Essentials in Metagenomics (Part I)

Authors: Angela Peña-González[1] and Alejandro Reyes Muñoz[2]

Affiliations:

1-Angela Peña-Gonzalez, PhD, MS
Postdoctoral Researcher
Group in Computational Biology and Microbial Ecology
Max Planck Tandem Group in Computational Biology
Department of Biological Sciences
Universidad de los Andes
Bogotá, Colombia.
https://orcid.org/0000-0003-0684-8179

2-Alejandro Reyes Muñoz, PhD (corresponding author)
Associate Professor
Director BCEM (bcem.uniandes.edu.co)
Group Leader Max Planck Tandem Group in Computational Biology
Department of Biological Sciences
Universidad de los Andes
Bogotá, Colombia
https://orcid.org/0000-0003-2907-3265
Max Planck Tandem Group in Computational Biology, Department of Biological Sciences, Universidad de los Andes, Cra. 1 #18a 12, Bogotá, Colombia, 111711
Center for Genome Sciences and Systems Biology, Department of Pathology and Immunology, Washington University School of Medicine, St Louis, MO, USA, 63110

Overview

In this course we present a review of essential aspects in metagenomic studies and analysis. We discuss aspects related to the design and implementation of metagenomic studies and provide practical guidance in sample processing, sequencing and data analysis. We also present the general principles behind the most common analyses and tools used for manipulating metagenomic DNA and libraries.

Teaching goals and learning outcomes

This course is aimed at undergraduate students, graduate students and professionals at any career stage performing research in the fields of microbiology and bioinformatics. Also, molecular biologists and microbiologists (life scientists) interested in learning basics on metagenomics data

analysis and bioinformatics will benefit from this course. An undergraduate knowledge of a subject related to the life sciences (molecular biology) would be an advantage.

By the end of this CABANA e-Learning tutorial, the user will be able to:

- Recognize the potential that metagenomics holds for identifying uncultivable microorganisms and for understanding the structure and function of naturally occurring microbial communities.
- Build a basic experimental design for a metagenomic study and describe the main steps involved in sample acquisition and processing.
- Identify the main technical challenges associated with the design and implementation of a metagenomic study as well as many sources of variation that need to be controlled.

About this guide

This is a guide that you can read through at your own speed. The sections "***Material to study***" are additional material to expand your knowledge. Additionally, the sections "***Stop and Think***" include questions or situations for you to consider in more depth about the knowledge you are acquiring.

For general definitions in the field of biological sciences, the user can query the following link: *https://www.ebi.ac.uk/training/online/glossary*. In addition, we suggest the user read the editorial material entitled 'The Vocabulary of Microbiome Research: A proposal' [11] written by Julian Marchesi and Jacques Ravel and published in the BMC journal Microbiome PMID:26229597. In this publication, authors proposed clear definitions of terms related to microbiome studies.

Key terms

Metagenomics, experimental design, sample acquisition and processing, next generation sequencing (NGS)

1-Introduction

Microbial communities play a central role in geochemical cycles, agricultural systems and human health. The metabolic activities of microorganisms and their aggregated diverse communities can have a critical impact on the biosphere and in localized micro-niches. Understanding the structure and function of complex microbial communities requires a system-view to define and quantify the mechanisms that control individual cellular organisms, their interaction in a community of heterogeneous cells and an ecological niche on a broader scale [1].

Classical microbiology has historically relied on the culturing of 'pure' isolates obtained from environmental samples in the laboratory. However, researchers have estimated that cultivable

microorganisms represent only a small fraction (~1%) of the total microbes within a given habitat [2]. The main limitation is that we still lack knowledge of the appropriate cultivation conditions for most microbial species [3]. The availability of high-throughput sequencing techniques to study environmental and clinical DNA (and mRNA) samples and the growing reference sequence databases allow us to study otherwise inaccessible genes and genomes from a variety of environments. **Metagenomics is a strategy for capturing and analyzing the genomes of an entire microbial community present within a given habitat (e.g. the metagenome), based on the culture-independent extraction of the total DNA.** By 2006, only 18.9% of the bacterial sequences and 6.8% of the archaeal sequences were obtained from isolated organisms [4], numbers that are most likely dropping daily.

Certainly, the field of metagenomics generated a revolution in microbial ecology studies, allowing researchers to gain new insights into the diversity, dynamics and evolution of naturally occurring microbial communities. Now it is possible to identify not only the microorganisms present in a variety of environments such as hot springs or glaciers, but it is also feasible to infer their metabolic potential.

Some of the main applications of metagenomics include:

- Identification of new species
- Characterization of structure and diversity in (complex) microbial communities
- Comparison of populations and potential functional activities present in different sites and metabolic states
- Bioprospecting for new sequences and their functional applications
- Reconstruction of metabolic pathways present in the community
- Pathogen detection, typing and surveillance

Several types of ecosystems have been studied so far using metagenomic approaches, including extreme environments such as areas of volcanism [5], with high or low temperature [6], alkalinity [7], acidity [8], low oxygen [9], heavy metal concentrations [10] and the human gut. This invaluable resource provides an infinite capacity for bioprospecting and enhancing the discovery of novel enzymes capable of catalyzing reactions of medical and/or biotechnological importance.

2- Metagenomics defined

The term ‘Metagenomics’ was originally defined as **‘the direct genetic analysis of genes and genomes contained in an environmental sample without the prior need for cultivating clonal populations’** [12]. Initially, the term was only used for functional and sequence-based analysis of the collective microbial genomes contained in an environmental sample, known as ‘**shotgun metagenomics**’ [2], but currently it is also widely applied to studies performing polymerase chain reaction (PCR) amplification of certain genes of interest such as the *16S rRNA* gene. The later approach is commonly known as ‘**marker gene metagenomics’** or **‘metabarcoding**’

Therefore, as explained by Mitchell et al., (2018) [13], strictly speaking metagenomics refers to whole genome shotgun sequencing of an environmental or clinical sample. However, you might find studies involving sequencing of amplicon marker genes such as *16S rRNA* gene (small subunit (16S) ribosomal RNA) and other marker genes as ITS1 o COX1, published under the umbrella of metagenomics.

All these methodologies allow a much faster genomic/genetic profiling of an environmental sample at a reduced cost compared to culturing. **Shotgun metagenomics** has the potential to fully sequence the majority of available genomes within a microbial community. This enables the characterization of the taxonomic profile that can be further associated with functional capabilities of known and unknown lineages. In other words, shotgun metagenomics has evolved to address the questions of ***'who is present'*** in a given environmental or clinical sample, ***'what are they doing'*** (or are capable of doing), and ***'how do these microorganisms interact'*** to sustain a balanced ecological niche.

Marker gene metagenomics is a relatively fast approach to elucidate the taxonomic composition of the microbial community by using PCR amplification and sequencing of evolutionary-conserved marker genes. The observed taxonomic profile (**composition and distribution of phylogenetic groups or, phylogroups**) can subsequently be associated with metadata derived from the samples in the study. The goal is to identify differentially present and/or abundant phylogroups among samples and describe the overall microbial community diversity and structure.

The most common challenges in metagenomics include:

- Many sequences coming from different organisms are analyzed at the same time
- Many sequences might not annotated in reference databases (unknown source)
- Short sequences are usually difficult to characterize
- Assembly of short reads can be difficult, especially in highly diverse environments

Other difficulties include:

- Different sample processing steps can generate different sequence results
- There is no perfect tool for analysis and different tools can generate different results
- Even the same tool used with the same parameters can provide different results if different reference databases are used

The **Figure 1** shows an example of how different bioinformatic tools used to characterize the taxonomic composition (at the genus level) of microbes living in different environments can produce varying results [14].

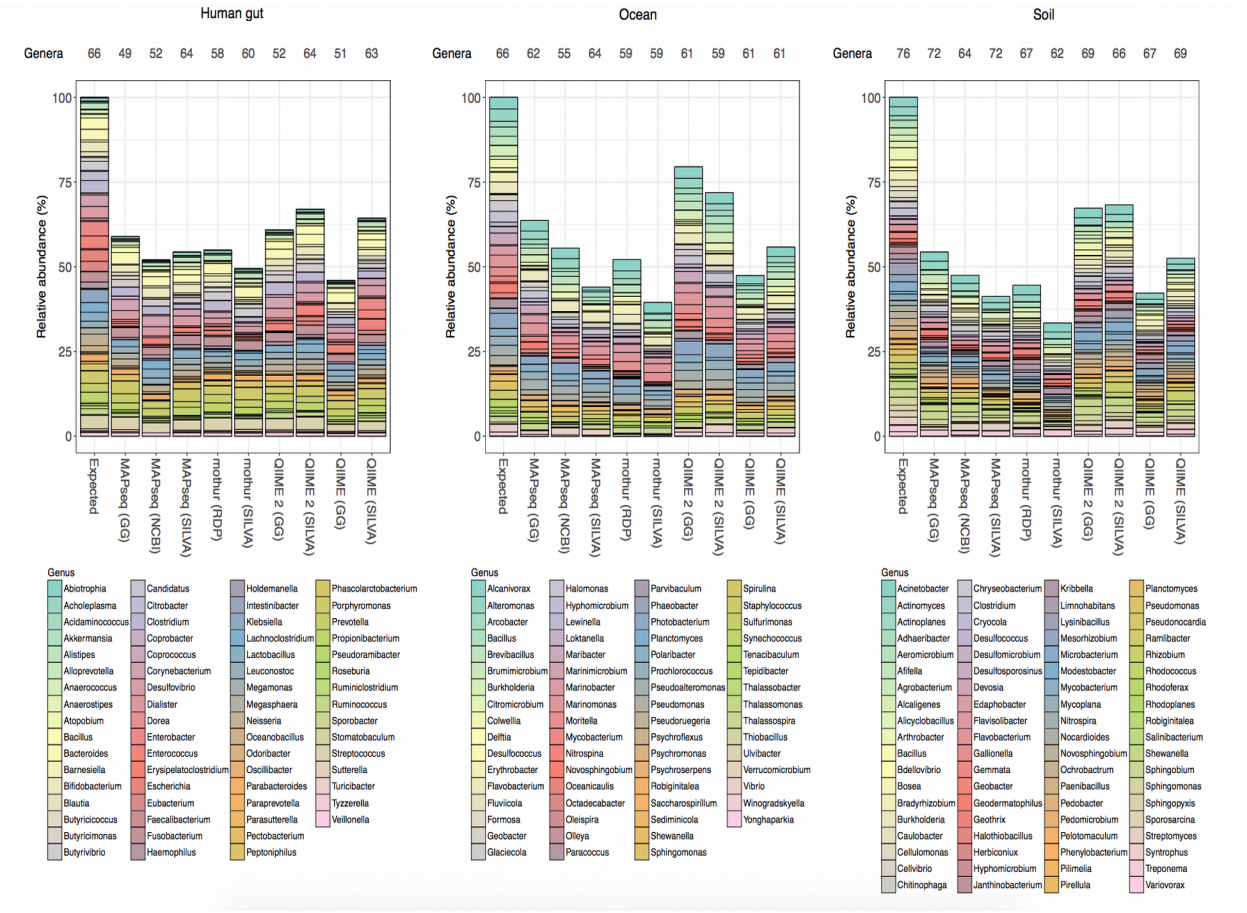

Figure 1. Genus-level classification of three environments (human gut, ocean and soil) with different tools and databases of the *16S rRNA* gene. Each bar corresponds to a different tool and the number above each bar indicates the number of correctly identified genera. In the figure we can observe that different bioinformatic tools can generate varied results and there does not exist a single perfect tool. Figure taken from [14]

3- Distinguishing microbiome from microbiota

According to Marchesi & Ravel (2015), the **microbiota** refers to the community of microorganisms found in an environment. In other words, it **is the assemblage of microorganisms present in a defined environment** [11]**.** This term is frequently used to describe the microbial census established using molecular methods relying predominantly on the analysis of *16S rRNA* genes, *18S rRNA* genes, or other marker genes and genomic regions, amplified and sequenced from a given biological samples. The tool can be used in different environments such as the soil, plant, animal or human-associated environments.

