NOMENCLATURE ERRORS IN PUBLIC 16S rDNA GENE DATABASES:

STRATEGIES TO IMPROVE THE ACCURACY OF SEQUENCE ANNOTATIONS

by

Kyle Lesack

BHSc Bioinformatics, University of Calgary, 2013

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE
in
THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES
(Bioinformatics)

THE UNIVERSITY OF BRITISH COLUMBIA
(Vancouver)

July 2017

Abstract

Obtaining an accurate representation of the microorganisms present in microbial ecosystems presents a considerable challenge. Microbial communities are typically highly complex, and may consist of a variety of differentially abundant bacteria, archaea, and microbial eukaryotes. The targeted sequencing of the 16S rDNA gene has become a standard method for profiling membership and biodiversity of microbial communities, as the bacterial and archaeal community members may be profiled directly, without any intermediate culturing steps. These studies rely upon specialized 16S rDNA gene reference databases, but little systematic and independent evaluation of the annotations assigned to sequences in these databases has been performed.

This project examined the quality of the nomenclature annotations provided by the 16S rDNA sequences in three public databases: The Ribosomal Database Project, SILVA, and Greengenes. To do that, first three nomenclature resources – the List of Prokaryotic Names with Standing in Nomenclature, Integrated Taxonomic Information System, and Prokaryotic Nomenclature Up-to-Date – were evaluated to determine their suitability for validating prokaryote nomenclature. A core-set of valid, invalid, and synonymous organism names was then collected from these resources, and used to identify incorrect nomenclature in the public 16S rDNA databases. To assess the potential impact of misannotated reference sequences on microbial gene survey studies, the misannotations identified in the SILVA database were categorized by sample isolation source. Methods for the detection and prevention of nomenclature errors in reference databases were examined, leading to the proposal of several quality assurance strategies for future biocuration efforts. These included phylogenetic methods for the identification of anomalous taxonomic placements, database design principles and technologies for quality control, and opportunities for community assisted curation.

Lay Summary

16S ribosomal DNA (rDNA) gene sequencing is a method for studying entire microbial communities. Because the 16S gene is present in all prokaryotes, and contains species-specific sequences, it may be used to identify unknown bacteria and archaea based upon their 16S sequence. Three specialized databases provide 16S rDNA reference sequences and their corresponding nomenclature annotations: Greengenes, RDP, and SILVA. During the curation of each of these databases, quality control steps were performed with the goal of ensuring that only high-quality 16S sequences were included. However, little systematic and independent evaluation of the annotations assigned to sequences in these databases has been performed.

During this project, lists of valid and invalid prokaryote names were curated. Validation using these lists, revealed the presence of invalid names assigned to sequences in Greengenes, RDP, and SILVA. These methods may be adapted in future database curation efforts to improve annotation quality.

Preface

The sections in this work have not yet been published. Chapter 2 and 3 are in the process of being submitted to peer reviewed scientific journals in the coming months.

The work presented in Chapter 2 and 3 was designed and conducted by Kyle Lesack in consultation with Dr. Inanc Birol. Kyle Lesack wrote the Python scripts used to collect prokaryote nomenclature information and to evaluate 16S rDNA reference gene database annotations. The main text was written by Kyle Lesack with input from Dr. Inanc Birol.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

BMSB: Bergey’s Manual of Systematic Bacteriology

BLAST: Basic Local Alignment Search Tool

CSS: Cascading Style Sheet

DNA: Deoxyribonucleic acid

DSMZ: Deutsche Sammlung von Mikroorganismen und Zellkulturen

FTP: File Transfer Protocol

HTML5: HyperText Markup Language

HVR: HyperVariable Region

ICN: International Code of Nomenclature for Algae, Fungi, and Plants

IJSEM: International Journal of Systematic and Evolutionary Microbiology

INSDC: International Nucleotide Sequence Database Collaboration

ITIS: Integrated Taxonomic Information System

JSON: JavaScript Object Notation

LPSN: List of Prokaryotic Names with Standing in Nomenclature

NCBI: National Center for Biotechnology Information

OTU: Operational Taxonomic Unit

PCR: Polymerase chain reaction

RDP: Ribosomal Database Project

RESTful: Representational State Transfer

rDNA: Ribosomal DNA

RNA: Ribonucleic acid

SSU: Small subunit

TOBA: Taxonomic Outline of Bacteria and Archaea

XML: Extensible Markup Language (XML)

Glossary

16S Gene Survey: Molecular method used to identify prokaryotes present in environmental samples using PCR amplification and sequencing of the 16S rDNA gene. In this approach, universal primers complementary to conserved regions of the 16S rDNA gene are used to target specific hypervariable regions. Differences in the sequences of the hypervariable regions allow for the assignment of taxonomic ranks to the reads in a sample.

16S rDNA: A section of prokaryotic DNA that codes for the 16S ribosomal RNA subunit.

Co-identical: Clones of a single parental strain contained in different culture collections around the world.

Chimera: Sequences composed from two or more parent DNA sequences that have formed during the PCR process. Chimeric amplicons may occur due to a number of causes, with incomplete extension being the most common.

Clustering: The assignment of a set of objects into groups such that objects in the same group are more similar to each other, and more dissimilar to those in other groups.

Gene Conversion: A unidirectional process in which one DNA sequence is converted, partially or completely, to that of another homologous sequence through DNA recombination. As a consequence, the degree of intragenomic sequence identity is increased between different copies of a gene within a genome.

Gene Marker Survey: See 16S gene survey.

Heterotypic Synonym: Synonym that occurs when a distinct species is reclassified under another existing group, resulting in a group containing multiple different strains.

Homotypic Synonym: Synonym that occurs when multiple names are associated to a taxon containing a single strain.

Intersection: The intersection of X ∩ Y is the set containing all elements if and only if they are contained in both X and Y.

Metagenome: The collection of microbial genes and genomes present in a given environment. For data obtained using shotgun sequencing of DNA extracted from a sample, see Shotgun Metagenomics.

Microbiome: The entire population of microorganisms that inhabit a particular habitat, along with their combined genetic material. Some authors consider this definition to include the environmental conditions of the habitat.

Monophyletic: A group consisting of all organisms descended from a single ancestor.

Operational Taxonomic Unit (OTU): A group consisting of clustered reads that roughly corresponds to a taxon. In 16S rDNA gene surveys, OTUs are frequently used as pragmatic approximations of the prokaryote genera or species being studied.

PCR Chimera: See Chimera

Polyphyletic: A group consisting of organisms descended from multiple ancestors. Groups discovered to be polyphyletic are usually reclassified to more accurately represent their evolutionary history.

Rhizoplane: The plant root surface and adhering soil particles.

Rhizosphere: The region of soil directly bordering plant roots. Numerous biochemical processes occur in the rhizosphere, many of which serve functions that contribute to plant health and microbial growth.

Shotgun Metagenomics: The application of shotgun sequencing to metagenomics. In this method, the metagenome DNA is randomly fragmented and sequenced. Because this method does not target a single genetic locus, the sample gene content can be quantified in addition to community membership. Drawbacks to this method include increased cost and difficulty of subsequent bioinformatics analyses.

Type Strain: A strain that serves as a reference point for a named species or subspecies with official standing in the nomenclature. Type strains play a crucial role in taxonomy, as they provide a reference point that can be used when classifying newly described species.

Acknowledgements

I would like to thank my supervisor Dr. Inanc Birol for his support throughout this project. I would also like to thank my committee members Dr. Phil Hieter and Dr. Mel Krajden for the feedback and suggestions provided during our committee meetings. I would like to acknowledge the Canadian Institute of Health Research (CIHR) for providing funding for two years of my studies.

Chapter 1: Introduction

Advances in the field of metagenomics have enabled complex microbial communities to be studied directly, without the need for the isolation and culturing of the individual community members. Two main molecular methods are used in metagenomics: shotgun metagenome sequencing, and targeted marker gene surveys. For shotgun metagenome sequencing, DNA isolated from the target community (e.g., host tissue biopsy or excretion, soil, water) is randomly sheared into fragments that are sequenced independently [1]. The analysis of these data typically includes mapping the sequencing reads to phylogenetically informative regions [2], gene prediction [3], and metagenome reconstruction [4]. This allows both the presence of organisms in the community and their combined metabolic functions to be inferred. Although the shotgun approach has the advantage of probing entire genomes, it has the drawback of increased cost, and requires more complex methods for analyzing the generated data.

Targeted gene surveys are an alternative approach for the analysis of complex microbial communities. Here, specific phylogenetically informative loci (e.g., 16S rDNA) are amplified and sequenced. Microbial community structure may be inferred by mapping the sequencing reads from a sample to a database containing both reference sequences and taxonomic annotations for the marker gene. Although targeted gene surveys are limited to profiling the community structure of the given sample, they have the benefit of providing higher sequencing depths compared to shotgun metagenome sequencing. Therefore, they may be preferable for situations that require accurate community diversity estimates, or that aim to detect low abundance community members. Although these methods are continually being improved upon and have already led to considerable advances in our understanding of the composition and potential functions of microbial communities, the molecular and computational methods currently used are prone to artifacts known to decrease the accuracy of the inferred community membership [5], [6].

Errors and biases that are introduced during the PCR (e.g., amplification bias, chimeras), and sequencing stages are well described, but the full validation of public reference database nomenclature remains incomplete. Nomenclature (i.e., the organism name) annotations provided for reference marker gene sequences allow unknown sequencing reads to be mapped to known taxa. Therefore, misannotated reference sequences may introduce errors, such as inaccurate inferred community membership, and biodiversity. A thorough evaluation of 16S rDNA database nomenclature quality is difficult due to the limitations of the resources available for the validation of prokaryote nomenclature. Valid and invalid prokaryote nomenclature has been collected by several resources [7], [8], but the data are provided in formats convenient for the validation of individual prokaryote names, and are not suitable for the validation of entire databases. There is also a need for the quantification of the nomenclature provided by these resources in terms of taxonomic coverage and nomenclature accuracy.

With these problems in mind, a validated set of prokaryote nomenclature was collected, and used to evaluate the quality of nomenclature in public 16S ribosomal DNA (rDNA) gene databases. From this research, considerable misannotations in 16S rDNA gene reference databases were identified. As the metagenomics research community uses these reference sequence databases extensively, confidence in the quality of these resources is critical for future research. The aim of this project was to quantify and characterize the accuracy of prokaryote nomenclature in publicly available 16S rDNA gene reference databases, and to propose strategies for improving the accuracy of nomenclature in future biocuration projects.

1.1 Project Aims

1.1.1 Primary

1. To evaluate the accuracy of nomenclature annotations assigned to reference sequences in public 16S rDNA marker gene databases
2. To characterize the biological sources of the misannotated sequences
3. To collect a validated set of valid and invalid prokaryote names

1.1.2 Secondary

1. To propose strategies to improve the quality of future marker gene databases

1.2 Background and Rationale

The characterization of microbial communities and their combined genome content (frequently referred to as the microbiome) has proved to be a popular research topic in recent years. Strong evidence exists for a beneficial role of the commensal microbiota in a number of human physiological processes, such as those involving nutrient metabolism and immune function [9]. Dysbiosis (abnormal composition or function of the commensal microbiota) is associated with several disease states, such as inflammatory bowel disease, obesity, and type 2 diabetes [10]. Microbial communities also play a key role in a number of environmental processes, with biogeochemical cycling and biodegradation being important examples [11]. Despite these advances, much work remains to bring clarity to the interaction between microorganisms and the hosts or environments that they inhabit.

1.2.1 16S rDNA Gene as a Proxy for Microbial Identification

Obtaining an accurate representation of the microorganisms contained in metagenomic samples presents a considerable challenge. Microbial communities are typically highly complex, and consist of a variety of differentially abundant organisms. Our inability to culture most of these organisms presents a further complication. Metagenomic techniques that sequence marker genes have become popular, as the community may be profiled directly without intermediate culturing steps. In general, if one wants to identify the organisms present in a community, and does not require an estimate of their metabolic potential, a **gene marker survey** is performed via targeted sequencing of a phylogenetic marker gene. Sample population structure may be inferred by mapping marker gene sequences to known taxa. Accurate taxonomic classification relies on the choice of a suitable genetic marker that contains enough information content to draw inferences on evolutionary relationships among the organisms. Equally important, a suitable molecular marker should be distributed ubiquitously among the organisms of interest, and contain a degree of conservation that allows for the calculation of the “molecular clock” [12]. The gene encoding the 16S rDNA small subunit (SSU) has become the standard marker for identifying bacteria and archaea present in complex environmental samples. The gene is present in all prokaryotes [13], [14], is typically inherited vertically, and has a high degree of conservation, thus making it a suitable molecular marker [15]. Nine hypervariable regions (HVRs) in the gene contain enough differences for delineating major prokaryote groups.

