Tackling the Current Limitations of Bacterial Taxonomy with Genome-Based Classification and Identification on a Crowdsourcing Web Service

Long Tian

Dissertation submitted to the faculty of the Virginia Polytechnic Institute and State University in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
In
Genetics, Bioinformatics, and Computational Biology

Boris A. Vinatzer, Chair
Lenwood S. Heath, Chair
Paul Marek, Member
Liqing Zhang, Member

August 16, 2019
Blacksburg, VA

Keywords: Bacterial taxonomy, average nucleotide identity, ANI, min-wise independent permutations, locality sensitive hashing, MinHash, Web service, crowdsourcing

Tackling the current limitations of bacterial taxonomy with genome-based classification and identification with a crowdsourcing Web service

Long Tian

ABSTRACT

Bacterial taxonomy is the science of classifying, naming, and identifying bacteria. The scope and practice of taxonomy has evolved through history with our understanding of life and our growing and changing needs in research, medicine, and industry. As in animal and plant taxonomy, the species is the fundamental unit of taxonomy, but the genetic and phenotypic diversity that exists within a single bacterial species is substantially higher compared to animal or plant species. Therefore, the current “type”-centered classification scheme that describes a species based on a single type strain is not sufficient to classify bacterial diversity, in particular in regard to human, animal, and plant pathogens, for which it is necessary to trace disease outbreaks back to their source. Here we discuss the current needs and limitations of classic bacterial taxonomy and introduce LINbase, a Web service that not only implements current species-based bacterial taxonomy but complements its limitations by providing a new framework for genome sequence-based classification and identification independently of the type-centric species. LINbase uses a sequence similarity-based framework to cluster bacteria into hierarchical taxa, which we call LINgroups, at multiple levels of relatedness and crowdsources users’ expertise by encouraging them to circumscribe these groups as taxa from the genus-level to the intraspecies-level. Circumscribing a group of bacteria as a LINgroup, adding a phenotypic description, and giving the LINgroup a name using the LINbase Web interface allows users to instantly share new taxa and complements the lengthy and laborious process of publishing a named species. Furthermore, unknown isolates can be identified immediately as members of a newly described LINgroup with fast and precise algorithms based on their genome sequences, allowing species- and intraspecies-level identification. The employed algorithms are based on a combination of the alignment-based algorithm BLASTN and the alignment-free method Sourmash, which is based on k-mers, and the MinHash algorithm. The potential of LINbase is shown by using examples of plant pathogenic bacteria.

Tackling the current limitations of bacterial taxonomy with genome-based classification and identification with a crowdsourcing Web service

Long Tian

GENERAL AUDIENCE ABSTRACT

Life is always easier when people talk to each other in the same language. Taxonomy is the language that biologists use to communicate about life by 1. classifying organisms into groups, 2. giving names to these groups, and 3. identifying individuals as members of these named groups. When most scientists and the general public think of taxonomy, they think of the hierarchical structure of “Life”, “Domain”, “Kingdom”, “Phylum”, “Class”, “Order”, “Family”, “Genus” and “Species”. However, the basic goal of taxonomy is to allow the identification of an organism as a member of a group that is predictive of its characteristics and to provide a name to communicate about that group with other scientists and the public. In the world of micro-organism, taxonomy is extremely important since there are an estimated 10,000,000 to 1,000,000,000 different bacteria species. Moreover, microbiologists and pathologists need to consider differences among bacterial isolates even within the same species, a level, that the current taxonomic system does not even cover. Therefore, we developed a Web service, LINbase, which uses genome sequences to classify individual microbial isolates. The database at the backend of LINbase assigns Life Identification Numbers (LINs) that express how individual microbial isolates are related to each other above, at, and below the species level. The LINbase Web service is designed to be an interactive web-based encyclopedia of micro-organisms where users can share everything they know about micro-organisms, be it individual isolates or groups of isolates, for professional and scientific purposes. To develop LINbase, efficient computer programs were developed and implemented. To show how LINbase can be used, several groups of bacteria that cause plant diseases were classified and described.

Acknowledgement

My time with the GBCB program, the School of Plant and Environmental Sciences, and Virginia Tech in the town of Blacksburg, Virginia has been the most beautiful four years in my life as far as I can remember. I came here for pursuing a Ph.D. degree, and what I have gained is much beyond that. Just as what was written on the cookie my advisor Dr. Boris Vinatzer gave me for my last lab meeting, “this is home,” and it is everyone who has been with me on this journey that makes it home. I owe them my endless gratitude.

First and foremost I would like to thank my advisors and committee members, Dr. Boris Vinatzer, Dr. Lenwood Heath, Dr. Paul Marek, and Dr. Liqing Zhang for their wisdom and support in every aspect of the project. LINbase is an excellent idea and project with profound meaning to science, and I’m glad that I have been a part of it. Boris, you are the best boss that I have ever had, a great mentor, and a true friend. I hope we will never fall out of touch.

Secondly, I want to thank Dr. David Bevan, Dr. Christopher Lawrence, and Dennie Munson of the GBCB program for taking me in.

LINbase is inspired by a lot of brilliant scientists and students. Thank Dr. Hathaim Merakeby and Dr. Alex Weisberg for the prototype of LIN assignment. Thank Grant Hughes, who is the first CS undergraduate student that joined the project and helped me started the Web service. Thank Chengjie (Nick) Huang, who has done a fantastic job to the user interface, hope you love your life as a Ph.D. student at Waterloo University. And thank everyone who has given me advice on LINbase.

I want to thank my mentor during my internship at Mayo Clinic, Dr. Zhifu Sun. As well as everyone I have met in Rochester, MN. It was a great experience.

I thank my lab members, Dr. Kevin Failor, Dr. Noam Eckshtain-Levi, Haijie Liu, Marco Mechan, Marcela Aguilera, and Parul Sharma. It has been a great pleasure to work with you all.

I also want to acknowledge my fellow GBCB students. Thanks for having my back all the time.

Finally, I sincerely thank my mother Feng Tian, for understanding and supporting me in pursuing my dream even after I had already got a job. Last but not least, my dearly beloved wife, Jiayi Liu, thank you for having faith in me and taking care of me when I was off from work-life balance, and I’m ready to spend the rest of my life with you.

Table of Contents

List of Figures

List of Tables

Chapter 1 Current Methods for Genome-Based Microbial Classification and Identification and Their Implementation in Software and Online Platforms

Long Tian[1], Lenwood S. Heath[2], Boris A. Vinatzer[1]

[1]School of Plant and Environmental Sciences, Virginia Tech, Blacksburg, VA 24061, USA,
[2]Department of Computer Science, Virginia Tech, Blacksburg, VA 24061, USA

Attribution

The following chapter is based upon the manuscript in preparation: Tian, L., Heath, L.S., Vinatzer, B.A. (2019). Current methods for genome-based microbial classification and identification and their implementation in software and online platforms.

Abstract

Taxonomy is the science of classifying, naming, and identifying organisms. While nomenclature has not changed significantly for hundreds of years and continues to use binomial species names, the methods used to classify and identify organisms, in particular microbes, have changed considerably with the introduction of new DNA sequencing technologies. In particular, the advent of Next Generation Sequencing technologies, which dramatically increased throughput and reduced cost, allowed the transition from gene-based classification and identification methods to whole genome-based methods providing resolution and precision that were impossible to achieve with past morphological, biochemical, and even single gene-based methods. Here we describe the new opportunities for microbial classification and identification at the species and strain level that have been created by the combination of whole genome sequencing, databases accessible through the Internet, new efficient algorithms, and cloud computing. We then compare current genome-based classification and identification methods and their implementations in tools and platforms with respects to resolution, accuracy, accessibility, and scalability. We conclude with an outlook on methods, tools, and platforms that we expect to see in the near future.

Introduction

The goal of taxonomy is to identify organisms as members of named classes, also called taxa (singular: taxon), to accurately predict their characteristics, also referred to as phenotypes (Identification). To reach this goal, each taxon needs to be composed of organisms with phenotypes that are conserved among all its members whereby at least some of these phenotypes need to be absent from the organisms outside of the same taxon (Classification). Finally, broadly agreed upon and unique names need to be used to name taxa to allow for organization and clear communication about them (Nomenclature).

With the introduction of an evolutionary framework to taxonomy, it became obvious that taxa should correspond to monophyletic groups (i.e., groups that consist of members that all evolved from the same most recent common ancestor [MRCA]) and that taxa need to be placed into a hierarchical system with the taxa that evolved from more recent ancestors being situated with taxa instead of those that evolved from more distant ancestors (*i.e.*, species are placed inside genera and genera are placed inside families, and so on). It also became clear, much earlier that the more recent an ancestor of a taxon is, the more phenotypes its members share with each other.

Before the introduction of DNA sequencing technologies, the main way to circumscribe microbial taxa and to identify unknown isolates as members of taxa was necessarily to use laboratory-based tests to determine phenotypes. The more phenotypic data were included in the description of taxa, the more reliably could taxa reflect evolutionary relationships and the more accurately could unknowns be identified as members of taxa. Importantly, the use of phenotypic data in taxa descriptions required phenotypic tests to be performed in the presence of reference strains of related taxa for comparison purposes. To make sure always the same reference strains were used and reference strains were available to the scientific community, a single strain needed to be designated as reference for each taxon and made available through culture collections. These reference strains are known as type strains. Type strains are not only the official representatives of each taxon when phenotypically comparing taxa with each other, they also are the official name carriers of each taxon.

This phenotype-based taxonomic system of microbes necessarily used taxa with stable and distinguishable phenotypes as ultimate units. These ultimate units are called species and nomenclatural rules make sure that each species has a unique name to communicate about them. Isolates within species often do not have stable phenotypes that distinguish them from each other. Distinguishing isolates from each other is thus not considered part of taxonomy. It is usually referred to as strain typing and investigates the epidemiological relationships among isolates to answer the main question of which outbreak strain a certain pathogen isolate belongs to.

Over time, phenotypic descriptions of species were integrated with measurements from which genetic distance between isolates could be inferred. The first such measurement was based on DNA-DNA hybridization (DDH) experiments, which experimentally determine how different the genomic DNA of any two bacteria is [1]. A 70% DDH value was chosen as the minimal threshold to consider two bacteria as members of the same species [2]. In spite of its potential to discriminate all taxonomic ranks, DDH is limited by experimental accuracy and the requirement for physical samples for each comparison, limiting portability of results [3, 4].

With the invention of polymerase chain reaction (PCR) and Sanger sequencing of DNA, it became possible to sequence the 16S rRNA gene, which is present in all microbes, and infer from its sequence, the evolutionary relationships among all microbes. Importantly, 16S rRNA sequencing made it possible to store sequence data in databases and compare the 16S rRNA sequence of any new isolates to all other previous sequences without having to physically move any bacteria creating the first truly portable classification and identification method. However, the 16S rRNA gene has the disadvantage that it can be 100% identical between closely related species and that it does not always accurately reflect evolutionary relationships [5]. Using Sanger sequencing to analyze a small number of housekeeping genes in the multi-locus sequence typing (MLST) method, increased the resolution of gene-based classification and identification methods allowing classification and identification from the genus level to the strain level [6]. However, the sequence diversity of housekeeping genes is too high to permit the use of the same genes for all microbes thereby limiting its applicability to classification at higher taxonomic ranks.

Finally, starting around 2005, the introduction of low cost, high throughput, Next Generation Sequencing (NGS) technology started to allow the use of whole genome sequencing (WGS) to classify and identify bacteria [7]. While NGS has changed the way we practice biological sciences, the most important effect may be on taxonomy and strain typing. In fact, WGS makes it now possible to precisely classify and identify microbes at all taxonomic ranks from kingdom to species and to type stains to identify pathogen isolates as belonging to individual outbreaks [8-10] . Phenotypic tests are not needed anymore for either describing taxa or identifying isolates as members of taxa [11]. This makes it even possible to classify and identify microbes without culturing them after assembling their genomes from metagenomic sequences, referred to as metagenome-assembled genomes (MAGs) [12-14] . A value of 95% average nucleotide identity (ANI) directly computed from WGS was found to correspond to the 70% DDH threshold for strains to belong to the same species [15] . The only limitation when using WGS for classification and identification is recombination between genetic lineages, which in some cases may make it impossible to unequivocally establish evolutionary relationships below the species level [16-19] . However, compared to gene-based methods, which are heavily influenced by recombination, the use of whole genome sequences in phylogenetic reconstruction generally makes recombination a minor issue [20].

Importantly, as with 16S rRNA databases, WGS databases are accessible by every lab over the Internet and make it possible to compare any isolate to every other isolate ever sequenced (as long as its sequence was deposited in a public database) to precisely identify isolates as members of taxa without the need to have type stains in the lab in order to do comparisons. On the flip side, the challenge with WGS-based classification and identification is the need for very efficient algorithms and considerable computing resources needed to make precise comparisons in a reasonable amount of time at all taxonomic ranks. Before describing current WGS-based classification and identification methods and how they handle these challenges, we review types of taxa that are used in WGS-based classification and identification.

Bacterial species and other units of bacterial classification

Taxa in classical taxonomy

Kingdom (domain), phylum, class, order, family, genus, and species are the ranks used in classical taxonomy, where the species is the fundamental unit [21] . Although each taxon should be monophyletic, there are cases in the current bacterial taxonomy where taxa are misclassified and inconsistent with evolutionary relationships. For example, the genus *Clostridium* consists of bacteria that phylogenetically belong to other taxonomic groups due to limitations of the historical phenotypic classification scheme [22]. To correct such misclassifications, in 2018, the Genome Taxonomy Database (GTDB) developed a phylogeny inferred from the alignment of 120 ubiquitous single-copy proteins across 94,759 bacterial genomes and proposed a taxonomy that more accurately represents evolutionary relationships [23].

The concept of "species" as the smallest unit of classification is still controversial in microbial taxonomy. A species was originally defined as a group of breeding or potentially inter-breeding groups of organisms in animal and plant taxonomies [24]. However, this definition cannot be used for microbes because microbes are asexual. For a period of time, taxonomists even claimed that species do not exist in bacteria [25]. In 2001, a pragmatic species concept was proposed by Rosselló-Mora and Amann that defined a species as "a monophyletic and genomically coherent cluster of individual organisms that show a high degree of overall similarity in many independent characteristics, and is diagnosable by a discriminative phenotypic property" [21]. This species concept has been widely accepted and is in general use today. This species concept is based on a "type"-centered scheme whereby each validly published species name is associated with a single bacterial isolate that constitutes the type strain of the named species [26]. To validly publish a new named species, the designated type strain of that species has to have lower than 70% DDH/95% ANI compared to the type strain of the most closely related already named species and experimentally determined phenotypes that distinguish it from the type strains of already named closely related species. However, the breadth of genomic and phenotypic diversity is not consistent between species. For example, *Bacillus anthracis* is a monomorphic species of which all members are very closely related, phenotypically very similar, and occupy the same ecological niche [27, 28]. while *Escherichia coli* is a heteromorphic species of which members are more distantly related, phenotypically diverse, and adapted to many different ecological niches [29] . Therefore, using a type-centered taxonomic scheme may not be problematic for monomorphic species, but it is not adequate for species with high genetic and phenotypic diversity.

Intraspecific (Infraspecific) taxa

Taxonomy includes one type of taxon below the species level, which is called “sub-species”. Sub-species have higher than 95% ANI to each other but experimentally distinguishable phenotypes.

Within species and sub-species, there are several types of taxa that are mostly used for pathogenic species, for example, pathovar (a group of plant pathogenic strains that cause the same disease on the same range of hosts [30]), sequevar (a group of plant pathogenic strains characterized by a specific DNA sequence), biovar (a group of strains that are physiologically and/or biochemically distinct [31]), or serovar (a group of strains characterized by the same set of antigens [32]).

Besides these taxa, there are purely phylogeny-based intraspecific taxa that simply circumscribe any monophyletic group of organisms. Commonly used terms include clade, phylogroup, and phylotype.

Non-classical taxa used as alternatives to species

The current pragmatic species concept encourages polyphasic classification based on a combination of genetic and phenotypic data [21]. Because it is much easier and faster today to sequence microbial genomes than to phenotypically characterize microbes, purely bioinformatics approaches based on different measurements of genome similarities have been proposed to cluster bacteria into groups independently of phenotypic data. These groups sometimes correspond to validly published named species but often reveal groups that have not yet been described as species.

In 2015, Varghese *et al*. proposed the concept of "clique" for delineating microbial taxa based on a combined measurement of whole genome average nucleotide identity (gANI) and alignment fraction (AF) between a pair of bacterial genomes. Species-level AF and gANI cut-offs cluster bacterial genomes into "cliques", which are complete graphs where all genomes are connected with each other, and "clique groups", which are cliques connected to each other with genomes in common. Based on the homogeneity and heterogeneity of the bacteria in a clique or a clique group, species are split into four categories: single homogeneous species, multiple homogeneous species, single heterogeneous species, and multiple heterogeneous species. A single homogeneous species is a clique or a clique-group containing members of the same species. A multiple homogeneous species refers to the case when the members of a species are spread throughout multiple cliques and/or clique-groups and every clique and/or clique-group only has genomes of that species. Single heterogeneous species is a clique or a clique-group that has members of multiple species. Multiple heterogeneous species refers to a group of cliques or clique-groups having genomes from multiple species in each of them. The clique concept has been validated to be consistent with species using the majority of examined genomes from 3082 well-classified species. However, genomes from 17.7% of the examined species disagree with clique-groups [33].

Another non-classical taxon concept is LINgroup. LINgroup is a derivative of the Life Identification Numbers (LIN) system that labels every bacterial genome with a unique LIN [34, 35] . A LIN consists of 20 positions where each of them represents an ANI cutoff, from as low as 70% at the leftmost position, A, to as high as 99.999% at the rightmost position, T. The genomic coherence of a group of bacteria can be interpreted by the degree of similarity represented by the length of LIN prefix they share, and this group is called a LINgroup, denoted by the shared LIN prefix [34, 35]. Unlike the classical species description which defines a species by a type strain, a LINgroup can described a species based on the pairwise genomic relatedness among all members of the species with the members sharing the same LIN prefix. The same principle also applies to LINgroups that correspond to other groups of related bacteria that have characteristics in common, for example, genera, intraspecific taxa, and even strains that cause a single disease outbreak.

Next, we describe the current methods used in WGS-based classification and identification.

Genome-based classification and identification methods of bacteria

Average Nucleotide Identity (ANI)

Average Nucleotide Identity, developed in 2004, is a measurement of bacterial whole genome similarity that approximates DNA-DNA Hybridization (DDH), which was the gold-standard measurement of species classification and identification since the 1960s with a DDH value of 70% being the speciation threshold [15]. For a pair of bacterial genomes, one genome, regarded as the query genome, is cut into consecutive 1020nt fragments. These fragments are aligned to the other genome, the subject, with BLASTN [36] . Alignments with ≥30% identity and ≥70% coverage are retained, and the average percent identity of these alignments is calculated as the ANI. ANI correlates well with evolutionary distances in the range from 70% to 100% whereby 95%-96% ANI corresponds to 70% DDH. However, it does not suggest clear genus boundaries for higher ranks [37].

Average Amino acid Identity (AAI)

Average Amino acid identity measures the genetic relatedness by comparing the conserved protein-coding genes of a pair of bacterial genomes with TBLASTN [15, 38]. AAI is able to differentiate more distantly related microbial populations (genus-level and higher) instead of ANI because nucleotide level measurements do not provide enough resolution at these levels [37].

MLST, ribosomal MLST (rMLST), and core genome MLST (cgMLST)

Before the advent of next generation sequencing, MLST was the gold-standard for conducting phylogeny-based bacterial classification and identification [6, 39]. rMLST is a variation of MLST that expands the number of analyzed genes to 53 ribosomal protein-coding genes conserved among all microbes allowing to use a single scheme to construct the Tree of Life of microbes [40]. However, due to the relatively small number of genes used to construct the phylogeny, the resolution of rMLST-based classification and identification is limited to the species level. Similarly to rMLST, the abovementioned Genome Taxonomy Database (see GTDB below) uses a single set of genes shared by most microbes but uses protein sequences instead of DNA sequences to identify phylogenetic relationships [23].

cgMLST instead determines phylogenies based on the core genome, *i.e.*, all conserved protein-coding genes shared by the analyzed genomes [41]. cgMLST thus provides high resolution at the intraspecific level but phylogenies constructed from the core genomes of different genomes cannot be directly compared with each other.

Min-wise independent permutations locality sensitive hashing (MinHash)

MinHash was initially developed to detect duplicated Web pages and deployed in the search engine AltaVista [42]. It estimates how similar two documents are by first hashing each document into a subset of lowest occurrence words and then calculating the Jaccard similarity as the proportion of the co-occurrence words in the union of the subsets. In 2016, MinHash was implemented to determine similarity between genomes [43]. It transforms genome sequences into a number (*n*) of fixed-length (*k*) k-mers, which occur the least time in a genome and calculates Jaccard similarity between the k-mer sets of genome sequences to approximate their genome similarity. It has been shown that, with appropriate selection of *n* and *k*, MinHash is able to cluster bacteria at the species level. Furthermore, as an alignment-free method, MinHash outperforms other alignment methods such as ANI, AAI, rMLST, and cgMLST in speed significantly [43]. In 2018, another implementation of MinHash was developed to approximate ANI that replaces BLAST in ANI calculations (see FastANI [36, 44, 45] below).

Single Nucleotide Polymorphism (SNP)-based methods

SNP-based methods are usually used for strain typing closely related bacteria from the same species [46]. The relationship among the bacteria can be determined by aligning their DNA sequences to a reference genome to detect variable positions, which form a SNP matrix and hence infer their phylogeny. SNP-based methods are accurate for differentiating bacteria with a limited number of SNPs, for example, strains of the same species that caused different outbreaks [8] . But SNP-based methods are not suitable for comparing distantly related bacteria with more than a few hundred SNPs. The basic procedure used in SNP-based methods is similar: 1) map raw reads to a single reference genome, 2) extract mapping information 3) summarize SNPs at different positions, 4) create an SNP matrix, and 5) infer the phylogenetic tree. The selection of the reference sequence is critical in SNP-based methods in order to achieve the optimal consensus phylogenetic tree. The choice of aligner brings up performance differences between different methods [47-49].

Tools, databases, and platforms for genome-based classification and identification

Here we discuss 1. a list of standalone tools, which need to be installed on a local machine with or without dependencies and do not require a network connection to server-side databases, 2. a series of platforms and databases that serve as sources of genomic data and metadata, and 3. Web services that process and compute data uploaded by users without the need to install dependencies.

National Center for Biotechnology Information (NCBI)

NCBI's prokaryotic RefSeq genomes are a collection of selected and high-quality genome assemblies from GenBank [50]. In May 2019, there were more than 31 million nucleotide accessions in the RefSeq and almost 1 billion WGS sequences in GenBank. Although NCBI has a large pool of genomic data, no whole genome classification system or identification pipeline is embedded. Still, NCBI plays an important role in data collection and data dissemination and facilitates the development of third party tools, databases, and platforms for classification and identification.

Integrated Microbial Genomes and Microbiomes (IMG/M)

IMG/M is a platform that aims to support the annotation, analysis and distribution of the microbial data sequenced at the Joint Genome Institute (JGI) as well as microbial data submitted in agreement with the IMG/M data release policy and submission standard [51]. At the time of writing, there were more than 70,000 bacterial genome assemblies including finished assemblies, draft assemblies, and permanent draft assemblies. Besides the classical taxonomic information submitted along with genome assemblies, IMG/G provides information about bacterial groups based on their genomic relatedness by classifying them into "cliques" and "clique-groups" based on their pre-computed pairwise alignment fraction

(AF) and average nucleotide identity (ANI). However, the “cliques” and “clique-groups” cannot be described or associated with any known taxa, and there is no identification function or pipeline implemented in IMG/M, so the “clique” system cannot be directly used to benefit the classification and identification of bacteria. JGI also provides a bioinformatics software suite called BBTools [52] to process genomic data for the purpose of comparing genome similarity and identifying bacterial isolates (See BBTools below).

Genome-to-genome distance calculator (GGDC)

GGDC uses a BLAST-based method to simulate DDH of a pair of bacterial genomes. Like traditional DDH, this digital-DDH (dDDH) uses a speciation threshold of 70%, and a recommended subspecies threshold of 79%-80% [53, 54].Traditional wet lab-based DDH, although limited by its portability and laborious process, was thought to have the potential to infer the phylogeny of all taxonomic ranks [55]. dDDH is an in-silico DDH replacement that shows a high correlation with DDH and overcomes the weaknesses of lab-based DDH.

JSpecies and JSpeciesWS

JSpecies is a Java-based standalone software that computes the pairwise ANI of bacterial genomes [56]. It provides both a command line interface (CLI) and a graphical user interface (GUI). Given a collection of whole genome sequences, JSpecies calculates the pairwise ANI with either BLASTN [36] or MUMmer [57-59] based on the user’s choice. JSpeciesWS is a Web server that implements JSpecies for pairwise genome comparison [60]. Besides calculating pairwise ANI of up to 15 bacterial genomes, either uploaded by the user or selected from the JSpecies reference database called GenomesDB, JSpecies also performs species-level identification with Tetra Correlation Search (TCS) based on the Tetra-nucleotide signature correlation index, an alignment-free approximation of ANI.

ANI calculator

ANI calculator is a Web-based tool that calculates one-way ANI and two-way ANI between two user-submitted bacterial genomes [61].

pyANI

pyANI is a standalone Python module that calculates ANI and other measurements generated alongside ANI, such as alignment length and coverage [62]. pyANI uses the same parameters as JSpecies to calculate ANI with either BLAST or MUMmer. Although it has been shown that there is no significant difference between reciprocal two-way ANI and one-way ANI [15], pyANI provides pairwise reciprocal two-way ANI of the provided set of bacterial genomes. pyANI also allows parallelization of pairwise ANI calculation that speeds up the analysis. Compared with JSpecies, pyANI is a faster tool to explore the genomic relationship of a large number of bacterial genomes with the built-in parallelization option, and it visualizes the genomic relationships with hierarchical clustering and heatmaps based on ANI.

Mash

Mash is a MinHash implementation to compare genomic sequences [43]. For genome comparison, it first transforms each genome into a “sketch”, which is a set of k-mers of the genome with the lowest occurrence. A mash distance derived from the Jaccard similarity and the k value is calculated to represent the dis-similarity of a pair of genomes. With pre-computed sketches, Mash allows real-time genome identification with the input of either genome assemblies or reads sequenced by a variety of platforms including Illumina MiSeq, PacBio RSII, and Oxford Nanopore MinION. As an alignment-free method with genomes hashed into smaller k-mer subsets, Mash is able to calculate the pairwise distance of 54,118

RefSeq genomes in 33 CPU hours, and has the ability for species-level classification of microbes [43]. However, the selection of sketch size and k value has a profound impact on computing speed, memory usage, and the accuracy of the computed genome relationships. There is no consensus in regard to the combination of sketch size and k value to achieve high-level accuracy and speed of classification and identification.

BBTools

BBTools is a suite of Java tools for processing DNA sequences developed by JGI, including tools to compare bacterial genome similarity with a MinHash procedure similar to Mash [52] . Other than the functions to compare local genomes, BBTools also enables genome identification in the online IMG/G database by submitting the query genome to the JGI server. The tool will return the similarities of the query genome to a list of reference genomes from the database. The identification result computes the similarity of the query genome to individual strains but does not provide an identification in regard to the named taxon the query may belong to. As a standalone software without a graphical user interface, BBTools is not sufficiently documented. Not enough guidance is provided to make the best use of its functionalities.

Sourmash

Sourmash is another MinHash implementation that also transforms bacterial genomes to “signatures” with provided k value and signature size, and then calculates Jaccard similarity of signatures to infer genomic relatedness [63]. Similar to Mash, Sourmash also takes either genome assemblies or DNA sequencing reads as the input to compare them with genome assemblies. Further, Sourmash provides the functionality to identify bacterial isolates by finding the last common ancestor (LCA) from the pre-built GenBank database or a custom database using user-selected bacterial genomes. With the GenBank database and custom databases, Sourmash also enables the identification of metagenomics data by first classifying the taxonomy in the database with the presence of marker k-mers and then quering metagenomics reads that are gathered as individual genomes.

FastANI

FastANI computes an approximation of ANI between bacterial genomes using the Mashmap algorithm based on MinHash and winnowing algorithms instead of BLASTN alignment [44, 45]. Like ANI, FastANI delineates species boundaries at 95% and is able to perform pairwise comparison 50 to 4600 times faster than BLASTN-based ANI. However, unlike the Jaccard similarity of MinHash that is able to provide high-level accuracy with higher k value and bigger sketch size, the value of FastANI is no longer consistent with BLASTN-based ANI if the k value is adjusted according to the formula provided in [45]. Therefore, the ability of FastANI to classify microbes beyond species level requires further investigation.

Bacterial Isolate Genome Sequence Database (BIGSdb)

BIGSdb is a MLST database system that is installed on a local or cloud server that can be accessed through a Web interface [64]. By uploading annotated genomes to BIGSdb, a user can identify loci to infer the evolutionary relationship of targeted bacterial isolates with MLST. Besides its use in determining phylogenetic relationships, BIGSdb is also useful as a Laboratory Information Management System (LIMS) to organize laboratory microbial genome data.

Genome Taxonomy Database (GTDB)

As mentioned above, MLST can be used to infer the phylogeny of analyzed bacteria with a selection of housekeeping genes with the limitation that the phylogeny is largely influenced by how many and which housekeeping genes are selected [6]. Therefore, GTDB uses a universal set of 120 ubiquitous genes as the MLST scheme to achieve accurate phylogeny when building the Tree of Life (bac120 tree) of 94,759 RefSeq release 80 genomes [23]. The bac120 tree implements bacterial taxonomy with respect of classification and nomenclature by discovering and validating the presence of polyphyletic groups, misclassifications, and inconsistent names. Although GTDB provides a standalone tool, GTDB-tk, to classify and identify bacterial isolates with the reference of the bac120 tree, it is not suitable for implementation in individual laboratories since it requires 90 GB memory to run, 25 GB disk space for the database, and 64 CPUs to enable the analysis of 1,000 genomes in 1 hour.

The Microbial Genome Atlas (MiGA)

MiGA is a Web server that classifies and identifies unknown isolates with the reference of existing taxonomy using a hierarchical hAAI/AAI/ANI scheme [65]. Since ANI has a speciation threshold of 95% and AAI provides resolution in differentiating distantly related bacterial genomes, MiGA uses the hierarchical hAAI/AAI/ANI scheme to predict in real-time to which taxonomic lineage a query belongs. Although the pipeline has been optimized to avoid all-to-all comparison, the speed is still compromised by the need of genome annotation for AAI calculation and is further slowed down by the process of genome assembly when DNA sequencing reads are uploaded instead of assembled genomes.

LINbase

LINbase is a Web service that implements the bacterial taxonomy in regard to classification, nomenclature, and identification with a genome similarity-based framework, the Life Identification Number® (LIN®) [35]. It collects bacterial genome assemblies published in public databases and uploaded by users along with the corresponding metadata and assigns each genome a LIN based on its ANI to its most similar genome in the database with a fast seed-and-extend algorithm with a combination of MinHash and ANI calculation. LINbase provides a user interface for users to describe LINgroups as taxa or groups of bacteria that share common phenotypes and unknown bacterial isolates can be identified as members of described LINgroups with fast algorithms. However, since LIN is assigned based on ANI with a limited working range from 70%-99.99%, which indicates inter-genus to almost strain-level, LINgroups cannot currently be used to either describe higher taxa, i.e. family, order, class, phylum, and domain, or individual disease outbreaks when outbreak strains only differ by a small number of SNPs.

EzBioCloud

EzBioCloud is an online database of 16S rRNA gene sequences and WGS that provides species-level genome identification of bacteria and archaea [66]. For a submitted whole genome assembly of an isolate, EzBioCloud first compares it with the type strains in the database by means of tetra-nucleotide, 16S rRNA, *recA,* and *gyrB* after genome annotation to find a list of phylogenetic neighbors. It then calculates ANI with all type strains in the list and finally determines whether the query belongs to an existing species or a novel species with the 95% ANI speciation threshold. With genomic data pre-structured by species, EzBioCloud achieves high speed identification by only comparing the isolate with the type strains. However, as discussed above about species described by one single type strain, the identification result may be less informative when the submitted isolate belongs to species with high diversity.

GenomeTrakr

GenomeTrakr is an open source foodborne pathogen genomic database hosted by the United States Food and Drug Administration, aiming to facilitate pathogen identification by collecting whole genome sequences and metadata such as geographical information of foodborne pathogen outbreaks [67]. GenomeTrakr enables public health officials and researchers to access the database and build up monitoring processes to detect pathogens and trace their sources. The aim is to eventually facilitate a real-time global surveillance system of foodborne pathogen outbreaks. Despite the ultimate goal of detection and traceback of foodborne pathogen outbreaks, GenomeTrakr does not provide an official or recommended bioinformatics pipeline or benchmarking reference genomes for the SNP-based methods.

Summary of strengths of available tools, databases, and platforms

With the development of Next Generation Sequencing technology, bacterial taxonomy can now use high-quality genomic data for classification and identification at a reduced cost. The aforementioned standalone tools are either using fast algorithms or taking advantage of powerful hardware to calculate genome similarity to quickly and precisely cluster bacterial genomes as taxa so that bacterial isolates can be identified as members of classified taxa with customized databases. Online databases and platforms are able to host large data sets with fast algorithms implemented for genomic data processing on high-performance servers and allow users to access the data and computational resources without the necessity of installing prerequisite dependencies. The identification functions, usually implemented with the discussed standalone tools, are able to identify an isolate within a few minutes. For databases and platforms with classification functions, the graphical user interfaces visualize genomically clustered taxa (cliques and clique-groups on IMG/G and LINgroup on LINbase, and the bac120 tree on GTDB) with a more straightforward view of the genomic relatedness among the members. Furthermore, LINbase offers the functionality for users to describe LINgroups as taxa based on shared phenotypes so that isolates can not only be identified as members of already published species or intra-specific taxa but also as members of any taxon newly described by the user.

It is promising that tools, databases, and platforms allow bacterial species classification and identification by replacing the lengthy biochemical tests with computational comparisons of genomic data.

Summary of weaknesses of available tools, databases, and platforms

There are three main aspects limiting the tools, databases, and platforms developed for bacterial taxonomy. One results from the inconsistency between classical taxonomy and groupings based on phylogeny and genomic similarity. Misclassifications and misnaming exist mostly due to historical reasons, especially the limits in technology that, back in time, did not allow researchers to differentiate different species that show similar phenotypes. Even though some of the misclassified bacteria can be re-assigned to the correct taxa, their names can hardly be corrected because they have been used in many fields, such as in medical practice, for many years. Correction would thus cause confusion.

The second limitation is in the genomic similarity measurement. As mentioned above, the computational measurement of genome similarity varies between alignment and alignment-free methods. Alignment-free methods cannot be used to reflect the phylogenetic and evolutionary relationship among organisms because they do not take the genetic variation brought by mutations of single nucleotides into consideration. Methods that involve sequence alignment such as MLST, ANI, and dDDH, although they reflect the phylogeny, are computationally too slow. Additionally, neither alignment-based nor alignment-free methods can be used as a universal measurement of genome similarity with a clear boundary at each taxonomic rank. Therefore, the tools, databases, and platforms are usually only able to classify and identify bacteria at the species level although resolution at the intraspecies level and higher ranks are necessary.

