Published online 30 March 2020 *Nucleic Acids Research, 2020, Vol. 48, Web Server issue* ***W529–W537***
doi: 10.1093/nar/gkaa190

LINbase: a web server for genome-based identification of prokaryotes as members of crowdsourced taxa

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

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

Received February 06, 2020; Revised March 04, 2020; Editorial Decision March 12, 2020; Accepted March 16, 2020

ABSTRACT

High throughput DNA sequencing in combination with efficient algorithms could provide the basis for a highly resolved, genome phylogeny-based and digital prokaryotic taxonomy. However, current taxonomic practice continues to rely on cumbersome journal publications for the description of new species, which still constitute the smallest taxonomic units. In response, we introduce LINbase, a web server that allows users to genomically circumscribe any group of prokaryotes with measurable DNA similarity and that uses the individual isolate as smallest unit. Since LINbase leverages the concept of Life Identification Numbers (LINs), which are codes assigned to individual genomes based on reciprocal average nucleotide identity, we refer to groups circumscribed in LINbase as LINgroups. Users can associate with each LINgroup a name, a short description, and a URL to a peer-reviewed publication. As soon as a LINgroup is circumscribed, any user can immediately identify query genomes as members and submit comments about the LINgroup. Most genomes currently in LINbase were imported from GenBank, but users can upload their own genome sequences as well. In conclusion, LINbase combines the resolution of LINs with the power of crowdsourcing in support of a highly resolved, genome phylogeny-based digital taxonomy. 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 (1). Precise identification of prokaryotes is also important in many other fields, e.g. when regulating commercial probiotics for human consumption (2) or biopesticides to control plant diseases in agriculture (3). While we often associate the process of identification with assigning an unknown organism to a named group, i.e. a taxon, the pragmatic goal of identification is to predict the characteristics, i.e. the phenotype, of the unknown organism, e.g. to answer a question such as: does the unknown microorganism cause a certain disease in a specific animal species? The prerequisite for such precise identification is precise classification (4). In fact, identification of an unknown as a member of a taxon can only lead to precise prediction of its phenotype when the taxon consists of members that are derived from a most recent common ancestor (MRCA), i.e. the taxon is a monophyletic group, and its members share a phenotype absent from organisms outside of that same taxon. For example, *Bacillus anthracis* is such a taxon since identifying a bacterial isolate as a member of this species predicts that the unknown will cause the disease anthrax. *Escherichia coli* is instead a taxon that is poorly predictive of the characteristics of its members. Members of the species *E. coli* may cause no disease at all, they may cause urinary tract infections or they may cause bloody diarrhea and cause death due to hemolytic-uremic syndrome.

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 (5). Classification and identification then transitioned to more precise gene-based methods, in particular, sequencing of the 16S rRNA gene (6). 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 to identify them as members of these groups based on identification of single nucleotide polymorphisms (7), construction of phylogenetic trees based on conserved genes (8,9) or measures of genome similarity at the whole genome level expressed as average nucleotide identity (ANI) (10). Conceptually, a taxon could now consist of microbes that share nothing other than a single mutation inherited from their MRCA

*To whom correspondence should be addressed. Tel: +1 540 231 2126; Email: vinatzer@vt.edu

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. For example, identifying an isolate as a member of a taxon characterized by a single mutation that confers antibiotic resistance can predict that the isolate has antibiotic resistance.