The **microbiome** in the other hand, refers to the entire habitat, including the microorganisms (bacteria, archaea, lower and higher eukaryotes, viruses), their genomes (i.e., genes), and the surrounding environmental conditions. This definition is based on that of **"biome", the biotic and abiotic factors of given environments**. Others in the field limit the definition of microbiome to the collection of genes and genomes of members of a microbiota [15]. It is argued that this is the definition of metagenome, which combined with the environment constitutes the microbiome (**Figure 2**). In general, the microbiome can be characterized by the application of one or more omics techniques that aim to study the collection of biomolecules (genes, proteins, metabolites, etc), combined with clinical or environmental metadata, which are contained within a niche.

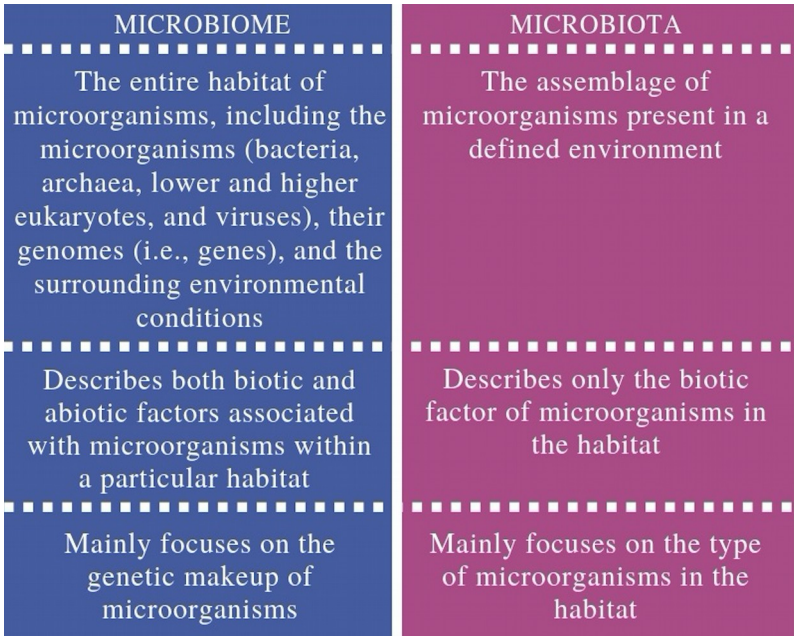

Figure 2. Chart summarizing the differences between microbiome and microbiota as suggested by Marchesi and Ravel (2015). Image modified from https://pediaa.com/what-is-the-difference-between-microbiome-and-microbiota/

Material to study

The following link will point you towards an entertaining reading about the origin of the word microbiome written by Dr. Jonathan Eisen (Professor at the University of California, Davis). Dr Eisen is the editor of ***microBEnet****, the microbiology of the Built Environment network, a blog dedicated to enhancing communication and collaboration in research related to the microbiology of the built environment. microBEnet*

4- Considerations when designing a metagenomic experiment

In metagenomic studies multiple aspects must be taken into consideration when designing the experiment. Ideally, the design stage should include a bioinformatic and statistical advisor, in addition to the molecular biologists, to define aspects that will be critical to obtain precise and significant results [16]. A common limitation in microbiome studies is that there are many confounding factors that can obscure patterns in the analyses, such as variability between individuals (often driven by differences in age, gender, diet, life style, among others) and environments (seasonal change, sampling time), preservation of samples during transport, among many others (**Figure 3**).

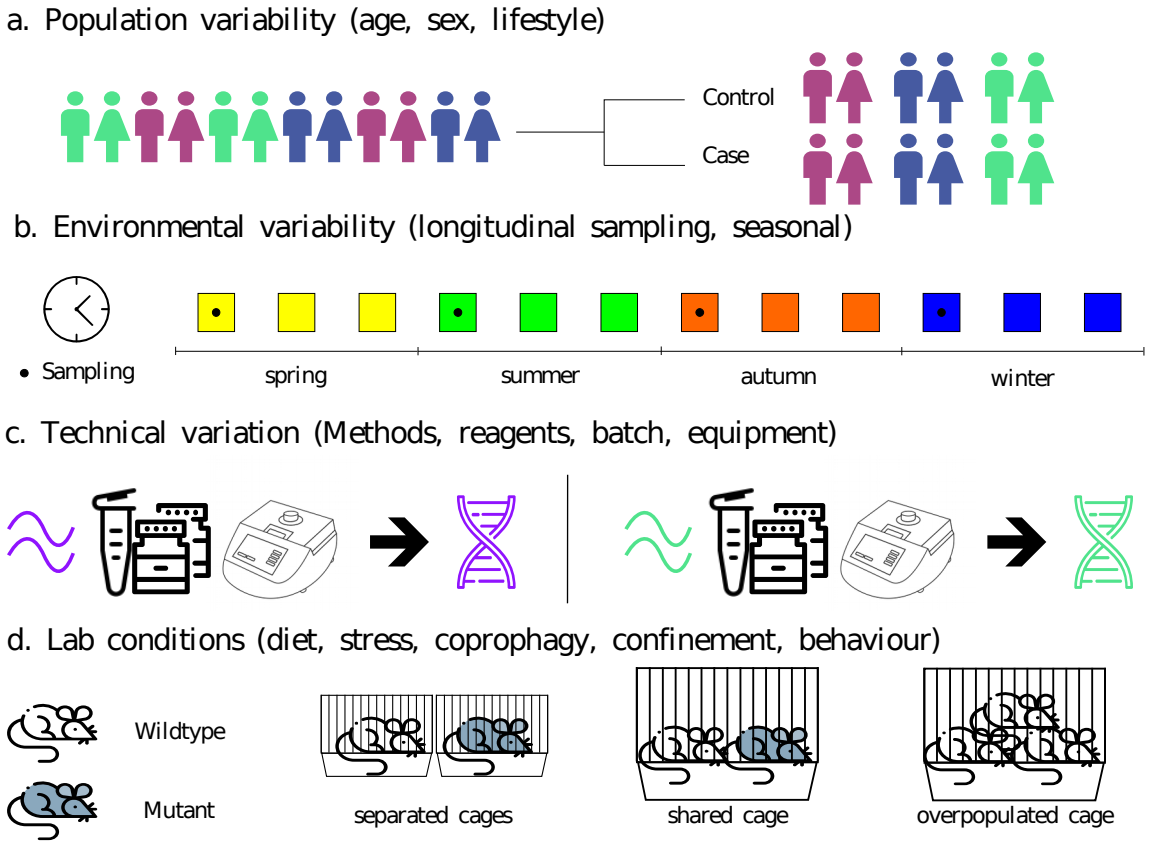

Figure 3. Some aspects to consider when designing a microbiome experiment. a) There are factors that can generate stratification among individuals and knowing them helps to avoid incorrect conclusions. b) Longitudinal studies have more power to assess the stability of communities. c) It is necessary to have control of all technical steps and ensure replicability of these. d) In animal studies, characteristics of behavior must be taken into account, for example coprophagy.

Probably the most important aspects to define in a metagenomic study are: 1) the scope of the experiment and 2) the question to be answered. Depending on this, a cross-sectional, longitudinal or interventional study can be designed, especially for association studies of microbiomes and disease. Once the plan and the questions have been defined, the sample size can be established, as well as the methods to determine the statistical power. This is done by discerning the variability inherent in the techniques with respect to the biology of the samples. In the case of microbiome studies, this is still a challenge, although multivariate analyses, such as PERMANOVA, Dirichlet multinomial and prediction methods, such as random forest can be used. However, it is advisable to consult a statistician, review the design of other successful experiments and evaluate the possibility of following a similar design [16]. This will also allow the comparison of studies in meta-analyses.

Another relevant factor to define when considering the experimental design is the **number of biological replicates**. The optimal number of replicates will depend on the biological variability of the target organism/environment and the technical variability of the methodologies. The number of replicates also influences statistical power. Accordingly, each experiment should include at least three biological replicates (ideally 5 replicates) to evaluate the reproducibility of the results [17]. Similarly, another key aspect to consider is **consistency.** It is critical to use consistent methodologies throughout the study to minimize potential biases which could lead to spurious results.

Material to study

For more information regarding statistical methods for the design and analysis of microbiome data, please review the following publications:

- *Hypothesis testing and statistical analyses of microbiome link*
- *Microbial diversity in clinical microbiome studies: sample size and statistical power considerations link*
- *Rigorous statistical methods for rigorous microbiome science link*

4.1 Sample processing

Sampling involves the collection of samples and the appropriate documentation of the associated metadata [18]. Metadata is a fancy word commonly used by researchers to make reference to the data that describes and provides information about other data. In other words, **metadata is data about the data.**

In the sampling step it is important to take into account that several technical factors could generate variability among samples: transport, sample preservation and metadata [19]. Sampling variation can have an effect on abundance measurements and comparison [20]. This step (sample collection) could introduce additional challenges for researchers as some samples must be

delivered anaerobically. Exposure to oxygen or freezing can change the dynamic composition of a given microbial community. For example, freezing, thawing and subsequent bead-beating can affect the cell wall of Gram-positive bacteria, and introduce artifacts compared to extraction performed on fresh samples [21]. Ideally, fresh samples should be fast frozen at -80°C immediately after collection or stored in stabilizing substances such as RNAlater (ThermoFisher). The goal is to reduce the degradation of RNA and DNA, which is a natural process that occurs rapidly in cells, and to avoid overgrowth of certain microbial species due to the storage conditions, changing the real composition of the community.

4.2 DNA extraction

The extraction of nucleic acids from the sample is probably one of the most crucial steps in any metagenomic project. The reason is that the DNA extracted should be representative of all cells present in the sample and sufficient amounts of high-quality nucleic acids must be obtained for subsequent library preparation and sequencing. The amount and integrity of the DNA required may vary depending on the sequencing platform to be used, in particular if long read technology is to be used. Many DNA extraction methods have been described in the literature and often they have been developed for specific cell or sample types, however they will usually share some common steps: cell lysis, purification and elution/precipitation. Before the nucleic acids are extracted, the sample must be homogenized so that the cells/tissues are broken down and the cell contents are in solution. This process is commonly done mechanically with ceramic pearls in special homogenizers or with equipment similar to a small blender.

As mentioned before, DNA extraction methods commonly include the following steps: cell lysis, protein removal, DNA precipitation or alternatively binding to membranes or filter made of fiberglass, silica or ion exchange (or with magnetic beads) and finally, DNA re-suspension in buffer solutions. The simplest way to get started is to use a DNA extraction kit. Typically kits offer a high level of consistency and excel for low-input sample types [22]. They are however more expensive than manual methods, typically costing around $5 per sample in the US.

Kits frequently used for DNA extraction from environmental and clinical samples include:

- MoBIO PowerSoil® DNA Extraction Kit, now from Qiagen
- Qiagen DNA Microbiome Kit
- Epicentre Metagenomic DNA Isolation Kit for water
- Epicentre Meta-G-Nome DNA Isolation Kit
- Among others...

Material to study

To learn more about DNA extraction strategies for sequencing, please visit the following GitHub repository (nickloman.github.com).

Stop and think

Keep in mind that fragment length for common extraction kits will be limited to around ~60 Kb which could be problematic when the extraction of longer reads is desired. In addition, several studies [23-25] have reported that contaminating DNA is ubiquitous in commonly used DNA extraction kits and other laboratory reagents, which varies greatly in composition between different kits and batches. This contamination might critically impact results obtained from samples containing a low microbial biomass. For this reason it is always good practice to include negative controls during DNA extraction.

