA, including non-specific reaction products.	Detects specific amplification products only.
Sensitivity	Depends on template quality and primer design and optimization.	1-10 copies.
Advantages	• Can detect the amplification of any double-stranded DNA sequence. • No probe is required, which can reduce the assay setup and running costs. • Multiple dyes can bind to a single amplified molecule, which increases the sensitivity for detecting amplification products. • Relative low cost of primers.	• Specific hybridization between the probe and target is required to generate a fluorescent signal, significantly reducing background and false positives. • Probes can be labelled with different, distinguishable reporter dyes. This makes it possible to have multiplex qPCR in one reaction tube. • Post-PCR processing is eliminated, which reduces the assay labour and material costs.
Disadvantages	• Because SYBR Green I dye binds to any double-stranded DNA (including non-specific double-stranded DNA sequences) it may generate false positive signals. • Cannot be used for multiplex qPCR (more then one primer set is used simultaneously).	• A different probe has to be synthesized for each unique target sequence. • Relative high cost of labelled probe.

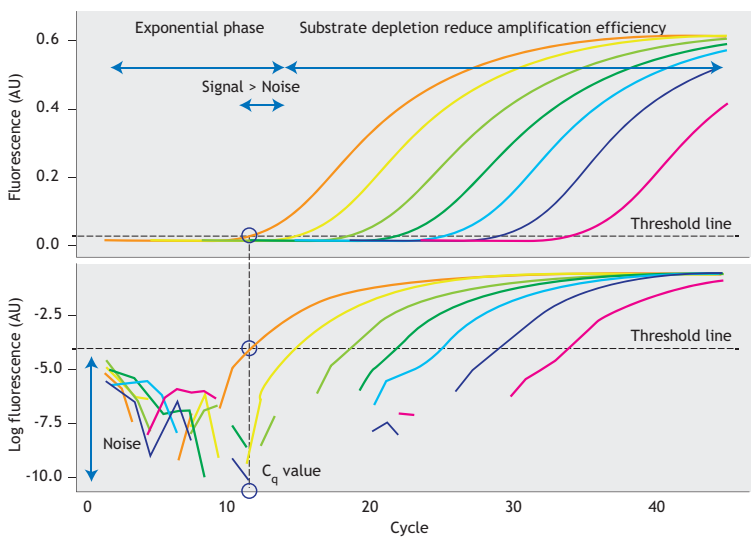

Figure 8.2 qPCR amplification curves. The upper panel shows traditional amplification plots and the lower shows the log-scaled data. The C_q value is determined from the qPCR cycle where the amplification curve intersects a threshold line placed where the fluorescence signal is significantly above the noise level. The threshold line can be determined manually from the log-scaled amplification plot (see the lower plot), but more often it is calculated by the qPCR software. An example is shown for the red amplification curve.

Before starting working with qPCR, it is recommended to read the minimal information for publication of quantitative real-time PCR experiment (MIQE) guidelines (Bustin *et al*., 2009). The aim of this section is to introduce the reader to the basic theory behind qPCR and provide details on how to adapt qPCR assays from the literature for wastewater research. Special emphasis will be on the required controls and the pitfalls in respect to data interpretation.

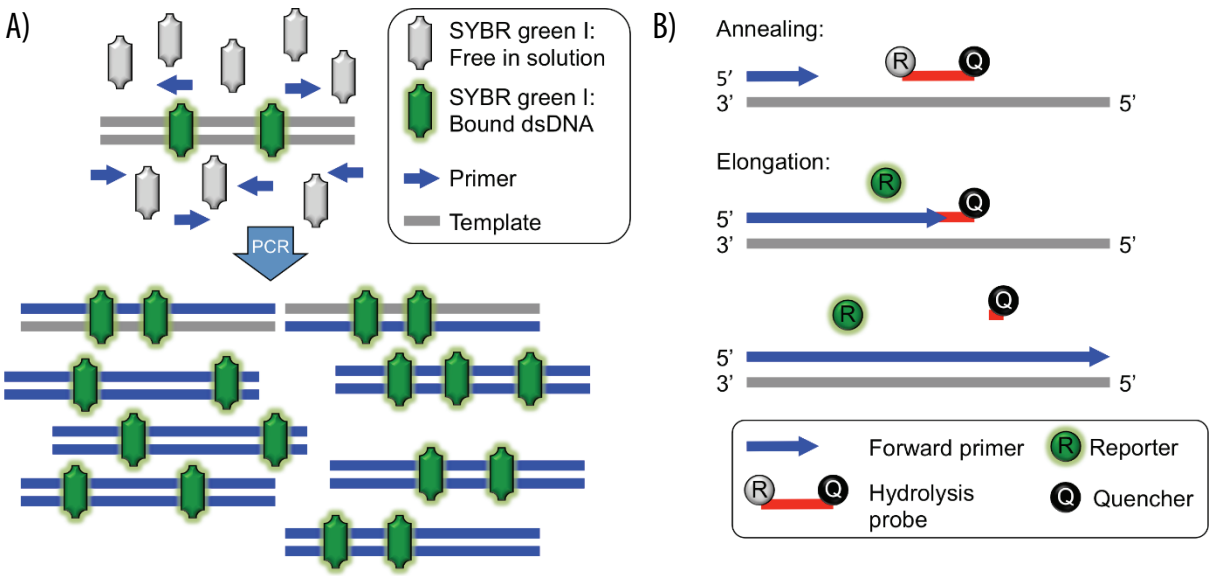

Figure 8.3 Mechanisms of common reporters used in qPCR. (A) SYBR Green I assay. The fluorescent reporter SYBR Green I is an asymmetric cyanine with two aromatic systems. When free in solution, SYBR Green I has virtually no fluorescence due to vibrations engaging both aromatic systems, which convert electronic excitation energy into heat that dissipates to the surrounding solvent. However, when SYBR Green I interacts with dsDNA, the vibrations are restricted and the asymmetric cyanine becomes highly fluorescent (Nygren *et al.,* 1998). SYBR Green I fluorescence consequently reflects the total abundance of dsDNA. (B) Hydrolysis probe-based assay. The hydrolysis probe contains a reporter fluorescent dye in the 5' end and a quencher dye in the 3´ end that greatly reduces the fluorescence of the reporter of the intact probe due to fluorescence resonance energy transfer (FRET). The probe anneals to the coding strand between the forward and reverse primer-binding sites. During each elongation step, the DNA polymerase cleaves the reporter dye from the probe. Once separated from the quencher, the reporter dye emits its characteristic fluorescence (Holland *et al.,* 1991). The reporter fluorescence consequently reflects the total number of target amplifications.

8.3.2 Materials

- **Primers**

A description of some generally applicable qPCR assays relevant for wastewater treatment is provided in Table 8.2. Primers and probes should be ordered at a concentration of 100 µM as desalted and HPLC-purified stocks, respectively. Probes designed with the combination of a 5' 6-FAM reporter and a 3' TAMRA quencher are suitable for most applications. However, higher sensitivity can be obtained by replacing TAMRA with a non-fluorescent quencher, such as the black hole quencher-1 (BHQ-1) (Biosearch Technologies, USA). The design of new qPCR assays requires advanced bioinformatics skills and will not be covered here. For instructions on how to custom-design qPCR assays, consult the following literature (Basu, 2015; Brzoska and Hassan, 2014).

Table 8.2 Examples of generally applicable qPCR assays for wastewater treatment.

Target and chemistry[1]	Description	Reference
Most bacteria (HP)	A universal qPCR assay for the amplification of the 16S rRNA gene from the domain bacteria was designed and evaluated. The assay allows quantification of the total bacterial abundance within a sample.	(Nadkarni *et al*., 2002)
Most archaea (HP)	The coverage of multiple primers for the 16S rRNA gene from the domain Archaea and bacteria was evaluated. Specific primers allow the relative abundance of Archaea and bacteria to be determined in microbial communities from various habitats.	(Wang and Qian, 2009)
Most fungi (HP)	A universal qPCR assay for the amplification of the fungal 18S rRNA gene was designed and evaluated *in silico* and *in vitro*. The assay allows quantification of the total fungal abundance within a sample.	(Liu *et al*., 2012)
Individual strains (HP)	A general technique for the development of strain-specific qPCR assays was presented and used to design a qPCR assay for the bioaugmentation strain *Pseudomonas monteilii* SB3074. The assay was subsequently used to evaluate the persistence of the strain in activated sludge.	(Dueholm *et al*., 2015)
Nitrification (HP)	A qPCR assay that targets part of the ammonia-monooxygenase sub-unit alpha gene (*amoA*), which is a key enzyme in ammonia oxidation by ammonia-oxidizing bacteria (AOB), was developed and used to estimate the population size of AOB in soil samples.	(Okano *et al.,* 2004)
Nitrate reduction (SG)	qPCR assays that targets membrane-bound (*narG*) and periplasmic (*napA*) protobacterial nitrate reductases were developed and used to determine their relative abundance in various environments.	(Bru *et al*., 2007)
Anaerobic ammonia oxidation (anammox) (SG)	A qPCR assay that specifically targets the 16S rRNA gene of all known anammox bacteria was designed and used to determine their abundance in wetland soils.	(Humbert *et al*., 2012)
Degradation of aromatic hydrocarbons (HP)	A qPCR assay was developed that targets *bssA*, which encodes the α-subunit of benzylsuccinate synthase, a key enzyme associated with anaerobic toluene and xylene degradation. The assay was used to study how gasohol releases from leaking underground storage tanks affected the indigenous toluene-degrading bacteria.	(Beller *et al.,* 2002)
Antibiotic resistance (HP)	qPCR assays were designed that targeted the antibiotic-resistance genes *vanA*, *ampC*, and *mecA*, which are related to vancomycin-resistant enterococci (VRE), β-lactam-resistant *Enterobactreiaceae*, and methicillin-resistant *Staphylococcus aureus* (MRSA), respectively. The assays were used to detect the resistance genes in municipal and clinical wastewater.	(Volkmann *et al*., 2004)
Waterborne pathogenic viruses (HP)	qPCR assays were used to determine the relative abundance of 11 different viruses in the influent and effluent of two wastewater treatment plants.	(Kitajima *et al*., 2014)

[1] SG: SYBR Green I; HP: Hydrolysis probe.

- **Real-time thermal cycler**

Real-time thermal cyclers can be obtained from a large range of suppliers, including Agilent Technologies (USA), Applied Biosystems Inc. (USA), Bio-Rad (USA); Eppendorf International (Germany), and Roche Applied Science (Switzerland). Good experience was reported using the Agilent Mx3005P qPCR system from Agilent Technologies.

- **qPCR reagents**

There is a wealth of commercial qPCR kits available. We have good experience with the Brilliant III Ultra-Fast SYBR Green qPCR Master Mix (Agilent Technologies) for SYBR Green I-based assays and the EXPRESS qPCR Supermix (Life Technologies, USA) for hydrolysis probe-based assays.

- **Equipment for measuring DNA concentration**

Authors recommend the use of a Qubit fluorometer (Life Technologies, USA) or a similar probe-based technique for the determination of DNA concentrations. A NanoDrop spectrophotometer (Thermo Scientific, USA) may also be used, but this technique is more sensitive to sample impurities (see Section 8.2.4).

8.3.3 Methods

- **Preparation of qPCR standards**

The absolute concentration of the target DNA sequence is determined by comparing the C_q value of the sample to a standard dilution series with known concentrations of the target sequence (amplicon) (Figure 8.4).

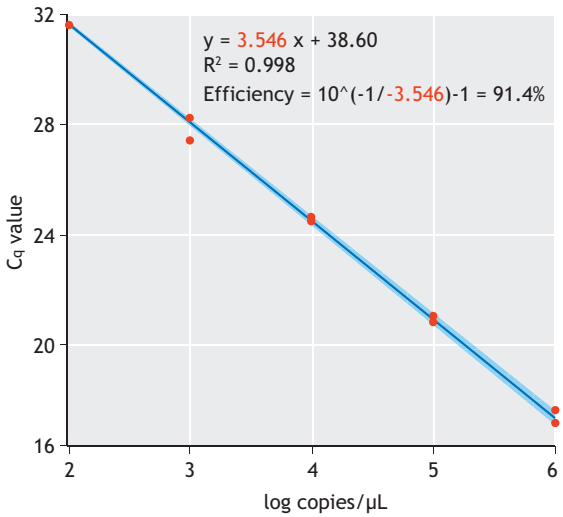

Figure 8.4 Evaluation of amplification efficiency from the slope of the linear regression of a standard dilution series.

The standards can be made from genomic DNA, plasmids, or PCR products. Although PCR products of the target sequence are easily obtained by PCR with the qPCR primers, these products often lead to a poor standard dilution series, as the small size makes it difficult to create reproducible dilutions. For routine qPCR assays, authors therefore recommend the use of linearized plasmids containing the amplicon. Linearization is important as the supercoiled circular confirmation of plasmid DNA may suppress PCR amplification (Hou *et al.*, 2010). Linearized plasmid qPCR standards are easily made as described below. Kits and enzymes should be used according to manufacturers' recommendations unless otherwise stated.

1. Amplify the target sequence using the qPCR primers and a standard Taq-polymerase.
2. Clone the PCR product into the pCR4-TOPO plasmid using the TOPO TA cloning kit for sequencing and *E. coli* One Shot TOP10 cells (Life Technologies, USA).
3. Inoculate 10 mL LB medium containing 50 µg mL^{-1} kanamycin with a positive clone and grow the culture overnight (37 °C, 200 rpm).
4. Purify plasmids from the culture using the QIAprep Spin Miniprep Kit (Qiagen, USA).
5. Linearize the plasmids using FastDigest ScaI or FastDigest SspI (Thermo Scientific, USA).
6. Purify the linearized plasmids using the QIAEX II Suspension kit (Qiagen, USA).
7. Determine the DNA concentration using a Qubit fluorometer or a NanoDrop spectrophotometer.
8. Calculate the molecular weight of the linearized plasmid with an insert using the equation below.

   MW = final plasmid size in bp × 607.4 g mol^{-1}

9. Calculate the target sequence abundance in the sample using the equation below.

   Target sequence copies per µL = concentration in ng $\mu L^{-1} \times 10^{-9} \times 6.022 \times 10^{23}$ / MW

10. Dilute the amplicon stock to 10^9 copies per µL with 10 mM tris buffer, pH 8.5.
11. Create a 10-fold dilution series ranging from 10^8 to 10^1 copies per µL with 10 mM tris buffer, pH 8.5. Use a new pipette tip and vortex the sample after each dilution.
12. Transfer 100 µL of the standards to 200 µL PCR 8-tubes strips and store at -18 °C until used.

- **Sample preparation**

1. Purify DNA from the samples as described in 'DNA extraction' in Section 8.2.5.
2. Determine the DNA concentration using a Qubit fluorometer or a NanoDrop spectrophotometer.

- **qPCR reaction setup**

1. Prepare the qPCR master mix according to the protocol supplied with the qPCR kit. The master mix is usually prepared so that it is suited for 5 µL samples.
2. Load the master mix into a qPCR assay plate.
3. Centrifuge the assay plate at 2,200 × g for 5 min.
4. Add duplicates of the standard dilution series into the first two columns of the qPCR plate.
5. Add duplicates of the samples to the qPCR plate.
6. Add the appropriate controls described below to the qPCR plate.
   a. NTC (No template control): NTC is prepared with DNA-free water instead of a DNA template. It serves as a general control for extraneous nucleic acid contamination. When using SYBR Green chemistry, it also serves as an important control for primer dimer formation.
   b. NAC (No amplification control): NAC is prepared without the DNA polymerase. It functions as a control for background fluorescence that is not a function of the PCR. Such fluorescence is typically caused by the use of partly degraded hydrolysis probes.

Consequently, the NAC is unnecessary in SYBR Green assays.

c. Diluted sample controls: These are used to determine whether the sample contains PCR inhibitors. This is the case if the diluted sample yields a significantly higher copy number than the sample after correction for the dilution factor.

d. Amplicon-spiked samples: Selected samples are spiked with a known high concentration of the amplicon and serve as controls for the presence of PCR inhibitors.

7. Centrifuge the assay plate at 2,200 × g for 5 min.
8. Run the qPCR according to the protocol supplied with the qPCR kit. Adjust the annealing temperature and elongation time according to the assay description.
9. If using a SYBR Green assay, end the qPCR run with a melting curve analysis. This may identify primer dimer formation and the production of unspecific products. Both can be seen as additional peaks in plots of the first derivatives of melting curves. Primer dimers have considerable lower melting temperatures than the target amplicon.
10. When applying a new qPCR assay for the first time, it is always a good idea to validate the assay. To do this, purify the produced PCR product and send a small aliquot and either the forward or reverse primer to a company that performs Sanger sequencing. From the sequencing data, confirm that the amplified product is indeed the target sequence. The purified product may also be analysed on an agarose gel. A single band at the predicted length should be observed.

8.3.4 Data handling

- **Determination of sample copy numbers**

Most real-time thermal cyclers are able to carry out the data handling automatically and provide the copy number for each sample. However, the copy number may also be calculated by comparing the C_q values of the sample to those of the standard dilution series manually. To do this, plot the C_q values of the standard dilution series against the logarithm ($\log_{10}$) of the target sequence abundance and then perform linear regression. The obtained equation may subsequently be used to determine the target abundance from the samples C_q values (Figure 8.4).

- **Evaluation of PCR efficiency**

The PCR efficiency describes how the amplification deviates from the ideal situation, where the amplicon concentration doubles after each PCR cycle. PCR efficiencies below 90 % may signify a sub-optimal PCR primer/probe design, the presence of PCR inhibitors or inaccurate sample or reagent pipetting, whereas efficiencies above 100 % always result from the inaccurate pipetting. It has been proposed as a guideline that the PCR efficiency should be between 80 and 115 % for environmental samples (Zhang and Fang, 2006).

The PCR efficiency can be determined from the slope of the linear regression of the standard dilution series described above (Eq. 8.1, Rasmussen, 2001). The calculated efficiency assumes that all the standards and samples have the same amplification efficiency (Souazé *et al.*, 1996). This is confirmed using the diluted sample or amplicon-spiked sample controls described earlier.

$$\text{Efficiency} = 10^{\frac{-1}{\text{slope}}} - 1 \qquad \text{Eq. 8.1}$$

8.3.5. Data output and interpretation

The final output of a qPCR analysis is a list of target sequence abundances for each sample. However, there are some important considerations to bear in mind when analysing the data that will significantly affect the final conclusions (Kim *et al.*, 2013).

