LINgroups as a Principled Approach to Compare and Integrate Multiple Bacterial Taxonomies

Reza Mazloom
Virginia Tech
Blacksburg, Virginia, USA
rmazloom@vt.edu

Leighton Pritchard
University of Strathclyde
Glasgow, Scotland
leighton.pritchard@strath.ac.uk

C. Titus Brown
University of California, Davis
Davis, California, USA
ctbrown@ucdavis.edu

Boris A. Vinatzer
Virginia Tech
Blacksburg, Virginia, USA
vinatzer@vt.edu

Lenwood S. Heath
Virginia Tech
Blacksburg, Virginia, USA
heath@vt.edu

ABSTRACT

Traditional taxonomy provides a hierarchical organization of bacteria and archaea across taxonomic ranks from kingdom to subspecies. More recently, bacterial taxonomy has been more robustly quantified using comparisons of sequenced genomes, as in the Genome Taxonomy Database (GTDB), resolving down to genera and species. Such taxonomies have proven useful in many contexts, yet lack the flexibility and resolution of a more fine-grained approach. We apply our Life Identification Number (LIN) approach as a common, quantitative framework to tie existing (and future) bacterial taxonomies together, increase the resolution of genome-based discrimination of taxa, and extend taxonomic identification below the species level in a principled way. We utilize our existing concept of a LINgroup as an organizational concept for microorganisms that are closely related by overall genomic similarity, to help resolve some of the confusions and unforeseen negative effects of nomenclature changes of microbes due to genome-based reclassification. Our results obtained from experimentation demonstrate the value of LINs and LINgroups in mapping between taxonomies, translating between different nomenclatures, and integrating them into a single taxonomic framework.

CCS CONCEPTS

• **Applied computing** → **Bioinformatics**; **Computational genomics**.

KEYWORDS

Bacteria, Archaea, taxonomy, genomics, *k*-mers, average nucleotide identity, Jaccard similarity

ACM Reference Format:
Reza Mazloom, Leighton Pritchard, C. Titus Brown, Boris A. Vinatzer, and Lenwood S. Heath. 2022. LINgroups as a Principled Approach to Compare and Integrate Multiple Bacterial Taxonomies. In *13th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics (BCB '22), August 7–10, 2022, Northbrook, IL, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3535508.3545546

1 INTRODUCTION

Taxonomy is the science of classifying and organizing biological organisms into named units to facilitate their identification via some notion(s) of similarity. Traditional taxonomy utilizes a hierarchical organization into taxonomic ranks, where each node in the hierarchy is a taxon. The lowest formal ranks are species and subspecies, and the latinate genus-species nomenclature for plants and animals is well established and is employed even in common parlance. The species concept for prokaryotes is much more problematic [1, 4, 13, 18] and is the subject of debate primarily because of reticulate evolution or horizontal gene transfer [4, 7, 11, 16] and the challenge of defining principles for establishing prokaryotic species boundaries. Indeed, it is more appropriate to view the hierarchy of prokaryotic taxa as a network rather than a tree [3, 6]. A pragmatic approach to identify species boundaries through whole genome sequence similarity began with the introduction of Average Nucleotide Identity (ANI), where a threshold ANI of approximately 95% or greater is often taken to characterize the boundary of a single species [10].

The two most widely-used taxonomic schemes for bacteria are the NCBI taxonomy, and the Genome Taxonomy Database (GTDB). The NCBI taxonomy organizes prokaryotes into two trees (bacteria and archaea) using the taxonomic nomenclature developed over the years by traditional methods [19] and collected in the List of Prokaryotic Names with Standing in Nomenclature (LPSN; https://www.bacterio.net/). More recent efforts have sought to organize taxa through genomic sequence, primarily of cultivated species, although some fastidious or difficult to cultivate organisms have only been identified through metagenomics. GTDB organizes its taxonomic hierarchy by construction of a phylogenetic tree from 120 marker genes and the use of Relative Evolutionary Divergence (RED, a normalized measure of branch length) to establish taxonomic rank thresholds, followed by application of ANI to establish similarity between sequenced genomes and place additional organisms at species level, within that framework [2, 14]. For the future, we may expect revisions of existing taxonomies and creation of new taxonomies based on new principles. This raises the challenge of this paper, namely, to integrate our knowledge of genomic sequence across multiple taxonomies in a fashion most useful to the biologist.