The third limitation is the maintenance and follow-up of the published tools, databases, and platforms. As mentioned above, taxonomy will be updated when misclassified bacteria are re-assigned to other taxa, hence the databases need to be kept up-to-date to avoid mis-identification of unknown isolates. Errors and bugs are often reported by users after a tool or online platform is published and its developers are responsible for fixing them. However, it is not rare that some tools and platforms are left unattended due to developers leaving or even abandonment of the project. Another common case is that online databases and platforms can no longer be visited, because their URLs have changed or become invalid without notifying the research community.

The future of genome-based taxonomy

Now that WGS has made phenotypic testing obsolete for identification purposes and WGS can more precisely assign isolates to taxa and outbreaks than any number of phenotypic tests, the phenotypic description of taxa could now focus on those phenotypes that are truly relevant to the biology of a taxon. In other words, instead of spending time describing dozens of microbial phenotypes that only a small number of taxonomists care about, we could spend time studying the phenotypes of new taxa that are relevant to the ecological niche the members of new taxa occupy, their evolution, adaptation, and life history, and, if applicable, their roles in human, animal, and plant health, and/or biotechnological potential. However, we are still in the middle of this paradigm shift and, so far, long lists of phenotypic data are still required for the purpose of publishing a new named species. Even more surprising is the fact that new named species are still published as traditional manuscripts instead of entries into a searchable species database. Such a database, the Digital Protologue Database (DPD), has been established but is neither broadly accepted by the taxonomic community nor commonly cited by microbe-related publications [68, 69].

Therefore, we agree with several taxonomists, such as [69, 70], and envision a future in which taxonomy becomes truly genome-based and database-driven.

Chapter 2 LINflow: A Method for Sustainable Microbial Whole Genome Similarity Calculation at Multiple Levels of Resolution

Long Tian[1], Lenwood S. Heath[2], Boris A. Vinatzer[1]

[1]School of Plant and Environmental Sciences, Virginia Tech, Blacksburg, VA 24061, USA,
[2]Department of Computer Science, Virginia Tech, Blacksburg, VA 24061, USA

Attribution

The following chapter is based upon the manuscript in preparation: Tian, L., Heath, L.S., Vinatzer, B.A. (2019). LINflow: A Method for Sustainable Microbial Whole Genome Similarity Calculation at Multiple Levels of Resolution.

Abstract

To learn how closely or distantly microbial strains are related to each other and whether they belong to the same taxon, genome similarity analysis is often performed on thousands of genomes by computing pairwise Average Nucleotide Identity (ANI) values. Pairwise comparisons are computationally expensive and each new genome that is added to a data set requires comparison to all other genomes. Here we introduce a method to approximate the genomic distances among a set of microbial genomes without the necessity of computing all pairwise ANI calculations. The genomic relationships are represented and stored in the form of Life Identification Numbers, a genome similarity-based system. Importantly, the addition of genomes to the data set does not require a reanalysis of the whole data set. This system in its current implementation can be scaled up to represent the genomic relatedness of tens of thousands of microbes with more than 70 times faster in runtime compared with pairwise ANI calculation. LINflow can be downloaded at https://github.com/LongTianPy/LINflow.git.

Introduction

The number of microbial genomes available at the National Center for Biotechnology Institute (NCBI) is growing rapidly and has reached 190,000 in 2019. It can be anticipated that many more genome assemblies will be published in the near future because of the reduced cost and the improved quality that can be obtained through next generation sequencing of microbial genomes. The large collection of microbial genomes provides the opportunity to explore relationships between species, boundaries of species, and the genetic diversity within species.

DNA-DNA hybridization (DDH) was the first method that incorporated genome content in microbial classification and was accepted as the gold standard. However, its low resolution, laborious experimental procedures, and very limited portability of results represent serious limitations [71]. Gene sequence-based methods have largely replaced DDH, since they are accurate and portable. 16S rRNA based phylogeny reconstruction determines the phylogenetic relationship of microbes by comparing their 16S rRNA gene sequences, which is a conserved gene in microbes. The conserved nature of the 16S rRNA gene enables its application in the classification at higher taxonomic ranks [72]. However, the cutoff for speciation is under debate and because of limited sequence variation at the species level and below the species level, 16S rRNA sequences are not informative of evolutionary relationships at lower taxonomic ranks [73]. Multi-locus Sequence Typing (MLST) uses a selection of house-keeping genes for phylogenetic reconstruction. Because of inconsistency in the selection of genes and the absence of sequence similarity thresholds for species delimitation, MLST is not used in taxonomy. It is rather used to reveal evolutionary relationships between strains of the same species. However, a variation of MLST, called rMLST, in which genes coding for conserved ribosomal proteins are used, has been explored for use in taxonomy [40, 41].

Average Nucleotide Identity (ANI) is a measure of the genomic similarity between microbes based on the comparison of whole genome sequences [15]. For a pair of microbial genomes, the query genome is cut into consecutive 1020 nt-long fragments, and each fragment is aligned to the subject genome using BLAST. Fragment alignments that have over

30% coverage and 70% identity are retained, and ANI is the the average identity of these alignments. 95% - 96% ANI has been suggested to be the speciation threshold corresponding to 70% DDH. ANI also provides the resolution necessary to differentiate phylogenetic groups of microbes within the same species *[15, 38]*. Although several variations of ANI calculation have been proposed, no breakthrough has been made to speed up the process as ANI is a sequence alignment-based metric [62, 66]. Therefore, using ANI alone to explore the relationship among all available microbial genomes will consume a large amount of computational power, which seldom laboratories or research institutes can afford.

MinHash was initially developed to detect duplicated Web pages, used by the search engine AltaVista [42]. In 2015, Ondov et al. published the first implementation of MinHash in microbial genome comparison, Mash [43]. It down-sizes a microbial genome to a sketch of a set of k-mers with the lowest appearance and compares two genomes by calculating the Jaccard similarity of their sketches [43]. Not only was it possible with this approach to process and calculate the pairwise similarity of 54,118 microbial genomes from NCBI RefSeq release 70 in 33 CPU hours, but the Jaccard similarity is also able to correlate with ANI almost linearly at the range of ANI >= 90%. Therefore, it can be used to cluster microbes as species [43]. FastANI, partly based on MinHash, was developed to approximate ANI. As another alignment-free method like MinHash, FastANI can be calculated 50 to 4000 times faster than ANI and shows high correlation to ANI when ANI>80% [44, 45]. In spite of the advantages of the aforementioned alignment-free methods, neither Jaccard similarity by MinHash nor FastANI indicates clear species boundaries. The k value, i.e. the length of k-mer, and the sketch size have significant impact on both Jaccard similarity and FastANI. Not only the values of both indices change along with different combinations of parameters, but their working ranges are also affected. A smaller k value enables the detection of low ANI but loses correlation to high

ANI, however, a higher k provides a high sensitivity when ANI is high, while low ANI is not detectable. A larger sketch size allows higher resolution when the microbial genomes that are being compared are very similar to each other, although it sacrifices computational time and space.

The Life Identification Number (LIN) system is a genome similarity-based system to classify individual microbes based on reciprocal ANI and to directly inform of the genomic relatedness between microbes based on the similarity between their LINs [34, 74, 75]. A LIN consists of a series of positions, where each position indicates an ANI threshold, from low to high, starting from the leftmost position. The LIN of a genome is assigned based on the ANI to its most similar genome whose LIN has been already assigned. Therefore, the more similar two microbial genomes are to each other, the longer their LINs are identical starting from the leftmost position. A group of microbes sharing the same leading part of LINs is called a LINgroup, denoted by the shared part of their LINs. It has been shown that LINgroups cannot only be used to describe microbes at the species level but also at intra-species levels [35].

To analyze the microbial diversity of a collection of microbial genomes, pairwise comparisons cannot be avoided by any of the above methods and implementations. However, when dealing with a large number of genomes, pairwise comparisons are computationally expensive and time-consuming. Furthermore, according to the growing trend of the number of genome assemblies in NCBI, frequent additions of new genomes to the existing data sets whose microbial diversities have been analyzed are inevitable, and the pairwise comparisons between the newly added genomes to all existing genomes need to be performed. The time and computational power consumption to do this is a challenge.

Here we alleviate the bottleneck of pairwise and unsustainable microbial diversity analysis by developing LINflow that uses LIN to represent and store the genomic relatedness of microbes and uses MinHash for fast identification of the most similar genome while assigning LIN. This method provides a chance to look into the phylogeny of a large scale of microbes not only at the species level but also at intra-specific levels with a minimized number of pairwise ANI calculations. LINflow can be downloaded at https://github.com/LongTianPy/LINflow.git.

Material and Methods

Overview

LIN here is the backbone of LINflow that represents and stores the relationship of bacterial genomes, therefore the assignment of LIN is the key part of LINflow. The LIN of a new genome is based on its most similar genome whose LIN has already been assigned. Here we use an implementation of MinHash, Sourmash [76], to quickly identify the most similar genome, and use pyani [62] to calculate average nucleotide identity. LIN has the nature of straightforwardly representing the genomic relatedness among a group of bacteria based on the number of positions shared by their LINs as a LINgroup. Such a LINgroup is denoted by the shared part of the LINs of its members, and the last position of the shared part of the LINs indicates the degree of the similarity between the members. In light of this, we use a seed-and-extend scheme to identify the most similar genome following a 2-step procedure (**Figure 2.1**). This 2-step procedure involves first identifies the LINgroup that the new genome belongs to using Sourmash and then identifies the most similar genome in the LINgroup. By default, LINflow uses the 95%-level LINgroup, and this can be modified by users based on the need of analyzing a specific data set.

Additionally, to make the result re-usable and easily accessible in terms of reading and writing, a relational database managed by SQLite is used to store data, with the schema shown in **Figure 2.2**. This relational database connects tables with primary and foreign keys, and the connections between tables are represented by arrows. The genome table stores the locations of the genome sequences. The taxonomy table stores the taxonomic information corresponding to each genome in the database. LIN schemes, based on which LINs are assigned, are kept in the Scheme table. Besides two default LIN schemes, new schemes can be added by users so that LINs can be assigned according to the users' needs in resolution. LINs are assigned with the two default schemes and with one user-defined scheme, if there is any.

Generation of signatures

The lowest occurrence k-mers of a genome consititute the signature in sourmash. It is generated with a given length of k-mer (*k*) and the size of signature (*n*). Sourmash allows a selection of multiple *k*s, here the parameters used are k=21 and 51 and n=2000.

Signature file system

The physical signature files of genomes are saved. The first member of each 95%-level LINgroup is regarded as the representative genome, and a copy of its signature file is also saved in a separate directory with signature files of all other representative genomes.

Initiation of LINflow

LINflow, by default, uses a 20-position LIN scheme that ranges from 70% ANI to 99.999% ANI to cope with genus to strain level differentiation (**Table 2.4**).A random genome will be selected to assign the first LIN with 0 in each of the 20 positions. Its signature will be generated and saved as the representative of the LINgroup of $0_A0_B0_C0_D0_E0_F$ both in the directories for representative genomes and this LINgroup.

LIN assignment

The new genome's (G_{Query}) signature S_{Query} will be first queried against the representative genomes of all existing 95%-level LINgroups (or F LINgroups) with k=21 using Sourmash. Based on analysis of more than 6000 bacterial genomes from different families, the Jaccard similarity of 0.2475 when k=21 was discovered to associate to 70% ANI (data not shown). If the highest Jaccard similarity J of one of the representative to S_{Query} is bigger than 0.2475, then the corresponding genome represents the 95%-level LINgroup ($L_{95\%}$) belongs to. If 0.0025 < J <= 0.2475, then the corresponding genome is at least 70% similar to G_{Query}, which means they share at least the A position in the current LIN scheme. If J <= 0.0025, the corresponding genome is lower than 70% similar to G_{Query}.

If J > 0.2475, S_{Query} will be queried against all the members of $L_{95\%}$ by Sourmash with k = 51. The most similar genome $G_{Subject}$ according to Jaccard similarity in $L_{95\%}$ is identified as the most similar genome to G_{Query} in the whole database.

If 0.0025 < J <= 0.2475, S_{Query} will be queried against all the members of $L_{95\%}$ by Sourmash with k = 21. The most similar genome $G_{Subject}$ according to Jaccard similarity in $L_{95\%}$ is identified as the most similar genome to G_{Query} in the whole database.

For the above cases, ANI between G_{Query} and $G_{Subject}$, ANI_{Query}, will be calculated with pyANI. To assign LIN, $G_{Subject}$'s LIN, $LIN_{Subject}$, will be used as the reference from A to the last position that the ANI threshold is lower than or equal to ANI_{Query}, the first position that ANI threshold is larger than ANI_{Query} will be assigned with a number that has never used with the prefix in the database, and the rest of the positions will be filled with 0s. For example, ANI_{Query} = 95.4575%, it is over 95% at F position but lower than 96% at G position, so it will use $LIN_{Subject}$ from A to F as LIN_{Query}'s A to F. At LIN_{Query} 's G position, a number that has never been used together with $LIN_{Subject}$'s prefix from A to G will be assigned. Each of LIN_{Query}'s H to T positions will be assigned with a 0.

If J <= 0.0025, no genome in current database is over 70% ANI to G_{Query}, so that a new number that has never been used in A position will be assigned to LIN_{Query}'s A position, and the rest of will be filled with 0s.

Update of database and signature file system

G_{Query}, $G_{Subject}$, ANI_{Query} and LIN_{Query} will all be written in the database. If LIN_{Query} creates a new 95%-level LINgroup, a new directory for this LINgroup will be created and S_{Query} will be saved in this directory as a member and as a representative genome with other representative genomes, otherwise S_{Query} will be only saved in the existing 95%-level LINgroup it belongs to.

Data

Here we compared the performances of sourmash, pyANI, FastANI, and LINflow by evaluating their runtime and accuracy when computing pairwise genomic similarity of 248 whole genome sequences belonging to the genus *Pseudomonas* (**Supplementary Table 1**) with parameters shown in **Table 2.1.** Among the 248 genomes, 222 are from 46 named species, including *Pseudomonas aeruginosa* and *Pseudomonas syringae sensu lato*, which is a group of species that can be classified as *Pseudomonas syringae*, and the rest 26 genomes are from *Pseudmonas sp..* These 248 genomes are split into 2 data sets: **data set A** consists of 247 genomes and was used to evaluate the computational speed, memory usage, and accuracy of the aforementioned software analyzing large scale of genomes, and **data set B** with 1 genome was used to evaluate the above software's computational speeds of adding a new genome to already-analyzed data set A.

Result

Computational speed and memory usage

The CPU time of each software analyzing 247 genomes from data set A is shown in **Table 2.2**. “Sourmash compute” extracts k-mer signatures of the genomes, and the “compare” sub-command computes the pairwise Jaccard similarity of the pre-computed signatures. FastANI is split into indexing phase and compute phase, but no sub-command is provided to execute the 2 phases separately. Therefore, the execution time of FastANI here includes both phases. Sourmash shows the best speedup for genome comparison with more than 4000x faster than pyANI including the execution of “compute” sub-command. FastANI and LINflow are 90x and 73x speedup compared to pyANI, respectively. Although LINflow does not show advantages in speed when analyzing large amount of genomes of data set A compared with Sourmash and FastANI, it outperforms them by using the least memory. Furthermore, LINflow’s speed of adding a new genome is relatively consistent with the growth of database, suggested by its average processing time per genome (3.51 min) in **Table 2.2** is similar to the time cost shown in **Table 2.3**.

Accuracy

Since ANI provides the speciation threshold at 95% ANI, the results of Sourmash, FastANI and LINflow were compared with that of pyANI and each other. The two k values used in the Sourmash component of LINflow, k=21 and k=51, were used separately here by Sourmash to calculate genome similarities at different levels. pyANI calculated two-way ANI values for each pair of genomes, and reciprocal ANIs were computed with a customized script. LINflow assigned LINs to these genome with both default schemes: one is the same scheme LINbase is using [35] (see **Table 2.4** for the scheme and **Supplementary Table 1** for the result), the other has 300 positions ranging from 70% to 100% with 0.1% interval between neighboring positions. The LINbase scheme was used to assign LINs to the genomes and classify them as LINgroups. The latter scheme was used to approximate an ANI similarity matrix. With similarity matrices prepared, heatmaps were generated to reflect the genomic relatedness among the analyzed genomes.

Heatmaps derived from ANI matrices calculated or approximated by pyANI (**Figure 2.3**), FastANI (**Figure 2.4A**) and LINflow (**Figure 2.4B**) show the same species level (ANI ≥ 95%) clustering of the 247 Pseudomonas genomes as red square blocks along the diagonal. 5 major clusters are highlighted. Cluster 1 consists of genomes belong to *P. aeruginosa*, cluster 2 represents the species *P. chlororaphis*, clusters 3, 4 and 5 together with other genomes constitutes *P. syringae sensu lato*. LINflow not only classfied the genomes as species, but also distinguished intraspecific groups as LINgroups. The LIN prefixes denotes LINgroups show both intergroup relationship and intragroup genomic relatedness.

Sourmash is also able to perform species-level clustering with both k values of 21 (**Figure 2.5A**) and 51 (**Figure 2.5B**). It is suggested from the result that Jaccard similarity calculated with k=51 has a weaker association with low ANI values compared to k=21, for example, cluster 2, 4 and 5. For genomes highly similar to each other, for example genomes in cluster 5, k=51 shows a better accuracy than k=21 to differentiate them from each other.

To further evaluate the accuracy of LINflow, hierarchical clustering was performed on each of the similarity matrices calculated by pyANI, FastANI, LINflow, and Sourmash using k=21 and k=51. The resulted dendrograms of hierarchical clustering were then compared with each other and measured by Cophenetic correlation (**Table 2.5**) [77, 78]. Their correlations were also visualized by a heatmap (**Figure 2.6**). LINflow shows the highest correlation to pyANI, outperforming FastANI and Sourmash. Surprisingly, FastANI has the lowest correlation to pyANI, and tuning to higher k value in Sourmash did not lead to a higher correlation to pyANI.

Discussion

Here we have compared LINflow's performance of analyzing a large data set and adding a new genome to an already-analyzed data set with pyANI, Sourmash, and FastANI. These two tasks represent the research needs of exploring the phylogenetic relationship of large amount of microbial genomes and updating the previous result with newly available data.

LINflow's speed of analyzing each genome is consistent regardless of the size of data set. This is because for each genome, the LINflow algorithm involves only a one-time signature creation, at most two times of fast Sourmash signature comparisons, and a one-time two-way ANI calculation between the new genome and its most similar genome identified by Sourmash in the database. It can be estimated that LINflow will outperform FastANI in terms of speed and memory usage if a larger data set is analyzed, and the speedup compared with pyANI will be more significant.

Although all of the software are able to discover the intraspecific groups among the 247 genomes analyzed, only

pyANI and LINflow measure the genomic similarity by calculating or approximating ANI, and Sourmash approximates genomic similarity between bacteria with Jaccard similarity. Although all of these software and implementations have been used to assess the relationship between bacteria, especially at species level, little is known about whether they represent the same or similar phylogeny by comparing the phylogenetic or distance trees derived from the abovementioned software. The LIN framework, reflecting ANI between microbial genomes, has been proven to correlate perfectly with the core genome phylogenetic tree of *Pseudomonas syringae sensu lato* [75]. Here we have performed the hierarchical clustering algorithm to each of the similarity matrices and compared the resulted dendrograms by measuring their pairwise Cophenetic correlation and LINflow has the highest correlation to pyANI among the software and implementations. Since the result investigated the resemblance of different dendrograms and LINflow has the highest correlation to pyANI result, we can conclude that LINflow is able to reveal more accurate microbial genomic relationships than FastANI and Sourmash, not only at species level. Subspecies or intraspecies exist, for bacteriologists to describe groups of microbes within a species that shows distinctive phenotypes. LINflow, by classifying genomes as LINgroups with the current LIN assignment scheme, shows the ability of straightforwardly distinguish them from other intraspecific groups belonging to the same species and suggest the microbial diversity within each group. Therefore, LINflow has proven its ability to precisely analyze microbes for various purposes at multiple taxonomic levels, for example, species description and pathogen identification.

Additionally, LINflow stores data in an SQLite relational database that organizes genomic data and the corresponding metadata. SQLite is a SQL database that can be accessed from its command line interface (CLI), various application programming interfaces (APIs) of different programming languages, or graphical user interfaces (GUIs) such as SQLiteManager. Users can easily retrieve whole genome sequences for other analyses, *e.g.* comparative genomics or customized reference database, by querying the database with filters of taxonomic information and/or LINs.

After all, LINflow is a precise, fast, and memory-efficient tool to study microbial genomes at a large scale by giving comparable result with pyANI, Sourmash, and FastANI, and it also responds well to the emergence of WGS by minimizing the pairwise comparisons between the query genome and the whole existing database when newly sequenced or published genomes are ready to be added to an already-analyzed data set. Therefore, we expect that LINflow is able to facilitate all topics of microbiology research and help the management of laboratory data with its database.

Tables

Table 2.1 Software, sub-commands and parameters used to analyze 247 P*seudomonas* genomes. Multiprocessing is enabled in pyANI and FastANI.

Software	Parameters
pyANI	-m ANIb –worker 100
Sourmash	-k 21, 51 -n 2000
FastANI	-k 16 -t 100
LINflow	As in Methods

Table 2.2 Runtime and memory usage of each software and sub-command used to analyze 236 *Pseudomonas* genomes. Total CPU time is listed for pyANI and FastANI.

Software and sub-command	Total CPU time	Memory usage (GiB)
pyANI	60862 min 6 sec	5.8
Sourmash compute	15 min 3 sec	0.05
Sourmash compare	0 min 14 sec	1.4
FastANI	673 min 18 sec	12.3
LINflow	829 min 39 sec	0.6

Table 2.3 Runtime of each software and sub-command used to add a new genome to analyzed data set A.

Software and sub-command	CPU time
pyANI	258 min 23 sec
Sourmash compute	0 min 4 sec
Sourmash search	0 min 11 sec
FastANI	4 min 4 sec
LINflow	3 min 36 sec

Table 2.4 LIN assignment scheme of LINbase used to assign LINs by LINflow in this study.

ANI	70%	75%	80%	85%	90%	95%	96%	97%	98%	98.5%	99%	99.25%	99.5%	99.75%	99.9%	99.925%	99.95%	99.975%	99.99%	99.999
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T

Table 2.5 Pairwise Cophentic correlations of dendrograms based on similarity matrices calculated by pyANI, fastANI, LINflow, and Sourmash

	pyANI	LINflow	FastANI	Sourmash k=21	Sourmash k=51

pyANI	1				
LINflow	0.9954	1			
FastANI	0.6190	0.6216	1		
Sourmash k=21	0.8541	0.8467	0.4740	1	
Sourmash k=51	0.7877	0.7716	0.4321	0.9813	1

Figures

Figure 2.1 Workflow of LINflow. The flowchart of LIN assignment algorithm of LINflow.

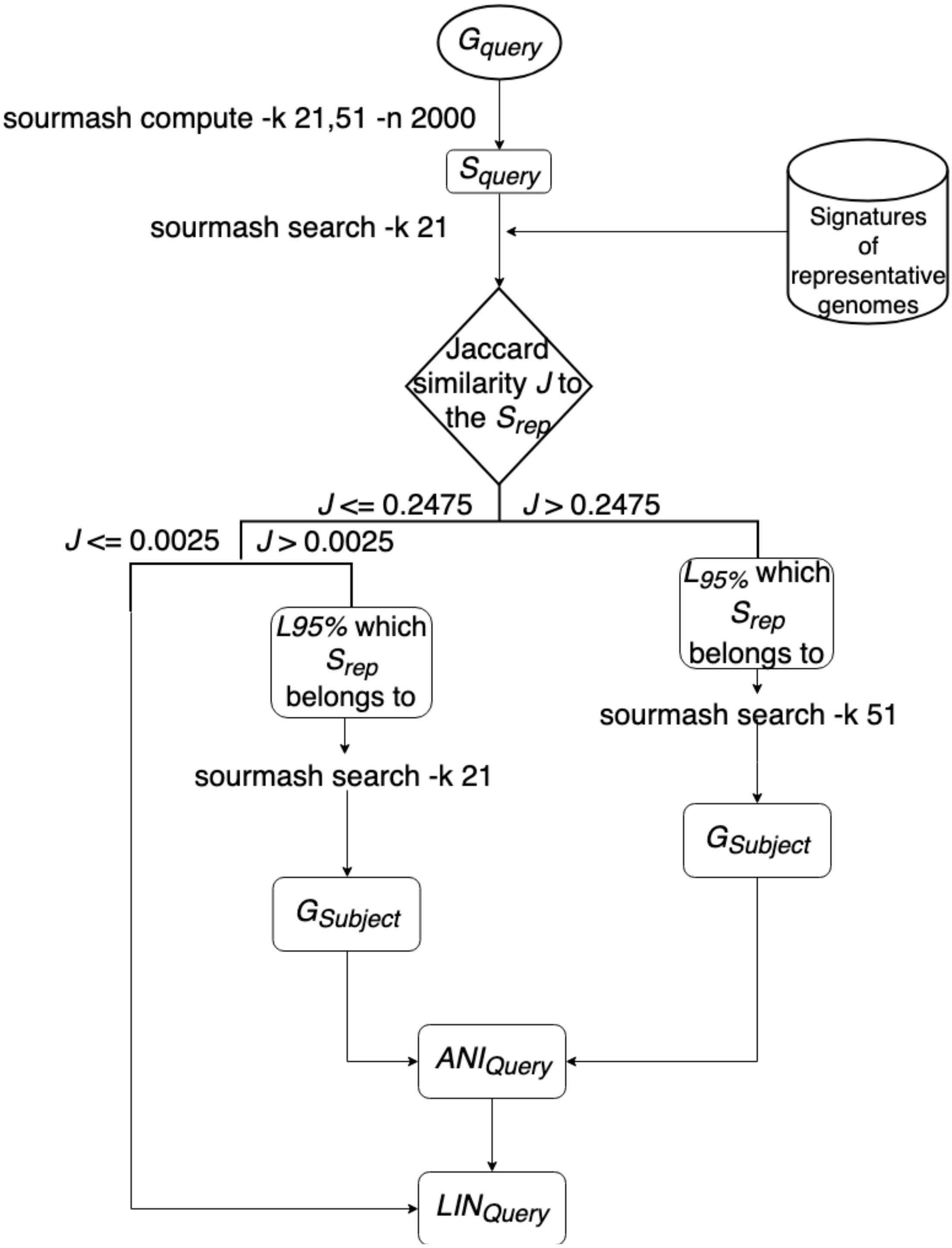
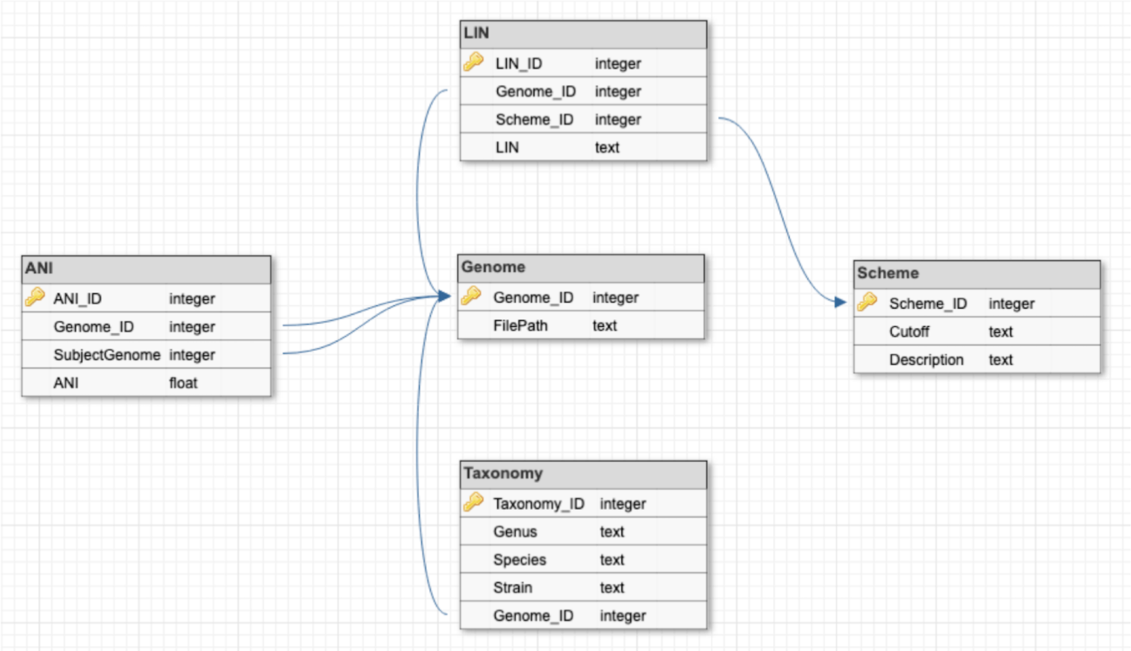


Figure 2.2 Database schema of LINflow.



LIN	
LIN_ID	integer
Genome_ID	integer
Scheme_ID	integer
LIN	text

ANI	
ANI_ID	integer
Genome_ID	integer
SubjectGenome	integer
ANI	float

Genome	
Genome_ID	integer
FilePath	text

Scheme	
Scheme_ID	integer
Cutoff	text
Description	text

Taxonomy	
Taxonomy_ID	integer
Genus	text
Species	text
Strain	text
Genome_ID	integer

Figure 2.3 Heatmap based on ANI matrix.

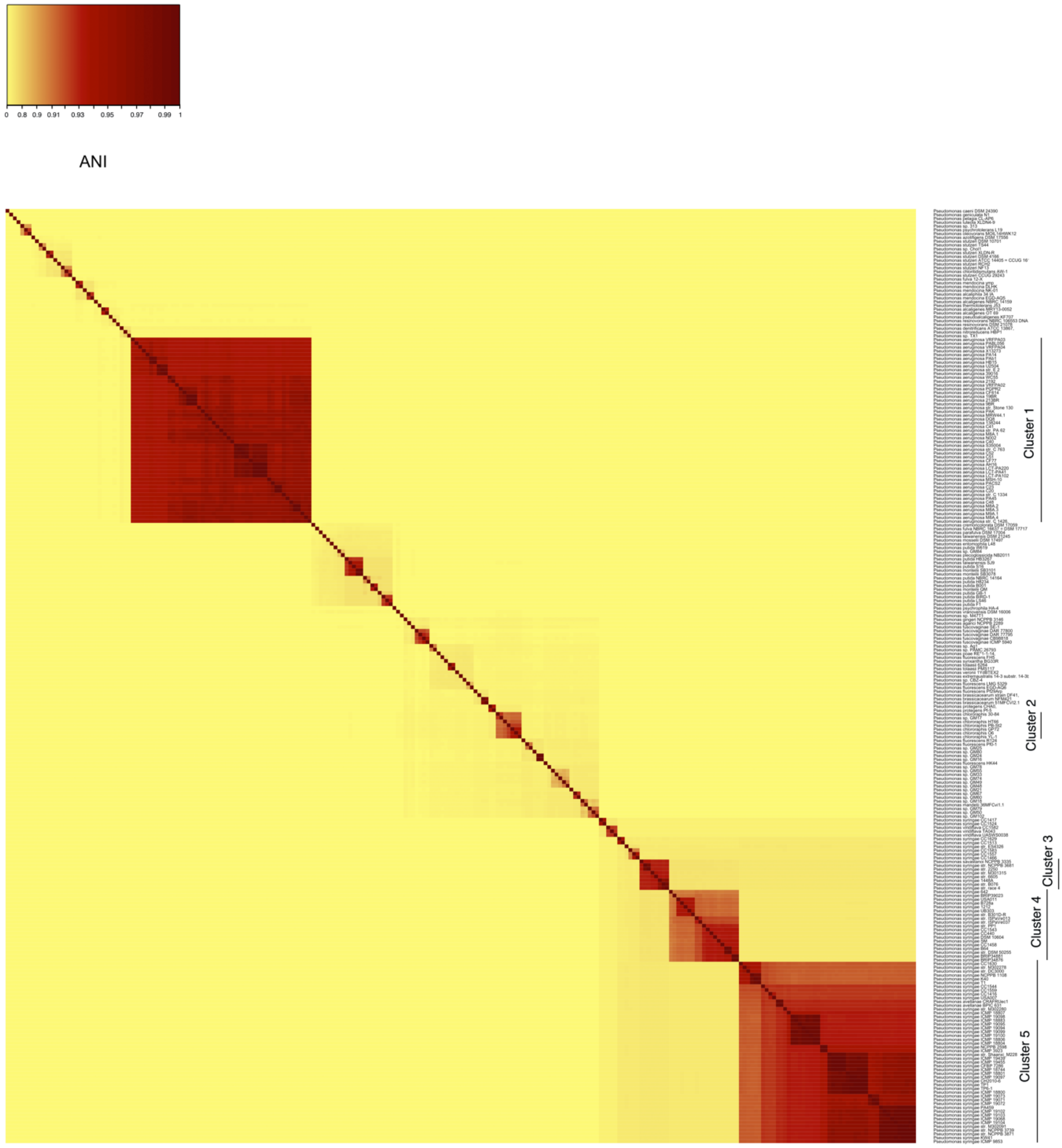
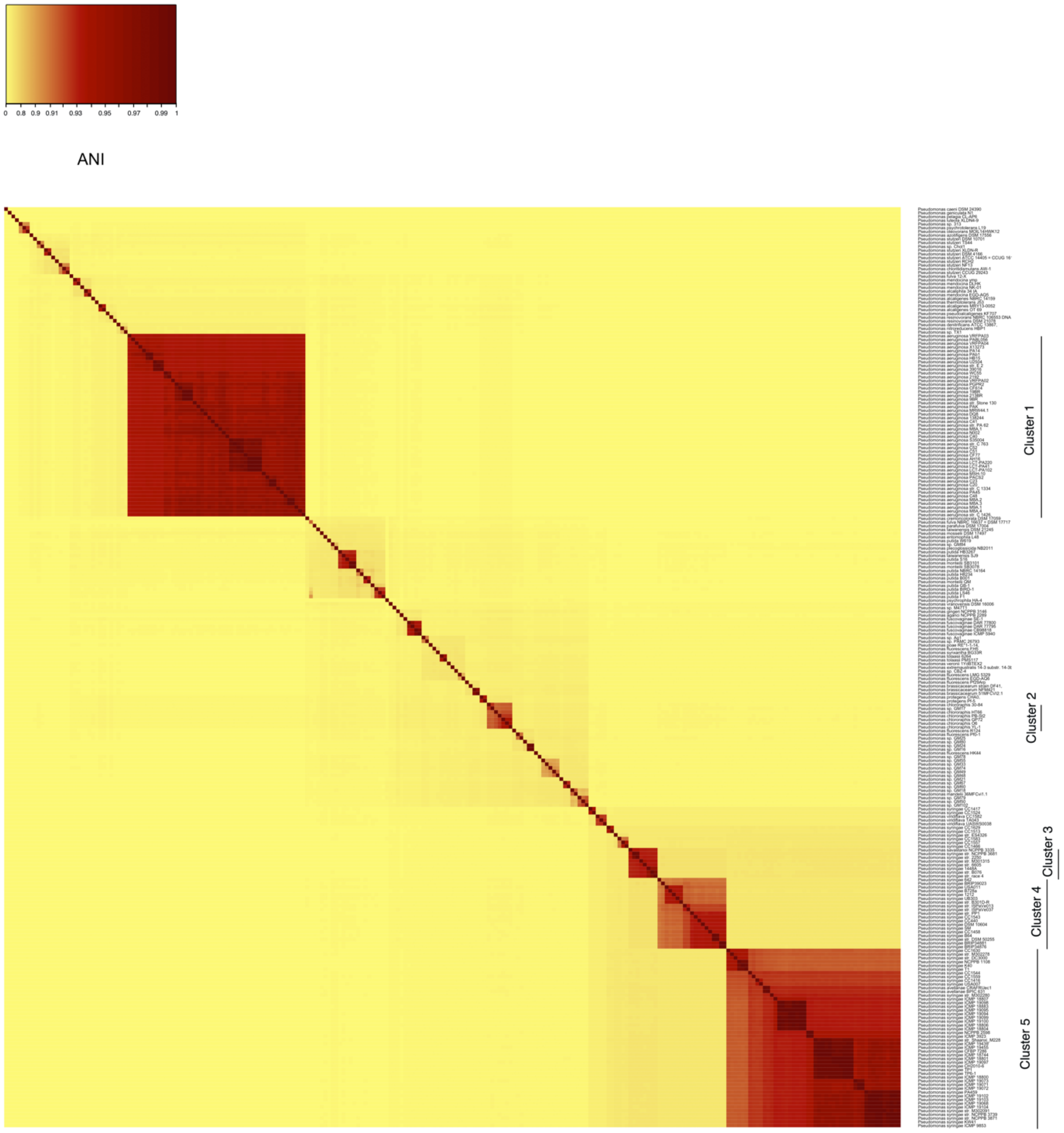


Figure 2.4 Heatmap based on approximated ANI matrix. A. ANI matrix calculated by FastANI. **B.** ANI matrix approximated by LINflow.