- **Extraction of nucleic acids is biased**

Samples from wastewater treatment systems contain a large diversity of microorganisms, which display considerable variation in their cell wall architecture (Saunders *et al*., 2015). Some of these are easy to lyse, whereas others are more difficult. There is consequently a significant bias introduced by the choice of nucleic acid extraction procedure (Albertsen *et al*., 2015). Accordingly, it may be very difficult to compare absolute quantification across studies. The DNA extraction protocol described in Section 8.2.5.2 provides results that are comparable to those of quantitative FISH analysis for wastewater treatment samples.

- **Quality of the template DNA**

Environmental samples, such as those from wastewater treatment systems, often contain compounds that have adverse effects on the PCR amplification (Bessetti, 2007). These may be humic acids, heavy metals, polysaccharides, phenolic compounds, or urea. Such inhibitors can be removed by sample polishing using adsorbent compounds, chemical washing or gel purification (Schriewer *et al*., 2011). However, it is important to always evaluate the removal of inhibitory compounds empirically as described in the Section 8.3.3 (Stults *et al.*, 2001).

- **Specificity of broad-range qPCR assays**

The qPCR assays used in wastewater treatment often target microbial groups rather than individual species or strains. The assays consequently use generic primers and probes that are designed based on the known degeneracy of the target sequence. However, the known degeneracy might not always reflect that what is seen in nature, resulting in over or under estimation of the target sequence. The use of highly degenerate primers and probes also poses another problem. If the microbial community is heavily enriched in specific organisms, the perfect matching primers for these organisms will be quickly depleted, whereas primers for low-abundant organisms will be present for longer. The amplification will consequently be biased toward the low-abundant organism, resulting in an underestimation of the target sequence abundance. Finally, there may be differences in the amplification efficiency for each organism due to the variation in their GC content (Kim *et al.*, 2013).

- **Amplification of extracellular DNA (eDNA)**

Biological processes such as wastewater treatment rely on the active population of microorganisms. However, qPCR is unable to distinguish between DNA originating from active bacteria and extracellular DNA (eDNA) originating from dead and lysed cells. As wastewater treatment samples contain considerable amounts of eDNA, this may bias the data (Dominiak *et al*., 2011). Care should therefore be taken when making conclusions about activity based on qPCR results.

- **Variation in the gene copy number**

Microbial genomes show a large variation in the copy number of metabolically important genes such as the 16S rRNA gene (Větrovský and Baldrian, 2013). This can bias quantification of the specific bacterial number unless the copy number is known. In addition, the number of whole genomes per cell may vary depending on the growth state of the bacteria (Ludwig and Schleifer, 2000). If the relative abundance of a specific bacterial species needs to be investigated, it is recommended to apply 16S rRNA amplicon data (see Section 8.4).

8.3.6 Troubleshooting

- **The sample contains PCR inhibitors**

There are three ways to circumvent PCR-inhibitor effects. The simplest option is dilution of the sample. PCR inhibitors are only effective above a certain concentration. However, dilution of the sample will also reduce the signal, leading to a less sensitive assay. The second option is to polish the DNA by applying an additional purification. This requires a concentrated sample, as material is always lost during purification. Large-scale sample polishing can be carried out in 96-well PCR plates using magnetic bead-based purification kits. The third and final solution is to purify the DNA from the original sample using another purification kit that is optimised for the given inhibitor.

- **The primer or probe design is not optimal**

A qPCR assay based on poor primers and probes should never be used. Instead, design and evaluate new primers and a probe set. Guidelines are given by several authors (Basu, 2015; Brzoska and Hassan, 2014).

- **Inaccurate sample and reagent pipetting**

Inaccurate calibration of pipettes is detrimental to qPCR. It is therefore recommended to keep a dedicated set of pipettes for qPCR that are regularly checked. The use of a multi-dispersal pipette is also recommended as it simplifies sample handling. Finally, it may be a good idea to review your pipetting techniques before carrying out qPCR.

8.3.7 Example

Enhanced degradation of specific pollutants can be achieved by the addition of catabolically relevant bacterial strains to the activated sludge in WWTPs (El Fantroussi and Agathos, 2005). This is known as bioaugmentation. Successful bioaugmentation requires that the introduced strains are able to thrive in the new environment (Thompson *et al.,* 2005). qPCR may be used to evaluate the persistence of bioaugmentation strains *in situ*. A strain-specific qPCR assay can be developed based on unique genomic sequences in the bioaugmentation strains. The abundance of the strains can subsequently be determined using DNA extracted from activated sludge at various time points after the addition of the bioaugmentation strain (Dueholm *et al.*, 2015). Here we show an example of how qPCR has been used to evaluate the persistence of the bioaugmentation strains *Pseudomonas monteilii* SB3078 and SB3101, which are used for the degradation of aromatic hydrocarbons (Dueholm *et al*., 2014; 2015).

8.3.7.1 Samples

The bioaugmentation strain *P. monteilii* SB3078 or SB3101 was introduced into 100 mL fresh activated sludge obtained from Aalborg East WWTP in an abundance of 1 % based on the cell number. It was roughly estimated that an suspended solids (SS) = 1 g L^{-1} (activated sludge) and an $OD_{600\ nm}$ = 1 (*Pseudomonas* pure culture) both correspond to approximately 10^9 cells

mL^{-1} as previously shown (Frølund *et al*., 1996). Benzene was added to 10 µg mL^{-1}. The culturing flasks were crimp-sealed using butyl rubber stoppers and incubated at 25 °C, 150 rpm, for 4 days. The flasks were opened every 12 h for 30 min to allow evaporation of the remaining trace levels of benzene and settling of the sludge particles. 50 mL of effluent water was then removed and replaced by 50 mL of primarily settled wastewater, simulating a hydraulic retention time of 24 h. Samples for DNA extraction were collected and benzene reintroduced to 10 µg mL^{-1}. The flasks were then sealed and the incubations continued (Dueholm *et al.*, 2015). DNA was essentially extracted as described in Section 8.2.

8.3.7.2 qPCR reaction setup

1. Prepare the qPCR master mix as described below. The primers and probe target both SB3078 and SB3101.

Item	Final concentration	Per reaction (20µL)	100 × reactions
EXPRESS qPCR Supermix	1 ×	10 µL	1,000 µL
ROX (25 µM)	50 nM	0.04 µL	4 µL
Forward primer (100 µM)	500 nM	0.10 µL	10 µL
Reverse primer (100 µM)	500 nM	0.10 µL	10 µL
Hydrolysis probe (100 µM)	200 nM	0.04 µL	4 µL
DEPC water	-	4.72 µL	472 µL
Aliquot	-	15 µL	-

2. Load the master mixes into a qPCR assay plate and add 5 µL of the samples and controls (see above).
3. Centrifuge the assay plate at 2,200 × g for 5 min.
4. Run the qPCR according to the following program:
   - 50 °C for 2 min (UDG incubation).
   - 95 °C for 2 min.
   - 45 cycles of:
     - 95 °C for 15 s.
     - 60 °C for 1 min.

8.3.7.3 Results

We started by evaluating the amplification efficiency. The C_q values of the standard dilution series were plotted against the logarithm of the copy number and linear regression was then performed. This yielded the following fitting equation:

$$y = -3.349 \cdot x + 39.80;\ R^2 = 1$$

The efficiency was then calculated from the slope as:

$$\text{Efficiency} = 10^{(-1/\text{slope})} - 1 = 10^{(-1/-3.349)} - 1 = 98.9\ \%$$

This was well within the acceptable regime for environmental samples of 80-115 % (Zhang and Fang, 2006).

Next, we evaluated the controls. The NTC and NAC control did not amplify within the 45 cycles. This confirmed that there were no amplifiable contaminants present in the reagents and that the probes were stable, respectively. A sample containing DNA extracted from untreated wastewater was also analysed. This control did not amplify either, confirming the specificity of the qPCR assay. Finally, we investigated the amplification of a few diluted samples. These produced similar results to the undiluted samples, confirming no significant effect of inhibitors. And we had a look at the experimental data (Figure 8.5).

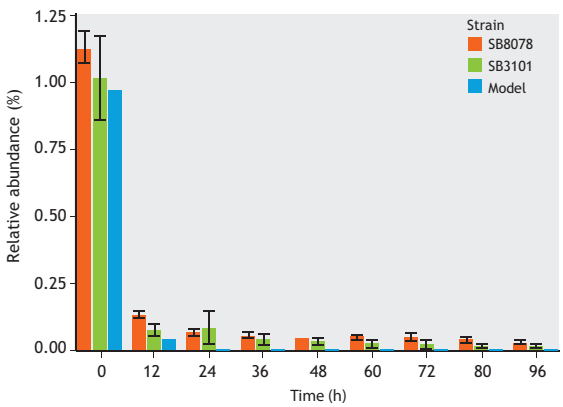

Figure 8.5 Persistence of *P. monteilii* SB3078 and SB3101 in activated sludge treated as a sequential batch reactor. Relative abundance of the bioaugmentation strains was determined using the strain-specific qPCR assay. The modelled data represents the theoretical decrease in cells that would be observed if there were no net growth and all bioaugmentation cells were planktonic. The volume occupied by solid material was calculated based on the diluted sludge volume index and this information was used together with the hydraulic retention time to calculate the rate by which planktonic cells were washed away.

The qPCR assay showed that approximately 90 % of the added bioaugmentation strains were lost within the first 24 h. This was probably due to the removal of planktonic cells with the effluent water, as it is replaced

by fresh wastewater every 12 h. The remaining bioaugmentation cells were stabilised within the sludge and were able to survive throughout the experiment (4 d). The data furthermore showed that SB3078 was more persistent than SB3101 in the activated sludge settings.

8.4 AMPLICON SEQUENCING

8.4.1 General considerations

The first step of trying to understand how the bacteria in activated sludge impact the performance of a wastewater treatment plant is to get an overview of the bacterial community. This involves identifying the bacteria, their abundance and knowledge about what they are doing. Advances in DNA sequencing have made it possible to identify bacteria with high resolution and throughput by reading the 16S ribosomal RNA (rRNA) genes of the bacteria and using them as 'fingerprints'.

The approach is called 16S rRNA amplicon sequencing and consists of a number of steps, which are depicted in Figure 8.6.

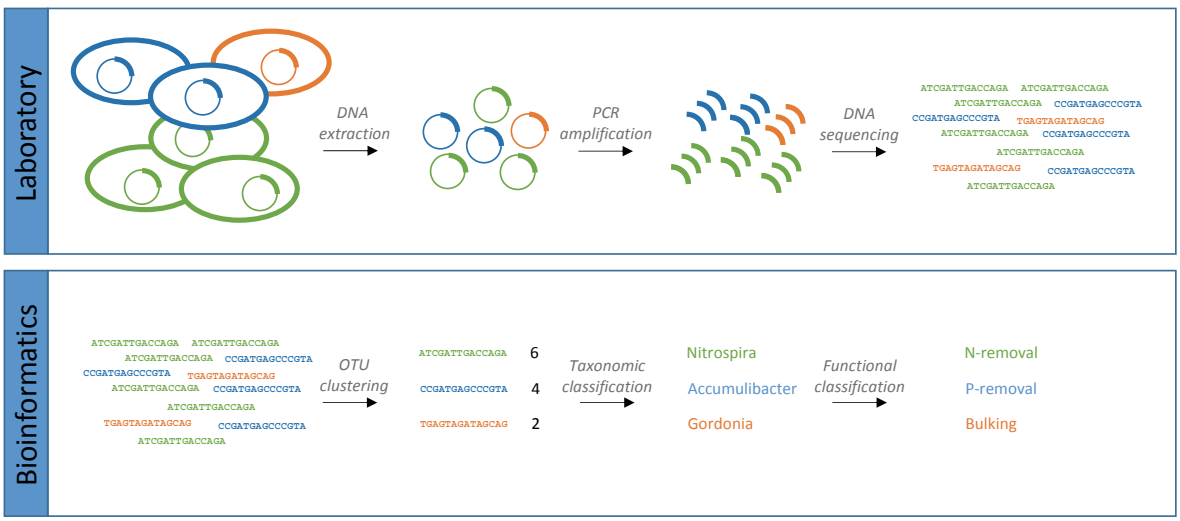

Figure 8.6 Overview of the basic steps in the analysis of microbial communities using 16S rRNA amplicon sequencing.

The first step is '*DNA extraction*', whereby genomic DNA from all the bacteria in a sample is extracted and purified (see Section 8.2). Afterwards, the 16S rRNA genes are amplified and prepared for sequencing by 'Polymerase chain reaction (PCR)'. The 16S rRNA gene amplicons are then read, or 'DNA sequenced', on a next-generation sequencing instrument. The output is the sequences of all the 16S rRNA genes in the sample. These sequences are then clustered into groups that represent a species, and the relative number of 16S rRNA genes belonging to each group is counted. Each group of species is identified through 'Taxonomic classification' by comparing the representative 16S rRNA gene sequence of each group to a database of known bacteria. The result is a table containing the name of each species in the sample and their relative abundance. This table is the basis for visualizing and analysing the bacterial community. The name of the species can also be used to link to functional information found in the literature or public databases such as MiDAS, midasfieldguide.org (McIlroy *et al.*, 2015), e.g. if some of the identified species are known foam formers or nitrifiers.

In the Sections 8.4.2 to 8.4.7, each step of the 16S rRNA amplicon sequencing approach will be described in detail. The descriptions are based on 16S rRNA amplicon sequencing using the Illumina sequencing platform. However, the basic idea is the same for all sequencing platforms.

8.4.2 The 16S rRNA gene as a phylogenetic marker gene

A phylogenetic marker gene encodes an essential function, which is shared by all the organisms that are to be targeted and which have not been subjected to lateral

gene transfer. In addition, the marker gene also has to have both evolutionary highly conserved positions as well as highly variable positions in its nucleotide sequence. The conserved parts make it possible to target the gene with a PCR and are needed for correct phylogenetic analysis, while the variable parts enable us to distinguish between different organisms and to investigate their relatedness (phylogeny).

Ribosomal genes have been the choice for phylogenetic analysis since Woese and Fox used them to show the division of life into three separate kingdoms in 1977 (Woese and Fox, 1977; Pace *et al*., 2012). Today the 16S rRNA gene is by far the most applied phylogenetic marker gene in environmental studies of bacterial diversity.

The 16S rRNA gene encode for a piece of RNA that makes up a functional part of the bacterial ribosome. Ribosomes are the protein factories of all cellular life forms and developed early in evolution. Conserved regions are crucial for correct ribosome structure and function, which means that most mutations in these regions are strongly selected against. Variable regions have more freedom for change, and mutations happen much more frequently (Madigan and Martinko, 2006). Therefore, the 16S rRNA gene contains several conserved islands with variable regions in between, called variable regions 1 to 9 (V1 to V9) (Ashelford *et al.,* 2005); see Figure 8.7.

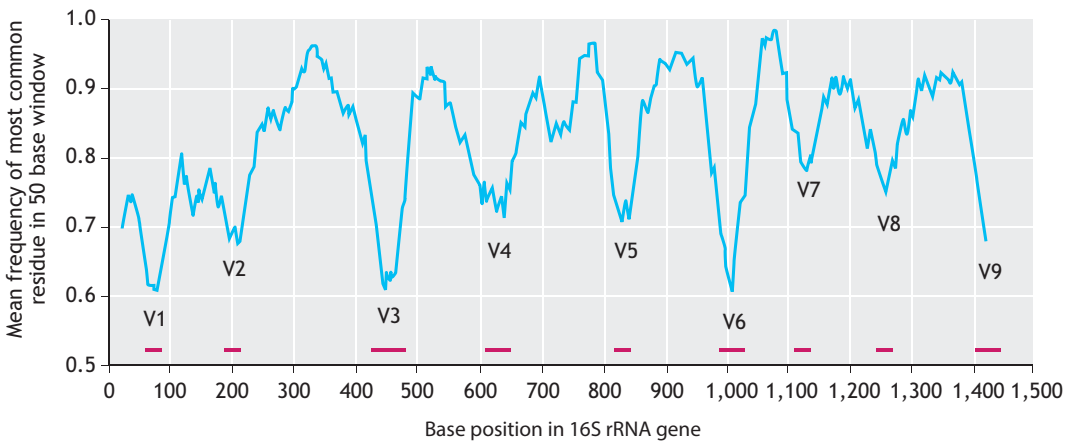

Figure 8.7 Variability in sequence composition across the 16S rRNA gene (adopted from Ashelford *et al.,* 2005).

This makes the 16S rRNA gene an excellent marker gene as it can be targeted in its entirety or in fragments, which provides some technical flexibility. The 16S rRNA gene has been used as a marker gene for many years, and the accumulated knowledge is compiled in extensive databases that are used for comparing 16S rRNA gene data and determining the phylogenetic affiliation of new 16S rRNA genes and assigning them to a group within the bacterial taxonomy. Usually it is possible to determine the taxonomy of bacteria down to the species level by its 16S rRNA gene sequences. The ubiquity and the resolution of the 16S rRNA gene, together with the database resources, make it the preferred marker gene for bacterial community analysis.

Other marker genes are used for bacterial community analysis as well but to a lesser extent, and often to obtain higher phylogenetic resolution e.g. to strain level. It is common for these marker genes to be more variable in their sequence composition, which provides higher phylogenetic resolution. However, this also means that they only target specific subgroups of bacteria. Examples of such marker genes are the *amoA* to target ammonium-oxidizing organisms (AOO) and the *mcrA* in methanogens; see also Section 8.3 about qPCR. The principles for analysing these other marker genes are similar to the 16S rRNA gene, although protein-coding genes such as *amoA* can also be analysed at the level of amino acid sequence. For strain resolution, nucleotide sequences are needed, but they must be aligned codon-wise based on an amino-acid alignment (Juretschko *et al.*, 2000).

8.4.3 PCR amplification

The first step in 16S rRNA amplicon sequencing is amplification of the 16S rRNA genes by PCR.

8.4.3.1 PCR reaction

The PCR is used to selectively amplify the 16S RNA gene from the background genomic DNA, so there is enough material that it is practical to analyse with the DNA sequencer (Figure 8.8).

