The global microbial biomass constitutes nearly one half of all biomass on Earth (Whitman et al., 1998). They number as the most common living organisms, and are responsible for driving the metabolic engines that enable all other life to exist. In addition, since the discovery of microbial entities as pathogens that cause disease, they have become key subjects towards the study of illness. They are the world’s oldest ancestors of life, having existed long before any human or animal has, and hold the key to unlocking the grand question of how we came to be. As such, great importance has been placed on the need to study and classify microbes to better understand the lines that tie our existence and history in with them. Yet, despite the importance of classifying the grand diversity of microbes, there remain significant challenges in doing so. In this essay, in fulfilling the broader aim to generate a synthesis of what we have learned in MICB 425, I will be detailing the challenges involved in defining a microbial species, in the context of our current reliance on 16S rRNA sequencing for classification of Operational Taxonomic Units (OTUs). Following this, I will be discussing the phenomenon of Horizontal Gene Transfer (HGT), specifically how it hinders our ability to trace the origin of species, and along with that, the origin of metabolic pathways. Finally, I will explain why it is necessary to have a clear definition, even in the context of a shifting definition with the advent of amplicon sequence variants (ASVs) about to replace OTUs as the standard unit of a species.

The advent of genome sequencing technology has enabled humanity the ability to study microbes in ways that have previously been impossible through culture-based identification methods alone. By-passing the cultivation limitation in detecting microbial life has allowed us to estimate that we had only been able to capture about 1% of all microbial diversity up to that point (Chen et al., 2013). 16S rRNA sequencing seeks to compare the sequences of RNA found within the 16S portion of the small ribosomal subunit. This sequence is used because all translationally active microbial life contains such a sequence. This has two primary benefits, the first being that sequences within this region are conserved for taxonomically similar organisms, allowing us to construct a phylogenetic tree that links all microbial life together through tracing the lineage of inheritance towards the common ancestor of all life on Earth. The second benefit is that only studying organisms that contain translationally active machinery allows us to narrow the scope of study to microbes that contribute towards the global metabolic engines that drive Earth’s biogeochemical cycles (Falkowski, Fenchel and Delong, 2008). From this, the concept of OTUs representing species came to be. The current standard for sequence similarity between OTUs is 97%. This currently defines what a species is in the status quo. To derive the OTUs present in each location, samples of the environment are gathered, and genetic material needs to be isolated from individual cells. PCR amplification is applied to extracted DNA, and sequences are processed through bioinformatics pipelines for compositional analysis of samples (Finotello, Mastrorilli and Di Camillo, 2016).

Challenges in defining a species within the microbial world thus stem from this definition of a species. The primary concern in defining a species through any measure is if that definition of a species accurately reflects what we would recognise as a species in other context. This is illustrated by the rift that we see when we compare the microbial ecology demarcation of a species against other taxonomic disciplines. If we were to apply the same standard of 97% sequence similarity to animal classification, all primates would be considered a single species (Lumen learning, 2018). As such, there has been suggestion for updating the current 97% identity threshold to define 16s rRNA OTUs (Edgar, 2017). This links towards the challenge in defining a species in the phenotypic differences within what we would recognise as a species. Could microbes of the same species display traits that are dissimilar enough that under different contexts would be considered a different species without much controversy? The answer appears to be an affirmative when we examine strains. There is an enormous amount of strain to strain variation, specifically when we look at pathogenic strains vs commensal ones. For example, when the genome of *E. coli* CFT073, a pathogen, EDL933 an enterohemorrhagic strain and MG1655 laboratory strain were compared, only 39.2% of their combined protein set, representing the functional elements within these lifeforms, were common between all three strains (Welch et al., 2002). This presents a case towards an update of our definition for an OTU to be aligned towards strains, with the justification being that this model would be more biologically informative (Edgar, 2017). However, there are also problems with this definition update. Some strains have similar phenotypes to each other, thus creating a situation where microbes that belong to the same OTU would be incorrectly assigned to a separate taxonomic unit. The next challenge lies in the use of 16S rRNA sequences for deriving our definitions of a species. It is clear, that our current definition of a microbial species lacks the biological relevance that is needed to exert translational change from our study of ecology, but on a broader level even the use of 16s rRNA has limitations. 16s rRNA makes up a small part of an organism’s entire genome, and apart from recognition of mRNA has no direct impact on the catalytic activity in a cell. It is therefore no surprise that a presentation of multi-omic sequence information lends a great deal more information towards the activity levels of individuals within a specific taxonomic unit than the use of 16S rRNA (Hawley et al., 2017). A similar challenge in the definition of a species is shown in differences that emerge in the methods employed to receive such sequences. To this end, the sequencing depth, sample preparation and storage, bioinformatics pipelines and even the sequencing platformed used to derive sequences impact how many species are observable in the world. This is evidenced by the fact that diversity estimates differ in varying degrees according to variations in any step of the study (Allali et al., 2017). Furthermore, diversity estimates employed for use in microbial ecology were originally developed for macroecology. Errors in sequencing techniques such as pyrosequencing, originally introduced as a means of increasing sequencing depth, lead to inflation of diversity estimates (Kunin et al., 2010). As such, we currently are largely ignorant of the biases and errors that are present in such estimates (Finotello, Mastrorilli and Di Camillo, 2016). What this illustrates is that our definition of a species is not yet fixed, and will depend on the methods used to study the 16s rRNA of microbes. Further complicating matters is the fact that pipelines fundamentally differ in the ways in which they classify OTUs. Some approaches make use of a reference sequence to cluster OTUs based on similarity to the reference, while others cluster OTUs according to sequence distance (Chen et al., 2013, Callahan, McMurdie and Holmes, 2017). When OTU clustering is not done referencing a database, there is no ability to cross-compare data between studies. These factors in combination make it difficult not only to classify species of microbes, but to also derive the ancestry and therefore trace the origins of metabolic activity in biogeochemical cycles.

