2014

Genome Jigsaw: Implications of 16S Ribosomal RNA Gene Fragment Position for Bacterial Species Identification

Jennifer Mitchell
Wilfrid Laurier University, jenny.dobson.mitchell@gmail.com

GENOME JIGSAW: IMPLICATIONS OF 16S RIBOSOMAL RNA GENE FRAGMENT POSITION FOR BACTERIAL SPECIES IDENTIFICATION

By

Jennifer Mitchell

(HBSc. Biomedical Science, University of Waterloo, 2012)

THESIS

Submitted to the Department of Biology

Faculty of Science

in partial fulfillment of the requirements for the

Master of Science in Integrative Biology

Wilfrid Laurier University

2014

Abstract

The 16S rRNA gene is present within all bacteria, and contains nine variable regions interspersed within conserved regions of the gene. While conserved regions remain mostly constant over time, variable regions can be used for taxonomic identification purposes. Current methodologies for characterizing microbial communities, such as those used to study the human microbiome, involve sequencing short fragments of this ubiquitous gene, and comparing these fragments to reference sequences in databases to identify the microbes present. Traditionally, whole 16S rRNA sequences with more than 97% sequence identity (id) are assigned to a single operational taxonomic unit (OTUs); each OTU being a proxy for a single species. However, because of the short sequence lengths produced by next generation sequencing, a recent trend has been to instead sequence small fragments spanning one or more of the gene's variable regions, and still cluster them as OTUs at 97% id.

This work evaluated the effectiveness of utilizing short fragments for OTU generation at different id thresholds compared to the complete 16S rRNA gene. Whole gene analysis may be effective for measuring diversity; however, the variable region source of these small fragments may require higher or lower id thresholds. How precisely should the pieces of this 'genomic jigsaw' be characterized and distinguished? Two algorithms, UCLUST and CD-HIT-EST, were used to cluster complete 16S rRNA sequences, as well as fragments spanning the V1-3 and V3-5 regions due to their widespread use in human microbiome research. These sequences were obtained from SILVA's Living-Tree-Project (LTP) database. These clusters were produced at several id thresholds to evaluate how closely fragment clusters would resemble those obtained using complete genes. It was revealed that clustering small fragments, as well as fragment position, impacts OTU generation. However, results have suggested more appropriate id thresholds for these fragments to perhaps help us better assemble this microbial jigsaw puzzle. Clustering at 94% and 96% id for the V1-3 and V3-5 regions, respectively, generates similar results to whole gene clustering at 97%.

Acknowledgements

I want to thank professor Gabriel Moreno-Hagelsieb (aka SuperGabo) for his help, patience and support during my two years in his lab. His ongoing support and guidance not only played a critical role in my research, it also enhanced my learning every step of the way and for that I'm forever grateful. I came to this field with no experience, but I will always appreciate his willingness to give this newbie a chance. I would also like to thank my committee members for their invaluable feedback: to Dr. Allison McDonald, you helped make me feel less nervous about public speaking, and to Dr. Tristan Long, you R a great professor, and made stats one of my favourite classes ever. A special thank you to the members of the Computational ConSequences lab, who have offered their support and made lab meetings something to look forward to. Of special note: Scott Dobson-Mitchell, thanks for the Ducks and Starbucks; and to my mom, Kathy Dobson, thank you for reading the *whole* thing, and being the best editor ever.

Table of Contents

1. Introduction

1.1. Looking at a new environment: the human microbiome

A key part of genomic research includes characterizing microbial communities using sequencing technologies. These technologies give researchers the ability to sequence the genes of microbes present in different environments, providing a snapshot of the microbial diversity within these environments. Traditionally, microorganisms were isolated in pure culture and their genes sequenced to combine knowledge of physiology with underlying genetics (Sabree *et al*., 2009). While this provided new information for understanding these microbes, this also required studying them outside their natural environments. Microbes were removed from their natural surroundings, grown in isolation in artificial media at optimal conditions, essentially removing the context of the microbes (Sabree *et al*., 2009). This also limited study to cultivable microorganisms. Metagenomics involves the direct isolation and examination of DNA from an environment, rather than attempting to isolate a single organism in pure culture for further study (National Research Council, 2007). This allows researchers to study microbes at the population level, examining the genes and/or genomes present in an environment to aid in understanding the diversity of that environment. This also allows for the study of microorganisms that as of yet have not been cultivable (National Research Council, 2007).

Metagenomes can provide important information to numerous fields of research. For instance, microorganisms are responsible for many essential processes in soils, such as nutrient cycling, nitrogen fixation and suppressing disease in plant life (George *et al*.,

2010). Metagenomics provides a means of studying the interactions of microbes and ecology of many diverse environments; aquatic systems, biofuel and environmental remediation are all fields of study that can, and have benefitted from metagenomic research (George *et al*., 2010; Li *et al*., 2009; National Research Council, 2007). An environment of particular interest today is the human microbiome. It is estimated that there are 10 times more bacteria than cells in the human body (Willey *et al*., 2011). Certain metabolic traits exhibited by humans are products of the activity of these bacterial communities, rather than the result of human evolution (Turnbaugh *et al*., 2007). Together, these bacterial communities comprise a bacterial genome that might partially dictate human genetic, metabolic and physiological diversity (Turnbaugh *et al*., 2007).

Human microbial communities are associated with a variety of diseases, and relating the genotypes of microbial communities to the phenotypes expressed in human hosts is an emerging field with increasing importance in the management of human health (Kuczynski *et al*., 2012). For instance, Crohn's disease (CD) is an inflammatory bowel disease that causes swelling and irritation of the human intestinal tract. While the definitive cause of CD is unknown, it is thought that CD is caused by abnormal responses of the immune system (Rosenfeld and Bressler, 2010). However, recent studies have shown that infectious bacteria may cause CD. *Mycobacterium avium paratuberculosis* is frequently associated with the inflammatory response observed in CD (Rosenfeld and Bressler, 2010). However, due to the difficult nature of isolating and growing *M. avium paratuberculosis* in culture, further studies need to be conducted to fully understand the relationship of this bacterium and CD (Rosenfeld and Bressler, 2010). To understand an ecosystem of bacteria, a method of determining what bacteria are present is needed.

Sequencing technologies may provide the long-needed bridge to span this gap, providing a culture-independent means of determining how many different species, and what species, are present in an environment.

A recent advance in treating *Clostridium difficile* infections has also highlighted the importance of a cultivation-independent means of microbial community characterization. *C. difficile* infections are a frequent nosocomial illness (Rohlke and Stollman, 2012), causing diarrhea and often serious intestinal health problems, such as colitis (inflammation). Patients undergoing an antibiotic regime are at increased risk for infection, as antibiotics reduce normal intestinal flora and *C. difficile* can dominate in the intestine, resulting in illness. Recently, novel strains of antibiotic resistant *C. difficile* have been appearing, which has increased rates of infection in healthy individuals (Rohlke and Stollman, 2012). A new treatment method for *C. difficile* infections is fecal microbiota transplantation, which replenishes the natural flora of the protective colonic microbiome (Rohlke and Stollman, 2012). Essentially, samples of fecal bacterial are taken from a healthy individuals are transplanted to the recipient. Once reestablished via fecal transplantation, the protective colonic microbiome suppresses *C. difficile* alleviating symptoms and curing >90% of infected individuals (Rohlke and Stollman, 2012). Metagenomic analysis allows for the study of colonic microbiomes in healthy individuals, which can then be used instead of fecal transplantation.

1.2. The *species* conundrum

Sequencing technologies as a means of determining the diversity and identity of bacterial species on and within the body has major implications in human health management. By

sequencing the genes of these microbes, researchers could elucidate a link between bacterial genotype and human phenotype (Clayton *et al*., 2009; Kuczynski *et al*., 2012). However, there is no general consensus on what defines a species. The first working definition of 'species' was Ernst Mayr's Biological Species Concept (BSC), grouping organisms that can produce fertile offspring via mating (Cohan, 2002; de Queiroz, 2005). While popular, an obvious difficulty arises when considering bacteria; how should species be designated for asexual organisms?

Controversy arises when attempting to apply the term species to bacteria because one must decide what set of criteria to use to define a bacterial species. Changing these criteria can group organisms in vastly different ways, with the end result being an arbitrary collection of organisms based on cutoffs that can generate ambiguous boundaries (Doolittle and Zhaxybayeva, 2009). This controversy is not limited to prokaryotic species designation (the BSC is inapplicable to some vertebrates, invertebrates, fungi, etc.), and, as a result, numerous species concepts have been developed over time to try and 'solve' this species conundrum (Doolittle and Zhaxybayeva, 2009).

While some researchers believe that there is a single concept that, once found, will be applicable to all organisms, others believe that it must be accepted that different concepts will have to be used for different organisms (Mishler and Brandon, 1987). Theories on the genetic and/or ecological processes contributing to the rise of separate groups of similar organisms have been used to determine how individuals could/should be categorized, and as a result numerous species concepts have been suggested (Doolittle and Zhaxybayeva, 2009). For instance, the ecological concept (organisms adapted to the

same niche), phenetic concept (organisms are phenotypically similar, looking different from other), and phylogenetic concept (favors evolutionary relationships among organisms, examining factors such as a common ancestor) are just a few examples of alternative concepts (de Queiroz, 2005). Many bacteria today have been named for their human interest, for instance *Neisseria meningitidis* and *Mycobacterium tuberculosis* are named after diseases they cause (Gevers *et al*., 2005). While useful, this categorization logic collapses as other factors are considered; not all bacteria have a direct relationship with humans that can be used for categorization.

Part of defining 'species' in applicable terms to bacteria involves theorizing how speciation might occur in bacterial populations. How do new species arise over time? Perhaps evolutionary novelties provided through mutant alleles could provide the means of generating novel bacterial species; favored mutant alleles that use a niche's resources more effectively can sweep to fixation, driving diversity (Cohan and Perry, 2007; Doolittle and Papke, 2006; Doolittle, 2012). However, while this suggests focusing on lineages to assign species and generating trees of life, bacteria frequently acquire new genetic material through horizontal gene transfer, which suggests the notion of a web of life, rather than a tree of life. A single population of bacteria could be divided into two subpopulations through the acquisition of a novel plasmid, which could for instance confer pathogenicity (Doolittle and Papke, 2006). However, certain critical genes have been found to exhibit low levels of horizontal gene transfer, in theory making the majority of variations in these genes the result of neutral substitutions (Doolittle and Papke, 2006), which could make differences in the gene a more accurate measure of time and therefore be used to distinguish species (Janda and Abbot, 2007).

While advances in assaying phenotypes have aided in generating bacterial phenotypic clusters as a means of separating species, genomic approaches have gained an appeal for grouping bacteria (Cohan, 2002). Advances in molecular technology led to a new 'gold standard' in species identification, that overall genomic similarity could play a role in species assignment, with fine scale differentiation using phenotypic differences (Doolittle and Zhaxybayeva, 2009). This gold standard became the widely accepted phylogenetic species concept, that bacterial strains with approximately 70% or greater DNA–DNA hybridization would constitute a bacterial 'species,' since at this level there is a high degree of phenetic similarity (Rosselló-Mora, 2006; Stackebrandt *et al*., 2002; Stackebrandt and Goebel, 1994). However, due to cost and time constraints, hybridization techniques are not an effective means for identifying species in the field.

1.3. 16S rRNA and Operational Taxonomic Units (OTUs)

As an alternative to DNA–DNA hybridization, certain molecular marker genes can be examined. A potential molecular marker gene should be present in all bacteria being studied, and have a function that has not changed over time, suggesting that random changes in its sequence provide a more accurate measure of time and possibly evolution (Janda and Abbott, 2007, Ueda *et al*., 1999). Target amplicon studies are a DNA sequencing technique used to create multiple copies (amplicons) of a specific region of a gene (Bybee *et al*., 2011). From a sample of bacterial genomes, target amplicon studies can generate multiple amplicons of variable regions of a specific molecular marker gene. The degree of [evolutionary] change in these variable regions can be assessed to identify

and define the taxonomic relationships between the bacteria in a sample (Bybee *et al*., 2011).

The small ribosomal subunit (16S rRNA) gene is a commonly used molecular marker gene in many studies, including those of the human microbiome. Found in all prokaryotes, the transcription of this gene produces an rRNA that makes up part of the ribosome. The ribosome is made up of a large and a small subunit, which together translate mRNA into proteins; the16S rRNA gene produces the rRNA of the small subunit of the ribosome. In addition to its essential function, this gene is approximately 1500 base pairs (bp) long, making it faster and cheaper to sequence than it would cost to sequence the 23S rRNA gene (Mizrahi-Man *et al*., 2013). It also contains nine variable regions interspersed along the gene; these regions vary between bacteria and are used to separate bacteria into species (see Figure 1.1A for diagram of 16S rRNA with variable regions highlighted).

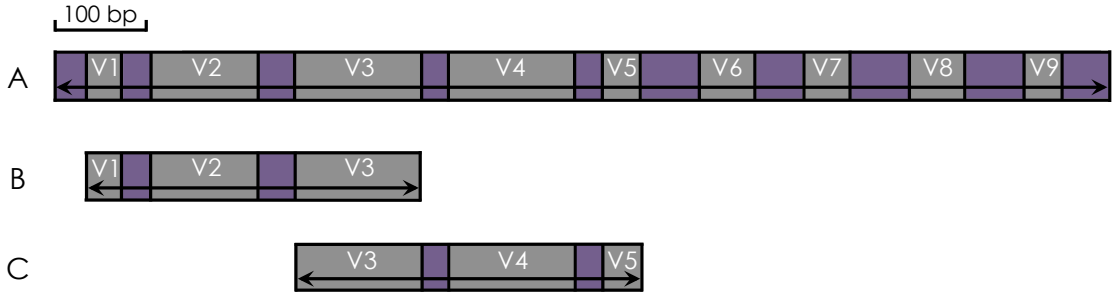


Figure 1.1: The 16S rRNA gene. Conserved regions of the 16S rRNA gene are violet and variable regions are numbered from V1 to V9. Illustration (A) shows the whole gene with all nine variable regions. Illustrations (B) and (C) represent fragments spanning the V1-3 and V3-5 regions respectively. These regions are sequenced in such studies as the Human Microbiome Project.

16S rRNA has a complex and highly conserved secondary structure that is preserved among bacteria, much more so than the primary sequence, due to their critical

biological functions (Gardner *et al*., 2005; Mizrahi-Man *et al*., 2013; Smit, *et al*., 2007). Functionality of this gene is highly dependent on the secondary structure, as this dictates how 16S rRNA interacts with ribosomal proteins (Kitahara *et al*., 2012). Figure 1.2 illustrates the general stem and loop appearance of the 16S rRNA secondary structure; variable regions are found in both stems and loops (helices), however it has been noted that changes in the primary sequence of these variable regions do not greatly change the overall structure of 16S rRNA (Van de Peer *et al*., 1996; Marchandin *et al*., 2003).