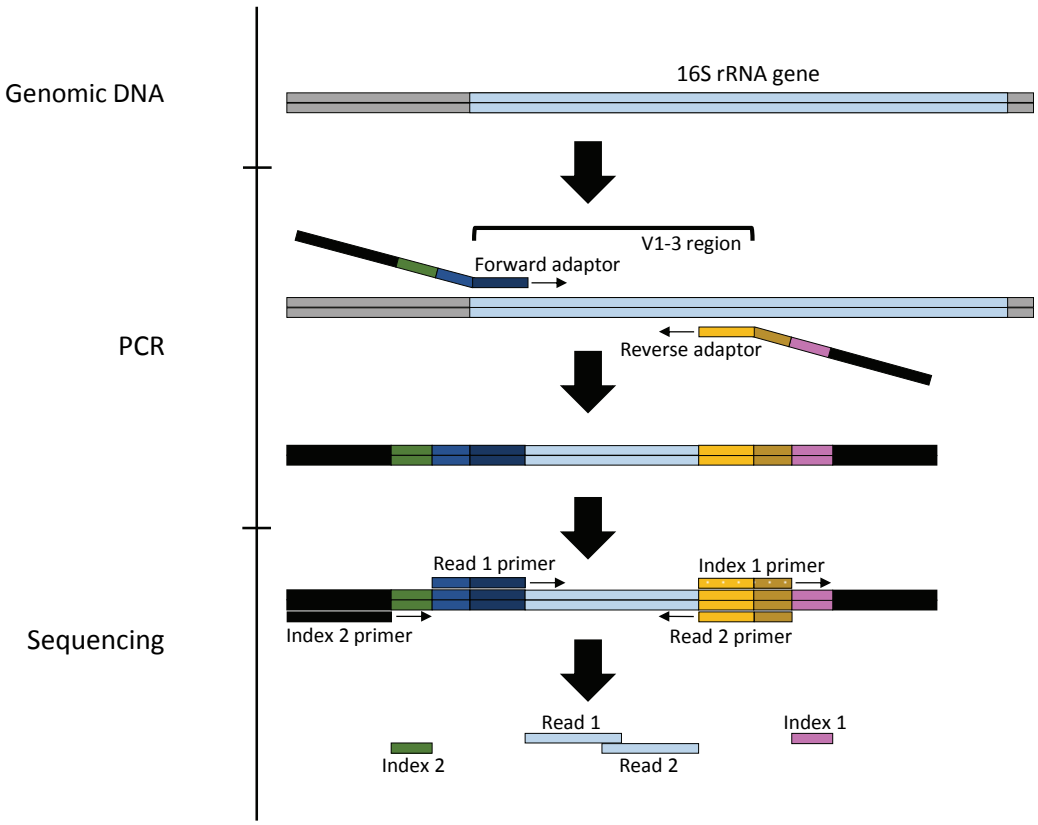

Figure 8.8 The basic steps in PCR amplification and sequencing of the 16S rRNA gene. Grey represents genomic DNA and light blue represents the 16S rRNA gene. The coloured parts of the adaptors and primers represent different functional sequences, and these are described in Section 8.4.11. The arrows signify the elongation/sequencing direction.

PCR uses a thermostable polymerase enzyme to copy DNA. The polymerase copies the DNA using small pieces of synthetic DNA (15-25 base pair: bp), called primers, as the starting point for the amplification. When analysing the 16S rRNA genes, the primers are designed to selectively match the conserved parts of the 16S rRNA gene, which enables a selective amplification of the region in between. During the PCR, cycles of heating and cooling are used to enable the amplification. First, the dsDNA is denatured by heating, which splits the two strands from each other. Secondly, the reaction is cooled to enable the primers to anneal to their target sites on the DNA. Thirdly, the temperature is increased again to provide an optimal temperature for the polymerase activity, which starts the copying of the 16S rRNA gene by extending the primer. One cycle creates two copies of the 16S rRNA gene from one original copy, and the copies can themselves function as templates. The cycle is repeated 25-35 times during the PCR, which results in an exponential amplification of the 16S rRNA gene. The copied gene product from PCR amplification is called an amplicon, hence the name 16S rRNA amplicon sequencing. Other than the DNA, primers and polymerase, the reaction contains the nucleotides that are

incorporated into the new DNA along with buffers and other additives (Mg^{2+} salts etc.) that provide the optimal conditions for primer annealing and polymerase activity. For a detailed discussion of the principles behind PCR see Green and Sambrook (2012).

Another important role of the PCR is to enable sequencing of the 16S rRNA amplicons. This is achieved by having a sequencing adaptor attached to the end of the primers used in the PCR. Hereby, the adaptor is attached to all the 16S rRNA amplicons that are produced. The adaptor consists of a synthetic piece of DNA (ca. 50 bp), which has different 'active' components (see Section 8.4.11). These components enable the sequencing machine to catch the 16S rRNA amplicons, to start reading them and to recognize which sample each 16S rRNA amplicon originated from, which is referred to as barcoding or indexing. The idea is to tag all the 16S rRNA amplicons from one sample with the same barcode. This enables a large number of different samples to be mixed together (multiplexing) and to be read at the same time on the DNA sequencer, and still to be able to separate the data from the individual samples afterwards (de-multiplexing) (Illumina Inc., 2015; Caporaso *et al.,* 2010).

The final product after PCR is called a 16S rRNA amplicon sequencing library. There are different strategies that can be used when preparing 16S rRNA amplicon sequencing libraries, but the overall idea is the same. The strategy described above uses a single step to amplify and attach adaptors; other strategies use two separate PCRs. These strategies have pros and cons in respect to cost, time and sequencing requirements. Currently, it is only feasible to sequence long 16S rRNA gene fragments with the one-step PCR strategy described. This includes the V1-3 fragment commonly used in activated sludge (Albertsen *et al*., 2015).

8.4.3.2 PCR biases

Different types of biases can be introduced in the PCR step, which will affect the final observed community structure. Primer bias is one of the most significant ones, and this will be addressed in the following paragraph (Albertsen *et al*., 2015). PCR drift is a bias introduced by stochastic events in the first cycles of the PCR, where relatively few molecules are involved in the replication or simply by reagent/sample handling variations (pipetting, position in thermocycler etc.). PCR selection bias is due to varying amplification efficiencies caused by physical properties of the 16S rRNA gene nucleotide sequence (Polz *et al*., 1998; Kennedy *et al.,* 2014). To reduce the impact of PCR drift and selection, replicate PCR reactions are performed, the number of PCR cycles is kept to a minimum and the amount of template DNA should be around 10 ng. Most standard amplicon sequencing protocols have optimized these parameters.

8.4.3.3 Primer choice

As mentioned above, the primer set matches conserved parts of the 16S rRNA gene. However, it is unavoidable to have some variability in the 'conserved' part of the 16S rRNA gene. Therefore, primers will match some bacteria better than others and for some they will not match at all (Klindworth *et al*., 2013). This introduces a significant primer bias to the whole analysis, which is important to be aware of.

When analysing the 16S rRNA gene it would be ideal to sequence the entire gene (approximately 1,600 bp) as this provides maximum phylogenetic resolution. However, due to limitations in the Illumina sequencing technology, it is currently only possible to sequence fragments up to 550 bp of the 16S rRNA gene.

As a consequence of the primer bias and the limitation in reading length, many primer sets have been designed that target different variable regions of the bacterial 16S rRNA gene. The most commonly used primer sets target the V1-3, V4 and V3-4 regions (Albertsen *et al*., 2015). The primer sets have different biases, and when selecting a primer set several things should be considered.

a. The primer set should have least possible bias against the bacteria in your samples that you are most interested in. While it is possible to get an idea of the primer bias via *in silico* analysis (Klindworth *et al*., 2013), it is always recommended to test the primers by sequencing.
b. The primer set should be the same as the studies you want to compare to. The MiDAS database, which attempts to summarize all the current knowledge about important bacteria in activated sludge, is based on primers targeting the V1-3 region. This is due to a good resolution and broad coverage of bacteria that are responsible for the processes of interest in the activated sludge community (Albertsen *et al*., 2015).

For samples from specialized activated sludge systems it can be a good idea to test other primer sets. For example, the most common anammox bacteria are not targeted very well by the V1-3 primer set, and it is better to use V4 primer sets instead (Laureni *et al*., 2015; Gilbert *et al.,* 2014). It is also possible to design new

primers, but usually this is not recommended, since it requires expert knowledge of microbial ecosystems and phylogeny as well as extensive laboratory time for optimization and validation.

8.4.4 DNA sequencing

8.4.4.1 Sequencing platform

After the 16S rRNA amplicon libraries have been prepared, there are a number of different options for DNA sequencing. However, each method employs markedly different strategies for sequencing, which means the resulting data is suited for different purposes and needs. In respect to 16S rRNA amplicon sequencing, the most important criteria are sequencing length (> 200 bp), sequencing quality (< 1 % errors), data yield (> 10,000 reads per sample), turnover time, cost and how easy library preparation is. Currently, early 2016, the Illumina MiSeq platform is the method of choice, when compromising between these criteria. The Illumina MiSeq enables analysis of up to 400 16S rRNA amplicon libraries (50,000 reads per sample) in a single sequencing run (56 h). The MiSeq is currently able to sequence 301 bp from each end of the 16S rRNA amplicons. This is also termed paired-end (PE) sequencing and each of the two 301 bp sequences are termed 'reads'. During data processing the two reads are merged together at the overlapping ends to obtain a maximum length of approximately 550 bp. For more specialized use, which requires longer read length or shorter turn-around time, other platforms are better suited e.g. Pacbio RS II (Pacific Biosciences) or the Ion Proton System (Thermo Fisher Scientific Inc), and soon the MinION (Oxford Nanopore Technologies). Be aware that different library preparation protocols have to be used for the different sequencing platforms.

8.4.4.2 Sequencing depth

When performing amplicon sequencing, it is important to have a rough estimate of the sequencing depth needed (number of reads per sample) in order to answer the questions posed through the experimental design. For general bacterial community analysis targeting the V1-3 of 16S rRNA gene in activated sludge, 50,000 raw PE reads per sample are routinely used. This is done from the rationale that it is often rather similar communities that are compared and it is usually important to get robust estimations of the abundance of the individual community members. A rule of thumb is not to have less than 100 reads from the bacteria of interest. Below 100 reads the final results become very uncertain, due to biological and technical variation (Albertsen *et al*., 2015). If higher resolution is needed, the best option is to include more biological replicates. However, if this is not an option, deeper sequencing can also be used. Activated sludge contains thousands of different bacteria, and the most important 100 of these account for more than 70 % of the total community abundance (Saunders *et al*., 2015). On average each of these 100 species make up > 0.5 % of the total community. To obtain > 100 reads from each of these species at least 20,000 bioinformatic processed reads per sample are needed or > 30,000 raw PE reads, depending on the sequencing quality. It should be noted that the sequencing cost is usually not the most expensive part of the analysis and hence it is often preferred to make sure that enough reads are sequenced.

8.4.5 Bioinformatic processing

8.4.5.1 Available software

A rigorous standard procedure for the handling of 16S rRNA sequencing data has still not been defined, and probably will not be anytime soon due to a rapidly developing field. However, the general idea remains the same and is depicted in Figure 8.9.

Many research groups make custom workflows and code performing processing of the data, and some have made comprehensive software bundles of these that can perform almost completely automatically. The most popular are QIIME (Caporaso *et al*., 2010), Mothur (Schloss *et al*., 2009) and UPARSE (Edgar, 2013). Their settings and underlying assumptions differ, also between versions, which will produce somewhat different results even from the same sequence data. Therefore, results from different software packages and versions should not be compared. Generally, it is advised to analyse all the data from an experiment with one software package in one session and to re-run all the analysis if some data is added or if settings are changed. In the following paragraphs, some general observations are made regarding the bioinformatic processing.

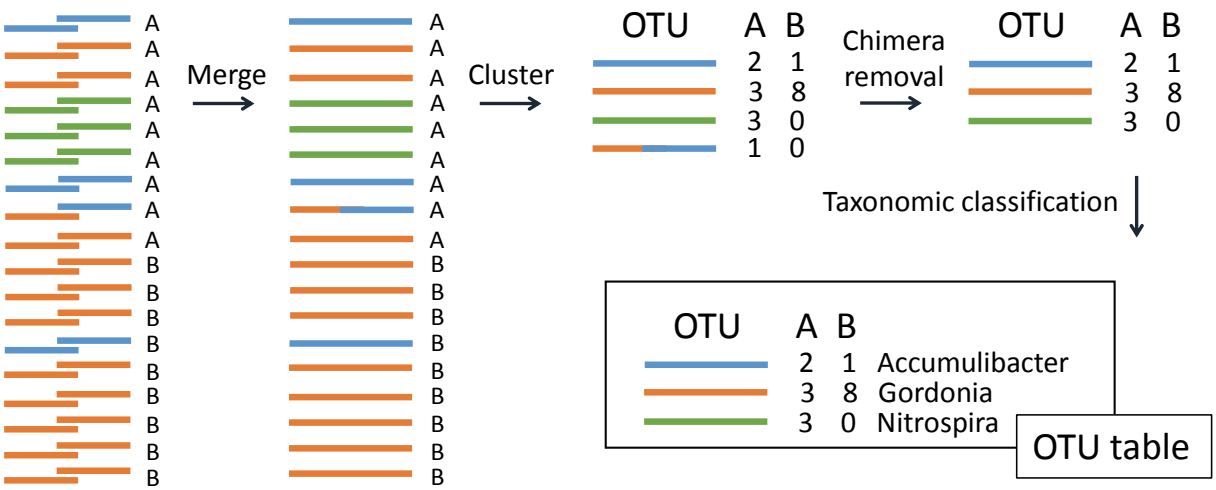

Figure 8.9 The basic steps in bioinformatic processing of 16S rRNA data from the Illumina MiSeq platform. The colours signify sequences that are > 97 % identical. The letters A a dn B signify what sample a sequence is from.

8.4.5.2 Raw data

The output from the Illumina MiSeq sequencer is a pair of files for each sample. Each file pair contains all Read1 of the paired reads in the first file and all Read2 in the second file. The files are fastq files (.fastq or .fq), which are simple text files with specific rules for listing sequence information (Figure 8.10).

Each read in a fastq file takes up four lines: *i*) The ID of the specific read starting with '@', *ii*) The raw sequence letters of the specific read, *iii*) '+' which works as a description field, but most often blank, and *iv*) Phred quality scores denoted by ASCII letters for each of the bases in the read (Cock *et al.,* 2010).

```
1  @M00878:94:000000000-AGD1E:1:1101:18594:12300 1:N:0:19
2  ATAGTACGTACGGTGTGCCAGCAGCCGCGGTAATCGGGTACGTACGTCGATCCGTA
3  +
4  CCCCCGGGGGGGGEFGGGGGGGFGCCFGEFGGGGGFGGFDGGGGGGGGDFCFCCDF
5  @M00878:94:000000000-AGD1E:1:1101:5140:12419 1:N:0:19
6  GACGAACGCTGGCGGCGTGCCTAATACATGCAAGTCGAACGAACGGATAAAGAGCT
7  +
8  CCCCCGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGFFGG
9  @M00878:94:000000000-AGD1E:1:1101:4381:16236 1:N:0:19
10 GACGAACGCTGGCGGCGTGCCTAATACATGCAAGTCGAGCGAGCGGAACTAACAGA
11 +
12 CCCCCGGGGGGCGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGEGGGGGGGG
```

Figure 8.10 Example of a fastq file with three reads. The reads are truncated for readability.

8.4.5.3 Quality scores and filtering

The Phred quality score is a measure of the probability that a specific base contains an error (Q10 = 10 %, Q20 = 1 % and Q30 = 0.1 %) (Cock *et al.,* 2010). For 16S rRNA amplicon sequencing, Q20 and above is preferred. The quality can be assessed by plotting the Phred quality scores, and the error rate of an internal standard. For 2 × 301 bp paired-end sequencing on a MiSeq of high-quality amplicon libraries, the average quality is usually above Q30 for the first 250 bp of Read1 and 200 bp of Read2, and after that the quality of the reads usually deteriorates (Figure 8.11). If the majority of the read quality is below Q20, something might have gone wrong during the sequencing. The bad quality data is removed through trimming and filtering. Often, if long stretches of a read have below Q20, the stretches are trimmed off from the 3' end. When sequencing V1-3 16S rRNA amplicons, reads that < 275 bp after trimming are discarded, since they are not suited for merging.

8.4.5.4 Merging paired-end reads

After quality filtering, the paired reads in the remaining high quality data are merged to create one continuous 16S rRNA gene sequence of a length of 250-550 bp depending on the targeted variable regions. After the filtering and merging there are usually around 70 % reads left, depending on the quality of the specific sequencing run. However, it is critical to be careful in the merging step as large biases can be introduced here. The length of the variable regions differ between species, e.g. the V1-3 region of the 16S rRNA gene can be between 425 and 525 bp. If the read quality is poor, 2 × 300 bp paired reads might be trimmed to 2 × 245 bp (490 bp), which is too short for merging of read pairs originating from species with large V1-3 fragments. Read pairs that doesn't merge are always discarded and thereby a bias is introduced. This can be prevented by discarding all reads < 275 bp before merging as mentioned before.

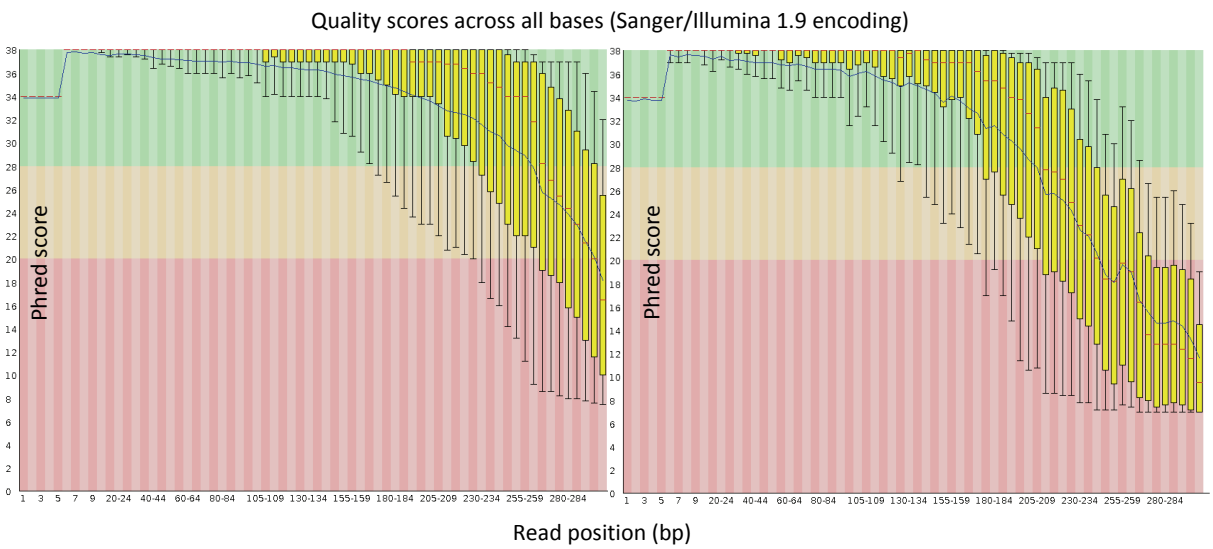

Figure 8.11 Example of Phred quality scores for 2 × 301 bp paired end sequencing of V1-3 16S amplicon libraries on an Illumina MiSeq. To the left is a typical plot for Read1 and to the right is a typical plot for Read2. Read position is plotted versus Phred score. The blue line represents the mean score and the boxplots visualize the score spread. The green area indicates good quality, yellow is reasonable and red is poor. The quality drops towards the end of the read. The quality overview was generated using the FastQC software v0.11.3.