Spin column kits are the most common type of DNA extraction kit you will probably find in a molecular biology laboratory. Spin columns are so called because reagents are added to the top of the tube and then forced through a binding matrix when revolved in a centrifuge. In some cases, columns include cell lysis reagents. Binding DNA, washing and eluting the DNA can be done rapidly in this way, with the whole process taking around an hour. In addition, you can perform many extractions in parallel, by using more positions in the centrifuge rotor. Spin columns are based on chemistry developed in the 1990s [26,27] using either silica or anion exchange resins to reversibly bind DNA allowing them to be separated from cellular proteins and polysaccharides [57].

How do spin column methods work? Most spin column kits use high concentrations of guanidinium hydrochloride in the lysis buffer [28]. Guanidinium hydrochloride is a chaotropic agent that disrupts the hydrophobic interactions between water and other molecules. This is a good choice because it both lyses cells by denaturing membrane proteins and precipitates DNA by disrupting its hydration shell which maintains its solubility in aqueous conditions. Under these conditions DNA binds to the binding matrix in the column allowing proteins and other contaminants to pass through. The DNA bound to the silica resin membrane can be washed using 70% ethanol to remove contaminating proteins and salts, including the lysis buffer itself. DNA is then eluted from the column by adding a low ionic concentration buffer such as 10 mM Tris and incubating for a few minutes [29]. The DNA re-solubilizes in the aqueous solution and the purified DNA is eluted from the column by centrifugation (**Figure 4**). A common limitation is that DNA is sheared during binding and elution due to the large physical forces experienced during centrifugation and when is forced to pass through the porous resin.

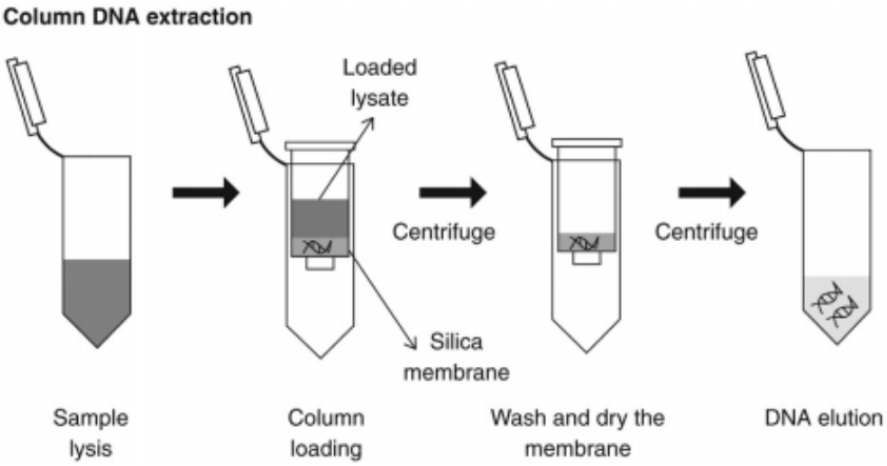

Figure 4. Schematic representation of a spin column-based nucleic acid purification. Prior to column loading, the sample is lysed according to the manufacturer's instructions. The lysate is then loaded into the column in the presence of a chaotropic salt allows the DNA to bind to the silica membrane. The addition of a washing agent as ethanol 70% allows the purification step. Finally, the spin column is dried and the DNA is eluted with an appropriate buffer. Image taken from https://doi.org/10.1016/B978-0-12-416999-9.00007

Stop and think

What if the target community is tightly associated with a host, (e.g. human or plant)? *If this is the case, then physical fractionation or selective lysis can be employed to ensure host DNA is kept to a minimum [30, 31] . This is particularly important in clinical samples obtained from wounds or infected/necrotic tissues where the host genome is large and hence might reduce the signal of the microbial community in the subsequent sequencing effort. Host material can also be removed during bioinformatics filtering. Regardless of the approach used, it's important to remember that extraction and isolation methods can introduce bias in terms of microbial diversity, DNA yield and overall fragment lengths. It's highly recommended that the exact same extraction method be used when comparing samples.*

What if the target community yields only very small amounts of DNA? *In this case amplification of the starting material might be required. Library generation for most sequencing technologies requires high-nanogram or microgram amounts of DNA. One method that can be employed to increase DNA yields from low biomass environments is the Multiple Displacement Amplification (MDA) using random hexamers and phage Phi29 polymerase [32]. As with any amplification method, there are potential problems associated with reagent contamination, chimera formation and sequence bias in the amplification. Their impact will depend on the amount and type of starting material and the required number of amplification rounds to produce sufficient amounts of nucleic acids.*

Material to study

-Additional material regarding Multiple Displacement Amplification can be found here: Rhee et al. 2016
-A post from the microBEnet in an interesting discussion about the best practices for sample processing and storage prior to microbiome DNA analysis. Link

4.3 Assessing DNA concentration and purity

Typically the first thing a molecular biologist wants to know about a DNA preparation is its concentration and purity. Both can be determined using spectrophotometric (absorption of ultraviolet light) and fluorometric (dyes that bind specifically to DNA or RNA) methods.

Historically, DNA and RNA molecules have been quantified using spectrophotometry to measure absorbance at 260 nm. Although this method is commonly used (and we will discuss the principles behind it), it can be unreliable and sometimes inaccurate. The reason is that UV absorbance measurements are not selective and can not reliably distinguish between DNA, RNA and proteins. Nonetheless, it's a commonly used method that can give a quick, initial assessment of DNA yield and purity. In light of these drawbacks, the use of fluorescent dyes that bind single and double stranded molecules of DNA and RNA selectively, has become a common alternative. Other alternatives employed in the quantification of nucleic acids include qPCR and capillary electrophoresis systems that are chip-based. These two processes are mainly used for specific tasks (e.g. qPCR: exact quantification of specific target sequences; capillary electrophoresis: exact determination of fragment size and integrity) as both methods are relatively costly as well as time-consuming.

4.3.1 Principles behind spectrophotometric approaches

The main principle behind this method is that the DNA absorbs UV light more or less strongly depending upon the wavelength. For example, DNA absorbs 1.8 times as much UV at 260 nm than it does at 280nm (**Figure 5**). A common device used to estimate DNA concentration based spectrophotometry is the NanoDrop ND-1000 (ThermoFisher Scientific). The nanodrop takes measurements at wavelengths of 260 and 280 nm, and in addition determines an absorption spectrum from 220 – 350 nm (**Figure 6**).

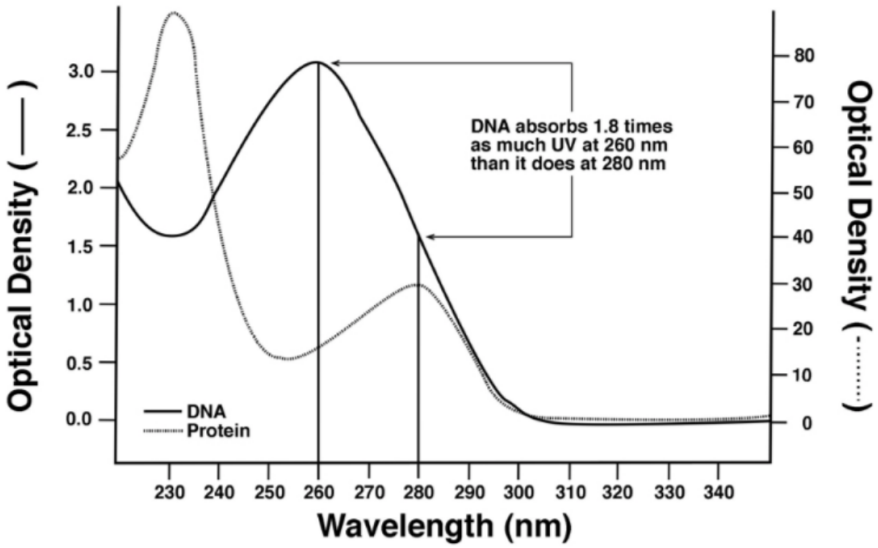

Figure 5. Absorbance of DNA and proteins at different wavelengths. The maximum absorbance for DNA occurs at 260 nm. At 280 nm DNA absorbs about half as much as it does at 260 nm. Image Taken from https://es.scribd.com/document/355068017/09-UV-Absorption-pdf

Although 260 nm is considered to be the *de facto* peak for DNA, the actual peak absorbance varies somewhat from DNA to DNA. UV absorption is a property of the bases, and each base absorbs differently [59]. The actual peak absorbance of a particular DNA, then, depends on its base composition.

The ratio of absorbance at 260 nm and 280 nm (260/280) is used to assess the purity of DNA and RNA (Figure 5). A ratio of ~1.8 is generally accepted as 'pure' for DNA; a ratio of 2.0 is generally accepted as 'pure' for RNA. If the ratio is appreciably lower in either case, it may indicate the presence of proteins, phenol or other contaminants that absorb strongly at or near 280 nm.

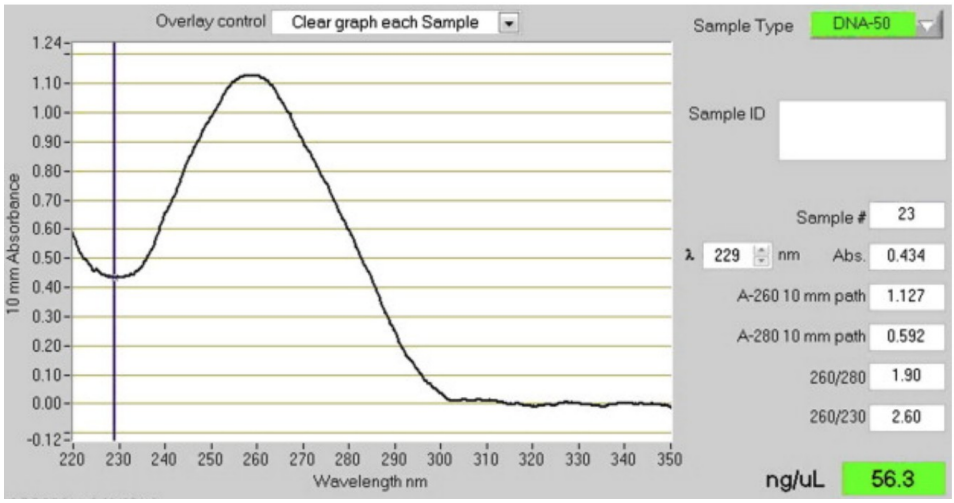

Figure 6. Typical spectral pattern for DNA nucleic acids generated by Nanodrop ND-1000. Image taken from http://stmichaelshospitalresearch.ca/wp-content/uploads/2015/09/NanoDrop-ND-1000-Spectrophotometer.pdf

Some DNA extraction protocols require the use of phenol. While protein contamination is not necessarily a critical problem, phenol contamination most definitely is. Phenol absorbs maximally at 270 nm and can have an impact on the A260/A280 ratio. The ratio A260/A230 is a secondary measure of nucleic acid purity. The ratio A260/A230 for pure nucleic acids are often higher than the respective A260/A280 values. Expected values are commonly in the range of 2.0 to 2.2. If the ratio is significantly lower than expected, it may indicate the presence of contaminating salts that absorb at 230 nm.

Stop and think

When assessing DNA purity it is important to understand that while the A260/A280 ratio is easy and economical to determine and is a widely used method, it is not particularly robust. DNA absorbs so strongly at 260 nm that it takes significant protein contamination to have a noticeable effect on the A260/A280 ratio. In addition, values are easily affected by other contaminants (free nucleotides, salts, organic compounds) and variations in base composition. Furthermore, the sensitivity of spectrophotometry is often limited, prohibiting quantification of molecules at low concentrations. In addition, keep in mind that the purity and absence of salts or other molecules that bind DNA is essential for sequencing methods such as Oxford Nanopore.