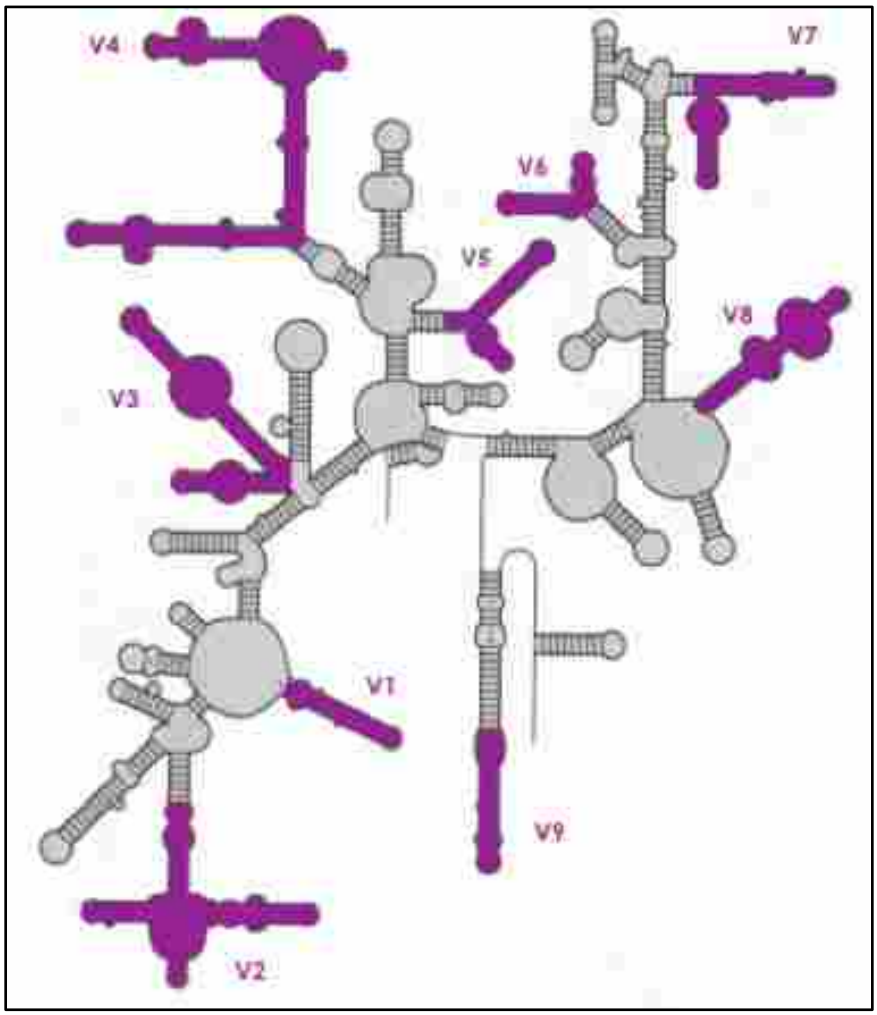


Figure 1.2 illustrates the general secondary structure of 16S rRNA, with variable regions highlighted in purple (adapted from Van de Peer, 1996; Tortoli, 2003).

Previous studies have shown that a 70% or greater DNA–DNA hybridization correlates with 16S rRNA gene sequence similarity (id) equal to or greater than 97% (Doolittle and Zhaxybayeva, 2009; Gevers *et al*., 2005; Janda and Abbott, 2007). This correlation is used to generate Operational Taxonomic Units (OTUs), 16S rRNA sequences grouped by a defined level of similarity, with each OTU being used as a proxy for a single 'species.' Using the term OTU obviates the controversy of using the "species" term.

For identification purposes, 16S rRNA sequences with 97% id or greater are assigned to the same OTU (representing a 'species' as defined by 97% sequence identity), while sequences with >95% and >85% sequence identity are assigned to the same genus and phylum respectively (Janda and Abbott, 2007; Schloss and Handelsman, 2005; Stackebrandt and Goebel, 1994).

1.4. Bacterial identification using 16S rRNA V regions

High-throughput sequencing technologies describe a range of technologies that can sequence DNA orders of magnitude faster and more cheaply than older technologies (Rodríguez-Ezpeleta *et al*., 2011). With the advent of high-throughput, or next generation sequencing technologies (NGS), short DNA fragments can be quickly sequenced and compared to reference sequences from databases to identify the bacterium associated with the fragment (Janda and Abbott, 2007). However, these new technologies sequence shorter fragments of the gene; the Illumina MiSeq sequencer can generate a maximum read length of 250bp (very recently 300bp), 454 sequencing technologies can produce fragments averaging 400-450bp long (recent models have attempted fragments up to

800bp), while older technologies such as the Sanger Method can generate much larger sequences, upwards of 1000bp long (Gevers *et al*., 2012). Studies of various bacterial communities use these new technologies for target amplicon studies of the 16S rRNA gene (Chakravorty *et al*., 2007). While earlier studies sequenced the whole 16S rRNA gene, there has been a shift towards using the short sequence reads generated from new high-throughput sequencing technologies (Mizrahi-Man *et al*., 2013). There are many advantages to this approach, as the new technologies are faster and cheaper. However, at present there is a lack of consensus over the most effective variable (V) region of the 16S rRNA gene to sequence, with many studies opting to examine more than one region as no single region has been shown to optimally differentiate among bacteria (Chakravorty *et al*., 2007; Mizrahi-Man *et al*., 2013). Bacterial species could show diverse levels of variation in the nine V regions of the gene, and an important step in 16S rRNA gene sequencing includes deciding what region(s) to sequence, as classification bias (depending on the V region used) has been previously observed (Li *et al*., 2009; Vilo and Dong, 2012)

Despite the aforementioned limitations, certain V regions of the 16S rRNA gene have been found to be useful in identifying a wide range of bacteria, and studies tend to examine more than one region in an attempt to increase the detection of sequence diversity (Chakravorty *et al*., 2007). Amplicons spanning V regions 1, 2 and 3 (denoted V1-3) and regions 3, 4 and 5 (denoted V3-5), are among the most commonly used in studies using this gene (see Figure 1.1B and 1.1C for diagram of 16S rRNA gene fragments spanning V1-3 and V3-5 respectively), particularly in studies of the human microbiome (Gevers *et al*., 2012; Huse *et al*., 2012). The V1-3 region is approximately

428bp long and the V3-5 region is approximately 446bp long, putting these two regions within the limits of NGS technologies.

1.5. Research objectives and significance

This project seeks to evaluate how effective 16S rRNA fragments are in representing diversity in comparison to whole 16S rRNA sequences, and to determine whether the variable region source of these small fragments may require higher or lower sequence similarity (id) thresholds for accurate (comparable to whole-gene sequence) measurements. Figure 1.3 provides an objectives flowchart.

Whole 16S rRNA gene analysis clustered at 97% id have been used for measuring diversity (Stackebrandt and Goebel, 1994). However, sequence fragments spanning one or more variable regions, such as V1-3 and V3-5, may require different thresholds to produce OTUs comparable to those produced with whole gene sequences, and new values may need to be suggested. In doing so, suggested thresholds should produce OTU clusters closer to what whole-gene clusters would provide. This would also allow for studies examining the same bacterial communities with different variable regions to have comparable results. Chapter 2 gives an overview of how whole-gene sequences versus gene-fragment sequences impact OTU generation. Chapter 3 will focus on suggesting new id thresholds for gene-fragment sequences (spanning V1-3 and V3-5).

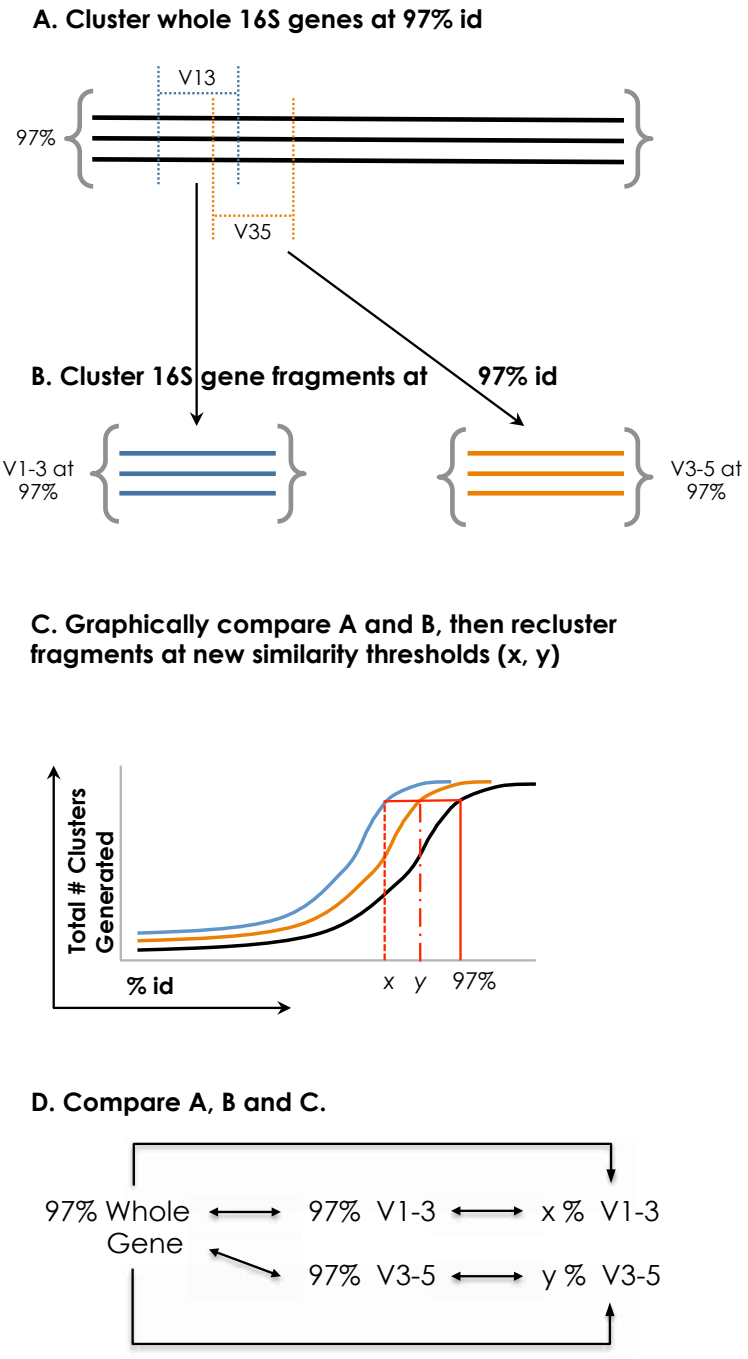


Figure 1.3: Flowchart summarizing how objectives will be achieved; **A.** depicts initial whole-gene clustering at 97% id; **B.** depicts clustering gene-fragments spanning variable regions one through 3 (V1-3) and three through five (V3-5); **C.** depicts graphically determining alternative id thresholds (x and y) for the V1-3 and V3-5 regions respectively through comparison to whole-gene clustering at 97% id; **D.** depicts comparing the results of **A.**, **B.** and **C.** to understand how gene-fragment region and id threshold affects OTU generation and composition.

1.6. Integrative nature: biology meets technology

Computational biology integrates biology and computer science (computational analyses) to study biological data sets. These data sets include the vast number of DNA sequences generated from high-throughput sequencing technologies. Computational methods are used to examine data generated by molecular analyses and identify patterns.

For instance, computational methods aid in ascertaining the taxonomic relationships between large sequencing datasets of bacterial genomes from sample sites of the human microbiome. While computational methods require molecular data to analyze, molecular methods, such as genome sequencing, can in turn be influenced by the findings of computational analyses. For example, determining whether short sequence fragments are of an adequate length to identify bacterial species could potentially alter how molecular methods are used to examine sample sites. This illustrates how computational and molecular methods are complementary.

While bioinformatics can be considered multidisciplinary, the wide variety of settings in which it can be used, such as studying human biological systems, also demonstrates its integrative nature. High-throughput sequencing technologies can be used to examine the microbial ecology of soil and marine environments, and computational methods are being used to examine the relationships among and between the genomic data generated by these studies (Gilbert and Dupont, 2011). Computational studies can examine multiple bacterial species and/or genera across a diverse range of environments.

2. Genome Jigsaw: fragment evaluation

Current methodology for characterizing bacterial communities involves sequencing short fragments of the ubiquitous 16S rRNA gene, and comparing these fragments to reference sequences in databases to identify the bacteria present (Bybee *et al*., 2011; Janda and Abbott, 2007). However, while whole 16S rRNA gene analysis may be effective for measuring diversity, a leap was made to use the same clustering 97% id threshold to group whole 16S rRNA gene sequences and gene fragments. This chapter will focus on an evaluation of how well 16S rRNA fragments represent diversity in comparison to whole 16S rRNA sequences when clustered at the same identity threshold. Since whole-gene 16S rRNA sequences clustered at 97% id has historically been considered the 'standard,' then gene-fragments spanning the widely used V1-3 and V3-5 regions should produce similar results. The null hypothesis is that the gene-fragments will produce clustering results that do not deviate significantly from results produced from whole-gene sequences.

2.1. Materials and methods

The methodology for this study is outlined in Figure 2.1. Bacterial sequences to be clustered were taken from the SILVA Living-Tree Project comprehensive ribosomal database. Whole gene sequences were used, as well as gene-fragments created by trimming whole gene sequences to span the V1-3 and V3-5 regions. The V1-3 region was identified as the region between 69 – 497 nucleotides in the aligned 16S rRNA sequences, while the V3-5 as the region between 433 – 879 nucleotides as suggested by

Vilo and Dong, 2012 and Chakravorty *et al*., 2007. For each gene region (whole-gene, V1-3 and V3-5), a PYTHON program was written to cluster sequences into OTUs using clustering algorithms UCLUST and CD-HIT-EST at 97% id. These clustering results were analyzed (with a PYTHON program) by determining the number of 'broken *species*' and 'contaminated clusters' generated by each gene region. Results were analyzed and visualized using the program R to illustrate the differences in OTU generation (i.e. the number of clusters produced) by each gene region (and by clustering algorithm). I will now elaborate on each step of the methodology.

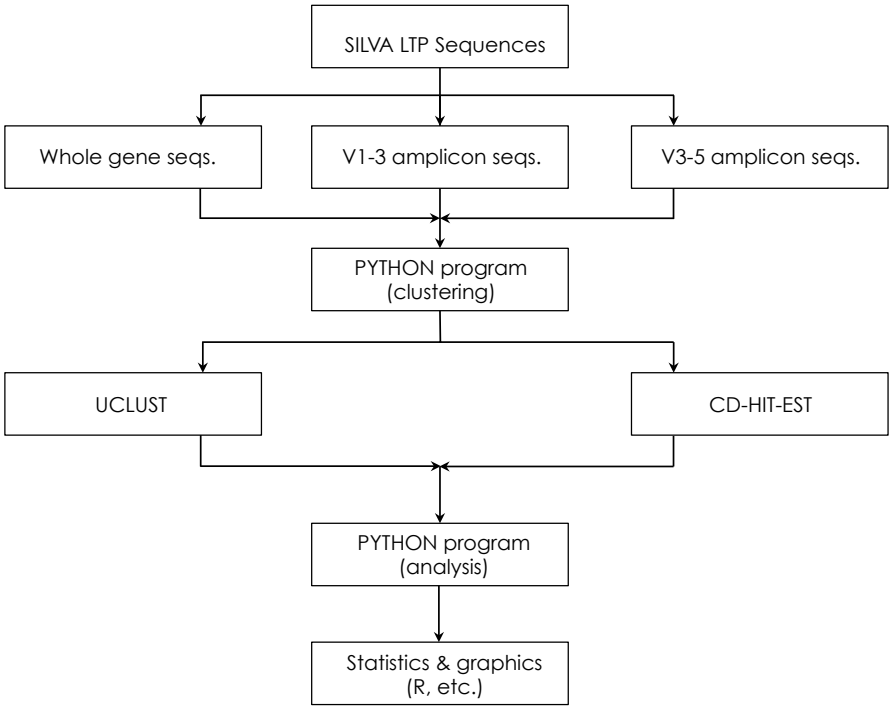


Figure 2.1: Flowchart of methodology followed for chapter 2. Full-length (or near full length) bacterial 16S rRNA gene sequences were taken from the SILVA LTP database. These sequences were analyzed for three gene regions: spanning the whole-gene, and then trimmed to span the V1-3 and V3-5 regions. A PYTHON program was used to cluster each gene region using clustering algorithms UCLUST and CD-HIT-EST (with % id specified). A second PYTHON program analyzed numerous aspects of the OTUs generated by each gene region. Finally, statistics were produced using the statistical computing program R.