8.4.5.5 OTU clustering

The pre-processed reads are then grouped by sequence identity. These groups are called operational taxonomic units (OTU) and the grouping process is called OTU clustering. The sequence variability in each group may arise from the presence of very closely related strains or because some sequencing errors have been introduced (Huse *et al.,* 2010). Even with a very low error rate of 0.1 % per base, the chance of a perfect 500 bp read is only 61 % (0.999^{500}). As sequencing errors are randomly distributed, sequencing 1,000 reads from a single 16S rRNA gene would result in (1 - 0.61) × 1,000 = 390

different reads. To circumvent this inflated diversity the reads are clustered by similarity into OTUs. When analysing 16S rRNA gene fragments from bacteria, the clustering criterion is a sequence identity of 97 %, which very roughly translates to species level depending on which variable region is used. It is important to understand the often-cited 97 % sequence identity threshold. The 97 % criterion is used for full-length 16S rRNA sequences to state that sequences with less than 97 % similarity belong to different species. The criterion cannot be used in the opposite direction, e.g. sequences that are 97 % similar or more do not necessarily belong to the same species (Janda *et al.,* 2007). The choice of clustering algorithm will alter the resulting OTUs (Edgar, 2013; Flynn *et al*., 2015). Although it is still most popular to use a single threshold for defining OTUs, several algorithms are emerging that take advantage of the error profile in order to use variable clustering thresholds so that maximum resolution can be obtained (Mahé *et al*., 2014). After clustering, the number of reads belonging to each OTU is counted and a representative sequence from the OTU cluster is chosen. This is usually the most frequently observed read in the cluster.

8.4.5.6 Chimera detection and removal

During the PCR step chimeric sequences, which are artificial sequences consisting of multiple different 16S rRNA genes, will be generated. Chimeras arise when two incomplete 16S rRNA gene fragments hybridize and are elongated by the polymerase. Chimeras can artificially increase the diversity of the samples and need to be identified and removed during the bioinformatic processing (Quince *et al*., 2011; Edgar, 2013).

8.4.5.7 Taxonomic classification

The sequences representing the 16S rRNA OTUs are taxonomically classified by comparing them with a database of existing sequences. The classification is very dependent on the algorithm and the database. There are three large universal databases that are commonly used: SILVA (Quast *et al*., 2013), RDP (Cole *et al*., 2014) and Greengenes (McDonald *et al*., 2012) that all attempt to cover all presently known microbes. However, given their broad scope they are not curated for specific ecosystems. The MiDAS database is an expert-curated version of the SILVA database, which provides a genus-level name for most of the abundant species in the activated sludge ecosystem (McIlroy *et al*., 2015). Names are important as they provide a link to other studies and to the literature where functional information may be found.

The algorithm used to compare the obtained OTUs with the database can use different strategies for classification. Some of the common algorithms use different versions of the lowest common ancestor (LCA) approach (Pruesse *et al.,* 2012). This approach takes into account that an OTU sequence can be similar to more than one sequence in the database. If so, the algorithm assigns the lowest shared taxonomy between these database sequences to the OTU.

8.4.5.8 The OTU table

The final result of the bioinformatic processing is an OTU table. The table rows represent the different OTUs and the columns represent each sample in the analysis. The table cells show the count values for the respective OTUs in the respective samples. Each OTU also has a taxonomic classification. The classification is often a delimited (separated by comma or similar) text string containing the classification at each taxonomic rank (kingdom, phylum, class, order, family, genus and species). If a classification at a certain rank is missing, this means robust classification at those levels was impossible. In addition, a fasta file is obtained which contains the DNA sequences of the reference OTUs.

8.4.6 Data analysis

8.4.6.1 Defining the goal of the data analysis

The theoretical possibilities within data analysis of 16S rRNA amplicon are plentiful. However, the attainable scope of the analysis is defined by the experimental design and the variability within the data generated. Therefore, it is highly recommended to perform a pre-study, where the goal is to determine the variation within the type of samples that are to be studied. By sequencing a number of biological replicates, it is possible to make informed decisions regarding the experimental design and the replicates needed to answer the questions posed.

In the following sections, short examples of different types of data analysis will be presented with links to exemplary studies that can serve as further inspiration. For more specific and data-driven examples it is recommended to consult the online documentation of QIIME (Carporaso *et al*., 2010), Mothur (Schloss *et al*., 2009), PhyloSeq (McMurdie and Holmes, 2013), vegan (Oksanen *et al*., 2015) and ampvis (Albertsen *et al*., 2015).

8.4.6.2 Data validation and sanity check

Before starting the main analysis, it is recommended to make a small exploratory analysis to validate the data and detect possible processing errors. This is especially important if the sequencing and bioinformatic processing has been outsourced. The amplicon-based workflow has numerous steps where errors can be introduced. These are typically sample mix-up, cross-contamination, wrong bioinformatic processing or poor sequencing quality. They can often be detected easily by getting an objective overview of the data with general statistics and simple overview plots.

For all the samples the number of raw reads before and after bioinformatic processing should be examined. If a sample generally has few reads this might indicate that there was something wrong with the library preparation. If many reads are lost during bioinformatic processing it might indicate that the data quality is poor. Sample mix-up can be detected by making principal component analysis plots (PCA) based on OTU counts of all the samples. In PCA plots samples with similar microbial communities will cluster together. Hence, use common sense and a visual inspection of whether the samples group as expected. For example, do the replicates group together?

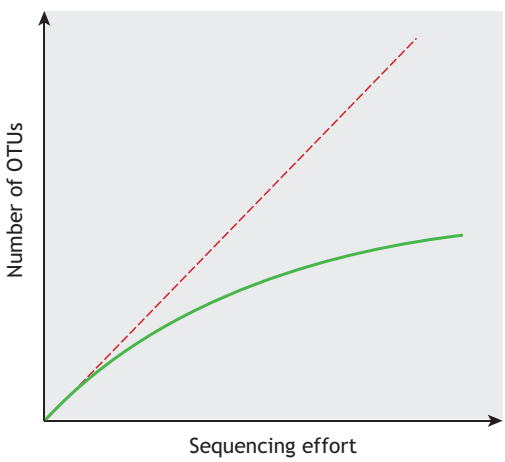

Figure 8.12 Evaluation of the sequencing effort using rarefaction curves. The red line indicates that every new read results in identification of a new OTU. The green line indicates that the discovery of new OTUs decreases with sequencing effort. Hence, the green sample is sufficiently sampled.

When the data has been sanity-checked it is advisable to check if the number of sequenced reads were enough to cover the observed diversity in the samples. This is done by a rarefaction analysis, where a curve is generated that depicts the number of identified OTUs at different sequencing depths through subsampling. The curve should flatten as a function of increasing sequencing depth (Figure 8.12), which indicates that the majority of diversity has been captured (Schloss and Handelsman, 2005).

8.4.6.3 Communities or individual species?

The unit, or perspective, for analysis can roughly be divided into two categories: communities and individual species. In the community-based analysis the parameter for comparison is whether there is an overall difference in community structure or diversity between the samples, while the individual species perspective tries to understand the role and effect of the individual species in the system.

- **The community perspective**

Community-based analysis can be divided into analysis of either alpha-diversity (within samples) or beta-diversity (between samples).

Alpha-diversity analyses are often used to investigate if a particular treatment has an effect on the number of different species observed (also called richness) or on the evenness in abundance of the species in the sample. To compare samples, a single metric is calculated for each sample, which can then be compared across all the samples (Magurran, 2004; Lozupone and Knight, 2008).

Beta-diversity analyses are used to compare the shared diversity between samples, either as the number of shared species or as the amount of shared phylogenetic diversity (Lozupone and Knight, 2008).

- **The species perspective**

While many highly complicated statistical analyses can be performed on 16S rRNA amplicon data, the vast majority of analyses are simply to identify the most abundant bacteria and link them with functional information, e.g. to work out the abundance of nitrifiers or filamentous bacteria in the sample (Figure 8.13). If the MiDAS taxonomy (McIlroy *et al*., 2015) is used for taxonomic assignment, the MiDAS Field Guide (midasfieldguide.org) can be used to manually look up functional information of the particular species in relation to activated sludge. Alternatively, the ampvis R package (Albertsen *et al*., 2015) can be used to link the 16S rRNA amplicon data directly with the functional information in the MiDAS Field Guide.

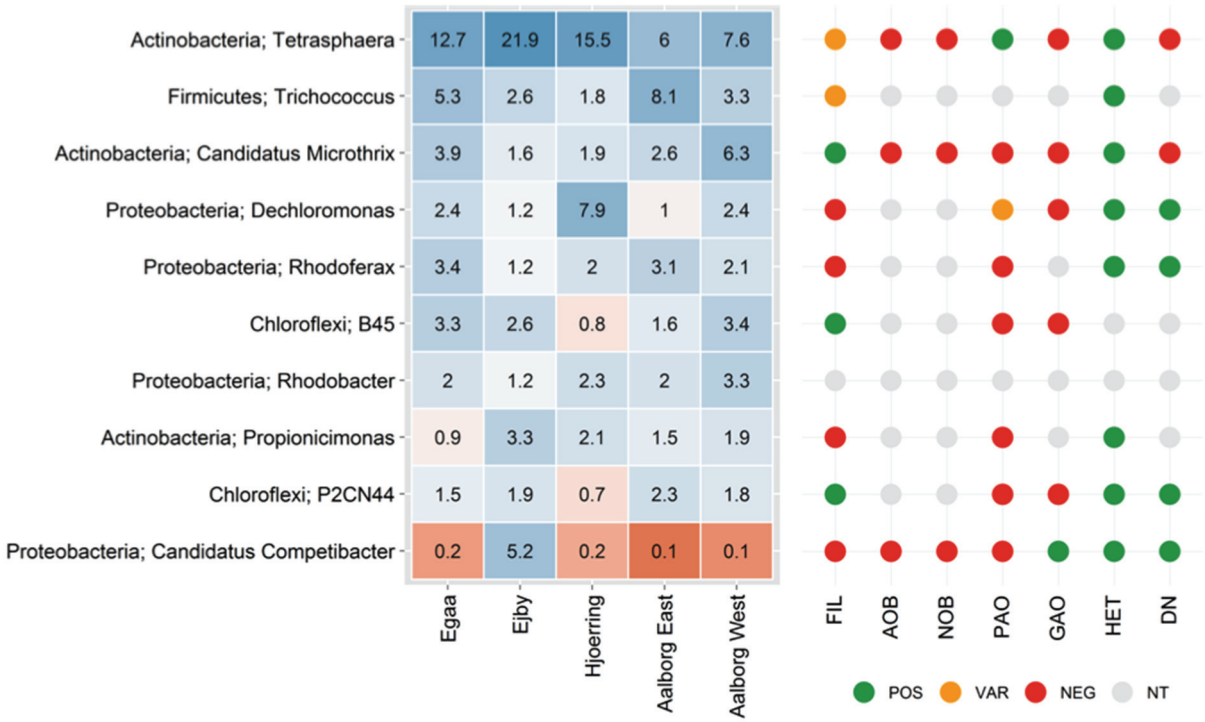

Figure 8.13 Combining species abundance with functional information is a starting point for most analysis. The figure was made through the ampvis R package (Albertsen *et al.*, 2015), which directly links genus names with functional information in the MiDAS Field Guide (www.midasfieldguide.org).

8.4.6.4 Identifying core and transient species

When investigating a specific system it is often of interest to identify the species that are important for the processes in the system and those that might cause problems. A starting point for this analysis is to identify core and transient species (Grime, 1998; Gibson *et al.*, 1999). This analysis needs multiple samples and could either be carried out in a single wastewater treatment plant (time series) or across a number of different wastewater treatment plants.

The traditional definition of core species is those species that are present in all the samples investigated (Saunders *et al.*, 2015). However, given the high sensitivity of the 16S rRNA amplicon sequencing approach, this also includes a high number of very low-abundant species as core species. These species are present in all the samples, but presumably do not contribute significantly to the main ecosystem process. Hence, often a more practical definition of the abundant core community is used, which determines a bacterium to be abundant if it, in eight out of ten samples, is included in the group of abundant species that make up 80 % of the community (Saunders *et al.*, 2015). In systems with a high degree of seeding from influent, it might also be necessary to investigate if the core organisms are simply coming with the influent or if they are actively growing. A reasonable estimate of this can be achieved by analysing the bacterial composition in the influent and activated sludge and combining it with a mass-balance of the system (Saunders *et al.*, 2015).

8.4.6.5 Explorative analysis using multivariate statistics.

The datasets generated by 16S rRNA amplicon sequencing can be extremely large and even with ten samples (each with 1,000s of species) it becomes difficult to get an overview of the data and to identify interesting patterns. Hence, often different varieties of ordination methods are applied, e.g. PCA. These methods can be used to visually group samples according to how similar their microbial communities are and to identify which species are responsible for the observed sample grouping (Figure 8.14).

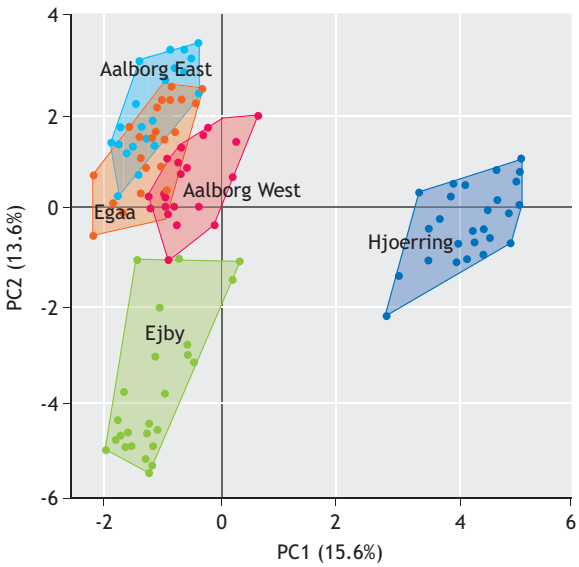

Figure 8.14 Explorative analysis using PCA. Samples are seen to group by five wastewater treatment plants in Denmark. The plot was made using the ampvis R package (Albertsen *et al.*, 2015).

Explorative plots, where the samples are coloured by different environmental variables, are often a good place to start before formal statistical hypothesis testing is applied. However, the specific choice of the ordination method and data transformations is highly dependent on the question posed, the number of samples and the distribution of the abundance. Hence, it is recommended to consult specific literature to understand the applications of different methods (Legendre and Gallagher, 2001; Ramette, 2007; Zuur *et al.*, 2007).

8.4.6.6 Correlation analysis

Another possibility with 16S rRNA amplicon data is to investigate correlations between different species or between species and environmental factors. Are some species co-occurring, are there keystone species that are needed for eco-system functioning or is the abundance of specific species correlated with environmental factors such as temperature?. Due to the number of possible correlations, the results are often displayed in network graphs that depict positive or negative correlations between species and environmental variables (Faust *et al.*, 2012; 2015). However, there are many pitfalls and choices to be made in correlation analysis of 16S rRNA amplicon data. For example, the correlation analysis is often made on time-series data and correlations might be time-lagged. A sudden increase in ammonium concentration might result in an increase of nitrifiers over a period of several weeks.

8.4.6.7 Effect of treatments on individual species

To statistically test the impact of specific treatments on individual species, a number of different statistical methods have been developed that take into account the nature of the count-based amplicon data. Most methods have been modified or directly imported from statistical frameworks developed for analysis of gene expression (mRNAseq) data (Robinson *et al.*, 2010; Love *et al.*, 2014).

8.4.7 General observations

8.4.7.1 A relative analysis

It is important to notice that microbial community analysis using 16S rRNA amplicon sequencing is a relative analysis. This means that the abundance of the reads representing the individual species is given as a percentage of the total. Hence, there is no link from the percentage to the absolute count of cells. If needed, such a link may be established by qPCR (see Section 8.3).

8.4.7.2 Copy number bias

While the 16S rRNA gene is universal among bacteria, different bacteria can have from one to over fifteen copies of the 16S rRNA gene in their genome (Farrelly *et al.*, 1995; Angly *et al.*, 2014). Hence, a bacterial species with ten copies of the 16S rRNA gene would appear ten times more abundantly than a species with a single 16S rRNA gene, if they were present in equal cell counts. In addition, some bacteria have more than one copy of their genome, which will further bias the abundance estimation (Mendell *et al.*, 2008; Pecoraro *et al.*, 2011). Hence, instead of referring to the percentages obtained from 16S rRNA amplicon data as 'abundances' they should be referred to as 'read abundances' to imply the inherent biases.

8.4.7.3 Primer bias

As touched upon earlier, the choice of primer and PCR conditions also has a large effect on the observed microbial community. The 'universal' primers are designed by comparing the sequence of conserved regions of 16S rRNA genes from thousands of different bacteria, and then making a consensus sequence that targets most bacteria. However, it is not possible to design primers that target all bacteria equally well. In addition, biases are often taxon-specific, which means

that whole taxonomic groups are under-represented or even overlooked completely. Many of the new candidate phyla that are being discovered using primer-independent methods, such as metagenomics, show large deviations in the conserved primer sites or even large sequence insertions in the 16S rRNA gene itself, which is the main reason for them having escaped our attention for decades (Brown *et al*., 2015).

8.4.7.4 Standardization

Many attempts have been made to eliminate the biases described above, but with a complex system such as activated sludge, this is probably never going to succeed. Furthermore, it is very difficult to validate whether the biases have been removed, as it is often difficult to make proper controls. At first sight this seems to undermine the analysis method. However, it just sets some limitations to what questions can be answered, and how the experiments should be designed.

The core message is that if the biases are the same for all the samples analysed, 16S rRNA amplicon sequencing is a very powerful tool for relative comparisons and observations of the presence or absence of specific bacteria. To ensure all biases are the same, the samples in one experiment have to be treated the exact same way throughout processing. It is preferable that they are sampled and stored, DNA is extracted, sequencing libraries are prepared and the data is processed in the exact same way.

8.4.7.5 Impact of the method

Despite some limitations, 16S rRNA amplicon sequencing is one of the most essential tools in microbial ecology today. The main reasons are the resolution and the throughput it enables, at which the preceding techniques were far inferior. In 2010, clone libraries and DGGE were still the standard techniques supplemented with FISH for *in situ* identification and quantification. A large study back then would include 100-1,000 16S rRNA sequences divided into 10 samples. Today, for the same cost, hundreds of samples can be analysed with thousands of reads per sample, in a single week. However, this also sets new requirements for the skill of the microbial ecologists in order to handle highly complex experiments and enormous amounts of data.