A



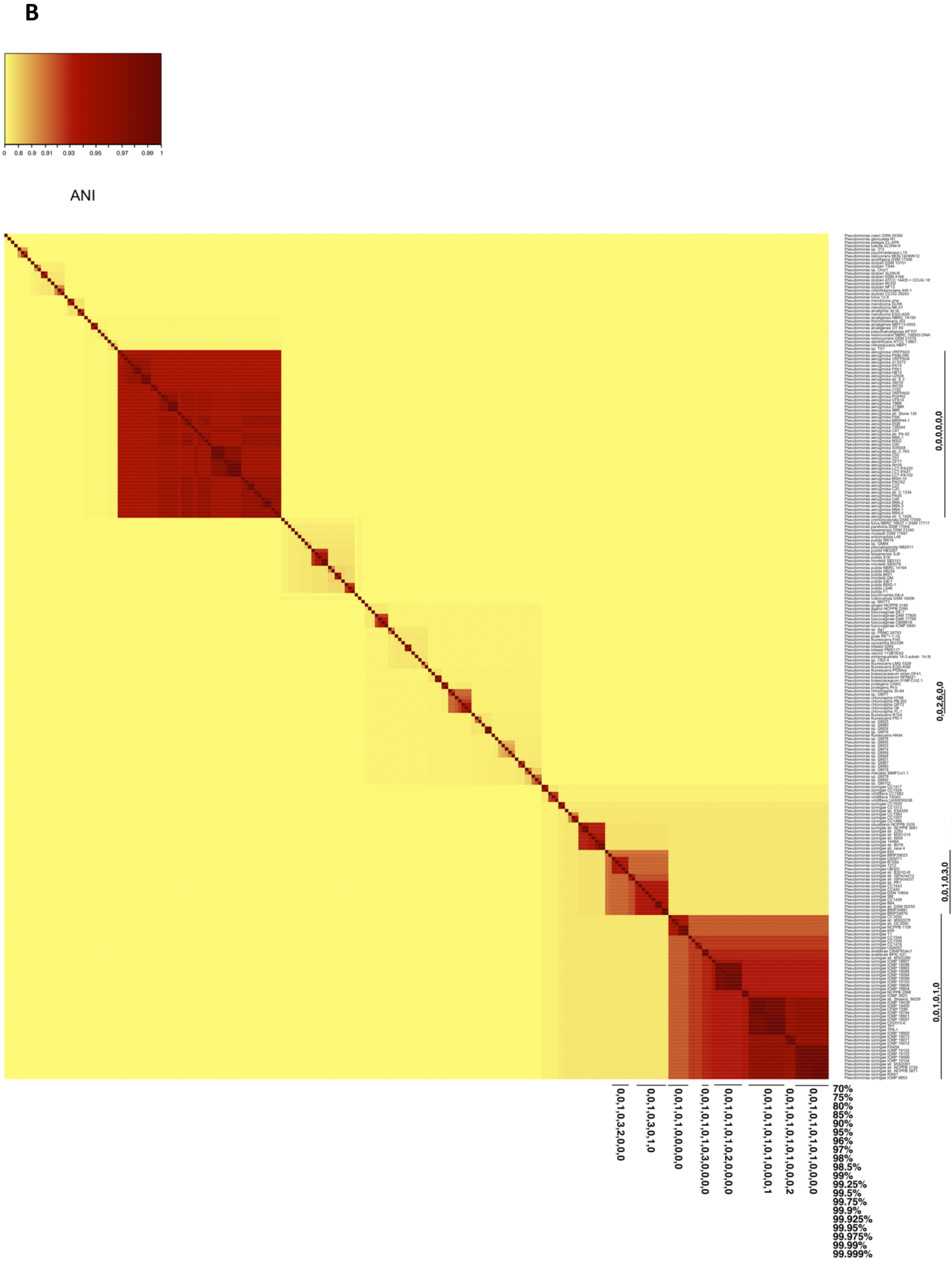
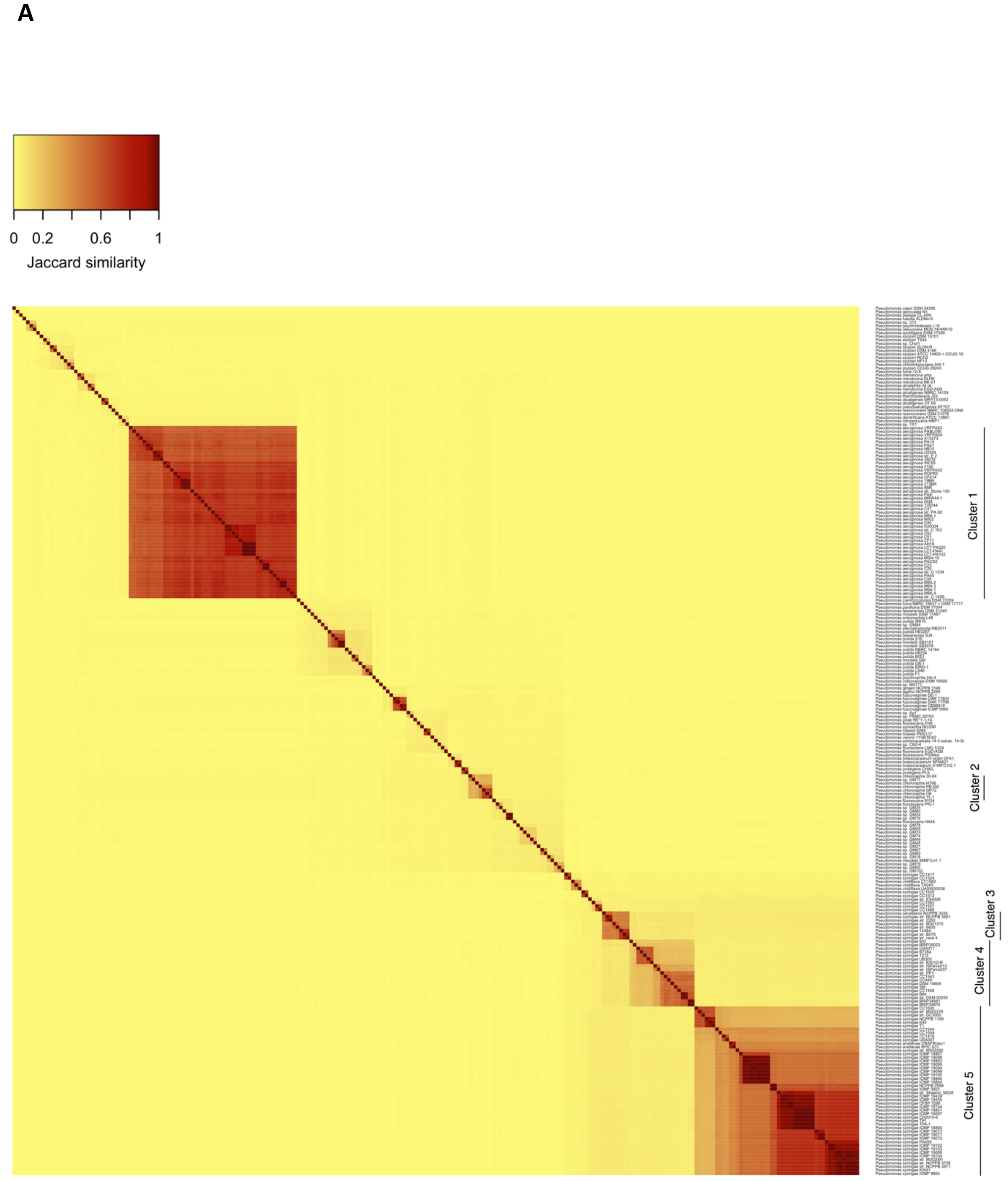

B
0
0.8
0.9
0.91
0.93
0.95
0.97
0.99
1
ANI
Pseudomonas caeni DSM 24390
Pseudomonas geniculata N1
Pseudomonas pelagia CL-AP6
Pseudomonas luteola XLDN4-9
Pseudomonas sp. 313
Pseudomonas psychrotolerans L19
Pseudomonas oleovorans MOIL14HWK12
Pseudomonas azotifigens DSM 17556
Pseudomonas stutzeri DSM 10701
Pseudomonas stutzeri TS44
Pseudomonas sp. Chol1
Pseudomonas stutzeri XLDN-R
Pseudomonas stutzeri DSM 4166
Pseudomonas stutzeri ATCC 14405 = CCUG 16'
Pseudomonas stutzeri RCH2
Pseudomonas stutzeri NF13
Pseudomonas chloritidismutans AW-1
Pseudomonas stutzeri CCUG 29243
Pseudomonas fulva 12-X
Pseudomonas mendocina ymp
Pseudomonas mendocina DLHK
Pseudomonas mendocina NK-01
Pseudomonas alcaliphila 34 IA
Pseudomonas mendocina EGD-AQ5
Pseudomonas alcaligenes NBRC 14159
Pseudomonas thermotolerans J53
Pseudomonas alcaligenes MRY13-0052
Pseudomonas alcaligenes OT 69
Pseudomonas pseudoalcaligenes KF707
Pseudomonas resinovorans NBRC 106553 DNA
Pseudomonas resinovorans DSM 21078
Pseudomonas denitrificans ATCC 13867,
Pseudomonas nitroreducens HBP1
Pseudomonas sp. TX1
Pseudomonas aeruginosa VRFPA03
Pseudomonas aeruginosa PABL056
Pseudomonas aeruginosa VRFPA04
Pseudomonas aeruginosa X13273
Pseudomonas aeruginosa PA14
Pseudomonas aeruginosa PAb1
Pseudomonas aeruginosa HB15
Pseudomonas aeruginosa U2504
Pseudomonas aeruginosa str. E 2
Pseudomonas aeruginosa 39016
Pseudomonas aeruginosa WC55
Pseudomonas aeruginosa 2192
Pseudomonas aeruginosa VRFPA02
Pseudomonas aeruginosa PGPR2
Pseudomonas aeruginosa CF614
Pseudomonas aeruginosa 19BR
Pseudomonas aeruginosa 213BR
Pseudomonas aeruginosa 9BR
Pseudomonas aeruginosa str. Stone 130
Pseudomonas aeruginosa PAK
Pseudomonas aeruginosa MRW44.1
Pseudomonas aeruginosa DQ8
Pseudomonas aeruginosa 138244
Pseudomonas aeruginosa C41
Pseudomonas aeruginosa str. PA 62
Pseudomonas aeruginosa M8A.1
Pseudomonas aeruginosa N002
Pseudomonas aeruginosa C40
Pseudomonas aeruginosa S35004
Pseudomonas aeruginosa str. C 763
Pseudomonas aeruginosa C52
Pseudomonas aeruginosa C51
Pseudomonas aeruginosa CF77
Pseudomonas aeruginosa AH16
Pseudomonas aeruginosa LCT-PA220
Pseudomonas aeruginosa LCT-PA41
Pseudomonas aeruginosa LCT-PA102
Pseudomonas aeruginosa MSH-10
Pseudomonas aeruginosa PACS2
Pseudomonas aeruginosa C23
Pseudomonas aeruginosa C20
Pseudomonas aeruginosa str. C 1334
Pseudomonas aeruginosa PA45
Pseudomonas aeruginosa C48
Pseudomonas aeruginosa M8A.2
Pseudomonas aeruginosa M8A.3
Pseudomonas aeruginosa M9A.1
Pseudomonas aeruginosa M8A.4
Pseudomonas aeruginosa str. C 1426,
Pseudomonas cremoricolorata DSM 17059
Pseudomonas fulva NBRC 16637 = DSM 17717
Pseudomonas parafulva DSM 17004
Pseudomonas taiwanensis DSM 21245
Pseudomonas mosselii DSM 17497
Pseudomonas entomophila L48
Pseudomonas putida W619
Pseudomonas sp. GM84
Pseudomonas plecoglossicida NB2011
Pseudomonas putida HB3267
Pseudomonas taiwanensis SJ9
Pseudomonas putida S16
Pseudomonas monteilii SB3101
Pseudomonas monteilii SB3078
Pseudomonas putida NBRC 14164
Pseudomonas putida H8234
Pseudomonas putida B001
Pseudomonas monteilii QM
Pseudomonas putida GB-1
Pseudomonas putida BIRD-1
Pseudomonas putida LS46
Pseudomonas putida F1
Pseudomonas psychrophila HA-4
Pseudomonas vranovensis DSM 16006
Pseudomonas sp. M47T1
Pseudomonas gingeri NCPPB 3146
Pseudomonas agarici NCPPB 2289
Pseudomonas fuscovaginae SE-1
Pseudomonas fuscovaginae DAR 77800
Pseudomonas fuscovaginae DAR 77795
Pseudomonas fuscovaginae CB98818
Pseudomonas fuscovaginae ICMP 5940
Pseudomonas sp. Ag1
Pseudomonas sp. PAMC 26793
Pseudomonas poae RE*1-1-14,
Pseudomonas fluorescens FH5
Pseudomonas synxantha BG33R
Pseudomonas tolaasii 6264
Pseudomonas tolaasii PMS117
Pseudomonas veronii 1YdBTEX2
Pseudomonas extremaustralis 14-3 substr. 14-3b
Pseudomonas sp. CBZ-4
Pseudomonas fluorescens LMG 5329
Pseudomonas fluorescens EGD-AQ6
Pseudomonas fluorescens Pf29Arp
Pseudomonas brassicacearum strain DF41,
Pseudomonas brassicacearum NFM421
Pseudomonas brassicacearum 51MFCVI2.1
Pseudomonas protegens CHA0,
Pseudomonas protegens Pf-5
Pseudomonas chlororaphis 30-84
Pseudomonas sp. GM17
Pseudomonas chlororaphis HT66
Pseudomonas chlororaphis PB-St2
Pseudomonas chlororaphis GP72
Pseudomonas chlororaphis O6
Pseudomonas chlororaphis YL-1
Pseudomonas fluorescens R124
Pseudomonas fluorescens Pf0-1
Pseudomonas sp. GM25
Pseudomonas sp. GM80
Pseudomonas sp. GM24
Pseudomonas sp. GM16
Pseudomonas fluorescens HK44
Pseudomonas sp. GM78
Pseudomonas sp. GM55
Pseudomonas sp. GM33
Pseudomonas sp. GM74
Pseudomonas sp. GM49
Pseudomonas sp. GM48
Pseudomonas sp. GM21
Pseudomonas sp. GM67
Pseudomonas sp. GM60
Pseudomonas sp. GM18
Pseudomonas mandelii 36MFCvi1.1
Pseudomonas sp. GM79
Pseudomonas sp. GM50
Pseudomonas sp. GM102
Pseudomonas syringae CC1417
Pseudomonas syringae CC1524
Pseudomonas viridiflava CC1582
Pseudomonas viridiflava TA043
Pseudomonas viridiflava UASWS0038
Pseudomonas syringae CC1629
Pseudomonas syringae CC1513
Pseudomonas syringae str. ES4326
Pseudomonas syringae CC1583
Pseudomonas syringae CC1557
Pseudomonas syringae CC1466
Pseudomonas savastanoi NCPPB 3335
Pseudomonas syringae str. NCPPB 3681
Pseudomonas syringae str. 2250
Pseudomonas syringae str. M301315
Pseudomonas syringae str. 6605
Pseudomonas syringae 1448A
Pseudomonas syringae str. B076
Pseudomonas syringae str. race 4
Pseudomonas syringae 642
Pseudomonas syringae BRIP39023
Pseudomonas syringae USA011
Pseudomonas syringae B728a
Pseudomonas syringae 1212
Pseudomonas syringae UB303
Pseudomonas syringae str. B301D-R
Pseudomonas syringae str. ISPaVe013
Pseudomonas syringae str. ISPaVe037
Pseudomonas syringae str. PP1
Pseudomonas syringae CC1543
Pseudomonas syringae CC440
Pseudomonas syringae DSM 10604
Pseudomonas syringae SM
Pseudomonas syringae CC1458
Pseudomonas syringae B64
Pseudomonas syringae str. DSM 50255
Pseudomonas syringae BRIP34881
Pseudomonas syringae BRIP34876
Pseudomonas syringae CC1630
Pseudomonas syringae str. M302278
Pseudomonas syringae str. DC3000
Pseudomonas syringae NCPPB 1108
Pseudomonas syringae K40
Pseudomonas syringae T1
Pseudomonas syringae CC1544
Pseudomonas syringae CC1559
Pseudomonas syringae CC1416
Pseudomonas syringae USA007
Pseudomonas avellanae CRAFRUec1
Pseudomonas avellanae BPIC 631
Pseudomonas syringae str. M302280
Pseudomonas syringae ICMP 18807
Pseudomonas syringae ICMP 19098
Pseudomonas syringae ICMP 18883
Pseudomonas syringae ICMP 19095
Pseudomonas syringae ICMP 19094
Pseudomonas syringae ICMP 19099
Pseudomonas syringae ICMP 19100
Pseudomonas syringae ICMP 18806
Pseudomonas syringae ICMP 18804
Pseudomonas syringae NCPPB 2598
Pseudomonas syringae ICMP 3923
Pseudomonas syringae str. Shaanxi_M228
Pseudomonas syringae ICMP 19438
Pseudomonas syringae ICMP 19455
Pseudomonas syringae CFBP 7286
Pseudomonas syringae ICMP 18744
Pseudomonas syringae ICMP 18801
Pseudomonas syringae ICMP 19097
Pseudomonas syringae CH2010-6
Pseudomonas syringae TP1
Pseudomonas syringae TP6-1
Pseudomonas syringae ICMP 18800
Pseudomonas syringae ICMP 19073
Pseudomonas syringae ICMP 19071
Pseudomonas syringae ICMP 19072
Pseudomonas syringae PA459
Pseudomonas syringae ICMP 19102
Pseudomonas syringae ICMP 19103
Pseudomonas syringae ICMP 19068
Pseudomonas syringae ICMP 19104
Pseudomonas syringae str. M302091
Pseudomonas syringae str. NCPPB 3739
Pseudomonas syringae str. NCPPB 3871
Pseudomonas syringae KW41
Pseudomonas syringae ICMP 9853
0,0,0,0,0,0
0,0,2,6,0,0
0,0,1,0,0,0
0,0,1,0,3,0
0,0,1,0,1,0
0,0,1,0,3,2,0,0,0
0,0,1,0,3,0,1,0
0,0,1,0,1,0,0,0,0,0
0,0,1,0,1,0,1,0,3,0,0,0,0
0,0,1,0,1,0,1,0,2,0,0,0,0
0,0,1,0,1,0,1,0,1,0,0,0,1
0,0,1,0,1,0,1,0,1,0,0,0,2
0,0,1,0,1,0,1,0,1,0,0,0,0
70%
75%
80%
85%
90%
95%
96%
97%
98%
98.5%
99%
99.25%
99.5%
99.75%
99.9%
99.925%
99.95%
99.975%
99.99%
99.999%

Figure 2.5 Heatmap based on Jaccard similarity matrix calculated by Sourmash. A. Jaccard similatity calculated with k=21. **B.** Jaccard similatity calculated with k=51.



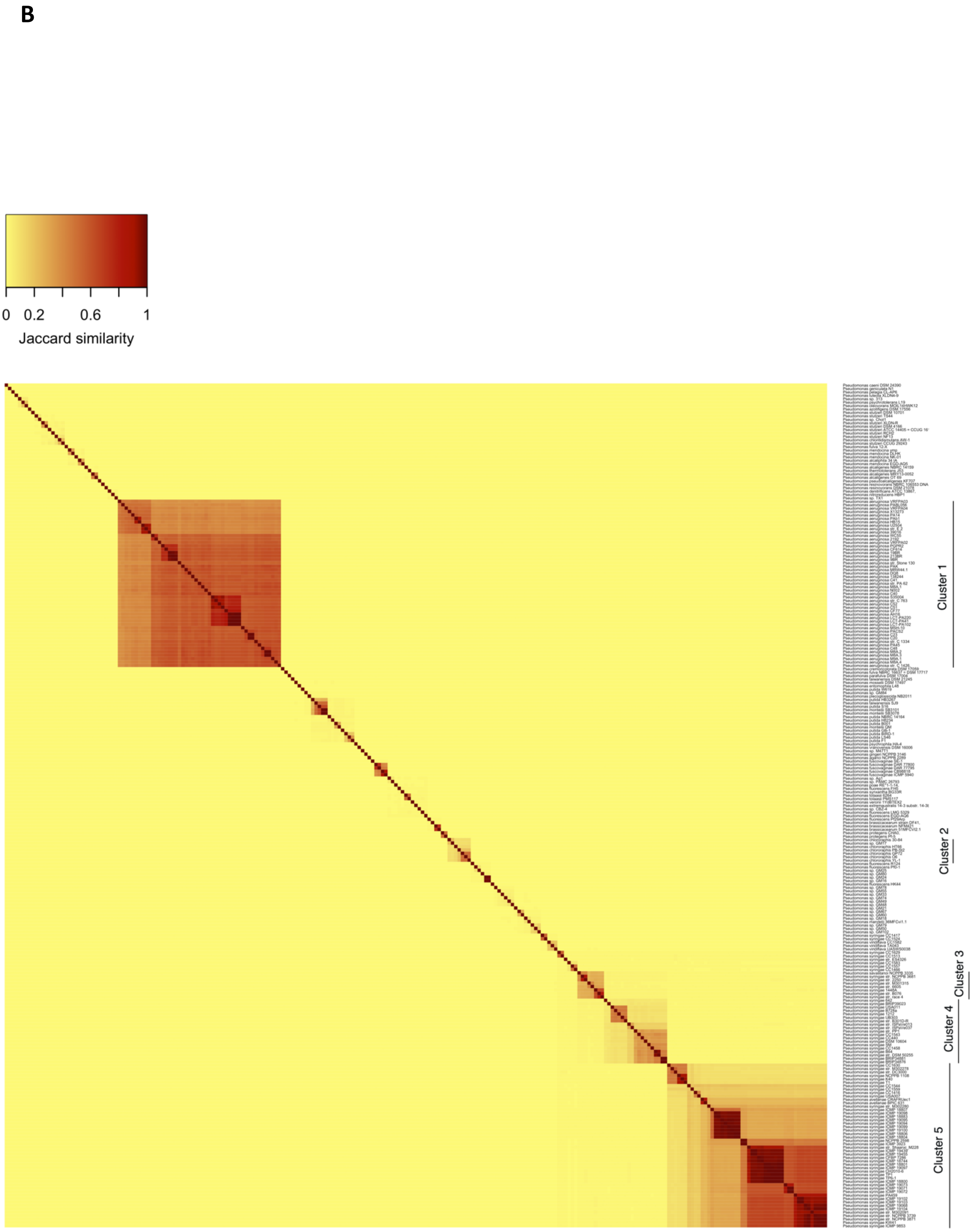
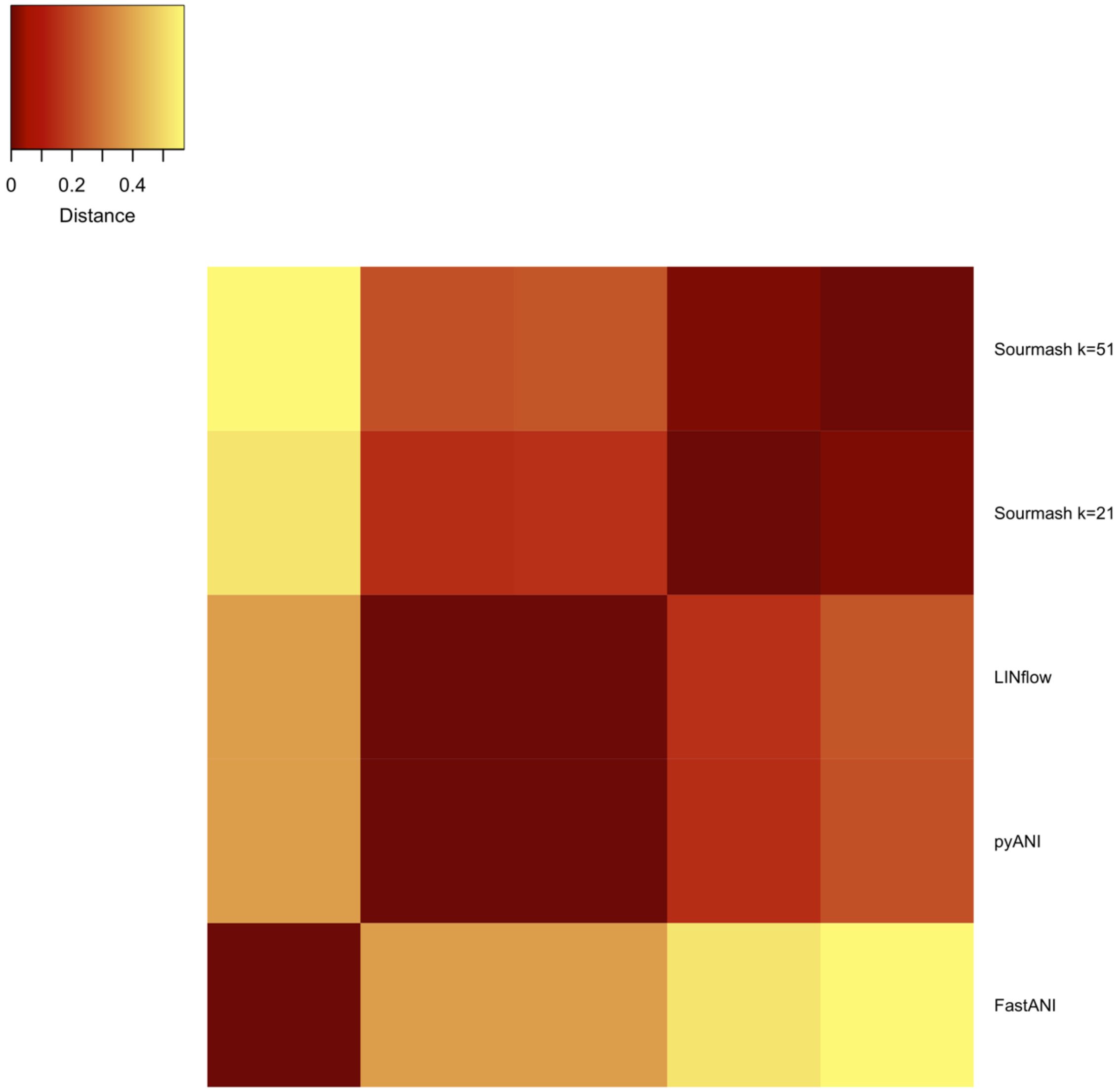

B
0
0.2
0.6
1
Jaccard similarity
Pseudomonas caeni DSM 24390
Pseudomonas geniculata N1
Pseudomonas pelagia CL-AP6
Pseudomonas luteola XLDN4-9
Pseudomonas sp. 313
Pseudomonas psychrotolerans L19
Pseudomonas oleovorans MOIL14HWK12
Pseudomonas azotifigens DSM 17556
Pseudomonas stutzeri DSM 10701
Pseudomonas stutzeri TS44
Pseudomonas sp. Chol1
Pseudomonas stutzeri XLDN-R
Pseudomonas stutzeri DSM 4166
Pseudomonas stutzeri ATCC 14405 = CCUG 16'
Pseudomonas stutzeri RCH2
Pseudomonas stutzeri NF13
Pseudomonas chloritidismutans AW-1
Pseudomonas stutzeri CCUG 29243
Pseudomonas fulva 12-X
Pseudomonas mendocina ymp
Pseudomonas mendocina DLHK
Pseudomonas mendocina NK-01
Pseudomonas alcaliphila 34 IA
Pseudomonas mendocina EGD-AQ5
Pseudomonas alcaligenes NBRC 14159
Pseudomonas thermotolerans J53
Pseudomonas alcaligenes MRY13-0052
Pseudomonas alcaligenes OT 69
Pseudomonas pseudoalcaligenes KF707
Pseudomonas resinovorans NBRC 106553 DNA
Pseudomonas resinovorans DSM 21078
Pseudomonas denitrificans ATCC 13867,
Pseudomonas nitroreducens HBP1
Pseudomonas sp. TX1
Pseudomonas aeruginosa VRFPA03
Pseudomonas aeruginosa PABL056
Pseudomonas aeruginosa VRFPA04
Pseudomonas aeruginosa X13273
Pseudomonas aeruginosa PA14
Pseudomonas aeruginosa PAb1
Pseudomonas aeruginosa HB15
Pseudomonas aeruginosa U2504
Pseudomonas aeruginosa str. E 2
Pseudomonas aeruginosa 39016
Pseudomonas aeruginosa WC55
Pseudomonas aeruginosa 2192
Pseudomonas aeruginosa VRFPA02
Pseudomonas aeruginosa PGPR2
Pseudomonas aeruginosa CF614
Pseudomonas aeruginosa 19BR
Pseudomonas aeruginosa 213BR
Pseudomonas aeruginosa 9BR
Pseudomonas aeruginosa str. Stone 130
Pseudomonas aeruginosa PAK
Pseudomonas aeruginosa MRW44.1
Pseudomonas aeruginosa DQ8
Pseudomonas aeruginosa 138244
Pseudomonas aeruginosa C41
Pseudomonas aeruginosa str. PA 62
Pseudomonas aeruginosa MBA,1
Pseudomonas aeruginosa N002
Pseudomonas aeruginosa C40
Pseudomonas aeruginosa S35004
Pseudomonas aeruginosa str. C 763
Pseudomonas aeruginosa C52
Pseudomonas aeruginosa C51
Pseudomonas aeruginosa CF77
Pseudomonas aeruginosa AH16
Pseudomonas aeruginosa LCT-PA220
Pseudomonas aeruginosa LCT-PA41
Pseudomonas aeruginosa LCT-PA102
Pseudomonas aeruginosa MSH-10
Pseudomonas aeruginosa PACS2
Pseudomonas aeruginosa C23
Pseudomonas aeruginosa C20
Pseudomonas aeruginosa str. C 1334
Pseudomonas aeruginosa PA45
Pseudomonas aeruginosa C48
Pseudomonas aeruginosa MBA,2
Pseudomonas aeruginosa MBA,3
Pseudomonas aeruginosa MBA,1
Pseudomonas aeruginosa MBA,4
Pseudomonas aeruginosa str. C 1426,
Pseudomonas cremoricolorata DSM 17059
Pseudomonas fulva NBRC 16637 = DSM 17717
Pseudomonas parafulva DSM 17004
Pseudomonas taiwanensis DSM 21245
Pseudomonas mosselii DSM 17497
Pseudomonas entomophila L48
Pseudomonas putida W619
Pseudomonas sp. GM84
Pseudomonas plecoglossicida NB2011
Pseudomonas putida HB3267
Pseudomonas taiwanensis SJ9
Pseudomonas putida S16
Pseudomonas monteilii SB3101
Pseudomonas monteilii SB3078
Pseudomonas putida NBRC 14164
Pseudomonas putida HB234
Pseudomonas putida B001
Pseudomonas monteilii QM
Pseudomonas putida GB-1
Pseudomonas putida BIRD-1
Pseudomonas putida LS46
Pseudomonas putida F1
Pseudomonas psychrophila HA-4
Pseudomonas vranovensis DSM 16006
Pseudomonas sp. M47T1
Pseudomonas gingeri NCPPB 3146
Pseudomonas agarici NCPPB 2289
Pseudomonas fuscovaginae SE-1
Pseudomonas fuscovaginae DAR 77800
Pseudomonas fuscovaginae DAR 77795
Pseudomonas fuscovaginae CB98818
Pseudomonas fuscovaginae ICMP 5940
Pseudomonas sp. Ag1
Pseudomonas sp. PAMC 26793
Pseudomonas poae RE*1-1-14,
Pseudomonas fluorescens FH5
Pseudomonas synxantha BG33R
Pseudomonas tolaasii 6264
Pseudomonas tolaasii PMS117
Pseudomonas veronii 1YdBTEX2
Pseudomonas extremaustralis 14-3 substr. 14-3t
Pseudomonas sp. CBZ-4
Pseudomonas fluorescens LMG 5329
Pseudomonas fluorescens EGD-AQ6
Pseudomonas fluorescens Pt29Arp
Pseudomonas brassicacearum strain DF41,
Pseudomonas brassicacearum NFM421
Pseudomonas brassicacearum 51MFCVI2.1
Pseudomonas protegens CHA0,
Pseudomonas protegens Pf-5
Pseudomonas chlororaphis 30-84
Pseudomonas sp. GM17
Pseudomonas chlororaphis HT66
Pseudomonas chlororaphis PB-St2
Pseudomonas chlororaphis GP72
Pseudomonas chlororaphis O6
Pseudomonas chlororaphis YL-1
Pseudomonas fluorescens R124
Pseudomonas fluorescens Pf0-1
Pseudomonas sp. GM25
Pseudomonas sp. GM80
Pseudomonas sp. GM24
Pseudomonas sp. GM16
Pseudomonas fluorescens HK44
Pseudomonas sp. GM78
Pseudomonas sp. GM55
Pseudomonas sp. GM33
Pseudomonas sp. GM74
Pseudomonas sp. GM49
Pseudomonas sp. GM48
Pseudomonas sp. GM21
Pseudomonas sp. GM67
Pseudomonas sp. GM60
Pseudomonas sp. GM18
Pseudomonas mandelii 36MFCvi1.1
Pseudomonas sp. GM79
Pseudomonas sp. GM50
Pseudomonas sp. GM102
Pseudomonas syringae CC1417
Pseudomonas syringae CC1524
Pseudomonas viridiflava CC1582
Pseudomonas viridiflava TA043
Pseudomonas viridiflava UASWS0038
Pseudomonas syringae CC1629
Pseudomonas syringae CC1513
Pseudomonas syringae str. ES4326
Pseudomonas syringae CC1583
Pseudomonas syringae CC1557
Pseudomonas syringae CC1466
Pseudomonas savastanoi NCPPB 3335
Pseudomonas syringae str. NCPPB 3681
Pseudomonas syringae str. 2250
Pseudomonas syringae str. M301315
Pseudomonas syringae str. 6605
Pseudomonas syringae 1448A
Pseudomonas syringae str. B076
Pseudomonas syringae str. race 4
Pseudomonas syringae 642
Pseudomonas syringae BRIP39023
Pseudomonas syringae USA011
Pseudomonas syringae B728a
Pseudomonas syringae 1212
Pseudomonas syringae UB303
Pseudomonas syringae str. B301D-R
Pseudomonas syringae str. ISPaVe013
Pseudomonas syringae str. ISPaVe037
Pseudomonas syringae str. PP1
Pseudomonas syringae CC1543
Pseudomonas syringae CC440
Pseudomonas syringae DSM 10604
Pseudomonas syringae SM
Pseudomonas syringae CC1458
Pseudomonas syringae B64
Pseudomonas syringae str. DSM 50255
Pseudomonas syringae BRIP34881
Pseudomonas syringae BRIP34876
Pseudomonas syringae CC1630
Pseudomonas syringae str. M302278
Pseudomonas syringae str. DC3000
Pseudomonas syringae NCPPB 1108
Pseudomonas syringae K40
Pseudomonas syringae T1
Pseudomonas syringae CC1544
Pseudomonas syringae CC1559
Pseudomonas syringae CC1416
Pseudomonas syringae USA007
Pseudomonas avellanae CRAFRUec1
Pseudomonas avellanae BPIC 631
Pseudomonas syringae str. M302280
Pseudomonas syringae ICMP 18807
Pseudomonas syringae ICMP 19098
Pseudomonas syringae ICMP 18883
Pseudomonas syringae ICMP 19095
Pseudomonas syringae ICMP 19094
Pseudomonas syringae ICMP 19099
Pseudomonas syringae ICMP 19100
Pseudomonas syringae ICMP 18806
Pseudomonas syringae ICMP 18804
Pseudomonas syringae NCPPB 2598
Pseudomonas syringae ICMP 3923
Pseudomonas syringae str. Shaanxi_M228
Pseudomonas syringae ICMP 19439
Pseudomonas syringae ICMP 19455
Pseudomonas syringae CFBP 7286
Pseudomonas syringae ICMP 18744
Pseudomonas syringae ICMP 18801
Pseudomonas syringae ICMP 19097
Pseudomonas syringae CH2010-6
Pseudomonas syringae TP1
Pseudomonas syringae TP6-1
Pseudomonas syringae ICMP 18800
Pseudomonas syringae ICMP 19073
Pseudomonas syringae ICMP 19071
Pseudomonas syringae ICMP 19072
Pseudomonas syringae PA459
Pseudomonas syringae ICMP 19102
Pseudomonas syringae ICMP 19103
Pseudomonas syringae ICMP 19068
Pseudomonas syringae ICMP 19104
Pseudomonas syringae str. M302091
Pseudomonas syringae str. NCPPB 3739
Pseudomonas syringae str. NCPPB 3871
Pseudomonas syringae KW41
Pseudomonas syringae ICMP 9853
Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster 5

Figure 2.6 Heatmap showing the distances among dendrograms based on the similarity matrices calculated by pyANI, FastANI, LINflow, and Sourmash. The hierarchical clustering algorithm is performed to each similarity matrix and dendrogram was generated. After calculating pairwise Cophenetic correlations of the dendrograms, their pairwise distances was derived by 1-Cophenetic correlation, and the heatmap visualizes the distances of the dendrograms.