The use of the 16S rDNA gene as a phylogenetic marker has some limitations that should be noted. Different mutation rates in the HVRs have been described between different groups of bacteria, thereby limiting its use as a molecular clock [16]. In fact, incongruence between phylogenies of the 16S rDNA gene and other methods of phylogenetic reconstruction have been described in the literature [17]. Because current high-throughput sequencing methods are usually limited to sequencing a subset of the 16S HVR regions, it usually is not possible to delineate closely related species. Poor resolution has also been described for certain genera [18]. Problems resulting from the limited phylogenetic resolution of the 16S gene are reflected in the bioinformatics literature, as no single sequence identity threshold has been agreed upon for methods that rely upon **clustering** algorithms to group sequencing reads into prokaryotic species or genera [19].

With these limitations in mind, the use of the 16S gene as a molecular marker still has broad applicability in metagenomics, as no other prokaryotic genetic marker has been shown to be equally comprehensive or to provide superior classification performance. When compared with other phylogenetic methods, the 16S rDNA gene can be considered to provide a good approximation of the actual evolutionary history [20]. Finally, it should be noted that the limitations mentioned above result in part from the resources available to analyze these data. For example, the 16S rDNA sequence identity shared between organisms within a taxon may be miscalculated if nomenclature errors (e.g., assignment of an invalid name to a species) result in the exclusion of taxon members in the analysis. The placement of taxa into **polyphyletic groups** could result in a lower minimal 16S rRDNA sequence identity, compared to a taxonomy limited to **monophyletic** groupings.Thus, despite these concerns, key opportunities exist for improving the 16S rDNA gene reference databases available to the research community.

1.3 Taxonomy and Systematics

Taxonomy is defined as the classification and naming of organisms. Although the terms, taxonomy and systematics are frequently treated as synonyms, the latter has a further requirement of grouping organisms according to their evolutionary history [21]. From the early days, the placement of microorganisms into existing classification systems has presented major challenges to microbiologists, and has been the topic of considerable debate [22], [23]. Many of these problems remain today, as no universally accepted standards exist for the systematic identification of novel bacterial species, or for their placement within an existing taxonomic system [24]. Even the merit of classifying bacteria into species has been questioned [21], [25]. It is clear that the existing methods for categorizing microorganisms are in part subjective, however, for pragmatic reasons, they remain in use for describing the diversity of the prokaryotic world.

1.3.1 Rules for Prokaryotic Nomenclature

The *International Code of Nomenclature of Prokaryotes* (formerly the *International Code of Nomenclature of Bacteria*) governs naming of prokaryote species [21]. The first edition, *the International Code of Nomenclature of Bacteria* was published in 1980, along with the *Approved Lists of Bacterial Names*. The latter designated existing bacterial names from the literature that would continue to be valid and would remain in use going forward. To obtain official standing in the nomenclature, new bacterial names must be published in the International Journal of Systematic and Evolutionary Microbiology (IJSEM; formerly International Journal of Systematic Bacteriology). It should be noted that these rules only apply to the naming and renaming of the lower taxa (order, family, genus, species) in prokaryotes. They govern neither the naming of the higher taxa, nor the placement of taxa into taxonomic hierarchies.

1.3.2 Sources of Information on Prokaryote Taxonomy

No single authority exists for prokaryote taxonomy, and the classification of an organism is left to the discretion of the authors. The classification framework provided in the *Bergey's Manual of Systematic Bacteriology* (BMSB) is widely used, and considered by many authors to be the best resource for categorizing bacteria and archaea [8], [21]. The *National Center for Biotechnology Information* (NCBI) taxonomy is also frequently used, but its accuracy has previously been questioned [25], [26]. In fact, the NCBI taxonomy group have stated that their taxonomy should not be considered an authoritative resource (http://www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html), and noted that "GenBank and the taxonomy group do not (and cannot) attempt to verify the taxonomic identification that is provided by the submitter, unless the sequence itself points to an egregious misidentification" [27].

1.3.3 Other Resources on Prokaryote Nomenclature and Taxonomy

The *List of Prokaryotic Names with Standing in Nomenclature* (LPSN) website (http://www.bacterio.net) was established in 1997 and continues to provide lists of valid names published in the IJSEM. Records present in the LPSN also contain taxonomic information. However, due to the lack of a single taxonomic standard, a number of sources are consulted. These include: the original publication, the *Taxonomic Outline of Bacteria and Archaea* (TOBA), the NCBI taxonomy database, the *Bergey's Manual of Systematic Bacteriology*, and the *All-Species Living Tree Project*. The LPSN author may also revise these classifications based on phylogenetic models [8]. The *Deutsche Sammlung von Mikroorganismen und Zellkulturen* (DSMZ) provides a similar resource, Prokaryotic Nomenclature Up-to-Date (https://www.dsmz.de/bacterial-diversity/prokaryotic-nomenclature-up-to-date. html). A current list of all names published in the IJSEM is available for download from their website.

Obtaining a taxonomic framework that is both accurate and up to date, while also being comprehensive in the coverage of known microorganisms, presents a major challenge. As mentioned earlier, the accuracy of the NCBI taxonomy database has been questioned. On the other hand, NCBI provides updates as the database is edited (changes are uploaded to their FTP site on an hourly basis). No other resource is updated with this frequency. For example, TOBA is still used as a taxonomic authority, but was last updated in 2007. The Bergey's taxonomy is typically considered to be the preferred taxonomy. However, past updates to the manual have only been provided over several years, and newly named or reclassified taxa may not be present. For example, the 2nd edition of the Bergey's Manual of Systematic Bacteriology consists of five volumes, published between 2001 and 2012. In 2015, the group responsible for continuing the publication of the *Bergey's Manual* released the *Bergey's Manual of Systematics of Archaea and Bacteria*, an online collection of Bergey's Manual articles. The collection currently provides access to articles from the 2nd edition of the BMSB, but will be updated with new content going forward. Although the authors state that the collection will be kept up to date, the planned frequency of updates was not mentioned.

Issues around the naming of prokaryotes that cannot be grown in pure culture should also be mentioned. To receive an official name, the bacteriological code requires that a pure culture representing a new species be deposited in two separate **type strain** collections. Organisms that cannot be cultured may be described using the provisional *Candidatus* designation. Seeing that most microorganisms are not culturable at present, a researcher interested in classifying organisms lacking type strains may be further limited in the choice of taxonomy [28].

1.4 Gene Marker Databases

Analyses of gene survey sequencing data rely upon annotated reference databases that contain marker gene sequences and their corresponding taxonomic assignments. These databases serve two main functions: to provide a means of mapping the sequencing reads present in a metagenomic sample to those of known (or similar) organisms, and to provide a multiple sequence alignment for clustering algorithms.

1.4.1 16S rDNA Gene Reference Databases

Three specialized databases that contain a broad coverage of 16S rDNA gene sequences are available for download: the Ribosomal Database Project (RDP), Greengenes, and SILVA. Although these databases provide similar functionality, several key differences exist, reflecting the goals of each database provider. The statistics and quality assurance methods for each database are summarized in Tables 1 and 2, and are briefly described below.

Table 1. 16S rDNA Gene Reference Databases

Name	Version	16S rDNA Sequences (count)	Monophyletic Groups	Taxonomy Format	Lowest Taxon
RDP	Release 11, Update 4; May 2015	3,224,600	Not enforced, used to resolve discrepancies	1º - Bergey's 2º - LPSN	Genus
Greengenes	13_5; May 2013	1,262,986	Strictly enforced; based on *de novo* tree	NCBI, and CyanoDB	Species for certain taxa[1]
SILVA (SSU Parc)	123; July 2015	4,332,731	Not enforced, used to resolve discrepancies	1º - Bergey's	2º -LPSN and manual curation
SILVA (SSU Ref)		1,636,081 (Non-redundant: 536,224)			

1 – Species level labels provided in metadata for certain records

Table 2. 16S rDNA Gene Reference Database Quality Control Methods

Name	Version	Minimum Sequence Length	Chimera Checking	Alignment
RDP	Release 11, Update 4; May 2015	500bp	Detected (UCHIME) and removed	Infernal, poorly aligned sequences discarded
Greengenes	13_5; May 2013	1250bp	Detected (UCHIME, ChimeraSlayer) and removed	Infernal, poorly aligned sequences manually adjusted
SILVA SSU Parc	123; July 2015	300bp	Flagged only (Pintail)	SINA, poorly aligned sequences discarded
SILVA SSU Ref		Archaea: 900bp; Bacteria: 1200bp		

The Ribosomal Database Project (RDP) provides a quality controlled database containing reference 16S rDNA sequences[26]. The most recent version (Release 11, Update 4; May 26 2015) contains 3,224,600 aligned sequences, of which 3,070,243 are of bacterial origin, and 154,357 are of archaeal origin. The sequences were collected from the International Nucleotide Sequence Database Collaboration (INSDC), and were primarily of environmental origin (85% for bacteria, 97% for archaea). Candidate sequences that clustered with known bacterial and archaeal sequences were required to be at least 500bp in length. PCR chimeras were screened for and removed using UCHIME [29]. The remaining sequences were aligned using the Infernal aligner [30], and poorly aligned sequences were discarded as a further quality control step. The RDP taxonomy backbone is based-upon the Bergey's taxonomy, and was constructed using type strains cross-referenced through the INSDC databases. Due to poor annotation quality, the LPSN and All-Species Living Tree Project were also consulted. Any discrepancies between these sources were resolved by retaining the taxonomic label that was best supported by a phylogenetic assessment. Due to the limited taxonomic resolution of the 16S rDNA gene, taxonomic labels are only provided down to the genus level.

The Greengenes database is another source for quality controlled 16S rDNA gene reference sequences [25]. The most recent version (13_5; May 2013) contains 1,262,986 sequences, representative of 1,242,330 bacteria and 20,656 archaea. This release was based upon a *de novo* phylogenetic tree, where the NCBI taxonomy was back-propagated onto it. Initially, **operational taxonomic units** (OTUs) were generated by clustering the 16S sequences at a 99% identity threshold, and representative sequences from each OTU were used to construct the phylogenetic tree. The Greengenes taxonomy was then back-propagated onto this tree, while enforcing a strict naming scheme to obtain only monophyletic groups. Of the 16S databases mentioned here, strict monophyly was only enforced by Greengenes. Similar to RDP, the NCBI taxonomy provided the primary taxonomic source. The CyanoDB database (http://www.cyanodb.cz/) was used as a secondary source for sequences representing the Cyanobacteria. Species level labels were assigned to nodes when possible. Sequences were aligned with the Infernal aligner, and poor alignments were adjusted manually. Quality control consisted of the exclusion of sequences below 1250bp in length, the identification and removal of mitochondrial and 18S rDNA sequences using Megablast, and chimera detection with UCHIME and ChimeraSlayer [31].

The SILVA curators provide three separate 16S reference databases [32]. The current version (123) of the Parc database contains 4,985,791 sequences in total, and is comprised of both 16S and 18S sequences. Although this database is the most comprehensive, it has limited utility due to minimal quality filtering. SILVA also provides two “reference” databases comprised of both 16S and 18S sequences. Of the 1,756,783 sequences contained in the SSURef database, 1,575,088 are of bacterial origin, and 60,993 of archaeal origin. The remaining 120,702 sequences represent either mitochondria or chloroplasts. A “non-redundant” version of this database containing OTUs representative of 513,311 bacteria and 22,913 archaea, was created by clustering the 16S rDNA gene sequences at 99% identity. For quality assurance, minimum sequence lengths of 1200bp and 900bp were required for bacteria and archaea respectively. Sequences that aligned poorly using the SINA aligner [33] were also removed. Although chimera detection was performed using the Pintail program [34], suspect sequences were not removed in order to avoid discarding unknown 16S sequences of biological origin. Users using the SILVA web interface have the option of flagging these sequences.

1.4.2 Impact of the Choice of 16S Reference Database on Sequence Classification

The presence of artifacts, such as sequencing errors and PCR chimeras, are known to bias the diversity estimates obtained from gene survey sequencing data [31]. Although the detection and removal of these errors is regularly included in bioinformatics pipelines, few quality assessments of the reference databases have been performed. One study published in 2005 suggested that at least 5% of the sequences deposited in a previous release of the RDP database (release 9, update 22) [34] contained errors. Due to the lack of recent estimates, it is unclear to what extent these artifacts are present in current versions of the public 16S reference databases. Comparisons between these databases for the classification of 16S reads have also been performed. The results obtained from these studies suggest that larger and more diverse reference databases allow for a higher proportion of reads to be classified [35], [36]. However, these studies used comprehensiveness (in terms of the highest proportion of classified reads) as the main criterion, and were incapable of measuring accuracy. Without knowing the accuracy of the taxonomic assignments, it is unclear if these results are representative of the actual communities under study. Although the creation of mock microbial communities in vitro may be infeasible for these studies, computer generated data could be used to estimate the classification accuracy.

Chapter 2: Nomenclature in 16S rDNA Gene Databases

2.1 Nomenclature Errors Present in 16S rDNA Reference Gene Databases

A preliminary evaluation of the quality of sequence annotations present in the RDP, Greengenes, and SILVA databases consisted of confirming the correct usage of the binomial nomenclature and for the screening of sequences of probable eukaryote origin. The results obtained revealed the presence of many major errors in the SILVA (v. 123), RDP (release 9, update 22) and Greengenes (v. 13_5) databases. The SILVA database contained 251 records with names that failed to meet the binomial species name requirement (e.g., the *Staphylococcus aureus* species was placed under the *Bacillus* genus). An examination of the NCBI taxonomic assignments corresponding to the Greengenes records revealed 3,659 sequences labeled as bacteria under the Greengenes taxonomy, but as eukaryotes under NCBI. BLAST searches were performed on a selection of these sequences, and the results obtained confirmed their probable eukaryote origin.