2.1.1. Sequence data: SILVA 16S rRNA database

In this study, the SILVA 'All-Species Living Tree' (LTP) project database of 16S rRNA datasets was used for clustering and OTU generation. Specifically, the LTP_release_108 was used. This release was available at <http://www.arb-silva.de/no_cache/download/archive/living_tree/>. The LTP was created with the goal of creating a reference dataset of 16S rRNA sequences spanning all sequenced type strains of classified bacteria (Quast *et al*., 2013). Therefore, the LTP only contains cultivated sequences of bacteria, and can provide a comparison between sequence clustering at a user designated identity level (i.e. 97% id) and the actual 'known' species of the sequence. SILVA uses the term *species* to refer to bacteria named according to the Bacteriological Code and appearing in the validation and/or notification lists of the International Journal of Systematic and Evolutionary Microbiology (IJSEM). For the remainder of this document, *species* will be italicized and will refer to how SILVA utilizes the term. The LTP contains approximately 9700 16S rRNA sequences, all at least 1200 bp long. This length ensures that all sequences in the database cover the V1-3 region and V3-5 region, and can be trimmed to cover these regions for further clustering.

2.1.2. Sequence clustering programs

PYTHON is a programming language that is known for being easy to learn and use, as well as for being multi-platform; it can be used with many operating systems (Bassi, 2007). PYTHON was used to write programs that utilize two clustering algorithms to sort bacterial sequences, in additional to other programs that helped perform OTU analysis. Many clustering programs exist which allow users to group gene sequences at user-

defined sequence similarity (id) thresholds to generate OTUs. Two prominent and widely used program suites are CD-HIT and USEARCH, which provide sequence-clustering algorithms, CD-HIT-EST and UCLUST. These two algorithms were used to cluster whole-gene and gene-fragment sequences at 97% id.

2.1.3. CD-HIT-EST

Originally a protein clustering program (CD-HIT), CD-HIT-EST is a variant of the original CD-HIT algorithm and uses a greedy incremental algorithm to cluster RNA and DNA sequences (Li *et al*., 2001). Sequences are first ordered by decreasing length, with the longest sequence becoming the representative sequence of the first cluster. Each subsequent sequence is then compared to this representative using a short word filtering system. Short word filtering reduces the number of pairwise alignments that must be made between two sequences, which greatly speeds up the process of sequence comparison (Li *et al*., 2006). At varying levels of sequence identity, both sequences should have at least a certain amount of dinucleotides, trinucleotides, etc., in common. CD-HIT-EST uses these common nucleotides, 'short words,' to compare sequences. A sequence will become a new cluster representative if it does not have enough words in common with a previous cluster representative. Short word size varies depending on the level of sequence identity at which a user is clustering. CD-HIT-EST calculates sequence identity as the number of identical nucleotides in the alignment divided by the length of the shorter sequence. Sequence gaps are not counted as differences (Li *et al*., 2006).

A drawback of greedy incremental sorting is that sequences are sorted into the first cluster that meets the set requirements, not necessarily the best match (a cluster with a higher percent sequence similarity). However, a greedy clustering algorithm has been

shown to provide good results with vastly improved clustering speeds (Li *et al*., 2012). Additionally, CD-HIT provides a recommended procedure for iterated runs to reduce errors that can be caused by one-step greedy clustering. A basic CD-HIT-EST command would be:

```
cd-hit-est –i filein –o fileout –c 0.97 –n 8
```

Where: `–i` is used to indicate the input file name, `–o` to indicate the output file name, `–c` to indicate the sequence identity threshold (here at 97%) and `–n` is used to indicate the word size.

2.1.4. UCLUST

UCLUST comes with two variants, `cluster_fast` and `cluster_smallmem`. The former works in a similar fashion to CD-HIT-EST, utilizing a greedy algorithm to sort sequences by length and a short word filtering system to cluster them (Edgar, 2010). The `cluster_smallmem` variant does not sort the sequences; sequences are clustered based on the order of the input file. UCLUST has the option of addressing the possible drawback of greedy algorithms by providing the option of constructing consensus sequences for each cluster using the `–consout` command, which allows the user to use the dominant/most abundant sequence as the cluster representative (Edgar, 2010). This would aid in ensuring sequences with the greatest sequence similarity end up in the OTU (not just one above the given threshold). UCLUST defines sequence identity as the number of identities (an alignment column with identical nucleotides) divided by the number of columns (Edgar, 2010). Internal sequence gaps are included in column counts, which differs from the identity definition provided by CD-HIT-EST. A basic UCLUST command would be:

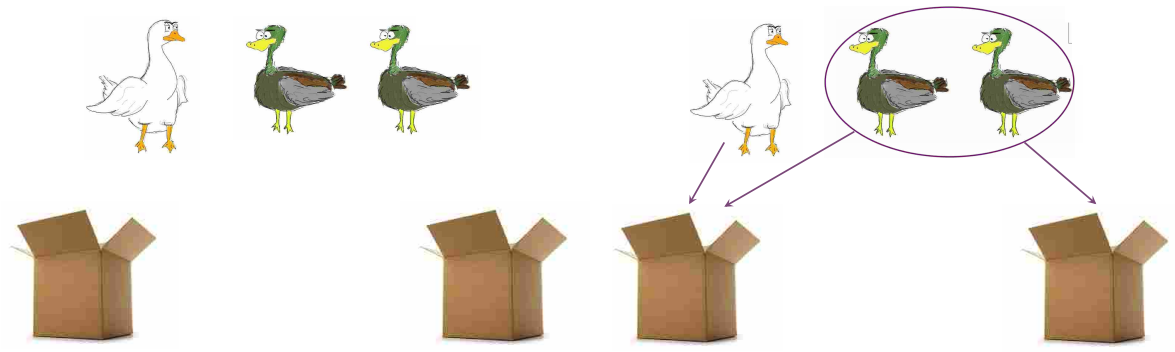
```
usearch7 –cluster_smallmem filein.txt –id 0.97 –uc
clusterfile.uc –centroids centroidfile.fasta
```

Where: `–id` is used to indicate the sequence identity threshold (here at 97%), `–uc` requests the production of a cluster output file, and `–centroids` produces a representatives fasta file.

2.1.5. Cluster analysis: 'broken *species*' and 'contaminated clusters'

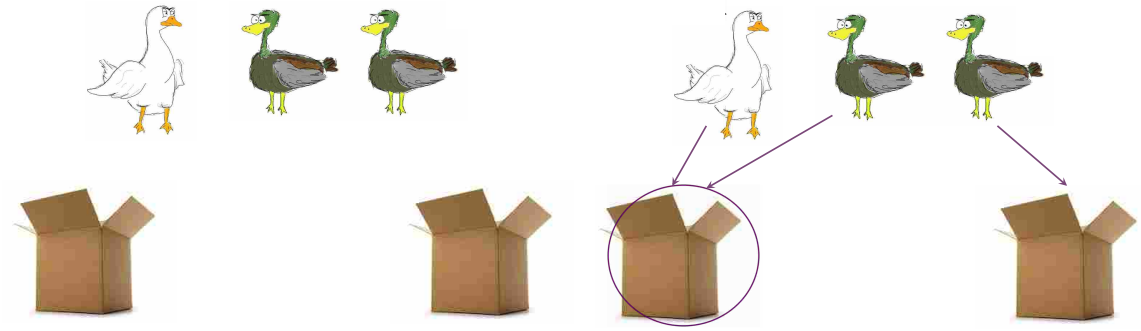
PYTHON was used to write a script to examine the differences between OTUs generated using whole 16S rRNA genes and those using gene-fragments spanning V1-3 and V3-5. Variables that were examined included: how many total OTUs were generated, how these OTUs differ depending on the gene region used for clustering, and the number of 'broken *species*' and 'contaminated clusters' produced.

Figure 2.2 illustrates the broken *species* concept; broken *species* counts show how many bacterial *species* (defined and provided by the SILVA LTP database) have been placed into two or more different OTUs – therefore 'breaking' the *species* into different groups. Figure 2.3 illustrates the contaminated clusters concept; counts of contaminated clusters show how many OTUs contain two or more different *species*. OTUs generated at 97% id should in theory contain only one species; therefore, this value indicates the effectiveness of the gene region (either whole, V1-3 or V3-5) at accurately separating different *species* into different OTUs.



(A) Before OTU generation (B) After OTU generation

Figure 2.2: An analogy to explain the broken *species* concept; OTUs can be considered 'boxes,' with each housing one species. Before clustering (2.2A), it's expected that the goose would go into one box (OTU), and the two ducks would go into the other, since they are from the same species. However, a *species* can be 'broken' (2.2B) if, after clustering, not all the organisms from the same *species* end up in the same box.



(A) Before OTU generation (B) After OTU generation

Figure 2.3: A continuation of the goose and duck box analogy, a box (OTU) should contain only one *species* (2.3A). However, a box can be 'contaminated' by containing more than one *species* (2.3B).

2.1.6. Statistics and graphics

Statistics were produced using the statistical computing program R. Chi-square tests were used to detect statistically significant ($p < 0.05$) deviances in broken *species* and contaminated cluster counts between whole-gene clustering and gene-fragment clustering (Fowler *et al.*, 1998). Graphs were produced using R's ggplot2 libraries (R Core Team, 2014; Wickham, 2009).

2.2. Results and discussion

Results for each clustering algorithm will be presented separately, starting with UCLUST and then CD-HIT-EST. This will be followed by an evaluation of using gene fragments when clustering 16S rRNA gene sequences.

2.2.1. UCLUST fragment evaluation

Figure 2.4 provides an overview of the number of OTUs generated by whole-gene and gene-fragment (V1-3 and V3-5) clustering at 97% id. Clustering whole 16S rRNA sequences at this id threshold generated 5427 OTUs in total. Of these, 3767 (~69%) were 'singleton' OTUs, containing a single sequence, with the remaining 1660 OTUs (~31%) containing two or more sequences.

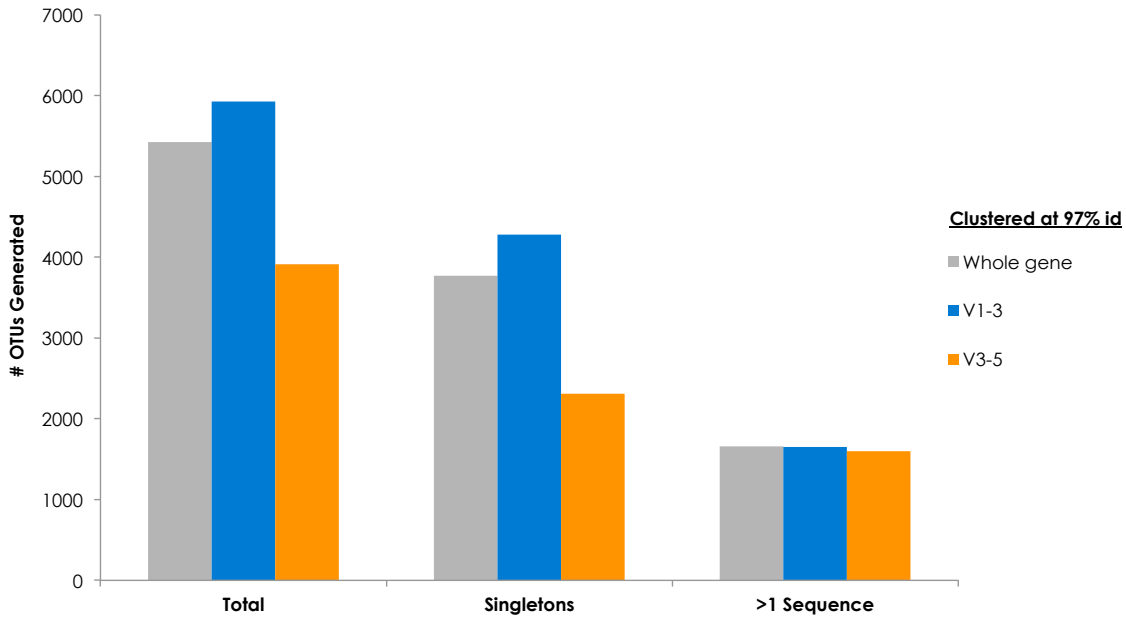
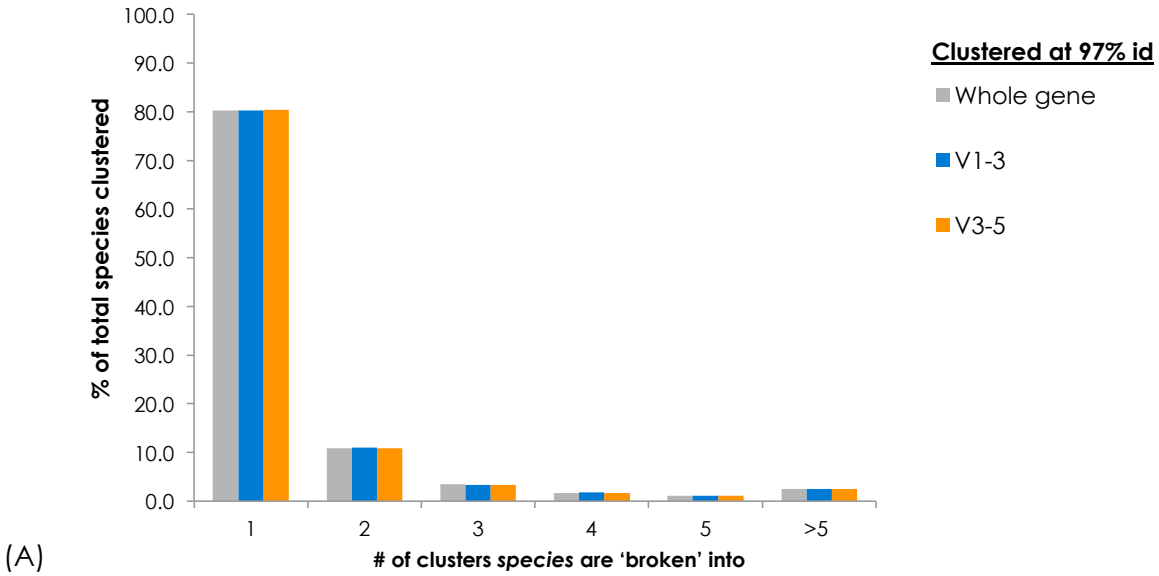


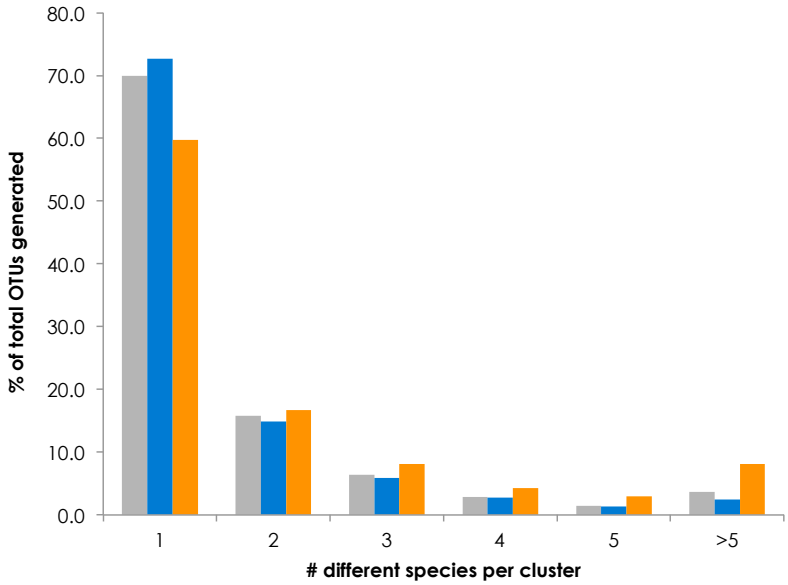
Figure 2.4: Illustration of how whole-gene sequences and gene-fragment sequences spanning V1-3 and V3-5 were clustered into OTUs by UCLUST at 97% id. 'Total' represents a count of the total OTUs (clusters) produced, 'Singletons' represents a count of OTUs contained a single sequence, and '>1 Sequence' represents a count of OTUs containing more than 1 sequence.