Chapter 3 LINbase: A Web Service for Genome-Based Identification of Microbes as Members of Crowdsourced Taxa

Long Tian[1], Chengjie Huang[2], Lenwood S. Heath[2], Boris A. Vinatzer[1]

[1]School of Plant and Environmental Sciences, Virginia Tech, Blacksburg, VA 24061, USA,
[2]Department of Computer Science, Virginia Tech, Blacksburg, VA 24061, USA

Attribution
The following chapter is based upon the manuscript in preparation: Tian, L., Huang, C., Heath, L.S., Vinatzer, B.A. (2019). LINbase: a Web service for genome-based identification of microbes as members of crowdsourced taxa.

Abstract

The development of next generation and third generation DNA sequencing technologies in combination with new efficient algorithms allows scientists to economically, quickly, and precisely identify microbes at all taxonomic levels and even attribute pathogen isolates to specific disease outbreaks. However, current taxonomic practice has not kept up with the sequencing revolution and still relies on cumbersome journal publications to describe new species. Here we introduce a Web service that allows any user to genomically circumscribe any monophyletic group of bacteria as a taxon and associate with each taxon a name and short description. Any other user can immediately identify their unknown microbe as a member of any of these crowdsourced taxa using gene or genome sequences. The Web service is called LINbase. It leverages the previously described concept of Life Identification Numbers (LINs), which are codes assigned to individual organisms based on genome similarity. Most genomes currently in LINbase were imported from GenBank but users have the option to upload their own genome sequences as well. Importantly, LINbase allows users to share the precise identity of their sequenced genomes without sharing the actual genome sequences, making not yet published or private genome sequences discoverable by the

scientific community stimulating collaboration between academia and industry. LINbase is available at http://www.LINbase.org.

Introduction

Fast and precise pathogen identification is crucial in human, animal, and plant disease diagnosis to identify the most effective treatment and to limit disease spread [79]. Precise identification of microbes is also important in many other fields, for example, when regulating commercial probiotics for human consumption [80] or biopesticides to control plant diseases in agriculture [81]. While we often associate the process of identification with giving an unknown organism a name, the ultimate goal of identification is to predict the characteristics of the unknown organism independently of what its name is, for example, to answer a question such as: does the unknown microorganism cause a certain disease in a specific animal species? The prerequisite to obtaining such precise identification is precise classification [82]. Only if microbes are classified into groups (called taxa) in which all members are derived from a recent common ancestor (MRCA) (*i.e.,* constitute a monophyletic group) and share a phenotype absent from organisms outside of that same taxon, can identification of an unknown as a member of such taxon lead to precise prediction of its phenotype.

Before the advent of DNA sequencing, classification and identification necessarily relied on phenotypic tests . Therefore, taxa were restricted to groups of microbes that could be phenotypically distinguished from other microbes based on relatively simple lab-based assays [83] . Classification and identification then transitioned to more precise gene-based methods, in particular, sequencing of the 16S rRNA gene [71]. With the development of ever faster and cheaper high throughput DNA sequencing methods, the entire genome of organisms can now be used to classify organisms into monophyletic groups and identify them as members of these groups based on identification of single nucleotide polymorphisms (SNPs) [84], construction of phylogenetic trees based on conserved genes [41, 85], or measures of genome similarity at the whole genome level expressed as average nucleotide identity (ANI) [38]. Conceptually, a taxon can now consist of microbes that share nothing else than a single mutation inherited from their most recent common ancestor compared to organisms outside of that taxon. If that single mutation changed the phenotype of the microbes belonging to the taxon, then identifying an unknown as a member of that taxon could predict the phenotype of the unknown. Or, in another example, if all microbes with a specific SNP were isolated during a specific disease outbreak, identifying an unknown as a member of the corresponding taxon would be informative of a transmission event and would become epidemiologically important to stop the further spread of a disease.

However, the fundamental unit of current taxonomy is not the smallest distinguishable unit based on genome sequencing but it is the species, whereby each named species is associated with a "type" strain, which is considered the name-bearing strain of the species [86]. Since current taxonomy is grounded in microbiological history, the valid publication of a new named microbial species requires much more than sequencing the genome of a type strain and reporting a distinctive phenotype. Besides showing that the type strain of the newly named species has less than 95% ANI compared to genomes of type strains of already named species, valid publication requires a long list of results derived from laboratory-based phenotypic tests [87]. Moreover, the process of validly publishing a named species still relies on publication of a traditional manuscript even though the key genomic and phenotypic information of a new species could be easily reduced to a simple database entry, similar to what has been proposed for the Digital Protologue Database [68]. Another limitation with using the species as the smallest unit of bacterial taxonomy is that members of the same species sometimes still vary considerably in regard to some phenotypes, for example, a single plant pathogen species may include many different strains with many of them having a different host range [88]. Although classification schemes at intraspecific levels exist to take into account phenotypic differences between strains belonging to the same species, they are not consistent across species making it difficult to interpret identification results based on a particular scheme for scientists who do not have familiarity with a particular species-specific scheme.

To address the above-listed limitations of current taxonomy and to take full advantage of genome sequencing for precise classification, the Life Identification Number (LIN) system was introduced [34, 75]. The LIN system classifies bacteria based on reciprocal ANI. In its current implementation, LINs consist of 20 positions, each representing a different ANI threshold. ANI thresholds range from 70% at the left-most position to 99.999% at the right-most position (**Fig. 1**). Importantly, LINs are assigned to individual genomes, whereby genomic relatedness between genomes is represented by the length of the longest common prefix of their LINs: the longer the LIN prefix is that is shared by two genomes, the more similar the genomes are to each other. To assign a LIN to a newly sequenced genome, the most similar genome that already has a LIN is identified in a database of genomes and the LIN of the new genome is computed based on its ANI to that most similar genome [89].

Any group of bacteria that share a LIN prefix of any length is called a LINgroup [35]. If the members of a LINgroup share a phenotype of interest, that single phenotype can be associated with that LINgroup. Therefore, if a microbe is identified as a member of a LINgroup based on its genome sequence, the unknown can be inferred, with high likelihood, to have the same phenotype as all the other members of that LINgroup. Validly published named species and genera can also be correlated with LINgroups. For example, if all known members of a named species share a certain LIN prefix, for example, $0_A1_B0_C0_D0_E4_F$, then the LINgroup $0_A1_B0_C0_D0_E4_F$ can be associated with that named species and unknowns can be precisely identified as a member of that named species based on their genome sequence (**Figure 1**).

Here we introduce LINbase, a Web service that implements the LIN and LINgroup concepts using an SQL database, efficient algorithms, and an intuitive website. Registered users can genomically circumscribe LINgroups and associate them with any phenotype based on their subject knowledge. Users can also associate a LINgroup with any validly published named species or genus based on their taxonomic expertise. This crowdsourcing approach is expected to provide precise genome-based circumscriptions and phenotypic descriptions of taxa, *i.e.*, LINgroups. To precisely identify microbes, users can query LINbase with genome sequences to determine if an unknown microbe is a member of any circumscribed LINgroup. Users can also upload their own genome sequences to LINbase. Importantly, genome sequences are not shared with other users but the assigned LINs reveal their precise similarity to all other genomes in LINbase and make them discoverable by all other users allowing even industry to share their repertoire of genomes without having to share actual DNA sequences. LINbase is fully functional but improvements in regard to speed, resolution, and functionality are ongoing.

Web Server infrastructure

Web server

LINbase is built with the LAMP (**L**inux, **A**pache server, **M**ySQL and **P**HP) stack with a RESTful API written in JavaScript and a job scheduler written in the Go programming language. All code is structured in an MVC (Model-View-Controller) framework named CodeIgniter. The analytical parts of LINbase are written in Python. The server currently runs on an Intel Xeon

16-core CPU at 1.90GHz, with 64GB RAM and the CentOS 7 operating system. The Web site can be accessed at http://linbase.org.

Database management

MySQL 5.6 is used to manage the database and store all relevant metadata. The schema is shown in **Figure 2**. Each table has a primary key and is connected to other tables with a foreign key. There are 4 main tables storing data related to uploaded genomes: the genome table stores the locations of the genome assemblies on the server, the taxonomy table stores the taxonomic information, the MetadataValue table stores associated metadata, and LINs of all uploaded genomes are recorded in the LIN table. The remaining tables serve the purpose of smoothing the data transfer and task management of LINbase. All tables are indexed for optimized query speed.

Summary of LINbase Functions

LINbase assigns LINs to microbial genomes uploaded by LINbase administrators and users. Users are encouraged to describe microbial taxa that correspond to genera or species or intraspecific groups with distinctive phenotypes as LINgroups. Users can also comment on LINgroups described by other users, search genomes and LINgroups by keyword, and query LINbase using gene sequences and genome sequences.

At the time of writing, 6204 bacterial genomes have been uploaded to LINbase and 4079 of them have been identified as members of 50 circumscribed LINgroups.

Genome upload function

The goal of LINbase administrators is to add all microbial genome sequence assemblies of NCBI's Genbank database to LINbase as long as assemblies satisfy minimal quality standards, such as having fewer than 500 contigs. However, if users are interested in Genbank genome sequences that have not been added to LINbase yet, they can upload Genbank genome sequences. Users can also upload their own unpublished genome sequences. When a user attempts to upload a genome sequence assembly that is already in LINbase, users will be redirected to that genome sequence.

If the user's genome sequence is not yet in LINbase, a LIN will be assigned using the LINflow procedure described in detail elsewhere (see Chapter 2). In short, a k-mer signature is computed using sourmash [76] with parameters k=21, 51. The computed signature is then compared with the signatures of representative genomes that are already in LINbase at the 95% ANI level (LIN position F) using k=21. If a genome sequence is found to have a Jaccard similarity of J >= 0.2475 (which corresponds to 95% ANI) compared to the uploaded genome, the uploaded genome is identified as a member of the represented LINgroup and the signature of the new genome is then compared with the signatures of all the members of this LINgroup using k=51. If instead, the LINgroup with the highest Jaccard similarity has a J<0.2475, the signature of the new genome is compared with the members of that LINgroup using k=21. In both cases, ANI is then calculated between the uploaded genome and the genome with the highest Jaccard similarity using pyANI [62]. The computed ANI value is then used to assign a LIN to the new genome based on the LIN of the genome with the highest Jaccard similarity. This is done by keeping the prefix of the reference LIN up to the LIN position at which the ANI threshold is smaller than the computed ANI value. At the next LIN position (*i.e.*, at which the computed ANI value is smaller than the ANI threshold), a number is assigned that has not yet been used at that position. The following positions are filled with 0's. The average time for one LIN assignment is 3 minutes and 54 seconds.

When uploading a genome, the user has to enter a strain name as the only required metadata value. Genus, species, and information on intraspecific classification are optional. Other metadata can be entered based on a user's selection of "Interest" (**Table 1** and **2**). Currently, the following interests are available (but additional interests and additional metadata options can be added upon contacting LINbase administrators): Undefined interest, Plant pathogens, Environmental bacteria, Uncultured bacteria, Foodborne pathogens, and Archaea (**Figure 3**).

After the genome is successfully uploaded, the result page will return the LIN assigned to the new genome, the most similar genome based on which the LIN was assigned, and the respective ANI value. The genome's membership in LINgroup(s) that have been described in LINbase by the same user or any other user are also reported. A description of how LINgroups are described follows below.

LINgroup description function

A group of genomes can be selected from any result page and described as a LINgroup by highlighting with the mouse the LINprefix shared by the group of genomes and clicking on the link "add a description". The user chooses the type of LINgroup (either a taxonomic rank or a non-taxonomic group within a species that share the same characteristics), adds a name (which can be a species name, if the LINgroup corresponds to a species, or any other name the user chooses), a description giving more information about the LINgroup (for example, the phenotype that is shared by its members), and a URL or DOI to a peer-reviewed publication about the LINgroup (**Figure 4**).

We expect users to generally choose the longest LIN prefix shared by a group of genomes when describing a LINgroup. For example, if all genomes of the genus *Pseudomonas* in LINbase share the LIN prefix 50_A, then a user could describe the LINgroup 50_A as genus *Pseudomonas*. Instead of choosing the maximum length of the LINprefix shared by a group of genomes, a user can also choose to describe a group of genomes by the minimum length of the LINprefix that distinguishes the group from members outside of the group. For example, if there are only two genomes in LINbase that belong to an intraspecific group, this may be the better approach since more diverse members of the group may be added later. Finally, a LINgroup can also be described based on a single genome choosing the LINprefix up to position F or G, which correspond to the broadly accepted ANI thresholds for speciation at 95%-96%. This will be typically done if only the genome of the type strain of a species has been added to LINbase.

As soon as a new LINgroup description has been added to LINbase, any newly uploaded genome will be automatically identified as a member of that LINgroup if its LIN includes the LINprefix of the LINgroup.

Genome and LINgroup search function

Both, individual genomes and described LINgroups, can be searched in LINbase. Entered parameters will form one single query so that query time is minimized. Searching by either genome or LINgroup takes less than 1 second to return the result.

When searching for genomes, users can use any LIN position(s), area of interest, taxonomic information, and isolation metadata as filters in the query (**Figure 5A**). The genome-search result page will list the genomes that match the query as well as the described LINgroups that include these genomes as members (**Figure 6**).

When searching for described LINgroups, users can search by LIN position(s), the name of the user who described the LINgroup, and words used in the LINgroup name and description (**Figure 5B**). The result page will list the described LINgroups that match the query.

Identify function using gene or genome sequences as query

Users can identify an unknown microbe either using a genome sequence or a gene sequence as the query.

When using a genome sequence as the query, the most similar genome in the database is identified using a workflow similar to the one described above for the genome upload function (**Figure 7A**). However, to achieve higher speed with only moderate reduction in accuracy, FastANI [45] replaces pyANI when computing ANI between the query genome and the most similar genome identified by sourmash. On the result page, the most similar genome and its LIN, the ANI value between the query genome and the most similar genome, and any LINgroup that the query genome is a member of are reported (**Figure 8A**).

Users who do not have the whole genome sequence of an unknown microbe can also use a single gene sequence as the query in combination with BLASTn (**Figure 7B**) [36]. However, accuracy is of course largely reduced since multiple genomes, which may even belong to different LINgroups, may align with a short gene sequence with 100% identity. To minimize the risk of misidentification, only genomes with low e-values are returned on the result page along with LINgroup(s) that these genomes belong to (**Figure 8B**).

Comment function

A commenting system is implemented in the LINgroup profile page to facilitate communication and potential collaboration among LINbase users. Users can add comments to any LINgroup (described or undescribed) to discuss the LINgroup with other users. Posted comments can be edited or deleted by the original poster. At this time, users are not automatically notified of comments posted to LINgroups they described. However, this function is planned for the future.

Data security and dissemination

Genome assemblies in LINbase, either sourced from public databases, such as NCBI, or uploaded by users, are securely saved on the server and cannot be viewed or downloaded by any user. Gene and genome sequences uploaded as part of the identification function are deleted along with intermediate data immediately after the identification process is finished. The data that are shared in LINbase are genome metadata (including taxonomic and isolation information), LINs, LINgroups, LINgroup descriptions, and comments. Therefore, LINbase is ideally suited for sharing the precise identity of sequenced genomes as soon as they are generated while keeping the actual genome sequences private until submission to a public database.

Discussion

Here we introduced LINbase, a Web service that implements bacterial taxonomy based on whole genome similarity and supported by fast and accurate algorithms. LINbase complements functionalities offered by other online Web services for genome-based microbial identification, such as MiGA [65] or EzBioCloud [66] as follows: 1. it labels individual genomes with LINs, which reflect the precise genomic relatedness among strains in the database, 2. it automatically gathers genomically similar bacteria into taxa (LINgroups), 3. it provides a user-friendly interface to genomically circumscribe validly published named taxa at the genus and species rank and at intraspecific levels as LINgroups permitting precise genome-based identification, 4. it uses crowdsourcing to incorporate informal taxa/LINgroups independently of published named taxa, 5. it encourages scientific exchange and early sharing of data by providing an avenue to share the precise identity of sequenced genomes without sharing the genome sequences themselves, and 6. it allows users to interact with each other by commenting on LINgroup circumscriptions and descriptions.

Despite the aforementioned advantages of LINbase, there are limitations in its current version in regard to the classification of bacteria at higher ranks (family, order, class, and phylum), which can currently not be circumscribed as LINgroups, and for bacteria with very recent common ancestors, *e.g.*, differentiating foodborne pathogens from different outbreaks is currently only possible when high quality genome assemblies are available. If assemblies are of low quality, the correlation between phylogeny and LINs fails at the right-most LIN positions. Also, genome upload is currently managed by a scheduler that only allows one process at a time. This limits the ability to batch upload genomes and does not allow multiple users to upload genomes at the same time.

Future implementations of LINbase will focus on increasing the speed of the identification function when using a genome sequence as the query and of LIN assignment. Parallelization is a promising solution to speed up LIN assignment when genomes are uploaded by different users at the same time. Parallelization would also allow batch uploading, which can further accelerate identification and LIN assignment. We are also planning to expand LIN positions to the left up to the phylum level by using algorithms to detect low-level genome similarity and to improve assignment of isolates to outbreaks by integrating additional algorithms to precisely identify phylogenetic relationships among very similar genomes. Finally, our aim is to automatically add all genome sequences in Genbank to LINbase, to automatically circumscribe all monophyletic taxa as LINgroups by integrating LINbase with the Genome Taxonomy Database [23], and to integrate LINbase with other platforms to improve genome-based classification and identification of microbes at all taxonomic ranks.

Tables

Table 3.1 The "Metadata" table of LINbase. Each Metadata ID is associated with a category of metadata (Metadata_Item).

Interest_ID	**InterestName**	**Metadata_IDs**
1	Plant pathogens	1,2,3,4,5,6,7,8,9,10,11,12,13,14
2	Foodborne pathogens	1,16,17,2,3,4,5,6,7,8
3	Environmental bacteria	1,2,3,4,5,6,7,8,15
4	Uncultured bacteria	1,2,3,4,5,6,7,8,15
5	Archaea	1,2,3,4,5,6,7,8,15
6	Unidentified interest	1,2,3,4,5,6,7,8,15

Table 3.2 The "Interest" table of LINbase. The current interests in LINbase and their corresponding metadata categories represented as lists of Metadata IDs (Metadata_IDs).

Metadata_ID	**Metadata_Item**
1	Type strain
2	NCBI Taxonomy ID
3	NCBI Accession Number
4	Date of isolation
5	Country
6	Region
7	GPS Coordinates
8	Link to peer-reviewed paper
9	Host of isolation
10	Secondary host
11	Disease
12	Symptom
13	Phenotype
14	Fluorescence
15	Environmental source
16	Source of isolation
17	Outbreak

Figures

Figure 3.1 The LIN and LINgroups concept. Each LIN position (A - T) represents an ANI threshold ranging from 70% at position A to 99.999% at position T. LINgroups are used to describe groups of microbes sharing the same phenotype(s). LINgroups are denoted by their shared LIN prefix (from position A to the right-most position that all members of a LINgroup share). For example, using the ANI speciation threshold of 95%, LINgroup $0_A1_B0_C0_D0_E3_F$ corresponds to the species G1 S2, and LINgroup $0_A1_B0_C0_D0_E4_F$ corresponds to the species G1 S3.

			70%	75%	80%	85%	90%	95%	96%	97%	98%	98.5%	99%	99.25%	99.5%	99.75%	99.9%	99.925%	99.95%	99.975%	99.99%	99.999%
Genus	Species	Strain	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
G1	S1	X1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
G1	S2	X2	0	1	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0
G1	S2	X3	0	1	0	0	0	3	0	0	0	0	1	0	0	0	0	0	0	0	0	0
G1	S3	X4	0	1	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0
G1	S3	X5	0	1	0	0	0	4	1	0	0	0	0	0	0	0	0	0	0	0	0	0
G1	S3	X6	0	1	0	0	0	4	1	0	0	0	0	0	0	0	1	0	0	0	0	0

Figure 3.2 Database schema of the SQL database of LINbase. All relevant data of LINbase are saved in a relational database with the shown schema. Each table is connected with other tables with a primary key and foreign key(s). The first row of each table is the primary key, and the arrow pointing to another table indicates the connection with the foreign key.

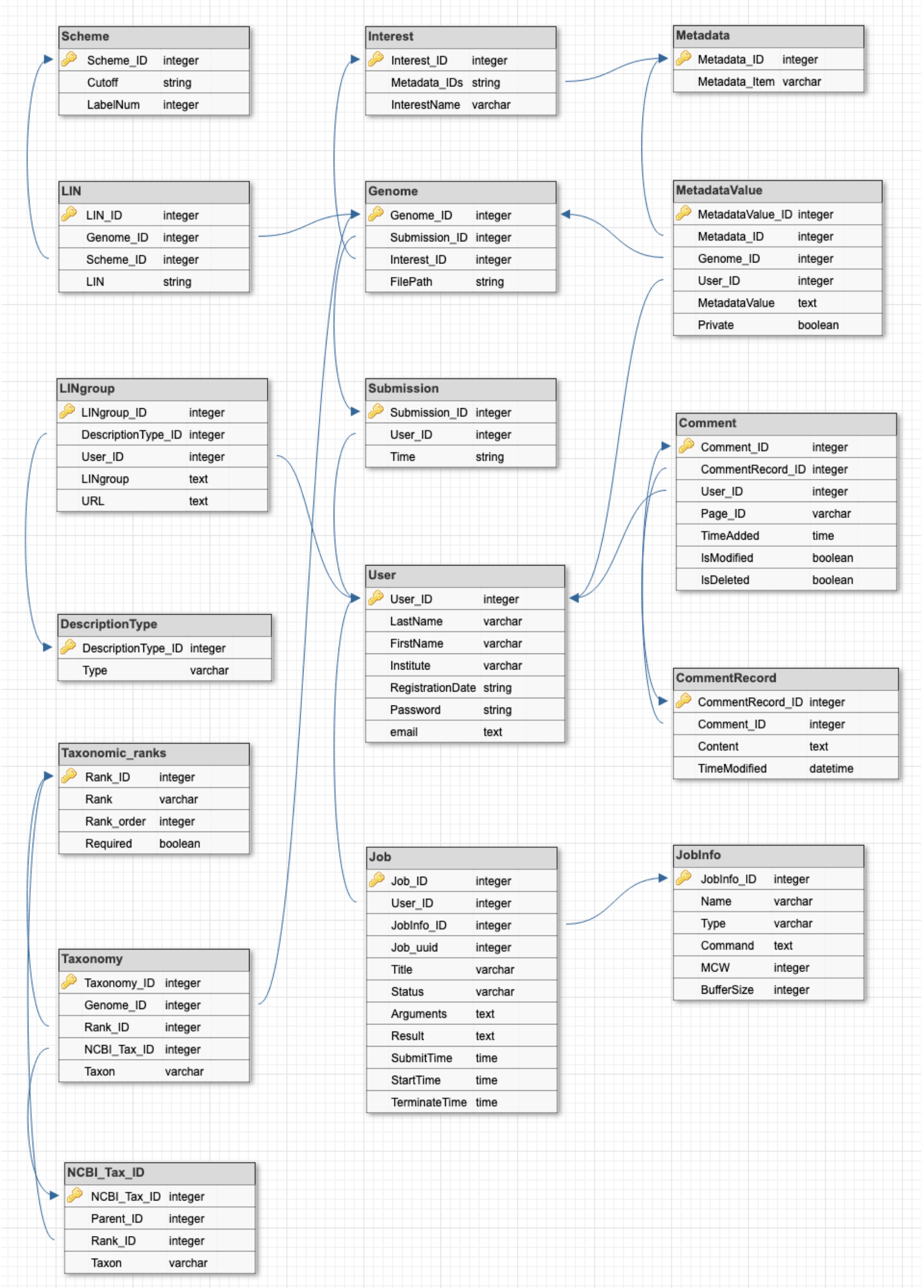
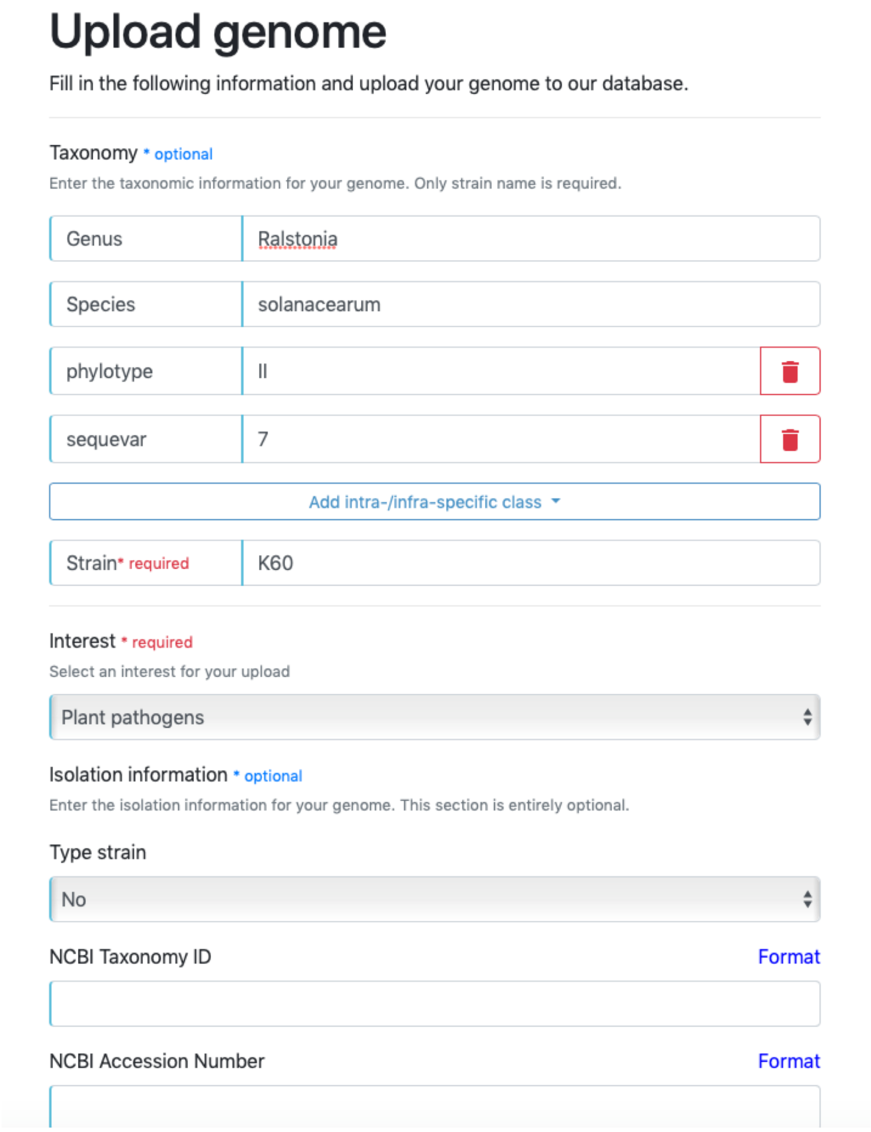


Figure 3.3 “Upload genome” form. Users are asked to enter taxonomic information and isolation metadata before uploading microbial genome sequences. In the taxonomic information section, only strain name is required as an identifier of the uploaded genome since taxonomic information may not be available. Users are required to choose an area of interest to associate the uploaded genome with a research area, e.g. Plant pathogen, Foodborne pathogen, Environmental bacteria, etc. This allows the form to change dynamically in regard to the available metadata fields. For example, the field “Host of isolation” only becomes available when choosing “Plant pathogens” but not when choosing “Environmental bacteria”.



Upload genome

Fill in the following information and upload your genome to our database.

Taxonomy * optional

Enter the taxonomic information for your genome. Only strain name is required.

Genus	Ralstonia
Species	solanacearum
phylotype	II
sequevar	7

Add intra-/infra-specific class

Strain* required	K60

Interest * required

Select an interest for your upload

Plant pathogens

Isolation information * optional

Enter the isolation information for your genome. This section is entirely optional.

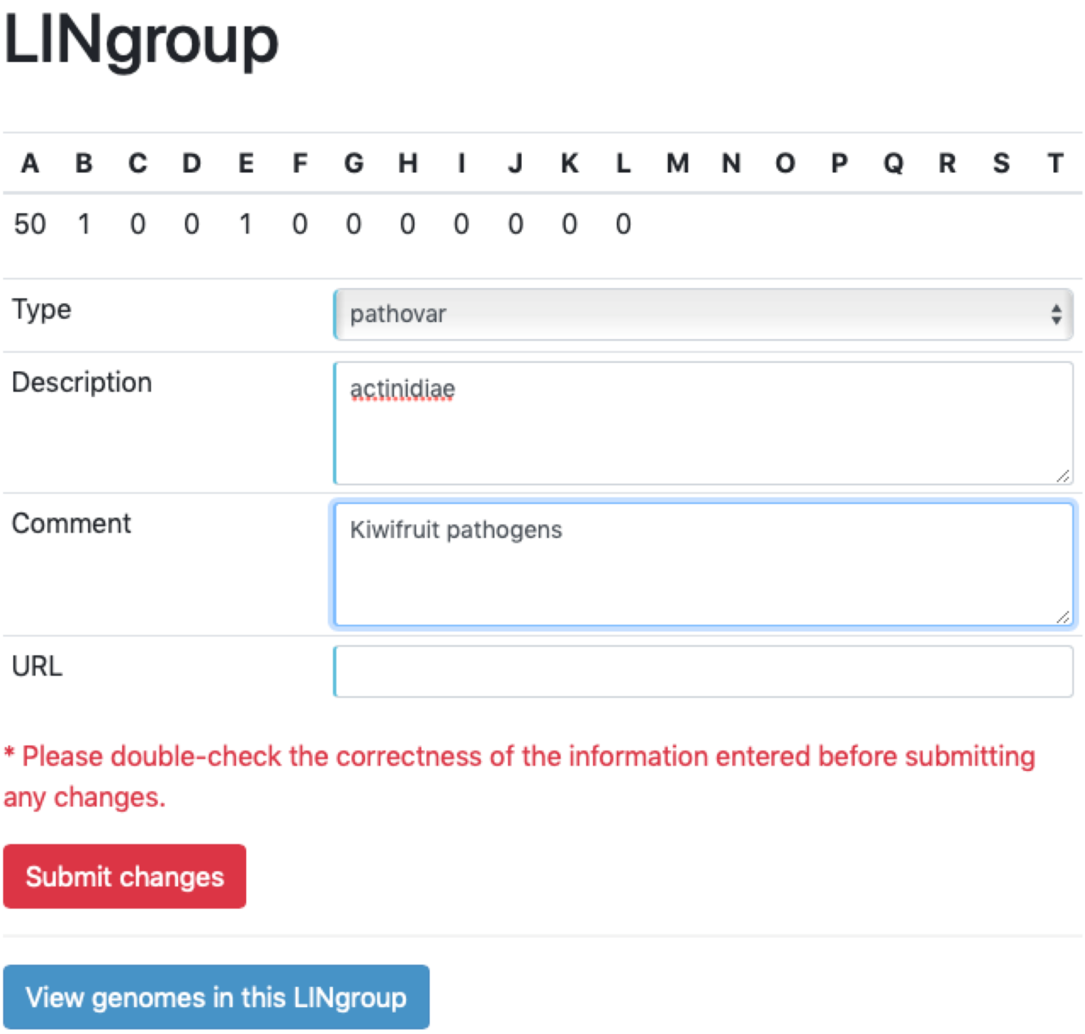
Type strain

No

NCBI Taxonomy ID Format

NCBI Accession Number Format

Figure 3.4 "Add a Lingroup description" form. After an undescribed LINgroup is selected, the user can describe the LINgroup at a taxonomic rank or as a group of microbes within a species that share a phenotype. This is done by choosing the type of taxon from the "Type" dropdown menu and entering a name and an optional comment and/or optional link to a peer-reviewed publication.