4.3.2 Principles behind fluorometric approaches

Quantification of nucleic acids via fluorescence is based on the use of fluorescent dyes (called fluorophores) which bind to the molecule of interest [33]. Only the complex consisting of nucleic acid and dye is excited by light at a specific wavelength (dependent on the dye) and will subsequently emit light of a slightly longer wavelength. The level of fluorescence is proportional to the DNA concentration which can be extrapolated from the fluorescence level of standards of known concentration [34]. For example, the Qubit 4 fluorometer (Life Technologies, Qubit4) is a fluorescence spectrophotometer commonly used for single samples and different kits are available for different sample types and concentration ranges. The most useful kit for metagenomics is the dsDNA HS Assay (Life Technologies) which measures concentrations between 0.01 - 100 ng/µl.

The main advantages of this method include:

- **High sensitivity** (1000 fold higher for the fluorescent dye PicoGreen® compared to absorbance measurements). This means that low concentration samples can be quantified accurately.
- **High robustness**. Requires only very small amounts of sample (as little as 1µL sample).
- **High accuracy**. Fluorescent dyes bind in a highly specific manner so that possible contaminations such as salt or protein will not lead to artificially elevated readings. In addition, some fluorophores bind specifically to one type of nucleic acid (for example dsDNA) thus practically eliminating the impact of possible ssDNA or RNA present in the sample. However, because of this property, this method will prevent the user from knowing if there are other potential contaminants that might affect downstream reactions.

After you make sure your DNA is of high quality, and prior to the sequencing step, DNA libraries need to be made, protocols for which can vary greatly depending on the technology to be employed. For that reason, we will review first the most common sequencing technologies currently being used.

5. Sequencing technologies

Several technologies have been used over the past decade: Ion Torrent PGM, Roche 454, PacBio and, most recently, Oxford Nanopore Technologies. But it is predominantly data from Illumina machines, from the MiniSeq, MiSeq, NextSeq, or HiSeq platforms, that are used for environmental and clinical metagenomic studies as evidenced by the vast amounts of Illumina data deposited in databases (>90% at the Short Read Archive - SRA).

Shotgun metagenomic sequencing might be challenging in the sense that a large diverse pool of microbes are sequenced, each with a different genome size, often mixed with host DNA. Current sequencing technologies offer a wide variety of read lengths and outputs. Illumina sequencing technology offers short reads, 2x150 up to 2x300 bp but generates high sequencing depth.

Longer reads are preferred as they overcome short repeats and other difficulties during assembly, resulting in longer contigs. However instruments that offer longer reads, e.g. PacBio and Oxford Nanopore are accompanied by higher error rates, lower sequencing depth and higher costs. PacBio error rates can be reduced using circular consensus sequencing (CCS), which involves repeat sequencing of a circular template and generation of a DNA insert consensus. High quality 500-4000 bp can be generated with >99% Q20 accuracy.

5.1 Second generation sequencing technology: Illumina/Solexa system

Illumina/Solexa systems have been extensively applied to metagenomic studies. In this sequencing platform, DNA fragments are immobilized on a solid surface (glass flow cell) and then PCR amplified, resulting in clusters of identical DNA fragments. These clusters are then sequenced with reversible terminators in a process referred to as '*sequencing-by-synthesis*' or SBS [35]. The SBS technology was originally developed by Shankar Balasubramanian and David Klenerman at the University of Cambridge. They founded the company Solexa in 1998 to commercialize their sequencing method. Illumina purchased Solexa in 2007 and has rapidly improved the original technology.

How does it work?

Illumina sequencing system uses modified dNTPs: fluorescently labeled adenine (A), cytosine (C), guanine (G) and thymine (T) containing a group terminator that blocks further polymerization, such that only a single base can be added by a polymerase to each growing DNA strand. The four bases compete for binding sites on the template DNA, and the molecules not incorporated are washed away. After each synthesis cycle, a laser is used to excite the dyes, and a high-resolution camera scans and detects the last base that was incorporated. This chemistry uses "*reversible terminators*" that are chemically unblocked at each step to ensure removal of the 3' terminal group that terminates polymerization, as well as the dye (**Figure 7**). This sequencing reaction is conducted simultaneously on a very large number (millions) of different template molecules spread on a solid surface.

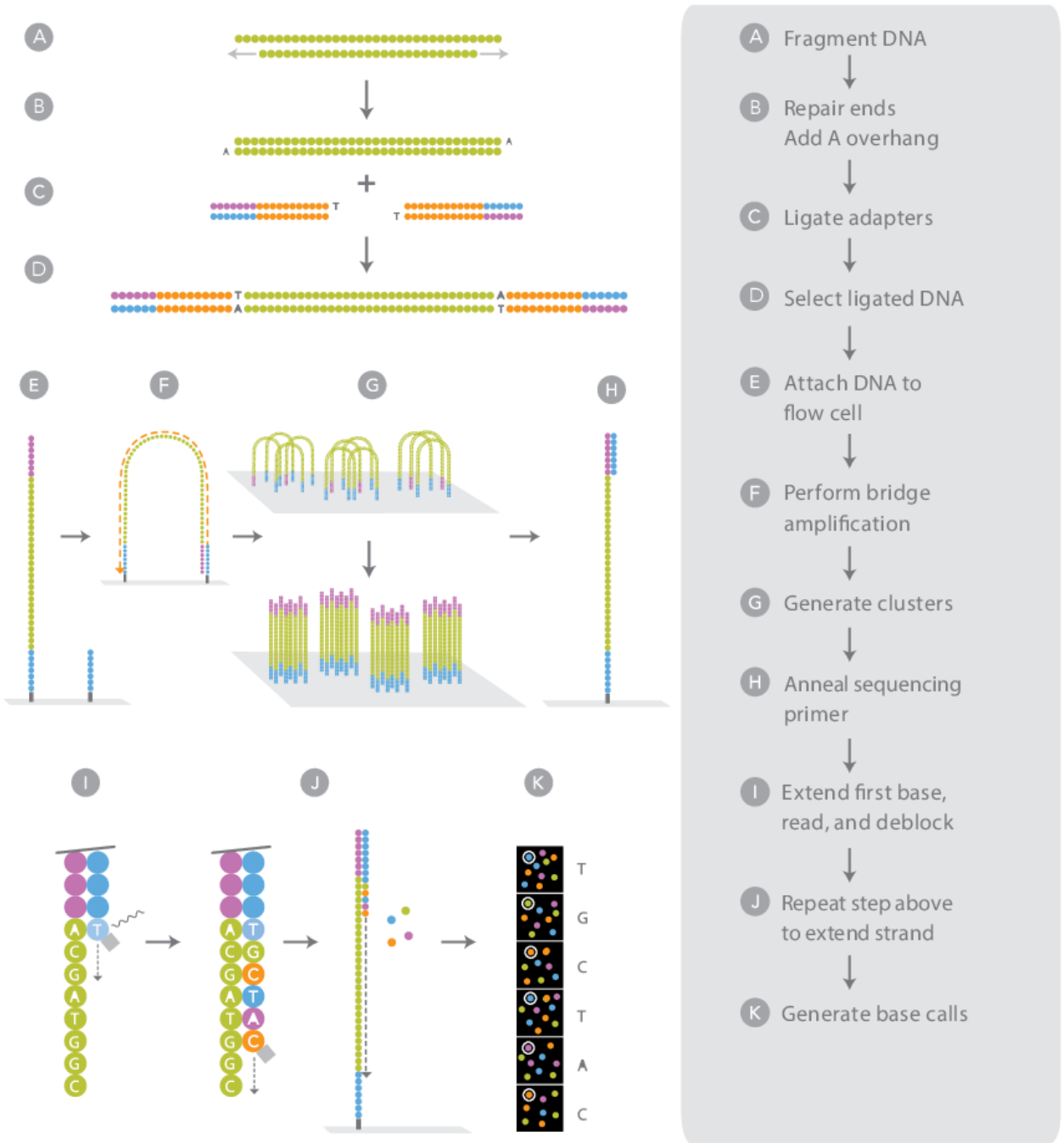

Figure 7. Typical Illumina sequencing by synthesis (SBS) workflow. The procedure for Illumina sequencing involves three major steps: 1) library preparation (steps A-E), 2) cluster generation (F-G) and 3) sequencing by synthesis (H-K). After universal double stranded adapter sequences are ligated to dsDNA (C), the library is spotted onto a glass slide, called a "flow cell", with millions of both forward and reverse immobilized adapter sequences (steps A-E). Cluster generation is initiated by the hybridization of denatured template DNA with its adapter sequences to the fixated complementary adapter oligonucleotides. In a process called bridge amplification, the single stranded protruding adapters of the DNA molecules hybridize to their nearby immobilized complements, shaping a bridge-like

structure (step F-G). Starting from a universal primer sequence in the adapters, the complementary strand of the DNA fragment is synthesized by DNA polymerases. Through repeated denaturing, bridging to nearby adapters and polymerization to dsDNA, one initial DNA molecule can produce a dense cluster of hundreds of clonal copies. After annealing a common primer to the protruding adapters, the sequencing of each cluster of DNA fragments begins (step H). In each sequencing 'cycle', the four dNTPs, are each reversibly labeled by a base-specific fluorescent dye and bound to a terminator molecule and are added to the flow cell. DNA polymerases add the respective nucleotides to complementary bases of the input DNA at the 3' end of the newly synthesized strands, while the attached terminator molecule ensures that only a single nucleotide is incorporated by preventing further strand elongation (step J). Lasers excite the different newly added fluorophores, and CCD chips image the characteristic signals that identify the incorporated base (step K). Due to the DNA amplification into clusters, the intensified signal of the whole cluster facilitates the correct dye identification. Before the addition of the next nucleotide, the terminating labels are enzymatically removed. Thus, each cycle consists of the addition of exactly one nucleotide to each DNA molecule. Image was taken from [58].

Although the fluorescent imaging system used in Illumina sequencers is not sensitive enough to detect the signal from a single template molecule, the major innovation of the Illumina method is probably the amplification of template molecules on a solid surface. The solid phase PCR process is called '*bridge amplification*' which creates approximately one million copies of each original template fragment. Illumina has improved its image analysis technology dramatically which allows for higher cluster density on the surface of the flow cell.

Illumina offers a variety of sequencing instruments (iSeq, MiniSeq, MiSeq, NextSeq, HiSeq, NovaSeq) and the selection of a particular instrument should be carefully evaluated to determine which would be the expected yield (sequencing depth) that will meet the requirements for the intended application. MiSeq for example can produce up to 15 Gb (~25 millions) of reads of 300 bp in length. Clustered fragments can be sequenced from both ends, which is known as '*paired-end sequencing'* (2x300 bp). HiSeq2500 can output up to 1000Gb per run but with a more limited fragment length (1x50 to 2x250 bp).

5.1.1 Other second generation sequencing technologies

Other technologies such as Roche/454 and Ion Torrent sequencing are still available, but are less frequently employed [36]. Roche/454 sequencing appeared on the market in 2005, commercialized by LifeSciences, and used the pyrosequencing technique which is based on the detection of pyrophosphate released after each nucleotide is incorporated in the newly synthesized DNA strand. The Roche/454 was able to generate relatively long reads which were easier to map to a reference genome. The main errors detected during sequencing were insertions and deletions due to the presence of homopolymeric regions. This platform was discontinued in 2013 when the technology became noncompetitive [37]. On the other hand, Life Technologies commercialized the Ion Torrent semiconductor sequencing technology in 2010. The principle behind this platform is similar to 454 pyrosequencing technology, instead of detecting pyrophosphate, it detects the hydrogen ion released during the sequencing process. The Ion

Torrent sequencers are capable of producing read lengths of 200 bp, 400 bp and 600 bp with throughput that can reach 10 Gb for the ion proton sequencer [38]. This platform is still commercialized but due to the ability of alternative third generation sequencing methods to achieve a greater read length, this technology may be best suited for small scale applications such as microbial genome sequencing, microbial transcriptome sequencing, targeted sequencing, etc.

