Antonie van Leeuwenhoek (2017) 110:1271–1279
DOI 10.1007/s10482-017-0849-z

REVIEW

A proposal for a portal to make earth's microbial diversity easily accessible and searchable

Boris A. Vinatzer · Long Tian · Lenwood S. Heath

Received: 8 December 2016 / Accepted: 22 February 2017 / Published online: 9 March 2017

Abstract Estimates of the number of bacterial species range from 10^7 to 10^{12}. At the pace at which descriptions of new species are currently being published, the description of all bacterial species on earth will only be completed in thousands of years. However, even if one day all species were named and described, these names and descriptions would still be of little practical value unless they could be easily searched and accessed, so that novel strains could be easily identified as members of any of these species. To complicate the situation further, many of the currently known species contain significant genotypic and phenotypic diversity that would still be missed if description of microbial diversity were limited to species. The solution to this problem could be a database in which every bacterial species and every intra-specific group is anchored to a genome-similarity framework. This ideal database should be searchable using complete or partial genome sequences as well as phenotypes. Moreover, the database should include functions to easily add newly sequenced novel strains, automatically place them into the genome-similarity framework, identify them as members of an already named species, or tag them as members of yet to be described species or new intra-specific groups. Here, we propose the means to develop such a database by taking advantage of the concept of genome sequence similarity-based codes, called Life Identification Numbers or LINs.

Keywords Genome sequences · Average nucleotide identity · Database · Bacterial species

B. A. Vinatzer (✉) · L. Tian
Department of Plant Pathology, Physiology and Weed Science, Virginia Tech, Latham Hall, Blacksburg, VA 24061, USA
e-mail: vinatzer@vt.edu

L. Tian
GeneticsBioinformatics, and Computational Biology Graduate Program, Virginia Tech, Blacksburg, VA, USA

L. S. Heath
Department of Computer Science, Virginia Tech, Blacksburg, VA, USA

Introduction

Considering that estimates of the number of bacterial species on earth range from 10^7 to 10^{12} (Curtis et al. 2002; Amann and Rossello-Mora 2016; Locey and Lennon 2016), considering that ~700 prokaryotic species are currently described per year (Kamau et al. 2015), and considering that fewer than 15,000 species have been validly named so far (Parte 2013), it would take thousands of years to validly publish names and descriptions for all bacterial species on earth at the current pace (Sutcliffe et al. 2012). Not only is this unacceptably slow, but, even more importantly, accumulating such a huge amount of information would be of little value if the collected information could not be

easily accessed. Therefore, even if eventually scientists finished describing every single bacterial species on earth, we would still only make little progress in understanding microbial diversity without a system to easily access these descriptions. Therefore, it is imperative to develop a tool that will make it fast and easy for scientists to (1) catalogue microbial diversity by depositing genome sequences and metadata of individual strains, (2) describe and name species based on the deposited information, (3) browse and search individual strains and named species, and (4) identify newly discovered strains as members of already named species or flag them as members of new species that need to be described and named. In this article, we propose such a tool. However, first, we need to introduce some basic concepts of current taxonomy, in particular, various nomenclature codes, since these concepts are essential in order to understand the importance of the tool we propose.

The need for species but why species are not sufficient

Many articles have been written about what a bacterial species is (Gevers et al. 2005). While taxonomists continue discussing this fundamental question, at the same time scientists have also continued describing and naming species using a pragmatic species concept that defines species as "monophyletic and genomically-coherent clusters of strains characterised by a high degree of similarity in several independent characteristics" (Rossello-Mora and Amann 2001). Interestingly, although the International Code of Nomenclature of Prokaryotes (Parker et al. 2015) provides strict rules on how to name species, there are no agreed-upon rules on what exactly constitutes "monophyletic and genomically-coherent clusters of strains characterized by a high degree of similarity in several independent characteristics". It is thus left up to reviewers of manuscripts in which newly named species are proposed to accept whether these species are justified or not. However, even in the absence of rules, taxonomists have come to the consensus that one of the essential criteria for naming a new species is that the genome of the type strain of the new species is below a certain threshold of genome similarity compared to the genome of the type strains of any already named species (see more about type strains in the next section). Genome similarity can be measured experimentally using DNA:DNA hybridization (DDH) (Wayne et al. 1987), can be inferred from similarity of 16S rDNA sequences (Stackebrandt and Goebel 1994), or can be measured using bioinformatics by comparing genome sequences and determining the average nucleotide identity (ANI) between genomes (Konstantinidis and Tiedje 2005) or in silico DDH (Meier-Kolthoff et al. 2013). The agreed upon thresholds are 70% DDH, 98.2–99.0 % 16S rDNA sequence identity, and 95–96% ANI (Goris et al. 2007; Richter and Rosselló-Móra 2009; Meier-Kolthoff et al. 2013; Kim et al. 2014).