8.4.8 Protocol: Illumina V1-3 16S rRNA amplicon libraries

This protocol describes how to make Illumina 16S rRNA amplicon sequencing libraries of the variable regions 1-3 of the bacterial 16S rRNA gene. The libraries are suitable for sequencing on the Illumina MiSeq using 600-cycle reagent kits. The total time needed to complete this protocol is approximately 10 h.

8.4.8.1 Apparatus

- A ND-1000 Spectrophotometer (Thermo Scientific) or similar UV-vis spectrophotometers for measuring DNA concentration and estimating DNA purity.
- A Nanodrop user manual (NanoDrop Technologies Inc. 2007).
- A Qubit 2.0 Fluorometer (Life Technologies), Infinite M1000 PRO (Tecan) or similar fluorometer for precise measurement of DNA concentration by use of DNA binding dyes.
- Qubit assay user guides (Thermo Scientific 2015a; 2015b).
- A standard PCR thermocycler with heated lid.
- A magnetic stand for 96-well plates used for DNA purification e.g. MagneSphere Technology Magnetic Separation Stand (Promega) or Magnetic Stand-96 (AM10027, Ambion).
- A tapestation 2200 (Agilent) gel electrophoresis setup for checking sequencing library quality. Alternatively use a conventional gel electrophoresis setup (Green and Sambrook, 2012).
- Tapestation 2200 manuals (Agilent Technologies, 2012; 2013; 2015).
- Pipettes (range 1 µL to 1,000 µL).

8.4.8.2 Materials

PCR Plate Spinner

- DNAse-free tips (10 µL, 300 µL and 1,000 µL).
- DNAse-free tubes (1.5 mL).
- DNAse-free thin-wall, clear, PCR tubes (500 µL).
- 96-well PCR plates (#82006-664, VWR).
- PCR strip caps.
- An OptiPlate-96 Black microplates (Perkin Elmer).
- Nuclease-free H_2O (Qiagen).
- Fluorescent DNA-binding dyes e.g. a Qubit dsDNA HS assay kit (Life Technologies), a Quant-iT dsDNA assay kit, broad range (Life Technologies), a Quant-iT dsDNA assay kit, high sensitivity (Life Technologies).
- A Platinum Taq DNA Polymerase High Fidelity kit (Life Technologies).

- dNTP mix.
- Barcoded V1-3 16S rRNA gene adaptor mixes (5 µM of each forward and reverse adaptor), see Section 8.4.11.
- Agencourt AMPure XP (Beckman Coulter).
- Ethanol, 99 %.
- D1000 Screentape (Agilent) and Genomic DNA Screentapes (Agilent). Alternatively use reagents for conventional gel electrophoresis setup (Green and Sambrook, 2012).

8.4.8.3. Protocol

- **Sample DNA quality control and dilution (2.5 h)**

In this section, the quality of the extracted genomic DNA is checked and the DNA is diluted to a concentration suitable for PCR. See Section 8.4.9 Interpretation and troubleshoot for further information.

1. Fluorescence DNA concentration measurement
   a. Use a Qubit dsDNA BR assay or a Quant-iT dsDNA broad range assay following the protocol recommended by the manufacturer.
   b. Use 2 µL of each sample per reaction.
   c. Perform one measurement per sample.
2. UV-vis quality check (optional)
   a. Use a Nanodrop1000 following the protocol recommended by the manufacturer (Nanodrop Technologies Inc., 2007).
   b. For large sample batches consider measuring a subset of samples (e.g. 8 out of 96).
   c. Initialize and blank the instrument with the same buffer that the samples are eluted in.
   d. Use 2 µL sample per measurement.
   e. Perform one measurement per sample.
3. Gel electrophoresis (optional)
   a. Use a Tapestation 2200 following the protocol recommended by the manufacturer.
   b. Use Genomic Screentapes with reference DNA ladder.
   c. Perform one measurement per sample.
   d. For large sample batches consider measuring a subset of samples (e.g. 7 samples out of 96 + 1 ladder).
4. Sample dilution
   a. Based on fluorescence DNA concentration measurements, calculate how much nuclease-free water is required for each of the samples to dilute them to 5 ng μL^{-1}. The formula used is:

$$\frac{V_{sample} \cdot C_{sample}}{C_{final}} - V_{sample} = V_{H_2O} \rightarrow \frac{5\mu L \cdot C_{sample}}{5\ \frac{ng}{\mu L}} - 5\mu L = V_{H_2O} \qquad \text{Eq. 8.2}$$

   b. Transfer a 5 µL sample to an empty well in a 96-well PCR plate.
   c. Dilute the sample with the calculated amount of nuclease-free water.

▲ Critical step If the sample concentration is 5 ng μL^{-1} or less, use the sample undiluted or discard the sample. If the dilution requires > 150 µL nuclease water, a pre-dilution might be necessary.

   d. Repeat for all the samples.
   e. Seal the plate with PCR strip caps.

■ Pause Point The diluted samples can be stored at -20 °C for at least a month.

- **Library PCR (2.0 h)**

This section concerns the preparation for sequencing libraries by PCR amplification. The template input is genomic DNA (5 ng μL^{-1}) and the output is V1-3 16S rRNA amplicons with a size of approximately 614 bp.

1. Preparation
   a. Library PCR reaction is run in duplicate for each sample.
   b. Remember to include a negative control (nuclease-free H_2O) and positive control (microbial community DNA known to amplify with 16S rRNA PCR).
   c. Note which V1-3 adaptors with unique barcodes are assigned to which samples.
2. Mix the PCR reaction
   a. Prepare the mastermix for (samples + controls) × 2 + 3. Any spare mastermix will make up for loss during pipetting.
   b. Add the reagents in the given order to produce a mastermix.
   c. Transfer 13 µL of the mastermix to the wells of a new 96-well PCR plate.
   d. Add 10 µL of assigned barcoded V1-3 16S rRNA adaptor mix (1 µM) to each well and pipette up and down 10 times to mix. The final adapter concentration is 400 nM.

▲ Critical step High risk of mixing up samples and adaptors. Stay alert.

   e. Add 2 µL of template DNA (total of 10 ng of DNA) and mix by pipetting up and down. The final volume is 25 µL.

▲ Critical step High risk of mixing up samples.

   f. Seal the 96-well PCR plate with PCR strip caps.
   g. Spin the 96-well PCR plate to settle the PCR reaction mix at the bottom of the plate.

Reagents	Final conc. in 25 µL reaction	Volume (µL) for 1 r×n
Nuclease-free water	-	7.65
×10 buffer Platinum High Fidelity	×1	2.5
dNTP (5 mM)	400 uM	2
$MgSO_4$ (50 mM)	1.5 mM	0.75
Platinum Taq DNA Polymerase High Fidelity (5 U μL^{-1})	0.02 U μL^{-1}	0.1
Total volume		13

3. Run PCR incubation
   a. Program the thermocycler with the following program:

Step	Temperature	Time
Denaturation	95 °C	2 min
	30 cycles	
Denaturation	95 °C	20 s
Annealing	56 °C	30 s
Amplification	72 °C	60 s
Amplification	72 °C	5 min
Storage	4 °C	Forever

   b. After the PCR reaction spin down the PCR reaction mix again.
   c. Remove the PCR strip caps and pool the duplicate PCR reactions for each individual sample. The final volume is 50 µL.
   d. After the PCR the samples are now referred to as 'sequencing libraries' or 'short libraries'.
   ■ Pause Point The libraries can be stored at -20 °C for at least a month.

- **Library cleanup (2.0 h)**

This section involves the cleanup of the PCR reactions. The aim here is to remove leftover reagents and possible short (< 200 bp) unspecific PCR products. The output is clean sequencing libraries that only consist of the 16S rRNA amplicons (ca. 614 bp).

1. Preparation
   a. Gently shake the Agencourt AMPure XP bottle to re-suspend the beads, remove the required volume 40 µL beads × [n(samples) + 3] and let it equilibrate to room temperature.
   b. Prepare a fresh solution of 80 % ethanol by transferring 20 mL of ethanol (99 %) to a greiner (or falcon) tube (50 mL) and add 5 mL Nuclease-free water. Mix by inverting the tube.
2. Bind the libraries to the beads
   a. Transfer 40 µL of the bead solution per well in a new 96-well PCR plate corresponding to the number of samples.
   b. Add 50 µL of library to each well with beads and mix by pipetting up and down 10 times.
   ▲ Critical step It is important that the bead to sample ratio is 4:5. If for some reason there is less/more than the 50-µL sample, adjust the bead volume used accordingly.
   c. Incubate for 5 min at room temperature.
3. Wash the bound libraries
   a. Place the 96-well PCR plate on the magnetic stand and wait until the liquid is clear (2-4 min). All the subsequent steps are performed with the plate on the magnetic stand.
   b. Remove and discard as much liquid as possible with a pipette.
   ▲ Critical step Take care not to transfer the beads (brown pellets).
   c. Wash the bead-pellets with 200 µL ethanol (80 %) by gently dispensing it over the beads with a pipette. Let it rest for 30 s and then remove the liquid.
   d. Repeat the above step (3c).
   e. Ensure no excess liquid is left after the washes. If there is, remove it with a 10 µL pipette.
   f. Let the pellets dry for 7 min.
   ▲ Critical step Avoid excessive drying by heating or long drying times as this will make the DNA elution difficult.
4. Elute the library
   a. Remove the 96-well PCR plate with the dried bead pellets from the magnetic stand.
   b. Add 33 µL of nuclease-free water and mix by pipetting up and down 10 times to re-suspend the beads.
   c. Incubate for 2 min at room temperature.
   d. Place the 96-well PCR plate back on the magnetic stand and wait until the liquid clears (1-2 min).
   e. Transfer 30 µL of the liquid to an empty well in a new 96-well plate.
   ▲ Critical step Take care not to transfer any of the beads.
   ■ Pause Point The purified libraries can be stored at -20 °C for at least 6 months (see the paragraph on Storage and transport below).

- **Library quality control (1.5 h)**

This section concerns the DNA concentration measurement of the cleaned libraries and subsequent quality check using gel electrophoresis. It is necessary to confirm that only the target 16S rRNA amplicon is present in the libraries.

1. Fluorescence DNA concentration measurement
   a. Use a Qubit dsDNA High sensitivity assay or a Quant-iT dsDNA High sensitivity assay following the protocol recommended by the manufacturer.
   b. Use 2 µL of each sample per reaction.
   c. Perform one measurement per sample.
2. Gel electrophoresis
   a. Use Tapestation 2200 following the protocol recommended by the manufacturer.
   b. Use D1000 Screentapes with the reference DNA ladder.
   c. Perform one measurement per sample.
   d. For large sample batches consider measuring a subset of samples (e.g. 15 out of 96). Always include the negative and the positive control in the analysis.

- **Library pooling (2.0 h)**

This section concerns the pooling of all the libraries. The aim is to obtain a final single sample containing equimolar concentrations (the same number of molecules) of each library. The volumes are calculated from the DNA concentrations of the libraries obtained above.

1. Calculate the required volume of each sample
   a. Libraries with a concentration of less than 1 ng μL^{-1} should be excluded (either leave out or re-run PCR).
   b. Detect the sample with the lowest concentration and multiply this concentration with 15 µL (e.g. 1 ng μL^{-1} × 15 µL = 15 ng). This is the amount of library wanted from each library.
   c. Calculate the volumes required to obtain the same amount of library for each of the other libraries.
   d. If volumes less than 1 µL are required for some libraries, consider diluting the libraries and re-calculate the required volumes.
2. Pool libraries
   a. Use a new tube (1.5 mL).
   b. Transfer the calculated volume of each sample to the tube.
   c. Mix well after all the samples have been added.
3. Fluorescence DNA concentration measurement
   a. Use a Qubit dsDNA High sensitivity assay following the protocol recommended by the manufacturer.
   b. Use 2 µL of the library pool per reaction.
   c. Perform the measurement in triplicate.
   d. Calculate the average concentration.
   e. Based on the concentration, calculate the nanomolar concentration with the following formula.

$$c_{nM} = \frac{c_{ng/\mu L} \cdot 1{,}000{,}000 \ \frac{\mu L}{L}}{650 \ \frac{g/mol}{bp} \cdot 614 \text{ bp}} \qquad \text{Eq. 8.3}$$

   f. If the concentration is less than 4 nM, concentrate the sample with ampure bead purification (see the section on library cleanup earlier in this chapter). Be sure to use a bead to sample ratio of 4:5 (e.g. if you have 100 µL sample pool you should use 80 µL ampure bead solution). Make sure to calculate how much nuclease-free water is required for elution to get a concentration of 4 nM or above. Also anticipate the loss of up to 50 % of the product. So, if a 2× concentrate is necessary then a 100 µL input library pool should be eluted in 25 µL nuclease-free water). Repeat the DNA concentration measurement of the concentrated pool.

- **Storage and transport**

This section concerns the storage of libraries and the library pool.

1. If storage of the library pool is planned do not dilute it. DNA withstands storage better if kept in its concentrated form (> 5 ng μL^{-1}).
2. For short-term storage or transport (< 14 d), purified library DNA can be kept at ambient temperature.
3. For medium-term storage (< 12 months), the libraries can be kept at -20 °C.
4. For long-term storage (> 12 months), the libraries should be kept at -80 °C.

8.4.9 Interpretation and troubleshooting

8.4.9.1 Sample DNA quality control and dilution

There are three things to consider regarding the input DNA: (*i*) the amount, (*ii*) the quality and (*iii*) potential contaminants. These characteristics are investigated using a fluorescence DNA concentration assay, UV-vis spectrophotometry and gel electrophoresis.

The recommended amount of microbial community DNA for a PCR is 1-100 ng total DNA, where 10 ng (approximately 2×10^6 cells) is often used. If more DNA is used then the risk of amplifying random pieces of DNA is increased and inhibition of the PCR reaction may occur. Random amplicons are problematic, since they result in decreased sequencing data yield and poor quality. Low DNA concentration increases the risk of PCR failure and introduces variance for low abundant members of the community (Kennedy *et al.,* 2014). UV-vis spectrophotometry can be used to measure DNA concentration, but the estimate can be uncertain as it is influenced by reagent contamination, the presence of nucleotides and RNA. Fluorescence-based methods are always better, and UV-vis spectrophotometry should only be used as backup (Li *et al.,* 2014).

The quality of the DNA is important since it influences the effective amount of DNA available for the PCR reaction. If the DNA is heavily degraded (most DNA < 5,000 bp), the risk of the marker genes being broken increases, which makes them unavailable for PCR (Beers *et al.,* 2006; Wilson *et al.,* 1997).

Non-DNA contaminants refer to chemicals or organic molecules introduced from the sample (i.e. humic acids and complex sugars) or during the DNA extraction (i.e. SDS, alcohols, chaotropic salts), and they can inhibit the PCR reaction, reducing efficiency (Wilson *et al.,* 1997). A UV-vis spectrum of clean DNA has a very distinct curve. Anomalies in this curve are indications of contamination and they are often screened for by looking at the ratio of the absorbance at 260 nm and 280 nm (A260/280) and the ratio at 260 nm and 230 nm (A260/230). A260/280 of clean DNA is usually around 1.8, but if it is very different (e.g. ± 0.4) then this might indicate protein contamination or residual reagents such as alcohols from the extraction kit. A260/230 of clean DNA is usually between 2.0-2.2, and if it is very different this might indicate residual carbohydrates or DNA extraction reagents. UV-vis spectrophotometry is a very sensitive method, and the type of buffer and pH can cloud the results (Thermo Scientific, 2015).

DNA contaminants can be DNA from life forms not of interest when targeting bacteria i.e. eukaryotic DNA from fungi or plants. The presence of contaminating DNA reduces the effective amount of target DNA in the sample and reduces PCR efficiency (Tebbe *et al.,* 1993). This cannot be measured beforehand, but clues might be obtained by visually inspecting the biomass before DNA extraction (e.g. is there visible plant material?).

If there is a suspicion of contamination, it is a good idea to perform a test PCR with a few samples to see if it is a problem. Often PCR still works despite contamination, however if the PCR fails, sometimes extra purification steps can be the solution. Remember to use a clean-up method that can handle high molecular genomic DNA (> 10,000 bp) e.g. Agencourt AMPure XP beads. Many column-based purification kits are designed for DNA < 10,000 bp and will therefore remove genomic DNA.

8.4.9.2 Library PCR

When making a PCR the composition of the mastermix and the incubation settings will influence the observed community composition in the final data. In a research setting, it is common to optimize the PCR conditions to increase the amount of target amplicon produced and to reduce the amount of unwanted, random PCR products. However, for 16S rRNA amplicon sequencing, PCR settings should be kept the same for all samples that are to be compared. To ensure consistency it is recommended to use the standard settings from tested protocols. For completeness a short overview of what is usually changed during PCR optimization is given below.

The Mg^{2+} concentration and the annealing temperature will influence how specific the primers are when finding their targets. If they are too stringent a negative bias against certain species will be observed. If the stringency is too low, there will be an increased risk of amplifying random pieces of DNA. The number of PCR cycles determines how much product is produced, and many times apparently failing PCRs can succeed simply by increasing the number of cycles. However, an increased number of PCR cycles is also reported to introduce PCR selection bias (Polz *et al.,* 1998; Kennedy *et al.,* 2014). Also, if too many cycles are run and the primers and other reagents are depleted there is an increased risk of chimeric products (Qiu *et al.,* 2001).

8.4.9.3 Library cleanup

The library cleanup steps removes leftover reagents (primers, nucleotides, polymerase etc.) as well as small (< 200 bp), random DNA pieces. All these contaminants can interfere with the DNA sequencing, either reducing the quality of the data or making the sequencing fail completely.

The purification method applied in the protocol is based on precipitation of the DNA onto small, magnetic plastic balls called SPRI beads (Solid Phase Reversible Immobilization). The principle is not well described in

the literature (DeAngelis *et al.,* 1995), but different blogs present some hypotheses (Hadfield, 2012). Due to its chemical composition, DNA has an overall negative charge and is easily dissolved in water where it electrostatically interacts with the polar water molecules. To precipitate the DNA from the solution NaCl salt is added. Na^+ ions are formed, which shield the negative charges of the DNA and makes it less soluble thereby promoting precipitation. The crowding agent, PEG, increases the effective concentration of Na^+ and drastically increases the shielding effect. When the DNA precipitates it prefers the surface of the SPRI beads. This preference is counter intuitive since the surface of the beads is negatively charged and therefore should repel DNA in negatively charged or shielded state. The topic is much debated and no clear explanation exist, however some propose the DNA/SPRI bead interaction is mediated by a layer of water or Na+ ions coating the SPRI beads. Hence, by mixing dissolved DNA and beads in a solution containing crowding agents the result will be that the DNA precipitates on the beads. The crowding solution can be removed and the beads with the DNA can be washed using 80 % ethanol. The composition of the washing solution is optimized to dissolve small contaminants such as salt, nucleotides etc. but ensures that the DNA is not re-dissolved. After washing, the DNA is re-dissolved in nuclease-free water or buffer and is ready to be used.