Material to study

In this link you can learn and compare the different sequencing platforms offered by Illumina
https://www.illumina.com/systems/sequencing-platforms.html

5.2 Third Generation Sequencing Technologies

The evolution of DNA sequencing maintains an accelerated pace and the development of new platforms marked the arrival of the third generation at the end of 2011. **What is the main difference between second and third generation sequencing technologies?** Second-generation sequencing depends on the amplification of the template DNA and produces short reads of a few hundred base pairs in length. Third-generation single-molecule technologies can generate over 10,000 bp reads or map over 100,000 bp molecules because it does not depend on PCR amplification of the target molecule.

Here are some other characteristics that distinguish third generation sequencing technologies:

1) They work in real time
2) Sequencing is performed on individual molecules
3) Read lengths are in the order of tens of thousands of bases
4) They have higher error rates (~ 10%)
5) Can be used directly for DNA, RNA and proteins (Oxford Nanopore)

Third generation sequencing technologies maintain the downward trend in sequencing costs, although in some cases the prices of the sequencers or their commercial availability, is not as good as that of second generation platforms.

Now, let's review the principles behind some of the most commonly used third generation sequencing technologies.

5.2.1 Pacific Biosciences (PacBio) Single Molecule Real Time (SMRT) sequencing

This is probably the most established of the third generation sequencing technologies. The PacBio SMRT technology was commercially introduced in 2010 [39]. The SMRT technology sequences DNA using the approach *sequencing-by-synthesis*, and optically monitors fluorescently tagged nucleotides as they are incorporated into individual template molecules. The current instrument, the PacBio RS II, produces read lengths of up to ~100,000 bp with the greatest throughput (~8GB/day) of the currently available long-read technologies [40].

How does it work?

PacBio-SMRT sequencing captures sequence information during the replication process of a target DNA molecule. The template is called SMRTbell. A SMRTbell is a closed, single-stranded circular DNA that is created by ligating hairpin adaptors to both ends of a target double-stranded DNA molecule. In this technology, a polymerase enzyme is immobilized at the bottom of a ZMW unit (Zero Mode Waveguide) (**Figure 8A**). A ZMW is a nanophotonic visualization chamber that enables observation of individual molecules against the required background of labeled nucleotides while maintaining high signal to noise output. ZMW is a cylindrical metallic chamber approximately 70 nm wide, which is illuminated through a glass support creating a very small detection volume. When a sample of the SMRTbell is loaded to the chip, this diffuses into the ZMW with the single polymerase immobilized at the bottom, which binds to either end of the hairpin adaptor of the template DNA and starts the replication. Nucleotides diffuse in and out of the ZMW in microseconds. When the polymerase encounters the correct nucleotide, it takes several milliseconds to incorporate, time during which its fluorescent label is excited, emitting light (pulse) that is captured by a sensitive detector [41] (**Figure 8B**). After the incorporation, the label is clipped out and diffuses away. SMRT sequencing exploits the natural ability of polymerases to synthesize ~10 or more bases per second over thousands of continuous incorporations, resulting in high speed and long bead sequencing.

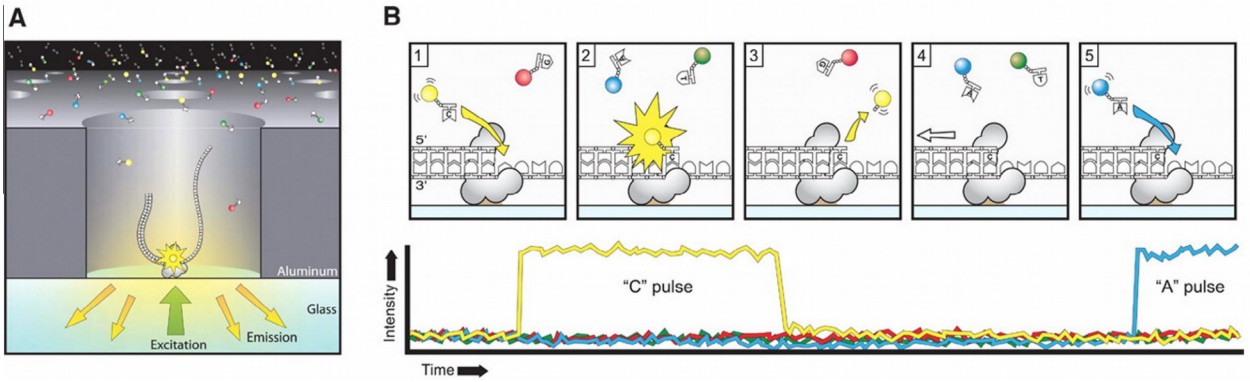

Figure 8. Single Molecule Real Time (SMRT) sequencing approach. A) A template DNA (SMRTbell) diffuses into a ZMW unit, and one of the adaptors binds the polymerase enzyme that is immobilized at the bottom of the chamber. B) Each of the four nucleotides have different fluorescent dyes and have different emissions when excited. As a nucleotide is maintained in the detection volume by the polymerase, a light pulse is produced that identifies the base that was incorporated. This figure was taken from [41]

The recently released PacBio Sequel instrument (2016) generates ~8 times more data than the original Sequel system. Reads have a raw error rate of 10% to 15% but nevertheless, several algorithmic techniques have been developed to improve the per-nucleotide accuracy to over 99.99% or more with sufficient coverage. Approximately 50X long read coverage is required for "self-correction" approaches, while lower coverage can be effectively used with hybrid error correction algorithms that leverage additional high coverage of Illumina short read sequencing to error correct the long reads.

The main limitation with PacBio sequencing is the cost relative to second generation approaches, which has limited its application to analysis of genomes. Nevertheless, to date hundreds of projects have successfully used PacBio sequencing, including nearly perfect assemblies or very high-quality genomes of microbes, fungi, plant and animal species, as well as very high-quality *de novo* assemblies of entire human genomes. In addition, given that the PacBio library preparation process does not utilize amplification techniques and the resulting library molecules are directly used as templates for the sequencing process, the quality of the DNA starting material is important. Obtaining high-quality, high-molecular-weight genomic DNA is critical for obtaining long read lengths and optimal sequencing performance [60].

5.2.2 Oxford Nanopore Technologies sequencing platform

The most recent third-generation technology was released by Oxford Nanopore Technologies in 2014. Their current instrument, the Oxford Nanopore MinION is a handheld device that sequences DNA by electronically measuring the minute disruptions to electric current as DNA molecules pass through a nanopore [42]. The read lengths of the currently available instruments are similar to those produced by PacBio, although to date the instrument has suffered from less accuracy and lower throughput which has limited its scope to sequencing small genomes, including *E. coli* (4.5Mbp) or yeast (12Mbp), or amplicons. Using error correction algorithms similar to those available for PacBio reads, the per-nucleotide accuracy of genomes sequenced using the MinION has been measured to be >99.95% [43]. Importantly, the instrument's small size and low cost have empowered it to be used for studies in very remote locations, including studying Ebola outbreaks in the field in West Africa [44, 45].

How does it work?

In the oxford nanopore system, a protein nanopore is inserted into an electrical resistant membrane created from synthetic polymers. A potential is applied across the membrane resulting in an ionic current flowing only through the aperture of the nanopore (**Figure 9a**). Single molecules that enter the pore cause characteristic disruptions in the current. By measuring that disruption, a molecule can be identified (**Figure 9b**). Oxford nanopore uses a '*strand sequencing method*' in which intact DNAs are passed through the nanopore and analyzed in real time. The strands of DNA to be sequenced are mixed with copies of an enzyme that carry the complex towards the nanopore [42]. The enzyme binds to a single-stranded leader at the end of the double stranded template molecule and unzips the template molecule feeding it through the nanopore [46].

As the DNA moves through the pore, a combination of nucleotides called k-mers within the narrow part of the pore creates a characteristic disruption in the electrical current. This information is used to determine the order of the bases on that DNA strand.

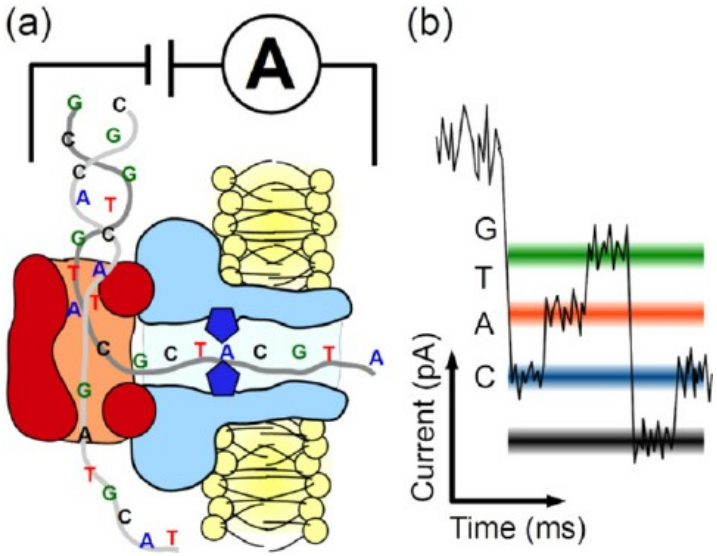

Figure 9. Cartoon depicting the sequencing technique of Oxford Nanopore Technology. In panel a) the enzyme in red unwinds a double stranded DNA molecule, feeding one strain through the protein pore. The nanopore contains a narrow site or constriction inside the channel (dark blue diamonds) which enables the unique shape of each nucleotide to produce a characteristic disruption in the electrical current. Panel b) shows a simplified scheme of the decoding method. Figure taken from:https://www.researchgate.net/figure/Scheme-depicting-the-sequencing-technique-of-Oxford-Nanopore-Technologies-a_fig3_271772842

Material to study

To learn more about Oxford Nanopore NGS Technology, please review the following publications:

- *The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community link*
- *Third-Generation sequencing in the clinical laboratory: exploring the advantages and challenges of nanopore sequencing link*

6.0 Library construction

After you make sure your DNA is of high quality, and prior to the sequencing step, DNA libraries need to be made. We will now review the most common library construction strategies currently being used.

6.1 *16S rRNA* gene amplicon metabarcoding

The 16S ribosomal RNA (*16S rRNA*) is a gene approximately 1,500 bp long that contains nine variable regions interspersed between conserved regions (**Figure 10**). Variable regions of the *16S rRNA* gene are frequently used for taxonomic profiling of diverse microbial populations. Which *16S rRNA* region to sequence is still an area of debate [47], and the selected region of interest might vary depending on aspects such as objectives, design, and sample type [48]. The principal difference lies between the length of the variable region, the phylogenetic resolution that can be achieved by sequencing the specific region and how conserved/biased the primers are for that region. Probably the most widely used region is the V4 (Earth microbiome project, http://www.earthmicrobiome.org/) since the primers have the least amplification bias among the different phyla; however, it has limited taxonomic resolution, often only reaching the genus/family level.