Importantly, there is also agreement that a newly named species should have one or more diagnosable phenotypes that distinguish it from other named species (Rossello-Mora and Amann 2001; Tindall et al. 2010; Vandamme and Peeters 2014). This leads to the basic motivation as to why we name species: the necessity to have an agreed-upon name for organisms with phenotypes of interest so that we can communicate about them. In other words, identifying an individual bacterial isolate as a member of a named species should be reliably predictive of at least some of its characteristics. Ideally, these characteristics would be the ones considered most relevant. For example, for a pathogenic species, it could be pathogenicity in a certain host plant or animal species or the type of disease caused in these host species. For an environmental bacterium, the phenotype may be the potential to synthesise certain natural products, or it may simply be the role the species plays in the ecosystem and/or the ecological niche it occupies.

However, one of the problems with many of the known species is that they are genetically and phenotypically too diverse to be predictive of the above-mentioned characteristics. The genetically and phenotypically highly diverse species *Escherichia coli* (Wirth et al. 2006; Meier-Kolthoff et al. 2014) and *Pseudomonas syringae* (Berge et al. 2014), which both include various pathogens (of humans and plants, respectively) as well as non-pathogenic commensals, are prime examples. Therefore, classifying, identifying, and naming diversity within species such as *E. coli* or *P. syringae* is very important due to the same basic motivation behind naming species: being able to communicate about intra-specific groups of *E. coli* or *P. syringae* using names that are predictive of the

distinctive phenotypic characteristics of the members of such groups.

The above-mentioned ideal tool to describe and name species, browse and search species, and identify newly discovered strains as members of species should thus go beyond species and allow us to classify, identify, and name intra-specific groups as well.

The "type", the central concept in nomenclature

While there is no agreement on what exactly constitutes a bacterial species, there is agreement on how to describe a newly discovered bacterial species. Moreover, although there are many differences between the ICNP and the various other international codes that regulate naming of viral, animal, and plant species, there is consensus about one thing: names of taxa need to be associated with name-bearing types (ICTV 2016; ICZN (2012); McNeill et al. 2012). In the case of bacteria, types consist in live cultures deposited in international culture collections. In fact, the ICNP states in Sect. 4, Rule 15 (ICSP 2008; Parker et al. 2015): "A taxon consists of one or more elements. For each named taxon of the various taxonomic categories, there shall be designated a nomenclatural type. The nomenclatural type, referred to in this Code as "type," is that element of the taxon with which the name is permanently associated". Even the biocode, a code that is under development to unify the naming of all diversity (bacteria, viruses, plants, algae, fungi, and animals), states in Principle III (Bionomenclature 2011): "The application of names of taxa is determined by means of name-bearing types". Therefore, although the pragmatic bacterial species definition refers to species as "clusters of strains", for the purpose of naming and describing prokaryotic species, a single type strain is sufficient. In fact, ca. 70% of prokaryotic species descriptions are based on single strains. Consequently, most current prokaryotic species descriptions neither capture the genotypic nor the phenotypic diversity that exists within a species. This is perhaps not an issue with genetically and phenotypically monomorphic species such as *Bacillus anthracis* (Van Ert et al. 2007) or *Yersinia pestis* (Cui et al. 2013), since they contain such a limited degree of diversity, but it is very problematic with species like *E. coli* or *P. syringae* and many other species, which comprise many intra-specific groups with many different genotypes and phenotypes.

The purpose of the proposed Phylocode is to name clades instead of types

The Phylocode (Cantino and de Queiroz 2004) provides a set of rules for a phylogenetic nomenclature system whereby clades in a phylogenetic tree are named instead of taxa; although a clade may represent a species (Dayrat et al. 2008). Phylogenetic nomenclature is more stable than other nomenclatural systems because of the absence of ranks, avoiding name changes when the rank of a taxon changes, for example, from species to genus (Cantino and de Queiroz 2004). However, clades themselves are dependent on the choice of genes, the choice of algorithms, and the number of sequences used in phylogenetic reconstruction. A certain organism may thus be placed in one clade in one tree but in a different clade in another tree. Therefore, although the Phylocode may improve stability of names within a phylogeny, it is not concerned with the problem that any gene tree is only an approximation of the "true" species tree. As long as there is no agreement on which genes and which algorithms to use in tree construction, phylogenetic trees will be in flux and clades may change over time.