2.2 Validation of Prokaryote Nomenclature

Obtaining an accurate estimate of the nomenclatural errors in 16S marker gene databases is challenging due to the lack of computer readable data that is accurate, up to date, and comprehensive in terms of known bacteria and archaea. Although listings of newly assigned names and name changes are published in IJSEM, a complete listing of these names is not provided by this resource. Therefore, the curation of nomenclature data provided in IJSEM would have required the collection of over three decades of data, and was infeasible for this project. Although other sources of prokaryote nomenclature exist, estimates of their reliability have not been performed.

In the interest of estimating the reliability of the annotations contained in the 16S rDNA gene reference databases, valid and invalid prokaryote names were collected from three resources: LPSN, Prokaryote Nomenclature Up-To-Date, and the Integrated Taxonomic Information System (ITIS). The comprehensiveness and agreement between these results were compared, with the aim of curating a "core-set" to be used in evaluating 16S rDNA gene reference database nomenclature. Although, no gold-standard is available to assess the accuracy of the nomenclature provided, agreement between multiple resources was used as an indicator of nomenclature reliability.

Historically, bacteria belonging to the Cyanobacteria phylum were classified as algae, and assigned names according governed by the *International Code of Nomenclature for Algae, Fungi, and Plants* (ICN)[37]. To date, few bacteria belonging to the Cyanobacteria phylum have been assigned names according to the *International Code of Nomenclature of Prokaryotes*. Therefore, a separate nomenclature resource, CyanoDB, was consulted for Cyanobacteria nomenclature.

2.3 Collection of Nomenclature Data

The data provided by LPSN and CyanoDB are provided in the HTML format, and the Prokaryote Nomenclature Up-To-Date and ITIS data are provided in XLSX, and SQL formats respectively. While these formats are suitable for the lookup of individual prokaryote names, they are impractical for verifying the large quantities of names contained in the 16S rDNA gene reference databases. Therefore, custom Python and R scripts were created to collect the names provided by each resource.

2.3.1 Collection of LPSN Nomenclature Data

Prokaryote nomenclature (valid and invalid names, *Candidatus* names) collected by the LPSN author is publicly available on the LPSN website (http://www.bacterio.net/). While data listed on individual webpages are convenient for manually verifying small sets of prokaryote names, the manual collection of the entire nomenclature is infeasible due to the quantity of data. To facilitate the automated collection of LPSN nomenclature, custom Python scripts were created to collect the required data based upon the underlying webpage structure. These scripts were used to: (1) download the html source code from the LPSN website, (2) extract the nomenclature using CSS selectors, and (3) perform string processing to remove non-ascii characters. The Beautiful Soup Python module (https://www.crummy.com/software/BeautifulSoup/bs4/doc/) was used to parse the html data (accessed September 2015) using the CSS selectors shown in Table 3. These scripts are available for download from http://www.bcgsc.ca/downloads/klesack/ (see Appendix for more information).

Table 3. LPSN URLs and CSS Selectors

Category	**URLs**	**CSS Selector**
Valid species names	http://www.bacterio.net/-allnamesac.html http://www.bacterio.net/-allnamesdl.html http://www.bacterio.net/-allnamesmr.html http://www.bacterio.net/-allnamessz.html	"span.genusspecies + span.genusspecies"
Valid genus names	http://www.bacterio.net/-allnamesac.html http://www.bacterio.net/-allnamesdl.html http://www.bacterio.net/-allnamesmr.html http://www.bacterio.net/-allnamessz.html	"br + span.genusspecies"
Commonly used invalid names	http://www.bacterio.net/-nonvalid.html	".genusspecies-quote + .genusspecies"

Candidatus	http://www.bacterio.net/-candidatus.html	".candidatus-designation + .candidatus-name"

2.3.2 Collection of Prokaryote Nomenclature Up-To-Date Nomenclature Data

Lists of valid and invalid prokaryote names were obtained from an XLSX file (accessed September 2015) available from the Prokaryote Nomenclature Up-To-Date website (https://www.dsmz.de/bacterial-diversity/prokaryotic-nomenclature-up-to-date.html). The XLSX file was first converted to CSV format using the *in2csv* command from the CSVKit tool (http://csvkit.readthedocs.io/en/1.0.1/scripts/in2csv.html). Annotations contained in the STATUS field were used to collect valid (VP or VL) and invalid prokaryote names (illegitimate name, rejected name, nom. illeg., nom. rej., orthographically incorrect name).

2.3.3 Collection of ITIS Nomenclature Data

A SQLite formatted copy of the ITIS database (accessed September 2015) was downloaded from the ITIS website (https://www.itis.gov/). Data contained in the *taxonomic_units* table were imported into R using the RSQLite library (https://cran.r-project.org/web/packages/RSQLite) and converted into a dplayr table data frame using the dplyr library (https://cran.r-project.org/web/packages/dplyr). Prokaryote records were extracted based on the kingdom_id field (kingdom_id=1 for bacteria; kingdom_id=1 for archaea), and names were categorized according to the *name_usage* field (name_usage=valid for valid names; name_usage =invalid for invalid names).

2.3.4 Collection of CyanoDB Nomenclature Data

Cyanobacteria nomenclature is available on the CyanoDB website. To automate the collection of CyanoDB nomenclature, a custom Python script were created to collect the required data based upon the underlying webpage structure. The script was used to: (1) download the html source code from the CyanoDB website, (2) extract the nomenclature using CSS selectors, and (3) perform string processing to remove non-ascii characters. The Beautiful Soup Python module (https://www.crummy.com/software/BeautifulSoup/bs4/doc/) was used to parse the html data using the CSS selectors shown in Table 4. These scripts are available for download from http://www.bcgsc.ca/downloads/klesack/ (see Appendix for more information).

Table 4. CyanoDB URLs and CSS Selectors

Category	URL	CSS Selector
Valid genus names	http://www.cyanodb.cz/valid_genera	"p a"
Valid species names	http://www.cyanodb.cz/valid_genera	"em , span"
Invalid names	http://www.cyanodb.cz/excludenda	"p"
Synonymous names	http://www.cyanodb.cz/synonyms	"p"

2.4 Validation of Collected Nomenclature

Lists of invalid and valid names were successfully collected from each resource, and were compared at the genus and species levels (Table 5). LPSN and Up-To-Date both contained similar numbers of valid names at both levels, while ITIS was less comprehensive at the species level. A list of 269 valid cyanobacteria genera and their representative type species were collected from CyanoDB. A considerable number of invalid names were collected from ITIS

(118 genera; 2070 species), while fewer invalid names were collected from LPSN (143 genera; 722 species), Up-To-Date (31 genera; 194 species) and CyanoDB (160 genera). These results are summarized in Figure 1 and Figure 2.

Table 5. Collected Invalid and Valid Prokaryote Names

Resource	Valid Names		Invalid Names	
	Genus	Species	Genus	Species
LPSN	2,361	13,711	144	722
Up-To-Date	2,521	13,599	29	201
ITIS	2,111	10,354	118	2,070
CyanoDB	269	269	144	0

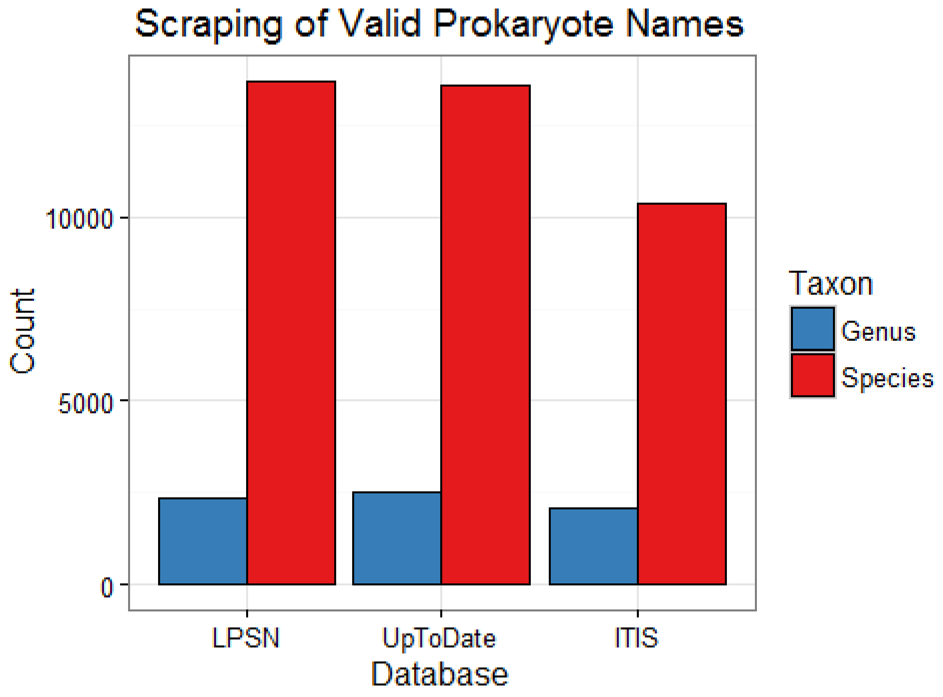

Figure 1. Valid Prokaryote Names Contained in Nomenclature Databases

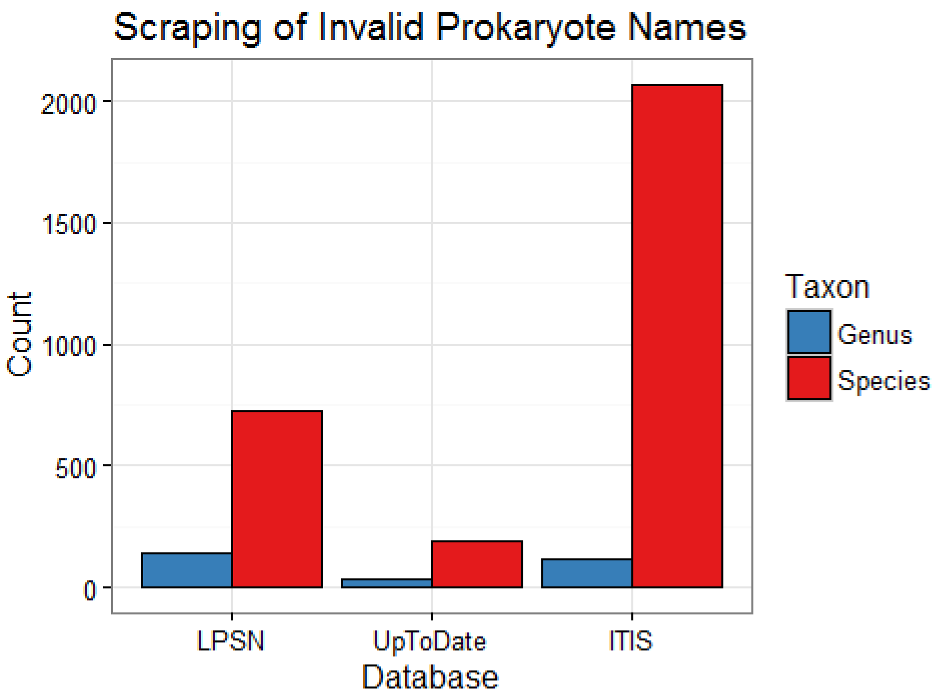

Figure 2. Invalid Prokaryote Names Contained in Nomenclature Databases

2.4.1 Comparison of Collected Nomenclature for LPSN, ITIS, and Up-To-Date

The pairwise **intersection** of valid genus and species (Table 6) names were used as the measurement of agreement between the separate nomenclature databases (excluding CyanoDB, as it was the sole provider of Cyanobacteria nomenclature). At both taxonomic levels, the highest agreement occurred between LPSN and Prokaryote Nomenclature Up-To-Date (genus = 0.99; species = 0.98).

Table 6. Valid Genus/Species Names Agreement

Database	ITIS	LPSN	Up-To-Date
ITIS		0.79 / 0.77	0.74 / 0.74
LPSN	0.79 / 0.77		0.99 / 0.98
Up-To-Date	0.74 / 0.74	0.99 / 0.98	

The agreement is represented by the proportion of the intersection of valid genus / species names between each database pair.

2.4.2 Final Nomenclature "Core-Set"

The final list of valid names was based upon the set intersection of genus and species names provided by LPSN, Prokaryote Nomenclature Up-To-Date, and CyanoDB. The nomenclature contained in ITIS was not used, as lower agreement occurred between ITIS and the other resources. By definition, there is no official list of invalid prokaryote names, and the invalid names provided by the LPSN and Prokaryote Nomenclature Up-To-Date authors cannot be expected to represent all invalid names that have been used in the literature. Nonetheless, the combined list of known invalid names from LPSN and Prokaryote Nomenclature Up-To-Date were collected for the purposes of providing a conservative estimate of incorrect nomenclature in public 16S rDNA databases. A list of synonyms was collected from the 2004 edition of Bergey's manual, Prokaryote Nomenclature Up-To-Date, and CyanoDB. A set of proposed names that remain to be validly published (candidatus) were collected from LPSN. The final "core-set" is summarized in Table 7.