At 97% id, the V1-3 region generated 5929 OTUs in total, and the V3-5 region at 97% id generated 3911 OTUs in total (refer to Table A.1 in Appendix for Singleton and >1 Sequence values). To further understand and compare these differences in OTU generation, the number of broken *species* and contaminated clusters were calculated. As Figure 2.5A shows, whole gene and gene fragment sequences (spanning V1-3 and V3-5) generated similar counts of broken species; approximately 80% of all *species* remained whole and 'unbroken' in a single cluster. (See Tables A.2 and A.3 for broken *species* counts and percentages respectively). There was no significant deviance in broken *species* counts between whole-gene sequences and the V1-3 region, and between whole-gene sequences and the V3-5 region (V1-3: $X^2 = 0.563$, df = 14, $p > 0.05$, V3-5: $X^2 = 1.2381$, df = 14, $p > 0.05$).

As Figure 2.5B illustrates, an examination of the number of contaminated clusters produced by whole-gene clustering found that approximately 70% of OTUs contained a single *species*. Gene fragment sequences spanning V1-3 created slightly fewer contaminated clusters, with approximately 73% of OTUs containing a single *species*. The gene fragment spanning V3-5 produced the greatest contamination; over 54% of OTUs generated by this region were contaminated. (See Tables A.4 and A.5 for contaminated OTU counts and percentages respectively). A chi-square test found a significant difference in the number of contaminated clusters produced by the gene fragment sequences spanning the V1-3 region in comparison to the expected amounts produced by the whole gene sequences ($X^2 = 27.0241$, df = 9, $p = 0.001386$). This was also found for the V3-5 region ($X^2 = 177.887$, df = 9, $p < 2.2e\text{-}16$).



(A)



(B)

Figure 2.5: Using the different gene regions, all *species* were clustered (6246 in total) at 97% id by UCLUST; (2.5A) shows the percentage of *species* that were clustered into a single OTU. The first set of bars represents the percentage of total *species* that were not broken into different OTUs (~80% for all three gene regions), the second set of bars represents the percentage broken into two clusters, and so on. (2.5B) shows what percentage of the total OTUs generated by each gene region that are contaminated (containing more than one *species*). The first set of bars represent OTUs that are not contaminated, containing only one *species*. The second set of bars represents the percent of OTUs generated that contained two different *species*, and so on.

2.2.2. CD-HIT-EST fragment evaluation

The V1-3 region generated the most OTUs, followed by whole-gene sequences and then the V3-5 region. Clustering the V1-3 region (Figure 2.6) produced 7519 OTUs (6391 singletons, 1128 OTUs with >1 sequence). Clustering whole 16S rRNA sequences at 97% id generated 4978 OTUs in total (Figure 2.6). Clustering the V3-5 region (Figure 2.6) at 97% id produced 5932 total OTUs (refer to Table A.6 for Singleton and >1 Sequence counts).

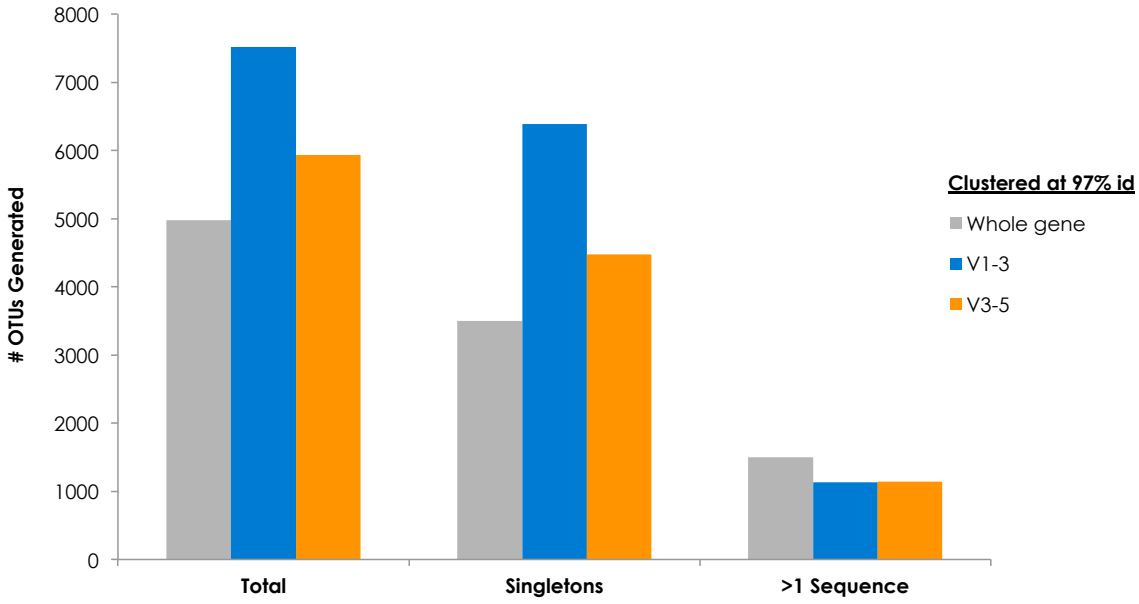
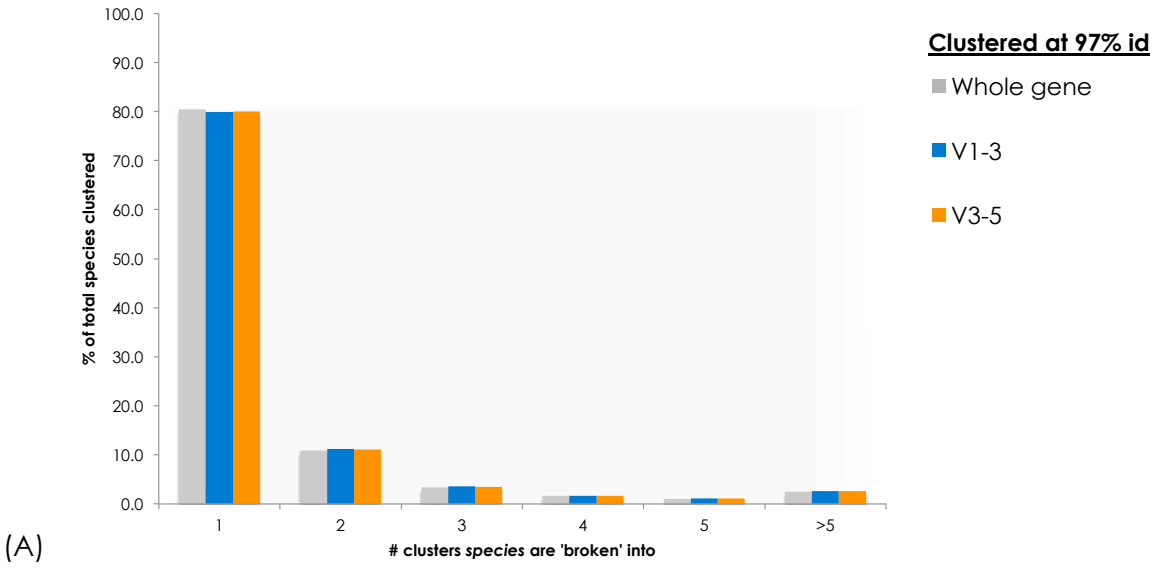


Figure 2.6: Illustration of how whole-gene sequences and gene-fragment sequences spanning V1-3 and V3-5 were clustered into OTUs by CD-HIT-EST at 97% id. 'Total' represents a count of the total OTUs (clusters) produced, 'Singletons' represents a count of OTUs contained a single sequence, and '>1 Sequence' represents a count of OTUs containing more than 1 sequence.

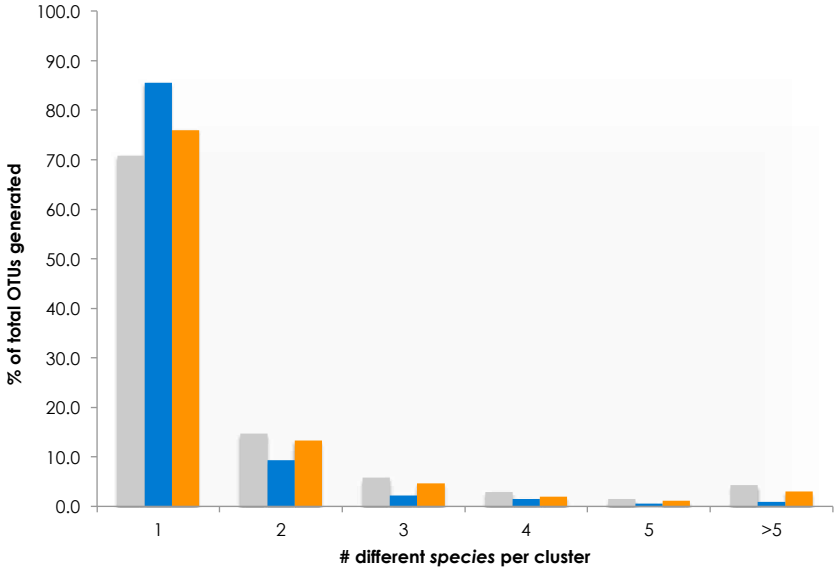
Again, to better understand and compare these differences in OTU generation, the number of broken *species* and contaminated clusters were determined. As Figure 2.7A shows, whole gene and gene-fragment sequences spanning V1-3 and V3-5 generated

similar amounts of broken species; approximately 80% of all species remained whole and 'unbroken' in a single cluster (refer to Tables A.7 and A.8 for broken *species* count and percentages). These results are also similar to those produced by UCLUST. There was no significant difference in broken *species* between whole gene sequences and the V1-3 and V3-5 regions (V1-3: $X^2 = 1.0634$, df = 14, $p > 0.05$, V3-5: $X^2 = 1.1276$, df = 14, $p > 0.05$).

For the number of contaminated clusters, it was found that approximately 71% of OTUs (3528 out of 4978) generated by clustering whole-gene sequences contained a single *species*. As Figure 2.7B illustrates, the gene fragment sequences spanning V1-3 created fewer contaminated clusters, with approximately 85% of OTUs generated by this gene region containing a single *species* (6428 OTUs out of 7519). Refer to Tables A.9 and A.10 for more contaminated clusters values and percentages. The gene fragment spanning V3-5 produced the greatest contamination, with approximately 53% of OTUs (4507 out of 5932) containing a single *species*. A chi square test found that there was a significant difference in the number of contaminated clusters produced by the gene fragment sequences spanning the V1-3 region in comparison to the expected amounts produced by the whole gene sequences ($X^2 = 470.2228$, df = 9, $p < 2.2e-16$). This was also found for the V3-5 region ($X^2 = 49.7968$, df = 9, $p = 1.176e-07$).



(A)



(B)

Figure 2.7: Using the different gene regions, all *species* were clustered (6246 in total) at 97% id by CD-HIT-EST; (2.7A) shows the percentage of *species* that were either clustered into a single OTU, 'broken' into two OTUs, and so on. The first set of bars represents the percentage of total *species* that were not broken into different OTUs (~80% for all three gene regions), the second set of bars represents the percentage broken into two OTUs, and so on. (2.7B) shows what percentage of the total OTUs generated by each gene region that are contaminated (containing more than one *species*). The first set of bars represent OTUs that are not contaminated, containing only one *species*. The second set of bars represents the percent of OTUs generated that contained two different *species*, and so on.

2.2.3. Fragment evaluation: conclusion

Clustering small fragments, as well as fragment position (i.e. fragments spanning V1-3 or V3-5), impacts OTU generation. Except for the V3-5 region clustered by UCLUST, all clustering of gene-fragments resulted in more OTUs being generated than whole-gene clusters (see Figures 2.5A and 2.7A). As reported in section 2.2.1 and 2.2.2, there was no significant difference in how many broken species were produced by the gene-fragments versus whole-gene sequences; all gene regions had a similar pattern, with approximately 80% of species remaining unbroken, approximately 10-11% being broken into two clusters, and the remaining 10% being broken into three or more clusters. However, significant difference was seen in the number of contaminated clusters produced for all three gene regions when using either UCLUST or CD-HIT-EST. This rejects the null hypothesis, and an examination of the contaminated clusters produced by the gene-fragments suggests the impact of this difference. Following the same pattern seen with the total OTUs produced (Figure 2.6), only the V3-5 region, when clustered by UCLUST, had a greater percent of generated OTUs be contaminated by two or more species than whole-gene clustering (see Tables A.6 and A.10 for contaminated cluster percentages).

Interestingly, the V1-3 region showed the lowest incidence of contaminated clusters; when clustered with UCLUST, approximately 73% were not contaminated (whole-gene using UCLUST was ~70%), and when clustered with CD-HIT-EST, 85% were not contaminated (whole-gene using CD-HIT-EST was ~71%). Based on these results, it would appear that these fragments are not only as effective as the whole-gene region in measuring diversity; they may be more effective. As previously mentioned, shorter gene-fragments are faster and cheaper to sequence; these findings support the use

of gene-fragments in 16S rRNA clustering, which has important implications in genomic research.

However, as whole-gene clustering at 97% id has traditionally been used and found to be effective in measuring diversity in microbial environments, finding id thresholds for these gene-fragments that generate OTUs with similar results to whole-gene clustering would be valuable. Chapter 3 covers a determination and examination of alternative id thresholds for these regions. If found to be equivalent in their OTU production, this would mean that different studies utilizing different gene-fragments when analyzing the same environment could be compared if using the id threshold appropriate for said gene-fragment despite the difference in gene region use.

3. Genome Jigsaw: id threshold evaluation

As mentioned previously, it has been assumed that short fragments of the 16S rRNA gene can be clustered at 97% id to produce similar results to whole-gene sequences. As discussed in chapter 2, short gene regions might require a new threshold to generate results that are more representative of those obtained from whole-gene clustering. This chapter will focus on an evaluation of alternative id thresholds so that 16S rRNA fragments could be more representative of whole 16S rRNA sequence clustering. The null hypothesis is that the gene-fragments will produce clustering results at their new respective id thresholds that do not deviate significantly from results produced form whole-gene sequences at 97% id. I will now elaborate on the methodology.