Importantly for the subject of this article, the Phylocode does not mention "types". Therefore, while the ICNP provides rules on how to name *individual* type strains as representatives of groups or clusters of strains, the Phylocode provides rules on how to name clades in a tree, which ultimately consist of groups or clusters of strains with a common ancestor. We shall see why we think that a comprehensive description of microbial diversity should do both: name species types and circumscribe the group or cluster of strains to which a species name applies.

Life identification numbers (LINs), a code for genome sequences

We have proposed genome codes (Marakeby et al. 2014), later re-named as LINs (Vinatzer et al. 2016; Weisberg et al. 2015), as a way to systematically name individual organisms based on their overall genome sequence similarity to related organisms. In practice, LINs are codes that consist of a series of positions whereby each position reflects a different ANI threshold going from the lowest ANI threshold at the

left-most position of the LIN to the highest ANI threshold at the right-most position of the LIN. Organisms that are very different from each other are thus different at the very left-most position of their LINs, organisms with intermediate similarity share the same LIN from the left up to an intermediate position, and almost identical organisms share the same LIN to almost the right-most position (Fig. 1). Therefore, looking at the LINs of two organisms, the reader can immediately tell the level of similarity between the two organisms.

We have previously described how LINs are assigned (Marakeby et al. 2014; Vinatzer et al. 2016; Weisberg et al. 2015), and Fig. 1 includes a summary of the process as well. Here, we want to emphasise that LINs are assigned sequentially as genomes are added to a LIN database, which we propose to call LINbase. Importantly, to avoid duplication of LINs, there can only be a single LINbase in which LINs are assigned to all organisms. While LINs could conceptually be assigned to any life form, we have focused our recent efforts on bacteria, and we have shown that LINs are highly informative of phylogenetic relationships and can even reflect association of bacterial isolates with individual genetic lineages of plant pathogens that have caused specific plant disease epidemics (Vinatzer et al. 2016).

A key difference of LINs compared to both species and clades, is that LINs are assigned to individual organisms sequentially. Therefore, while discovery of new diversity or refinement of phylogenetic reconstruction will require changes to existing named species and to previously constructed clades, they will not require changes to LINs of individual organisms. Therefore, LINs can provide a stable—or even "static"—genome-similarity framework as an anchor for "dynamic" species descriptions and phylogenetic clades, which necessarily change over time (Fig. 2).

Finally, we propose here to make a slight change to the LIN concept compared to our previous publications. While we have previously stated that LINs are assigned to organisms, we now think that it would be more appropriate to specify that LINs are assigned to genome sequences. In similarity to the concept of "gene trees" and "species trees" whereby phylogenetic trees based on genes are a proxy for the "true" phylogenetic relationships between the actual taxa, LINs assigned to genome sequences are a proxy for the "true" similarity between the actual genomes of the organisms.

Therefore, we would like to propose that taxonomy is a system to identify, classify, and name species, the Phylocode is a system to name clades, and LINs represents a system to name genome sequences, i.e., LINs are a form of genome code.

How LINs can complement type strain-based species descriptions

As stated above, LINs are assigned sequentially to individual genome sequences. However, this does not mean that LINs cannot be used to name groups of genome sequences as well. In fact, any group of genome sequences can be named by the LIN positions shared by its members. We call such a group of genome sequences a LINgroup (Vinatzer et al. 2016). For example, all genomes sequences that share the same LIN up to LIN position C, would be named based on the symbols that they share at these three left-most LIN positions. A LINgroup "name" is thus predictive of the genomic similarity of its members: the more LIN positions it encompasses, the more similar its members are. For example, a LINgroup that encompasses only LIN position A is much more diverse than a LINgroup that encompasses LIN positions up to position D (see also Fig. 2).

As described above, a type strain-based named species description is predictive of phenotypes shared by the type strain with the other members of the same species. However, the knowledge that two organisms belong to the same species does not say anything about how similar they are to each other genetically, besides that they have a minimum level of genome similarity corresponding to a DDH value of at least 70% or an ANI value of at least 95%. Therefore, LINgroups complement type strain-based species descriptions: if the LINs of all genome-sequenced members of a species are known, then the number of LIN positions these members share with the type strain of the species immediately visualises how genetically diverse the species is, i.e., the level of genome similarity among the members of the species.