However, the fundamental unit of current taxonomy is not the smallest distinguishable unit based on genome sequencing, but it is the species, whereby each species is associated with a 'type' strain, which is considered the name-bearing strain of the species (11). Since current taxonomy is grounded in microbiological history, the valid publication of a new microbial species generally includes much more than sequencing the genome of a type strain and reporting a distinctive phenotype. Besides showing that the type strain of the new species has <95% ANI compared to genomes of type strains of already named species, valid publication entails a long list of results derived from laboratory-based phenotypic tests (12). 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, as has been proposed for the Digital Protologue Database (13). The necessity of a cultured type strain for descriptions of named species has also been questioned, and using genome sequences as type material for unculturable organisms has been proposed (14). 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, e.g. a single plant pathogen species may include many strains with different host ranges (15). Although classification schemes at intraspecific levels exist, they are not consistent across species. This makes it difficult for scientists who do not have familiarity with a particular species-specific scheme to interpret identification results based on that 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 (16,17). 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 (Figure 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 that is shared by two genomes, the more similar the genomes are to each other. The numbers at each LIN position are used as symbols and not as values. For example, two genomes that have a reciprocal ANI value of over 70% but <75%, only share the same symbol at the A position (corresponding to the 70% ANI Threshold) but are different at the B position (corresponding to the 75% ANI threshold). Two genomes that instead have a reciprocal ANI value of over 99% but <99.25%, share the same symbols from position A to position K (corresponding to the 99% ANI threshold), but are different at the L position (corresponding to the 99.25% ANI threshold). 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 (16).

Any group of bacteria that share a LIN prefix of any length is called a LINgroup (18). 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. In this way, LINgroups allow for classification and identification independently of named species and take full advantage of the precision of genome sequencing. However, validly published named species and genera can be circumscribed as LINgroups as well. For example, if all known members of a named species share a certain LIN prefix, e.g. $0_A1_B0_C0_D0_E4_F$, 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 server that implements the LIN and LINgroup concepts using an SQL database, efficient algorithms and an intuitive web server. 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, uploaded 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 capacity, 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. All code can be accessed through Virginia Tech's GitLab repository at https://code.vt.edu/linbaseproject/linbase_back. The server currently runs on an Intel Xeon 16-core CPU at 1.90GHz,

			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 1. The LIN and LINgroups concept. Each LIN position (A–T) represents a different ANI threshold, ranging from 70% at position A to 99.999% at position T. Therefore, the more similar two genomes are, the further to the right their LINs overlap. Note that numbers are used as symbols and not as values. For example, strains X1 and X2 are identical at LIN position A but are different at LIN position B because they are over 70%, but <75%, identical to each other. LINgroups are used to describe groups of microbes based on how similar their members are to each other and are denoted by the LIN prefix their members share. For example, strains X2 and X3 belong to the species S2 since they have a reciprocal ANI of over 95%, the commonly accepted ANI threshold for species. Species S2 is thus denoted by the LIN prefix up to position F (corresponding to the 95% ANI threshold): $0_A1_B0_C0_D0_E3_F$. All genomes that share this LIN prefix belong to species S2. However, since in our example strains X2 and X3 are between 98.5 and 99% identical to each other, they also belong to the LINgroup $0_A1_B0_C0_D0_E3_F0_G0_H0_I0_J$. Therefore, LINgroups precisely reflect the reciprocal genome similarity of their members and can be used to describe any group of bacteria with reciprocal ANI values between 70 and 99.999% while species are limited to a single ANI threshold of 95% ANI.

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 Supplementary Figure S1. Each table has a primary key and is connected to other tables with a foreign key. There are four 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 FUNCTIONS

Access

Users can either access LINbase as guest without registering or with a personal account after registration. In both cases, the history of user activities will be saved in the user's browser and not be visible to other users. A link to a quick start guide with instructions on how to use LINbase is available on the home page. The start guide also includes a link to a sample genome for testing purposes. While all functions are available without registering a personal account, a personal account is advised when uploading genomes and describing LINgroups so that genomes and LINgroups can be traced to the submitter.

Genome upload function

The goal of LINbase administrators is to add all prokaryotic 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 (manuscript in preparation). In short, *k*-mer signatures are computed using sourmash (19), with parameters $k = 21$ and $k = 51$. The computed signatures are 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 $J \geq 0.2475$ (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 https://github.com/widdowquinn/pyani. The computed ANI value is then

Table 1. The 'Metadata' table of LINbase

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

Each Metadata ID is associated with a category of metadata (Metadata_Item).