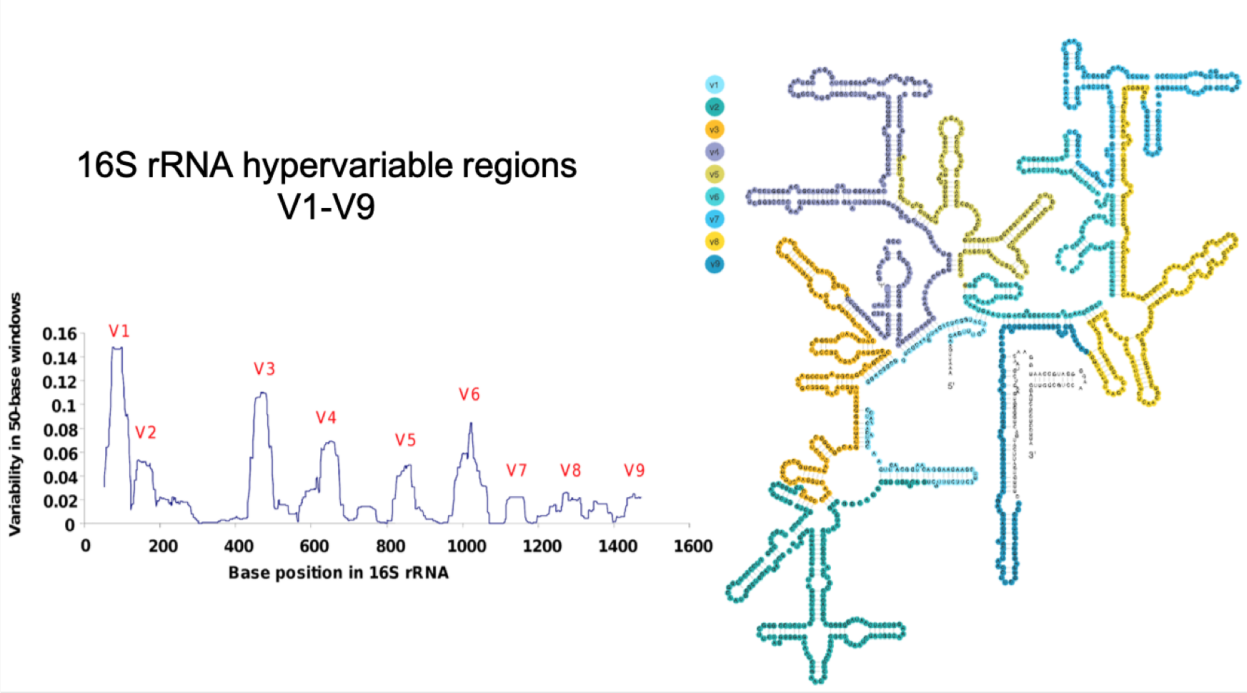

Figure 10. Hypervariable regions within the *16S rRNA* gene in Pseudomonas spp (left) and *E. coli* (right). The plotted line reflects fluctuations in variability amongst aligned *16S rRNA* gene sequences of 79 strains [50]. The right *16SrRNA* figure showing the different hypervariable regions highlighted. Figure was taken from http://themicrobiome.com/media/16S_viewer.cfm

The *16S rRNA* gene is probably the genetic marker most frequently used by microbial ecologists to study bacterial phylogeny and taxonomy. There are several reasons for this:

1. It is present in all bacteria, often existing as an operon
2. The function of *16S rRNA* gene has not changed over time, suggesting that random sequence changes are a more accurate measure of evolutionary time
3. Contains hypervariable regions interposed between highly conserved regions which allows the design of conserved primers that can target variable small regions of interest

In this tutorial we will discuss the sequencing strategy initially developed by Gregory Caporaso (2011) [55] and subsequently modified by James Kozich (2013) [49], targeting the V4 hypervariable region specifically and taking advantage of Illumina's MiSeq sequencing platform. The main distinguishing feature of this strategy is that researchers can use dual-indexes to multiplex and sequence a large number of samples in parallel. Indexes are unique sequences of 6-8 nucleotides that are used to identify the inserts (reads) that belong to each sample. **Dual indexing is a sequencing strategy that has distinct, unrelated index sequences for the i5 and i7 Index Read.**

Material to study

For more information regarding dual indexes (UDI) and associated library prep kits, please access the following link Understanding unique dual indexes (UDI)

Until recently, one of the most significant problems with the Illumina platforms was the ability to sequence samples with low genetic diversity, such as that commonly found with *16S rRNA* gene amplicons. To artificially increase the genetic diversity, it is a common practice to mix in a control library genomic DNA from the phage PhiX, such that 10-50% of the total loaded DNA is from the phage PhiX control library.

The PhiX Control library is derived from the small, well characterized bacteriophage PhiX genome. It is a concentrated Illumina's library (10 nM in 10 µl) that has an average size of 500 bp and consists of balanced base composition at ~45% GC and ~55% AT. The PhiX library serves as a calibration control to examine the overall performance of the Illumina sequencing platforms, e.g. cluster generation, sequencing and alignment. Also, for low diversity libraries as *16S rRNA* amplicons, the PhiX control library provides balanced fluorescent signals at each cycle to improve the overall run quality [56].

Libraries for *16S rRNA* gene amplicon sequencing are commonly amplified using specific primers that target one or several hypervariable regions in the gene. For example, the primers 515F (5′-GTGCCAGCMGCCGCGGTAA-3′) and 806R (5′-GGACTACHVGGGTWTCTAAT-3′) target the V4 region of the gene (**Figure 11**) which produces fragments of ~254 bps.

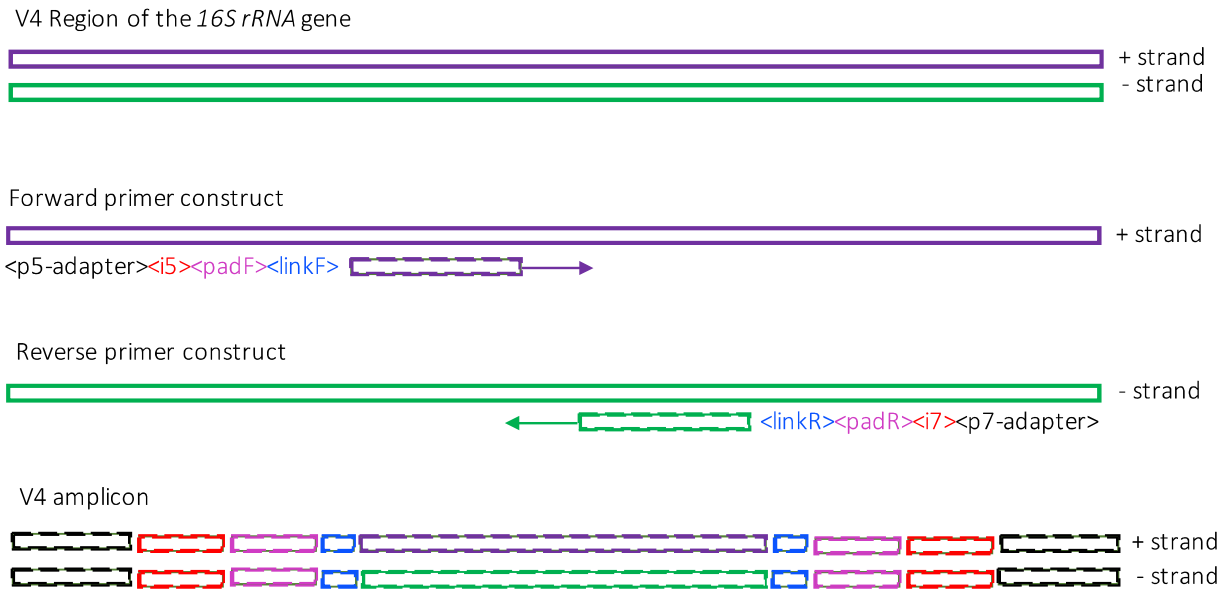

Figure 11. Dual-index, paired-read *16S rRNA* sequencing strategy. The primers specific to the V4 region of *16S rRNA* gene are shown in boldface black text, linkers are in blue, pads are in green, the sample-specific index region is in red, and Illumina's adapters (p5 and p7) are underlined. Image adapted from [49].

The dual-index paired-end sequencing approach uses primers consisting of:

1) The appropriate Illumina adapter (p5 and p7)
2) Two 8-nt index sequences (sample-specific indexes)
3) A 10-nt pad sequence
4) A 2-nt linker and
5) The *16SrRNA* gene-specific primer

PCR Amplicons are commonly generated using a high-fidelity polymerase (as AccuPrime; Invitrogen) and performed in duplicates or triplicates to minimize the effect of PCR bias. Specific amplification is usually verified by agarose gel electrophoresis and duplicated samples are pooled and purified using a DNA purification kit, such as the magnetic bead capture kit Ampure (by Agencourt). Purified amplicons are then quantified and pooled in equimolar concentration (usually 2nM) and sequenced together with 5% PhiX control DNA in Illumina sequencing platforms.

Material to study

- *A standard operating procedure from University of Michigan with detailed primer and index sequences for 16S rRNA gene amplicon library preparation and sequencing in a MiSeq Illumina sequencing platform is provided in the following Link*
- *The Earth Microbiome Project protocols and standards can be accessed in this site link*

6.2 Mock bacterial communities

A mock community is an artificially prepared bacterial community. As part of any *16S rRNA* gene microbiota study, it is important to include a control -mock community composed of predetermined ratios of DNA derived from a mixture of bacterial species [51]. This quality control step is important because it not only allows the quantification of sequencing error but also allows identification of bias introduced during the library preparation processes [52-54]. For example, a mock community containing bacterial taxons which are of specific interest to the research group can be used to calculate whether these taxons are likely to be over- or underrepresented in samples. Similar to mock communities, spike-in standards can also be used to analyze bias and

the reproducibility of methodologies. Pre-prepared bacterial communities could be available in two different ways:

- DNA mock communities
- Whole-cell mock communities

The whole-cell mock communities are useful for establishing the efficiency of the DNA extraction protocols, whereas DNA mock communities will only help assessing the efficiency of PCR, clean-up, sequencing, and analysis steps [48]. Mock communities are commonly available from the American Type Culture Collection (ATCC) and Zymo Research (ZymoBIOMICS). When planning a *16S rRNA* gene metabarcoding study, the inclusion of a control mock community is strongly encouraged.

6.3 Whole shotgun library preparation

There are different methodologies to prepare libraries, the product of these are thousands of inserts of ~500 bp for each sample that will then be sequenced. Generally, library preparation methods include either a mechanical or enzymatic fragmentation step of total DNA into small pieces, to which artificial generated sequences are added in order to fix these on the sequencing platform and process multiple samples in a single run (multiplexing). The starting material is usually around 1ng of total DNA.

There are several steps when preparing samples for sequencing. Broadly, these include:

- Library generation and indexing
- Purification and quality control
- Normalization and pooling
- Quantification
- Sequencing

6.3.1 Library generation and indexing

One of the most popular chemistries used currently is the Nextera XT DNA Library preparation protocol [61]. The Nextera XT Library Prep kit uses an engineered transposome system (**Figure 12A**) to tagment genomic DNA. **Tagmentation is a process aimed at fragmenting genomic DNA and then tagging the generated fragments with adapter sequences, all in one step** (**Figure 12B**). This step is done under heated conditions (5 minutes at 55°C) using 5µl of 1ng DNA input. Next, a limited-cycle PCR uses the index-adapters to amplify the insert DNA. The PCR step adds the index 1(i7) adapters, index 2 (i5) adapters, and sequences required for sequencing cluster generation (**Figure 12C**).