The same applies to clades. If there is a named clade in a phylogenetic tree, the name of the clade does not reveal how genetically diverse the taxa in that clade are. However, if the genomes of the taxa in a clade are

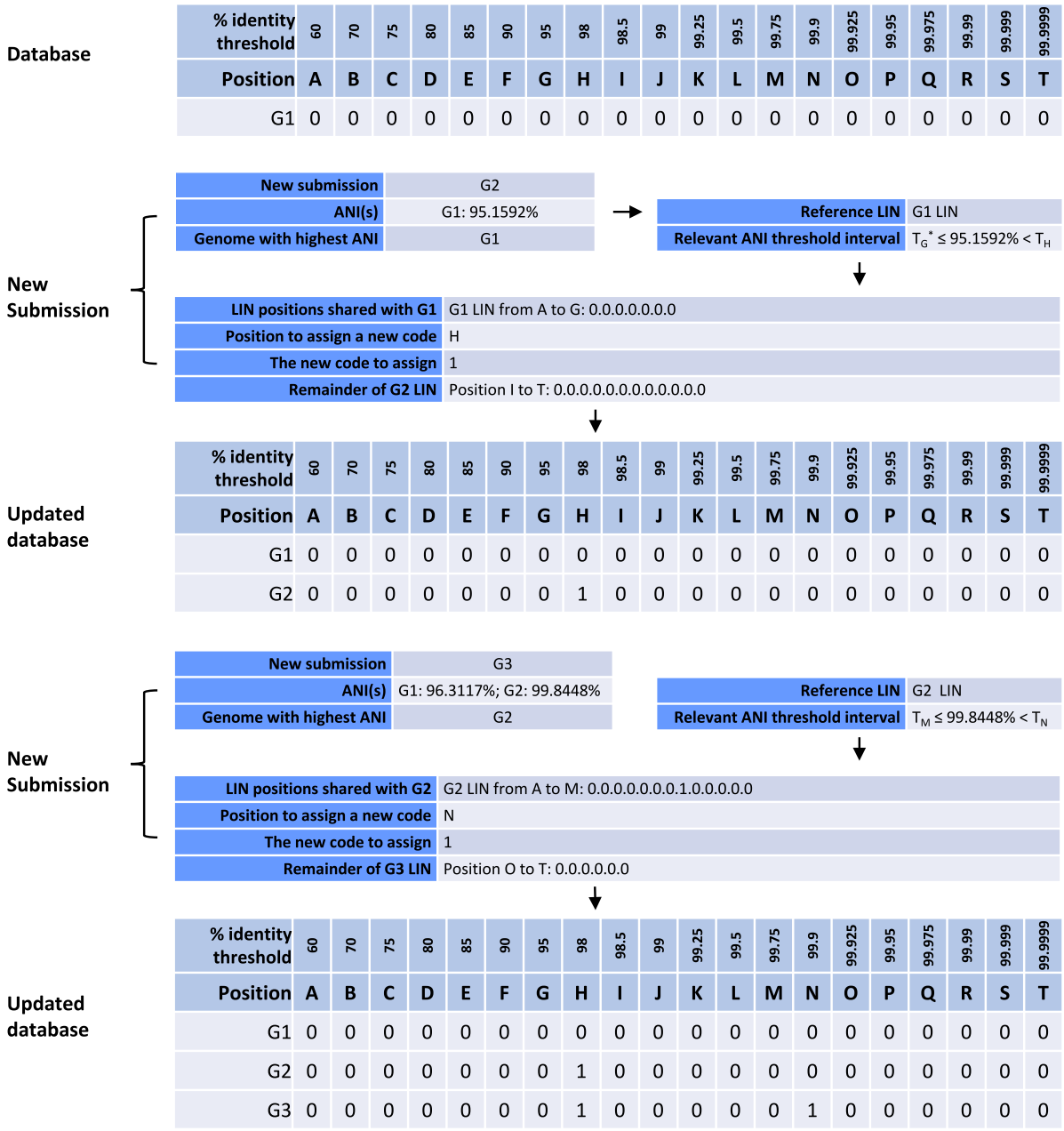

Fig. 1 LINs and their assignment. The assignment of a LIN to a newly submitted genome sequence (for example, genome sequence G2 or genome sequence G3 in this figure) is based on the LIN of the genome in the database that has the highest ANI compared to the newly submitted genome sequence. In the depicted example, when G2 is submitted, the only genome sequence already in the database is G1 (which has all zeros at all positions since it was the very first genome added to the database). Therefore, the genome in the database with the highest ANI compared to G2 is necessarily G1. Consequently, the LIN of G2 is assigned based on the LIN of G1. By the time G3 is submitted, G1 and G2 are already in the database. After ANI calculation, the ANI between G3 and G2 is calculated and found to be higher than that the ANI between G3 and G1. Therefore, the LIN of G3 is assigned based on the LIN of G2. *T_G ANI threshold at LIN position G

Genomes	T*	A	B	C	D	E	F	G	H	I
G1	No	0	0	0	0	0	0	0	0	0
G2	No	0	0	0	0	0	0	1	0	0
G3	No	0	0	0	0	0	0	2	0	0
G4	No	0	0	0	0	0	1	0	0	0
G5	No	0	0	0	0	0	2	0	0	0
G6	No	0	0	0	1	0	0	0	0	0
G7	**Yes**	**0**	**0**	**0**	**1**	**0**	**1**	**0**	**0**	**0**
G8	No	0	0	0	2	0	0	0	0	0
G9	No	0	0	1	0	0	0	0	0	0

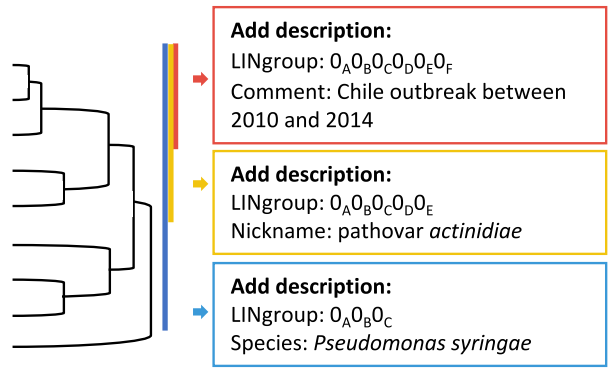