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 LIN of the genome with the highest Jaccard similarity 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 LIN assignment to a new genome is currently 3 min and 54 s.

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 (since the user may not have that information). Other metadata can be entered based on a user's selection of 'Interest' (Tables 1 and 2). Currently, the following interests are available (but additional interests and additional metadata options can be added upon contacting LINbase administrators at LINbase@vt.edu): Undefined interest, Plant pathogens, Environmental bacteria, Uncultured bacteria, Foodborne pathogens and Archaea (Figure 2).

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 circumscribed 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, e.g. a clade or phylogroup within a species), 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 3).

We expect users to generally choose the longest LIN prefix shared by a group of genomes when describing a LINgroup. For example, since all genomes of the genus *Pseudomonas* in LINbase share the LIN prefix 50_A, a user would describe the LINgroup 50_A as genus *Pseudomonas*. However, instead of choosing the maximum length of the LINprefix shared by a group of genomes, a user could 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. At last, for circumscriptions of named species, a LINgroup can also be described based on the genome of the type strain of the species and choosing the LINprefix up to position F, which correspond to the broadly accepted ANI thresholds for speciation at 95%.

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. However, only LINgroup circumscriptions submitted by registered users with validated credentials will be stored in LINbase. LINgroups submitted by unknown guest users will be deleted.

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 <1 s 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 (Supplementary Figure S2A). 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 (Supplementary Figure S3A).

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 (Supplementary Figure S2B). The result page will list the described LINgroups that match the query (Supplementary Figure S3B).

Identify function using gene or genome sequences as query

Users can identify an unknown prokaryote 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 4A). However, to achieve higher speed with only moderate reduction in accuracy, FastANI (20) 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