3.1. Materials and methods: id evaluation

The methodology for this chapter is outlined in Figure 3.1. The same sequence database as outlined in section 2.1.1 was used, as were the same clustering algorithms described in section 2.1.2. However, instead of just clustering at 97% id, all three gene regions were clustered over a range of id thresholds starting at 85% id to form a baseline, and then spanning 90-99%, at 1% increments. The total number of OTUs generated at each of these % id increments were then plotted to graphically propose new id thresholds for these two variable regions that could be equivalent to whole-gene clustering at 97% id. A graph of this concept can be seen in Figure 1.3C. Once new id thresholds were determined for both gene-fragments (for both clustering algorithms), the same steps as outlined in sections 2.1.3 and 2.1.4 were followed.

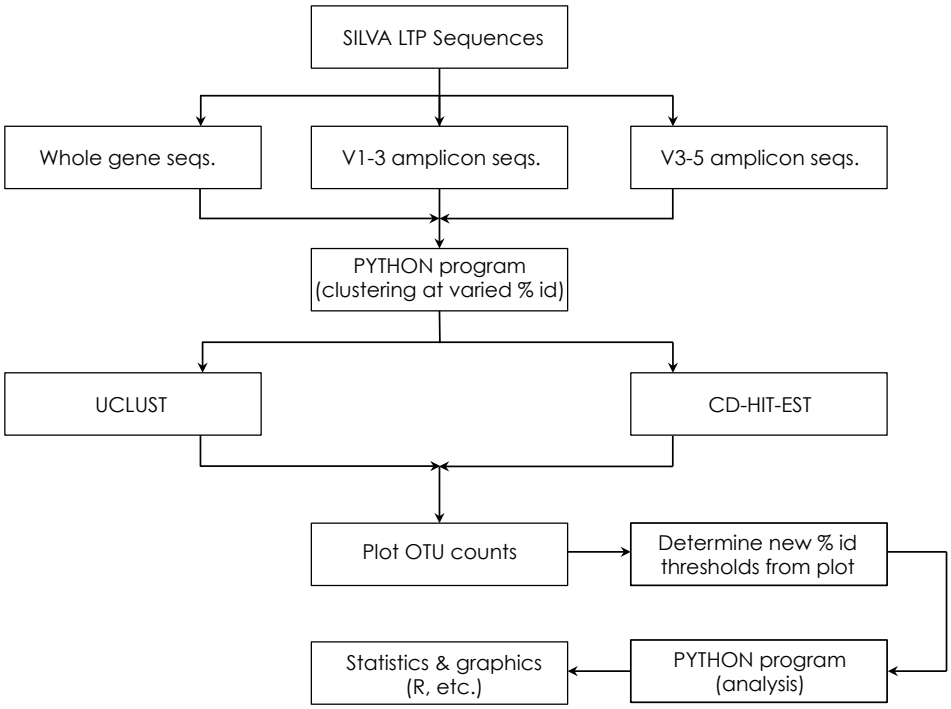


Figure 3.1: Flowchart of methodology followed for chapter 3. The same sequences and gene-fragments used in chapter 2 were used again, and clustered at varied % id thresholds. The total OTUs generated from each % id was plotted, and new id thresholds were determined from this plot. Numerous aspects of the OTUs generated by these new id thresholds were examined for each gene-fragment; total OTUs produced, broken *species* counts, contaminated cluster counts, etc. Finally, statistics and graphs were produced using R.

3.2. Results and discussion

Results for each clustering algorithm will be presented separately starting with UCLUST and then CD-HIT-EST. This will be followed by an evaluation of using the proposed alternative id thresholds when clustering 16S rRNA gene sequences.

3.2.1. UCLUST id threshold: evaluation

Figure 3.2 is a plot of the total OTUs produced by whole-gene and gene-fragments (spanning V1-3 and V3-5) clustered across a range of id thresholds. From this plot, it

appears that clustering gene-fragments spanning the V1-3 region at 96.5%, and gene-fragments spanning the V3-5 region at 98.5% produce similar amounts of OTUs; 5362 and 5438 total OTUs for the V1-3 and V3-5 regions respectively *versus* 5427 OTUs for whole-genes at 97% id (values listed in Table A.1).

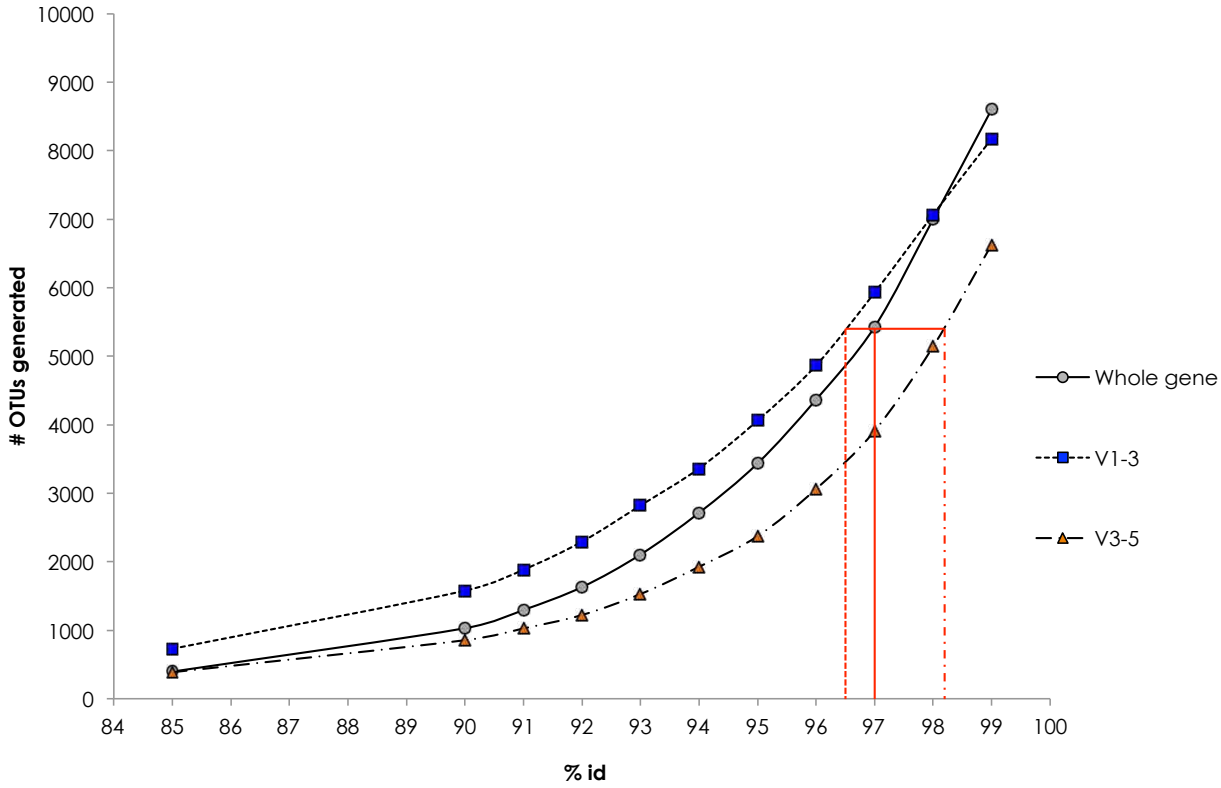


Figure 3.2: Plot of total OTUs generated by UCLUST using whole-gene sequences, V1-3 sequences and V3-5 sequences across a range of % id values. Solid red lines have been used to highlight where the V1-3 and V3-5 curves line up with the whole-gene curve at 97% id. Dashed red lines (matching the dashed lines used for each V region curve), show what % id on the x-axis appears to produce results similar to whole-gene clustering at 97% id.

An analysis of the broken *species* was conducted comparing whole-gene sequences clustered at 97% id versus the gene-fragments clustered at their new respective thresholds (V1-3 at 96.5%, V3-5 at 98.5%). As Figure 3.3A shows, results were similar to those found in section 2.2.1, with all three gene regions showing a similar trend in how

species were broken; approximately 80% of species were not broken; 11% were broken into two different OTUs; and the remaining 10% of *species* were broken into three or more different OTUs (values listed in Table A.2, percentages in Table A.3). There was no significant deviance in broken *species* between whole gene sequences and the V1-3 and V3-5 regions (V1-3: $X^2 = 0.563$, df = 14, $p > 0.05$; V3-5: $X^2 = 1.2381$, df = 14, $p > 0.05$).

Figure 3.3B illustrates how the gene-fragments, at their new respective id thresholds, contaminated clusters in comparison to whole-gene sequences clustered at 97% id. Interestingly, in contrast to the results in section 2.2.1 (UCLUST fragment evaluation), the V3-5 region produced the fewest contaminated clusters, followed by the V1-3 region (refer to Table A.9 for contaminated cluster counts). A chi square test found that there was no significant deviance in the number of contaminated clusters produced by the gene fragment sequences spanning the V1-3 region in comparison to the expected amounts produced by the whole gene sequences ($X^2 = 10.1254$, df = 9, $p = 0.3404$). This was also found for the V3-5 region ($X^2 = 5.4364$, df = 9, $p = 0.7947$).

Additionally, an examination of the differences between the OTUs generated by these variable regions at new id thresholds and whole-gene sequences at 97% id was carried out. The V1-3 region generated 5929 OTUs when clustered at 97% id by UCLUST. Approximately 60% of these OTUs (3566) were found to be identical (i.e. containing the same collection of sequences within each OTU) to OTUs generated by whole-gene sequences, and approximately 40% (2363 OTUs) were unique. When the V1-3 region was clustered at 96.5% id, 5362 OTUs were generated, and approximately 62% of these were identical to those produced by whole-gene sequences clustered at 97%.

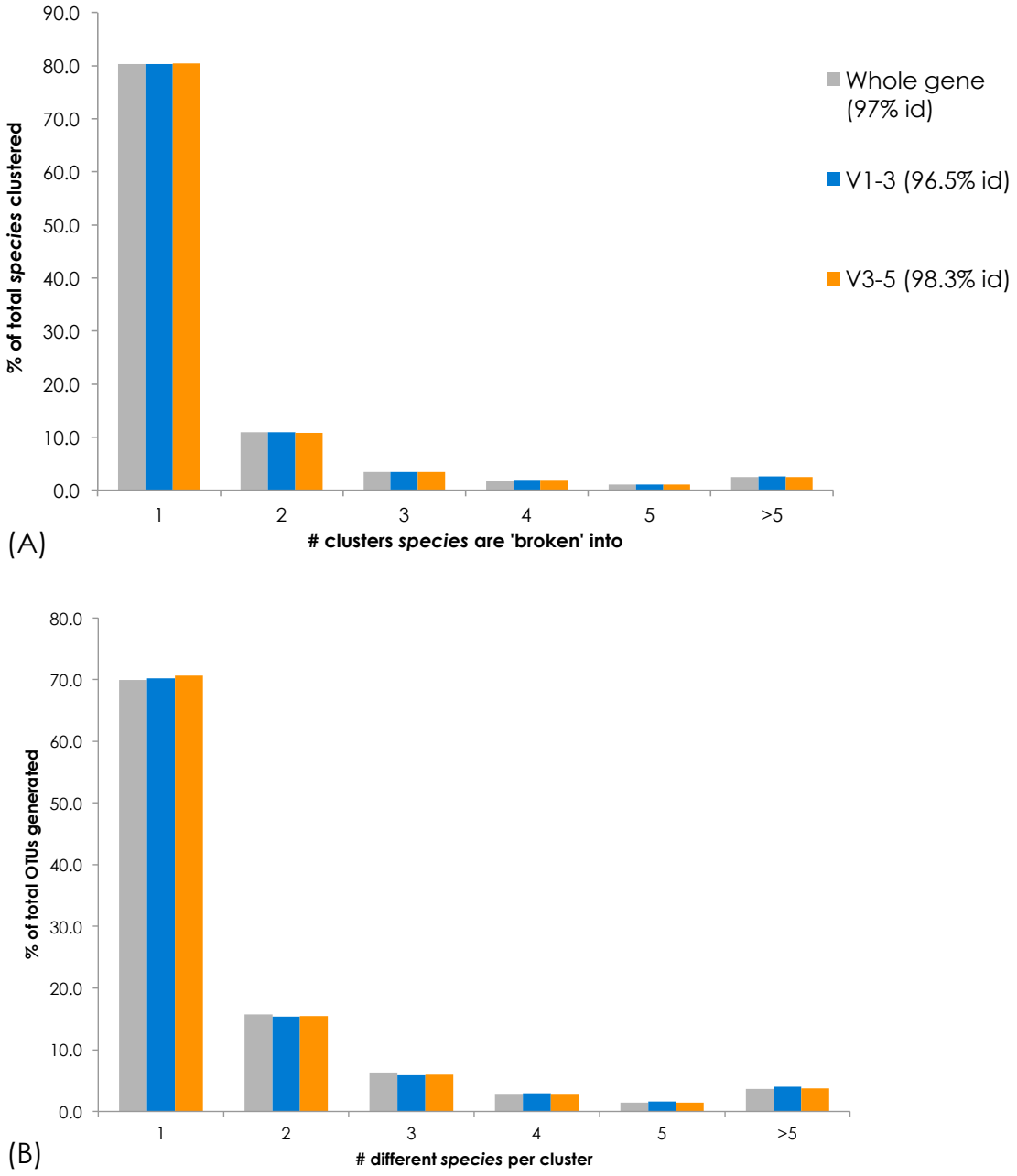


Figure 3.3: Whole-gene sequences and gene-fragments spanning V1-3 and V3-5 were clustered using UCLUST at 97%, 96.5% and 98.3% id respectively. (3.3A) shows the percentage of total *species* (6246) that were either clustered into a single OTU, 'broken' into two OTUs, and so on. The first set of bars represents the percentage of total *species* that were not broken into different OTUs (~80% for all three gene regions), the second set of bars represents the percentage broken into two clusters, and so on. (3.2B) Shows what percentage of the total OTUs generated by each gene region are contaminated (containing more than one species). The first set of bars represent OTUs that are not contaminated, containing only one *species*. The second set of bars represents the percent of OTUs generated that contained two different *species*, and so on.

The V3-5 region generated 5428 OTUs when clustered at 97% id, with approximately 61% (2410 OTUs) being identical to whole-gene sequence generated OTUs at 97% id. When the V3-5 region was clustered at 98.3% id, 5438 OTUs were generated, and approximately 61% of these were identical to those produced by whole-gene sequences clustered at 97%.

These results indicate that the new id thresholds used produce clustering results that provide a similar snapshot of diversity, and could be used as a proxy for 97% id. However, these values are very close to the original 97% id value, and the new V3-5 region id value is higher than 97%, making this less practical than 97%. Additionally, the percent of OTUs in common (between the variable regions at new id thresholds and whole-gene sequences at 97% id) remained fairly constant. These new id values did not increase the effectiveness of gene-fragments generating OTUs that are equivalent to whole-gene generated OTUs. Therefore, when using UCLUST, the most appropriate action to take when clustering V1-3 and V3-5 regions would be to continue using the traditional 97% id threshold.

3.2.2. CD-HIT-EST id threshold evaluation

Similar to Figure 3.2, Figure 3.4 is a plot of the total OTUs produced by whole-gene and gene-fragments clustered across a range of id thresholds, now using CD-HIT-EST. From this plot, it appeared that clustering gene-fragments spanning the V1-3 region at 94%, and gene-fragments spanning the V3-5 region at 96% produced similar amounts of OTUs; 4894 and 4938 total OTUs for the V1-3 and V3-5 regions respectively versus 4978 OTUs for whole-genes at 97% id (values listed in Table A.4).