Table 7. Validated Nomenclature “Core-Set”

Category	Source	Count
Valid	UpToDate/LPSN/CyanoDB	15,223
Invalid	UpToDate/LPSN/CyanoDB	1,238
Synonymous	UpToDate/Bergey’s/CyanoDB	1,564
Candidatus	LPSN	956

2.4.3 Valid and Invalid Names in 16S rDNA Gene Reference Databases

Each 16S rDNA database provides a delimiter-separated file containing the taxonomic rank assigned to each reference sequence, and the genus and species names (the “core-set” was limited to these ranks) were obtained using the Linux *cut* command. The Greengenes names contained prefixes designating the taxonomic rank (e.g., g__ for genus; s__ for species), which were also removed using the *cut* command. The nomenclature “core-set” was then used to categorize the extracted names using the Linux *grep* command (Table 8). Only exact matches (ignoring case) were used to classify names as either valid or invalid, and the remaining names were left uncategorized. The percentage of valid names ranged from 76.7% (SILVA) to 83.8% (Greengenes) and 90.0% (RDP). From 3.7% (RDP) to 5.1% (SILVA) and 5.5% (Greengenes) were classified as invalid. This is likely a conservative estimate, as many sequences were not classified as either valid or invalid (RDP = 4.8%, Greengenes = 7.0%, SILVA = 16.8%).

Table 8. 16S Database Nomenclature Validation

Database	Total Sequences	Valid Names	Invalid Names	Uncategorized
Greengenes	1,262,986	83.8%	5.5%	7.0%
RDP	1,636,081	90.0%	3.7%	4.8%
SILVA (NR)	3,224,600	76.7%	5.1%	16.8%

2.5 Isolation Source of Misannotated Sequences

A rigorous description of the environmental origins of metagenomic samples is non-trivial. Although an “isolation_source” field is available for submissions to Genbank, the environmental origin is frequently omitted for many sample submissions or may be indiscernible due to ambiguous or imprecise terminology [38]. The need for the consistent representation of the terms describing the environmental origins of metagenomic samples, as well as their attributes and relationships led to the development of the Environment Ontology (ENVO) [39]. The ontology hierarchy contains classes primarily relevant to metagenomics, and includes terms representing biomes, environmental features, and ecological processes. Because multiple different terms can be used to describe single entities, the ontology classes are cross-referenced to synonymous terms. With these capacities in mind, ENVO was used in the curation of key terms that describe typical metagenomics sampling environments. To assess the metagenomics projects most likely to be impacted by misannotated sequences, the collected terms were used to categorize the sampling origins of misannotations in SILVA. The sample sources of misannotated sequences in the RDP and Greengenes databases were not characterized due to the lack of necessary metadata.

2.5.1 Sampling Environment Metadata Collection and Categorization

A database containing sequence alignments, phylogenic analyses, and sequence metadata for the SILVA reference gene sequences is provided by the ARB project [40]. To describe the environment that misannotated sequences were isolated from, metadata contained in the "isolation_source" field were extracted from the ARB database (version 128). Of the 98,820 misannotations identified in SILVA, an ARB record containing metadata was available for 85,269 16S sequences.

In order to categorize the misannotated samples by sample isolation source, terms describing metagenomics environment types were extracted from the ENVO ontology (release date 2017-01-08) using the Protégé (version 5.1.0) [41] ontology editor. The ontology is hierarchical in structure, which facilitated grouping the terms into several broad categories, each containing more specific sub-categories. An initial assessment of the categorization of the ARB “isolation_source” metadata based upon these category terms resulted in numerous unclassified records. Therefore, further manual refinement (e.g., addition of categories, recognition of synonymous terms) was necessary. Ultimately, six broad categories were created, which were further grouped into 16 subcategories (Table 9).

Table 9. Description of Sample Isolation Source Categories

Category		**Description**	**Examples**
Host-associated			
	Animal	Samples isolated from a host animal. Includes samples isolated from anatomical entities and substances produced by an excretory or secretory process.	▪ respiratory secretion ▪ intestinal mucosal biopsy from crohn's patient
	Plant	Samples isolated from a host plant. Includes samples isolated from anatomical entities, and plant surfaces. Excludes samples isolated from the plant **rhizoplane** or **rhizosphere** soil.	▪ sugarcane leaves ▪ phyllosphere
	Fungus	Samples isolated from a host fungus.	▪ fungal ascocarp
Environmental material (Non-aqueous)			
	Soil	Samples isolated from material explicitly described as soil. Excludes soil isolated from rhizosphere.	▪ soil sample ▪ soil crust
	Inorganic material	Samples isolated from non-organic environmental material. Commonly used for environmental material that did not fit any other category.	▪ beach sand ▪ sediment
	Rhizosphere	Samples isolated from rhizosphere soil.	▪ maize rhizosphere soil
	Organic material	Samples isolated from environmental material derived from living organisms. Contains predominantly biofilm samples.	▪ biofilm surrounding a gastropod
	Biome	Samples isolated from a biome or habitat. Here, a biome is defined as an ecological zone to which resident species have evolved adaptations.	▪ forest ▪ alpine grassland
Built environment			
	Waste	Samples isolated from anthropogenic waste material. Includes, wastewater, household waste, and sewage.	▪ municipal wastewater plant ▪ landfill leachate
	Industrial	Samples isolated from material produced by an industrial process. Includes both industrial goods and waste products.	▪ oil sludge ▪ quinoline-degrading bioreactor

Category		**Description**	**Examples**
	Building	Samples isolated from human-made buildings and structures. Includes predominantly residential and hospital environments.	▪ hospital environment ▪ house dust
	Human settlement	Samples isolated from human settlements. Excludes samples that can be categorized under building.	▪ urban aerosol
Aquatic			
	Marine	Samples isolated from the ocean and other saline aquatic environments.	▪ mediterranean sea ▪ estuarine sediment
	Freshwater	Samples isolated from freshwater environments such as rivers and lakes.	▪ lake water ▪ river water
	Polar	Samples isolated from frozen material or from polar environments.	▪ collins glacier ▪ ice core
	Other	Samples that elude classification in other aquatic categories. Frequently used when the sample environmental salinity is unclear.	▪ wetland water ▪ cold seep sediment
Food and agriculture		Samples isolated from food products, or from land used for agricultural production.	▪ fermented dairy products ▪ agriculture soil
Gaseous		Samples isolated from gases or aerosols.	▪ biogas z8aerogenic ▪ biogas pond sample

2.5.2 Sampling Environments of Misannotated 16S Sequences in SILVA

The placement of misannotated 16S sequences from SILVA into categories describing sample isolation environments is shown in Table 10. Of the 85,269 misannotated sequences with corresponding ARB isolation source metadata, 80,127 (94%) were successfully classified into one of the six broad categories. Further classification into a subcategory was possible for 78,780 sequences. Of the 1,347 sequences that were not placed into a subcategory, only 83 were placed under a parent category containing subcategories (subcategories were not created for Food and agriculture, and Gaseous). The metadata for these sequences described aquatic sampling environments, allowing for placement into the broad aquatic category, however, it was not clear whether these were isolated from marine or freshwater sources.

Among the sequences assigned to a broad category, 85% (67,819 out of 80,127)) were isolated from a host-associated environment, with animals (n = 66,710) being the predominant host. Although the host species was not indicated for most samples (only 747 sequences included the terms "human" or "homo sapiens"), this trend likely results from an early focus on studying the human microbiome. Samples isolated from environmental material (6.5%) and built environments (4.5%) accounted for most of the remaining misannotated sequences. This also likely reflects substantial sampling of these environments.

Table 10. Summary of SILVA Misannotations by Sample Isolation Source Category

Category		**Number of Sequences**
Host-associated		
	Animal	66,710
	Plant	986
	Fungus	123
Host-associated total		**67,819**
Environmental material (Non-aqueous)		
	Soil	3,318
	Inorganic material	703
	Rhizosphere	659
	Organic material	283
	Biome	264
Environmental material total		**5,227**
Built environment		
	Waste	2,370
	Industrial	913
	Building	258
	Human settlement	26
Built environment total		**3,567**
Aquatic		
	Marine	1,518
	Freshwater	525
	Polar	124
	Other	83
Aquatic total		**2,250**
Food and agriculture		**1,152**
Gaseous		**112**
Unclassified		5,142
Total classified		**80,127**

Chapter 3: Strategies to Improve the Accuracy of Marker Gene Databases

The considerable misannotations identified in public 16S ribosomal RNA gene reference databases demonstrate a clear need for improving the quality nomenclature annotations. That is, the sequences contained in these databases should be free of technical artifacts and contaminants, while the associated annotations need to be up to date and valid according to the nomenclature. While providing a completely error-free database may not be a realistic goal, our evaluation of the existing 16S databases suggests that considerable room for improvement exists. In this section strategies aimed at improving the quality of microbial gene marker reference databases have been proposed. Although these strategies were designed with the validation of a 16S rDNA gene database in mind, they are broadly applicable, and would be suitable for other microbial gene marker databases (e.g., fungal ITS, eukaryote 18S rDNA).

3.1 Marker Gene Sequence Quality Assurance

High-throughput DNA sequencing technologies employ complex series of molecular and computation methods, each capable of introducing errors that may impact downstream analyses. Several studies have identified errors common to different sequencing platforms and library preparation methods, and have been recently reviewed in the literature [42]–[44]. Errors can arise during any point in the sequencing workflow, such as sample preparation, library preparation, sequencing, and bioinformatics analysis. Therefore, accounting for errors arising from the DNA sequencing process should be an important consideration for future curators of microbial reference gene databases. Fortunately, these errors can be in part mitigated by several bioinformatics methods. The data collection, cleaning and quality control steps described below are limited to those applicable to candidate gene sequences for a typical reference gene database.

3.1.1 Collection of Marker Gene Sequences

Candidate marker gene sequences and their associated annotations can be easily acquired from the public INSDC nucleotide sequence databases. However, the results obtained would be limited to sequences containing annotations for the marker gene of interest. Unannotated metagenomic data containing sequences from uncultured organisms may be used to obtain a more comprehensive set of candidate sequences. For the prediction of 16S rDNA genes several tools are available, including rRNASelector, REAGO, and RNAmmer [45]–[47].

3.1.2 Sequence Filtering and Trimming

DNA sequencing data is inherently "noisy", and may contain PCR and sequencing errors. It is therefore necessary to subject the candidate sequences to quality assurance methods aimed at trimming low quality regions, the removal of adapter or primer sequences, and filtering of reads that contain **PCR chimeras** or fail to meet minimum length requirements. PCR chimeras [48], [49] are especially problematic, and are known to inflate biodiversity estimates [50]. Failing to remove chimeras may also lead to the classification of spurious taxa that do not correspond to actual novel lineages. Several methods exist for chimera detection, including tools designed for specific marker genes (e.g., Pintail [34], ChimeraSlayer [31] and Mallard [51] for 16S rDNA; ChimeraChecker [52] for ITS rDNA), or for general use (Bellerophone [53], UCHIME [29], DECIPHER [54]). Most methods (e.g., Pintail, ChimeraChecker, Mallard, DECIPHER) screen putative chimeras against databases considered to be of high-quality and free of chimeras. However, few chimera-free 16S reference sequence databases are available, and algorithm performance using a given database can vary for different input data [55]. Most reference-based chimera detection approaches have used a database provided by Haas and colleagues [31]. The dataset is expected to be mostly chimera-free [31], and contains 5,181 16S sequences obtained from both type strains (4,468) and complete or draft genome assemblies (713). However, because the sequences are limited to well-characterized organisms, the database is best suited for data sampled from well-studied environments [55]. For samples containing under-represented (e.g., archaea) and uncultured groups, larger chimera-screened reference datasets, such as RDP or SILVA, are recommended [56]. However, despite screening, chimeric sequences are still identified in these databases [54]. To overcome this difficulty, UCHIME, Perseus [57], and ChimeraSlayer provide *de novo* chimera detection modes that do not require a reference database.

The performance of a given algorithm depends largely on the characteristics of the data being analyzed. For example, the results published for the ChimeraSlayer algorithm demonstrated superior performance compared to the Bellerephon and Pintail algorithms for the detection of shorter chimeric sequences [31]. ChimeraSlayer also outperformed Bellerephon and Pintail for chimeras resulting from closely related parental sequences (less than 10% divergence), but performed poorly compared to UCHIME2 for sequences with higher parental divergence [56]. Other factors that may contribute to performance include the proportion of the chimeric sequence contributed by each parent sequence [54], ratio of chimeric and non-chimeric reads in the dataset [55], taxonomic representation [31], and the presence of complex (greater than two parents) chimeras [31], [54]. A comparison of the relative strengths and weaknesses of these tools is therefore difficult, as the published performance results may be tailored to specific data, and a given algorithm may perform poorly on other data.

