Textbook chapter in the use of Metagenomics for the Identification and Classification of Uncultivable Bacterium

**The Problem:**

The identification of these uncultivable bacteria can be extremely difficult, and getting sequence data can be even harder. The traditional way to identify and sequence bacteria is to grow them in the lab, view under the microscope, and then extract DNA for sequencing. This is not possible with bacteria that are deemed uncultivable, as they have extremely complex growth requirements that are unreasonable or impossible to recreate in a laboratory setting. It is often that most of the growth requirements are unknown until after genetic sequencing, or extensive laboratory testing. It is estimated that only about two percent of all bacteria can be grown under standard laboratory conditions. While many of these bacteria are found in the soil, or in extreme environments, many of them can be found in human beings. Only about fifty percent of the bacteria located in the mouth can be cultivated in the laboratory, and even some bacteria that cause diseases, still cannot be cultivated. For example the cause of syphilis, Treponema pallidum, could not be cultivated in the laboratory until very recently in 2018, even though it was discovered in 1905. Many of these bacteria can also require complex substrates that can only be received from other bacteria, and are this way codependent on other bacterium.

**Introduction to Metagenomics:**

One of the way to sequence and identify these bacteria is to conduct environmental samples, of where the organism of interest might be found. This type of sampling and sequencing is called metagenomics. **Metagenomics is the study of genetic material recovered directly from environmental samples.** While this method is nearly the only way to identify there uncultivable bacteria, there are still several problems that need to be addressed. The first is that there is extremely little control of what bacteria ends up in your sample, as there is no way to view the bacteria in your sample, it is mostly based off of where the sample is collected from. Also, there is often an extremely low concentration of bacteria present in these samples, and can be extremely difficult to obtain a substantial amount of DNA, and can lead to sequencing problems. **Genomic Sequencing is the use of highly technical machines and techniques in which a organisms genome is chopped up into pieces, and then these shorter pieces have their nucleotide sequence read.** Once these hurdles have been dealt with, there is still the issue that this is a metagenomics sample. This can be a problem because there could be thousands of different bacteria in the sample, and are not differentiated while sequencing, so their genomic data is all mixed together.

**Metagenomic Filtering:**

The complexity of a Metagenomics Assembly can be resolved by a process called binning. An **Assembly is a term used to describe the pieces of DNA that were put back together after being sequenced.** In the past the higher cost of sequencing combined with the relative randomness of this sequencing made metagenomics sequencing of environmental samples to large of a risk. There was also the problem of computing power historically, and until recently obtaining the computing power to work on bioinformatical projects were extremely expensive. **Bioinformatics is the science of collecting and analyzing complex biological data such as genetic codes.** There was the problem of database quality, which is a huge factor in the field of bioinformatics, as it allows for the identification of genes and other sequences, without de novo analysis. **De novo** is a common term used in the field of bioinformatics, and mean to **starting from the beginning.** Today the price of sequencing has become relatively affordable, along with access to computing power as more and more universities build supercomputers, and with the lowered cost of sequencing databases are getting more complete as more and more people conduct sequencing efforts.

**Recent Advancements of Metagenomics**

 While many of the problems of metagenomic sequencing has been solved my modern technology, the biggest of the problems still exists, which is choosing which bacteria you sample. Samples of most often taken form environments that may be extreme, or abnormal, or from environments that would just be near impossible to recreate in the laboratory. These studies most often have the goal of identifying novel bacteria, and finding out some of the biochemical pathways, in the effort to categorize all bacteria. One other main purpose of this is to find bacteria that may have novel characteristics that may be useful to humans in some way. This may include some pathway that is extremely useful, or a new type of antibiotic.

Conquering Low DNA Concentration:

The next hurdle that has to be overcome is that of the quantity of DNA that is present. The amount of total DNA present may be extremely high, but the concentration of specific bacteria may be quite low. This means that it can be extremely difficult to sequence, and get high enough coverage to obtain full genomes of some species. This is sometimes overcome by amplification in the laboratory, or just sequencing at a much higher coverage level, to try to get enough data for the underrepresented bacteria. As the cost of sequencing goes down this option becomes more and more viable, especially as amplification can be biased. Sequencing technology has come so far that it is even possible now to obtain a full genome from a single cell.