Small DNA molecules are more likely to stay in solution compared to larger DNA molecules, when adding a certain amount of crowding agent and salt. This means that small pieces of DNA can be removed by adding the correct amount of crowding solution. In this protocol, 40 µL Agencourt AMPure XP bead solution (beads + crowding solution) is used for 50 µL sample. This bead to sample ratio of 4:5 favours the binding of DNA > 200 bp to the beads and smaller pieces remain in the solution and are subsequently removed. Care has to be taken when dispensing the sample or bead solution. If the bead/sample ratio is < 0.5, the 16S rRNA amplicon is also removed. If the ratio is > 1.0 then the contamination will not be removed properly.

When drying the residual ethanol from the beads after washing, take care not to overdo it. Do not use heat or dry them for too long. If the bead pellet becomes too dry (many cracks in the pellet), re-dissolving the DNA will be difficult and the product will be lost.

Other purification methods can be used instead of beads, such as column-based methods e.g. QIAquick PCR Purification Kit (Qiagen). The overall purification principles are similar.

8.4.9.4 Library quality control

The DNA concentration is measured to provide the basis for library pooling. The concentration measurement does not need to be very precise (± 20 % is appropriate), hence only one measurement is made.

Gel electrophoresis is run to ensure that all the contaminating DNA (primers and randomly amplified DNA) is removed and only the target 16S rRNA amplicon is present (614 bp). Generally, only a subset of all the samples is run since gel electrophoresis is expensive and time consuming, and a random subset should give an indication of the purification success. When picking the samples for the subset, always include the negative and the positive samples. Also, based on the concentration measurement, include the samples that have low concentration, to see if there is V1-3 16S rRNA amplicon product present. If no V1-3 16S rRNA amplicon product is present then re-run the PCR on these samples.

The V1-3 16S rRNA amplicon product has an average size of 614 bp, however the size can vary as much as ± 100 bp between the different species. For bacterial communities with many different species this usually manifests itself as a broad peak with a peak maximum around 614 bp when using Tapestation electrophoresis. If a community with only a few members is analyzed, multiple peaks between 500 and 700 bp might show up on the Tapestation electropherogram.

There should be no DNA pieces left below 300 bp. If there are, consider revising the purification step and redo the purification. If the contaminant make up < 1 % of the total DNA in the sample, sequencing should still produce a useful result and redoing the purification can be omitted. Rarely, unknown contamination above 700 bp can be seen. If this is present, try to redo the PCR. If it remains, try sequencing the library anyway. It will be possible to filter out the contaminant using bioinformatics.

The positive sample should show a clear product around 614 bp. The negative sample should show no product, or a faint product that is < 1 ‰ of the total DNA of the other samples. It is often difficult to completely avoid contamination, but as long as the level of contamination is relatively low it is acceptable. If the negative control has a relatively high level of contamination, it is likely that all the samples are

contaminated. Revise your PCR setup and redo the PCR reactions for all the samples, possibly with new batches of reagent.

8.4.9.5 Library pooling

Equimolar library pooling is performed to ensure that each sample gets equal amounts of data during sequencing. Errors in pooling will have a direct effect on the amount of data obtained. If mistakes are made in pooling, start again if possible.

8.4.9.6 Pool quality control and dilution

The concentration measurement of the library pool is very important since this is used to determine how much of the library is to be loaded on the sequencer. If the concentration is measured as higher than it really is, less data will be obtained from the sequencing run. If the measured concentration is lower than it actually is, the sequencer can be overloaded, which results in poor data quality or the sequencer crashing completely.

Therefore, the library pool should be measured in triplicate and the concentrations averaged. Be sure to calibrate the fluorometer before use with reference samples that are known not to be contaminated or degraded.

8.4.9.7 Storage

The DNA should be stored in ultra-pure water or TE buffer (pH 7-8). There have been reports of degradation of DNA after long-term storage in ultrapure water, which is probably due to a fragile pH balance, as no buffer is present. Pure, dissolved DNA is very stable even at room temperature and can be safely stored or transported at ambient temperature for a short time. For medium or long-term storage, freezing the DNA at -20 or -80 °C is recommended (Smith and Monn, 2005). Avoid repeated freeze-thaw cycles of the DNA as this degrades the DNA. If a sample is to be used multiple times, prepare aliquots.

8.4.10 Protocol: Illumina V1-3 16S amplicon sequencing

This protocol describes important protocol additions when preparing sequencing of Illumina V1-3 16S rRNA Amplicon Libraries on the Illumina MiSeq. For details about the steps following the standard procedure, refer to the MiSeq System User Guide (Illumina Inc. 2014b). The time needed for the execution of the protocol is approximately 3.5 h, however sequencing requires 56 h.

8.4.10.1 Apparatus

For the execution of the protocol one needs the following apparatus:

- A MiSeq (Illumina).
- Micropipettes (range 1 uL to 1,000 uL).
- A MiSeq System User Guide (Illumina Inc. 2014b).
- An Illumina Experimental Manager v1.9 (illumina.com).

8.4.10.2 Reagents

For the execution of the protocol one needs the following reagents:

- Ice.
- A MiSeq Reagent kit v3, 600 cycles (Illumina). Includes a reagent cartridge and HT1 buffer.
- 2 M NaOH, molecular grade.
- Sequencing primers (Read1, Read2, and Index), see Section 8.4.11.
- DNAse-free tips (10 µL, 300 µL and 1,000 µL).
- DNAse-free tubes (1.5 mL).
- Nuclease-free water (Qiagen).
- PhiX control library v3, 10 nM (Illumina).
- Ethanol, 70 % (molecular grade).
- Laboratory wipes, lint-free.
- Microscope lens cleaning wipes.

8.4.10.3 Protocol

- **Prepare MiSeq (2.0 h)**

This section describes the thawing of reagents, and preparation of the MiSeq instrument and the Sample sheet file.

1. Instrument washing
   a. According to the recommendations in the User Guide.
2. Reboot the MiSeq to reset the memory
   a. Under *Manage Instrument* in the MiSeq control software, press *Reboot* and wait (this might take 10 min).
   b. The MiSeq control software will start initializing the instrument after reboot. When done the *Control interface* will appear.
3. Thaw the reagents
   a. Place the reagent cartridge and HT1 buffer in the water bath at room temperature for 1 h. After thawing store at 4 °C until use.

   ■ Pause Point The thawed MiSeq reagent cartridge can be stored for a week at 4 °C.

   b. Reagent cartridge inspection: invert the cartridge ten times, inspect for precipitates, and then tap the cartridge on the table to remove any bubbles.
   c. Thaw the sequencing primers (Read1, Read2 and index) and 2 M NaOH at room temperature. Place the primers on ice after thawing, and leave 2 M NaOH at room temperature until use.
4. Prepare the Sample Sheet
   a. Open a MiSeq 'SampleSheet.csv' template in Notepad++ or another simple text editor.
   ▲ Critical step The 'SampleSheet.csv' is a comma-separated value text file (.csv) but should not be opened in Microsoft Excel! Excel might corrupt the formatting of the file.
   b. Change the project and sample-specific info: [Header] Investigator, Project name, Experiment Name, Date; [Data] Sample_ID, Sample_Name, index and index2.
   ▲ Critical step Important information is: [Header] Chemistry, [Reads] and [Data] index and index2 columns. This information has significant impact on how the run is performed. All the other information can be changed later if it is wrong.
   c. After filling out the sample sheet, check the integrity of the 'SampleSheet.csv' by loading it into the Illumina 'Experimental Manager'. If the sheet can be loaded then it should be compatible with the MiSeq.
   d. Transfer 'SampleSheet.csv' to MiSeq with a USB-stick.

- **Prepare sequencing libraries (1.0 h)**

This section describes the denaturation and dilution of the sequencing libraries.

1. Sample overview
   a. Control library: PhiX control library v3, 10 nM.
   b. Library pool (up to 400 samples), > 4 nM.
   c. Follow the following steps for both the PhiX control library and the library pool.
2. Thaw the libraries and store on ice.
3. Dilute the sequencing libraries to 4 nM with nuclease-free water.
4. Prepare 0.1 M NaOH solution.
   a. 475 µL DNA H_2O and 25 µL 2 M NaOH
5. Denature the sequencing libraries
   a. Mix 5 µL library + 5 µL 0.1 M NaOH. The final library concentration is 2 nM.
   b. Pipette up and down 10 times to mix.
   c. Incubate for 5 min at room temperature.
6. Dilute the denatured libraries (2 nM) to 20 pM.
   a. Mix 10 µL denatured library with 990 pre-chilled HT1. The concentration is 20 pM.
7. Mix the PhiX library (20 pM) with the library pool (20 pM) so they make up 20 % and 80 %, respectively, of the final mix.
   a. Mix 120 µL PhiX library with 480 µL of the library pool.
   b. Place the PhiX/pool mix on ice until use.
   ■ Pause Point Diluted libraries can be stored at -20 °C for up to a month. Longer storing times might result in reduced concentration, and consequently reduced sequencing output.

- **Load the sample and primers on the reagent cartridge**

1. Adding custom-sequencing primers to the reagent cartridge.
   a. Primer destinations:
   Read1 = well 12.
   Index = well 13.
   Read2 = well 14.
   ▲ Critical step Well numbering on the reagent cartridge can be confusing. Take the time to make sure the correct wells are used.
   b. For each primer: Pierce the tinfoil covering the target well with the tip of the 1,000 µL pipette, and then transfer 100 µL of the well content to a spin tube. Add 3.4 µL of the respective primer and mix well. Transfer the solution back to the well it originated from and mix. Repeat for all the primers.
2. Adding the sample to the reagent cartridge
   a. Pierce well no. 17 with a tip and add 600 µL PhiX/pool mix.

- **Sequencing (0.5 + 56 h)**

This section describes the startup of the sequencing run.

1. Press *Sequence* on the MiSeq Control software interface and follow the instructions in the MiSeq System User Guide for preparing/loading the flow cell, loading the reagent cartridge, referring to the sample sheet and starting the sequencing run.
2. During sequencing the progress can be monitored by opening the run folder in the Sequence analysis Viewer software and looking at the run.

8.4.10.4. Interpretation and troubleshooting

- **Prepare MiSeq and metadata**

Making sure a wash has been performed on the MiSeq prior to sequencing is important. Cross-contamination between runs can be a problem. Studies have shown that there is a bleed-over of samples from run-to-run. Usually

the bleed-over will not impact 16S amplicon analysis from activated sludge, and therefore the default *Post Run Wash* is acceptable between washes. However, for delicate samples consider performing one or two *Maintenance Washes* in between runs and/or a dedicated wash of the sample line. These washes are more thorough and leftover contamination is diluted. Otherwise follow the wash instructions in the MiSeq System user guide.

```
[Header]
IEMFileVersion,4
Investigator Name,SMK
Project Name, DNASense-NDJ-RHK
Experiment Name,J214
Date,08/11/15
Workflow,GenerateFASTQ
Application,FASTQ Only
Assay,TruSeq HT
Description
Chemistry,Amplicon

[Reads]
301
301

[Settings]

[Data]
Sample_ID,Sample_Name,index   ,index2  ,Description
LIB-CP034,16SAMP-6380,ACGTGTAC,GAGCTCTC,bV13fr-337
```

Figure 8.15 SampleSheet example. The SampleSheet.csv file has been opened in Notepad++. The section titles are marked by [ ].

A MiSeq reboot helps reset the instrument computer, and makes it less likely to crash during sequencing.

Take care when thawing the MiSeq reagents. The quality of the reagents impacts the sequencing quality, especially in the read 3' ends. Experience shows that sequencing can still succeed with reagents that have been stored for 24 h at room temperature but no guarantee is given.

The 'SampleSheet.csv' file tells the MiSeq instrument how the sequencing run should be performed, how to demultiplex the samples and some basic metadata related to the samples. The metadata sheet cannot be prepared by the Illumina Experiment Manager when using adaptors and barcodes not purchased from Illumina. The 'SampleSheet.csv' file has to be prepared in a text editor such as Notepad++. Find a template online or create your own from the Illumina Experiment Manager. For detailed information about the SampleSheet.csv see Illumina Inc. (2013).

Explanation of critical information in the 'SampleSheet.csv' file:

Chemistry, Amplicon
The amplicon setting allows use of two indexes (index1 and index2)
[Reads]
301
301
Orders the MiSeq to perform paired end sequencing, where each read is 300 bp long.

Sample_ID, Sample_Name, index, index2, Description
Sample_ID column: The ID of your samples.
Sample_Name column: The name of the samples. The data output will be called by these names.
Index column: The first read barcode of your samples. Type in the sequence of the barcode. The presence and length of a barcode lets the sequencer know you want to sequence the barcode.
Index2 column: Similar to above.
Description column: Notes about samples.

- **Prepare sequencing libraries**
  In this step the sequencing libraries (the PhiX control library and the library pool) are denatured with high pH (NaOH) to obtain the library amplicons in single-stranded form. The library amplicons need to be single-stranded in order for them to be captured by the MiSeq.

The pH is very important, and both too high and too low pH will prevent the library amplicons from being captured. Preferably use 2 M molecular grade NaOH.

After denaturation the libraries are diluted. The concentration is extremely important as it directly determines the amount of data produced. 20 pM produces approximately 18-25 million reads. If the concentration is lower or higher the output will be proportionally lower or higher. Outside the range of 2-25 pM there is a great risk of the sequencing run crashing completely.

The PhiX control library is used to estimate the error rate and to assist in calibration of the instrument during sequencing. This is especially important for library pools that have low complexity. Low complexity means the sequences analysed have similar sequence composition. As the 16S rRNA gene has conserved regions this is the case for 16S amplicon libraries.

- **Load the sample and primers on the reagent cartridge**

Adding sequencing primers to the MiSeq reagent cartridge is needed when sequencing libraries prepared

with adaptors not bought from Illumina. The primers initiate the reading of Read1, Read2 and Index1. The run will fail if they are not present.

- **Sequencing**

After the sequencing has started the first stats will appear after approximately 4 h. Yield per sample can be obtained after 32 h and the run will complete in 56 h.

The cluster density reveals the estimated output of the run. An average run prepared by the above protocols will produce between 700-1,000 k mm^{-2}, which produces 17-25 million PE reads.

Cluster PF reveals how much of the data meets basic internal quality requirements. With sequencing of V1-3 16S rRNA Amplicon libraries, a value of > 90 % is standard. This might change a little during sequencing.

% ≥ Q30 reveals the number of bases in the whole run expected to have a quality score above Q30. This changes a lot during sequencing. For read1 the average is usually > 70 % and for Read2 the average is usually > 60 %.

The error rate explains the actual measured error rate in the sequencing of the PhiX control library. During sequencing the MiSeq recognizes library amplicons from PhiX and compares them to a reference PhiX genome, to detect and measure sequencing mistakes. In simple terms, the quality score is the theoretical quality where the error rate is the empirical quality.

The aligned statistics explains how much the PhiX control library makes up of the whole run. It should be close to 20 % if the above protocol was followed.

8.4.11 Design of Illumina 16S amplicon sequencing adaptors

The section describes the adaptor/primer designs and the functions of the different parts.

For preparing 16S rRNA amplicon libraries, so-called adaptors are used. They come in pairs consisting of a forward and a reverse adaptor. Each adaptor consists of an adaptor part and a primer part. During the library PCR, the primer parts of the forward and reverse adaptors are used to specifically amplify the variable region 1 to 3 (V1-3) of the 16S rRNA gene (Figure 8.7). The final library amplicons contain the V1-3 16S rRNA sequences and the adaptor parts. The primers are adopted from the Human microbiome project (HMP, 2010) and are called 27F and 534R. The adaptor parts are adopted from Caporaso *et al.* (2011; 2012) and from Illumina Inc. (2014a).

During sequencing, sequencing primers attach to the adaptors and initiate sequencing (Figure 8.16).

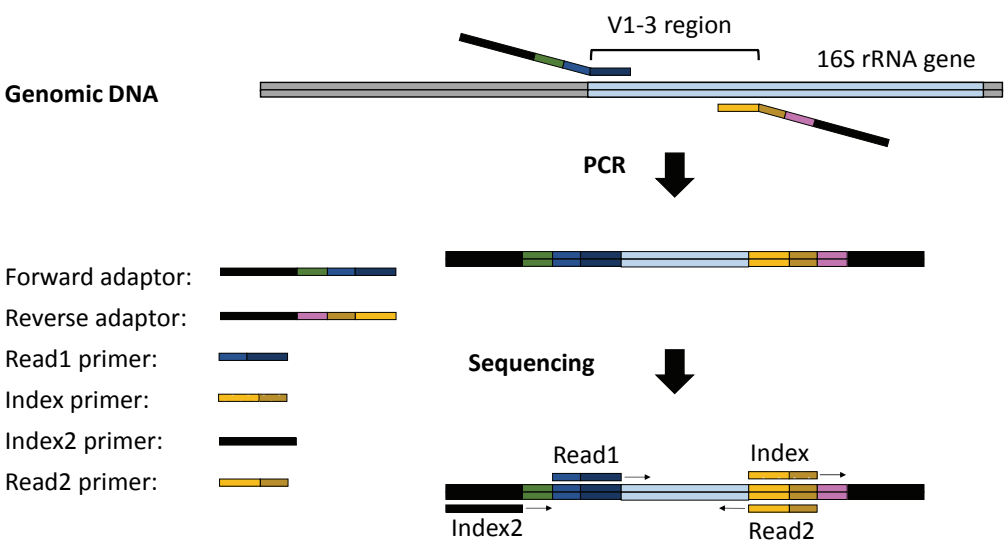

Figure 8.16 Conceptual overview of the oligos used in 16S amplicon sequencing. The adaptors are introduced during PCR and the rest of the oligos are used for sequencing. The coloured parts of the adaptors/primers represent different functional sequences.

In total, four sequences are read: Read1 and Read2, which span the V1-3 16S rRNA gene part and make up the sequences of the paired end reads, and index and index2, which make up the barcoding part which is used to identify what sample each library amplicon originates from. The barcoding part is processed directly on the MiSeq and is not a part of the data output.