How does HGT influence our understanding of this problem? HGT creates an evolutionary dynamic that is different from our typified understanding of phylogeny. At the core of this idea lies our construction of the microbial phylogenetic tree. The way this tree is drawn indicates that all microbial life originates from a singular common ancestor, and by that same logic, a single cell thought to be the first living form on Earth. HGT disrupts this by forcing us to consider the metabolic implications of such organisms solely bearing the burden to survive an extremely hostile environment which was early Earth. To do this, metabolic machinery to draw energy from a nutrient source must have existed in its entirety to enable a fully functional cell to exist from the onset of life. It is more likely that the earliest organisms evolved from what would tread the boundary of being alive and not, to quickly split the burden of metabolic processing to derive energy from an elemental source through the mechanism of HGT. As such, the earliest lifeforms need to be considered not as a single cell but a community of cells, promiscuously sharing information through HGT (Nisbet and Sleep, 2001). Therefore, even at the most logistically simple definition of a species: that of the universal common ancestor, HGT complicates our definition of the term “species”. Further down with increasing evolutionary complexity, HGT links parts of the tree that would otherwise occur far apart and clearly distinct from each other (Nisbet and Sleep, 2001). Between strains of what would otherwise be considered the same species, genomic islands residing within the common backbone are acquired via HGT (Welch et al., 2002). These occurrences provide indication towards the existence of genetic reservoirs, which are responsible for the storage and distribution of genes across species (Sogin et al., 2006). Together, these dynamics force us to imagine the true global phylogenetic distribution as more of a mangrove forest or a delta as opposed to a tree (Nisbet and Sleep, 2001), where mechanisms for both vertical and horizontal gene transfer exist, which further limit our ability to state a firm definition of a species.

Given the challenges that currently exist from the use of OTUs as the definitional markers of microbial species, and taking into consideration the increased complexity of this issue presented by HGT, we need to question the importance of maintaining a clear definition of species. To address this, we must first consider why we require a definition of microbial species, and how we apply this definition towards practical applications. Part of the reason behind why classification of microbes is important is the sheer number of them. It would be physically impossible to analyse each cell we encounter daily to derive their nature and what they are. Classification, especially relating to metabolic pathways that reside within cells, allows us to derive the nature of microbes residing within an ecological niche (BBC, 2018). From the generalized knowledge of an environmental niche, we can then expedite the translation of knowledge pertaining to specific microbes toward actions. This saves time, which can be illustrated to be important in two major fields, environmental engineering and medicine. For the former, to effectively mediate the effects of climate change on specific regions of agricultural importance, a sound understanding of the microbial composition within such an area will alert us to the fact that a disruption in the food chain from the primary production level is occurring (Torres-Beltrán et al., 2017, Hawley et al., 2017). In medicine, understanding the differences in presence between a pathogenic, enterohemorrhagic, or commensal strain can provide more insight as to mechanisms behind resistance to therapy (Welch et al., 2002). With the need to have a clear definition of microbial species established, we also find optimism in the fact that the development of new tools allows us to better form this definition to more accurately reflect reality. The creation of ASVs overcomes some of the limitations imposed by OTUs. De novo OTUs rely on emergent features of a data set defined for each, while closed reference OTUs rely on referencing a database to compare percentage similarity for classification of microbes. ASVs do not rely arbitrarily by difference in percentage to assign similarity to sequences. They instead infer biological sequences in samples prior to the introduction of amplification and sequencing error (Callahan, McMurdie and Holmes, 2017). As such, they carry the advantages present in biological relevance present in de-novo methods, whilst allowing for cross reference of studies. This will allow for future studies in microbial ecology to adopt a more standardized comparison, with improvements to biological relevance in the classification of rare species (Sogin et al., 2006).