Ensemble methods are a popular approach for tackling problems considered too complex for individually trained classifiers [58]. By combining the predictions of multiple separate classifiers, ensembles frequently outperform individually trained classifiers. For an ensemble to perform well, the individual classifiers must be both diverse (disagreement among the classifier predictions), and accurate [59]. If both criteria are achieved, improved performance may be achieved by exploiting the strengths of the individual classifiers that perform best on different subsets of the input data [60]. A comparison between several reference-based (UCHIME, ChimeraSlayer, Pintail, and DECIPHER) and *de novo* (UCHIME, ChimeraSlayer, Perseus) chimera detection tools was recently performed, which formed the basis for the ensemble classifier CATCh [55]. A variety of mock datasets (chosen to include a range of chimera types) were used to benchmark the chimera detection performance provided by the CATCh ensemble compared to the individual classifiers. The results obtained demonstrated improved performance for both reference-based and *de novo* chimera detection. Despite these improvements, it should also be noted that bioinformatics methods alone are unlikely to result in complete chimera detection and removal. Most chimeras result from highly similar parent sequences, and many can be indistinguishable from naturally occurring sequences [48].

Contaminants resulting from off-target amplification may also decrease the quality of a reference sequence database. The high degree of homology between the prokaryote 16S rDNA gene with the chloroplast and mitochondrial 16S rDNA genes is a well described source of contamination in microbial gene survey studies [61]–[63]. Therefore, *in silico* contaminant screening is advisable. Currently available tools include UPARSE [64], DeconSeq [65], and BioBloom Tools [66].

Quality trimming and filtering entails the trimming of poor quality nucleotides from either end of a sequencing read, as well as the subtraction of entire reads failing to meet minimum quality thresholds. The importance of ensuring the quality of the initial DNA sequencing data has been repeatedly addressed by the bioinformatics community, leading to the development of numerous tools that serve this purpose. Due to the large amount of available tools, the reader will be referred to the literature for a reference on the choice of suitable methods [50], [67]–[70].

3.1.3 Sequence Alignment

The alignment of reference sequences provides an additional quality control step, as it can be used to ensure that all sequences are within the predicted gene boundaries [67], [71]. Sequence alignments are also required by distance based bioinformatics methods that are employed in most gene marker surveys. By identifying the similarities and differences between the sample and reference sequences, sample biodiversity and taxonomic classifications can be inferred. The alignment of marker genes containing regions with high degrees of divergence (e.g., HVRs in SSU rDNA genes) presents a considerable bioinformatics challenge. Current algorithms are computationally intensive, and do not necessarily recognize positional homology in variable regions containing insertions or deletions [67], [71], [72]. Therefore, the provision of a trustworthy pre-existing alignment of reference database sequences would provide an additional benefit to the user.

3.1.4 Alignment Approaches for the 16S rDNA Gene

Several strategies have been developed with the goal of overcoming the difficulty of accurately aligning variable regions. These include the incorporation of secondary structure information (e.g., Infernal, SINA) [30], [33], and the usage of filtering or weighting masks [40] to reduce the impact of problematic regions. Many algorithms depend upon a static reference multiple sequence alignment (e.g., NAST, SINA). This approach has the benefit of reduced computational requirements, and decreased susceptibility to alignment errors that result from poor-quality sequences. Choice of a suitable alignment strategy remains challenging as comprehensive performance comparisons are unavailable, and the value of incorporating secondary structure is debatable [73], [74].

3.1.5 Molecular and Computational Methods for Benchmarking

The curation of a new reference marker gene database would rely upon the retrieval and integration of data from multiple sources. Although multiple steps aimed at ensuring the inclusion of only high-quality data in the final database have been described, each stage of data collection and processing must be considered carefully. Errors introduced by sequencing technologies are well-known [6], [75], [76], and a selection of bioinformatics methods aimed at the detection and removal of these problems have been described [42], [69]. However, the choice of suitable tools is not without complications. Tool choice relies on several factors, including data type, the statistical analysis to be performed, and scalability. Software licensing and computational requirements (e.g., operating system, storage, memory, and CPU resources) may also be a limiting factor. Poor usability [77], [78], and software availability [79]–[81] or licensing may further limit the choice of software.

A major obstacle is the general shortage of comprehensive and independent benchmarks that enable end-users to identify the relative strengths and weaknesses of many bioinformatics tools. Most end-users must rely upon the results published by the developer of a new method to justify tool choice. However, two major problems have been described for bioinformatics research that lead to artificially inflated performance measurements [82]. The first problem is that the new algorithms and parameter settings are frequently optimized against the datasets used for benchmarking, leading to model over-fitting. A given method may in fact perform markedly different on other data. Secondly, specific datasets may be selected because they yield better results for the new method compared to existing algorithms. Thus, the possible unreliability of published performance metrics should be considered when choosing bioinformatics tools.

In the situation that independent benchmarks are lacking, the analyst may choose to compare different tools and methods themselves. One of the challenges for the benchmarking of quality control methods used in the curation of a gene marker database is the limited quality assessments in existing datasets. This limitation could be addressed in part by using sequencing data generated from mock communities composed of known organisms. Sequencing data from 11 separate mock communities (composed of known quantities of “spiked-in” bacteria) have been identified in the literature [83], [84], and are available to the research community. Because these datasets were generated using current sequencing technology, they have the advantage of providing performance on data similar to that obtained from a typical marker gene survey. Once again, being limited to specific test datasets may result in performance estimates adapted to specific data that do not necessarily reflect performance on other datasets. For this reason, the use of computer generated test data should also be considered. With the *in silico* generation of sequencing data, a diverse dataset of mock communities may be generated that would otherwise be infeasible using molecular methods.

3.2 Marker Gene Annotation Quality Assurance

Nomenclature validation strategies will depend largely on the type of error, the information available to the database curators, as well as requirements of end-users. Strategies for minimizing the most common types of error are discussed below.

3.2.1 Sequences Containing Synonymous or Invalid Annotations

Out of date names may be updated with relative ease if a comprehensive set of synonyms are available for the list of valid names. Incorrectly annotated sequences where a known valid name is either unavailable or unknown present a greater challenge. For invalid names that result from the misspelling of a valid name, it may be possible to obtain the correct name using a spell-checking tool. Classic spell checking tools typically employ two main similarity measures for the detection and resolution of spelling errors: the minimum edit distance, and phonetic similarity [85]–[87]. For the minimum edit distance approach, similarity is calculated as the minimum number of editing operations (i.e., character substitutions, deletions, or additions) required to transform a misspelled string to a word contained in a dictionary. Phonetic approaches employ similarity measures that are intended to identify spelling errors resulting from the substitution of equivalent or similar sounding syllables (e.g., abiss rather than abyss). Contemporary tools frequently employ a combination of both approaches to provide comprehensive spell checking. The inclusion of contextual information may also increase algorithm performance. For the resolution of organism names, name resolution can be complemented by verifying that names are in accordance with the nomenclature rules, such as the correct usage of Latin suffixes, binomial nomenclature, and Latin or Greek substantives. (Table 111)[88]. The Taxamatch [89] and Taxonomic Name Resolution Service [90] (a modified version of the Taxamatch algorithm) are two recently developed tools, both employing a combination of the phonetic, minimum edit distance, and contextual approaches.

Table 11. Nomenclature Naming Rules

Rank	**Suffix**	**Other Rules**	**Example**
Order	*-ales*	Substantive of Latin or Greek origin, feminine gender, the plural number, and written with an initial capital letter	Pseudomonadales
Family	*-aceae*		Pseudomonadaceae
Genus	Varies	Latin or Greek substantive, in the singular number and written with an initial capital letter	*Pseudomonas*
Species	Varies	Binary combination consisting of the name of the genus followed by a single specific epithet	*Pseudomonas aeruginosa*

Adapted from International Code of Nomenclature of Bacteria: Bacteriological Code, 1990 Revision. Washington (DC): ASM Press; 1992. Available from: https://www.ncbi.nlm.nih.gov/books/NBK8817/, Accessed 21 November 2016.

Finally, invalid names that cannot be resolved should also be considered. One approach would be to exclude these sequences altogether from the database. For sequences belonging to groups that are well represented by other sequences in the database, the curator may decide that excluding these sequences would not be a large detriment to the final database. Conversely, if comprehensiveness is desired, the curator may choose to label these sequences as having an “unknown” classification. However, a reference database containing a sizeable proportion unclassified sequences may not meet the requirements of end-users requiring the assignment of unclassified reads to known taxa.

3.2.2 Unclassified Sequences

Most sequences contained in the public 16S rDNA gene databases are from “unclassified” environmental samples obtained in previous gene marker survey experiments. While the inclusion of these sequences may be desired for comprehensiveness, they have the potential of overwhelming methods aimed at providing *de novo* taxonomic affiliations for unknown reads. For this reason, the RDP, Greengenes, and SILVA databases infer the taxonomic classification of unclassified sequences based on similarity to other known sequences. However, this approach should be applied with caution, as taxonomic classification would be inherently biased towards previously characterized organisms. Recently diverged species, containing nearly identical 16S sequences further compound this problem [91]. Given that the "uncertainties" of this approach have not been well quantified, a clear need exists for the research community to investigate the reliability of ecological inferences based upon this approach (e.g., diversity, community membership).

3.2.3 Valid Names Assigned to the Incorrect Sequence

Misannotations resulting from the assignment of a valid name to the incorrect sequence is also worth consideration. This type of error could result from several causes, including incorrect annotation during sequence submission, contamination, or the misidentification of a strain in the laboratory or culture collection. While the scale of the problem remains uncertain, a recent examination of the type material in the NCBI Taxonomy database revealed lower than expected average nucleotide identity between **co-identical** type-strains [92].

While it is unlikely that the correct taxonomic annotation may be assigned with complete certainty for all sequences, the internal consistency of a reference database may be evaluated at least in part. For example, putative misclassified sequences may be flagged by identifying the inconsistent phylogenetic placement of sequences belonging to a specific strain or species [92]. Following the identification of putative mislabeled sequences, the analyst is faced with a greater problem: are the taxonomic labels assigned to the sequences in fact incorrect, or do the inconsistencies result from actual biological phenomena? Several situations may result in a taxonomic label being mistakenly identified as incorrect. For example, inconsistent phylogenetic placement may result from internal variations of the marker gene rather than misannotations. Intragenomic sequence heterogeneity has been reported for bacteria containing multiple copies of the 16S rDNA gene [93]–[98]. In most cases, the divergent sequences are nearly identical (< 1%, typically 1 or 2 polymorphisms), and would not likely have a meaningful impact on phylogenetic placement. It has been hypothesized that gene conversion plays a role in maintaining the high degree of sequence homogeneity in this gene [99]. Nonetheless, higher degrees of intragenomic heterogeneity (up to 8.3% nucleotide differences) have been described, indicating that exceptions to this trend occur occasionally [95], [97], [98].

Taxonomy can be described as a dynamic science, subject to frequent revisions, proposals, and conflicting opinions. The taxonomy of bacteria has been particularly susceptible to restructuring, as the introduction of newer molecular methods allowed taxonomists to classify bacteria according to genetic relatedness, rather than morphological and physiological properties. Adding to the problem, no single agreed upon species definition for bacteria exists, leading a variety of proposed methods for the recognition of bacterial species [100]. The analyst should therefore be aware of these challenges when screening a collection of sequences for taxonomic misclassifications. For instance, the analyst may choose to examine the taxonomic labels assigned to clusters of highly similar sequences. While the assignment of multiple labels may identify misclassified sequences, they may also result from polyphyletic groups awaiting reclassification. The genera *Escherichia* and *Shigella* are well described examples of polyphyletic groups awaiting revision [101]. The genera *Clostridia* and *Bacilli* are another noteworthy example [102].

3.3 Database Implementation

Multiple factors are important for the creation of a new biological database. The curator must consider the end-user requirements, and the technologies available to deliver the data in a responsible manner. Attention paid to the initial design requirements can result in considerable decreases of both duplicated work and debugging time. As the problem of inaccurate annotations in biological databases becomes further appreciated by the research community, databases curated with the initial data quality in mind will likely be a key consideration for the choice of a database. While the specific details will depend largely on the target audience, available resources, and curator goals, several design considerations are described below. These strategies were developed with the curation of a marker gene database in mind, with a clear emphasis on data quality assurance and usability.

3.3.1 Database User Interfaces

Microbial gene marker databases cater to two main groups of users whose requirements depend on their degree of programming experience. Users with limited programming knowledge, such as biologists or clinical researchers, would typically access the data through a simple web interface. For the most part, these users would access data using either search or browsing interfaces. Technologies such as HTML5, and CSS facilitate the design of responsive websites suitable for multiple types of browsers and viewing environments (e.g., desktop, tablet). JavaScript data visualization libraries, such as D3.js and BioJS can be used to create dynamic and interaction data visualizations for the presentation of complex data (e.g., phylogenetic trees).