Table 2. The 'Interest' table of LINbase

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

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

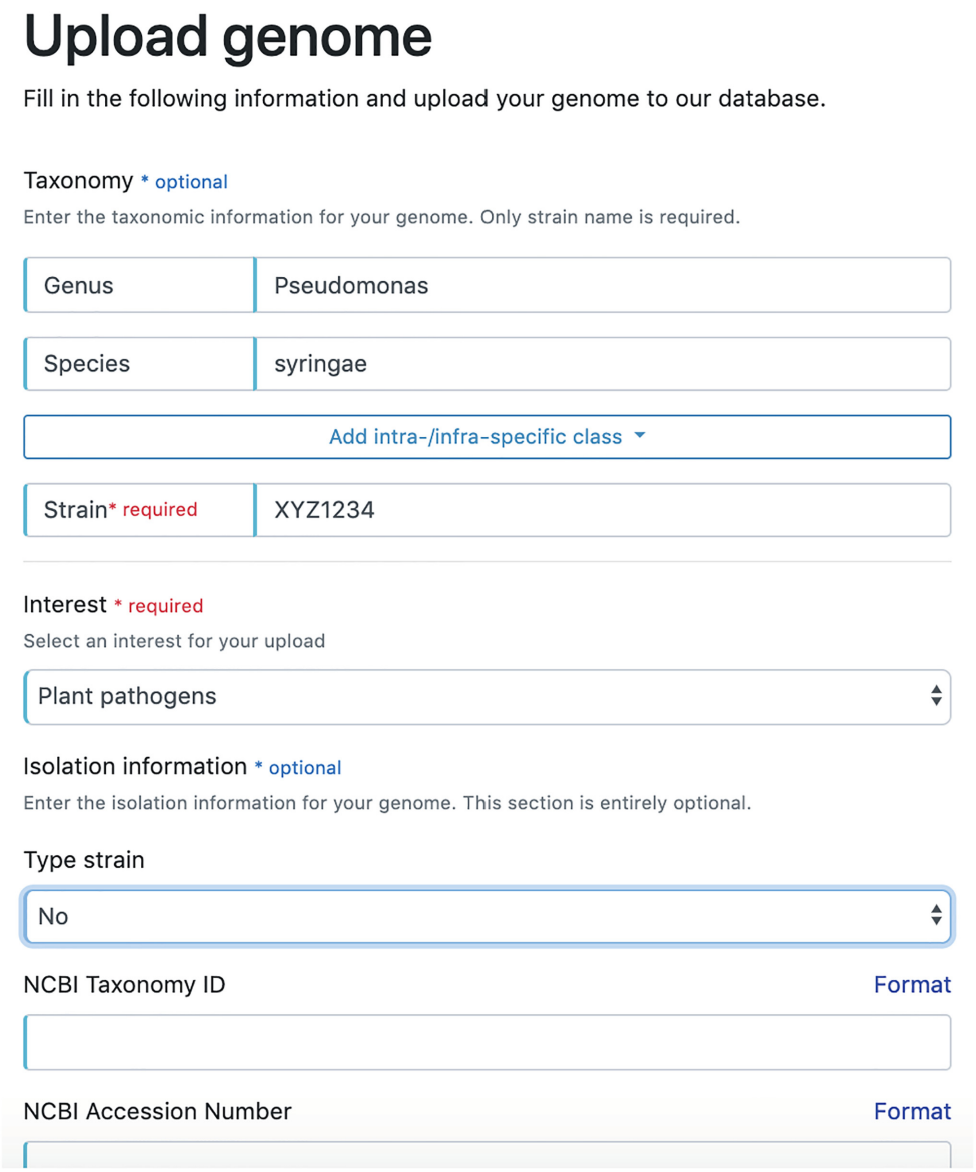

Figure 2. 'Upload genome' form. Users are asked to enter taxonomic information and isolation metadata before uploading microbial genome sequences. In the taxonomic information section, only strain name is required as an identifier of the uploaded genome since taxonomic information may not be available. 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'.

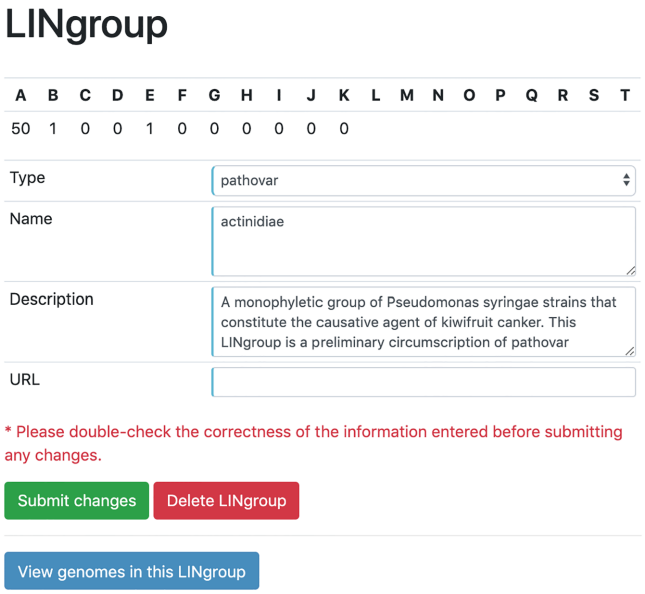

Figure 3. 'Add a Lingroup description' form. After an undescribed LINgroup is selected, the user can describe the LINgroup at a taxonomic rank or as a group of microbes within a species that share a phenotype. This is done by choosing the type of taxon from the 'Type' dropdown menu and entering a name and an optional comment and/or optional link to a peer-reviewed publication.

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 5A).

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 4B) (21). 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 5B).

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 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 DNA sequences private until submission to a public database.

CONCLUSION

Here we introduced LINbase, a web server that implements bacterial taxonomy based on whole genome similarity and supported by fast and accurate algorithms. LINbase complements functionalities offered by other online web servers for genome-based microbial identification, such as MiGA (22) or EzBioCloud (23), as follows: (i) it labels individual genomes with LINs, which reflect the precise genomic relatedness among strains in the database; (ii) it automatically gathers genomically similar bacteria into taxa (LINgroups); (iii) 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; (iv) it uses crowdsourcing to incorporate informal taxa/LINgroups independently of published named taxa; (v) 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 (vi) it allows users to interact with each other by commenting about LINgroup circumscriptions and descriptions.