In conclusion, the introduction of sequencing technology has allowed us to study the vast diversity of microbes in depth and accuracy never before possible. However, significant challenges persist in our current understanding of what defines a species. This, compounded by the fact that HGT plays a key role in the distribution of genes, means that it is unlikely that we will ever come to a perfect definition for what defines a species. However, the practical implications of refusing to define species altogether would be problematic. To this end, the introduction of new technologies that allow us to reach a clearer, more precise and more applicable understanding of what defines a species.

References

Whitman, W., Coleman, D. and Wiebe, W. (1998). Prokaryotes: The unseen majority. *Proceedings of the National Academy of Sciences*, 95(12), pp.6578-6583.

Chen, W., Zhang, C., Cheng, Y., Zhang, S. and Zhao, H. (2013). A Comparison of Methods for Clustering 16S rRNA Sequences into OTUs. *PLoS ONE*, 8(8), p.e70837.

Falkowski, P., Fenchel, T. and Delong, E. (2008). The Microbial Engines That Drive Earth's Biogeochemical Cycles. *Science*, 320(5879), pp.1034-1039.

Finotello, F., Mastrorilli, E. and Di Camillo, B. (2016). Measuring the diversity of the human microbiota with targeted next-generation sequencing. *Briefings in Bioinformatics*, p.bbw119.

Courses.lumenlearning.com. (2018). *Classification of Microorganisms | Boundless Microbiology*. [online] Available at: https://courses.lumenlearning.com/boundless-microbiology/chapter/classification-of-microorganisms/ [Accessed 22 Apr. 2018].

Edgar, R. (2017). Updating the 97% identity threshold for 16S ribosomal RNA OTUs. *Bioinformatics*.

Welch, R., Burland, V., Plunkett, G., Redford, P., Roesch, P., Rasko, D., Buckles, E., Liou, S., Boutin, A., Hackett, J., Stroud, D., Mayhew, G., Rose, D., Zhou, S., Schwartz, D., Perna, N., Mobley, H., Donnenberg, M. and Blattner, F. (2002). Extensive mosaic structure revealed by the complete genome sequence of uropathogenic Escherichia coli. *Proceedings of the National Academy of Sciences*, 99(26), pp.17020-17024.

Allali, I., Arnold, J., Roach, J., Cadenas, M., Butz, N., Hassan, H., Koci, M., Ballou, A., Mendoza, M., Ali, R. and Azcarate-Peril, M. (2017). A comparison of sequencing platforms and bioinformatics pipelines for compositional analysis of the gut microbiome. *BMC Microbiology*, 17(1).

Callahan, B., McMurdie, P. and Holmes, S. (2017). Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *The ISME Journal*, 11(12), pp.2639-2643.

Hawley, A., Torres-Beltrán, M., Zaikova, E., Walsh, D., Mueller, A., Scofield, M., Kheirandish, S., Payne, C., Pakhomova, L., Bhatia, M., Shevchuk, O., Gies, E., Fairley, D., Malfatti, S., Norbeck, A., Brewer, H., Pasa-Tolic, L., del Rio, T., Suttle, C., Tringe, S. and Hallam, S. (2017). A compendium of multi-omic sequence information from the Saanich Inlet water column. *Scientific Data*, 4, p.170160.

Kunin, V., Engelbrektson, A., Ochman, H. and Hugenholtz, P. (2010). Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates. *Environmental Microbiology*, 12(1), pp.118-123.

Nisbet, E. and Sleep, N. (2001). The habitat and nature of early life. *Nature*, 409(6823), pp.1083-1091.

L. Sogin, M., G. Morrison, H., A. Huber, J., Mark Welch, D., M. Huse, S., R. Neal, P., M. Arrieta, J. and J. Herndl, G. (2006). Microbial diversity in the deep sea and the underexplored ‘‘rare biosphere’’. *PNAS*, 103(32), pp.12115-12120.

Bbc.co.uk. (2018). *BBC - KS3 Bitesize Science - Variation and classification : Revision, Page 7*. [online] Available at: http://www.bbc.co.uk/bitesize/ks3/science/organisms\_behaviour\_health/variation\_classification/revision/7/ [Accessed 23 Apr. 2018].

Callahan, B., McMurdie, P. and Holmes, S. (2017). Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *The ISME Journal*, 11(12), pp.2639-2643.

Torres-Beltrán, M., Hawley, A., Capelle, D., Zaikova, E., Walsh, D., Mueller, A., Scofield, M., Payne, C., Pakhomova, L., Kheirandish, S., Finke, J., Bhatia, M., Shevchuk, O., Gies, E., Fairley, D., Michiels, C., Suttle, C., Whitney, F., Crowe, S., Tortell, P. and Hallam, S. (2017). A compendium of geochemical information from the Saanich Inlet water column. *Scientific Data*, 4, p.170159.