While simple web interfaces allow data to be accessed quickly, they are not suitable for more technical orientated users that require access to large quantities of data. Access to a Representational State Transfer (RESTful) API would be more suitable, as the user could access the required data using a standard set of operations. Furthermore, data returned from an API request is typically returned in a structured format (e.g., XML, JSON) supported by most modern programming languages, thereby decreasing the work required to use the data in a bioinformatics pipeline. Text-based data formats developed specifically for biological data are used extensively in bioinformatics. Existing marker gene databases typically provide the sequence and alignment data in the FASTA format.

3.3.2 Web Frameworks

A web framework is a collection of pre-made software packages designed to assist in web development by reducing time spent on low-level details. In general, a developer may create an entire website incrementally, by adding the required packages one after another. A number of web frameworks are currently available for the development of database-driven websites, including the Python framework Django (https://www.djangoproject.com/) and Shiny for R (https://shiny.rstudio.com/).

3.3.3 Database Technology

DNA sequences and their associated metadata are naturally suited for storage in a relational database. Relational databases have several benefits for the storage of tabular data, including complex querying, efficient storage (data in tables is usually stored once), and scalability. A key feature of relational databases is the elimination of data duplication: modifications made to a record are shown in other tables with links to that entry. In addition to improving storage efficiency, this can help ensure data consistency. Therefore, avoiding data duplication should be a major focus of curators of future reference gene databases. The consistency of data subject to revision, such as organism names, could then be ensured between different database relations. The schema diagram shown in Figure 3 provides an example database configuration suitable for a marker gene database.

Relational database management systems (RDMS) are applications that manage the interaction between a user and the database. Nearly all RDMSs in use are based upon the Structure Query Language (SQL), which allows the user to perform tasks such as the retrieval, addition, or deletion of data. Multiple RDMSs are currently available, including both commercial (e.g., Oracle, Microsoft SQL Server) and open-source (MySQL, SQLite) implementations.

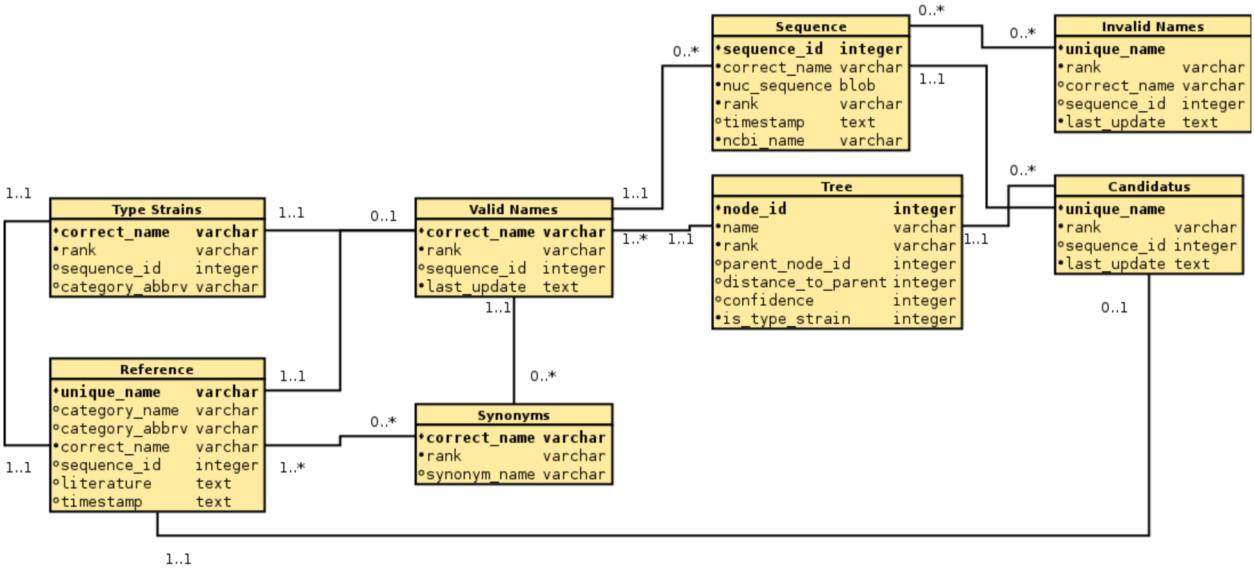

Figure 3. Example Database Schema. Primary keys are displayed in bold. The tables and relationships were designed to assist quality control and efficient data retrieval for a hypothetical marker gene database.

3.4 Community Assisted Curation

The importance of quality control measures during the initial data collection, and database design phases have been repeatedly noted [103]–[107]. However, the maintenance of an existing database involves further responsibilities, such as continual content and feature updates. For the most part, existing biological databases have relied upon small teams of specialists for curation. However, traditional curation approaches are increasingly unable to cope with the quantity of biological data generated by the research community [108]. This has led to proposals calling for the adoption of community assisted curation tools, such as wikis [108]–[113], or database content management systems [114].

Quality assurance was a key focus during the curation of the RDP, SILVA, and Greengenes databases. Common control methods (described in 1.4.1) included chimera-detection, and filtering based on sequence or alignment quality. The collected reference sequences made available by these databases clearly provide a valuable resource to the research community, and reduced the need for manual validation by the end users. However, the results presented here indicate that considerable nomenclature errors remain. Therefore, further quality assurance measures are necessary to ensure data quality. A strategy for the collection of nomenclature metadata and the identification of misannotations in 16S rDNA reference databases was described earlier, and could be adopted by the database curators. However, despite being largely automated, the adoption of this strategy may prove difficult for the small teams that curate these databases. The adoption of community assisted curation could be used to address this challenge by reducing the burden placed upon specialist curators. Several challenges and potential solutions are described as follows.

3.4.1 Barriers to Community Assisted Curation

Despite the clear need of improved curation efforts, community involvement remains low [115]. The lack of recognition for authorship is typically cited as the main obstacle for community participation [108]. Poor usability, requisite expertise, and uncertain annotation quality have also been suggested as other barriers to involvement. That aside, a number of solutions to these problems have been explored, leading to increased user participation in the curation of several biological databases [112], [113]. For example, editing privileges may be limited to registered users, in order ensure that changes are only performed by expert users. By tracking and displaying which author made changes to a given record, the perceived lack of credit for work performed could be addressed to some degree. Potential usability problems could also be identified through user testing on common curation tasks, such as taxonomic revisions. Online training (e.g., webinars, tutorials) could also enable users to use the system efficiently and effectively.

3.4.2 Quality Assurance During Data Entry

Although quality control measures should be performed throughout the various phases of database curation, special attention should be paid to quality assurance during the data upload stage. The automated collection of data is usually preferable, as the quantity of data may be infeasible for manual entry, and repetitive tasks can lead to human error. However, automated data entry may not be feasible for certain data, which would require manual entry. To ensure that data are entered consistently between different users, it is advisable to follow a standard operating procedure that provides a precise set of step-by-step instructions for manual data entry [116]. Many CMS tools and databases provide controls that may be used to enforce the type and formatting of data entered manually. Common controls include standardized data entry forms, validation rules, and controlled vocabularies. Similarly, field controls, such as input masks (a specific pattern that data contained in a field must adhere to; e.g., a date format), allowable values, and mandatory fields (i.e., Null or blank values not allowed). Validation between data contained in different fields or tables may also be used to ensure data consistency. For example, the genus epithet of a species name should match the genus name. Another example could be to check that the publication date of a taxonomic revision is later than any earlier references to that taxon in the database.

3.4.3 Data Provenance and Version Control

Data provenance, the process of tracking the origin and changes of data, is another key part of maintaining a high-quality database. By tracking when and where data came from, and any subsequent changes, database quality and consistency can be evaluated over time. Although data provenance is used mainly for quality assurance, it can also provide a basis for incentivizing community participation in curation by tracking and displaying user contributions to the database.

The subject of database release management highlights the importance of attention to data provenance. Maintaining a high-quality biological database requires considerable work to continually meet community needs: the information contained in the database must be kept up-to-date, and comprehensive. Furthermore, each database update made available to the public has the potential to introduce problems or downtime. A common approach to minimizing problems associated with deploying database changes is to employ different release environments. Numerous deployment strategies are available, but at a minimum would include separate development and production environments. Rather than introducing prospective changes directly to the public database (production version), a development version could be used for testing and evaluation. Therefore, any problems introduced by the changes may be identified and corrected without impacting the end-user. A similar strategy could be adopted to ensure the quality of user contributed changes to a database that allows for community curation. Rather than deploying user changes directly to the public, proposed edits could be subject to expert validation prior to acceptance.

Chapter 4: Conclusion

Reference marker gene databases are crucial tools for microbial ecology research: they allow community membership to be inferred by mapping unknown sequencing reads to known organisms, and provide reference gene multiple sequence alignments used in biodiversity calculations and discovery. However, the quality of the annotations contained in many 16S reference gene databases has been taken for granted. Although quality assurance methods were employed by the curators of the well-established 16S rDNA gene reference databases used in most targeted gene surveys, the results of this project have revealed the presence of considerable nomenclature and taxonomic errors. This in turn calls into question the reliability of published biological inferences made using these resources. The exact impact of this problem on routine downstream calculations remains to be quantified, but this analysis suggests these errors could impact the validity of reported analytic inferences.

As contemporary research continues to rely upon these databases, there is a need to quality assure reference data so these data can be used with confidence. This will require either a thorough review of the records contained in existing 16S rDNA databases, or the curation of a new database created with a more rigorous approach to the collection of nomenclature data. While correcting the annotations contained in the existing databases would avoid the duplication of considerable work, it would remain up to the curators to dedicate further time and resources to this undertaking. Therefore, the curation of a new 16S rDNA reference database can be justified if the existing databases cannot be corrected in a reasonable time. Several quality control strategies have been provided here, many having been implemented successfully in other disciplines and projects. While these strategies were designed with the curation of a 16S rDNA gene reference database in mind, they are broadly applicable, and would be suitable for the creation of other reference databases containing other marker gene sequences.

Community awareness of the problem can help drive the need to make important quality improvements. Community involvement in curation has been proposed as a method of decreasing the reliance on already overburdened database curators, but the perceived lack of incentive for participation is an important obstacle. This may be addressed in part by the tracking and attribution of user-contributed curation, but greater recognition of the need for genomics database quality assurance by academic institutions and funding agencies is sorely needed. Failure to address this problem will compromise the quality of future research efforts that depend on these resources, and the implementation of solutions in the future may become increasingly difficult the longer errors continue to accumulate.

Bibliography

[1] S. Creer *et al.*, "The ecologist's field guide to sequence-based identification of biodiversity," *Methods Ecol. Evol.*, vol. 7, pp. 1008–1018, 2016.

[2] S. W. Kembel, J. A. Eisen, K. S. Pollard, and J. L. Green, "The phylogenetic diversity of metagenomes," *PLoS One*, vol. 6, no. 8, 2011.

[3] J. Ma, A. Prince, and K. M. Aagaard, "Use of whole genome shotgun metagenomics : A practical guide for the microbiome-minded physician scientist," *Semin Reprod Med*, vol. 32, pp. 5–13, 2014.

[4] N. Sangwan, F. Xia, and J. A. Gilbert, "Recovering complete and draft population genomes from metagenome datasets," *Microbiome*, vol. 4, no. 1, p. 8, 2016.

[5] V. Kunin, A. Engelbrektson, H. Ochman, and P. Hugenholtz, "Wrinkles in the rare biosphere: Pyrosequencing errors can lead to artificial inflation of diversity estimates," *Environ. Microbiol.*, vol. 12, no. 1, pp. 118–123, 2010.

[6] M. Schirmer, U. Z. Ijaz, R. D'Amore, N. Hall, W. T. Sloan, and C. Quince, "Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform," *Nucleic Acids Res.* , vol. 43, no. 6, pp. e37–e37, Mar. 2015.

[7] V. Skerman, V. McGowan, P. Sneath, and H. Andrews, "Approved lists of bacterial names," *Int. J. Syst. Evol. Microbiol.*, vol. 30, no. 1, pp. 225–420, 1980.

[8] A. C. Parte, "LPSN-list of prokaryotic names with standing in nomenclature," *Nucleic Acids Res.*, vol. 42, pp. D613–D616, 2013.

[9] F. Guarner and J. R. Malagelada, "Gut flora in health and disease," *Lancet*, vol. 361, no. 9356, pp. 512–519, 2003.

[10] K. J. Pflughoeft and J. Versalovic, "Human microbiome in health and disease," *Annu. Rev.*

Pathol. Mech. Dis., vol. 7, pp. 99–122, 2012.

[11] P. G. Falkowski, T. Fenchel, and E. F. Delong, “The microbial engines that drive earth’s biogeochemical cycles,” *Science*, vol. 320, no. 5879, pp. 1034–1039, 2008.

[12] C. R. Woese, “Bacterial evolution,” *Microbiol. Rev.*, vol. 51, no. 2, pp. 221–271, 1987.

[13] J. E. Clarridge, “Impact of 16S rRNA gene sequence analysis for identification of bacteria on clinical microbiology and infectious diseases,” *Clin. Microbiol. Rev.*, vol. 17, no. 4, pp. 840–862, 2004.