Fig. 2 LINs as anchors for naming and describing taxa in the LINbase database. LINs are assigned sequentially to individual genome sequences in LINbase. LINs thus represent a stable reference framework for naming and describing any group of organisms with similar genome sequences, which we call LINgroups. This figure shows three hypothetical examples of names and descriptions assigned to LINgroups. LINgroup $0_A0_B0_C$ (*highlighted in blue*) includes the type strain of *P. syringae*, which has genome sequence G7 (in *bold*). If, for example, all *P. syringae* genomes in the LINbase database share the LIN $0_A0_B0_C$, then the species *P. syringae* is comprehensively described in LINbase by (1) the description associated with the type strain G7; (2) the genomic diversity associated with the LIN $0_A0_B0_C$. The LINgroup $0_A0_B0_C0_D0_E$ (*highlighted in yellow*) corresponds to an infra-specific group of *P. syringae* called pathovar *actinidiae*. Therefore, the LIN $0_A0_B0_C0_D0_E$ is predictive of this named group of pathogen strains that cause disease on kiwifruit although the name does not correspond to a named species and no type strain is associated with this LINgroup. Finally, the LINgroup $0_A0_B0_C0_D0_E0_F$ (*highlighted in red*) represents a sub-group within pathovar *actinidiae*; all its members were isolated in Chile between 2010 and 2014. Therefore, the LIN $0_A0_B0_C0_D0_E0_F$ is predictive of a genetic lineage of the canker of kiwifruit pathogen that caused a disease outbreak in this specific country in this specific time frame. *T* Type strain

sequenced, then the LINs assigned to the respective genome sequences demonstrate immediately how genetically diverse the clade is.

Importantly, if a LINgroup consists of genome sequences of organisms that do not correspond to a species but simply to an intra-specific group in which members share a common phenotype of interest, then this phenotype can be associated with the LINgroup and the LINgroup becomes predictive of this specific phenotype (Fig. 2).