Despite the aforementioned advantages of LINbase, there are limitations in its current version in regard to the classification of prokaryotes 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 or by basing the left-most LIN position on the similarity of conserved genes. At the other extreme, we are planning to improve assignment of isolates to outbreaks by integrating additional algorithms to precisely identify phylogenetic relationships among very similar genomes. At last, 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

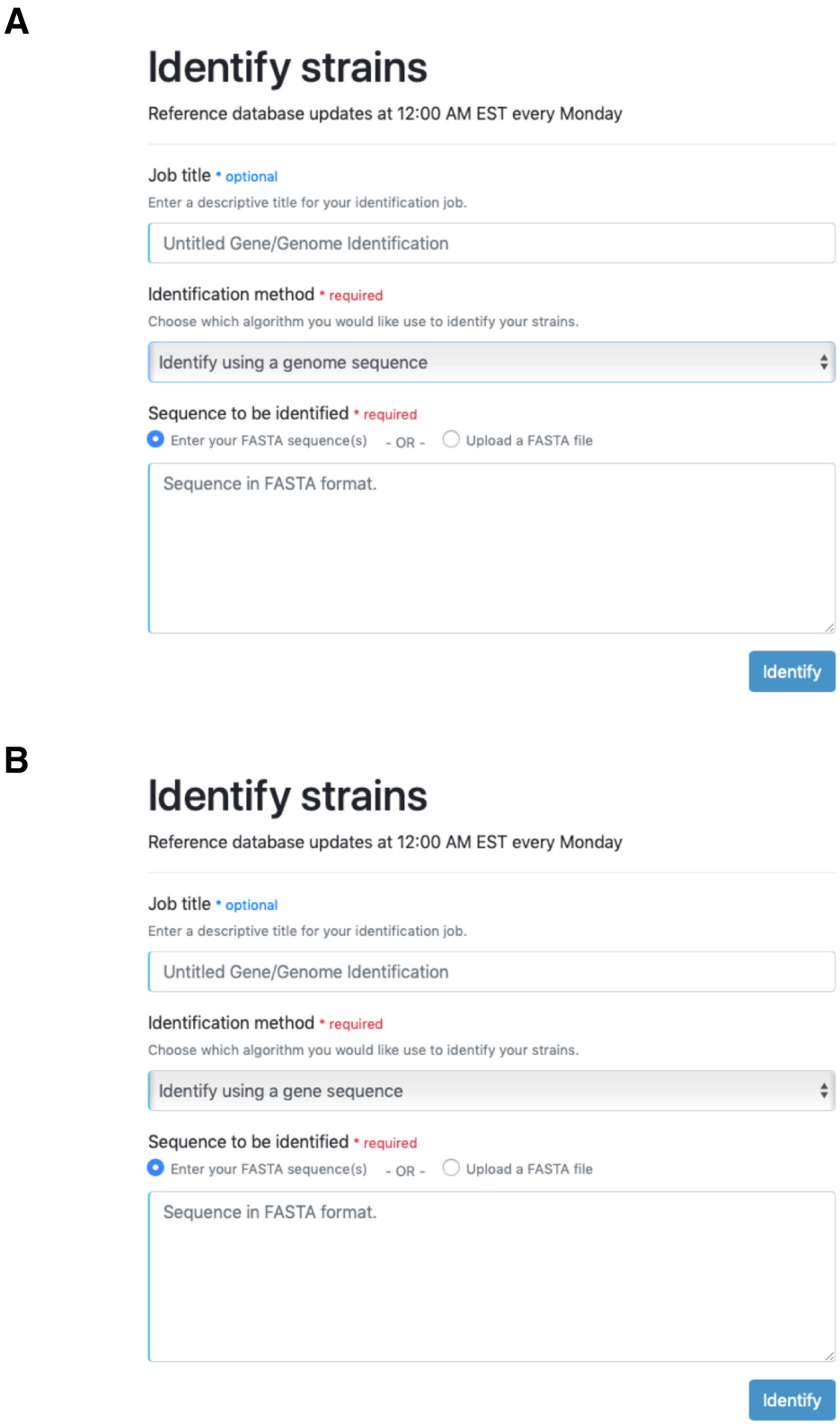

Figure 4. 'Identify' forms. (**A**) Identification with a genome assembly. (**B**) Identification with a gene sequence. Both functions accept gene or genome sequences uploaded as a FASTA-format file or entered in the textbox.