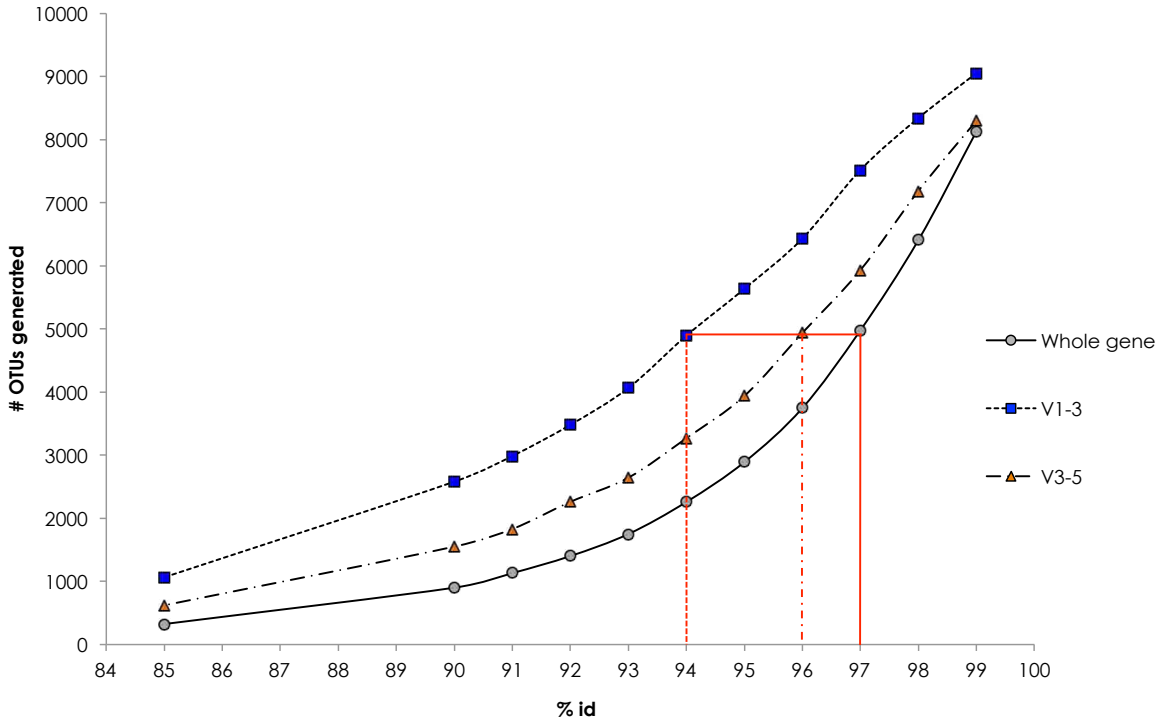


Figure 3.4: Plot of total OTUs generated by CD-HIT-EST using whole-gene sequences, V1-3 sequences and V3-5 sequences across a range of % id values. Solid red lines have been used to highlight where the V1-3 and V3-5 curves line up with the whole-gene curve at 97% id. Dashed red lines (matching the dashed lines used for each V region curve), show what % id on the x-axis appears to produce results similar to whole-gene clustering at 97% id.

An analysis of the broken *species* was conducted comparing whole-gene sequences clustered at 97% id versus the gene-fragments clustered at their new respective thresholds (V1-3 at 94%, V3-5 at 96%). As Figure 3.5A shows, results were similar to those found in section 2.2.2 (CD-HIT-EST fragment evaluation). All three gene regions showed a similar trend in how *species* were broken; approximately 80% of species were not broken, 11% were broken into two different OTUs, and the remaining 10% of *species* were broken into three or more different OTUs (Tables A.7 and A.8 list broken *species* values and percentages). There was no significant deviance in broken *species* between

whole gene sequences and the V1-3 and V3-5 regions (V1-3: $X^2 = 1.0634$, df = 14, $p > 0.05$; V3-5: $X^2 = 1.1276$, df = 14, $p > 0.05$).

Figure 3.5B illustrates how the gene-fragments, at their new respective id thresholds, contaminated clusters in comparison to whole-gene sequences clustered at 97% id. The number of uncontaminated clusters generated by all three gene regions were within 1% of each other at approximately 70%; 70% of OTUs contained a single species (refer to Tables A.9 and A.10 for contaminated cluster counts and percentages). A chi square test found that there was no significant deviance in the number of contaminated clusters produced by the V1-3 or V3-5 gene regions at their new id thresholds in comparison to numbers produced by the whole gene sequences at 97% id (V1-3: $X^2 = 0.0953$, df = 9, $p > 0.05$; V3-5: $X^2 = 0.1551$, df = 9, $p > 0.05$).

As was done in section 3.2.1 (UCLUST id evaluation), the differences between the OTUs generated by V1-3 and V3-5 at new id thresholds and whole-gene sequences at 97% id was examined. The V1-3 region generated 7519 OTUs when clustered at 97% id by CD-HIT-EST. Approximately 47% of these OTUs (3498) were found to be identical to OTUs generated by whole-gene sequences clustered at 97%. When the V1-3 region was clustered at 94% id, 4894 OTUs were generated, and approximately 55% of these were identical to those produced by whole-gene sequences clustered at 97%. The V3-5 region generated 5428 OTUs when clustered at 97% id, with approximately 51% (3040 OTUs) being identical to whole-gene sequence generated OTUs at 97% id. When the V3-5 region was clustered at 96% id, 4938 OTUs were generated, and approximately 52% (2554) of these were identical to those produced by whole-gene sequences clustered at 97%.

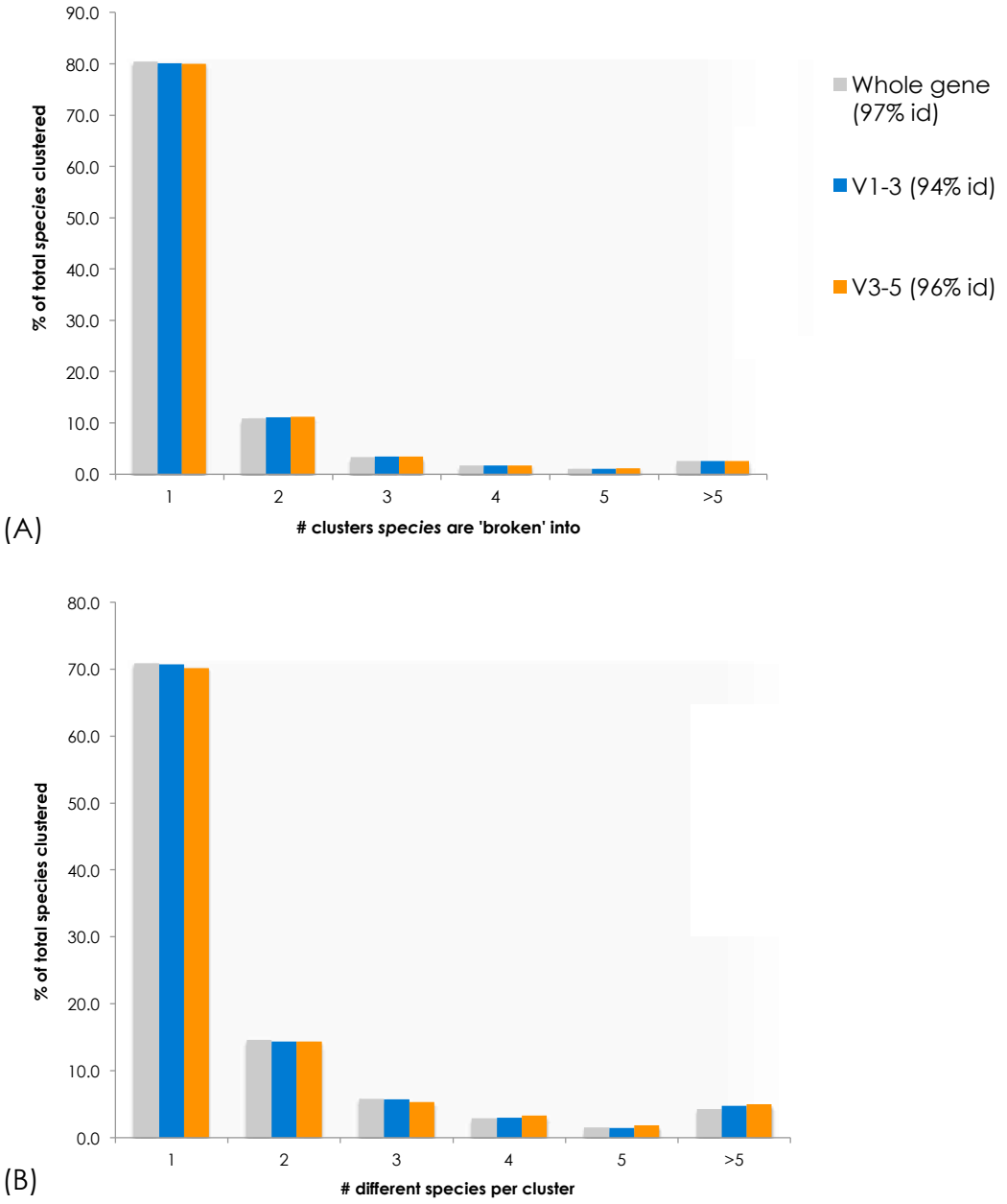


Figure 3.5: Whole-gene sequences and gene-fragments spanning V1-3 and V3-5 were clustered using CD-HIT-EST at 97%, 94% and 96% id respectively. (3.3A) Shows the percentage of total *species* (6246) that were either clustered into a single OTU, 'broken' into two OTUs, and so on. The first set of bars represents the percentage of total *species* that were not broken into different OTUs (~80% for all three gene regions), the second set of bars represents the percentage broken into two clusters, and so on. (3.5B) Shows what percentage of the total OTUs generated by each gene region are contaminated (containing more than one *species*). The first set of bars represent OTUs that are not contaminated, containing only one *species*. The second set of bars represents the percent of OTUs generated that contained two different *species*, and so on.

These results indicate that using alternative id thresholds when clustering with CD-HIT-EST can be effective at replicating the diversity measured by whole-gene sequences at 97% id; there was no significant difference in the number of broken *species* or contaminated OTUs between whole-gene sequences clustered at 97% id and the V1-3 and V3-5 regions clustered at 94% and 96% respectively. Additionally, these new id thresholds for V1-3 and V3-5 aided in generating OTUs that were more coherent with those generated by whole-gene clustering at 97% id. Of note is V1-3 region; from clustering at 97% id to 94% id, the number on OTUs in common with whole-gene clustering rose 8%. This suggests that, when clustering V1-3 gene-fragments with CD-HIT-EST, 94% id is a more effective threshold to obtain accurate results.

3.2.3. Identity evaluation: conclusion

This study has shown that alternative id (identity) thresholds can be used to generate OTUs that adequately represent what whole-gene clustering at 97% id would provide. CD-HIT-EST lowered this traditional id threshold by 3% and 1% for the V1-3 and V3-5 regions respectively, while UCLUST lowered this threshold by 0.5% for the V1-3 region, and required the threshold for the V3-5 region to be raised to 98.3% (for this region to adequately represent whole-gene clustering at 97% id). Lower thresholds allow clustering algorithms to function faster (less sequence comparisons required). Therefore the 3% id decrease from 97 to 94 for the V1-3 region (when using CD-HIT-EST) is of special note.

These two clustering algorithms use different definitions of id when clustering, which accounts for the variation in alternative id threshold values between the two. It has been noted that clustering algorithm suites often generate varied numbers of OTUs (as

well as variation in the content of these OTUs), and the results of this study support these findings (Huse *et al*., 2010; Bonder *et al*., 2012). Therefore, in addition to providing alternative id thresholds to aid in making studies examining different variable regions of the 16S rRNA gene comparable, this study could perhaps also aid in making the results of different clustering algorithms comparable as well.

Both clustering algorithms varied in the total OTUs generated. As it was known that 6246 different species were in the database being clustered, one would expect that the number of OTUs generated by these clustering algorithms would be close to this number. Contrast this expectation with the reality; CD-HIT-EST generated 4978 OTUs and UCLUST generated 5427 OTUs when clustering whole 16S rRNA genes at 97% id. An analysis of this finding indicates that both algorithms underestimated diversity. However, this finding actually highlights one of the major limitations of 16S rRNA cluster analysis; that similar *species* can have 16S rRNA sequences with >97% sequence identity, and be clustered into the same OTU. A prominent example of this is the *Shigella* spp./*E. coli* problem; while historically these two *species* have been considered separate, both belong to the family *Enterobacteriaceae* and biochemically there can be difficulty separating the two (Fukushima *et al*., 2002). Additionally, *Shigella* spp. are considered human pathogens, and certain stains of *E. coli* cause *Shigella*-like symptoms (e.g. diarrhea, etc.), making differentiation even more difficult. On the basis of DNA homology, many researchers consider them to be a single *species* (Brenner, 1984), and 16S rRNA clustering analysis places both in the same OTU at 97% id. Indeed, in this study, regardless of gene region or id threshold used, it was found that at least one strain of *E. coli* and *Shigella* were grouped in the same OTU. This example highlights again the

grey area that exists in grouping microbes at the *species* level; traditional taxonomist categories such as morphology or behavioral traits cannot be easily applied to microbes, but molecular methods (such as 16S rRNA analyses) do not always have the resolution to distinguish very similar *species*/strains (Fraser *et al*., 2009).

However, despite this lack of resolution for some *species*, the 16S rRNA gene still provides an accurate and rapid measure of diversity in microbial populations, and additional molecular analyses used in conjunction with 16S rRNA clustering may provide this needed resolution. For instance, studies have shown that the *gyrB* gene, which encodes the subunit B protein of DNA gyrase (relieves torsional strain caused by helicase during DNA replication), can show sharper separations between similar species such as *Shigella* and *E. coli,* and *species* within the *Bacillus subtilis* group (Fukushima *et al*., 2002; Wang *et al*., 2007). An interesting addition to this study would be to examine 16S rRNA clustering at alternative id thresholds in conjunction with *gyrB* gene clustering as a means of minimizing similar species grouping together.