Instead of a phenotype, we may simply be interested in the association of pathogenic isolates with a specific disease outbreak. Therefore, by including geographic location and time of isolation of pathogen isolates when uploading genome sequences to the proposed LINbase database, the LIN positions shared by isolates within a geographic range and time frame of isolation corresponding to a specific disease outbreak become predictive of association with the specific disease outbreak. The strain that caused the outbreak can then be named using the LIN positions shared by the isolates belonging to it. If a related strain causes a disease outbreak somewhere else at a different time, the LIN positions shared between the new outbreak strain and the previous outbreak strains are informative of how genomically similar the two outbreak strains are. Examples are our work on genome sequences of the Ebola virus disease epidemic in West Africa in 2014 (Weisberg et al. 2015) or our work on the recent canker of kiwifruit epidemic caused by an intra-specific group of *P. syringae*: *P. syringae* pathovar (pv.) *actinidiae* (Vinatzer et al. 2016). In both examples, neither current strain names nor clade names nor taxa designations inform about the similarity between epidemic clones whereas LINs do.

How to implement LINs in a genome similarity database

The conceptual framework described so far is, of course, of no practical benefit unless there is a tool to translate this framework into something that taxonomists and epidemiologists can easily use. Also, while many genome databases exist today, there is no database that allows the use of genome sequences to describe and name bacterial diversity. There is not even a searchable database for bacterial species descriptions since species descriptions are still

presented in traditional peer-reviewed manuscripts, although new initiatives such as the 'digital protologue' are emerging (Rosselló-Móra et al. 2017). Therefore, we are currently developing a LINbase prototype. A description of LINbase functions follows.

LINbase functions

1. Genome submission and LIN assignment
   A new genome sequence submitted by a user will be uploaded to the back end of the service along with metadata. After performing ANI calculations of the newly submitted genome against all genomes already in the database, the genome with the highest ANI to the newly submitted genome is identified and a LIN is assigned to the new genome based on the LIN of this genome. An email will be sent to the user after the new LIN is assigned and recorded in the database, informing the user that a result page has been generated including the new LIN, the LIN of the most similar genome that was already in the database, the ANI between the newly submitted genome and this genome, and a list of other similar genomes in LINbase sorted by their LINs. Alternatively, the user can view a list of named LINgroups to which the newly uploaded genome belongs (see below). The LINbase database grows per usage in real time and is managed by its own schema and is self-curated by LINs without any manual intervention.
2. Adding names and comments to a LINgroup
   This function will be embedded in the aforementioned result pages. Genomes displayed in the result pages will be sorted by their LINs whereby genomically similar genomes will be visually grouped together. Therefore, it will be easy for users to select any LINgroup of their interest by simply selecting the LIN positions that are shared by all its members. Users can then name the LINgroup and add comments to it. However, it is to be emphasised that proposed new species names will be expected to comply with the nomenclatural rules of the ICNP (Parker et al. 2015).
3. Searching genomes
   Users will be able to search LINbase by: metadata associated with individual genomes, LINs, LINgroup descriptions, genome sequences and gene sequences. If there are any entries in the database that match search terms or query sequences, the results page will display them in the browser.

How to describe, name, and search microbial taxa using LINbase

Next, we describe how LINbase can be used for: (1) taxa descriptions based on types; (2) taxa and clade descriptions without types, based exclusively on genome sequences of members; (3) comprehensive taxa descriptions including both type-based and member-based descriptions; and (4) searching species using genomes of unidentified strains.

1 Taxa descriptions based on types
   Species descriptions associated with a type strain can be directly submitted to the database together with the genome sequence of the type strain. The type strain will serve as an anchor to place the species into the genome-similarity framework based on its assigned LIN (Fig. 2).
2 Taxa and clade descriptions based on genome sequences of members
   All other genome sequences, i.e., all genome sequences that are not genome sequences of type strains, can be uploaded with a set of chosen metadata including phenotypic descriptions. Upon uploading, a LIN will be assigned, which will place the new genome into the genome-similarity framework. The user who uploaded the sequence, or any other user, can then group this new genome with genomes already in LINbase very easily. All that is needed is to select a number of genomes that share the same LIN up to a certain position and at least one value in at least one category of metadata, in other words, at least one common phenotype. This could simply be the location of where the organisms were isolated in order to describe, for example, a group of pathogen isolates associated with a disease outbreak; or it could be a shared phenotype, for example, pathogenicity on kiwifruit plants for a group of plant pathogens (Fig. 2). Although the group is automatically named by the shared LIN positions, the user can also give that group a "nickname", which can be anything other than a Latin binomial, which is

reserved for LINgroups that correspond to validly named species (see below).