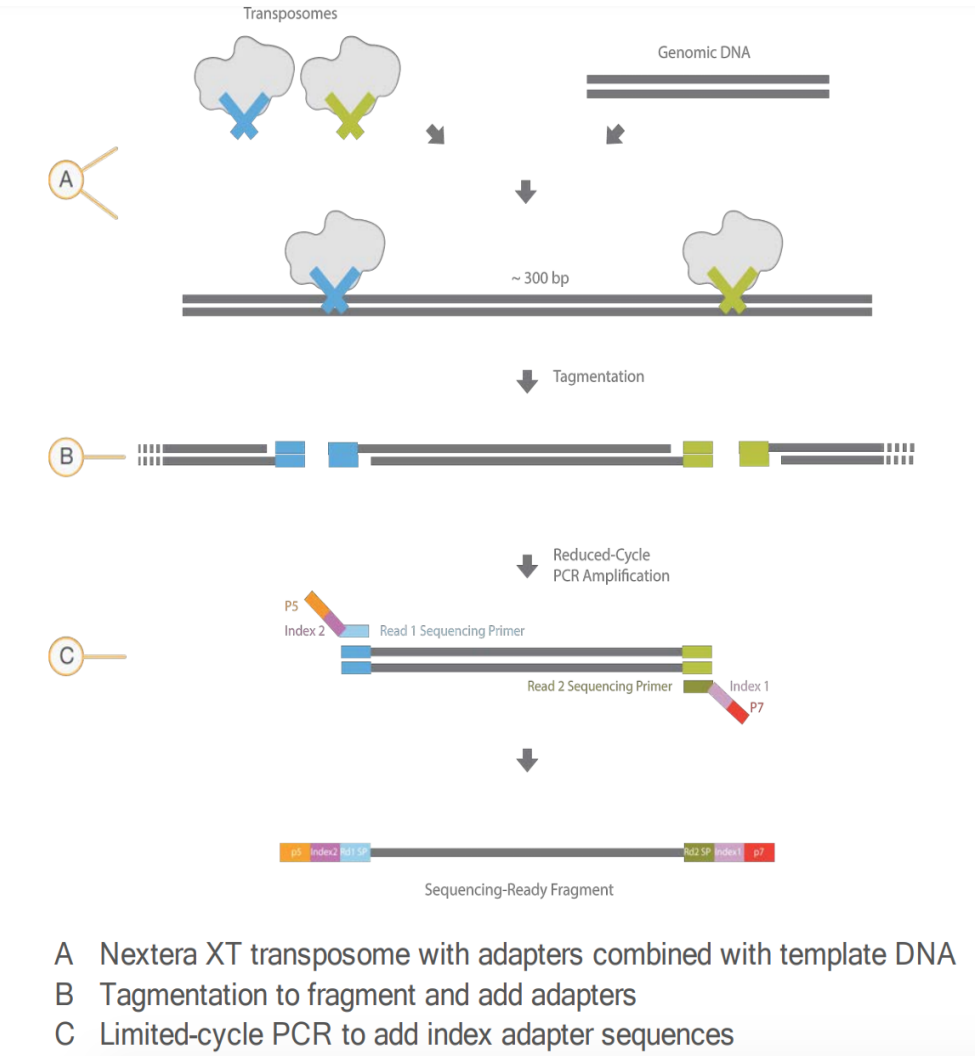

Figure 12. Cartoon showing how the Nextera XT assay works. In step A, genomic DNA is mixed with transposomes. A transposome is a system composed of engineered transposases (enzymes) that cleave and tag double-stranded DNA leaving a universal overhang. In step B, DNA is fragmented and adapter sequences are tagged simultaneously (process called tagmentation). In step C, index sequences are subsequently added on both ends of the DNA fragments through a limited-cycle PCR amplification. Index adapters are used to pool multiple libraries and sequence them in parallel. Image taken from https://genome.med.harvard.edu/documents/libraryPrep/IlluminaNexteraXTProtocol.pdf

The adapters (p5 and p7) are short oligonucleotides that contain artificial sequences complementary to the flow cell and are used to fix the inserts onto the surface (**Figure 10**). In addition, the adapters contain sequences complementary to the primers necessary for the sequencing process. Remember that **the indexes are unique sequences of 6-8 nucleotides that**

are used to identify the inserts (reads) that belong to each sample (Figure 13). Indexes are required for sequencing of pooled libraries on Illumina sequencing platforms.

Figure 13. Insert containing adapters and primers required for sequencing and demultiplexing. Image modified from https://genome.med.harvard.edu/documents/libraryPrep/IlluminaNexteraXTProtocol.pdf

6.3.2 PCR purification using paramagnetic beads

Once the PCR amplification is done, it is necessary to clean up and purify the libraries. In this step researchers commonly use the Agencourt AMPure XP beads (single-sided beads). The Agencourt AMPure XP utilizes an optimized buffer to selectively bind DNA fragments >=100bp to paramagnetic beads (beads magnetic only in magnetic fields). Excess primers, nucleotides, salts, and enzymes can be removed using a washing procedure with 70% ethanol (**Figure 14**).

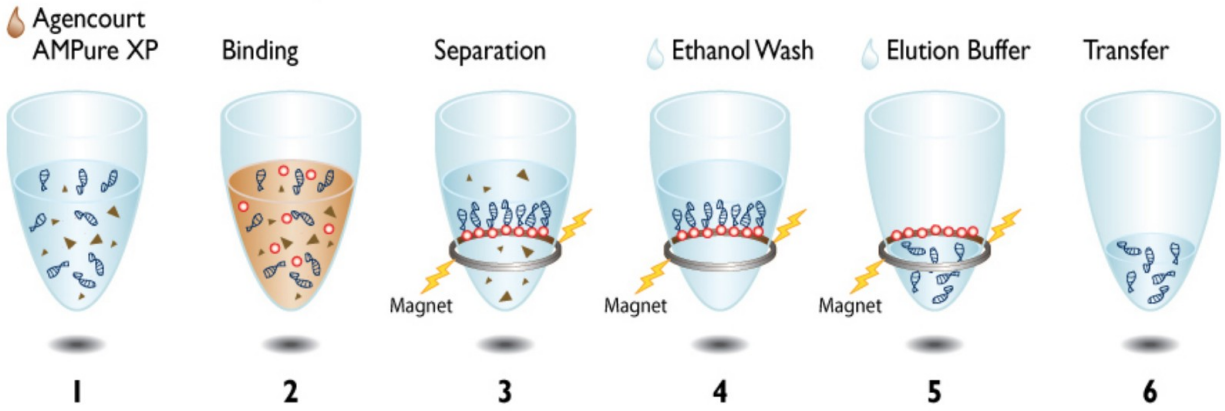

Figure 14. Cartoon showing the workflow for PCR purification with the Agencourt AMPure XP. PCR purification with Agencourt AMPure XP consist of six steps: 1) Adding 1.8 µL AMPure XP per each 1.0 µL of sample; 2) Binding DNA fragments to paramagnetic beads; 3) Separation of beads + DNA fragments from contaminants; 4) Washing beads + DNA fragments twice with 70% Ethanol to remove contaminants; 5) Eluting purified DNA fragments from beads; 6) Transferring clean library to a new tube. Image taken from https://www.beckmancoulter.com/wsrportal/techdocs?docname=B37419

Material to study

- *More details about AMPure XP can be found here link*
- *This guide is an overview of other library preparation applications and kits that are in common use for next-generation sequencing link*

6.3.3 Library quality control using chip-based capillary electrophoresis

To check the quality of the libraries produced for sequencing, researchers usually run 1 µl of undiluted amplified library on an Agilent Technology 2100 Bioanalyzer instrument (2100_Bioanalyzer) using a High_Sensitivity_DNA_kit. **The Bioanalyzer is a chip-based capillary electrophoresis machine to analyze RNA, DNA, and protein**.

In the Bioanalyzer instrument each sample is loaded onto a microfluidic chip along with reagents and standards. Each chip contains an interconnected set of microfluidic channels that are used for separation of the nucleic acid fragments by size using electrophoresis. The fragments will move through the channels at different speeds dependent on their size and charge with smaller fragments moving more quickly than longer fragments. Typical DNA libraries show a broad size distribution of ~250–1,000 bp (**Figure 15**). However, various libraries can be sequenced with average fragment sizes as small as 250 bp or as large as 1,500 bp.

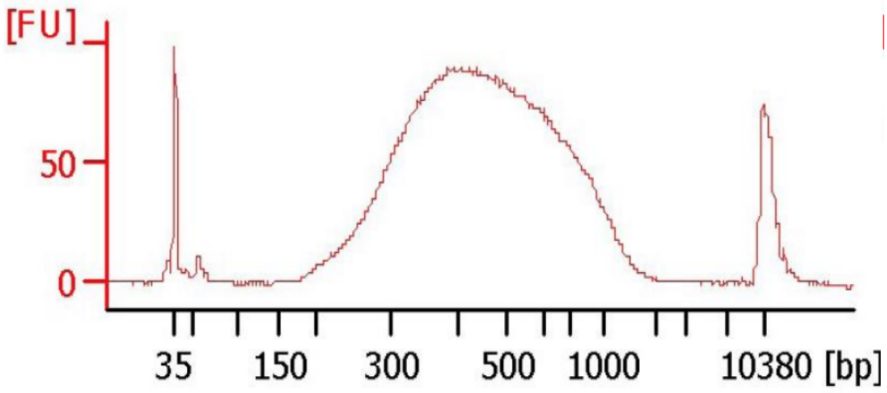

Figure 15. Insert size distribution for DNA fragments along with lower (35 bp) and upper (10,380 bp) molecular markers. The electropherogram displays a plot of fragment size in the x-axis (bp) versus fluorescence intensity (fluorescence units, FU) in y-axis. Peaks at 35 bp and 10380 bp represent lower and upper markers. Image taken from http://www.science.smith.edu/wp-content/uploads/sites/36/2015/09/LibraryQC_andTroubleshooting_June-2015.pdf

7. Conclusions

In this tutorial, we presented a concise overview of essential aspects in metagenomics, which included an introduction to the field, the considerations in the design and implementation of an experiment, sample processing, DNA extraction, sequencing technologies and library preparation. We reviewed the tremendous potential that metagenomics holds for identifying uncultivable microorganisms and for understanding the structure and function of naturally occurring microbial communities. By now, we know that there are many technical challenges associated with the design and implementation of a metagenomic study as well as many sources of variation that need to be controlled. In the coming years, new sequencing platforms will be more accessible to researchers producing a larger amount of data (in Terabytes), which will require the development of new approaches and applications capable of producing and analyzing this large amount of data.

Acknowledgements and funding

We thank Drs Maria Mercedes Zambrano (CorpoGen, Colombia), Rebecca Campos (Universidad de Costa Rica, Costa Rica), and Anisha Thanki (University of Leicester, UK) for valuable insights and edits on this document. We would also like to thank the members of the Group in Computational Biology and Microbial Ecology (BCEM) at Universidad de los Andes, Colombia for the evaluation and helpful discussions in the content of this tutorial. Funding for this tutorial was provided by the CABANA project, through an eLearning secondment.

Author Contributions

Dr. Alejandro Reyes was involved in conceptualization, funding acquisition, supervision, review and editing of this document. Dr. Angela Peña-Gonzalez was involved in formal structuring and writing of the tutorial.