The adaptors and primers are synthesized DNA oligos, which can be ordered from any large reagent company. The sequences of the respective DNA oligos can be found in Figure 8.17. The different parts of the oligos have names and a designated role in the library preparation and/or sequencing. The index part of the adaptors (NNNNNNNN) is different for each sample that is to be sequenced. For example, if 96 samples are to be sequenced then there will usually be 8 forward adaptors and 12 reverse adaptors, all of which have a unique index. Hence, a total of 96 unique combinations of the forward and reverse adaptors can be obtained.

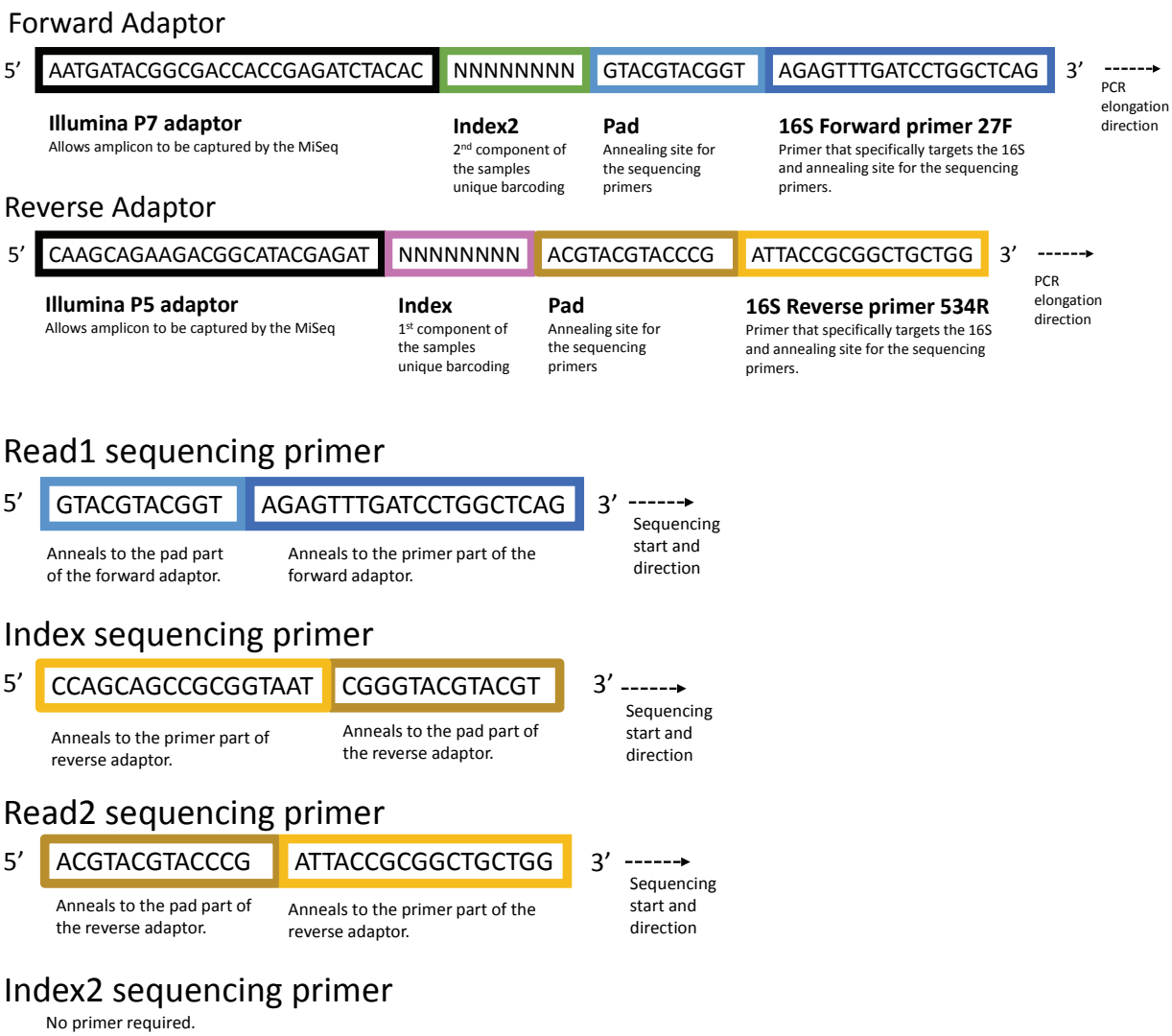

Figure 8.17 The adaptors and primers are synthesized DNA oligos. The sequences of the respective DNA oligos are shown. The different parts of the oligos have names and a designated role in the library preparation and/or sequencing. The index part of the adaptors (NNNNNNNN) is different for each sample to be sequenced.

8.5 OTHER METHODS

Fluorescence *in situ* hybridization (FISH) is an independent method to visualize microorganisms by fluorescently labelled 16S rRNA-targeted oligonucleotide probes. FISH is in itself a very powerful method for microbial community analysis, but it is also an excellent supporting technique for validating results from amplicon sequencing. FISH is described in detail in Chapter 7.

Advanced sequencing-based techniques, such as metagenomics, metatranscriptomics and metaproteomics, exist that enable the culture-independent analysis of functions in microbial communities. These techniques are starting to be applied to activated-sludge systems by the research community. However, these techniques are complicated and will not be mature enough to be valuable for widespread use in the foreseeable future. Therefore, the description of these techniques will be limited to a basic overview.

Metagenomics, or environmental genomics (Wooley *et al.,* 2010), is the study of all the community DNA recovered directly from environmental samples. Figure 8.18 shows the workflow.

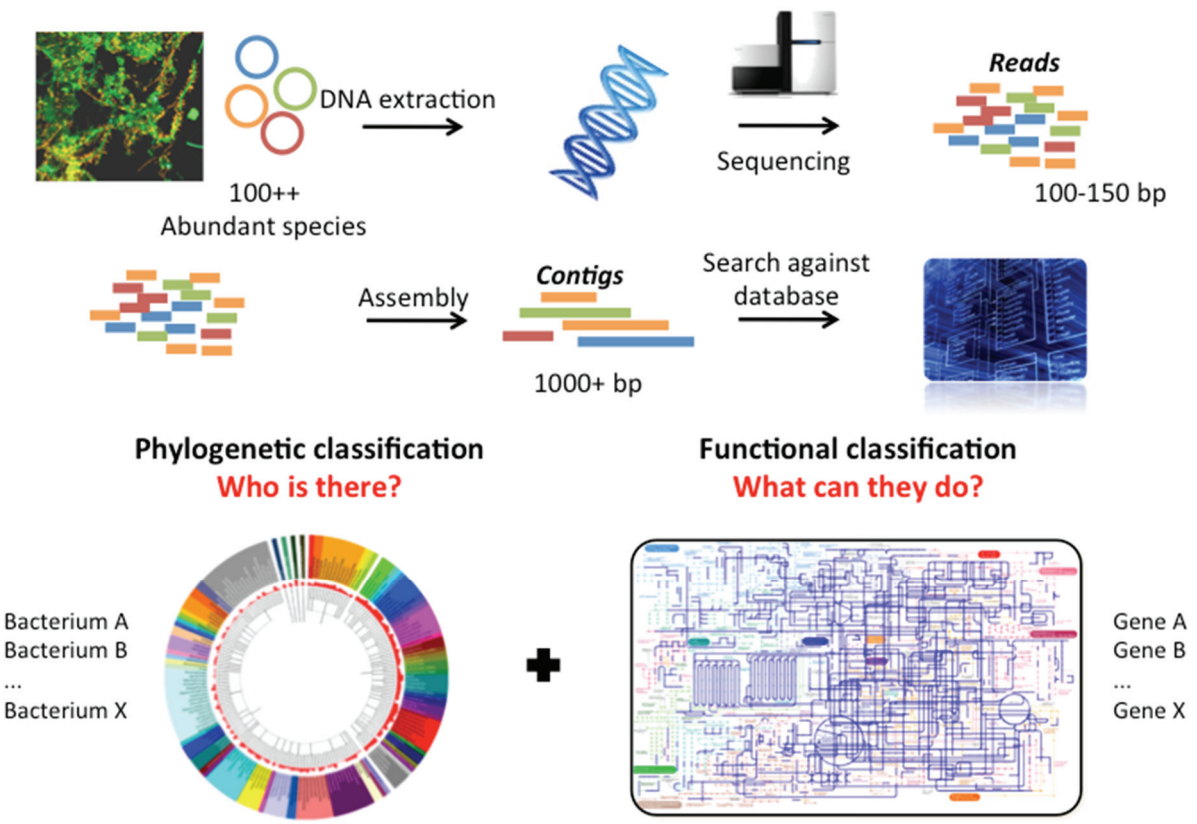

Figure 8.18 Overview of the workflow in metagenomics and the potential outcome.

Community DNA is extracted and purified before it is sequenced, here by the Illumina platform, providing short reads. The reads are assembled into progressively longer contiguous sequences (contigs). These can be applied to obtain information about the taxonomy and the function of the community members by comparing the sequence information to reference databases. However, the resulting information is unfortunately very biased. Proper taxonomic classification of DNA sequences requires closely related, annotated genomes in the

reference databases. Only relatively few genomes are available in the reference databases today, so often the resulting classification is unreliable. Furthermore, a detailed functional characterization of the genes is also difficult, again due to incomplete reference databases (Albertsen *et al.,* 2013).

Metatranscriptomics and metaproteomics describe the complete set of expressed genes and proteins in the microorganisms in an environmental sample. These techniques are rarely applied in systems related to wastewater treatment for several reasons. The main reason is the lack of good reference genomes that are a prerequisite for a reliable study of the expressed genes and proteins. Furthermore, for metaproteomics, it is difficult to obtain sufficient protein extraction efficiency (Seifert *et al.,* 2013; Jensen *et al.,* 2014).

Microarray technologies were regarded as very promising for taxonomic and functional analyses of communities, but the fast development in sequencing technologies has largely outpaced these techniques.

References

Agilent Technologies (2012). Agilent 2200 TapeStation User Manual edition 8. Manual Part number G2966-90001.

Agilent Technologies (2013). Agilent D1000 ScreenTape System Quick Guide edition 10. Manual Part number G2964-90032 Rev. B.

Agilent Technologies (2015). Agilent Genomic DNA ScreenTape System Quick Guide edition 6. Manual Part number G2964-90040 Rev. D.

Albertsen, M, Karst, S.M., Ziegler, A.S., Kirkegaard R.H., Nielsen P.H. (2015). Back to Basics – The influence of DNA extraction and primer choice on phylogenetic analysis of activated sludge communities. *PLoS ONE* 10:e0132783.

Albertsen M., Saunders, A.M., Nielsen K.L. and Nielsen P.H. (2013): Metagenomes obtained by "deep sequencing" - what do they tell about the EBPR communities? *Wat. Sci. Tech*., 68: 1959-1968.

Angly, F.E., Dennis, P.G., Skarshewski, A., Vanwonterghem, I., Hugenholtz, P., and Tyson, G.W. (2014). CopyRighter: a rapid tool for improving the accuracy of microbial community profiles through lineage-specific gene copy number correction. *Microbiome*, 2, 11.

Ashelford, K.E., Chuzhanova, N.A., Fry, J.C., Jones, A. J., and Weightman, A.J. (2005). At least 1 in 20 16S rRNA sequence records currently held in public repositories is estimated to contain substantial anomalies. *Appl. Environ. Microbiol.* 71: 7724-7736.

Basu, C. (2015). PCR primer design (New York: Humana Pr).

Beers, E.H. Van, Joosse, S.A., Ligtenberg, M.J., Fles, R., Hogervorst, F.B.L., Verhoef, S., and Nederlof, P.M. (2006). A multiplex PCR predictor for aCGH success of FFPE samples. *British J. Cancer* 94: 333-337.

Beller, H.R., Kane, S.R., Legler, T.C., and Alvarez, P.J.J. (2002). A real-time polymerase chain reaction method for monitoring anaerobic, hydrocarbon-degrading bacteria based on a catabolic gene. *Environ. Sci. & Technol.* 36: 3977–3984.

Bessetti, J. (2007). An introduction to PCR inhibitors. *J. Microbiol. Meth. 28*: 159-167.

Bollet C, Gevaudan M.J., de Lamballerie X., Zandotti C., de Micco P. 1991. A simple method for the isolation of chromosomal DNA from Gram positive or acid-fast bacteria. *Nucleic Acids Research* 19:1955.

Brown, C.T., Hug, L.A, Thomas, B.C., Sharon, I., Castelle, C.J., Singh, A., … Banfield, J. F. (2015). Unusual biology across a group comprising more than 15 % of domain Bacteria. *Nature*, 523: 208-211.

Bru, D., Sarr, A., and Philippot, L. (2007). Relative abundances of proteobacterial membrane-bound and periplasmic nitrate reductases in selected environments. *Appl. Environ. Microbiol.* 73: 5971-5974.

Brzoska, A.J., and Hassan, K.A. (2014). Quantitative PCR for detection of mRNA and gDNA in environmental isolates. In Environmental Microbiology, I.T. Paulsen, and A.J. Holmes, eds. (Totowa, NJ: Humana Press), pp. 25-42.

Bürgmann H, Pesaro M, Widmer F, Zeyer J. (2001). A strategy for optimizing quality and quantity of DNA extracted from soil. *J Microbiol. Methods* 45: 7-20.

Bustin, S.A., Benes, V., Garson, J.A., Hellemans, J., Huggett, J., Kubista, M., Mueller, R., Nolan, T., Pfaffl, M.W., Shipley, G.L., *et al*. (2009). The MIQE guidelines: minimum information for publication of quantitative real-time PCR experiments. *Clin. Chem.* 55: 611-622.

Caporaso, J.G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F.D., Costello, E.K., … Knight, R. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nature Methods*. 7: 335-6.

Caporaso, J.G., Lauber, C.L., Walters, W.A, Berg-Lyons, D., Huntley, J., Fierer, N., … Knight, R. (2012). Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *ISME Journal* 6: 1621-1624.

Caporaso, J.G., Lauber, C.L., Walters, W.A., Berg-Lyons, D., Lozupone, C.A., Turnbaugh, P.J., … Knight, R. (2011). Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *PNAS* 108. Suppl 1, 4516-4522.

Cole, J.R., Wang, Q., Fish, J. a., Chai, B., McGarrell, D.M., Sun, Y., … Tiedje, J.M. (2014). Ribosomal Database Project: Data and tools for high throughput rRNA analysis. *Nucleic Acids Research*, 42: 633–642.

Cock, P.J.A., Fields, C.J., Goto, N., Heuer, M.L., and Rice, P.M. (2010). The Sanger FASTQ file format for sequences with quality scores, and the Solexa / Illumina FASTQ variants, *Nucleic Acids Research* 38: 1767-1771.

Daims, H., Lebedeva, E.V., Pjevac, P., Han, P., Herbold, C., Albertsen, M., … and Wagner, M. (2015). Complete nitrification by *Nitrospira* bacteria. *Nature* Doi: 10.1038/nature16461.

DeAngelis M.M., Wang D.G., Hawkins T.L. (1995). Solid-phase reversible immobilization for the isolation of PCR products. *Nucleic Acids Research* 23: 4742-4743.

Dominiak, D.M., Nielsen, J.L., and Nielsen, P.H. (2011). Extracellular DNA is abundant and important for microcolony strength in mixed microbial biofilms. *Environ. Microbiol.* 13: 710-721.

Dueholm, M.S., Albertsen, M., D'Imperio, S., Tale, V.P., Lewis, D., Nielsen, P.H., and Nielsen, J.L. (2014). Complete genome sequences of *Pseudomonas monteilii* SB3078 and SB3101, two benzene-, toluene-, and ethylbenzene-degrading bacteria used for bioaugmentation. *Genome Announc.* 2: 524-14.

Dueholm, M.S., Marques, I.G., Karst, S.M., D'Imperio, S., Tale, V.P., Lewis, D., Nielsen, P.H., and Nielsen, J.L. (2015). Survival and activity of individual bioaugmentation strains. *Biores. Technol.* 186: 192-199.

Edgar, R. C. (2013). UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nature Methods,* 10: 996-8.

El Fantroussi, S., and Agathos, S.N. (2005). Is bioaugmentation a feasible strategy for pollutant removal and site remediation? *Curr Opin Microbiol* 8: 268-275.

Farrelly, V., Rainey, F.A, and Stackebrandt, E. (1995). Effect of genome size and rrn gene copy number on PCR amplification of 16S rRNA genes from a mixture of bacterial species. *Appl. Environ. Microbiol.* 67: 2798-2801.

Faust, K., and Raes, J. (2012). Microbial interactions: from networks to models. *Nature Reviews. Microbiology*, 10, 538–50.

Faust, K., Lahti, L., Gonze, D., de Vos, W. M., and Raes, J. (2015). Metagenomics meets time series analysis: unraveling microbial community dynamics. *Current Opin Microbiol.*, 25: 56-66.

Filippidou, S., Junier, T., Wunderlin, T., Lo, C-C., Li, P-E., Chain, P.S., Junier, P. (2015). Under-detection of endospore-forming Firmicutes in metagenomic data. *Comput. and Structural Biotechnol. J.* 13: 299-306.

Fisher, S., Barry, A., Abreu, J. (2011). A scalable, fully automated process for construction of sequence-ready human exome targeted capture libraries. *Genome Biology*, 12, R1 doi:10.1186/gb-2011-12-1-r1

Flynn, J. M., Brown, E. a., Chain, F. J. J., MacIsaac, H. J., and Cristescu, M. E. (2015). Toward accurate molecular identification of species in complex environmental samples: testing the performance of sequence filtering and clustering methods. *Ecol. Evol.*, 5: 2252-2266.

Frølund, B., Palmgren, R., Keiding, K., and Nielsen, P.H. (1996). Extraction of extracellular polymers from activated sludge using a cation exchange resin. *Wat. Res.* 30: 1749-1758.

Ge, S., Wang, S., Yang, X., Qiu, S., Li, B., and Peng, Y. (2015). Detection of nitrifiers and evaluation of partial nitrification for wastewater treatment: A review. *Chemosphere,* 140: 85-98.

Gibson, D.J., Ely, J.S., Collins, S.L. (1999). The core-satellite species hypothesis provides a theoretical basis for Grime's classification of dominant, subordinate, and transient species. *J. Ecol*. 87: 1064-1067.

Gilbert, E.M. ,S. Agrawal, S.M. Karst, H. Horn, P.H. Nielsen, S. Lackner (2014). Low temperature partial nitritation/anammox in a moving bed biofilm reactor treating low strength wastewater. *Environ. Sci. Technol.,* 48: 8784-8792.

Green, M.R., Sambrook, J. (2012). Molecular Cloning: A Laboratory Manual (Fourth Edition). Cold Spring Harbor Laboratory Press. Cold Spring Harbor, New York.