A key objective of the Human Microbiome Project (HMP) is to determine what bacterial species are considered standard constituents of the normal microbial composition of healthy humans (Kuczynski *et al.*, 2012). Using these data, further research can be done to examine how perturbing this standard composition can influence human health. Traditional diagnosis of disease followed the paradigm of 'one pathogen for one disease.' However, the diversity of communities among different sample body regions (the gut, oral cavity, genital and skin) varies immensely; even between consecutive samples from the same region because bacterial populations within an individual change over time for a variety of reasons, such as diet, health or age

(Turnbaugh *et al.*, 2007). The HMP has indicated that rather than identifying a single pathogen, observing how the microbiomes of individual body sites change during the course of an illness may be more important for health care. The ability to have species level differentiation in the medical setting may not be as critical as previously thought; relating a patient's complex medical history and symptoms with 16S rRNA analyses could collectively contribute to more effective diagnosis and treatment. 330 families of bacteria are represented in the 9700 sequences examined in this study. Interestingly, it was found that contaminated OTUs tended to contain *species* that had been broken into numerous OTUs; *species* that were broken into five or more OTUs were much more likely to be found in OTUs with other *species*. For example, *Shigella* was broken into five OTUs, and all five OTUs contained at least one sequence from another *species*; none of the *Shigella* sequences remained isolated. *Species* that had been broken into two, three and four OTUs rarely had other species in their cluster. Families had an average of 18 sequences in the LTP database, representing on average six *species*. The three families that had the highest incidence of broken *species* were *Enterobacteriaceae* (221 sequences), *Pseudomonadaceae* (138 sequences), and *Enterococcaceae* (50 sequences). However, this may partly be due to the greater selection of species within the database for these families; since more species were represented for these families, there were more chances for a species to be broken. As Table A.11 shows, the phylum *Proteobacteria* (to which *Enterobacteriaceae* and *Pseudomonadaceae* belong) and the phylum *Firmicutes & Tenericutes* (to which *Enterococcaceae* belongs) both had the greatest representation in the LTP database. Approximately 60% of the sequences and *species* in the database belong to these two phyla. Approximately 27% of the *species* that were broken belong to

Proteobacteria, and another 28% to *Firmicutes & Tenericutes* (see Table A.11). This is not surprising, as these two phyla are the most diverse and abundant on Earth (Kormas, 2011). However, it shows that the LTP database has a bias towards containing more abundant *species*.

Interestingly, the genus *Pseudomonas* (family *Pseudomonadaceae*) has undergone numerous taxonomic revisions in the past, as historically the genus was a 'catch-all' with a broad (and vague) phenotypic definition (Özen and Ussery, 2012). Recently the genus has become smaller due to a more refined definition, which led to numerous bacteria being moved to other genera (Özen and Ussery, 2012). However, according to Dworkin and Stanley, between species similarities of *Pseudomonas* ranges from 93 – 99.9% (2006). Additionally, it has been suggested that some genera, such as *Azotobacter,* may actually be *Pseudomonas* (Dworkin and Stanley, 2006; Özen and Ussery, 2012). The families *Streptococcaceae* (84 sequences), *Pasteurellaceae* (67 sequences) and *Vibrionaceae* (116 sequences) appeared to have the highest incidence of contaminated (mixed) OTUs. These finding are in accordance with previously published results, for instance the low variability of genera such as *Streptococcus* (family *Streptococcacea*e) would explain this high incidence of contaminated OTUs (Lal *et al*., 2011). S*treptococcaceae* belongs to the phylum *Firmicutes & Tenericutes*, and *Pasteurellaceae* and *Vibrionaceae* both belong to the phylum *Proteobacteria*. As mentioned previously, these two phyla have the greatest representation in the LTP database, and as shown in Table A.11, approximately 36% of contaminated (mixed) OTUs were generated by *species* from the phylum *Proteobacteria*. Approximately 28% of mixed OTUs were generated by *species* of the phylum *Firmicutes & Tenericutes*.

Of note is that many of these bacterial families have significance in the health setting: for instance the *Enterobacteriaceae* family, while encompassing many harmless symbionts, also contains numerous potential pathogens such as *Salmonella*, *E. coli* and *Shigella* (Gupta *et al.,* 2011). Not only does this family have numerous broken species, it also generated many contaminated OTUs, which could have negative implications in utilizing 16S rRNA analysis in medical settings. However, as discussed earlier, 16S rRNA analyses can be used in conjunction with other diagnostic tools, such as an evaluation of patient symptoms and history.

Another limitation of 16S rRNA clustering is the selection of the variable region(s) to examine. A span of variable regions is commonly used for clustering (i.e. fragments spanning the V1-3 and V3-5 regions as seen in human microbiome research, and as done in this study), but gene fragments spanning a single variable region can also be used. No single variable region can differentiate between all bacteria, and certain regions provide a greater advantage to detecting specific bacteria (Vilo and Dong, 2012). For instance, the V1 region has been found to best differentiate between *Staphylococcus aureus* strains, and the V2 and V3 regions have been found to have great differentiating capability for most bacteria to the genus level, except for bacteria closely related to enterobacteriaceae (Chakravorty *et al*., 2007). As mentioned previously, in this study *E. coli* and *Shigella spp*. were always found mixed among one or more OTUs, illustrating this finding for the V2 and V3 regions.

Previous studies have shown that clustering the same dataset using different V regions can generate varied results, indicating the limitations of any single 16S rRNA gene fragment in simulating whole-gene clustering (Chakravorty *et al*., 2007; Mizrahi-

Man *et al*., 2013; Schloss, 2009). Genetic diversity decreases in general along the length of the 16S rRNA gene, which would influence the genetic diversity observed depending on the variable region(s) examined (Schloss, 2009). However, this highlights the need for examining how gene fragments have been, and are currently, used in place of whole gene sequences and suggesting new workflows when utilizing these fragments. For instance, fewer sequence differences are needed at a 0.03% OTU threshold when considering shorter reads (Schloss, 2009). In combination with sequencing and alignment artifacts, this can lower the confidence in these OTUs (Schloss, 2009). Of note, however, is that the majority of studies cluster sequences at the traditional 97% id. This is a symptom of a larger problem, namely that microbial community assessment should change when considering whole genes versus gene fragments. For instance, it has been suggested that analyses of mock communities with biological samples aides in the selection of variable regions for study, which could greatly reduce errors in OTU generation (Kozlch, 2013). This study provides evidence that alternative id thresholds can be used to provide adequate representation of whole-gene clustering at 97% id, and this could allow different studies to compare results when examining the same environment using different variable regions of the 16S rRNA gene. A possible result of this would be that different V regions could be used to aid in identifying more *species* while possibly minimizing the wide variation in number of OTUs generated (and the variation in sequences within them).

Recommendations for further study would include examining different V regions; the ultimate goal of a study of this nature would be to provide a set of alternate id threshold guidelines for all nine variable regions when clustering. Another important

addition to this study would be to cluster a set of environmental sequences; this would provide a data set more representative of true sequencing results, rather than the highly curated and annotated dataset used in this study (not a true representation of the average genomic study).

4. Summary

This study examined if clustering16S rRNA gene-fragments at 97% id provides results comparable to whole-gene clustering as well as whether new id thresholds could and should be used in place of 97% id; higher or lower id thresholds may be required for gene-fragments to produce clusters that are more representative of whole-gene clustering. It was found that indeed, 16S rRNA gene-fragments can produce clusters comparable to whole 16S rRNA genes. Two clustering algorithms were used: UCLUST and CD-HIT-EST. While alternative id thresholds were proposed for clustering the V1-3 and V3-5 regions using UCLUST, these values are very close to the standard 97%, and these new thresholds were no more effective than 97% in aiding these gene-fragments to represent the OTUs generated by whole-gene sequences.

In contrast, two possible alternative id thresholds have been suggested for the V1-3 and V3-5 region of the 16S rRNA gene; results indicate that the V1-3 region clustered at 94% id by CD-HIT-EST not only produced OTUs representative of whole-gene clustering but that more OTUs at this new threshold were identical to OTUs generated by whole-gene sequences at 97% id than OTUs generated by the V1-3 region at 97% (currently considered a standard). Additionally, the V3-5 regions clustered at 96% appeared effective in representing the OTUs generated by whole-gene sequences.

References

Bassi, S. 2007. A Primer on Python for Life Science Researchers. PLoS Comput Biol, 3(11): e199. doi:10.1371/journal.pcbi.0030199

Bonder, M.J., Abeln, S., Zaura, E., and Brandt, B.W. 2012. Comparing clustering and pre-processing in taxonomy analysis. Bioinformatics (Oxford, England), 28: 2891-2897. doi: 10.1093/bioinformatics/bts552.

Brenner, D.J. 1984. Family I. Enterobacteriaceae, p.408–420. In N. R. Krieg and J. G. Holt (ed.), Bergey's manual of systematic bacteriology, vol. 1. Williams & Wilkins, Baltimore, Md.

Bybee, S.M., BrackenGrissom, H., Haynes, B.D., Hermansen, R.A., Byers, R.L., Clement, M.J., Udail, J.A., Wilcox, E.R., and Crandall, K.A. Targeted Amplicon Sequencing (TAS): A Scalable Next-Gen Approach to Multilocus, Multitaxa Phylogenetics. Genome Biol. Evol., 3:1312-1322.

Chakravorty, S., Helb, D., Burday, M., Connell, N., and Alland, D. 2007. A detailed analysis of 16S ribosomal RNA gene segments for the diagnosis of pathogenic bacteria. J Microbiol Methods, 69(2). doi: 10.1016/j.mimet.2007.02.005

Clayton, T.A., Baker, D., Lindon, J.C., Everett, J.R. and Nicholson, J.K. 2009. Pharmacometabonomic identification of a significant host-microbiome metabolic Interaction affecting human drug metabolism. Proc. Natl Acad. Sci. USA 106, 14728–14733.

Cohan, F.M. 2002. What are bacterial species? Annu. Rev. Microbiol, 56: 457-487.

Cohan, F.M. and Perry, E.B. 2007. A Systematics for Discovering the Fundamental Units of Bacterial Diversity. Current Biology, 17, 373-386.

de Queiroz, K. 1998. The General Lineage Concept of Species, Species Criteria, and the Process of Speciation: A Conceptual Unification and Terminological Recommendations. In D. J. Howard and S. H. Berlocher (Eds.), Endless Forms: Species and Speciation (57-75). New York: Oxford University Press.

de Queiroz, K. Ernst Mayr and the modern concept of species. Proc. Natl. Acad. Sci, 102: 660-6607.

Doolittle, W.F. 2012. Population Genomics: How Bacterial Species Form and How They Don't Exist. Current Biology, 22(11): 451-453.

Doolittle, W.F. and Papke, R.T. 2006. Opinion: Genomics and the bacterial species problem. Genome Biology, 7(1): 116. doi:10.1186/gb-2006-7-9-116.

Doolittle, W.F. and Zhaxybayeva, O. 2009. On the origin of prokaryotic species. Genome Res, 19: 744-756. doi:10.1101/gr.086645.108.

Dworkin, M. and Falkow, S. 2006. *The Prokaryotes: Vol. 6: Proteobacteria: Gamma Subclass*. Springer Science & Business Media, pg. 658.

Edgar, R.C. 2010. Search and clustering orders of magnitude faster than BLAST. Bioinformatics, 26(19), 2460-2461.doi: 10.1093/bioinformatics/btq461.

Fowler, J., Cohen, L., Jarvis P., and Wiley, J. 1998. *Practical statistics for field biology*. Wiley Chichester.

Fraser, C., Alm, E.J., Polz, M.F., Spratt, B.G., and Hanage, W.P. 2009. The Bacterial Species Challenge: Making Sense of Genetic and Ecological Diversity. Science, 323(6): 741-746.

Fukushima, M., Kakinuma, K., and Kawaguchi, R. 2002. Phylogenetic Analysis of *Salmonella, Shigella,* and *Escherichia coli* Strains on the Basis of the *gyrB* Gene

Sequence. J. Clin. Microbiol, 40(8), 2779–2785. doi: 10.1128/JCM.40.8.2779–2785.2002.

Gardner, P., Wilm, A., and Washietl, S. 2005. A benchmark of multiple sequence alignment programs upon structural RNAs. Nucleic Acids Research, 33(8):2433-2439.

George, I., Stenuit, B., and Agathos, S. 2010. Application of Metagenomics to Bioremediation. In D. Marco (Ed), *Metagenomics: Theory, Methods and Applications*. Caister Academic Press.

Gevers, D., Cohan, F.M., Lawrence, J.G., Spratt, B.G., Coenye, T., Feil, E.J., Stackebrandt, E., Van de Peer, Y., Vandamme, P., Thompson, F.L., and Swings, J. 2005. Re-evaluating prokaryotic species. Nature Reviews Biology, 3: 733-739.

Gevers, D., Knight, R., Petrosino, J.F., Huang, K., McGuire, A.L., Birren, B.W., Nelson, K.E., White, O., Methe´, B.A., and Huttenhowe, C. 2012. The Human Microbiome Project: A Community Resource for the Healthy Human Microbiome. PLoS Biol, 10(8): e1001377. doi:10.1371/ journal.pbio.1001377.

Gilbert, J.A., Dupont, C.L. 2011. Microbial metagenomics: beyond the genome. Ann Rev Mar Sci, 3: 347-71.

Gupta, N., Limbago, B.M., Patel, J.B., and Kallen, A.J. 2001. Carbapenem-Resistant *Enterobacteriaceae*: Epidemiology and Prevention. Clinical Infectious Diseases, 53(1): 60-67.

Huse, S.M., Welch, D.M., Morrison, H.G. Sogin, M.L. 2010. Ironing out the wrinkles in the rare biosphere through improved OTU clustering. Environ Microbiol, 12: 1889-1898. doi: 10.1111/j.1462-2920.2010.02193.x.

Huse, S.M., Yes, Y., Zhou, Y., and Fodor, A.A. 2012. A Core Human Microbiome as Viewed through 16S rRNA Sequence Clusters. PLoS ONE, 7(6): e34242. doi: 10.1371/journal.pone.0034242

Janda, M.J., and Abbott, S.L. 2007. 16S rRNA Gene Sequencing for Bacterial Identification in the Diagnostic Laboratory: Pluses, Perils, and Pitfalls. J. Clin. Microbiol., 45(9): 2761-2764.

Kitahara, K., Yasutake, Y., and Miyazaki, K. 2012. Mutational robustness of 16S ribosomal RNA, shown by experimental horizontal gene transfer in Escherichia coli. PNAS Early Edition, 109(47): 19220–19225. doi: 10.1073/pnas.1213609109

Kormas, K.A. 2011. Interpreting diversity of Proteobacteria based on 16S rRNA gene copy number. In: Sezenna ML (Ed.), *Proteobacteria: Phylogeny, metabolic diversity and ecological effects*. Hauppauge, NY: Nova Publishers.

Kozlch, J.J., Westcott, S.L., Baxter, N.T., Highlander, S.K., and Schloss, P.D. 2013. Development of a Dual-Index Sequencing Strategy and Curation Pipeline for Analyzing Amplicon Sequence Data on the MiSeq Illumina Sequencing Platform. Appl. Environ. Microbiol, 79(17), 5112-5120. doi: 10.1128/AEM.01043-13.

Kuczynski, J., Lauber, C.L., Walters, W.A., Parfrey, L.W., Clemente, J.C., Gevers, D., and Knight, R. 2012. Experimental and analytical tools for studying the human microbiome. Nature Reviews, 13:47-58.

Lal, D., Verma, M., and Lal, Rup. 2001. Exploring internal features of 16S rRNA gene for identification of clinically relevant species of the genus *Streptococcus*. Annals of Clinical Microbiology and Antimicrobials, 10(28): doi:10.1186/1476-0711-10-28.

Li, H., Zhang, Y., Li D., Xu, H., Chen, G., Zhang, C. 2009. Comparisons of different hypervariable regions of rrs genes for fingerprinting of microbial communities in paddy soils. Soil Biology & Biochemistry, 41: 954 – 968.

Li, L., McCorkle, S.R., Monchy, S., Taghavi, S., and van der Lelie, D. 2009. Bioprospecting metagenomes: glycosyl hydrolases for converting biomass. Biotechnology for Biofuels. 2(10): doi:10.1186/1754-6834-2-10.

Li, W., Jaroszewski, L., and Godzik, A. 2001. Clustering of highly homologous sequences to reduce the size of large protein databases. Bioinformatics Applications Note, 17(3): 282-283.

Li, W., and Godzik, A. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics Applications Note, 22(13): 1658-1659.