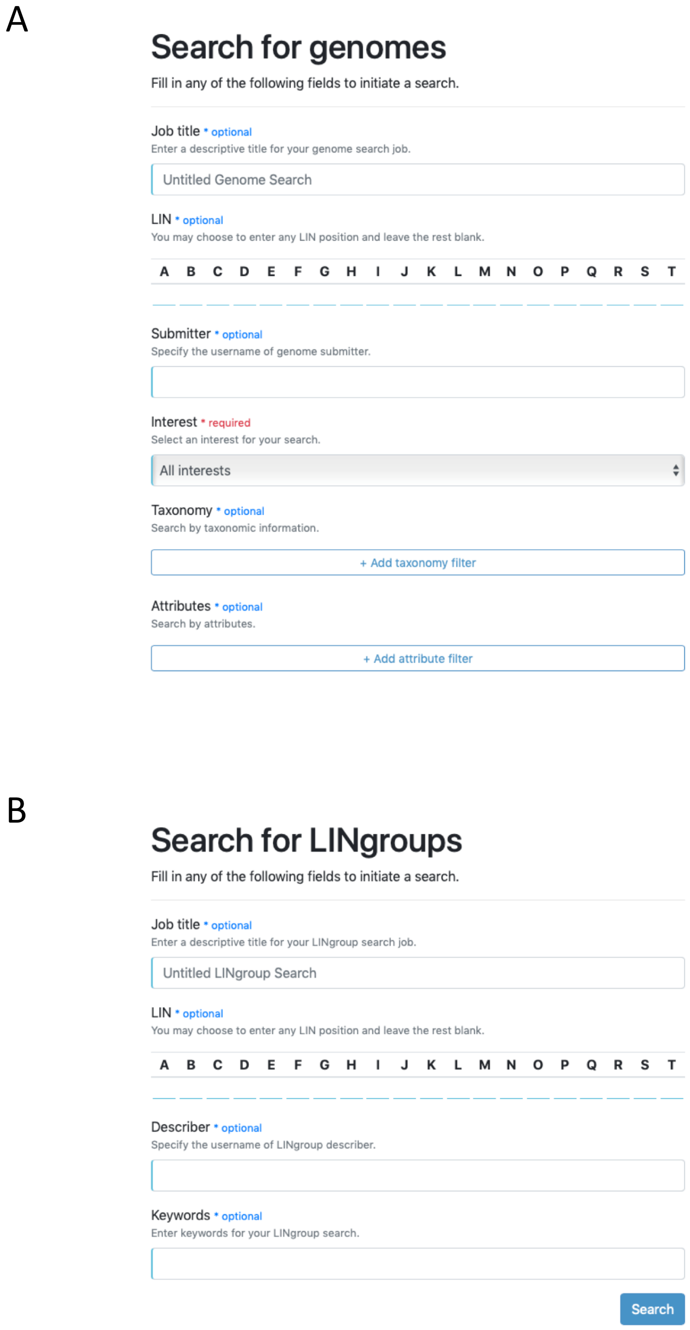
LINgroup

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
50	1	0	0	1	0	0	0	0	0	0	0								

Type	pathovar
Description	actinidiae
Comment	Kiwifruit pathogens
URL	

* Please double-check the correctness of the information entered before submitting any changes.

Figure 3.5 Forms for searching LINbase. (A) “Search for genomes” form. The user can search for genomes in LINbase by using any of the provided fields including LIN, submitter, interest, taxonomic information, and isolation information to narrow down the search. (B) “Search for LINgroups” form. The user can search for described LINgroups by describer and keywords. All filled fields will be passed to the backend as filters to query the database.



Search for genomes

Fill in any of the following fields to initiate a search.

Job title • optional
Enter a descriptive title for your genome search job.
Untitled Genome Search

LIN • optional
You may choose to enter any LIN position and leave the rest blank.

A B C D E F G H I J K L M N O P Q R S T

Submitter • optional
Specify the username of genome submitter.

Interest • required
Select an interest for your search.
All interests

Taxonomy • optional
Search by taxonomic information.

Attributes • optional
Search by attributes.

Search for LINgroups

Fill in any of the following fields to initiate a search.

Job title • optional
Enter a descriptive title for your LINgroup search job.
Untitled LINgroup Search

LIN • optional
You may choose to enter any LIN position and leave the rest blank.

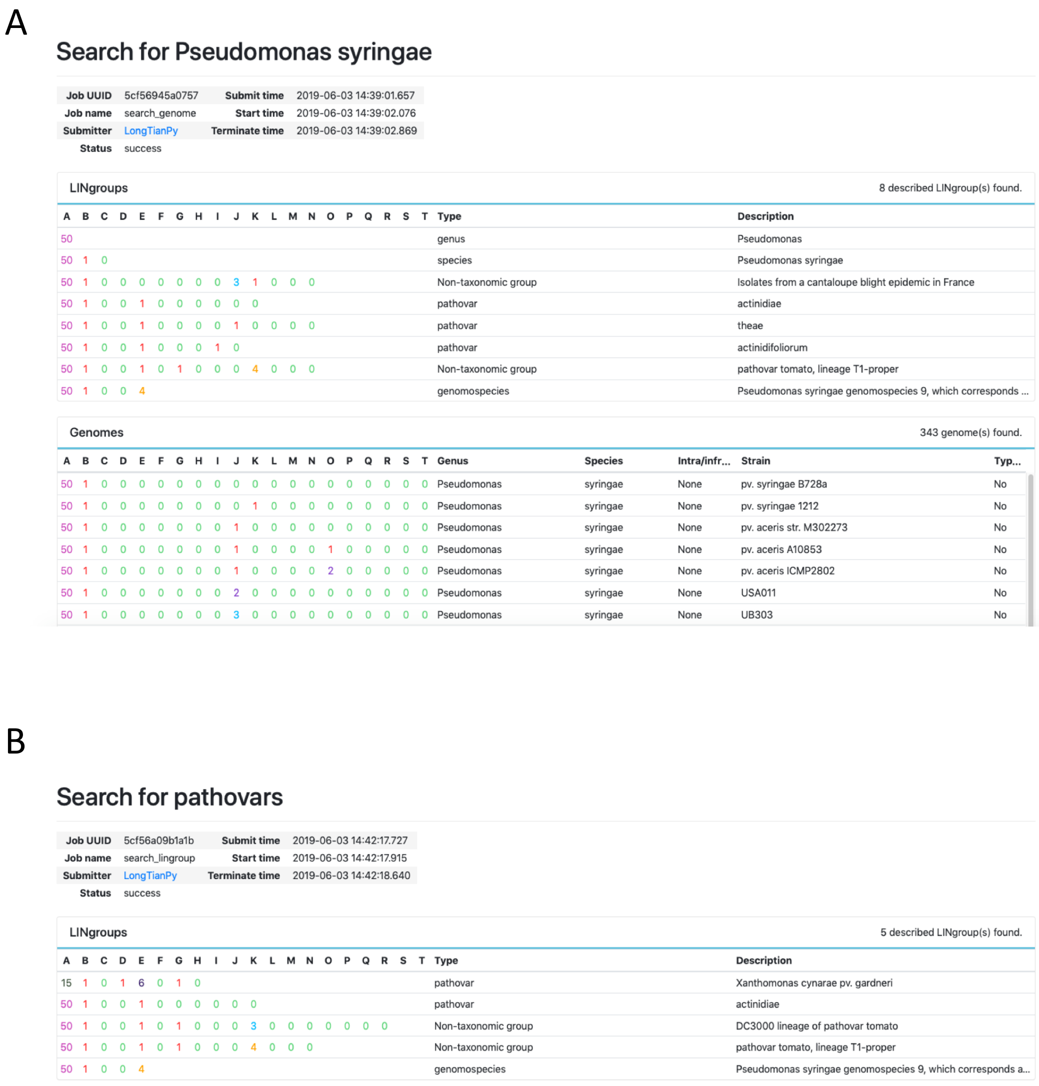
A B C D E F G H I J K L M N O P Q R S T

Describer • optional
Specify the username of LINgroup describer.

Keywords • optional
Enter keywords for your LINgroup search.

Figure 3.6 Search result pages. (A) The result page when searching, for example, for the species *Pseudomonas syringae*. All *Pseudomonas syringae* genomes and all associated

LINgroups are printed to the screen. (B) The result page when searching, for example, for the keyword “pathovar”. All LINgroups with “pathovar” in their description are returned.



Search for Pseudomonas syringae

Job UUID	5cf56945a0757	Submit time	2019-06-03 14:39:01.657
Job name	search_genome	Start time	2019-06-03 14:39:02.076
Submitter	LongTianPy	Terminate time	2019-06-03 14:39:02.869
Status	success		

LINgroups — 8 described LINgroup(s) found.

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	Type	Description
50																				genus	Pseudomonas
50	1	0																		species	Pseudomonas syringae
50	1	0	0	0	0	0	0	0	3	1	0	0	0							Non-taxonomic group	Isolates from a cantaloupe blight epidemic in France
50	1	0	0	1	0	0	0	0	0	0										pathovar	actinidiae
50	1	0	0	1	0	0	0	0	1	0	0	0	0							pathovar	theae
50	1	0	0	1	0	0	0	1	0											pathovar	actinidifoliorum
50	1	0	0	1	0	1	0	0	0	4	0	0	0							Non-taxonomic group	pathovar tomato, lineage T1-proper
50	1	0	0	4																genomospecies	Pseudomonas syringae genomospecies 9, which corresponds ...

Genomes — 343 genome(s) found.

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	Genus	Species	Intra/infr...	Strain	Typ...
50	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Pseudomonas	syringae	None	pv. syringae B728a	No
50	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	Pseudomonas	syringae	None	pv. syringae 1212	No
50	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	Pseudomonas	syringae	None	pv. aceris str. M302273	No
50	1	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	Pseudomonas	syringae	None	pv. aceris A10853	No
50	1	0	0	0	0	0	0	0	1	0	0	0	0	2	0	0	0	0	0	Pseudomonas	syringae	None	pv. aceris ICMP2802	No
50	1	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	Pseudomonas	syringae	None	USA011	No
50	1	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	Pseudomonas	syringae	None	UB303	No

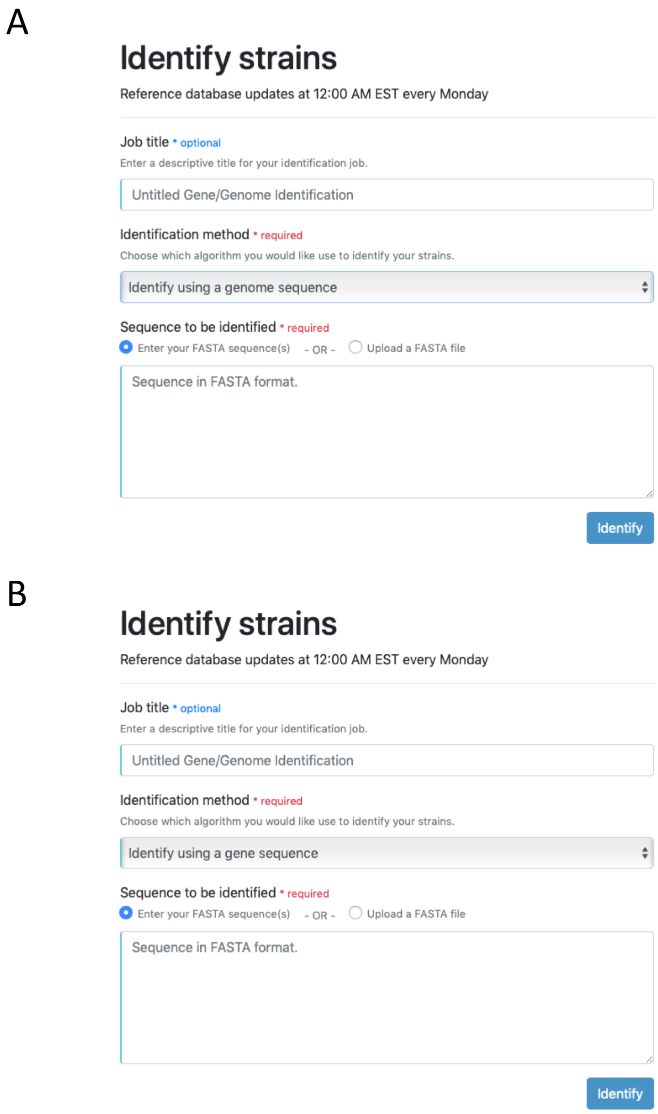
Search for pathovars

Job UUID	5cf56a09b1a1b	Submit time	2019-06-03 14:42:17.727
Job name	search_lingroup	Start time	2019-06-03 14:42:17.915
Submitter	LongTianPy	Terminate time	2019-06-03 14:42:18.640
Status	success		

LINgroups — 5 described LINgroup(s) found.

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	Type	Description
15	1	0	1	6	0	1	0													pathovar	Xanthomonas cynarae pv. gardneri
50	1	0	0	1	0	0	0	0	0	0										pathovar	actinidiae
50	1	0	0	1	0	1	0	0	0	3	0	0	0	0	0	0	0			Non-taxonomic group	DC3000 lineage of pathovar tomato
50	1	0	0	1	0	1	0	0	0	4	0	0	0							Non-taxonomic group	pathovar tomato, lineage T1-proper
50	1	0	0	4																genomospecies	Pseudomonas syringae genomospecies 9, which corresponds a...

Figure 3.7 “Identify” forms. (A) Identification with a genome assembly. (B) Identification with a gene sequence. Both functions accept gene or genome sequences uploaded as a FASTA-format file or entered in the textbox.



A

Identify strains

Reference database updates at 12:00 AM EST every Monday

Job title * optional

Enter a descriptive title for your identification job.

Untitled Gene/Genome Identification

Identification method * required

Choose which algorithm you would like use to identify your strains.

Identify using a genome sequence

Sequence to be identified * required

● Enter your FASTA sequence(s) - OR - ○ Upload a FASTA file

Sequence in FASTA format.

B

Identify strains

Reference database updates at 12:00 AM EST every Monday

Job title * optional

Enter a descriptive title for your identification job.

Untitled Gene/Genome Identification

Identification method * required

Choose which algorithm you would like use to identify your strains.

Identify using a gene sequence

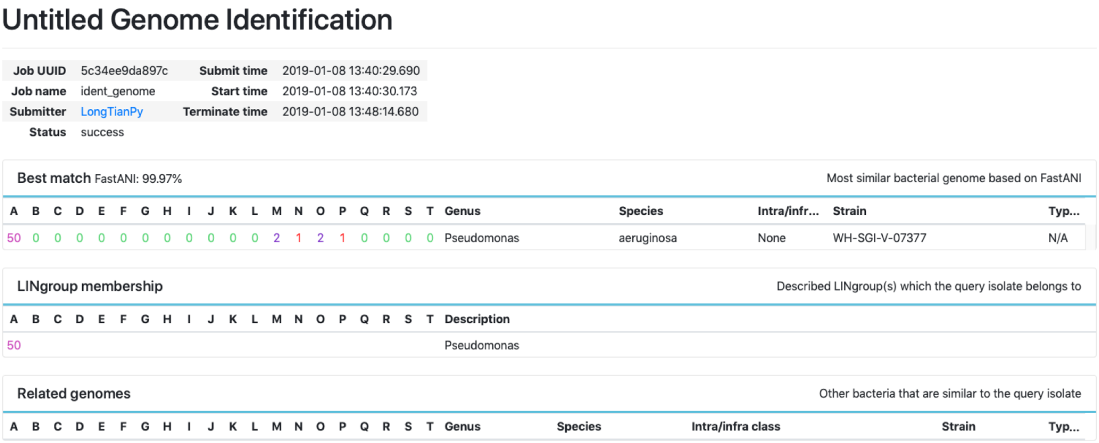
Sequence to be identified * required

● Enter your FASTA sequence(s) - OR - ○ Upload a FASTA file

Sequence in FASTA format.

Figure 3.8 “Identify” result page. (A) Result page for genome-based identification. The submitted genome is queried against LINbase genomes and the genome with the highest FastANI is returned. The LINgroups that the query genome belongs to, based on its ANI with the best match, are listed as well. **(B) Result page for gene-based identification.** The submitted gene sequence is queried against LINbase genomes with BLASTN Genomes with E-value=0 are listed as best matches. The LINgroups the best matches belong to are listed as well. For both types of identification, the submitted sequences will be deleted from the server once the query is completed.

A



Untitled Genome Identification

Job UUID	5c34ee9da897c	Submit time	2019-01-08 13:40:29.690
Job name	ident_genome	Start time	2019-01-08 13:40:30.173
Submitter	LongTianPy	Terminate time	2019-01-08 13:48:14.680
Status	success		

Best match FastANI: 99.97%

Most similar bacterial genome based on FastANI

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	Genus	Species	Intra/infr...	Strain	Typ...
50	0	0	0	0	0	0	0	0	0	0	0	2	1	2	1	0	0	0	0	Pseudomonas	aeruginosa	None	WH-SGI-V-07377	N/A

LINgroup membership

Described LINgroup(s) which the query isolate belongs to

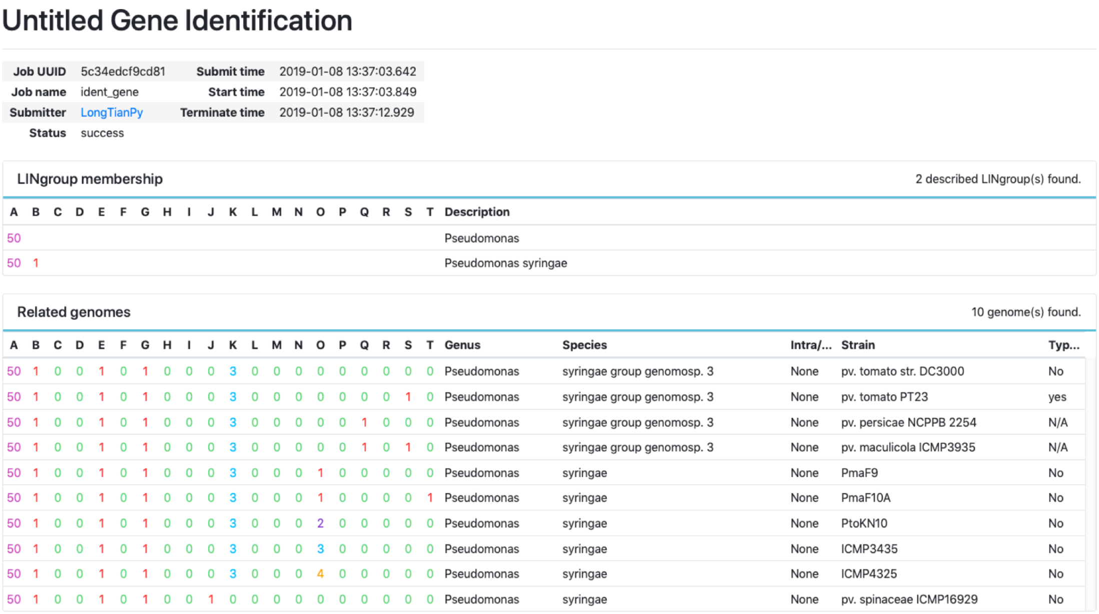
A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	Description
50																				Pseudomonas

Related genomes

Other bacteria that are similar to the query isolate

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	Genus	Species	Intra/infra class	Strain	Typ...

B



Untitled Gene Identification

Job UUID	5c34edcf9cd81	Submit time	2019-01-08 13:37:03.642
Job name	ident_gene	Start time	2019-01-08 13:37:03.849
Submitter	LongTianPy	Terminate time	2019-01-08 13:37:12.929
Status	success		

LINgroup membership

2 described LINgroup(s) found.

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	Description
50																				Pseudomonas
50	1																			Pseudomonas syringae

Related genomes

10 genome(s) found.

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	Genus	Species	Intra/...	Strain	Typ...
50	1	0	0	1	0	1	0	0	0	3	0	0	0	0	0	0	0	0	0	Pseudomonas	syringae group genomosp. 3	None	pv. tomato str. DC3000	No
50	1	0	0	1	0	1	0	0	0	3	0	0	0	0	0	0	0	1	0	Pseudomonas	syringae group genomosp. 3	None	pv. tomato PT23	yes
50	1	0	0	1	0	1	0	0	0	3	0	0	0	0	0	1	0	0	0	Pseudomonas	syringae group genomosp. 3	None	pv. persicae NCPPB 2254	N/A
50	1	0	0	1	0	1	0	0	0	3	0	0	0	0	0	1	0	1	0	Pseudomonas	syringae group genomosp. 3	None	pv. maculicola ICMP3935	N/A
50	1	0	0	1	0	1	0	0	0	3	0	0	0	1	0	0	0	0	0	Pseudomonas	syringae	None	PmaF9	No
50	1	0	0	1	0	1	0	0	0	3	0	0	0	1	0	0	0	0	1	Pseudomonas	syringae	None	PmaF10A	No
50	1	0	0	1	0	1	0	0	0	3	0	0	0	2	0	0	0	0	0	Pseudomonas	syringae	None	PtoKN10	No
50	1	0	0	1	0	1	0	0	0	3	0	0	0	3	0	0	0	0	0	Pseudomonas	syringae	None	ICMP3435	No
50	1	0	0	1	0	1	0	0	0	3	0	0	0	4	0	0	0	0	0	Pseudomonas	syringae	None	ICMP4325	No
50	1	0	0	1	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	Pseudomonas	syringae	None	pv. spinaceae ICMP16929	No

Chapter 4 Genome-based classification and identification of bacterial plant pathogens from the genus level to the intraspecific level based on genome similarity

Long Tian[1], Tiffany M. Lowe-Power[2], Lenwood S. Heath[3], Boris A. Vinatzer[1]

[1]School of Plant and Environmental Sciences, Virginia Tech, Blacksburg

[2]Plant Pathology Department, University of California, Davis

[3]Computer Science Department, Virginia Tech, Blacksburg

Attribution

The following chapter is based upon the manuscript in preparation: Tian, L., Lowe-Power, T.M., Heath, L.S., Vinatzer, B.A. (2019). Genome-based classification and identification of bacterial plant pathogens from the genus level to the intraspecific level based on genome similarity.

Abstract

Classification of many bacterial plant pathogens is challenging because of extensive variation in host range within the currently used species boundary of 95% average nucleotide identity (ANI). Since taxonomy uses the species as the smallest unit, current taxonomic practices cannot be used for these pathogens. The situation has been further complicated by a history of revisions that has seen the assignment of many pathogen strains to different named species or different intraspecific taxa over a short period of time. To facilitate classification and identification of plant pathogenic bacteria in this challenging situation, we have leveraged publicly available genome sequences, efficient algorithms, and the concept of genome-based Life Identification Numbers (LINs) in the Web service LINbase (linbase.org). Using LINbase, we have already circumscribed many taxa of plant pathogenic bacteria from the genus level to various intraspecific levels. Unknown plant pathogen strains can now be precisely identified as members of any of these taxa based on genome sequences or metagenome-assembled genome sequences (MAGs). Plant pathologists with expertise in genomics of any bacterial

plant pathogen taxon are invited to upload additional genome sequences, contribute additional pathogen circumscriptions, and endorse current circumscriptions (or suggest revisions) using the respective functions available at LINbase.

Introduction

While human medicine deals with pathogens of only one species, *i.e.*, *Homo sapiens*, plant pathologists deal with pathogens of an estimated 35,000 plant species of cultivated and domesticated plants [90] . Since every plant species is affected by several diseases, including many bacterial diseases, one would expect an even higher number of plant pathogens, including bacterial plant pathogens. However, there are only approximately 13 validly published names of plant pathogenic bacterial species [91]. The reason is that adaptation to different plant hosts mainly occurs within bacterial species, which are currently defined as monophyletic groups of bacteria that share a common evolutionary trajectory, are phenotypically distinct from other bacteria, and share over 70% reciprocal DNA-DNA hybridization values, corresponding to at least 95% average nucleotide identity (ANI) based on whole genome sequence comparisons [92]. Therefore, the current taxonomic framework for bacteria, which uses the species as the basic unit of classification, is not sufficient for the classification of the diversity of plant pathogenic bacteria.

In a small number of cases, intraspecific host range variants also have additional distinct phenotypic characteristics that allowed the descriptions of these host range variants as validly published subspecies. An example is the subspecies within the species *Acidovorax avenae* [93]. However, in most cases, host range variants cannot be distinguished by additional phenotypic characteristics. Therefore, other intraspecific classification systems were introduced. The most common system is the pathovar system [30, 94] that is in use for

Pseudomonas syringae and some *Xamthomonas* species. This system classifies strains based on their host range and the disease symptoms they cause, whereby the pathovar name is most often based on one of the plant species included in the host range, commonly the economically most import crop host. This system was introduced when the names of many plant pathogenic *Pseudomonas* and *Xanthomonas* species were not included in the approved list of bacterial names in 1980 because their names had never been validly published, i.e., the corresponding species had never been sufficiently described phenotypically and/or the descriptions were not associated with any deposited type strain [26]. Therefore, many plant pathogen strains were assigned to the validly published species *P. syringae* and *X. campestris* followed by a pathovar designation. When a pathovar is described based on thorough host range tests of several strains on several plant species, it usually corresponds to a monophyletic group of bacteria and it becomes a useful taxon for use in basic research and applied plant pathology. The problem is that, at least in many cases, strains were assigned to the same pathovar simply because they had been originally isolated from the same crop and incomplete host range testing did not reveal that these stains had only similar, but not identical, host ranges. In some of these cases, strains assigned to the same pathovar were later found to be genetically unrelated making these pathovars polyphyletic (and thus confusing), for example, strains originally identified as *P. syringae* pv. *maculicola* [95].

Two complementary systems were instead introduced for *Ralstonia solanacearum*: race and biovar [96]. Race designations are based on host range while biovar designations are based on biochemical tests. The problem is that as with the pathovar system, *R. solanacearum* races and biovars are not always monophyletic.

With the introduction of DNA sequencing of individual loci or even multiple loci, bacterial plant pathogen strains could for the first time be clearly classified based on evolutionary relationships and be assigned to different monophyletic clades. For example, this allowed researchers to clearly divide *R. solanacearum* strains into 4 different "phylotypes" and 23 different "sequevars" [31, 97] and to assign *P. syringae* strains to 13 different "phylogroups" [98]. While such phylogenetic clustering adds a lot of clarity, it can also create confusion whenever different publications define phylogenetic groups based on different genes and/or name the identified clades with different numbers and/or letters. Multilocus sequence typing (MLST) databases, such as the PAMDB database developed by us for plant-associated bacteria [99], can partially alleviate the above problem by reporting subspecies/pathovar/race/biovar designations together with allele designations from different MLST schemes.

Fortunately, with the development of next generation sequencing and third generation sequencing technologies and the dropping costs of these technologies, we can today replace the sequencing of individual loci with whole genome sequencing (WGS). WGS can be used to classify and identify bacteria at all taxonomic levels (for example, [10, 100]), and even at intraspecific levels down to the level of individual outbreak strains, as now routinely done for strain typing of foodborne human pathogens [67].

The current challenge is that genome sequencing provides unprecedented resolution for precise bacterial identification but we are still missing a classification system at the intraspecific level (above 95% ANI) to translate that resolution into precise classification. This situation is similar to a hypothetical optical instrument that may distinguish 100 different shades of blue based on precise wavelength measurements, while nobody has defined any shades of blue and the result of the highly precise wavelength measurement is still simply that the analyzed color is "blue". To alleviate this problem, we introduced the Life Identification Number (LIN) concept that, in its current implementation, allows assignment of bacteria, based on their genome sequences and reciprocal ANI measurements, to taxa (which we call LINgroups) that consist of members with reciprocal ANI as high as 99.999% ANI [75].

The LIN concept was described in detail in previous publications [34, 101]. In short, unique LINs are assigned to individual bacterial isolates as they are added to a single database whereby each LIN consists of 20 positions (in the current implementation) with each position representing a different ANI threshold with thresholds increasing from the left to the right (70% at position A to 99.999% at position T). As in classical hierarchical taxonomy, in which the taxonomic lineages of closely related bacteria overlap almost all the way from the phylum level to the species level while distantly related bacteria may only share the phylum level, the LIN of two distantly related bacteria may already be different at the A position (when the two bacteria share less than 70% ANI) but very closely related bacteria may share a LIN from the A position to the S position (**Figure 4.1**).

LINs are automatically assigned to bacteria based on their genome sequences without the need for human judgement. LINs simply express how similar bacteria are without deciding if any similarity threshold is more important than another one, thus avoiding any subjective judgement on what should constitute a boundary for a named species or for any other taxon. Therefore, any taxon (defined here as any monophyletic group of bacteria) can be precisely circumscribed as a LINgroup. Importantly for plant pathology, any monophyletic group of bacterial isolates with a distinctive host range (or even simply membership in the same disease outbreak) can be circumscribed, described, and named as a LINgroup, be it a species, a pathovar, a race/biovar combination, or an outbreak strain. Unknown isolates can then be identified as members of any of these LINgroups based on their genome sequences.

Here we used the Web server LINbase at linbase.org, which implements the LIN and LINgroup concepts (reference to the bioRxiv submission of chapter 3), to circumscribe, describe, and name a series of important bacterial plant pathogens as LINgroups (from the genus rank to various intraspecific groups, such as pathovars). Using a crowdsourcing approach, we encourage our plant pathology colleagues to add additional genome sequences and LINgroups to LINbase, to either endorse current LINgroups or propose changes to our circumscriptions and descriptions using the “Comment” function, and to identify unknown isolates using the “Identify” function.

Methods

Use of LINbase for LINgroup circumscriptions

LINbase is a Web service that implements bacterial taxonomy using a genome-based classification scheme. It accomplishes fast and precise identification with efficient algorithms (Chapter 2 & 3). Microbial genomes are uploaded to LINbase individually by users or in batch by the administrators along with corresponding metadata including taxonomic information and isolation metadata. Each uploaded genome is assigned a unique LIN which expresses the genomic relatedness to all other bacteria in LINbase. Users can hover the mouse over and select the highlighted LIN prefix shared by a group of genomes from the Website. Then users can describe these genomes as a LINgroup by clicking on the “Add description” button. With bacteria classified as LINgroups, unknown isolates can be identified as a member of a described LINgroup based on its assigned LIN.

Distance matrices and trees

Two-way pairwise ANI was calculated among the 55 genomes of the genus *Xylella* (**Supplementary Table 2**) with pyANI [62] and reciprocal ANI was calculated with a customized script to form the ANI matrices (**Figure 4.2A**). Trees were drawn based on the obtained ANI matrices with a hierarchical clustering algorithm (**Figure 4.2B**).

Results

Below we summarize plant pathogen LINgroup circumscriptions and descriptions entered in LINbase so far. Since we will continue to update these circumscriptions and plant pathologists with expertise in any bacterial plant pathogen are invited to contribute, this is simply a snapshot to show the utility of LINbase for bacterial plant pathogen classification and identification. It is by no means a complete list yet.

The genus *Xylella*

The species *Xylella fastidiosa* is an insect-transmitted xylem-inhabiting pathogen that is difficult (fastidious) to culture [102]. Because of the small genome size of bacteria in this species, it was the first plant pathogenic bacterial species to have its genome completely sequenced [103]. *X. fastidiosa* continues to exert its negative impact on crops world-wide, spreading from its center of origin in central America to more and more crops in more and more countries [104].

There are currently 55 genomes of the genus *Xylella* in LINbase. To illustrate how LINgroups correlate with ANI, we show in **Figure 4.2A** a distance matrix for these 55 genomes and in **Figure 4.2B** a distance tree based on the calculated ANI values. One can see in Figure 2A that all *Xylella* genomes are over 80% identical to each other. They thus share the same LIN prefix up to the C position ($15_A3_B0_C$), which corresponds to an ANI threshold of 80%. The genus *Xylella* was thus circumscribed as LINgroup $15_A3_B0_C$ (**Figure 4.2B**). The genus includes genomes of two species: *X. fastidiosa* and *X. taiwanensis*. The *X. fastidiosa* LINgroup was circumscribed based on the 53 *X. fastidiosa* genomes in LINbase that all share the $15_A3_B0_C0_D0_E0_F$ prefix (since they have pair-wise ANI values of at least 95% corresponding to the ANI threshold at position F). There are only two *X. taiwanensis* genomes in LINbase. They share over 99.95% ANI. However, instead of defining the *X. taiwanensis* LINgroup based on the LIN prefix shared by these two genomes ($15_A3_B0_C1_D0_E0_F0_G0_H0_I0_J0_K0_L0_M0_N0_O0_P0_Q$), we decided to use the standard species threshold of 95% ANI and circumscribe *X. taiwanensis* as $15_A3_B0_C1_D0_E0_F$. Our rationale is that two genomes are not representative of the diversity of an entire species and that it is likely that genetically more diverse *X. taiwanensis* genomes will be added to LINbase in the future.

Within *X. fastidiosa*, we circumscribed all current subspecies (**Figure 4.2A and 4.2B**) based on the manuscript by Denance and colleagues [105]. We also circumscribed the genetic lineage of *X. xylella* ssp. *pauca* that is spreading on olive in the Italian region of Puglia [106, 107]. We expect this LINgroup to be useful in identifying the source of new outbreaks in case this strain spreads to additional Italian regions or even other countries in the future.

The genus *Xanthomonas*

The genus *Xanthomonas* is closely related to the genus *Xylella* and comprises many economically important species of plant pathogenic bacteria that cause disease on dozens of crop plants [108]. Within some of the species, several subspecies and/or pathovars have been described as well.

At the time of writing, LINbase contained 615 genomes identified as members of the genus *Xanthomonas*. These genomes share the 15_A1_B prefix. Since *Xylella* and *Xanthomonas* are closely related genera that both belong to the family Xanthomoandaceae, it is not surprising that the LINgroup corresponding to the genus *Xylella*, $15_A3_B0_C$, and the LINgroup corresponding to the genus Xanthomonas, 15_A1_B, are identical at the A position. The overlap of their LIN prefixes at position A reveals that *Xylella* and *Xanthomonas* genomes are over 70% identical to each other.

Within the genus *Xanthomonas*, we circumscribed twelve species so far and we invite plant pathologists with expertise in *Xanthomonas* species to add additional circumscriptions. Within the *X. perforans* species, two intraspecific groups (1A and 2) were described based on a study by Schwartz and colleague [109].

To show the power of genome-based identification of plant pathogens using LINbase, we isolated a strain of *Xanthomonas* from a tomato seedling on the Eastern Shore of Virginia showing symptoms of bacterial spot but of unknown species affiliation. After sequencing and assembling the genome of this strain, the strain was precisely identified as *X. perforans* group 2 using the “Identify using a genome sequence” function of LINbase.

Pseudomonas syringae sensu lato

P. syringae sensu lato refers to all phytopathogenic Pseudomonads that are related to the validly published *P. syringae* species itself [110]. It includes some validly published species, such as *P. amygdali*, but also many strains that cluster into separate “genomospecies” based on DNA-DNA Hybridization (DDH) [111] and ANI [112] that have not been validly published because of absence of distinguishable phenotypic characteristics. *P. syringae* strains have also been grouped into pathovars based on their host range and the disease symptoms they cause [113] and into phylogroups based on MLST [98], which at least in part correspond to the earlier described genomospecies [110].

At the time of writing, 2964 genome sequences of the genus *Pseudomonas* were included in LINbase. These genomes all share the LIN prefix 50_A. *Pseudomonas* was thus circumscribed as LINgroup 50_A. Genomes of phytopathogenic Pseudomonads all share the LIN prefix $50_A1_B0_C$. *P. syringae sensu lato* was thus circumscribed as LINgroup $50_A1_B0_C$.

Various genomospecies and monophyletic pathovars within *P. syringae* have been circumscribed so far, for example, the three closely related pathovars *theae*, *actinidiae*, and *actinidifliorum*. As an example of how polyphyletic pathovars can be genomically circumscribed, we separately circumscribed the rare *P. syringae* pv. *tomato* DC3000 lineage and the common *P. syringae* pv. *tomato* T1-proper lineage based on our previous publications [114]. We finally circumscribed a genetic lineage that caused an outbreak of cantaloupe blight in France with isolates from cantaloupe and from various water sources [115].