Competing Interests

The authors declare no conflict of interest

References

1. Zengler K, Palsson BO. A road map for the development of community systems (CoSy) biology. Nat Rev Microbiol. 2012 Mar 27;10(5):366–72.
2. Riesenfeld CS, Schloss PD, Handelsman J. Metagenomics: genomic analysis of microbial communities. Annu Rev Genet. 2004;38:525–52.
3. Alain K, Querellou J. Cultivating the uncultured: limits, advances and future challenges. Extremophiles. 2009 Jul;13(4):583–94.
4. Schloss PD, Girard RA, Martin T, Edwards J, Thrash JC. Status of the archaeal and bacterial census: an update. MBio. 2016 May 17;7(3).
5. Danovaro R, Canals M, Tangherlini M, Dell'Anno A, Gambi C, Lastras G, et al. A submarine volcanic eruption leads to a novel microbial habitat. Nat Ecol Evol. 2017 Apr 24;1(6):144.
6. Chan CS, Chan K-G, Tay Y-L, Chua Y-H, Goh KM. Diversity of thermophiles in a Malaysian hot spring determined using 16S rRNA and shotgun metagenome sequencing. Front Microbiol. 2015 Mar 5;6:177.
7. Nolla-Ardèvol V, Strous M, Tegetmeyer HE. Anaerobic digestion of the microalga Spirulina at extreme alkaline conditions: biogas production, metagenome, and metatranscriptome. Front Microbiol. 2015 Jun 22;6:597.
8. Chen L, Hu M, Huang L, Hua Z, Kuang J, Li S, et al. Comparative metagenomic and metatranscriptomic analyses of microbial communities in acid mine drainage. ISME J. 2015 Jul;9(7):1579–92.
9. Tsementzi D, Wu J, Deutsch S, Nath S, Rodriguez-R LM, Burns AS, et al. SAR11 bacteria linked to ocean anoxia and nitrogen loss. Nature. 2016 Aug 11;536(7615):179–83.
10. Kerfahi D, Ogwu MC, Ariunzaya D, Balt A, Davaasuren D, Enkhmandal O, et al. Metal-Tolerant Fungal Communities Are Delineated by High Zinc, Lead, and Copper Concentrations in Metalliferous Gobi Desert Soils. Microb Ecol. 2019 Jul 4;
11. Marchesi JR, Ravel J. The vocabulary of microbiome research: a proposal. Microbiome. 2015 Jul 30;3:31.

12. Oulas A, Pavloudi C, Polymenakou P, Pavlopoulos GA, Papanikolaou N, Kotoulas G, et al. Metagenomics: tools and insights for analyzing next-generation sequencing data derived from biodiversity studies. Bioinform Biol Insights. 2015 May 5;9:75–88.

13. Mitchell AL, Scheremetjew M, Denise H, Potter S, Tarkowska A, Qureshi M, et al. EBI Metagenomics in 2017: enriching the analysis of microbial communities, from sequence reads to assemblies. Nucleic Acids Res. 2018 Jan 4;46(D1):D726–35.

14. Almeida A, Mitchell AL, Tarkowska A, Finn RD. Benchmarking taxonomic assignments based on 16S rRNA gene profiling of the microbiota from commonly sampled environments. Gigascience. 2018 May 1;7(5).

15. NIH HMP Working Group, Peterson J, Garges S, Giovanni M, McInnes P, Wang L, et al. The NIH human microbiome project. Genome Res. 2009 Dec;19(12):2317–23.

16. Knight R, Vrbanac A, Taylor BC, Aksenov A, Callewaert C, Debelius J, et al. Best practices for analysing microbiomes. Nat Rev Microbiol. 2018;16(7):410–22.

17. Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, et al. A survey of best practices for RNA-seq data analysis. Genome Biol. 2016 Jan 26;17(1):13.

18. Staley C, Sadowsky MJ. Practical considerations for sampling and data analysis in contemporary metagenomics-based environmental studies. J Microbiol Methods. 2018 Oct 1;154:14–8.

19. Fouhy F, Deane J, Rea MC, O'Sullivan Ó, Ross RP, O'Callaghan G, et al. The effects of freezing on faecal microbiota as determined using MiSeq sequencing and culture-based investigations. PLoS ONE. 2015 Mar 6;10(3):e0119355.

20. Rubin BER, Gibbons SM, Kennedy S, Hampton-Marcell J, Owens S, Gilbert JA. Investigating the impact of storage conditions on microbial community composition in soil samples. PLoS ONE. 2013 Jul 31;8(7):e70460.

21. Knudsen BE, Bergmark L, Munk P, Lukjancenko O, Priemé A, Aarestrup FM, et al. Impact of sample type and DNA isolation procedure on genomic inference of microbiome composition. mSystems. 2016 Oct 18;1(5).

22. Yang DY, Eng B, Waye JS, Dudar JC, Saunders SR. Technical note: improved DNA extraction from ancient bones using silica-based spin columns. Am J Phys Anthropol. 1998 Apr;105(4):539–43.

23. Salter SJ, Cox MJ, Turek EM, Calus ST, Cookson WO, Moffatt MF, et al. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. BMC Biol. 2014 Nov 12;12:87.

24. Evans GE, Murdoch DR, Anderson TP, Potter HC, George PM, Chambers ST. Contamination of Qiagen DNA extraction kits with Legionella DNA. J Clin Microbiol. 2003 Jul;41(7):3452–3.

25. Glassing A, Dowd SE, Galandiuk S, Davis B, Chiodini RJ. Inherent bacterial DNA contamination of extraction and sequencing reagents may affect interpretation of microbiota in low bacterial biomass samples. Gut Pathog. 2016 May 26;8:24.

26. Berthelet M, Whyte LG, Greer CW. Rapid, direct extraction of DNA from soils for PCR analysis using polyvinylpolypyrrolidone spin columns. FEMS Microbiol Lett. 1996 Apr

15;138(1):17–22.

27. Carter MJ, Milton ID. An inexpensive and simple method for DNA purifications on silica particles. Nucleic Acids Res. 1993 Feb 25;21(4):1044.

28. Chomczynski P, Sacchi N. The single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction: twenty-something years on. Nat Protoc. 2006;1(2):581–5.

29. Greenspoon SA, Scarpetta MA, Drayton ML, Turek SA. QIAamp spin columns as a method of DNA isolation for forensic casework. J Forensic Sci. 1998 Sep;43(5):1024–30.

30. Thoendel M, Jeraldo PR, Greenwood-Quaintance KE, Yao JZ, Chia N, Hanssen AD, et al. Comparison of microbial DNA enrichment tools for metagenomic whole genome sequencing. J Microbiol Methods. 2016 May 26;127:141–5.

31. Hunter SJ, Easton S, Booth V, Henderson B, Wade WG, Ward JM. Selective removal of human DNA from metagenomic DNA samples extracted from dental plaque. J Basic Microbiol. 2011 Aug;51(4):442–6.

32. Binga EK, Lasken RS, Neufeld JD. Something from (almost) nothing: the impact of multiple displacement amplification on microbial ecology. ISME J. 2008 Mar;2(3):233–41.

33. Gallagher SR. Quantitation of DNA and RNA with Absorption and Fluorescence Spectroscopy. Curr Protoc Immunol. 2017 Feb 2;116:A.3L.1-A.3L.14.

34. Singer VL, Jones LJ, Yue ST, Haugland RP. Characterization of PicoGreen reagent and development of a fluorescence-based solution assay for double-stranded DNA quantitation. Anal Biochem. 1997 Jul 1;249(2):228–38.

35. Kumar KR, Cowley MJ, Davis RL. Next-Generation Sequencing and Emerging Technologies. Semin Thromb Hemost. 2019 Oct;45(7):661–73.

36. Verma M, Kulshrestha S, Puri A. Genome Sequencing. Methods Mol Biol. 2017;1525:3–33.

37. Luo C, Tsementzi D, Kyrpides N, Read T, Konstantinidis KT. Direct comparisons of Illumina vs. Roche 454 sequencing technologies on the same microbial community DNA sample. PLoS ONE. 2012 Feb 10;7(2):e30087.

38. Cao Y, Fanning S, Proos S, Jordan K, Srikumar S. A Review on the Applications of Next Generation Sequencing Technologies as Applied to Food-Related Microbiome Studies. Front Microbiol. 2017 Sep 21;8:1829.

39. van Dijk EL, Jaszczyszyn Y, Naquin D, Thermes C. The third revolution in sequencing technology. Trends Genet. 2018 Jun 22;34(9):666–81.

40. English AC, Richards S, Han Y, Wang M, Vee V, Qu J, et al. Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. PLoS ONE. 2012 Nov 21;7(11):e47768.

41. Rhoads A, Au KF. Pacbio sequencing and its applications. Genomics Proteomics Bioinformatics. 2015 Oct;13(5):278–89.

42. Mikheyev AS, Tin MMY. A first look at the Oxford Nanopore MinION sequencer. Mol Ecol Resour. 2014 Nov;14(6):1097–102.

43. Leggett RM, Clark MD. A world of opportunities with nanopore sequencing. J Exp Bot. 2017 Nov 28;68(20):5419–29.

44. Hoenen T, Groseth A, Rosenke K, Fischer RJ, Hoenen A, Judson SD, et al. Nanopore sequencing as a rapidly deployable ebola outbreak tool. Emerging Infect Dis. 2016 Feb;22(2):331–4.

45. Quick J, Loman NJ, Duraffour S, Simpson JT, Severi E, Cowley L, et al. Real-time, portable genome sequencing for Ebola surveillance. Nature. 2016 Feb 11;530(7589):228–32.

46. Lu H, Giordano F, Ning Z. Oxford nanopore minion sequencing and genome assembly. Genomics Proteomics Bioinformatics. 2016 Oct;14(5):265–79.

47. Bukin YS, Galachyants YP, Morozov IV, Bukin SV, Zakharenko AS, Zemskaya TI. The effect of 16S rRNA region choice on bacterial community metabarcoding results. Sci Data. 2019 Feb 5;6:190007.

48. Pollock J, Glendinning L, Wisedchanwet T, Watson M. The madness of microbiome: attempting to find consensus "best practice" for 16S microbiome studies. Appl Environ Microbiol. 2018 Apr 1;84(7).

49. Kozich JJ, Westcott SL, Baxter NT, Highlander SK, Schloss PD. Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform. Appl Environ Microbiol. 2013 Sep;79(17):5112–20.

50. Bodilis J, Nsigue-Meilo S, Besaury L, Quillet L. Variable copy number, intra-genomic heterogeneities and lateral transfers of the 16S rRNA gene in Pseudomonas. PLoS ONE. 2012 Apr 24;7(4):e35647.

51. Salipante SJ, Kawashima T, Rosenthal C, Hoogestraat DR, Cummings LA, Sengupta DJ, et al. Performance comparison of Illumina and ion torrent next-generation sequencing platforms for 16S rRNA-based bacterial community profiling. Appl Environ Microbiol. 2014 Dec;80(24):7583–91.

52. Schloss PD, Gevers D, Westcott SL. Reducing the effects of PCR amplification and sequencing artifacts on 16S rRNA-based studies. PLoS ONE. 2011 Dec 14;6(12):e27310.

53. Parada AE, Needham DM, Fuhrman JA. Every base matters: assessing small subunit rRNA primers for marine microbiomes with mock communities, time series and global field samples. Environ Microbiol. 2016;18(5):1403–14.

54. Stämmler F, Gläsner J, Hiergeist A, Holler E, Weber D, Oefner PJ, et al. Adjusting microbiome profiles for differences in microbial load by spike-in bacteria. Microbiome. 2016 Jun 21;4(1):28.

55. Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Lozupone CA, Turnbaugh PJ, Fierer N, Knight R. Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. Proceedings of the national academy of sciences. 2011 Mar 15;108(Supplement 1):4516-22.

56. URL accessed on Feb 03, 2020. https://support.illumina.com/bulletins/2017/02/what-is-the-phix-control-v3-library-and-what-is-its-function-in-.html

57. URL accessed on December 13, 2019. https://patents.google.com/patent/US5808041A/en
58. URL accessed on January 24, 2020. https://pdfs.semanticscholar.org/02d1/75e45823ad464134ba31391e590f6db9ca47.pdf?_ga=2.15645043.493250237.1583161019-1795296374.1583161019
59. URL accessed on October 12, 2019. https://www.biotek.com/resources/application-notes/nucleic-acid-purity-assessment-using-a260/a280-ratios/
60. URL accessed on August 04, 2020. http://med.stanford.edu/content/dam/sm/gssc/documents/Pacbio-Guidelines-SMRTbell-Libraries-v1.0.pdf
61. URL accessed on August 04, 2020. https://support.illumina.com/content/dam/illumina-support/documents/documentation/chemistry_documentation/samplepreps_nextera/nextera-xt/nextera-xt-library-prep-reference-guide-15031942-05.pdf