Ultimately, one of the most important properties of a taxonomy is that the name of a group of related organisms is predictive of a phenotype of the members of that group. Therefore, LINgroups fulfill one of the primary purposes of taxonomy.

3 Comprehensive taxa descriptions based on types and genome sequences of members

If a LINgroup includes the genome of a type strain of a validly named species and all members of the same LINgroup share the description of the species, then the LINgroup can be named with the Latin binomial of the species that is associated with the type strain (Fig. 2).

Importantly, as pointed out earlier, such a LINgroup provides more information than a species description that is based exclusively on a type strain because combining description of a taxon based on a type strain with that based on genomic diversity of its members allows a comprehensive description of a taxon that is predictive of both phenotype and genomic diversity.

4. Searching LINbase with genome sequences of unidentified strains

Finally, if a genome of an unidentified strain is submitted to LINbase, the ANI compared to the most similar type strain genome will immediately tell the user if the strain is a member of an already named species or a member of any named LINgroup representing any particular intra-specific group with a particular phenotype, or if it possibly represents a new species that needs to be characterised phenotypically and named.

Moreover, any gene sequence and any keyword corresponding to any species name or any LINgroup name or any of the metadata or descriptions associated with individual strains or LINgroups or species in LINbase can be used to search LINbase as well. Every search result will lead back to a LIN, which places each search result precisely into the underlying genome similarity framework of LINbase.

Conclusions

Advances in genome sequencing technology and computer science make it possible today to revolutionise the discipline of bacterial taxonomy. We hope that the here-described proposal can be implemented soon and help accelerate this revolution and make cataloguing, describing, naming, and searching the immense genetic diversity of bacteria almost as easy as posting and sharing photos or searching for a friend on social media. The technology exists. The question is just how fast we are able to implement and deploy it. We expect to launch a Linbase prototype in 2017 for beta testing by the scientific community.

Acknowledgements LSH was supported by National Science Foundation Grant DBI-1062472. BAV and LT were supported by National Science Foundation grant IOS-1354215. Funding for work in the Vinatzer laboratory was also provided in part by the Virginia Agricultural Experiment Station and the Hatch Program of the National Institute of Food and Agriculture, US Department of Agriculture.

References

Amann R, Rossello-Mora R (2016) After all, only millions? MBio. doi:10.1128/mBio.00999-16

Berge O, Monteil CL, Bartoli C, Chandeysson C, Guilbaud C, Sands DC, Morris CE (2014) 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 9:e105547. doi:10.1371/journal.pone.0105547

Bionomenclature ICo (2011) Bionomenclature Across All Groups of Organisms. http://www.bgbm.org/biodivinf/docs/biocode2011/biocode2.html—Introduction. Accessed Nov 23 2016

Cantino PD, de Queiroz K (2004) The Phylocode. http://www.ohio.edu/phylocode/PhyloCode4c.pdf. 2013

Cui Y et al (2013) Historical variations in mutation rate in an epidemic pathogen, Yersinia pestis. Proc Natl Acad Sci 110:577–582. doi:10.1073/pnas.1205750110

Curtis TP, Sloan WT, Scannell JW (2002) Estimating prokaryotic diversity and its limits. Proc Natl Acad Sci 99:10494–10499. doi:10.1073/pnas.142680199

Dayrat B, Cantino PD, Clarke JA, de Queiroz K (2008) Species names in the phylocode: the approach adopted by the international society for phylogenetic nomenclature. Syst Biol 57:507–514. doi:10.1080/10635150802172176

Gevers D et al (2005) Opinion: re-evaluating prokaryotic species. Nat Rev Microbiol 3:733–739

Goris J, Konstantinidis KT, Klappenbach JA, Coenye T, Vandamme P, Tiedje JM (2007) DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. Int J Syst Evol Microbiol 57:81–91

ICSP ICoSoP (2008) International code of nomenclature of prokaryotes (2008 Revision) [DRAFT]. http://code.icsp.org/. Accessed 22 Sept 2014

ICTV ICoToV (2016) The international code of virus classification and nomenclature. http://www.ictvonline.org/codeOfVirusClassification.asp. Accessed 23 Nov 2016

ICZN ICoZN (2012) International Code of Zoological Nomenclature. http://www.nhm.ac.uk/hosted-sites/iczn/code/. Accessed 23 Nov 2016

Kamau EC, Winter G, Stoll P-T (2015) Research and development on genetic resources: public domain approaches in implementing the nagoya protocol. Routledge, Abingdon