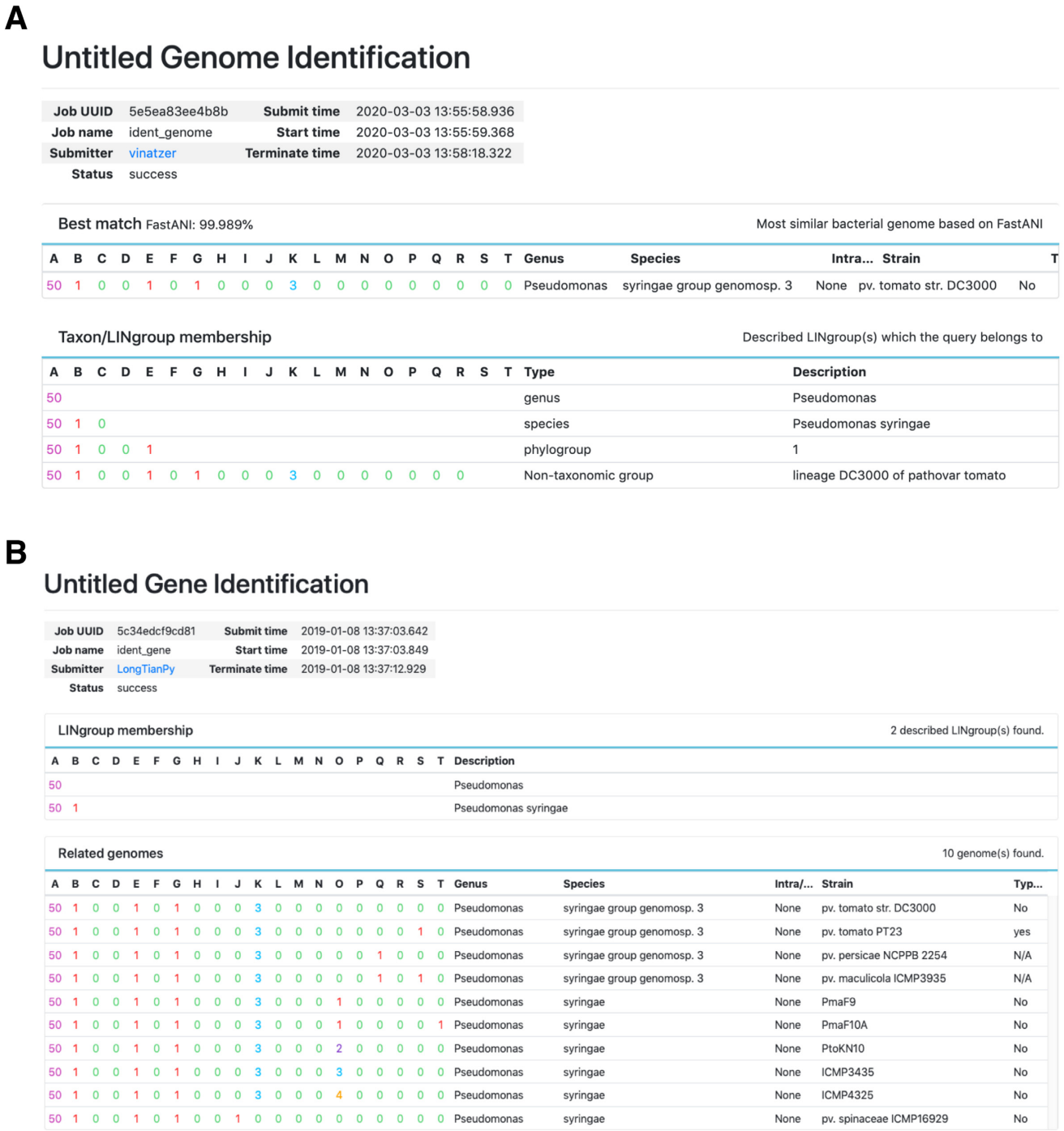

Figure 5. 'Identify' result pages. (**A**) Result page for genome-based identification. The submitted genome is queried against LINbase genomes and the genome with the highest FastANI is returned. The LINgroups that the query genome belongs to (based on its ANI with the best match) are listed as well. (**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.

(24), and to integrate LINbase with other platforms to improve genome-based classification and identification of microbes at all taxonomic ranks.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the undergraduate students of Computer Science Department at Virginia Polytechnic Institute and State University, Grant Hughes, Vincent Eastman and Teresa Paul for participating in the early design of the user interface.

FUNDING

National Science Foundation [IOS-1354215, in part]; College of Agriculture and Life Sciences, Virginia Polytechnic Institute and State University; Virginia Agricultural Experiment Station (to B.A.V.) (in part); Hatch Program of the National Institute of Food and Agriculture, US Department of Agriculture (to B.A.V.). Funding for open access charge: National Science Foundation; Institutional Funds.
Conflict of interest statement. Life Identification Number® and LIN® are registered trademarks of This Genomic Life Inc. L.S.H. and B.A.V. report in accordance with Virginia Tech policies and procedures and their ethical obligation as researchers, that they have a financial interest in This Genomic Life Inc that may be affected by the research reported in this manuscript. They have disclosed those interests fully to Virginia Tech, and they have in place an approved plan for managing any potential conflicts arising from this relationship.