[14] P. Londei, “Archaeal Ribosomes,” *eLS*.

[15] E. Sentausa and P. Fournier, “Advantages and limitations of genomics in prokaryotic taxonomy,” *Clin. Microbiol. Infect.*, vol. 19, no. 9, pp. 790–795, 2013.

[16] C. H. Kuo and H. Ochman, “Inferring clocks when lacking rocks: The variable rates of molecular evolution in bacteria,” *Biol Direct*, vol. 4, p. 35, 2009.

[17] M. Martens, P. Dawyndt, R. Coopman, M. Gillis, P. De Vos, and A. Willems, “Advantages of multilocus sequence analysis for taxonomic studies: A case study using 10 housekeeping genes in the genus Ensifer (including former Sinorhizobium),” *Int. J. Syst. Evol. Microbiol.*, vol. 58, no. 1, pp. 200–214, 2008.

[18] S. Mignard and J. P. Flandrois, “16S rRNA sequencing in routine bacterial identification: A 30-month experiment,” *J. Microbiol. Methods*, vol. 67, no. 3, pp. 574–581, 2006.

[19] A. Oren and G. M. Garrity, “Then and now: A systematic review of the systematics of prokaryotes in the last 80 years,” *Antonie Van Leeuwenhoek*, vol. 106, no. 1, pp. 43–56, 2014.

[20] X. Y. Zhi, W. Zhao, W. J. Li, and G. P. Zhao, “Prokaryotic systematics in the genomics era,” *Antonie Van Leeuwenhoek*, vol. 101, no. 1, pp. 21–34, 2012.

[21] A. Oren, “Systematics of archaea and bacteria,” *Biol. Sci. Fundam. Syst.*, vol. 2, 2009.

[22] R. Sharma, A. V. Polkade, and Y. S. Shouche, “‘Species concept’ in microbial taxonomy and systematics,” *Curr. Sci.*, vol. 108, no. 10, pp. 1804–1814, 2015.

[23] A. Oren, “Microbial Systematics,” in *Environmental Biotechnology*, L. K. Wang et al., Eds. Totowa, NJ: Humana Press, 2010, pp. 81–120.

[24] P. Kämpfer, “Systematics of prokaryotes: The state of the art,” *Antonie Van Leeuwenhoek*, vol. 101, no. 1, pp. 3–11, 2012.

[25] D. McDonald *et al.*, “An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea,” *ISME J.*, vol. 6, no. 3, pp. 610–618, 2012.

[26] J. R. Cole *et al.*, “Ribosomal Database Project: Data and tools for high throughput rRNA analysis,” *Nucleic Acids Res.*, vol. 42, pp. D633–D642, 2014.

[27] S. Federhen, “The NCBI taxonomy database,” *Nucleic Acids Res.*, vol. 40, pp. D136–D143, 2012.

[28] P. Yarza *et al.*, “Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences,” *Nat Rev Micro*, vol. 12, no. 9, pp. 635–645, Sep. 2014.

[29] R. C. Edgar, B. J. Haas, J. C. Clemente, C. Quince, and R. Knight, “UCHIME improves sensitivity and speed of chimera detection,” *Bioinformatics*, vol. 27, no. 16, pp. 2194–2200, 2011.

[30] E. P. Nawrocki, D. L. Kolbe, and S. R. Eddy, “Infernal 1.0: Inference of RNA alignments,” *Bioinformatics*, vol. 25, no. 10, pp. 1335–1337, 2009.

[31] B. J. Haas *et al.*, “Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons,” *Genome Res.*, vol. 21, no. 3, pp. 494–504, 2011.

[32] E. Pruesse *et al.*, “SILVA: A comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB,” *Nucleic Acids Res.*, vol. 35, no. 21, pp. 7188–7196, 2007.

[33] E. Pruesse, J. Peplies, and F. O. Glöckner, “SINA: Accurate high-throughput multiple sequence alignment of ribosomal RNA genes,” *Bioinformatics*, vol. 28, no. 14, pp. 1823–1829, 2012.

[34] K. E. Ashelford, N. A. Chuzhanova, J. C. Fry, A. J. Jones, and A. J. Weightman, “At least 1 in 20 16S rRNA sequence records currently held in public repositories is estimated to contain substantial anomalies,” *Appl. Environ. Microbiol.*, vol. 71, no. 12, pp. 7724–7736, 2005.

[35] J. J. Werner *et al.*, “Impact of training sets on classification of high-throughput bacterial 16s rRNA gene surveys,” *ISME J.*, vol. 6, no. 1, pp. 94–103, 2012.

[36] I. L. Newton and G. Roeselers, “The effect of training set on the classification of honey bee gut microbiota using the Naive Bayesian Classifier,” *BMC Microbiol.*, vol. 12, p. 221, 2012.

[37] A. Oren and S. Ventura, “The current status of cyanobacterial nomenclature under the ‘prokaryotic’ and the ‘botanical’ code,” *Antonie van Leeuwenhoek, Int. J. Gen. Mol. Microbiol.*, pp. 1–13, 2017.

[38] P. L. Buttigieg, N. Morrison, B. Smith, C. J. Mungall, S. E. Lewis, and ENVO Consortium, “The environment ontology: Contextualising biological and biomedical entities,” *J. Biomed. Semantics*, vol. 4, no. 1, p. 43, 2013.

[39] P. L. Buttigieg, E. Pafilis, S. E. Lewis, M. P. Schildhauer, R. L. Walls, and C. J. Mungall, “The environment ontology in 2016: Bridging domains with increased scope, semantic

density, and interoperation," *J. Biomed. Semantics*, vol. 7, p. 57, 2016.

[40] W. Ludwig *et al.*, "ARB: A software environment for sequence data," *Nucleic Acids Res.*, vol. 32, no. 4, pp. 1363–1371, 2004.

[41] M. A. Musen, "The Protégé project: A look back and a look forward," *AI Matters*, vol. 1, no. 4, pp. 4–12, 2015.

[42] D. Laehnemann, A. Borkhardt, and A. C. McHardy, "Denoising DNA deep sequencing data—high-throughput sequencing errors and their correction," *Brief. Bioinform.*, vol. 17, no. 1, pp. 154–179, 2016.

[43] K. Robasky, N. E. Lewis, and G. M. Church, "The role of replicates for error mitigation in next-generation sequencing," *Nat. Rev. Genet.*, vol. 15, no. 1, pp. 56–62, 2014.

[44] E. R. Mardis, "Next-generation sequencing platforms," *Annu. Rev. Anal. Chem*, vol. 6, pp. 287–303, 2013.

[45] J. H. Lee, H. Yi, and J. Chun, "rRNASelector: A computer program for selecting ribosomal RNA encoding sequences from metagenomic and metatranscriptomic shotgun libraries," *J. Microbiol.*, vol. 49, no. 4, pp. 689–691, 2011.

[46] C. Yuan, J. Lei, J. Cole, and Y. Sun, "Reconstructing 16S rRNA genes in metagenomic data," *Bioinformatics*, vol. 31, no. 12, pp. i35–i43, 2015.

[47] K. Lagesen, P. Hallin, E. A. Rødland, H. H. Stærfeldt, T. Rognes, and D. W. Ussery, "RNAmmer: Consistent and rapid annotation of ribosomal RNA genes," *Nucleic Acids Res.*, vol. 35, no. 9, pp. 3100–3108, 2007.

[48] R. P. Smyth *et al.*, "Reducing chimera formation during PCR amplification to ensure accurate genotyping," *Gene*, vol. 469, pp. 45–51, 2010.

[49] M. S. Judo, A. B. Wedel, and C. Wilson, "Stimulation and suppression of PCR-mediated

recombination,” *Nucleic Acids Res.*, vol. 26, no. 7, pp. 1819–1825, 1998.

[50] P. D. Schloss, D. Gevers, and S. L. Westcott, “Reducing the effects of PCR amplification and sequencing artifacts on 16S rRNA-based studies,” *PLoS One*, vol. 6, no. 12, 2011.

[51] K. E. Ashelford, N. A. Chuzhanova, J. C. Fry, A. J. Jones, and A. J. Weightman, “New screening software shows that most recent large 16S rRNA gene clone libraries contain chimeras,” *Appl. Environ. Microbiol.*, vol. 72, no. 9, pp. 5734–5741, 2006.

[52] R. H. Nilsson *et al.*, “An open source chimera checker for the fungal ITS region,” *Mol. Ecol. Resour.*, vol. 10, no. 6, pp. 1076–1081, 2010.

[53] T. Huber, G. Faulkner, and P. Hugenholtz, “Bellerophon: A program to detect chimeric sequences in multiple sequence alignments,” *Bioinformatics*, vol. 20, no. 14, pp. 2317–2319, 2004.

[54] E. S. Wright, L. S. Yilmaz, and D. R. Noguera, “DECIPHER, a search-based approach to chimera identification for 16S rRNA sequences,” *Appl. Environ. Microbiol.*, vol. 78, no. 3, pp. 717–725, 2012.

[55] M. Mysara, Y. Saeys, N. Leys, J. Raes, and P. Monsieurs, “CATCh, an ensemble classifier for chimera detection in 16S rRNA sequencing studies,” *Appl. Environ. Microbiol.*, vol. 81, no. 5, pp. 1573–1584, 2015.

[56] R. Edgar, “UCHIME2: Improved chimera prediction for amplicon sequencing,” *bioRxiv*, 2016.

[57] C. Quince, A. Lanzen, R. J. Davenport, and P. J. Turnbaugh, “Removing noise from pyrosequenced amplicons.,” *BMC Bioinformatics*, vol. 12, no. 1, p. 38, 2011.

[58] D. Optiz and R. Maclin, “Popular ensemble methods: An empirical study,” *J. Artifical Intell. Res.*, vol. 11, pp. 169–198, 1999.

[59] G. Brown, J. Wyatt, R. Harris, and X. Yao, “Diversity creation methods: A survey and categorisation,” *Inf. Fusion*, vol. 6, no. 1, pp. 5–20, 2005.

[60] L. I. Kuncheva and C. J. Whitaker, “Measures of diversity in classifier ensembles,” *Mach. Learn.*, vol. 51, no. 2, pp. 181–207, 2003.

[61] R. T. Jones, L. G. Sanchez, and N. Fierer, “A cross-taxon analysis of insect-associated bacterial diversity,” *PLoS One*, vol. 8, no. 4, p. e61218, 2013.

[62] C. W. Nossa *et al.*, “Design of 16S rRNA gene primers for 454 pyrosequencing of the human foregut microbiome,” *World J. Gastroenterol.*, vol. 16, no. 33, pp. 4135–4144, 2010.

[63] A. S. Hanshew, C. J. Mason, K. F. Raffa, and C. R. Currie, “Minimization of chloroplast contamination in 16S rRNA gene pyrosequencing of insect herbivore bacterial communities,” *J. Microbiol. Methods*, vol. 95, no. 2, pp. 149–155, 2013.

[64] R. C. Edgar, “UPARSE: Highly accurate OTU sequences from microbial amplicon reads,” *Nat. Methods*, vol. 10, no. 10, pp. 996–8, 2013.

[65] R. Schmieder and R. Edwards, “Fast identification and removal of sequence contamination from genomic and metagenomic datasets,” *PLoS One*, vol. 6, no. 3, p. e17288, Jan. 2011.

[66] J. Chu *et al.*, “BioBloom tools: Fast, accurate and memory-efficient host species sequence screening using bloom filters,” *Bioinformatics*, vol. 30, no. 23, pp. 3402–3404, 2014.

[67] P. D. Schloss, “The effects of alignment quality, distance calculation method, sequence filtering, and region on the analysis of 16S rRNA gene-based studies,” *PLoS Comput. Biol.*, vol. 6, no. 7, p. e1000844, 2010.

[68] Q. Zhou, X. Su, and K. Ning, “Assessment of quality control approaches for metagenomic

data analysis," *Sci. Rep.*, vol. 4, 2014.

[69] R. K. Patel and M. Jain, "NGS QC toolkit: A toolkit for quality control of next generation sequencing data," *PLoS One*, vol. 7, no. 2, 2012.

[70] R. Schmieder, Y. W. Lim, F. Rohwer, and R. Edwards, "TagCleaner: Identification and removal of tag sequences from genomic and metagenomic datasets," *BMC Bioinformatics*, vol. 11, p. 341, 2010.

[71] W. Ludwig and H. Klenk, "Overview: A phylogenetic backbone and taxonomic framework for procaryotic systematics," in *Bergey's Manual® of Systematic Bacteriology*, New York: Springer, 2005, pp. 49–66.

[72] J. Castresana, "Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis," *Mol. Biol. Evol.*, vol. 17, no. 4, pp. 540–552, 2000.

[73] P. D. Schloss, "Secondary structure improves OTU assignments of 16S rRNA gene sequences," *ISME J.*, vol. 7, no. 3, pp. 457–460, 2013.