Kim M, Oh H-S, Park S-C, Chun J (2014) Towards a taxonomic coherence between average nucleotide identity and 16S rRNA gene sequence similarity for species demarcation of prokaryotes. Int J Syst Evol Microbiol 64:346–351. doi:10.1099/ijs.0.059774-0

Konstantinidis KT, Tiedje JM (2005) Genomic insights that advance the species definition for prokaryotes. Proc Natl Acad Sci U S A 102:2567–2572

Locey KJ, Lennon JT (2016) Scaling laws predict global microbial diversity. Proc Natl Acad Sci 113:5970–5975. doi:10.1073/pnas.1521291113

Marakeby H et al (2014) A system to automatically classify and name any individual genome-sequenced organism independently of current biological classification and nomenclature. PLoS ONE 9:e89142. doi:10.1371/journal.pone.0089142

McNeill J et al. (2012) International code of nomenclature for algae, fungi, and plants (Melbourne Code) http://www.iapt-taxon.org/nomen/main.php. Accessed 23 Nov 2016

Meier-Kolthoff JP, Auch AF, Klenk H-P, Göker M (2013) Genome sequence-based species delimitation with confidence intervals and improved distance functions. BMC Bioinform 14:60. doi:10.1186/1471-2105-14-60

Meier-Kolthoff JP et al (2014) Complete genome sequence of DSM 30083T, the type strain (U5/41T) of Escherichia coli, and a proposal for delineating subspecies in microbial taxonomy. Stand Genom Sci. doi:10.1186/1944-3277-9-2

Parker CT, Tindall BJ, Garrity GM (2015) International Code of Nomenclature of Prokaryotes. Int J Syst Evolut Microbiol. doi:10.1099/ijsem.0.000778

Parte AC (2013) List of prokaryotic names with standing in nomenclature. http://www.bacterio.net/-number.html—total. Accessed 31 Jan 2017

Richter M, Rosselló-Móra R (2009) Shifting the genomic gold standard for the prokaryotic species definition. Proc Natl Acad Sci 106:19126–19131. doi:10.1073/pnas.0906412106

Rossello-Mora R, Amann R (2001) The species concept for prokaryotes. FEMS Microbiol Rev 25:39–67

Rosselló-Móra R, Trujillo ME, Sutcliffe IC (2017) Introducing a digital protologue: a timely move towards a database-driven systematics of archaea and bacteria. Antonie Van Leeuwenhoek. doi:10.1007/s10482-017-0841-7

Stackebrandt E, Goebel BM (1994) Taxonomic note: a place for DNA-DNA reassociation and 16S rRNA sequence analysis in the present species definition in bacteriology. Int J Syst Bacteriol 44(4):846–849

Sutcliffe IC, Trujillo ME, Goodfellow M (2012) A call to arms for systematists: revitalising the purpose and practises underpinning the description of novel microbial taxa. Antonie Van Leeuwenhoek 101:13–20. doi:10.1007/s10482-011-9664-0

Tindall BJ, Rosselló-Móra R, Busse H-J, Ludwig W, Kämpfer P (2010) Notes on the characterization of prokaryote strains for taxonomic purposes. Int J Syst Evol Microbiol 60:249–266. doi:10.1099/ijs.0.016949-0

Van Ert MN et al (2007) Global genetic population structure of Bacillus anthracis. PLoS ONE 2:e461. doi:10.1371/journal.pone.0000461

Vandamme P, Peeters C (2014) Time to revisit polyphasic taxonomy. Antonie Van Leeuwenhoek 106:57–65. doi:10.1007/s10482-014-0148-x

Vinatzer BA, Weisberg AJ, Monteil CL, Elmarakeby HA, Sheppard SK, Heath LS (2016) 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: PHYTO-07-16-0252-R doi: 10.1094/PHYTO-07-16-0252-R

Wayne LG et al (1987) Report of the ad hoc committee on reconciliation of approaches to bacterial systematics. Int J Syst Bacteriol 37:463–464

Weisberg AJ, Marakeby H, Heath LS, Vinatzer BA (2015) Similarity-based codes sequentially assigned to ebolavirus genomes are informative of species membership, associated outbreaks, and transmission chains. Open Forum Infect Dis. doi:10.1093/ofid/ofv024

Wirth T et al (2006) Sex and virulence in Escherichia coli: an evolutionary perspective. Mol Microbiol 60:1136–1151