Plant pathogenic *Enterobacteriaceae*

The Enterobacteriaceae are a very diverse family of non-pathogens and human, animal, and plant pathogens. The taxonomy of the Enterobacteriaceae underwent numerous revisions over the years because phenotypic characters were not able to distinguish strains that today are members of several different genera, which only became evident after the advent of 16S rRNA sequencing and MLST [116].

At the time of writing, LINbase contained 309 genomes of genera that belong to the plant pathogenic genera *Brenneria*, *Dickeya*, *Erwinia*, *Pantoea*, and *Pectobacterium* in the *Enterobacteriaceae* family (Lingroup 51_A). Each of these five genera were described as separate LINgroups. The case of *Brenneria* is a good example of how in some case a species needs to be described as a combination of two LINgroups because the diversity of the species is in between the ANI thresholds at two adjacent LIN positions and a second taxon is closely related. In the case of *Brenneria*, genomes of this species have the prefix $51_A4_B2_C$ or $51_A4_B3_C$ while genomes of the genus *Dickeya* have the prefix $51_A4_B1_C$. So far, LINbase is not set up to describe a LINgroup with an alternate position that would be needed here to circumscribe the genus *Brenneria*: $51_A4_B2/3_C$. Therefore, we described both LINgroups, $51_A4_B2_C$ and $51_A4_B3_C$, as genus *Brenneria*.

For some of the *Enterobacteriaceae* genera, we also described species. For example, in *Pantoea* we circumscribed the species *P. agglomerans*, *P. ananatis*, and *P. stewartii* as LINgroups. We invite plant pathologists with expertise in genomics of the *Enterobacteriaceae* to add cirumscriptions of additional species.

The genus *Ralstonia*

The species complex of *Ralstonia solanacearum* includes dozens of economically important pathogens of dozens of different crop species with partially overlapping host ranges [96]. A host range-based race classification system and a biochemistry-based biovar classification system were used in parallel in the past leading to strain identifications such as *R. solancearum* Race 3 Biovar 2, a cold-virulent tomato and potato pathogen that spread around the world out of South America beginning in the 1960s [117]. This pathogen represents a potential threat to the US potato industry. Since it is not established in the USA yet, it is considered as a biosecurity concern and is listed as a “Select Agent” announced by Center for Disease Control and Prevention (CDC) and United States Department of Agriculture (USDA) [118].

Based on phylogeny, the *R. solanacearum* species complex is divided into four phylotypes [97]. Two of these, phylotypes I and III, were recently validly published as a newly named species, *R. pseudosolanacearum* [119]. Since phylotype II includes the type strain of *R. solanacearum*, this phylotype corresponds to the validly published *R. solanacearum* species. Phylotype IV includes the type strain of the validly published species *R. syzygii* with three subspecies: *R. syzygii* subsp. *celebesensis*, *R. syzygii* subsp. *indonesiensis*, and *R. syzygii* subsp. *syzygii*.

At the time of writing, 99 genomes in LINbase were identified as members of the genus *Ralstonia*. Since these genomes share the prefix $14_A1_B0_C$, this prefix identifies the LINgroup corresponding to the genus *Ralstonia*. Within the *Ralstonia* LINgroup, we circumscribed all four *R. solanacearum* phylotypes as well as the species proposed by Safni and colleagues (same reference as above). While a thorough phenotypic characterization in regard to cold virulence is still in the works for the select agent *R. solanacearum* Race 3 Biovar 2, we provisionally circumscribed this pathogen as LINgroup by including genomically relatively divergent strains that may or may not be cold-virulent [117]. After completion of the cold virulence tests, this LINgroup will be updated if necessary.

Candidatus Liberibacter species and *Liberibacter crescens*

Since citrus greening emerged as a destructive disease first in Brazil [120] and the in Florida [121], the genus Liberibacter [122] has attracted much attention. The genus includes the so far unculturable *Candidatus* species Liberibacter asiaticus[123], causal agent of citrus greening in Florida and Brazil , *Candidatus* Liberibacter africanus [123], *Candidatus* Liberibacter americanus [124], *Candidatus* Liberibacter europeus [125], and *Candidatus*

Liberibacter solanacearum [126]. It also includes one culturable species, *Liberibacter crescens* [127].

At the time of writing, 31 genomes of members of the genus *Liberibacter* had been uploaded to LINbase. They all share the LIN prefix 8_A, which thus constitutes the LINgroup corresponding to the genus. Note that one genome that does not belong to the genus *Liberibacter* also has the LIN prefix 8_A by mistake. This probably happened because of problems when calculating ANI between genomes of small size, like the ones of *Liberibacter*, with genomes of larger size, like the one that was assigned the LIN prefix 8_A by mistake: *Phenylobacterium immobile*.

Most Liberibacter species only have a small number of genomes in LINbase and the corresponding LINgroups were thus circumscribed based on the 95% ANI threshold. However, there were enough genomes for two *Liberibacter* species to justify using the genomic diversity of members genomes for LINgroup circumscriptions. Sixteen genomes are of members of *Candidatus Liberibacter asiaticus*. These sixteen genomes are all over 99% identical to each other. Therefore *Candidatus* Liberibacter asiaticus was circumscribed as LINgroup $8_A1_B0_C0_D0_E0_F0_G0_H0_I0_J0_K$. *Candidatus* Liberibacter solanacearum has nine members that are over 97% identical to each other and this species was thus circumscribed as LINgroup $8_A0_B1_C0_D0_E0_F0_G0_H$.

Other plant pathogens are awaiting circumscriptions as LINgroups

Several other genomes of bacterial plant pathogens have already been uploaded to LINbase. As stated above, we invite the plant pathology community to circumscribe additional pathogens as LINgroups to make LINbase an ever more comprehensive resource for genome sequence-based identification of plant pathogenic bacteria.

Discussion

The ease of sequencing genomes of isolated bacterial plant pathogens or even the possibility to assemble entire pathogen genome sequences directly from metagenomes of symptomatic plants without the need of culturing is revolutionizing plant pathogen identification. However, to efficiently use the latest DNA sequencing technologies, whole genome databases that comprehensively circumscribe all plant pathogens are necessary for correct and precise genome sequenced-based identification of plant pathogens.

We have shown here how bacterial plant pathogens can be circumscribed precisely based on the genomes sequences of their members from the genus level all the way to the level of individual genetic lineages. The LINbase user interface makes it very easy for users to do this by selecting the LIN prefix shared by members of a plant pathogen taxon and to describe it as a LINgroup by adding a name, description, and URL. The commenting function allows other users to either endorse a LINgroup or to propose changes.

We have also shown here with an example of a sequenced genome of a *Xanthomonas* isolate from a symptomatic tomato seedling how straightforward it is to identify an unknown isolate as a member of a plant pathogen circumscribed as a LINgroup. A result can typically be obtained within a few minutes. Individual gene sequences can also be used as a query. However, the returned result is severely limited by the lack of precision when using a single gene sequence.

While LINbase has all the basic functions to be useful as a genome-based identification platform, there are limitations: 1. sometimes one LINgroup was not enough to circumscribe a genus or species and we had to add the same taxon description to two LINgroups, 2. sometimes not enough genomes to solidly circumscribe a group based on members were available and we had to describe a species based on a single strain using the standard 95% ANI species threshold, 3. taxon circumscriptions may need to be expanded by reducing the length of the LIN prefix of the corresponding LINgroup with the discovery of more diverse strains.

Also, additional improvement could be made to the web site by adding functions and the employed algorithms could be enhanced to increase speed. LINbase will remain a work in progress and its success will depend on the reception it receive from the plant pathology community. The more active users it will attract that will add taxon descriptions, the more useful LINbase will become. However, if the number of users increases, the current server will not be able to handle the traffic, and more investments will be needed to transfer the service to the cloud. The long term sustainability will of course depend on funding but if LINbase turns out to be useful to the community, we are confident that attracting funding will be possible.

Tables

Table 4.1 List of genera, species, and intraspecific groups of bacterial plant pathogens circumscribed as LINgroups in LINbase at the time of writing.

LINgroup	Type	Name
8	genus	*Liberibacter*
8,0,0,1,0,0	species	*Candidatus Liberibacter europaeus*
8,0,1,0,0,0,0,0	species	*Candidatus Liberibacter solanacearum*
8,0,2,0,0,0	species	*Candidatus Liberibacter americanus*
8,1,0,0,0,0,0,0,0,0,0	species	*Candidatus Liberibacter asiaticus*
8,1,0,1,0,0	species	*Candidatus Liberibacter africanus*
8,2,0,0,0,0	species	*Liberibacter crescens*
14,1,0	genus	*Ralstonia*
14,1,0,0,0,0	species	*Ralstonia pseudosolanacearum*
14,1,0,0,0,0,0,0,0,0	phylotype	I
14,1,0,0,0,0,1	phylotype	III
14,1,0,0,0,1	phylotype	II (validly published Ralstonia solanacearum)
14,1,0,0,0,1,0,0,0,0	Non-taxonomic group	preliminary race 3 biovar 2 circumscription
14,1,0,0,0,2,0,0	phylotype	IV (Ralstonia syzygii)
14,1,0,1,1	species	*Ralstonia pickettii*
14,1,0,1,2	species	*Ralstonia mannitolilytica*
15,1	genus	*Xanthomonas*
15,1,0,1,0,0,0,0,0,0	species	*Xanthomonas vasicola*
15,1,0,1,0,1,0,0	species	*Xanthomonas oryzae*
15,1,0,1,1,1,0,0,0,0,2,0,0	species	*Xanthomonas perforans*
15,1,0,1,1,1,0,0,0,0,2,0,0,0,2,0,0	Non-taxonomic group	*X. perforans* Group 2
15,1,0,1,1,1,0,0,0,0,2,0,0,0,4,0,0,0	Non-taxonomic group	*X. perforans* Group 1A
15,1,0,1,1,1,0,0,0,0,3,0	species	*Xanthomonas euvesicatoria*
15,1,0,1,1,2,0,0	species	*Xanthomonas phaseoli*
15,1,0,1,2,0	species	*Xanthomonas arboricola*
15,1,0,1,3,0,0,0,0,0,0,0,0	species	*Xanthomonas fragariae*
15,1,0,1,6,0,1,0	pathovar	*Xanthomonas cynarae pv. gardneri*
15,1,0,1,8,0,0,0,0,0	species	*Xanthomonas vesicatoria*
15,1,0,1,10,0,0	species	*Xanthomonas cannabis*
15,1,1,0,0,0	species	*Xanthomonas translucens*
15,1,1,1,0,0,0,0,0	species	*Xanthomonas albilineans*

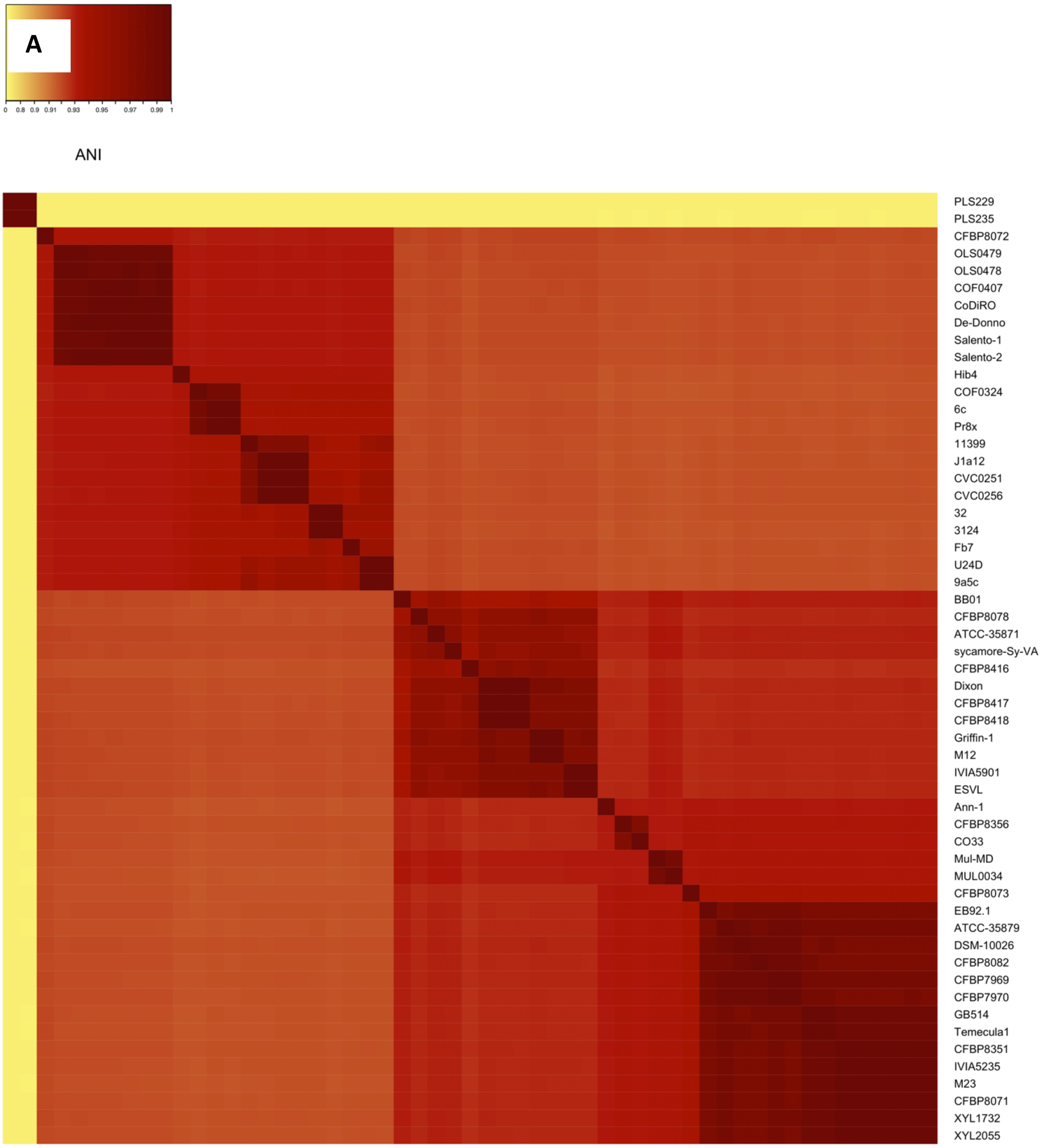
15,3,0	genus	*Xylella*
15,3,0,1,0,0	species	*Xylella taiwanensis*
15,3,0,0,0,0	species	*Xylella fastidiosa*
15,3,0,0,0,0,0,0,0,3,0,0,0,0	subspecies	*X.f.* ssp. morus
15,3,0,0,0,0,0,0,0,0	subspecies	*X.f.* ssp. fastidiosa
15,3,0,0,0,0,0,0,0,1,0,0,0	subspecies	*X.f.* ssp. sandyi-like
15,3,0,0,0,0,0,0,0,2,0,0,0	subspecies	*X.f.* ssp. sandyi
15,3,0,0,0,0,0,0,1,0	subspecies	*X.f.* ssp. multiplex
15,3,0,0,0,0,1,0,0	subspecies	*X.f.* ssp. pauca
15,3,0,0,0,0,1,0,0,1,0,0,0,0,0,0,0,0	Non-taxonomic group	*X.f.* ssp. pauca - Italy
50	genus	*Pseudomonas*
50,1,0	species	*Pseudomonas syringae*
50,1,0,0,1,0,0,0,1,0	pathovar	actinidifoliorum
50,1,0,0,1,0,0,0,0,0,0	pathovar	actinidiae
50,1,0,0,0,0,0,0,0,3,1,0,0,0	Non-taxonomic group	Isolates from a cantaloupe blight epidemic in France
50,1,0,0,1,0,1,0,0,0,3,0,0,0,0,0,0,0	Non-taxonomic group	lineage DC3000 of pathovar tomato
50,1,0,0,1,0,1,0,0,0,4,0,0,0	Non-taxonomic group	lineage T1-proper of pathovar tomato
50,1,0,0,4	genomospecies	*Pseudomonas syringae* genomospecies 9, which corresponds approximately to *P. cannabina*.
51	family	Plant-pathogenic *Enterobacteriaceae* of the genera *Pantoea, Erwinia, Pectobacterium, Dickeya*, and *Brenneria.*
51,0,0,1,0,0,0,0,0	species	*Erwinia amylovora*
51,4,0,0	genus	*Pectobacterium*
51,4,1	genus	*Dickeya*
51,4,2	genus	*Brenneria*
51,4,3	genus	*Brenneria*
51,6	genus	*Pantoea*
51,6,1,1,0,0,0,0,0,0	species	*Pantoea stewartii*
51,6,1,0,0,0,0	species	*Pantoea ananatis*
51,6,0,0,0,0,0,0	species	*Pantoea agglomerans*

Figures

Figure 4.1 The LINgroup concept. Taxa are generally circumscribed as LINgroups based on their member genomes. The name of a LINgroup is the LIN prefix shared by the members of a LINgroup. In cases where only a small number of genomes of members of a species have been sequenced, the standard 95% ANI threshold of bacterial species (corresponding to LIN position F) may be applied instead.

			70%	75%	80%	85%	90%	95%	96%	97%	98%	98.5%	99%	99.25%	99.5%	99.75%	99.9%	99.925%	99.95%	99.975%	99.99%	99.999%
Genus	Species	Strain	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
G1	S1	X1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
G1	S2	X2	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
G1	S2	X3	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
G1	S3	X4	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
G1	S3	X5	0	1	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
G1	S3	X6	0	1	0	0	0	1	1	0	0	0	0	0	0	0	1	0	0	0	0	0

Figure 4.2 Distance matrix, distance tree, LINs, and LINgroups for members of the genus *Xylella*. A The distance matrix of 55 *Xylella* genomes derived from their pairwise ANI values is visualized as a heatmap. Species of *X. taiwanensis* and *X. fastidiosa* constitute the 2 major blocks in the heatmap along the diagonal. Within the block of *X. fastidiosa*, there are blocks with higher similarity corresponding to the existences of subspecies. **B** A hierarchical clustering dendrogram based on the *Xylella* distance matrix. Each tip of the tree representing a genome corresponds to its LIN assigned by LINbase. 9 described LINgroups including 2 species highlighted by bold font and 7 subspecies of *Xylella* highlighted by different colors.



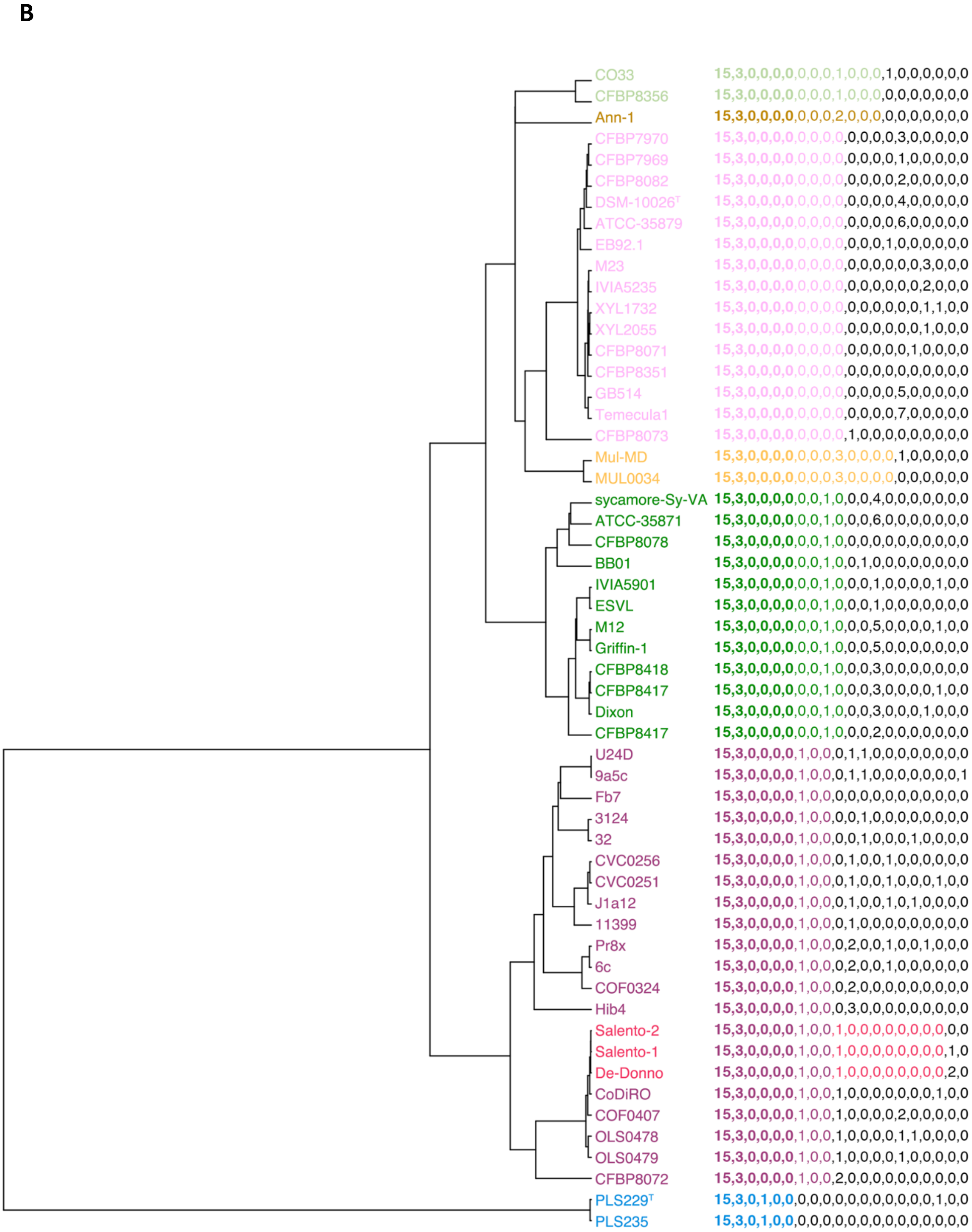
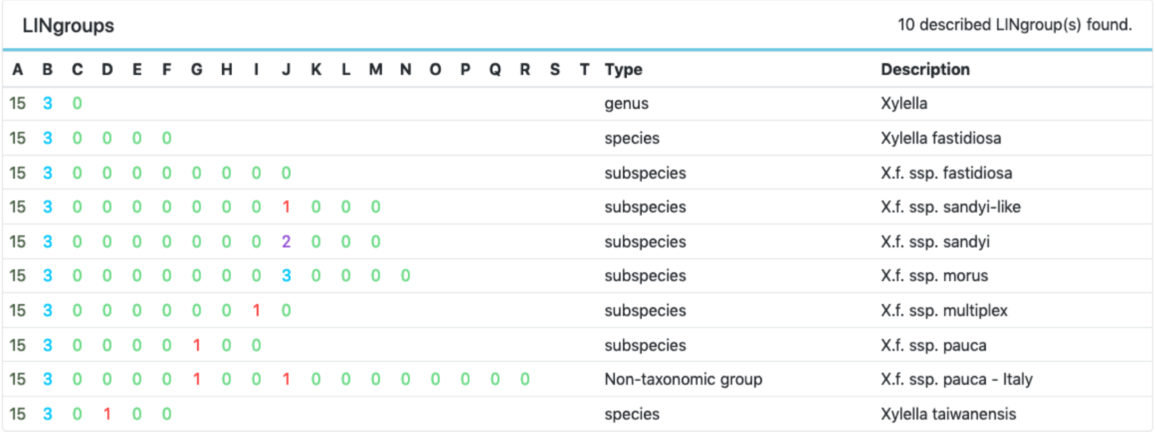

B
CO33 15,3,0,0,0,0,0,0,0,1,0,0,0,1,0,0,0,0,0,0
CFBP8356 15,3,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0
Ann-1 15,3,0,0,0,0,0,0,0,2,0,0,0,0,0,0,0,0,0,0
CFBP7970 15,3,0,0,0,0,0,0,0,0,0,0,0,0,3,0,0,0,0,0
CFBP7969 15,3,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0
CFBP8082 15,3,0,0,0,0,0,0,0,0,0,0,0,0,2,0,0,0,0,0
DSM-10026T 15,3,0,0,0,0,0,0,0,0,0,0,0,0,4,0,0,0,0,0
ATCC-35879 15,3,0,0,0,0,0,0,0,0,0,0,0,0,6,0,0,0,0,0
EB92.1 15,3,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0
M23 15,3,0,0,0,0,0,0,0,0,0,0,0,0,0,0,3,0,0,0
IVIA5235 15,3,0,0,0,0,0,0,0,0,0,0,0,0,0,0,2,0,0,0
XYL1732 15,3,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,1,0,0
XYL2055 15,3,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0
CFBP8071 15,3,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0
CFBP8351 15,3,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
GB514 15,3,0,0,0,0,0,0,0,0,0,0,0,0,5,0,0,0,0,0
Temecula1 15,3,0,0,0,0,0,0,0,0,0,0,0,0,7,0,0,0,0,0
CFBP8073 15,3,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0
Mul-MD 15,3,0,0,0,0,0,0,0,3,0,0,0,0,1,0,0,0,0,0
MUL0034 15,3,0,0,0,0,0,0,0,3,0,0,0,0,0,0,0,0,0,0
sycamore-Sy-VA 15,3,0,0,0,0,0,0,1,0,0,0,4,0,0,0,0,0,0,0
ATCC-35871 15,3,0,0,0,0,0,0,1,0,0,0,6,0,0,0,0,0,0,0
CFBP8078 15,3,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0
BB01 15,3,0,0,0,0,0,0,1,0,0,1,0,0,0,0,0,0,0,0
IVIA5901 15,3,0,0,0,0,0,0,1,0,0,0,1,0,0,0,0,1,0,0
ESVL 15,3,0,0,0,0,0,0,1,0,0,0,1,0,0,0,0,0,0,0
M12 15,3,0,0,0,0,0,0,1,0,0,0,5,0,0,0,0,1,0,0
Griffin-1 15,3,0,0,0,0,0,0,1,0,0,0,5,0,0,0,0,0,0,0
CFBP8418 15,3,0,0,0,0,0,0,1,0,0,0,3,0,0,0,0,0,0,0
CFBP8417 15,3,0,0,0,0,0,0,1,0,0,0,3,0,0,0,0,1,0,0
Dixon 15,3,0,0,0,0,0,0,1,0,0,0,3,0,0,0,1,0,0,0
CFBP8417 15,3,0,0,0,0,0,0,1,0,0,0,2,0,0,0,0,0,0,0
U24D 15,3,0,0,0,0,1,0,0,0,1,1,0,0,0,0,0,0,0,0
9a5c 15,3,0,0,0,0,1,0,0,0,1,1,0,0,0,0,0,0,0,1
Fb7 15,3,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0
3124 15,3,0,0,0,0,1,0,0,0,0,1,0,0,0,0,0,0,0,0
32 15,3,0,0,0,0,1,0,0,0,0,1,0,0,0,1,0,0,0,0
CVC0256 15,3,0,0,0,0,1,0,0,0,1,0,0,1,0,0,0,0,0,0
CVC0251 15,3,0,0,0,0,1,0,0,0,1,0,0,1,0,0,0,1,0,0
J1a12 15,3,0,0,0,0,1,0,0,0,1,0,0,1,0,1,0,0,0,0
11399 15,3,0,0,0,0,1,0,0,0,1,0,0,0,0,0,0,0,0,0
Pr8x 15,3,0,0,0,0,1,0,0,0,2,0,0,1,0,0,1,0,0,0
6c 15,3,0,0,0,0,1,0,0,0,2,0,0,1,0,0,0,0,0,0
COF0324 15,3,0,0,0,0,1,0,0,0,2,0,0,0,0,0,0,0,0,0
Hib4 15,3,0,0,0,0,1,0,0,0,3,0,0,0,0,0,0,0,0,0
Salento-2 15,3,0,0,0,0,1,0,0,1,0,0,0,0,0,0,0,0,0,0
Salento-1 15,3,0,0,0,0,1,0,0,1,0,0,0,0,0,0,0,0,1,0
De-Donno 15,3,0,0,0,0,1,0,0,1,0,0,0,0,0,0,0,0,2,0
CoDiRO 15,3,0,0,0,0,1,0,0,1,0,0,0,0,0,0,0,1,0,0
COF0407 15,3,0,0,0,0,1,0,0,1,0,0,0,0,2,0,0,0,0,0
OLS0478 15,3,0,0,0,0,1,0,0,1,0,0,0,0,1,1,0,0,0,0
OLS0479 15,3,0,0,0,0,1,0,0,1,0,0,0,0,1,0,0,0,0,0
CFBP8072 15,3,0,0,0,0,1,0,0,2,0,0,0,0,0,0,0,0,0,0
PLS229T 15,3,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0
PLS235 15,3,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0

Figure 4.3 Screenshot of the LINbase page showing the LINgroup circumscriptions of the genus Xylella. 10 LINgroups, including the genus *Xylella*, the species of *X. taiwanensi*s and *X. fastidiosa*, and 7 subspecies have been described on LINbase.



LINgroups

10 described LINgroup(s) found.

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	Type	Description
15	3	0																		genus	Xylella
15	3	0	0	0	0															species	Xylella fastidiosa
15	3	0	0	0	0	0	0	0	0											subspecies	X.f. ssp. fastidiosa
15	3	0	0	0	0	0	0	0	1	0	0	0								subspecies	X.f. ssp. sandyi-like
15	3	0	0	0	0	0	0	0	2	0	0	0								subspecies	X.f. ssp. sandyi
15	3	0	0	0	0	0	0	0	3	0	0	0	0							subspecies	X.f. ssp. morus
15	3	0	0	0	0	0	0	1	0											subspecies	X.f. ssp. multiplex
15	3	0	0	0	0	1	0	0												subspecies	X.f. ssp. pauca
15	3	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0			Non-taxonomic group	X.f. ssp. pauca - Italy
15	3	0	1	0	0															species	Xylella taiwanensis

Conclusion

Bacteria constitute about 15% of the Earth's total biomass and are distributed across all types of environments: terrestrial, marine, and deep subsurface [128]. Taxonomy, being the science of systematically classifying and naming organisms, faces challenges in bacteria. The decades-long debate on the definition of a bacterial species certainly hindered the development of bacterial taxonomy and caused many misclassifications due to the limitations of technology over these years. With the pragmatic species concept developed in 2001 and sequencing technology and computational power brought into taxonomy, we are now looking for solutions to (1) correct existing misclassifications, (2) describe all species on Earth, and (3) classify bacteria within species, the fundamental unit of taxonomy.

Names of some misclassified bacteria remain unchanged even though these bacteria have been reassigned to correct taxa. This causes confusions because their names no longer suggest their positions on the tree of life. In this sense, it would be the best if all species described in the future are classified correctly with stable names. Currently, to publish a new species, laborious biochemical tests are performed on cultivated bacterial colonies in a laboratory environment, and then the new species needs to be published as a manuscript. This process could take months, and among the 61 known bacterial phyla, 31 of them are unculturable with traditional methods [129]. It has been suggested that genomic data have the potential to replace the required biochemical tests with genome-based classification methods since the tested phenotypes are regulated by the genotypes, and the speed of publishing a new species can be accelerated. Furthermore, metagenomic sequencing technologies make it possible to harvest genome assemblies of unculturable bacteria, hence describing unculturable bacterial species has become easier. Subspecies and intra-specific groups exist within species and they are used to describe bacteria for different purposes using different criteria. While they are classified with different schemes, genome-based classification measurements, *e.g.* ANI, have the accuracy and resolution to differentiate them from each other. We introduced LINbase, together with its algorithm and applications in describing plant pathogenic bacteria, as the approach to reach all of the aforementioned goals. The genome sequence is thought to be the ultimate unique identifier for each individual organism, and LINbase uses the fast, precise, and memory-efficient LINflow algorithm and the Life Identification Number framework to "compress" each bacterial isolate genome to a unique 20-position long LIN that stably and straightforwardly represents the corresponding isolate's genomic relatedness to all other bacteria in LINbase. The LINgroup approach has proven itself as effective in describing taxa, from the genus level to subspecies level, based on, as the pragmatic bacterial species concept states, the genomic coherence of a cluster of bacteria rather than the characteristics of a single strain. The "crowdsourcing LINgroup description function in LINbase improves the accessibility of bacterial taxonomy and encourages scientists to share their knowledge and wisdom of bacteria. The commenting system, too, facilitates the communication of scientists as well as collaboration across laboratories globally.

At the time of writing, LINbase cannot be used to describe taxa at higher taxonomic ranks (family, class, order, phylum or kingdom) as a single LINgroup because ANI is not suitable for exploring inter-family relationship. It would also be difficult for LINbase to differentiate very closely related bacteria and their difference is beyond ANI's resolution. We hope that in the future, a new genomic relationship measurement, perhaps an extension of ANI will be developed and can be used as a universal measurement to classify bacteria at all levels. With that, LINbase could be used to describe the whole kingdom of bacteria comprehensively.

LINbase, with its potential of circumscribing all bacterial taxa and describing non-taxonomic intraspecific groups of bacteria with shared phenotypes, is able to contribute to multiple areas of bacteriology. It can be envisioned to be the digital encyclopedia of bacteriology for researchers and students who need easy access to correct, comprehensive, and stable knowledge of bacteria. With pathogenic groups circumscribed as LINgroups, LINbase can help to build up pathogen surveillance systems not only for plant pathogens but also for human and animal pathogens. Combined with cutting edge sequencing technologies such as Oxford Nanopore allowing on-site and real time sequencing of pathogenic isolates within a few hours after sample collection, LINbase, with its easy access on the Internet, can reduce the time to identify pathogens from outbreaks and traceback their sources. Metagenomic sequencing technologies provide us with the opportunity to investigate how microbial communities interact with the habitat or environment such as the animal gut, soil, ocean, and so on. The reference database is the key part to achieve precise identification of metagenomics data, however, public reference databases are using the current bacterial taxonomy with misclassified and mis-named taxa. LINbase can be used to build a reference database that is not only correct, but classified at high resolution as well. Last but not least, we hope that LINbase leads the evolution of bacterial taxonomy to recognize digital entries as valid publications and the unique LIN of each bacterium can be used across different databases and platforms as a bridge to connect taxonomic data, isolation data, and genomic data.