The Process Binning:

The next issue arises when **the sorting of the sequences begins, which is called binning**, as there may be thousands of different genomes represented in one metagenomics assembly. This process is done using a verity of techniques using a mixture of complex algorithms, and sequence identification. These algorithms are usually based around the principles of grouping sequences or **contigs, which is the name given to the pieces of DNA that were assembled,** based on sequence composition, or sequence alignment, or a combination of both. Sequence composition usually refers to guanine-cytosine content of the genome, also known as **GC content**, and other specific markers. Sequence alignment refers to piecing sequences together, or against one another, moving matches into more and more specific bins, or groups. Binning has a variable success rate, often based on how structurally different the genomes of the organisms are, or how complete the genomes are. This can lead to bins being extremely specific, sometimes almost down to the species, but poorer bins can be as high as phylum. Binning can also be done before or after assembly of contigs, but when using raw sequences it can be much harder to get good alignments, leading to poorer bins.

Annotation of Genomes:

After individual genomes are created then it comes down to the process of analyzing the genome to find biochemical pathways, and structural info on the bacteria, along with constructing phylogenetic trees. All of these things are examples of annotation, **Annotation is the process of identifying the locations of genes and all of the coding regions in a genome and determining what those genes do.** The phylogenetic tree is based off of highly conserved genes, like the gene that codes for the 16s ribosomal RNA. These trees are typically based off of more than just the 16s gene, as trees constructed on relatively few genes can contain inaccuracies. The phylogenetic tree allows for a starting point to know what is closely related to the bacterium, and if it is a candidate for a newly discovered bacterium. If a bacterium has promise as biochemical pathway analyses can be done to see what molecules it can utilize, what type of respiration occurs, what type of environments it most likely inhabits, its shape, size, and rate of reproduction, also which what molecules it can and cannot synthesize. This is mostly done by calling genes from the genome, which is relatively easy as bacteria have a single circular chromosome, with no introns. These genes that are called then can be compared to a database containing the common genes for pathways most commonly found in bacteria. While the sequences for the genes in the pathway may not be an exact match, when they have a extremely high similarity, it can be concluded that the bacterium is capable of utilizing these pathways. This allows us to determine what the bacterium may be useful for, or at least how to categorize it. As more and more bacterium are sequenced these databases will become more and more useful, as there will be more sequences to compare against. With all of this information a precise description of the bacterium can be used to categorize it, allowing it to be identified later, or if it was found to be useful for some other process. This process of metagenomics can also be used in the medical world to sequence bacterium that are uncultivable, or in the world of studying microbiomes. Microbiome studies usually focus on the diversity of the organisms present so will often focus their segueing on the 16s ribosomal RNA genes.

Refrences:

Wade W. Unculturable bacteria--the uncharacterized organisms that cause oral infections. *J R Soc Med*. 2002;95(2):81-3.

National Research Council (US) Committee on Metagenomics: Challenges and Functional Applications. The New Science of Metagenomics: Revealing the Secrets of Our Microbial Planet. Washington (DC): National Academies Press (US); 2007. 1, Why Metagenomics? Available from: <https://www.ncbi.nlm.nih.gov/books/NBK54011/>

 Youssef, Noha H. ; Farag, Ibrahim F. ; Hahn, C. Ryan ; Premathilake, Hasitha ; Fry, Emily ; Hart, Matthew ; Huffaker, Krystal ; Bird, Edward ; Hambright, Jimmre ; Hoff, Wouter D. ; Elshahed, Mostafa S.. (2019). “Candidatus Krumholzibacterium zodletonense gen. nov., sp nov, the first representative of the candidate phylum Krumholzibacteriota phyl. nov. recovered from an anoxic sulfidic spring using genome resolved metagenomics.” Systematic and Applied Microbiology, 42(1), 85-93.

Handelsman, Jo. “Metagenomics: application of genomics to uncultured microorganisms” Microbiology and molecular biology reviews : MMBR vol. 68,4 (2004): 669-85.