[74] X. Wang, Y. Cai, Y. Sun, R. Knight, and V. Mai, "Secondary structure information does not improve OTU assignment for partial 16s rRNA sequences," *ISME J.*, vol. 6, no. 7, pp. 1277–1280, 2012.

[75] R. D'Amore *et al.*, "A comprehensive benchmarking study of protocols and sequencing platforms for 16S rRNA community profiling," *BMC Genomics*, vol. 17, p. 55, 2016.

[76] S. J. Salipante *et al.*, "Performance comparison of Illumina and Ion Torrent next-generation sequencing platforms for 16S rRNA-based bacterial community profiling," *Appl. Environ. Microbiol.*, vol. 80, no. 24, pp. 7583–7591, 2014.

[77] T. Seemann, "Ten recommendations for creating usable bioinformatics command line software," *Gigascience*, vol. 2, p. 15, 2013.

[78] D. Bolchini, A. Finkelstein, V. Perrone, and S. Nagl, “Better bioinformatics through usability analysis,” *Bioinformatics*, vol. 25, no. 3, pp. 406–412, 2009.

[79] S. J. Schultheiss, M. C. Münch, G. D. Andreeva, and G. Rätsch, “Persistence and availability of web services in computational biology,” *PLoS One*, vol. 6, no. 9, p. e24914, 2011.

[80] J. D. Wren, “URL decay in MEDLINE—a 4-year follow-up study,” *Bioinformatics*, vol. 24, no. 11, pp. 1381–1385, 2008.

[81] J. D. Wren, “404 not found: The stability and persistence of URLs published in MEDLINE,” *Bioinformatics*, vol. 20, no. 5, pp. 668–672, 2004.

[82] A. L. Boulesteix, “Over-optimism in bioinformatics research,” *Bioinformatics*, vol. 26, no. 3, pp. 437–439, 2009.

[83] N. Bokulich *et al.*, “A standardized, extensible framework for optimizing classification improves marker-gene taxonomic assignments,” *PeerJ Prepr.*, vol. 3, p. e934v2, 2015.

[84] J. J. Kozich, S. L. Westcott, N. T. Baxter, S. K. Highlander, and P. D. Schloss, “Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform,” *Appl. Environ. Microbiol.*, vol. 79, no. 17, pp. 5112–5120, 2013.

[85] K. Kukich, “Technique for automatically correcting words in text,” *ACM Comput. Surv.*, vol. 24, no. 4, pp. 377–439, 1992.

[86] R. Mitton, “Ordering the suggestions of a spellchecker without using context,” *Nat. Lang. Eng.*, vol. 15, no. 2, pp. 173–192, 2009.

[87] M. Flor, “Four types of context for automatic spelling correction,” *TAL Trait. Autom. des Langues*, vol. 53, no. 3, pp. 61–99, 2012.

[88] S. P. Lapage, P. H. Sneath, E. F. Lessel, V. B. Skerman, H. P. Seeliger, and W. A. Clark, *International Code of Nomenclature of Bacteria: Bacteriological Code, 1990 Revision*. Washington, DC, USA: ASM Press, 1992.

[89] T. Rees, “Taxamatch, an algorithm for near (’Fuzzy’) matching of scientific names in taxonomic databases,” *PLoS One*, vol. 9, no. 9, p. e107510, 2014.

[90] B. Boyle *et al.*, “The taxonomic name resolution service: An online tool for automated standardization of plant names,” *BMC Bioinformatics*, vol. 14, no. 16, 2013.

[91] G. E. Fox, J. D. Wisotzkey, and P. Jurtshuk, “How close is close: 16S rRNA sequence identity may not be sufficient to guarantee species identity.,” *Int. J. Syst. Bacteriol.*, vol. 42, no. 1, pp. 166–170, 1992.

[92] S. Federhen, “Type material in the NCBI Taxonomy Database,” *Nucleic Acids Res.*, vol. 43, pp. D1086–D1098, 2015.

[93] Y. Wai Ho, Z. Zhang, and Y. Wang, “Distinct types of rRNA operons exist in the genome of the actinomycete Thermomonospora chromogena and evidence for horizontal transfer of an entire rRNA operon,” *J. Bacteriol.*, vol. 181, no. 17, pp. 5201–5209, 1999.

[94] W. H. Yap and Y. Wang, “Molecular cloning and comparative sequence analyses of rRNA operons in Streptomyces nodosus ATCC 14899,” *Gene*, vol. 232, no. 1, pp. 77–85, 1999.

[95] Y. Wang, Z. Zhang, and N. Ramanan, “The actinomycete Thermobispora bispora contains two distinct types of transcriptionally active 16S rRNA genes,” *J. Bacteriol.*, vol. 179, no. 10, pp. 3270–3276, 1997.

[96] S. Mylvaganam and P. P. Dennis, “Sequence heterogeneity between the two genes encoding 16S rRNA from the halophilic archaebacterium Haloarcula marismortui,”

Genetics, vol. 130, no. 3, pp. 399–410, 1992.

[97] T. P. Tourova, B. B. Kuznetzov, E. V Novikova, A. B. Poltaraus, and T. N. Nazina, “Heterogeneity of the nucleotide sequences of the 16S rRNA genes of the type strain of Desulfotomaculum kuznetsovii,” *Mikrobiologiya*, vol. 70, no. 6, pp. 788–795, 2001.

[98] C. R. Arias, O. Olivares-Fuster, and J. Goris, “High intragenomic heterogeneity of 16S rRNA genes in a subset of Vibrio vulnificus strains from the western Mediterranean coast,” *Int. Microbiol.*, vol. 13, no. 4, pp. 179–188, 2010.

[99] J. Hashimoto, B. Stevenson, and T. M. Schmidt, “Rates and consequences of recombination between ribosomal RNA operons,” *J. Bacteriol.*, vol. 185, no. 3, pp. 966–972, 2002.

[100] D. J. Brenner, J. T. Staley, and N. R. Krieg, “Classification of procaryotic organisms and the concept of bacterial speciation,” in *Bergey’s Manual of Systematic Bacteriology Volume Two: The Proteobacteria, Part A Introductory Essays*, Boston, MA: Springer US, 2005, pp. 27–32.

[101] E. A. Pettengill, J. B. Pettengill, and R. Binet, “Phylogenetic analyses of Shigella and enteroinvasive Escherichia coli for the identification of molecular epidemiological markers: Whole-genome comparative analysis does not support distinct genera designation,” *Front. Microbiol.*, vol. 6, pp. 1–11, 2016.

[102] M. D. Collins *et al.*, “The phylogeny of the genus Clostridium: Proposal of five new genera and eleven new species combinations,” *Int. J. Syst. Bacteriol.*, vol. 44, no. 4, pp. 812–826, 1994.

[103] M. Piattini, C. Calero, and M. Genero, “The organization’s most important data issues,” in *Information and Database Quality*, Springer Science & Business Media, 2002, pp. 13–14.

[104] M. V Patil and A. M. Yogi, “Importance of data collection and validation for systematic software development process,” *Int’l J. Comput. Sci. Inf. Technol.*, vol. 3, no. 2, pp. 260–278, 2011.

[105] S. Ambler and P. Sadalage, “Evolutionary database development,” in *Refactoring Databases: Evolutionary Database Design*, Boston, MA, USA: Addison Wesley Professional, 2006, pp. 16–25.

[106] Y. Jiang, B. Cukic, and T. Menzies, “Fault prediction using early lifecycle data,” in *18th IEEE Int. Symp. Software Reliability Engineering*, 2007, pp. 237–246.

[107] B. W. Boehm, “Software risk management: Principles and practices,” *IEEE Softw.*, vol. 8, no. 1, pp. 32–41, 1991.

[108] D. Howe and S. Yon, “The future of biocuration,” *Nature*, vol. 455, no. 7209, pp. 47–50, 2008.

[109] J. C. Hu *et al.*, “The emerging world of wikis,” *Science*, vol. 320, no. 5881, pp. 1289–1290, Jun. 2008.

[110] D. M. Bolser *et al.*, “MetaBase—the wiki-database of biological databases,” *Nucleic Acids Res.*, vol. 40, pp. D1250–D1254, 2012.

[111] S. L. Salzberg, “Genome re-annotation: A wiki solution?,” *Genome Biol.*, vol. 8, no. 1, p. 102, 2007.

[112] Z. Zhang *et al.*, “RiceWiki: A wiki-based database for community curation of rice genes,” *Nucleic Acids Res.*, vol. 42, pp. D1222–D1228, 2014.

[113] A. R. Pico, T. Kelder, M. P. Van Iersel, K. Hanspers, B. R. Conklin, and C. Evelo, “WikiPathways : Community-based curation of biological pathways,” *Nature Precedings*, 2010. [Online]. Available: http://dx.doi.org/10.1038/npre.2010.5361.1.

[114] I. A. Chen *et al.*, “Supporting community annotation and user collaboration in the integrated microbial genomes (IMG) system,” *BMC Genomics*, vol. 17, p. 307, 2016.

[115] L. Dai *et al.*, “AuthorReward: Increasing community curation in biological knowledge wikis through automated authorship quantification,” *Bioinformatics,* vol. 29, no. 14, pp. 1837–1839, 2013.

[116] R. Gliklich, N. Dreyer, and M. Leavy, “Data collection and quality assurance,” in *Registries for Evaluating Patient Outcomes: A User’s Guide*, 3rd ed., vol. 1, R. Gliklich et al., Eds. Rockville (MD): Agency for Healthcare Research and Quality, 2014, p. 360.

Appendices

Supplemental material for this thesis are available at http://www.bcgsc.ca/downloads/klesack/

Appendix A - Supplementary Data

A.1

Supplementary Data File

Description:

The accompanying csv file contains the list of valid prokaryote names that were obtained from CyanoDB, ITIS, LPSN, and Prokaryote Nomenclature Up-To-Date. The file contains three columns: (1) the organism name, (2) rank, and (3) source database(s).

Filename: valid_names.csv

A.2

Supplementary Data File

Description:

The accompanying csv file contains the list of invalid prokaryote names that were obtained from CyanoDB, ITIS, LPSN, and Prokaryote Nomenclature Up-To-Date. The file contains three columns: (1) the organism name, (2) rank, and (3) source database(s).

Filename: invalid_names.csv

A.3

Supplementary Data File

Description:

The accompanying csv file contains the list of valid prokaryote names included in the nomenclature Core-Set. The names were obtained from CyanoDB, LPSN, and Prokaryote Nomenclature Up-To-Date, and are considered accurate. The file contains three columns: (1) the organism name, (2) rank, and (3) source database(s).

Filename: valid_names_coreset.csv

A.4

Supplementary Data File

Description:

The accompanying csv file contains the list of invalid prokaryote names included in the nomenclature Core-Set. The names were obtained from CyanoDB, LPSN, and Prokaryote Nomenclature Up-To-Date, and are considered accurate. The file contains three columns: (1) the organism name, (2) rank, and (3) source database(s).

Filename: invalid_names_coreset.csv

A.5

Supplementary Data File

Description:

The accompanying csv file contains the list of synonymous prokaryote names included in the nomenclature Core-Set. The names were obtained from CyanoDB, LPSN, and Prokaryote Nomenclature Up-To-Date, and are considered accurate. The file contains three columns: (1) the organism name, (2) rank, and (3) source database(s).

Filename: synonyms_coreset.csv

A.6

Supplementary Data File

Description:

The accompanying csv files contains the lists of misannotated 16S rDNA gene sequences found in RDP (release 9, update 22), SILVA (v. 123), and Greengenes (v. 13_5) databases. Validity was assessed using the invalid and synonymous core-sets. Separate files are provided for each database. Each file contains three columns: (1) the organism name, (2) rank, and (3) type of misannotation.

Filenames:

rdp_misannotations.txt

silva_misannotations.csv

greengenes_misannotations.csv

A.7

Supplementary Data File

Description:

The accompanying csv file contains the sample environment for misannotated sequences found in SILVA (v. 123). Sample environment metadata were obtained from the "isolation_source" field in the ARB (v. 128) database. The file contains six columns: (1) ARB database ID, (2) INSDC accession number, (3) organism name, (4) isolate source metadata, (5) the broad sampling environment category, and (6) the sampling environment subcategory. Data provided in columns 1-4 were obtained from ARB. Columns 5 and 6 (if a subcategory was assigned) indicate the assigned categories from section 2.3.

Filename: misannotations_sample_env.csv

Appendix B - Python and R Scripts

Supplemental material for this thesis are available at http://www.bcgsc.ca/downloads/klesack/

B.1

Supplementary Script Files

Description:

The accompanying Python script files were used to collect lists of invalid, valid, and candidatus prokaryote names from LPSN.

Filenames: LPSN_scripts.tar.gz, LPSN_scripts.zip

B.2

Supplementary Script Files

Description:

The accompanying Python script files were used to collect lists of invalid, and valid Cyanobacteria names from CyanoDB.

Filenames: CyanoDB_scripts.tar.gz, CyanoDB_scripts.zip

B.3

Supplementary Script File

Description:

The accompanying R script files were used to collect lists of invalid, valid, and candidatus prokaryote names from ITIS.

Filenames: ITIS_scripts.tar.gz, ITIS_scripts.zip