References

1. Schildkraut, C.L., J. Marmur, and P. Doty, *The formation of hybrid DNA molecules and their use in studies of DNA homologies.* Journal of Molecular Biology, 1961. **3**(5): p. 595-IN16.
2. L. G. Wayne, D.J.B., R. R. Colwell, P. A. D. Grimont, O. Kandler, M. I. Krichevsky, L. H. Moore, W. E. C. Moore, R. G. E. Murray, E. Stackebrandt, M. P. Starr, H. G. Truper, *Report of the Ad Hoc Committee on Reconciliation of Approaches to Bacterial Systematics.* International Journal of Systematic Bacterialogy, 1987. **Oct. 1987**.
3. Sneath, P.H.A., *Analysis and Interpretation of Sequence Data for Bacterial Systematics: The View of a Numerical Taxonomist.* Systematic and Applied Microbiology, 1989. **12**(1): p. 15-31.
4. Stackebrandt, E., *The Richness of Prokaryotic Diversity: There Must be a Species Somewhere.* Food Technology and Biotechnology, 2003. **41**(1): p. 17-22.
5. Fox, G.E., J.D. Wisotzkey, and P. Jurtshuk, Jr., *How close is close: 16S rRNA sequence identity may not be sufficient to guarantee species identity.* Int J Syst Bacteriol, 1992. **42**(1): p. 166-70.
6. Maiden, M.C.J., J.A. Bygraves, E. Feil, G. Morelli, J.E. Russell, R. Urwin, Q. Zhang, J. Zhou, K. Zurth, D.A. Caugant, I.M. Feavers, M. Achtman, and B.G. Spratt, *Multilocus sequence typing: A portable approach to the identification of clones within populations of pathogenic microorganisms.* Proceedings of the National Academy of Sciences of the United States of America, 1998. **95**(6): p. 3140-3145.
7. Vandamme, P. and C. Peeters, *Time to revisit polyphasic taxonomy.* Antonie van Leeuwenhoek, 2014. **106**(1): p. 57-65.
8. Nielsen, R., J.S. Paul, A. Albrechtsen, and Y.S. Song, *Genotype and SNP calling from next-generation sequencing data.* Nature reviews. Genetics, 2011. **12**(6): p. 443-451.
9. Klemm, E. and G. Dougan, *Advances in Understanding Bacterial Pathogenesis Gained from Whole-Genome Sequencing and Phylogenetics.* Cell Host & Microbe, 2016. **19**(5): p. 599-610.
10. Chun, J. and F.A. Rainey, *Integrating genomics into the taxonomy and systematics of the Bacteria and Archaea.* International Journal of Systematic and Evolutionary Microbiology, 2014. **64**(2): p. 316-324.
11. Sutcliffe, I.C., *Challenging the anthropocentric emphasis on phenotypic testing in prokaryotic species descriptions: rip it up and start again.* Frontiers in Genetics, 2015. **6**(218).
12. Alneberg, J., C.M.G. Karlsson, A.-M. Divne, C. Bergin, F. Homa, M.V. Lindh, L.W. Hugerth, T.J.G. Ettema, S. Bertilsson, A.F. Andersson, and J. Pinhassi, *Genomes from uncultivated prokaryotes: a comparison of metagenome-assembled and single-amplified genomes.* Microbiome, 2018. **6**(1): p. 173.
13. Bowers, R.M., N.C. Kyrpides, R. Stepanauskas, M. Harmon-Smith, D. Doud, T.B.K. Reddy, F. Schulz, J. Jarett, A.R. Rivers, E.A. Eloe-Fadrosh, S.G. Tringe, N.N. Ivanova, A. Copeland, A. Clum, E.D. Becraft, R.R. Malmstrom, B. Birren, M. Podar, P. Bork, G.M. Weinstock, G.M. Garrity, J.A. Dodsworth, S. Yooseph, G. Sutton, F.O. Glöckner, J.A. Gilbert, W.C. Nelson, S.J. Hallam, S.P. Jungbluth, T.J.G. Ettema, S. Tighe, K.T. Konstantinidis, W.-T. Liu, B.J. Baker, T. Rattei, J.A. Eisen, B. Hedlund, K.D. McMahon, N. Fierer, R. Knight, R. Finn, G. Cochrane, I. Karsch-Mizrachi, G.W. Tyson, C. Rinke, C. Genome Standards, A. Lapidus, F. Meyer, P. Yilmaz, D.H. Parks, A.M. Eren, L. Schriml, J.F. Banfield, P. Hugenholtz, and T. Woyke, *Minimum information about a single amplified genome (MISAG) and a metagenome-assembled*

genome (MIMAG) of bacteria and archaea. Nature biotechnology, 2017. **35**(8): p. 725-731.

14. Wilkins, L.G.E., C.L. Ettinger, G. Jospin, and J.A. Eisen, *Metagenome-assembled genomes provide new insight into the microbial diversity of two thermal pools in Kamchatka, Russia.* Scientific Reports, 2019. **9**(1): p. 3059.
15. Konstantinidis, K.T. and J.M. Tiedje, *Genomic insights that advance the species definition for prokaryotes.* 2005. **102**(7): p. 2567-2572.
16. Stapley, J., P.G.D. Feulner, S.E. Johnston, A.W. Santure, and C.M. Smadja, *Recombination: the good, the bad and the variable.* Philosophical transactions of the Royal Society of London. Series B, Biological sciences, 2017. **372**(1736): p. 20170279.
17. Bohlin, J. and J.H.O. Pettersson, *Evolution of Genomic Base Composition: From Single Cell Microbes to Multicellular Animals.* Computational and structural biotechnology journal, 2019. **17**: p. 362-370.
18. Palmer, M., S.N. Venter, M.P.A. Coetzee, and E.T. Steenkamp, *Prokaryotic species are sui generis evolutionary units.* Systematic and Applied Microbiology, 2019. **42**(2): p. 145-158.
19. Bobay, L.-M. and H. Ochman, *Impact of Recombination on the Base Composition of Bacteria and Archaea.* Molecular biology and evolution, 2017. **34**(10): p. 2627-2636.
20. Didelot, X. and M.C.J. Maiden, *Impact of recombination on bacterial evolution.* Trends in microbiology, 2010. **18**(7): p. 315-322.
21. Rossello-Mora, R. and R. Amann, *The species concept for prokayotes.* FEMS Microbiol. Rev., 2001. **25**: p. 39-67.
22. Yutin, N. and M.Y. Galperin, *A genomic update on clostridial phylogeny: Gram-negative spore formers and other misplaced clostridia.* Environmental microbiology, 2013. **15**(10): p. 2631-2641.
23. Parks, D.H., M. Chuvochina, D.W. Waite, C. Rinke, A. Skarshewski, P.-A. Chaumeil, and P. Hugenholtz, *A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life.* Nature Biotechnology, 2018. **36**: p. 996.
24. Mallet, J., *A species definition for the modern synthesis.* Trends in Ecology & Evolution, 1995. **10**(7): p. 294-299.
25. Cowan, S.T., *Principles and Practice of Bacterial Taxonomy—a Forward Look.* Microbiology, 1965. **39**(1): p. 143-153.
26. Lapage SP, S.P., Lessel EF, et al., *International Code of Nomenclature of Bacteria: Bacteriological Code.* 1992, Washington (DC): ASM Press.
27. Van Ert, M.N., W.R. Easterday, L.Y. Huynh, R.T. Okinaka, M.E. Hugh-Jones, J. Ravel, S.R. Zanecki, T. Pearson, T.S. Simonson, J.M. U'Ren, S.M. Kachur, R.R. Leadem-Dougherty, S.D. Rhoton, G. Zinser, J. Farlow, P.R. Coker, K.L. Smith, B. Wang, L.J. Kenefic, C.M. Fraser-Liggett, D.M. Wagner, and P. Keim, *Global Genetic Population Structure of Bacillus anthracis.* PLOS ONE, 2007. **2**(5): p. e461.
28. Keim, P., J.M. Gruendike, A.M. Klevytska, J.M. Schupp, J. Challacombe, and R. Okinaka, *The genome and variation of Bacillus anthracis.* Molecular Aspects of Medicine, 2009. **30**(6): p. 397-405.
29. Lukjancenko, O., T.M. Wassenaar, and D.W. Ussery, *Comparison of 61 Sequenced Escherichia coli Genomes.* Microbial Ecology, 2010. **60**(4): p. 708-720.
30. J.M. Young, C.T.B., S.H. De Boer, G. Firrao, L. Gardan, G.E. Saddler, D.E. Stead and Y. Takikawa, *International Standards for Naming Pathovars of Phytopathogenic Bacteria.* 2001.
31. Fegan, M. and P. Prior, *How Complex is the "Ralstonia Solanacearum Species Complex.* 2005.

32. EJ, B., *Classification*, in *Baron's Medical Microbiology*. 1996, University of Texas Medical Branch.

33. Varghese, N.J., S. Mukherjee, N. Ivanova, K.T. Konstantinidis, K. Mavrommatis, N.C. Kyrpides, and A. Pati, *Microbial species delineation using whole genome sequences.* Nucleic Acids Res, 2015. **43**(14): p. 6761-71.

34. Marakeby, H., E. Badr, H. Torkey, Y. Song, S. Leman, C.L. Monteil, L.S. Heath, and B.A. Vinatzer, *A System to Automatically Classify and Name Any Individual Genome-Sequenced Organism Independently of Current Biological Classification and Nomenclature.* PLOS ONE, 2014. **9**(2): p. e89142.

35. Vinatzer, B.A., L. Tian, and L.S. Heath, *A proposal for a portal to make earth's microbial diversity easily accessible and searchable.* Antonie van Leeuwenhoek, 2017. **110**(10): p. 1271-1279.

36. Altschul, S.F., W. Gish, W. Miller, E.W. Myers, and D.J. Lipman, *Basic local alignment search tool.* Journal of Molecular Biology, 1990. **215**(3): p. 403-410.

37. Rodriguez-R, L. and K. Konstantinidis, *Bypassing Cultivation To Identify Bacterial Species.* Microbe Magazine.

38. Konstantinidis, K.T. and J.M. Tiedje, *Towards a Genome-Based Taxonomy for Prokaryotes.* 2005. **187**(18): p. 6258-6264.

39. Miyoshi-Akiyama, T., K. Hayakawa, N. Ohmagari, M. Shimojima, and T. Kirikae, *Multilocus Sequence Typing (MLST) for Characterization of Enterobacter cloacae.* PLOS ONE, 2013. **8**(6): p. e66358.

40. Jolley, K.A., C.M. Bliss, J.S. Bennett, H.B. Bratcher, C. Brehony, F.M. Colles, H. Wimalarathna, O.B. Harrison, S.K. Sheppard, A.J. Cody, and M.C.J. Maiden, *Ribosomal multilocus sequence typing: universal characterization of bacteria from domain to strain.* Microbiology (Reading, England), 2012. **158**(Pt 4): p. 1005-1015.

41. Maiden, M.C.J., M.J. Jansen van Rensburg, J.E. Bray, S.G. Earle, S.A. Ford, K.A. Jolley, and N.D. McCarthy, *MLST revisited: the gene-by-gene approach to bacterial genomics.* Nature reviews. Microbiology, 2013. **11**(10): p. 728-736.

42. Broder, A.Z. *On the resemblance and containment of documents*. in *Proceedings. Compression and Complexity of SEQUENCES 1997 (Cat. No.97TB100171)*. 1997.

43. Ondov, B.D., T.J. Treangen, P. Melsted, A.B. Mallonee, N.H. Bergman, S. Koren, and A.M. Phillippy, *Mash: fast genome and metagenome distance estimation using MinHash.* Genome biology, 2016. **17**(1): p. 132-132.

44. Jain, C., S. Koren, A. Dilthey, A.M. Phillippy, and S. Aluru, *A fast adaptive algorithm for computing whole-genome homology maps.* Bioinformatics, 2018. **34**(17): p. i748-i756.

45. Jain, C., L.M. Rodriguez-R, A.M. Phillippy, K.T. Konstantinidis, and S. Aluru, *High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries.* Nature Communications, 2018. **9**(1): p. 5114.

46. Schürch, A.C., S. Arredondo-Alonso, R.J.L. Willems, and R.V. Goering, *Whole genome sequencing options for bacterial strain typing and epidemiologic analysis based on single nucleotide polymorphism versus gene-by-gene–based approaches.* Clinical Microbiology and Infection, 2018. **24**(4): p. 350-354.

47. Davis, S., J.B. Pettengill, Y. Luo, J. Payne, A. Shpuntoff, H. Rand, and E. Strain, *CFSAN SNP Pipeline: an automated method for constructing SNP matrices from next-generation sequence data.* PeerJ Computer Science, 2015. **1**: p. e20.

48. Bertels, F., O.K. Silander, M. Pachkov, P.B. Rainey, and E. van Nimwegen, *Automated Reconstruction of Whole-Genome Phylogenies from Short-Sequence Reads.* Molecular Biology and Evolution, 2014. **31**(5): p. 1077-1088.

49. Pightling, A.W., N. Petronella, and F. Pagotto, *Choice of Reference Sequence and Assembler for Alignment of Listeria monocytogenes Short-Read Sequence Data Greatly Influences Rates of Error in SNP Analyses.* PLOS ONE, 2014. **9**(8): p. e104579.
50. Benson, D.A., M. Cavanaugh, K. Clark, I. Karsch-Mizrachi, D.J. Lipman, J. Ostell, and E.W. Sayers, *GenBank.* Nucleic acids research, 2013. **41**(Database issue): p. D36-D42.
51. Chen, I.M.A., V.M. Markowitz, K. Chu, K. Palaniappan, E. Szeto, M. Pillay, A. Ratner, J. Huang, E. Andersen, M. Huntemann, N. Varghese, M. Hadjithomas, K. Tennessen, T. Nielsen, N.N. Ivanova, and N.C. Kyrpides, *IMG/M: integrated genome and metagenome comparative data analysis system.* Nucleic acids research, 2017. **45**(D1): p. D507-D516.
52. Brian Bushnell, J.R., Esther Singer, *BBMap short read aligner and other bioinformatic tools*. 2017.
53. Meier-Kolthoff, J.P., A.F. Auch, H.-P. Klenk, and M. Göker, *Genome sequence-based species delimitation with confidence intervals and improved distance functions.* BMC Bioinformatics, 2013. **14**(1): p. 60.
54. Meier-Kolthoff, J.P., R.L. Hahnke, J. Petersen, C. Scheuner, V. Michael, A. Fiebig, C. Rohde, M. Rohde, B. Fartmann, L.A. Goodwin, O. Chertkov, T.B.K. Reddy, A. Pati, N.N. Ivanova, V. Markowitz, N.C. Kyrpides, T. Woyke, M. Göker, and H.-P. Klenk, *Complete genome sequence of DSM 30083T, the type strain (U5/41T) of Escherichia coli, and a proposal for delineating subspecies in microbial taxonomy.* Standards in Genomic Sciences, 2014. **9**(1): p. 2.
55. Wayne, L.G., D.J. Brenner, R.R. Colwell, P.A.D. Grimont, M.I. Krichevsky, and H.G. Truper, *Report of the Ad Hoc Committee on Reconciliation of Approaches to Bacterial Systematics.* 1987: p. 463-464.
56. Richter, M. and R. Rosselló-Móra, *Shifting the genomic gold standard for the prokaryotic species definition.* Proceedings of the National Academy of Sciences of the United States of America, 2009. **106**(45): p. 19126-19131.
57. Delcher, A.L., S. Kasif, R.D. Fleischmann, J. Peterson, O. White, and S.L. Salzberg, *Alignment of whole genomes.* Nucleic acids research, 1999. **27**(11): p. 2369-2376.
58. Delcher, A.L., A. Phillippy, J. Carlton, and S.L. Salzberg, *Fast algorithms for large-scale genome alignment and comparison.* Nucleic Acids Research, 2002. **30**(11): p. 2478-2483.
59. Kurtz, S., A. Phillippy, A.L. Delcher, M. Smoot, M. Shumway, C. Antonescu, and S.L. Salzberg, *Versatile and open software for comparing large genomes.* Genome Biology, 2004. **5**(2): p. R12.
60. Richter, M., R. Rosselló-Móra, F. Oliver Glöckner, and J. Peplies, *JSpeciesWS: a web server for prokaryotic species circumscription based on pairwise genome comparison.* Bioinformatics (Oxford, England), 2016. **32**(6): p. 929-931.
61. Goris, J., K.T. Konstantinidis, J.A. Klappenbach, T. Coenye, P. Vandamme, and J.M. Tiedje, *DNA–DNA hybridization values and their relationship to whole-genome sequence similarities.* International Journal of Systematic and Evolutionary Microbiology, 2007. **57**(1): p. 81-91.
62. Pritchard, L., *pyani: Python module for average nucleotide identity analyses*. 2014: https://github.com/widdowquinn/pyani.
63. C. Titus Brown, L.I., *sourmash: a library for MinHash sketching of DNA*. 2016.
64. Jolley, K.A. and M.C.J. Maiden, *BIGSdb: Scalable analysis of bacterial genome variation at the population level.* BMC Bioinformatics, 2010. **11**(1): p. 595.
65. Rodriguez-R, L.M., S. Gunturu, W.T. Harvey, R. Rosselló-Mora, J.M. Tiedje, J.R. Cole, and K.T. Konstantinidis, *The Microbial Genomes Atlas (MiGA) webserver:*

taxonomic and gene diversity analysis of Archaea and Bacteria at the whole genome level. Nucleic Acids Research, 2018. **46**(W1): p. W282-W288.

66. Yoon, S.-H., S.-M. Ha, S. Kwon, J. Lim, Y. Kim, H. Seo, and J. Chun, *Introducing EzBioCloud: a taxonomically united database of 16S rRNA gene sequences and whole-genome assemblies.* International journal of systematic and evolutionary microbiology, 2017. **67**(5): p. 1613-1617.
67. Allard, M.W., E. Strain, D. Melka, K. Bunning, S.M. Musser, E.W. Brown, and R. Timme, *Practical Value of Food Pathogen Traceability through Building a Whole-Genome Sequencing Network and Database.* Journal of clinical microbiology, 2016. **54**(8): p. 1975-1983.
68. Rosselló-Móra, R., M.E. Trujillo, and I.C. Sutcliffe, *Introducing a digital protologue: a timely move towards a database-driven systematics of archaea and bacteria.* Antonie van Leeuwenhoek, 2017. **110**(4): p. 455-456.
69. Stackebrandt, E. and D. Smith, *Expanding the 'Digital Protologue' Database (DPD) to 'Current Microbiology': An Offer to Scientists and Science.* Current Microbiology, 2017. **74**(9): p. 1003-1004.
70. Reimer, L.C., A. Vetcininova, J.S. Carbasse, C. Söhngen, D. Gleim, C. Ebeling, and J. Overmann, *BacDive in 2019: bacterial phenotypic data for High-throughput biodiversity analysis.* Nucleic Acids Research, 2018. **47**(D1): p. D631-D636.
71. Stackebrandt, E. and B.M. Goebel, *Taxonomic Note : A Place for DNA-DNA Reassociation and s rRNA Sequence Analysis in the Present Species Definition in Bacteriology.* 1994: p. 846-849.
72. Woese, C.R., O. Kandler, and M.L. Wheelis, *Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya.* Proceedings of the National Academy of Sciences of the United States of America, 1990. **87**(12): p. 4576-4579.
73. Yang, B., Y. Wang, and P.-Y. Qian, *Sensitivity and correlation of hypervariable regions in 16S rRNA genes in phylogenetic analysis.* BMC bioinformatics, 2016. **17**: p. 135-135.
74. Weisberg, A.J., H.A. Elmarakeby, L.S. Heath, and B.A. Vinatzer, *Similarity-Based Codes Sequentially Assigned to Ebolavirus Genomes Are Informative of Species Membership , Associated Outbreaks , and Transmission Chains.* 2015: p. 1-11.
75. Vinatzer, B.A., A.J. Weisberg, C.L. Monteil, H.A. Elmarakeby, S.K. Sheppard, and L.S. Heath, *A Proposal for a Genome Similarity-Based Taxonomy for Plant-Pathogenic Bacteria that Is Sufficiently Precise to Reflect Phylogeny, Host Range, and Outbreak Affiliation Applied to Pseudomonas syringae sensu lato as a Proof of Concept.* Phytopathology, 2016. **107**(1): p. 18-28.
76. Brown, C.T. and L. Irber, *sourmash: a library for MinHash scketching of DNA.* Journal of Open Source Software, 2016. **1**(5): p. 27.
77. Sokal, R.R. and F.J. Rohlf, *The Comparison of Dendrograms by Objective Methods.* Taxon, 1962. **11**(2): p. 33-40.
78. Galili, T., *dendextend: an R package for visualizing, adjusting and comparing trees of hierarchical clustering.* Bioinformatics (Oxford, England), 2015. **31**(22): p. 3718-3720.
79. Bird, B.H. and J.A.K. Mazet, *Detection of Emerging Zoonotic Pathogens: An Integrated One Health Approach.* Annual Review of Animal Biosciences, 2018. **6**(1): p. 121-139.
80. Huys, G., M. Vancanneyt, K. D'Haene, V. Vankerckhoven, H. Goossens, and J. Swings, *Accuracy of species identity of commercial bacterial cultures intended for probiotic or nutritional use.* Research in Microbiology, 2006. **157**(9): p. 803-810.

81. Velivelli, S.L.S., P. De Vos, P. Kromann, S. Declerck, and B.D. Prestwich, *Biological control agents: from field to market, problems, and challenges.* Trends in Biotechnology, 2014. **32**(10): p. 493-496.

82. Sutcliffe, I.C., M.E. Trujillo, and M. Goodfellow, *A call to arms for systematists: revitalising the purpose and practises underpinning the description of novel microbial taxa.* Antonie van Leeuwenhoek, 2012. **101**(1): p. 13-20.

83. Thompson, C.C., L. Chimetto, R.A. Edwards, J. Swings, E. Stackebrandt, and F.L. Thompson, *Microbial genomic taxonomy.* BMC Genomics, 2013. **14**(1): p. 913.

84. Pettengill, J.B., Y. Luo, S. Davis, Y. Chen, N. Gonzalez-Escalona, A. Ottesen, H. Rand, M.W. Allard, and E. Strain, *An evaluation of alternative methods for constructing phylogenies from whole genome sequence data: a case study with Salmonella.* PeerJ, 2014. **2**: p. e620.

85. Mellmann, A., D. Harmsen, C.A. Cummings, E.B. Zentz, S.R. Leopold, A. Rico, K. Prior, R. Szczepanowski, Y. Ji, W. Zhang, S.F. McLaughlin, J.K. Henkhaus, B. Leopold, M. Bielaszewska, R. Prager, P.M. Brzoska, R.L. Moore, S. Guenther, J.M. Rothberg, and H. Karch, *Prospective genomic characterization of the German enterohemorrhagic Escherichia coli O104:H4 outbreak by rapid next generation sequencing technology.* PloS one, 2011. **6**(7): p. e22751-e22751.

86. Tindall, B.J. and G.M. Garrity, *Proposals to clarify how type strains are deposited and made available to the scientific community for the purpose of systematic research.* International Journal of Systematic and Evolutionary Microbiology, 2008. **58**(8): p. 1987-1990.

87. Fournier, P.-E., D. Raoult, and M. Drancourt, *New Species Announcement: a new format to prompt the description of new human microbial species.* New microbes and new infections, 2016. **15**: p. 136-137.

88. Baltrus, D.A., H.C. McCann, and D.S. Guttman, *Evolution, genomics and epidemiology of Pseudomonas syringae.* Molecular Plant Pathology, 2017. **18**(1): p. 152-168.

89. Marakeby, H., E. Badr, H. Torkey, Y. Song, and S. Leman, *A System to Automatically Classify and Name Any Individual Genome-Sequenced Organism Independently of Current Biological Classification and Nomenclature.* 2014. **9**(2).

90. Khoshbakht, K. and K. Hammer, *How many plant species are cultivated?* Genetic Resources and Crop Evolution, 2008. **55**(7): p. 925-928.

91. Parte, A.C., *LPSN – List of Prokaryotic names with Standing in Nomenclature (bacterio.net), 20 years on.* International Journal of Systematic and Evolutionary Microbiology, 2018. **68**(6): p. 1825-1829.

92. Konstantinidis, K.T. and J.M. Tiedje, *Genomic insights that advance the species definition for prokaryotes.* Proceedings of the National Academy of Sciences of the United States of America, 2005. **102**(7): p. 2567.

93. Willems, A., M. Goor, S. Thielemans, M. Gillis, K. Kersters, and J. De Ley, *Transfer of several phytopathogenic Pseudomonas species to Acidovorax as Acidovorax avenae subsp. avenae subsp. nov., comb. nov., Acidovorax avenae subsp. citrulli, Acidovorax avenae subsp. cattleyae, and Acidovorax konjaci.* Int J Syst Bacteriol, 1992. **42**(1): p. 107-19.

94. Bull, B.A.V.a.C.T., *The Impact of Genomic Approaches on Our Understanding of Diversity and Taxonomy of Plant Pathogenic Bacteria*, in *Plant Pathogenic Bacteria: Genomics and Molecular Biology*, R.W. Jackson, Editor. 2009, Caister Academic Press. p. 37-61.

95. Bull, C.T., C. Manceau, J. Lydon, H. Kong, B.A. Vinatzer, and M. Fischer-Le Saux, *Pseudomonas cannabina pv. cannabina pv. nov., and Pseudomonas cannabina pv. alisalensis (Cintas Koike and Bull, 2000) comb. nov., are members of the emended*

species Pseudomonas cannabina (ex Šutič & Dowson 1959) Gardan, Shafik, Belouin, Brosch, Grimont & Grimont 1999. Systematic and Applied Microbiology, 2010. **33**(3): p. 105-115.

96. Alvarez, A.M., *Diversity and diagnosis of Ralstonia solanacearum*, in *Bacterial Wilt Disease and the Ralstonia solanacearum Species Complex*, P.P. Caitilyn Allen, and A. C. Hayward, Editor. 2005, American Phytopathological Society.

97. Prior, P. and M. Fegan. *RECENT DEVELOPMENTS IN THE PHYLOGENY AND CLASSIFICATION OF RALSTONIA SOLANACEARUM.* 2005. International Society for Horticultural Science (ISHS), Leuven, Belgium.

98. Berge, O., C.L. Monteil, C. Bartoli, C. Chandeysson, C. Guilbaud, D.C. Sands, and C.E. Morris, *A user's guide to a data base of the diversity of Pseudomonas syringae and its application to classifying strains in this phylogenetic complex.* PloS one, 2014. **9**(9): p. e105547-e105547.

99. Almeida, N.F., S. Yan, R. Cai, C.R. Clarke, C.E. Morris, N.W. Schaad, E.L. Schuenzel, G.H. Lacy, X. Sun, J.B. Jones, J.A. Castillo, C.T. Bull, S. Leman, D.S. Guttman, J.C. Setubal, and B.A. Vinatzer, *PAMDB, A Multilocus Sequence Typing and Analysis Database and Website for Plant-Associated Microbes.* Phytopathology, 2010. **100**(3): p. 208-215.

100. Paul, B., G. Dixit, T.S. Murali, and K. Satyamoorthy, *Genome-based taxonomic classification.* Genome, 2019. **62**(2): p. 45-52.

101. Weisberg, A.J., H.A. Elmarakeby, L.S. Heath, and B.A. Vinatzer, *Similarity-based codes sequentially assigned to ebolavirus genomes are informative of species membership, associated outbreaks, and transmission chains.* Open forum infectious diseases, 2015. **2**(1): p. ofv024-ofv024.

102. Purcell, A.H. and D.L. Hopkins, *FASTIDIOUS XYLEM-LIMITED BACTERIAL PLANT PATHOGENS.* Annual Review of Phytopathology, 1996. **34**(1): p. 131-151.

103. Simpson, A.J.G., F.C. Reinach, P. Arruda, F.A. Abreu, M. Acencio, R. Alvarenga, L.M.C. Alves, J.E. Araya, G.S. Baia, C.S. Baptista, M.H. Barros, E.D. Bonaccorsi, S. Bordin, J.M. Bové, M.R.S. Briones, M.R.P. Bueno, A.A. Camargo, L.E.A. Camargo, D.M. Carraro, H. Carrer, N.B. Colauto, C. Colombo, F.F. Costa, M.C.R. Costa, C.M. Costa-Neto, L.L. Coutinho, M. Cristofani, E. Dias-Neto, C. Docena, H. El-Dorry, A.P. Facincani, A.J.S. Ferreira, V.C.A. Ferreira, J.A. Ferro, J.S. Fraga, S.C. França, M.C. Franco, M. Frohme, L.R. Furlan, M. Garnier, G.H. Goldman, M.H.S. Goldman, S.L. Gomes, A. Gruber, P.L. Ho, J.D. Hoheisel, M.L. Junqueira, E.L. Kemper, J.P. Kitajima, J.E. Krieger, E.E. Kuramae, F. Laigret, M.R. Lambais, L.C.C. Leite, E.G.M. Lemos, M.V.F. Lemos, S.A. Lopes, C.R. Lopes, J.A. Machado, M.A. Machado, A.M.B.N. Madeira, H.M.F. Madeira, C.L. Marino, M.V. Marques, E.A.L. Martins, E.M.F. Martins, A.Y. Matsukuma, C.F.M. Menck, E.C. Miracca, C.Y. Miyaki, C.B. Monteiro-Vitorello, D.H. Moon, M.A. Nagai, A.L.T.O. Nascimento, L.E.S. Netto, A. Nhani, F.G. Nobrega, L.R. Nunes, M.A. Oliveira, M.C. de Oliveira, R.C. de Oliveira, D.A. Palmieri, A. Paris, B.R. Peixoto, G.A.G. Pereira, H.A. Pereira, J.B. Pesquero, R.B. Quaggio, P.G. Roberto, V. Rodrigues, A.J. de M. Rosa, V.E. de Rosa, R.G. de Sá, R.V. Santelli, H.E. Sawasaki, A.C.R. da Silva, A.M. da Silva, F.R. da Silva, W.A. Silva, J.F. da Silveira, M.L.Z. Silvestri, W.J. Siqueira, A.A. de Souza, A.P. de Souza, M.F. Terenzi, D. Truffi, S.M. Tsai, M.H. Tsuhako, H. Vallada, M.A. Van Sluys, S. Verjovski-Almeida, A.L. Vettore, M.A. Zago, M. Zatz, J. Meidanis and J.C. Setubal, *The genome sequence of the plant pathogen Xylella fastidiosa.* Nature, 2000. **406**(6792): p. 151-157.

104. Almeida, R.P.P., L. De La Fuente, R. Koebnik, J.R.S. Lopes, S. Parnell, and H. Scherm, *Addressing the New Global Threat of Xylella fastidiosa.* Phytopathology, 2019. **109**(2): p. 172-174.

105. Denancé, N., M. Briand, R. Gaborieau, S. Gaillard, and M.-A. Jacques, *Identification of genetic relationships and subspecies signatures in Xylella fastidiosa.* BMC genomics, 2019. **20**(1): p. 239-239.

106. Giampetruzzi, A., M. Saponari, R.P.P. Almeida, S. Essakhi, D. Boscia, G. Loconsole, and P. Saldarelli, *Complete Genome Sequence of the Olive-Infecting Strain Xylella fastidiosa subsp. pauca De Donno.* Genome announcements, 2017. **5**(27): p. e00569-17.

107. Marcelletti, S. and M. Scortichini, *Xylella fastidiosa CoDiRO strain associated with the olive quick decline syndrome in southern Italy belongs to a clonal complex of the subspecies pauca that evolved in Central America.* Microbiology, 2016. **162**(12): p. 2087-2098.

108. Jacques, M.-A., M. Arlat, A. Boulanger, T. Boureau, S. Carrère, S. Cesbron, N.W.G. Chen, S. Cociancich, A. Darrasse, N. Denancé, M. Fischer-Le Saux, L. Gagnevin, R. Koebnik, E. Lauber, L.D. Noël, I. Pieretti, P. Portier, O. Pruvost, A. Rieux, I. Robène, M. Royer, B. Szurek, V. Verdier, and C. Vernière, *Using Ecology, Physiology, and Genomics to Understand Host Specificity in Xanthomonas.* Annual Review of Phytopathology, 2016. **54**(1): p. 163-187.

109. Schwartz, A.R., N. Potnis, S. Timilsina, M. Wilson, J. Patané, J. Martins, G.V. Minsavage, D. Dahlbeck, A. Akhunova, N. Almeida, G.E. Vallad, J.D. Barak, F.F. White, S.A. Miller, D. Ritchie, E. Goss, R.S. Bart, J.C. Setubal, J.B. Jones, and B.J. Staskawicz, *Phylogenomics of Xanthomonas field strains infecting pepper and tomato reveals diversity in effector repertoires and identifies determinants of host specificity.* Frontiers in Microbiology, 2015. **6**(535).

110. Bull, C.T., C.R. Clarke, R. Cai, B.A. Vinatzer, T.M. Jardini, and S.T. Koike, *Multilocus Sequence Typing of Pseudomonas syringae Sensu Lato Confirms Previously Described Genomospecies and Permits Rapid Identification of P. syringae pv. coriandricola and P. syringae pv. apii Causing Bacterial Leaf Spot on Parsley.* Phytopathology, 2011. **101**(7): p. 847-858.

111. Gardan, L., H. Shafik, S. Belouin, R. Broch, F. Grimont, and P.A. Grimont, *DNA relatedness among the pathovars of Pseudomonas syringae and description of Pseudomonas tremae sp. nov. and Pseudomonas cannabina sp. nov. (ex Sutic and Dowson 1959).* Int J Syst Bacteriol, 1999. **49 Pt 2**: p. 469-78.

112. Gomila, M., A. Busquets, M. Mulet, E. García-Valdés, and J. Lalucat, *Clarification of Taxonomic Status within the Pseudomonas syringae Species Group Based on a Phylogenomic Analysis.* Frontiers in Microbiology, 2017. **8**: p. 2422.

113. Dye, D.W., J.F. Bradbury, M. Goto, A.C. Hayward, R.A. Lelliott, and M.N. Schroth, *International standards for naming pathovars of phytopathogenic bacteria and a list of pathovar names and pathotype strains.* Review of Plant Pathology, 1980. **59**(4): p. 153-168.

114. Cai, R., J. Lewis, S. Yan, H. Liu, C.R. Clarke, F. Campanile, N.F. Almeida, D.J. Studholme, M. Lindeberg, D. Schneider, M. Zaccardelli, J.C. Setubal, N.P. Morales-Lizcano, A. Bernal, G. Coaker, C. Baker, C.L. Bender, S. Leman, and B.A. Vinatzer, *The plant pathogen Pseudomonas syringae pv. tomato is genetically monomorphic and under strong selection to evade tomato immunity.* PLoS pathogens, 2011. **7**(8): p. e1002130-e1002130.

115. Monteil, C.L., K. Yahara, D.J. Studholme, L. Mageiros, G. Méric, B. Swingle, C.E. Morris, B.A. Vinatzer, and S.K. Sheppard, *Population-genomic insights into emergence, crop adaptation and dissemination of Pseudomonas syringae pathogens.* Microbial genomics, 2016. **2**(10): p. e000089-e000089.

116. Young, J.M. and D.C. Park, *Relationships of plant pathogenic enterobacteria based on partial atpD, carA, and recA as individual and concatenated nucleotide and peptide sequences.* Systematic and Applied Microbiology, 2007. **30**(5): p. 343-354.

117. Clarke, C.R., D.J. Studholme, B. Hayes, B. Runde, A. Weisberg, R. Cai, T. Wroblewski, M.-C. Daunay, E. Wicker, J.A. Castillo, and B.A. Vinatzer, *Genome-Enabled Phylogeographic Investigation of the Quarantine Pathogen Ralstonia solanacearum Race 3 Biovar 2 and Screening for Sources of Resistance Against Its Core Effectors.* Phytopathology, 2015. **105**(5): p. 597-607.

118. Program, F.S.A., *HHS and USDA Select Agents and Toxins*, C.f.D.C.a. Prevention, Editor. 2017.

119. Safni, I., I. Cleenwerck, P. De Vos, M. Fegan, L. Sly, and U. Kappler, *Polyphasic taxonomic revision of the Ralstonia solanacearum species complex: proposal to emend the descriptions of Ralstonia solanacearum and Ralstonia syzygii and reclassify current R. syzygii strains as Ralstonia syzygii subsp. syzygii subsp. nov., R. solanacearum phylotype IV strains as Ralstonia syzygii subsp. indonesiensis subsp. nov., banana blood disease bacterium strains as Ralstonia syzygii subsp. celebesensis subsp. nov. and R. solanacearum phylotype I and III strains as Ralstoniapseudosolanacearum sp. nov.* International Journal of Systematic and Evolutionary Microbiology, 2014. **64**(9): p. 3087-3103.