Li, W., Fu, L., Niu, B., Wu, S. and Wooley, J. 2012. Ultrafast clustering algorithms for metagenomic sequence analysis. Brief Bioinform., 13(6): 656-668.

Marchandin, H., Teyssier, C., de Buochberg, M.S., Jean-Pierre, H., Carriere, C., and Jumas-Bilak, E. 2003. Microbiology, 149(6): 1493-1501.

Mishler, B.D. and Brandon, R.N. 1987. Individuality, pluralism, and the phylogenetic species concept. Biol. Philos, 2: 397-414.

Mizrahi-Man, O., Davenport, E.R., and Gilad, Y. 2013. Taxonomic Classification of Bacterial 16S rRNA Genes Using Short Sequencing Reads: Evaluation of Effective Study Designs. PLoS ONE 8(1): e53608. doi: 10.1371/journal.pone.0053608.

Özen, A.I. and Ussery, D.W. 2012. Defining the *Pseudomonas* Genus: Where Do We Draw the Line with *Azotobacter*? Microbial Ecology, 63(2): 239-248.

National Research Council. 2007. *The New Science of Metagenomics: Revealing the Secrets of Our Microbial Planet*. Washington, DC: The National Academies Press.

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., Glockner, F.O. 2013. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Research, 41:590-596.

R Core Team. 2014. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. [http: //www.R-project.org].

Rodríguez-Ezpeleta N., Hackenberg, M., and Aransay, A.M. (Eds.). 2011. Bioinformatics for High Throughput Sequencing. New York, NY: Springer Science+Business Media.

Rohlke, F., and Stollman, N. 2012. Fecal microbiota transplantation in relapsing *Clostridium difficile* infection. Ther Adv Gastroenterol, 5(6): 403-420.

Rosenfeld, G., and Bressler, B. 2010. Mycobacterium avium paratuberculosis and the etiology of Crohn's disease: a review of the controversy from the clinician's perspective. Can J Gastroenterol, 24(10): 619-624.

Rosselló-Mora, R. 2006. 2 DNA–DNA Reassociation Methods Applied to Microbial Taxonomy and Their Critical Evaluation. In E. Stackbrandt (Eds.), *Molecular*

Identification, Systematics, and Population Structure of Prokaryotes (pp. 23-56). Berlin: Springer-Verlag.

Sabree, Z.L., M.R. Rondon, and J. Handelsman. 2009. Metagenomics. In, *Encyclopedia of Microbiology* (Third Edition). Academic Press, pp. 622-632.

Schloss, P.D. 2009. The Effects of Alignment Quality, Distance Calculation Method, Sequence Filtering, and Region on the Analysis of 16S rRNA Gene-Based Studies. PLoS Computational Biology, 6(7): e1000844. doi:10.1371/journal.pcbi.1000844.

Schloss, P.D., and Handelsman, J. 2005. Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness. Appl Environ Microbiol, 71(3):1501-1506.

Smit, S., Widmann, J., and Knight, R. 2007. Evolutionary rates vary among rRNA structural elements. Nucleic Acids Research, 35(10): 3339-3354.

Stackebrandt, E., Frederiksen, W., Garrity, G.M., Grimont, P., Kämpfer, P., Maiden, M., Nesme, X., Rosselló-Mora, R., Swings, J., Trüper, H.G., Vauterin, L., Ward, A.C., and Whitman. 2002. International Journal of Systemaitc and Evolutionary Microbiology, 52:1043-1047. doi: 10.1099/ijs.0.02360-0.

Stackebrandt, E., and Goebel, B.M. 1994. Taxonomic Note: A Place for DNA-DNA Reassociation and 16s rRNA Sequence Analysis in the Present Species Definition in Bacteriology. International Journal of Systematic Bacteriology, 44(4):846-849.

Tortoli, E. 2003. Impact of Genotypic Studies on Mycobacterial Taxonomy: the New Mycobacteria of the 1990s. Clinical Microbiology Reviews, 16(2): 319-354. doi: 10.1128/CMR.16.2.319-354.2003.

Turnbaugh, P. J., Ley, R.E., Hamady, M., Fraser-Liggett, C.M., Knight, R., and Gordon, J.I. 2007. The human microbiome project. Nature, 449(7164): 804-810.

Ueda, K., Seki, T., Kudo, T., Yoshida, T., and Kataoka, M. 1999. Two Distinct Mechanisms Causes Heterogeneity of 16S rRNA. J. Bacteriol, 181(1): 78-82.

Van de Peer, Y., Chapelle, S., and De Watcher, Rupert. 1996. A quantitative map of nucleotide substitution rates in bacterial rRNA. Nucleic Acids Research, 24(17), 3381-3391.

Větrovský, T., and Baldrian, P. 2013. The Variability of the 16S rRNA Gene in Bacterial Genomes and Its Consequences for Bacterial Community Analyses, PLoS ONE 8(2): e57923. doi:10.1371/journal.pone.0057923

Vilo, C. and Dong, Q. 2012. Evaluation of the RPD Classifier Accuracy Using 16S rRNA Gene Variable Regions. Metagenomics, 1: 5 pages. doi: 10.4303/mg/235551.

Wang, L.T., Lee, F.L., Tai, C.J., and Kasai, H. 2007. Comparison of gyrB gene sequences, 16S rRNA gene sequences and DNA-DNA hybridization in the Bacillus subtilis group. Int J Syst Evol Microbiol, 57: 1846-50.

Wickham, H. 2009. *ggplot2: elegant graphics for data analysis*. Springer Publishing Company, Incorporated.

Willey, J., Sherwood, L., and Woolverton, C. 2011. *Prescott's Microbiology* (8th ed.). New York: McGraw Hill. pp. 731–737.

Appendix

Table A.1: Number of clusters generated after clustering three gene regions using UCLUST at various id levels. Alternative id threshold values are highlighted in red.

Gene region	Clustering id	Total	Singletons	>1 Sequence
Whole gene	97	5427	3767	1660
V1-3	97	5929	4277	1652
V3-5	97	3911	2312	1599
V1-3	96.5	5362	3730	1632
V3-5	98.3	5438	3806	1632

Table A.2: Counts of broken *species* produced after clustering three gene regions using UCLUST. Alternative id threshold values are highlighted in red.

# OTUs *species* 'broken' into	Whole gene (97% id)	V1-3 (97% id)	V3-5 (97% id)	V1-3 (96.5% id)	V3-5 (98.3% id)
1	5015	5012	5026	5018	5018
2	683	689	676	684	685
3	216	210	211	211	211
4	106	109	108	110	105
5	68	66	69	64	69
6	33	33	29	35	33
7	25	27	28	24	26
8	25	25	23	25	24
9	14	13	15	13	14
10	11	12	11	12	11
11	7	7	7	7	7
12	5	4	5	5	4
13	4	5	3	3	5
14	4	4	6	6	4
15	30	30	29	29	30
>5	158	160	156	159	158
Total Clusters =	5427	5929	5362	3911	5438

Table A.3: Percent of total *species* that were 'broken' (based on Table A.3). Alternative id threshold values are highlighted in red.

# OTUs *species* 'broken' into	Whole gene (97% id)	V1-3 (97% id)	V3-5 (97% id)	V1-3 (96.5% id)	V3-5 (98.3% id)
1	80.29	80.24	80.47	80.34	80.47
2	10.93	11.03	10.82	10.95	10.82
3	3.46	3.36	3.38	3.38	3.38
4	1.70	1.75	1.73	1.76	1.73
5	1.09	1.06	1.10	1.02	1.10
6	0.53	0.53	0.46	0.56	0.46
7	0.40	0.43	0.45	0.38	0.45
8	0.40	0.40	0.37	0.40	0.37
9	0.22	0.21	0.24	0.21	0.24
10	0.18	0.19	0.18	0.19	0.18
11	0.11	0.11	0.11	0.11	0.11
12	0.08	0.06	0.08	0.08	0.08
13	0.06	0.08	0.05	0.05	0.05
14	0.06	0.06	0.10	0.10	0.10
15+	0.48	0.48	0.46	0.46	0.46
>5	2.53	2.56	2.50	2.55	2.50

Table A.4: Counts of clusters containing a given amount of different species after clustering three gene regions using UCLUST. Alternative id threshold values are highlighted in red.

# Species/Cluster	Whole gene (97% id)	V1-3 (97% id)	V3-5 (97% id)	V1-3 (96.5% id)	V3-5 (98.3% id)
1	3796	4312	2337	3765	3841
2	855	880	655	824	841
3	345	351	319	314	325
4	156	164	168	160	153
5	78	78	116	84	76
6	55	39	76	67	58
7	38	33	44	32	39
8	31	13	37	24	27
9	8	14	24	20	18
10+	65	45	135	72	60
>5	197	144	316	215	202
Total Clusters =	5427	5929	3911	5362	5438

Table A.5: Percent of total clusters generated containing a given amount of different species (based on Table A.5). Alternative id threshold values are highlighted in red.

# Species/Cluster	Whole gene (97% id)	V1-3 (97% id)	V3-5 (97% id)	V1-3 (96.5% id)	V3-5 (98.3% id)
1	69.95	72.73	59.75	70.22	70.63
2	15.75	14.84	16.75	15.37	15.47
3	6.36	5.92	8.16	5.86	5.98
4	2.87	2.77	4.30	2.98	2.81
5	1.44	1.32	2.97	1.57	1.40
6	1.01	0.66	1.94	1.25	1.07
7	0.70	0.56	1.13	0.60	0.72
8	0.57	0.22	0.95	0.45	0.50
9	0.15	0.24	0.61	0.37	0.33
10+	1.20	0.76	3.45	1.34	1.10
>5	3.63	2.43	8.08	4.01	3.71

Table A.6: Number of clusters generated after clustering three gene regions using CD-HIT-EST at various id levels. Alternative id threshold values are highlighted in red.

Gene Fragment	Clustering id	Total	Singletons	>1 Sequence
Whole gene	97	4978	3499	1497
V1-3	97	7519	6391	1128
V3-5	97	5932	4479	1143
V1-3	94	4894	3433	1461
V3-5	96	4938	3440	1498

Table A.7: Counts of broken *species* produced after clustering three gene regions using CD-HIT-EST. Alternative id threshold values are highlighted in red.

# OTUs *species* 'broken' into	Whole gene (97% id)	V1-3 (97% id)	V3-5 (97% id)	V1-3 (94% id)	V3-5 (96% id)
1	5024	4990	4997	5004	4999
2	683	698	694	694	697
3	210	221	219	216	214
4	105	106	105	108	105
5	65	69	70	66	71
6	36	35	33	32	32
7	24	26	27	26	28
8	24	25	25	25	24
9	14	14	14	14	15
10	11	12	12	11	11
11	7	7	7	7	7
12	5	4	4	5	4
13	4	4	4	3	6
14	4	5	5	7	3
15	30	30	30	28	30
>5	159	162	161	158	160
Total Clusters =	4978	7519	4894	5932	4938

Table A.8: Percent of total *species* that were 'broken' (based on Table A.7). Alternative id threshold values are highlighted in red.

# OTUs *species* 'broken' into	Whole gene (97% id)	V1-3 (97% id)	V3-5 (97% id)	V1-3 (94% id)	V3-5 (96% id)
1	80.44	79.89	80.00	80.12	80.04
2	10.93	11.18	11.11	11.11	11.16
3	3.36	3.54	3.51	3.46	3.43
4	1.68	1.70	1.68	1.73	1.68
5	1.04	1.10	1.12	1.06	1.14
6	0.58	0.56	0.53	0.51	0.51
7	0.38	0.42	0.43	0.42	0.45
8	0.38	0.40	0.40	0.40	0.38
9	0.22	0.22	0.22	0.22	0.24
10	0.18	0.19	0.19	0.18	0.18
11	0.11	0.11	0.11	0.11	0.11
12	0.08	0.06	0.06	0.08	0.06
13	0.06	0.06	0.06	0.05	0.10
14	0.06	0.08	0.08	0.11	0.05
15+	0.48	0.48	0.48	0.45	0.48
>5	2.55	2.59	2.58	2.53	2.56

Table A.9: Counts of clusters containing a given amount of different species after clustering three gene regions using CD-HIT-EST. Alternative id threshold values are highlighted in red.

# Species/Cluster	Whole gene (97% id)	V1-3 (97% id)	V3-5 (97% id)	V1-3 (94% id)	V3-5 (96% id)
1	3528	6428	4507	3461	3465
2	729	703	787	704	708
3	289	167	275	279	263
4	146	112	119	149	164
5	75	39	66	70	90
6	48	18	51	42	61
7	36	16	33	37	39
8	26	8	25	39	31
9	16	8	16	21	20
10+	85	20	53	92	97
>5	211	70	178	231	248
Total Clusters =	4978	7519	5932	4894	4938

Table A.10: Percent of total clusters generated containing a given amount of different species (based on Table A.9). Alternative id threshold values are highlighted in red.

# Species/Cluster	Whole gene (97% id)	V1-3 (97% id)	V3-5 (97% id)	V1-3 (94% id)	V3-5 (96% id)
1	70.87	85.49	75.98	70.72	70.17
2	14.64	9.35	13.27	14.38	14.34
3	5.81	2.22	4.64	5.70	5.33
4	2.93	1.49	2.01	3.04	3.32
5	1.51	0.52	1.11	1.43	1.82
6	0.96	0.24	0.86	0.86	1.24
7	0.72	0.21	0.56	0.76	0.79
8	0.52	0.11	0.42	0.80	0.63
9	0.32	0.11	0.27	0.43	0.41
10+	1.71	0.27	0.89	1.88	1.96
>5	4.24	0.93	3.00	4.72	5.02

Table A.11: Proportion of the total sequences and *species* represented by the different phyla of bacteria form the SILVA LTP database, as well as the proportion of broken *species* and mixed (contaminated) OTUs generated by the different phyla.

Phylum	Sequences	*Species*	Broken *species*	Mixed OTUs
Proteobacteria	0.3729	0.3519	0.2714	0.3559
Spirochaetes	0.0089	0.0054	0.0040	0.0034
Fusobacteria	0.0039	0.0022	0.0024	0.0021
Deferribacteres	0.0010	0.0006	0.0032	0.0028
Chrysiogenetes	0.0004	0.0003	0.0008	0.0007
Acidobacteria	0.0020	0.0008	0.0016	0.0014
Bacteroidetes	0.0956	0.0922	0.1825	0.1490
Chlorobi	0.0012	0.0018	0.0000	0.0000
Verrucomicrobia	0.0012	0.0019	0.0008	0.0007
Lentisphaerae	0.0002	0.0003	0.0000	0.0000
Chlamydiae	0.0013	0.0014	0.0008	0.0007
Planctomycetes	0.0015	0.0010	0.0016	0.0014
Fibrobacteres	0.0003	0.0002	0.0000	0.0000
Deinococcus-Thermus	0.0077	0.0090	0.0056	0.0048
Nitrospira	0.0008	0.0005	0.0008	0.0007
Actinobacteria	0.2673	0.2192	0.2354	0.1890
Chloroflexi	0.0024	0.0029	0.0016	0.0014
Firmicutes & Tenericutes	0.2214	0.2659	0.2762	0.2772
Aquificae	0.0030	0.0032	0.0040	0.0034
Thermodesulfobacteria	0.0008	0.0010	0.0008	0.0007
Synergistetes	0.0020	0.0027	0.0032	0.0028
Thermotogae	0.0040	0.0037	0.0032	0.0021