Grime, J. (1998). Benefits of plant diversity to ecosystems: immediate, filter and founder effects. *J. Ecol.* 86: 902-910.

Guillén-Navarro, K., Herrera-López, D., López-Chávez, M.Y., Cancino-Gómez, M., Reyes-Reyes, A.L. (2015). Assessment of methods to recover DNA from bacteria, fungi and archaea in complex environmental samples. *Folia Microbiol*. 60: 551-558.

Guo, F., Zhang, T. (2013). Biases during DNA extraction of activated sludge samples revealed by high throughput sequencing. *Appl. Microbiol. Biotechnol.* 97: 4607-4616.

Holland, P.M., Abramson, R.D., Watson, R., and Gelfand, D.H. (1991). Detection of specific polymerase chain reaction product by utilizing the 5'----3' exonuclease activity of *Thermus aquaticus* DNA polymerase. *PNAS* 88: 7276-7280.

Horz, H.P., Vianna, M.E., Gomes, B.P.F.A., and Conrads, G. (2005). Evaluation of universal probes and primer sets for assessing total bacterial load in clinical samples: General implications and practical use in endodontic antimicrobial therapy. *J. Clin. Microbiol.* 43: 5332-5337.

Hou, Y., Zhang, H., Miranda, L., and Lin, S. (2010). Serious overestimation in quantitative pcr by circular (supercoiled) plasmid standard: Microalgal pcna as the model gene. *PLoS ONE* 5, e9545.

Humbert, S., Zopfi, J., and Tarnawski, S.-E. (2012). Abundance of anammox bacteria in different wetland soils. *Environ. Microbiol. Rep.* 4: 484-490.

Huse, S. M., Welch, D. M., Morrison, H. G., and Sogin, M. L. (2010). Ironing out the wrinkles in the rare biosphere through improved OTU clustering. *Environ. Microbiol.* 12: 1889-1898.

Illumina Inc. (2013) MiSeq Sample Sheet, Quick Reference Guide, Part # 15028392 Rev. J

Illumina Inc. (2014a). Illumina Customer Sequence Letter. Oligonucleotide sequences © 2007-2013 Illumina, Inc. All rights reserved. Derivative works created by Illumina customers are authorized for use with Illumina instruments and products only. All other uses are strictly prohibited.

Illumina Inc. (2014b) MiSeq System User Guide, Part # 15027617 Rev. O

Illumina, Inc. (2015). An Introduction to Next-Generation Sequencing Technology, www.illumina.com

Janda, J. M., and Abbott, S. L. (2007). Minireview: 16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: Pluses, perils, and pitfalls. *J. Clin. Microbiol.* 45: 2761-2764.

Jensen S.H., A. Stensballe, P.H. Nielsen, F.-A. Herbst (2014). Metaproteomics: Evaluation of protein extraction from activated sludge. *Proteomics* 14(21-22), 2535-2539.

Hadfield, J. (2012). How do SPRI beads work? http://core-genomics.blogspot.dk/2012/04/how-do-spri-beads-work.html

Human Microbiome Project (HMP) (2010). Jumpstart Consortium Human Microbiome Project Data Generation Working Group (2010). 16S 454 Sequencing Protocol HMP Consortium. Version 4.2.2.

Juretschko, S., Purkhold, U., Pommerening-ro, A., Schmid, M. C., Koops, H., and Wagner, M. (2000). Phylogeny of all recognized species of ammonia oxidizers based on comparative 16S rRNA and amoA sequence analysis: Implications for molecular diversity surveys. *Appl. Environ. Microbiol.* 66, 5368–5382.

Kim, J., Lim, J., and Lee, C. (2013). Quantitative real-time PCR approaches for microbial community studies in wastewater treatment systems: Applications and considerations. *Biotechnology Advances* 31: 1358-1373.

Kitajima, M., Iker, B.C., Pepper, I.L., and Gerba, C.P. (2014). Relative abundance and treatment reduction of viruses during wastewater treatment processes identification of potential viral indicators. *Sci. Total Environ.* 488-489: 290-296.

Klindworth, A., Pruesse, E., Schweer, T., Peplies, J., Quast, C., Horn, M., and Glöckner, F. O. (2013). Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. *Nucleic Acids Research*, 41, e1.

Kubista, M., Andrade, J.M., Bengtsson, M., Forootan, A., Jonák, J., Lind, K., Sindelka, R., Sjöback, R., Sjögreen, B., Strömbom, L., *et al*. (2006). The real-time polymerase chain reaction. *Mol. Aspects Medicine* 27: 95-125.

Laureni, M., Weissbrodt, D.G., Szivák, I., Robin, O., Nielsen, J.L., Morgenroth, E., Joss, A. (2015). Activity and growth of anammox biomass on aerobically pre-treated municipal wastewater. *Wat. Res*., 80: 325-336.

Legendre, P., and Gallagher, E. (2001). Ecologically meaningful transformations for ordination of species data. *Oecologia*, 129: 271-280.

Li, X., Wu, Y., Zhang, L., Cao, Y., Li, Y., Li, J., … Wu, G. (2014). Comparison of three common DNA concentration measurement methods. *Analytical Biochemistry*, 451: 18-24.

Liu, C.M., Kachur, S., Dwan, M.G., Abraham, A.G., Aziz, M., Hsueh, P.-R., Huang, Y.-T., Busch, J.D., Lamit, L.J., Gehring, C.A., *et al*. (2012). FungiQuant: A broad-coverage fungal quantitative real-time PCR assay. *BMC Microbiol.* 12: 255.

Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology,* 15: 550.

Lozupone, C.A., and Knight, R. (2008). Species divergence and the measurement of microbial diversity. *FEMS Microbiol. Rev*., 32: 557-578.

Ludwig, W., and Schleifer, K.-H. (2000). How quantitative is quantitative PCR with respect to cell counts? *Syst. Appl. Microbiol.* 23: 556–562.

Madigan, M.T., Martinko, J.M. (2006). Brock Biology of Microorganisms, 11th ed. Pearson Education International. ISBN 0131968939.

Magurran, A. E. (2004). Measuring biological diversity. Oxford, UK: Blackwell Publishing.

Mahé, F., Rognes, T., Quince, C., de Vargas, C., and Dunthorn, M. (2014). Swarm: robust and fast clustering method for amplicon-based studies. *PeerJ*, 1-12.

Marsh, T.L., 1999. Terminal restriction fragment length polymorphism (T-RFLP): an emerging method for characterizing diversity among homologous populations of amplification products. *Curr. Opin. Microbiol.* 2: 323-327.

Marzorati, M., Wittebolle, T., Boon, N., Daffonchio, D., Verstraete., W. (2008). How to get more out of molecular fingerprints: Practical tools for microbial ecology. *Environ. Microbiol.* 10: 1571-1581.

Matsuda, K., Tsuji, H., Asahara, T., Kado, Y., and Nomoto, K. (2007). Sensitive quantitative detection of commensal bacteria by rRNA-targeted reverse transcription-PCR. *Appl. Environ. Microbiol.* 73: 32-39.

McDonald, D., Price, M. N., Goodrich, J., Nawrocki, E. P., DeSantis, T. Z., Probst, A., … Hugenholtz, P. (2012). An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *ISME Journal*, 6: 610-618.

McIlroy, S.J., Saunders, A.M., Albertsen, M., Nierychlo, M., McIlroy, B., Hansen A.A., Karst, S.M., Nielsen, J.L., Nielsen, P.H. (2015). MiDAS: the field guide to the microbes of activated sludge. *Database* bav062. doi: 10.1093/database/bav062.

McMaster, G.K., Carmichael, G.G. (1977). Analysis of single- and double-stranded nucleic acids on polyacrylamide and agarose gels by using glyoxal and acridine orange. *PNAS* 74: 4835-4838.

McMurdie, P.J., and Holmes, S. (2013). Phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PloS One,* 8, e61217.

Mendell, J.E., Clements, K.D., Choat, J.H., and Angert, E.R. (2008). Extreme polyploidy in a large bacterium. *PNAS* 105: 6730-6734.

Miller, D.N., Bryant, J.E., Madsen, E.L., Ghiorse, W.C. (1999). Evaluation and optimization of DNA extraction and purification procedures for soil and sediment samples. *Appl. Environ. Microbiol.* 65: 4715-4724.

Muyzer, G. (1999).DGGE/TGGE a method for identifying genes from natural ecosystems. *Curr. Opin. Microbiol*. 2: 317-322.

Nadkarni, M.A., Martin, F.E., Jacques, N.A., and Hunter, N. (2002). Determination of bacterial load by real-time PCR using a broad-range (universal) probe and primers set. *Microbiology* 148: 257-266.

Nanodrop Technologies, Inc. (2007). ND-1000 Spectrophotometer V3.3 User's manual, rev.3.

Nolan, T., Hands, R.E., and Bustin, S.A. (2006). Quantification of mRNA using real-time RT-PCR. *Nature Protocols* 1: 1559-1582.

Nygren, J., Svanvik, N., and Kubista, M. (1998). The interactions between the fluorescent dye thiazole orange and DNA. *Biopolymers* 46: 39-51.

Okano, Y., Hristova, K.R., Leutenegger, C.M., Jackson, L.E., Denison, R.F., Gebreyesus, B., Lebauer, D., and Scow, K.M. (2004). Application of real-time PCR to study effects of ammonium on population size of ammonia-oxidizing bacteria in soil. *Appl. Environ. Microbiol.* 70: 1008-1016.

Oksanen, J., Blanchet, G. F., Kindt, R., Legendre, P., Minchin, P. R., O'Hara, R. B., … Wagner, H. (2015). Vegan: *Community Ecology Package*. http://r-forge.r-project.org/projects/vegan/.

Pace, N. R., Sapp, J., and Goldenfeld, N. (2012). Phylogeny and beyond: Scientific, historical, and conceptual significance of the first tree of life. *PNAS*, 109: 1011-1018

Padmanaban, A., Inche, A., Gassmann, M., Salowsky, R.. 2013. High-Throughput DNA Sample QC Using the Agilent 2200 Tapestation System. *J. Biomol. Techn. JBT* 24:S41.

Panaro, N.J., Yuen, P.K., Sakazume, T., Fortina, P., Kricka, L.J., Wilding, P. 2000. Evaluation of DNA fragment sizing and quantification by the Agilent 2100 bioanalyzer. *Clinical Chemistry* 46: 1851-1853.

Pecoraro, V., Zerulla, K., Lange, C., and Soppa, J. (2011). Quantification of ploidy in proteobacteria revealed the existence of monoploid, (mero-)oligoploid and polyploid species. PLoS ONE, 6. e16392. doi:10.1371/journal.pone.0016392.

Polz, M.F., and Cavanaugh, C.M. (1998). Bias in template-to-product ratios in multitemplate PCR, *Appl Environ Microbiol.* 64(10): 3724-3730.

Pruesse, E., Peplies, J., and Glöckner, F. O. (2012). SINA: Accurate high-throughput multiple sequence alignment of ribosomal RNA genes. *Bioinformatics* 28(14): 1823-1829.

Qiu, X., Wu, L., Huang, H., Donel, P. E. M. C., Palumbo, A. V, Tiedje, J. M., and Zhou, J. (2001). Evaluation of PCR-generated chimeras, mutations, and heteroduplexes with 16S rRNA gene-based cloning, *Appl Environ Microbiol.* 67(2): 880-887.

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., … Glöckner, F. O. (2013). The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. *Nucleic Acids Research*, 41. doi:10.1093/nar/gks1219

Quince, C., Lanzen, A., Davenport, R. J., and Turnbaugh, P. J. (2011). Removing noise from pyrosequenced amplicons. *BMC Bioinformatics*, 12: 38.

Pruesse, E., Peplies, J., and Glöckner, F. O. (2012). SINA: Accurate high-throughput multiple sequence alignment of ribosomal RNA genes, *Bioinformatics* 28(14): 1823-1829.

Ramette, A. (2007). Multivariate analyses in microbial ecology. *FEMS Microbiol. Ecol*., 62: 142-60.

Rasmussen, R. (2001). Quantification on the LightCycler. In Rapid Cycle Real-Time PCR, P.D. med S. Meuer, P.D.C. Wittwer, and D.K.-I. Nakagawara, eds. (Springer Berlin Heidelberg), pp. 21-34.

Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010). EdgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26: 139-40.

Rotthauwe, J.H., Witzel, K.P., Liesack, W., (1997). The ammonia monooxygenase structural gene *amoA* as a functional marker: Molecular fine-scale analysis of natural ammonia-oxidizing populations. *Appl Environ Microbiol.* 63: 4704-4712.

Saiki, R.K., Scharf, S., Faloona, F., Mullis, K.B., Horn, G.T., Erlich, H.A., and Arnheim, N. (1985). Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. *Science* 230: 1350-1354.

Salonen, A., Nikkilä, J., Jalanka-Tuovinen, J., Immonen, O., Rajilić-Stojanović, M., Kekkonen, R. a., Palva, A., de Vos, W.M. (2010). Comparative analysis of fecal DNA extraction methods with phylogenetic microarray: effective recovery of

bacterial and archaeal DNA using mechanical cell lysis. *J. Microbiological Methods* 81: 127-34.

Saunders, A.M., Albertsen, M., Vollertsen, J., and Nielsen, P.H. (2015). The activated sludge ecosystem contains a core community of abundant organisms. *ISME J.,*doi:10.1038/ismej.2015.117

Schloss, P.D. and Handelsman, J. (2005). Metagenomics for studying unculturable microorganisms: cutting the Gordian knot. *Genome Biol.* 6: 229.

Schloss, P.D., Westcott, S.L., Ryabin, T., Hall, J.R., Hartmann, M., Hollister, E.B., … Weber, C.F. (2009). Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol.,* 75: 7537-41.

Schriewer, A., Wehlmann, A., and Wuertz, S. (2011). Improving qPCR efficiency in environmental samples by selective removal of humic acids with DAX-8. *J. Microbiol. Methods* 85: 16-21.

Seifert, J.F.-A. Herbst, P.H. Nielsen, F.J. Planes, M. Ferrer, M. von Bergen (2013) Bioinformatic progress and applications in metaproteogenomics for bridging the gap between genomic sequences and metabolic functions in microbial communities. *Proteomics* 13: 2786-2804. doi: 10.1002/pmic.201200566.

Singer, V.L., Jones, L.J., Yue, S.T., Haugland, R.P. (1997). Characterization of PicoGreen reagent and development of a fluorescence-based solution assay for double-stranded DNA quantitation. *Analytical Biochemistry* 249: 228-238.

Smith, S., Morin, P.A. (2005). Optimal storage conditions for highly dilute dna samples: a role for trehalose as a preserving agent. *J. Forensic Sci.*, 50: 1101-1108.

Souazé, F., Ntodou-Thomé, A., Tran, C.Y., Rostène, W., and Forgez, P. (1996). Quantitative RT-PCR: limits and accuracy. *BioTechniques* 21: 280-285.

Stults, J.R., Snoeyenbos-West, O., Methe, B., Lovley, D.R., and Chandler, D.P. (2001). Application of the 5 fluorogenic exonuclease assay (taqman) for quantitative ribosomal DNA and rRNA analysis in sediments. *Appl. Environ. Microbiol.* 67: 2781-2789.

Tebbe, C.C., and Vahjen, W. (1993). Interference of humic acids and DNA extracted directly from soil in detection and transformation of recombinant DNA from bacteria and a yeast. *Appl. Environ. Microbiol* 59(8): 2657-2665.

Thermo Scientific (2015). Assessment of Nucleic Acid Purity, T042-TECHNICAL BULLETIN, T042 Rev 1/11.

Thermo Scientific (2015a), Qubit dsDNA HS Assay Kits – User Guide, MAN0002326, P32851, Revision: B.0

Thermo Scientific (2015b), Qubit dsDNA BR Assay Kits – User Guide, MAN0002325, P32850, Revision: A.0

Thomas, T., Gilbert, J., Meyer, F. (2012). Metagenomics - a guide from sampling to data analysis. *Microb. Inform. Exp.* 2:3. 10.1186/2042-5783-2-3.

Thompson, I.P., Van Der Gast, C.J., Ciric, L., and Singer, A.C. (2005). Bioaugmentation for bioremediation: the challenge of strain selection. *Environ. Microbiol* .7: 909-915.

Tsai, Y., Olson, B. (1991). Rapid method for direct extraction of DNA from soil and sediments. *Appl. Environ. Microbiol.* 57: 1070-1074.

Tullis, R.H., Rubin, H. (1980). Calcium protects DNase I from proteinase K: a new method for the removal of contaminating RNase from DNase I. *Analyt. Biochem.*107: 260-264.

Větrovský, T., and Baldrian, P. (2013). The variability of the 16S rRNAS gene in bacterial genomes and its consequences for bacterial community analyses. PLoS ONE *8*. e57923. doi:10.1371/journal.pone.0057923.

Volkmann, H., Schwartz, T., Bischoff, P., Kirchen, S., and Obst, U. (2004). Detection of clinically relevant antibiotic-resistance genes in municipal wastewater using real-time PCR (TaqMan). *J. Microbiol. Methods* 56: 277-286.

Wang, Y., and Qian, P.-Y. (2009). Conservative fragments in bacterial 16S rRNA genes and primer design for 16s ribosomal DNA amplicons in metagenomic studies. PLoS ONE *4*, e7401. doi:10.1371/journal.pone.0007401.

Wilfinger, W.W., Mackey, K., Chomczynski, P. (1997). Effect of pH and ionic strength on the spectrophotometric assessment of nucleic acid purity. *BioTechniques* 22: 474-476.

Wilson, I.G. (1997). Inhibition and facilitation of nucleic acid amplification. *Appl. Environ. Microbiol.*, 63(10): 3741-3751.

Woese, C.R., and Fox, G.E. (1977). Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *PNAS*, 74: 5088-90.

Wooley, J.C., Godzik, A., Friedberg, I. (2010). A primer on metagenomics. *PLoS Comput. Biol.* 6(2): e1000667. doi:10.1371/journal.

Zhang, T., and Fang, H.H.P. (2006). Applications of real-time polymerase chain reaction for quantification of microorganisms in environmental samples. *Appl.Microbiol. Biotechnol.* 70: 281-289.

Zhou, J., Bruns, M.A., Tiedje, J.M. (1996). DNA recovery from soils of diverse composition. *Appl. Environ. Microbiol.* 62: 316-322.

Zipper, H., Brunner, H., Bernhagen, J., and Vitzthum, F. (2004). Investigations on DNA intercalation and surface binding by SYBR Green I, its structure determination and methodological implications. *Nucl. Acids Res.* 32: e103–e103.

Zuur, A.F., Ieno, E.N. and Smith, G.M. (2007) Analysing Ecological Data. Springer, New York.