120. Coletta-Filho, H.D., M.L.P.N. Targon, M.A. Takita, J.D. De Negri, J. Pompeu, M.A. Machado, A.M. do Amaral, and G.W. Muller, *First Report of the Causal Agent of Huanglongbing ("Candidatus Liberibacter asiaticus") in Brazil.* Plant Disease, 2004. **88**(12): p. 1382-1382.

121. Hallbert, S.E. *The discovery of huanglongbing in Florida*. in *2nd International Citrus Canker and Huanglongbing Research Workshop*. 2005. Orlando, FL, USA: Florida Citrus Mutual.

122. JAGOUEIX, S., J.-M. BOVE, and M. GARNIER, *The Phloem-Limited Bacterium of Greening Disease of Citrus Is a Member of the α Subdivision of the Proteobacteria.* International Journal of Systematic and Evolutionary Microbiology, 1994. **44**(3): p. 379-386.

123. Garnier, M., S. Jagoueix-Eveillard, P.R. Cronje, H.F. Le Roux, and J.M. Bové, *Genomic characterization of a liberibacter present in an ornamental rutaceous tree, Calodendrum capense, in the Western Cape Province of South Africa. Proposal of &apos;Candidatus Liberibacter africanus subsp. capensis&apos.* International Journal of Systematic and Evolutionary Microbiology, 2000. **50**(6): p. 2119-2125.

124. Texeira, D.C., J. Ayres, E.W. Kitajima, L. Danet, S. Jagoueix-Eveillard, C. Saillard, and J.M. Bové, *First Report of a Huanglongbing-Like Disease of Citrus in Sao Paulo State, Brazil and Association of a New Liberibacter Species, "Candidatus Liberibacter americanus", with the Disease.* Plant Disease, 2005. **89**(1): p. 107-107.

125. Raddadi, N., E. Gonella, C. Camerota, A. Pizzinat, R. Tedeschi, E. Crotti, M. Mandrioli, P. Attilio Bianco, D. Daffonchio, and A. Alma, *'Candidatus Liberibacter europaeus' sp. nov. that is associated with and transmitted by the psyllid Cacopsylla pyri apparently behaves as an endophyte rather than a pathogen.* Environmental Microbiology, 2011. **13**(2): p. 414-426.

126. Liefting, L.W., P.W. Sutherland, L.I. Ward, K.L. Paice, B.S. Weir, and G.R.G. Clover, *A New 'Candidatus Liberibacter' Species Associated with Diseases of Solanaceous Crops.* Plant Disease, 2009. **93**(3): p. 208-214.

127. Fagen, J.R., M.T. Leonard, J.F. Coyle, C.M. McCullough, A.G. Davis-Richardson, M.J. Davis, and E.W. Triplett, *Liberibactercrescens gen. nov., sp. nov., the first cultured*

member of the genus Liberibacter. International Journal of Systematic and Evolutionary Microbiology, 2014. **64**(7): p. 2461-2466.

128. Bar-On, Y.M., R. Phillips, and R. Milo, *The biomass distribution on Earth.* Proceedings of the National Academy of Sciences, 2018. **115**(25): p. 6506.

129. Vartoukian, S.R., R.M. Palmer, and W.G. Wade, *Strategies for culture of 'unculturable' bacteria.* FEMS Microbiology Letters, 2010. **309**(1): p. 1-7.

Appendix

Supplementary Table 1

Data set	Genus	Species	Strain	LIN
A	Pseudomonas	aeruginosa	PACS2	0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0
	Pseudomonas	aeruginosa	N002	0,0,0,0,0,0,0,0,0,0,0,1,1,0,0,0,0,0,0,0
	Pseudomonas	aeruginosa	C40	0,0,0,0,0,0,0,0,0,0,0,1,10,0,0,0,0,0,0,0
	Pseudomonas	aeruginosa	C23	0,0,0,0,0,0,0,0,0,0,0,1,11,0,0,0,0,0,0,0
	Pseudomonas	aeruginosa	C20	0,0,0,0,0,0,0,0,0,0,0,1,11,0,0,0,0,0,0,1
	Pseudomonas	aeruginosa	M8A.4	0,0,0,0,0,0,0,0,0,0,0,1,12,0,0,0,0,0,0,0
	Pseudomonas	aeruginosa	M8A.1	0,0,0,0,0,0,0,0,0,0,0,1,13,0,0,0,0,0,0,0
	Pseudomonas	aeruginosa	MRW44.1	0,0,0,0,0,0,0,0,0,0,0,1,2,0,0,0,0,0,0,0
	Pseudomonas	aeruginosa	DQ8	0,0,0,0,0,0,0,0,0,0,0,1,2,1,0,0,0,0,0,0
	Pseudomonas	aeruginosa	PA45	0,0,0,0,0,0,0,0,0,0,0,1,3,0,0,0,0,0,0,0
	Pseudomonas	aeruginosa	str. C 1334	0,0,0,0,0,0,0,0,0,0,0,1,4,0,0,0,0,0,0,0
	Pseudomonas	aeruginosa	str. PA 62	0,0,0,0,0,0,0,0,0,0,0,1,5,0,0,0,0,0,0,0
	Pseudomonas	aeruginosa	str. C 1426	0,0,0,0,0,0,0,0,0,0,0,1,6,0,0,0,0,0,0,0
	Pseudomonas	aeruginosa	PAK	0,0,0,0,0,0,0,0,0,0,0,1,7,0,0,0,0,0,0,0
	Pseudomonas	aeruginosa	MSH-10	0,0,0,0,0,0,0,0,0,0,0,1,8,0,0,0,0,0,0,0
	Pseudomonas	aeruginosa	C48	0,0,0,0,0,0,0,0,0,0,0,1,9,0,0,0,0,0,0,0
	Pseudomonas	aeruginosa	M8A.3	0,0,0,0,0,0,0,0,0,0,0,1,9,1,0,0,0,0,0,0
	Pseudomonas	aeruginosa	M8A.2	0,0,0,0,0,0,0,0,0,0,0,1,9,1,0,0,0,0,1,0
	Pseudomonas	aeruginosa	138244	0,0,0,0,0,0,0,0,0,0,0,2,0,0,0,0,0,0,0,0
	Pseudomonas	aeruginosa	9BR	0,0,0,0,0,0,0,0,0,0,0,3,0,0,0,0,0,0,0,0
	Pseudomonas	aeruginosa	19BR	0,0,0,0,0,0,0,0,0,0,0,3,0,0,0,0,0,1,0,0
	Pseudomonas	aeruginosa	213BR	0,0,0,0,0,0,0,0,0,0,0,3,0,0,0,0,1,0,0,0
	Pseudomonas	aeruginosa	PGPR2	0,0,0,0,0,0,0,0,0,0,0,3,1,0,0,0,0,0,0,0
	Pseudomonas	aeruginosa	CF614	0,0,0,0,0,0,0,0,0,0,0,3,1,1,0,0,0,0,0,0
	Pseudomonas	aeruginosa	LCT-PA102	0,0,0,0,0,0,0,0,0,0,0,4,0,0,0,0,0,0,0,0
	Pseudomonas	aeruginosa	LCT-PA220	0,0,0,0,0,0,0,0,0,0,0,4,0,0,0,0,0,0,1,0
	Pseudomonas	aeruginosa	LCT-PA41	0,0,0,0,0,0,0,0,0,0,0,4,0,0,0,0,0,0,2,0
	Pseudomonas	aeruginosa	AH16	0,0,0,0,0,0,0,0,0,0,0,4,0,0,0,0,0,1,0,0
	Pseudomonas	aeruginosa	CF77	0,0,0,0,0,0,0,0,0,0,0,4,0,1,0,0,0,0,0,0

	Pseudomonas	aeruginosa	str. C 763	0,0,0,0,0,0,0,0,0,0,0,4,1,0,0,0,0,0,0,0
	Pseudomonas	aeruginosa	C52	0,0,0,0,0,0,0,0,0,0,0,4,1,0,1,0,0,0,0,0
	Pseudomonas	aeruginosa	C51	0,0,0,0,0,0,0,0,0,0,0,4,1,0,2,0,0,0,0,0
	Pseudomonas	aeruginosa	S35004	0,0,0,0,0,0,0,0,0,0,0,4,1,0,4,0,0,0,0,0
	Pseudomonas	aeruginosa	C41	0,0,0,0,0,0,0,0,0,0,0,4,2,0,0,0,0,0,0,0
	Pseudomonas	aeruginosa	VRFPA02	0,0,0,0,0,0,0,0,0,0,0,5,0,0,0,0,0,0,0,0
	Pseudomonas	aeruginosa	M9A.1	0,0,0,0,0,0,0,0,0,0,0,6,0,0,0,0,0,0,0,0
	Pseudomonas	aeruginosa	2192	0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0
	Pseudomonas	aeruginosa	WC55	0,0,0,0,0,0,0,0,0,0,1,0,1,0,0,0,0,0,0,0
	Pseudomonas	aeruginosa	PAb1	0,0,0,0,0,0,0,0,0,0,2,0,0,0,0,0,0,0,0,0
	Pseudomonas	aeruginosa	HB15	0,0,0,0,0,0,0,0,0,0,2,0,0,0,1,0,0,0,0,0
	Pseudomonas	aeruginosa	PA14	0,0,0,0,0,0,0,0,0,0,2,0,1,0,0,0,0,0,0,0
	Pseudomonas	aeruginosa	X13273	0,0,0,0,0,0,0,0,0,0,2,0,3,1,0,0,0,0,0,0
	Pseudomonas	aeruginosa	39016	0,0,0,0,0,0,0,0,0,0,2,1,0,0,0,0,0,0,0,0
	Pseudomonas	aeruginosa	str. E 2	0,0,0,0,0,0,0,0,0,0,2,1,0,0,1,0,0,0,0,0
	Pseudomonas	aeruginosa	U2504	0,0,0,0,0,0,0,0,0,0,2,1,0,4,0,0,0,0,0,0
	Pseudomonas	aeruginosa	PABL056	0,0,0,0,0,0,0,0,0,0,2,2,0,0,0,0,0,0,0,0
	Pseudomonas	aeruginosa	VRFPA04	0,0,0,0,0,0,0,0,0,0,2,3,0,0,0,0,0,0,0,0
	Pseudomonas	aeruginosa	str. Stone 130	0,0,0,0,0,0,0,0,0,0,3,0,0,0,0,0,0,0,0,0
	Pseudomonas	aeruginosa	VRFPA03	0,0,0,0,0,0,0,0,0,0,4,0,0,0,0,0,0,0,0,0
	Pseudomonas	alcaligenes	OT 69	0,0,0,2,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
	Pseudomonas	alcaligenes	MRY13-0052	0,0,0,2,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0
	Pseudomonas	resinovorans	DSM 21078	0,0,0,3,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
	Pseudomonas	resinovorans	NBRC 106553 DNA	0,0,0,3,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
	Pseudomonas	pseudoalcaligenes	KF707	0,0,0,3,2,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
	Pseudomonas	nitroreducens	HBP1	0,0,0,4,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
	Pseudomonas	sp.	TX1	0,0,0,4,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0
	Pseudomonas	denitrificans	ATCC 13867,	0,0,0,4,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0

	Pseudomonas	thermotolerans	J53	0,0,0,5,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
	Pseudomonas	alcaligenes	NBRC 14159	0,0,0,6,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
	Pseudomonas	syringae	str. NCPPB 3681	0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
	Pseudomonas	syringae	str. 2250	0,0,1,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0
	Pseudomonas	savastanoi	NCPPB 3335	0,0,1,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0
	Pseudomonas	syringae	str. M301315	0,0,1,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0
	Pseudomonas	syringae	str. 6605	0,0,1,0,0,0,0,1,0,1,0,0,0,0,0,0,0,0,0
	Pseudomonas	syringae	1448A	0,0,1,0,0,0,0,1,1,0,0,0,0,0,0,0,0,0,0
	Pseudomonas	syringae	str. B076	0,0,1,0,0,0,0,1,1,0,0,1,0,0,0,0,0,0,0
	Pseudomonas	syringae	str. race 4	0,0,1,0,0,0,0,1,1,0,0,1,0,1,0,0,0,0,0
	Pseudomonas	syringae	K40	0,0,1,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0
	Pseudomonas	syringae	NCPPB 1108	0,0,1,0,1,0,0,0,0,0,0,0,0,1,0,0,0,0,0
	Pseudomonas	syringae	T1	0,0,1,0,1,0,0,0,0,0,0,0,0,2,0,0,0,0,0
	Pseudomonas	syringae	str. M302278	0,0,1,0,1,0,0,0,0,1,0,0,0,0,0,0,0,0,0
	Pseudomonas	syringae	str. DC3000	0,0,1,0,1,0,0,0,0,1,1,0,0,0,0,0,0,0,0
	Pseudomonas	syringae	CC1630	0,0,1,0,1,0,0,0,0,2,0,0,0,0,0,0,0,0,0
	Pseudomonas	syringae	str. M302280	0,0,1,0,1,0,1,0,0,0,0,0,0,0,0,0,0,0,0
	Pseudomonas	syringae	str. M302091	0,0,1,0,1,0,1,0,1,0,0,0,0,0,0,0,0,0,0
	Pseudomonas	syringae	str. NCPPB 3871	0,0,1,0,1,0,1,0,1,0,0,0,0,0,0,1,0,0,0,0
	Pseudomonas	syringae	str. NCPPB 3739	0,0,1,0,1,0,1,0,1,0,0,0,0,0,0,1,0,1,0,0
	Pseudomonas	syringae	KW41	0,0,1,0,1,0,1,0,1,0,0,0,0,0,0,1,0,2,0,0
	Pseudomonas	syringae	ICMP 9853	0,0,1,0,1,0,1,0,1,0,0,0,0,0,0,1,1,0,0,0
	Pseudomonas	syringae	PA459	0,0,1,0,1,0,1,0,1,0,0,0,0,0,1,0,0,0,0,0
	Pseudomonas	syringae	ICMP 19103	0,0,1,0,1,0,1,0,1,0,0,0,0,0,2,0,0,0,0,0
	Pseudomonas	syringae	ICMP 19068	0,0,1,0,1,0,1,0,1,0,0,0,0,0,3,0,0,0,0,0
	Pseudomonas	syringae	ICMP 19104	0,0,1,0,1,0,1,0,1,0,0,0,0,0,4,0,0,0,0,0
	Pseudomonas	syringae	ICMP 19102	0,0,1,0,1,0,1,0,1,0,0,0,0,0,5,0,0,0,0,0
	Pseudomonas	syringae	CFBP 7286	0,0,1,0,1,0,1,0,1,0,0,0,1,0,0,0,0,0,0,0

	Pseudomonas	syringae	ICMP 18744	0,0,1,0,1,0,1,0,1,0,0,0,1,0,0,0,1,0,0,0
	Pseudomonas	syringae	ICMP 19455	0,0,1,0,1,0,1,0,1,0,0,0,1,0,0,0,2,0,0,0
	Pseudomonas	syringae	ICMP 19439	0,0,1,0,1,0,1,0,1,0,0,0,1,0,0,0,2,0,1,0
	Pseudomonas	syringae	CH2010-6	0,0,1,0,1,0,1,0,1,0,0,0,1,0,1,0,0,0,0,0
	Pseudomonas	syringae	ICMP 19097	0,0,1,0,1,0,1,0,1,0,0,0,1,0,1,0,1,0,0,0
	Pseudomonas	syringae	ICMP 18800	0,0,1,0,1,0,1,0,1,0,0,0,1,0,1,0,2,0,0,0
	Pseudomonas	syringae	TP6-1	0,0,1,0,1,0,1,0,1,0,0,0,1,0,1,0,2,0,0,1
	Pseudomonas	syringae	TP1	0,0,1,0,1,0,1,0,1,0,0,0,1,0,1,0,2,0,1,0
	Pseudomonas	syringae	ICMP 18801	0,0,1,0,1,0,1,0,1,0,0,0,1,0,2,0,0,0,0,0
	Pseudomonas	syringae	str. Shaanxi_M228	0,0,1,0,1,0,1,0,1,0,0,0,1,1,0,0,0,0,0,0
	Pseudomonas	syringae	ICMP 19073	0,0,1,0,1,0,1,0,1,0,0,0,2,0,0,0,0,0,0,0
	Pseudomonas	syringae	ICMP 19071	0,0,1,0,1,0,1,0,1,0,0,0,2,1,0,0,0,0,0,0
	Pseudomonas	syringae	ICMP 19072	0,0,1,0,1,0,1,0,1,0,0,0,2,1,1,0,0,0,0,0
	Pseudomonas	syringae	ICMP 3923	0,0,1,0,1,0,1,0,1,0,1,0,0,0,0,0,0,0,0,0
	Pseudomonas	syringae	NCPPB 2598	0,0,1,0,1,0,1,0,1,0,1,0,0,0,0,1,0,0,0,0
	Pseudomonas	syringae	ICMP 19100	0,0,1,0,1,0,1,0,2,0,0,0,0,0,0,0,0,0,0,0
	Pseudomonas	syringae	ICMP 19099	0,0,1,0,1,0,1,0,2,0,0,0,0,0,1,0,0,0,0,0
	Pseudomonas	syringae	ICMP 19098	0,0,1,0,1,0,1,0,2,0,0,0,0,0,2,0,0,0,0,0
	Pseudomonas	syringae	ICMP 18883	0,0,1,0,1,0,1,0,2,0,0,0,0,0,3,0,0,0,0,0
	Pseudomonas	syringae	ICMP 19095	0,0,1,0,1,0,1,0,2,0,0,0,0,0,4,0,0,0,0,0
	Pseudomonas	syringae	ICMP 19094	0,0,1,0,1,0,1,0,2,0,0,0,0,0,5,0,0,0,0,0
	Pseudomonas	syringae	ICMP 18804	0,0,1,0,1,0,1,0,2,0,0,0,0,0,6,0,0,0,0,0
	Pseudomonas	syringae	ICMP 18806	0,0,1,0,1,0,1,0,2,0,0,0,0,0,6,0,0,0,1,0
	Pseudomonas	syringae	ICMP 18807	0,0,1,0,1,0,1,0,2,0,1,0,0,0,0,0,0,0,0,0
	Pseudomonas	avellanae	BPIC 631	0,0,1,0,1,0,1,0,3,0,0,0,0,0,0,0,0,0,0,0
	Pseudomonas	avellanae	CRAFRUec1	0,0,1,0,1,0,1,0,3,0,0,0,0,0,1,0,0,0,0,0
	Pseudomonas	syringae	USA007	0,0,1,0,1,0,1,1,0,0,0,0,0,0,0,0,0,0,0,0
	Pseudomonas	syringae	CC1416	0,0,1,0,1,0,1,1,0,1,0,0,0,0,0,0,0,0,0,0

	Pseudomonas	syringae	CC1544	0,0,1,0,1,0,1,1,1,0,0,0,0,0,0,0,0,0,0,0
	Pseudomonas	syringae	CC1559	0,0,1,0,1,0,1,2,0,0,0,0,0,0,0,0,0,0,0,0
	Pseudomonas	syringae	str. ES4326	0,0,1,0,2,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
	Pseudomonas	syringae	str. ISPaVe013	0,0,1,0,3,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
	Pseudomonas	syringae	str. ISPaVe037	0,0,1,0,3,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0
	Pseudomonas	syringae	BRIP34876	0,0,1,0,3,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0
	Pseudomonas	syringae	BRIP34881	0,0,1,0,3,0,1,0,0,0,0,0,0,0,0,0,0,1,0,0
	Pseudomonas	syringae	B64	0,0,1,0,3,0,1,0,0,0,1,0,0,0,0,0,0,0,0,0
	Pseudomonas	syringae	str. DSM 50255	0,0,1,0,3,0,1,0,0,0,1,0,0,0,1,0,0,0,0,0
	Pseudomonas	syringae	SM	0,0,1,0,3,0,1,0,0,1,0,0,0,0,0,0,0,0,0,0
	Pseudomonas	syringae	CC440	0,0,1,0,3,0,1,0,0,2,0,0,0,0,0,0,0,0,0,0
	Pseudomonas	syringae	CC1543	0,0,1,0,3,0,1,0,0,3,0,0,0,0,0,0,0,0,0,0
	Pseudomonas	syringae	CC1458	0,0,1,0,3,0,1,0,0,4,0,0,0,0,0,0,0,0,0,0
	Pseudomonas	syringae	DSM 10604	0,0,1,0,3,0,1,0,0,5,0,0,0,0,0,0,0,0,0,0
	Pseudomonas	syringae	str. PP1	0,0,1,0,3,0,1,0,1,0,0,0,0,0,0,0,0,0,0,0
	Pseudomonas	syringae	BRIP39023	0,0,1,0,3,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0
	Pseudomonas	syringae	1212	0,0,1,0,3,2,0,0,0,0,0,0,0,0,0,0,0,0,0,0
	Pseudomonas	syringae	B728a	0,0,1,0,3,2,0,0,0,0,1,0,0,0,0,0,0,0,0,0
	Pseudomonas	syringae	USA011	0,0,1,0,3,2,0,0,0,1,0,0,0,0,0,0,0,0,0,0
	Pseudomonas	syringae	UB303	0,0,1,0,3,2,0,0,0,2,0,0,0,0,0,0,0,0,0,0
	Pseudomonas	syringae	str. B301D-R	0,0,1,0,3,2,0,0,0,2,1,0,0,0,0,0,0,0,0,0
	Pseudomonas	syringae	642	0,0,1,0,3,3,0,0,0,0,0,0,0,0,0,0,0,0,0,0
	Pseudomonas	syringae	CC1629	0,0,1,0,4,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
	Pseudomonas	syringae	CC1513	0,0,1,0,4,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0
	Pseudomonas	syringae	CC1583	0,0,1,0,5,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
	Pseudomonas	syringae	CC1466	0,0,1,0,5,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0
	Pseudomonas	syringae	CC1557	0,0,1,0,5,1,0,0,0,0,1,0,0,0,0,0,0,0,0,0
	Pseudomonas	viridiflava	TA043	0,0,1,2,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0

	Pseudomonas	viridiflava	UASWS0038	0,0,1,2,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0
	Pseudomonas	viridiflava	CC1582	0,0,1,2,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0
	Pseudomonas	syringae	CC1524	0,0,1,2,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
	Pseudomonas	syringae	CC1417	0,0,1,2,1,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0
	Pseudomonas	sp.	M47T1	0,0,11,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
	Pseudomonas	fluorescens	HK44	0,0,2,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
	Pseudomonas	fuscovaginae	CB98818	0,0,2,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
	Pseudomonas	fuscovaginae	ICMP 5940	0,0,2,1,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0
	Pseudomonas	fuscovaginae	DAR 77795	0,0,2,1,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0
	Pseudomonas	fuscovaginae	DAR 77800	0,0,2,1,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0
	Pseudomonas	gingeri	NCPPB 3146	0,0,2,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
	Pseudomonas	agarici	NCPPB 2289	0,0,2,1,2,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
	Pseudomonas	fuscovaginae	SE-1	0,0,2,1,3,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
	Pseudomonas	fluorescens	R124	0,0,2,2,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
	Pseudomonas	fluorescens	Pf0-1	0,0,2,2,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
	Pseudomonas	sp.	GM25	0,0,2,2,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0
	Pseudomonas	sp.	GM24	0,0,2,2,3,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0
	Pseudomonas	sp.	GM16	0,0,2,2,3,1,0,0,0,0,0,0,0,0,0,1,0,0,0,0
	Pseudomonas	sp.	GM80	0,0,2,2,4,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
	Pseudomonas	fluorescens	Pf29Arp	0,0,2,3,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
	Pseudomonas	brassicacearum	51MFCVI2.1	0,0,2,3,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0
	Pseudomonas	brassicacearum	NFM421	0,0,2,3,1,1,0,0,0,0,0,0,2,0,0,0,0,0,0,0
	Pseudomonas	brassicacearum	strain DF41,	0,0,2,3,2,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
	Pseudomonas	fluorescens	LMG 5329	0,0,2,4,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0
	Pseudomonas	fluorescens	EGD-AQ6	0,0,2,4,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0
	Pseudomonas	sp.	CBZ-4	0,0,2,4,12,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
	Pseudomonas	veronii	1YdBTEX2	0,0,2,4,13,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
	Pseudomonas	synxantha	BG33R	0,0,2,4,14,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0

	Pseudomonas	fluorescens	FH5	0,0,2,4,2,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
	Pseudomonas	poae	RE*1-1-14,	0,0,2,4,6,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
	Pseudomonas	sp.	Ag1	0,0,2,4,7,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0
	Pseudomonas	sp.	PAMC 26793	0,0,2,4,7,2,0,0,0,0,0,0,0,0,0,0,0,0,0,0
	Pseudomonas	extremaustralis	14-3 substr. 14-3b strain 14-3	0,0,2,4,8,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
	Pseudomonas	tolaasii	PMS117	0,0,2,4,9,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
	Pseudomonas	tolaasii	6264	0,0,2,4,9,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0
	Pseudomonas	sp.	GM78	0,0,2,5,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0
	Pseudomonas	mandelii	36MFCvi1.1	0,0,2,5,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
	Pseudomonas	sp.	GM102	0,0,2,5,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0
	Pseudomonas	sp.	GM50	0,0,2,5,1,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0
	Pseudomonas	sp.	GM79	0,0,2,5,1,2,0,0,0,0,0,0,0,0,0,0,0,0,0,0
	Pseudomonas	sp.	GM18	0,0,2,5,1,3,0,0,0,0,0,0,0,0,0,0,0,0,0,0
	Pseudomonas	sp.	GM33	0,0,2,5,3,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0
	Pseudomonas	sp.	GM74	0,0,2,5,3,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0
	Pseudomonas	sp.	GM55	0,0,2,5,3,2,0,0,0,0,0,0,0,0,0,0,0,0,0,0
	Pseudomonas	sp.	GM49	0,0,2,5,3,3,0,0,0,0,0,0,0,0,0,0,0,0,0,0
	Pseudomonas	sp.	GM48	0,0,2,5,3,4,0,0,0,0,0,0,0,0,0,0,0,0,0,0
	Pseudomonas	sp.	GM67	0,0,2,5,4,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
	Pseudomonas	sp.	GM60	0,0,2,5,4,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0
	Pseudomonas	sp.	GM21	0,0,2,5,5,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
	Pseudomonas	chlororaphis	HT66	0,0,2,6,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
	Pseudomonas	chlororaphis	YL-1	0,0,2,6,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0
	Pseudomonas	chlororaphis	GP72	0,0,2,6,0,0,1,0,0,0,1,0,0,0,0,0,0,0,0,0
	Pseudomonas	chlororaphis	O6	0,0,2,6,0,0,1,0,0,0,2,0,0,0,0,0,0,0,0,0
	Pseudomonas	chlororaphis	PB-St2	0,0,2,6,0,0,1,1,0,0,0,0,0,0,0,0,0,0,0,0
	Pseudomonas	sp.	GM17	0,0,2,6,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0
	Pseudomonas	chlororaphis	30-84	0,0,2,6,0,1,0,1,0,0,0,0,0,0,0,0,0,0,0,0

	Pseudomonas	protegens	Pf-5	0,0,2,8,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
	Pseudomonas	protegens	CHA0,	0,0,2,8,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0
	Pseudomonas	putida	LS46	0,0,3,0,0,0,0,0,0,1,0,1,0,0,0,0,0,0,0,0
	Pseudomonas	putida	F1	0,0,3,0,0,0,0,0,0,2,0,1,0,0,0,0,0,0,0,0
	Pseudomonas	putida	BIRD-1	0,0,3,0,0,0,0,1,1,0,0,0,1,0,0,0,0,0,0,0
	Pseudomonas	putida	B001	0,0,3,0,0,2,0,0,0,0,0,0,0,2,0,0,0,0,0,0
	Pseudomonas	monteilii	QM	0,0,3,0,0,2,0,0,0,1,0,0,0,0,0,0,0,0,0,0
	Pseudomonas	putida	GB-1	0,0,3,0,0,3,0,0,0,0,0,0,0,0,0,0,0,0,0,0
	Pseudomonas	putida	H8234	0,0,3,0,0,4,0,0,0,0,0,0,0,0,0,0,0,0,0,0
	Pseudomonas	putida	NBRC 14164	0,0,3,0,0,5,0,0,0,0,0,0,0,0,0,0,0,0,0,0
	Pseudomonas	plecoglossicida	NB2011	0,0,3,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
	Pseudomonas	sp.	GM84	0,0,3,0,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0
	Pseudomonas	parafulva	DSM 17004	0,0,3,0,2,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
	Pseudomonas	taiwanensis	DSM 21245	0,0,3,0,3,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
	Pseudomonas	taiwanensis	SJ9	0,0,3,0,4,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
	Pseudomonas	monteilii	SB3078	0,0,3,0,4,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0
	Pseudomonas	monteilii	SB3101	0,0,3,0,4,0,0,0,0,0,1,0,0,0,0,0,0,0,0,1
	Pseudomonas	putida	S16	0,0,3,0,4,0,0,0,0,0,2,0,0,0,0,0,0,0,0,0
	Pseudomonas	putida	HB3267	0,0,3,0,4,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0
	Pseudomonas	mosselii	DSM 17497	0,0,3,0,7,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
	Pseudomonas	entomophila	L48	0,0,3,0,8,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
	Pseudomonas	putida	W619	0,0,3,0,9,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
	Pseudomonas	cremoricolorata	DSM 17059	0,0,3,3,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
	Pseudomonas	vranovensis	DSM 16006	0,0,3,4,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
	Pseudomonas	fulva	NBRC 16637 = DSM 17717	0,0,3,5,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0
	Pseudomonas	stutzeri	DSM 4166	0,0,5,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0
	Pseudomonas	stutzeri	XLDN-R	0,0,5,0,0,0,0,0,4,0,0,0,0,0,0,0,0,0,0,0
	Pseudomonas	chloritidismutans	AW-1	0,0,5,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0

	Pseudomonas	stutzeri	CCUG 29243	0,0,5,0,1,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0
	Pseudomonas	stutzeri	NF13	0,0,5,0,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0
	Pseudomonas	stutzeri	RCH2	0,0,5,0,3,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
	Pseudomonas	stutzeri	ATCC 14405 = CCUG 16156	0,0,5,0,4,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
	Pseudomonas	stutzeri	DSM 10701	0,0,5,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
	Pseudomonas	stutzeri	TS44	0,0,5,2,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
	Pseudomonas	sp.	Chol1	0,0,5,2,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0
	Pseudomonas	psychrophila	HA-4	0,0,6,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
	Pseudomonas	mendocina	EGD-AQ5	0,0,7,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
	Pseudomonas	alcaliphila	34 IA	0,0,7,0,1,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0
	Pseudomonas	mendocina	ymp	0,0,7,0,2,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
	Pseudomonas	mendocina	DLHK	0,0,7,0,2,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0
	Pseudomonas	mendocina	NK-01	0,0,7,0,3,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
	Pseudomonas	fulva	12-X	0,0,7,1,0,0,0,2,0,0,0,0,0,0,0,0,0,0,0,0
	Pseudomonas	azotifigens	DSM 17556	0,0,8,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
	Pseudomonas	oleovorans	MOIL14HWK12	0,0,9,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1
	Pseudomonas	psychrotolerans	L19	0,0,9,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0
	Pseudomonas	sp.	313	0,0,9,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0
	Pseudomonas	geniculata	N1	0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
	Pseudomonas	luteola	XLDN4-9	0,2,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
	Pseudomonas	pelagia	CL-AP6	0,3,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
	Pseudomonas	caeni	DSM 24390	0,4,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
B	Pseudomonas	aeruginosa	str. C3719	

Supplementary Table 2

Genus	Species	Subspecies	Strain	Note
Xylella	taiwanensis		PLS229	
Xylella	taiwanensis		PLS235	
Xylella	fastidiosa	ssp. pauca	CFBP8072	
Xylella	fastidiosa	ssp. pauca	OLS0479	

Xylella	fastidiosa	ssp. pauca	OLS0478	
Xylella	fastidiosa	ssp. pauca	COF0407	
Xylella	fastidiosa	ssp. pauca	CoDiRO	
Xylella	fastidiosa	ssp. pauca	De-Donno	Isolated in Italy
Xylella	fastidiosa	ssp. pauca	Salento-1	Isolated in Italy
Xylella	fastidiosa	ssp. pauca	Salento-2	Isolated in Italy
Xylella	fastidiosa	ssp. pauca	Hib4	
Xylella	fastidiosa	ssp. pauca	COF0324	
Xylella	fastidiosa	ssp. pauca	6c	
Xylella	fastidiosa	ssp. pauca	Pr8x	
Xylella	fastidiosa	ssp. pauca	11399	
Xylella	fastidiosa	ssp. pauca	J1a12	
Xylella	fastidiosa	ssp. pauca	CVC0251	
Xylella	fastidiosa	ssp. pauca	CVC0256	
Xylella	fastidiosa	ssp. pauca	32	
Xylella	fastidiosa	ssp. pauca	3124	
Xylella	fastidiosa	ssp. pauca	Fb7	
Xylella	fastidiosa	ssp. pauca	U24D	
Xylella	fastidiosa	ssp. pauca	9a5c	
Xylella	fastidiosa	ssp. multiplex	BB01	
Xylella	fastidiosa	ssp. multiplex	CFBP8078	
Xylella	fastidiosa	ssp. multiplex	ATCC-35871	
Xylella	fastidiosa	ssp. multiplex	sycamore-Sy-VA	
Xylella	fastidiosa	ssp. multiplex	CFBP8416	
Xylella	fastidiosa	ssp. multiplex	Dixon	
Xylella	fastidiosa	ssp. multiplex	CFBP8417	
Xylella	fastidiosa	ssp. multiplex	CFBP8418	
Xylella	fastidiosa	ssp. multiplex	Griffin-1	
Xylella	fastidiosa	ssp. multiplex	M12	
Xylella	fastidiosa	ssp. multiplex	IVIA5901	
Xylella	fastidiosa	ssp. multiplex	ESVL	
Xylella	fastidiosa	ssp. sandyi	Ann-1	
Xylella	fastidiosa	ssp. sandyi-like	CFBP8356	
Xylella	fastidiosa	ssp. sandyi-like	CO33	
Xylella	fastidiosa	ssp. morus	Mul-MD	
Xylella	fastidiosa	ssp. morus	MUL0034	
Xylella	fastidiosa	ssp. fastidosa	CFBP8073	
Xylella	fastidiosa	ssp. fastidosa	EB92.1	
Xylella	fastidiosa	ssp. fastidosa	ATCC-35879	
Xylella	fastidiosa	ssp. fastidosa	DSM-10026	
Xylella	fastidiosa	ssp. fastidosa	CFBP8082	
Xylella	fastidiosa	ssp. fastidosa	CFBP7969	

Xylella	fastidiosa	ssp. fastidosa	CFBP7970	
Xylella	fastidiosa	ssp. fastidosa	GB514	
Xylella	fastidiosa	ssp. fastidosa	Temecula1	
Xylella	fastidiosa	ssp. fastidosa	CFBP8351	
Xylella	fastidiosa	ssp. fastidosa	IVIA5235	
Xylella	fastidiosa	ssp. fastidosa	M23	
Xylella	fastidiosa	ssp. fastidosa	CFBP8071	
Xylella	fastidiosa	ssp. fastidosa	XYL1732	
Xylella	fastidiosa	ssp. fastidosa	XYL2055	