REFERENCES

1. Bird,B.H. and Mazet,J.A.K. (2018) Detection of emerging zoonotic pathogens: an integrated one health approach. *Annu. Rev. Anim. Biosci.*, **6**, 121–139.
2. Huys,G., Vancanneyt,M., D'Haene,K., Vankerckhoven,V., Goossens,H. and Swings,J. (2006) Accuracy of species identity of commercial bacterial cultures intended for probiotic or nutritional use. *Res. Microbiol.*, **157**, 803–810.
3. Velivelli,S.L.S., De Vos,P., Kromann,P., Declerck,S. and Prestwich,B.D. (2014) Biological control agents: from field to market, problems, and challenges. *Trends Biotechnol.*, **32**, 493–496.
4. Sutcliffe,I.C., Trujillo,M.E. and 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.
5. Thompson,C.C., Chimetto,L., Edwards,R.A., Swings,J., Stackebrandt,E. and Thompson,F.L. (2013) Microbial genomic taxonomy. *BMC Genomics*, **14**, 913.
6. Stackebrandt,E. and Goebel,B.M. (1994) Taxonomic note: a place for DNA-DNA reassociation and s rRNA sequence analysis in the present species definition in bacteriology. **44**, 846–849.
7. Robins-Browne,R.M., Holt,K.E., Ingle,D.J., Hocking,D.M., Yang,J. and Tauschek,M. (2016) Are escherichia coli pathotypes still relevant in the era of whole-genome sequencing? *Front. Cell Infect. Microbiol.*, **6**, 141–141.
8. Maiden,M.C.J., Jansen van Rensburg,M.J., Bray,J.E., Earle,S.G., Ford,S.A., Jolley,K.A. and McCarthy,N.D. (2013) MLST revisited: the gene-by-gene approach to bacterial genomics. *Nat. Rev. Microbiol.*, **11**, 728–736.
9. Mellmann,A., Harmsen,D., Cummings,C.A., Zentz,E.B., Leopold,S.R., Rico,A., Prior,K., Szczepanowski,R., Ji,Y., Zhang,W. *et al.* (2011) Prospective genomic characterization of the German enterohemorrhagic *Escherichia coli* O104:H4 outbreak by rapid next generation sequencing technology. *PLoS One*, **6**, e22751.
10. Konstantinidis,K.T. and Tiedje,J.M. (2005) Towards a genome-based taxonomy for prokaryotes. *J. Bacteriol.*, **187**, 6258–6264.
11. Tindall,B.J. and Garrity,G.M. (2008) Proposals to clarify how type strains are deposited and made available to the scientific community for the purpose of systematic research. *Int. J. Syst. Evol. Microbiol.*, **58**, 1987–1990.
12. Fournier,P.-E., Raoult,D. and Drancourt,M. (2016) New species announcement: a new format to prompt the description of new human microbial species. *New Microbes New Infect.*, **15**, 136–137.
13. Rosselló-Móra,R., Trujillo,M.E. and Sutcliffe,I.C. (2017) Introducing a digital protologue: a timely move towards a database-driven systematics of archaea and bacteria. *Antonie Van Leeuwenhoek*, **110**, 455–456.
14. Konstantinidis,K.T., Rosselló-Móra,R. and Amann,R. (2020) Advantages outweigh concerns about using genome sequence as type material for prokaryotic taxonomy. *Environ. Microbiol.*, **22**, 819–822.
15. Baltrus,D.A., McCann,H.C. and Guttman,D.S. (2017) Evolution, genomics and epidemiology of Pseudomonas syringae. *Mol. Plant Pathol.*, **18**, 152–168.
16. Marakeby,H., Badr,E., Torkey,H., Song,Y., Leman,S., Monteil,C.L., Heath,L.S. and Vinatzer,B.A. (2014) A system to automatically classify and name any individual genome-sequenced organism independently of current biological classification and nomenclature. *PLoS One*, **9**, e89142.
17. Vinatzer,B.A., Weisberg,A.J., Monteil,C.L., Elmarakeby,H.A., Sheppard,S.K. and Heath,L.S. (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*, **107**, 18–28.
18. Vinatzer,B.A., Tian,L. and Heath,L.S. (2017) A proposal for a portal to make earth's microbial diversity easily accessible and searchable. *Antonie Van Leeuwenhoek*, **110**, 1271–1279.
19. Pierce,N.T., Irber,L., Reiter,T., Brooks,P. and Brown,C.T. (2019) Large-scale sequence comparisons with sourmash [version 1; peer review: 2 approved]. *F1000Res*, **8**, 1006–1006.
20. Jain,C., Rodriguez-R,L.M., Phillippy,A.M., Konstantinidis,K.T. and Aluru,S. (2018) High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat. Commun.*, **9**, 5114.
21. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
22. Rodriguez-R,L.M., Gunturu,S., Harvey,W.T., Rosselló-Mora,R., Tiedje,J.M., Cole,J.R. and Konstantinidis,K.T. (2018) The microbial genomes atlas (MiGA) webserver: taxonomic and gene diversity analysis of archaea and bacteria at the whole genome level. *Nucleic Acids Res.*, **46**, W282–W288.
23. Yoon,S.-H., Ha,S.-M., Kwon,S., Lim,J., Kim,Y., Seo,H. and Chun,J. (2017) Introducing EzBioCloud: a taxonomically united database of 16S rRNA gene sequences and whole-genome assemblies. *Int. J. Syst. Evol. Microbiol.*, **67**, 1613–1617.
24. Parks,D.H., Chuvochina,M., Waite,D.W., Rinke,C., Skarshewski,A., Chaumeil,P.-A. and Hugenholtz,P. (2018) A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat. Biotechnol.*, **36**, 996–1004.