BCB '22, August 7–10, 2022, Northbrook, IL, USA

ACM ISBN 978-1-4503-9386-7/22/08.
https://doi.org/10.1145/3535508.3545546

In addition, we address the challenge of extending genome-based taxonomy both between current taxonomic ranks and well below the species level.

Then we argue the need for biologists to have access to a broader conceptual framework that merges, to the extent possible, multiple taxonomies and that supports taxonomy below the species level.

Finally, we briefly describe characteristics of such a framework and point to the next section for our approach.

2 THE LIN CONCEPT

To properly explicate the Life Identification Number (LIN) concept, it is essential to first discuss the term "genome". In biology, the genome of an organism is the complete collection of genetic material in the cells of the organism. Modern DNA sequencing technology provides us with some access to details concerning this genetic material in the form of large numbers of either short or long DNA reads, though these reads alone need not exactly match the genetic material due to incomplete coverage, sequencing errors, contamination, and the challenge of piecing together the actual completed genome. Consequently, the set of reads, after processing by a suitable genome assembly program, will result in a set of contigs, which we optimistically call the assembled genome. These assembled genomes constitute the basis for the LIN concept and are what we mean henceforth by the term "genome".

The LIN concept is a general mechanism for organizing sequenced genomes according to a measure of similarity. By sequenced genomes, we mean a sequence file containing assembled contigs from a collection of DNA reads obtained from some sequencing technology. The better the quality of the reads and the contigs, the more successful the application of the LIN concept will be. A number of our prior publications have proposed LINs, applied them to existing sequenced genomes of particular classes of organisms, and demonstrated their broad utility [12, 20–26]. Here we describe the LIN concept in full mathematical generality for the first time demonstrating their robustness.

We are interested in organizing a dynamic universe U of sequenced genomes, denoted $G_1, G_2, \ldots$; our notion is that the universe utilized is fixed at any instant in time but, of course, expands as we add additional sequenced genomes. This organization requires one or more genome similarity measures. We use functional notation $s(G_i, G_j)$ for the similarity between genomes as measured by the similarity measure s. We require $0 \leq s(G_i, G_j) \leq 1$ and often speak of a similarity as a percentage. Examples of similarity measures include: average nucleotide identity (ANI) as computed by pyani [17] or approximated by FastANI [8]; Jaccard similarity as approximated by sourmash [15]; split k-mer analysis for SNP-level similarity as computed by SKA [5]; and Average Amino Acid Identity (AAI) as computed by EzAAI [9]. It is essential to be aware of the diversity of these and other similarity measures in terms of characteristics and resolution; there is no "one size fits all" measure. In particular, while it would be ideal to have one similarity measure that satisfactorily covered the entire interval $[0, 1]$ and was efficient to compute, this is not the case: in practice, it is necessary to employ multiple similarity measures to cover multiple subintervals and/or enhance computational efficiency.

For illustration, we start with just one similarity measure s, which we will employ to resolve similarity for some subinterval $[t_0, 1]$ of the entire interval $[0, 1]$. (A subinterval is chosen because high similarity values are generally more accurate than low ones.) In the LIN concept, we subdivide that interval into $m + 1$ nonoverlapping, exhaustive subintervals by selecting a sequence $0 = t_0, t_1, t_2, \ldots, t_m, t_{m+1} = 1$ of thresholds such that

$$t_0 < t_1 < t_2 < \cdots < t_m < t_{m+1} = 1;$$

the sequence $t_1, t_2, \ldots, t_m$ is a LIN scheme. The subintervals defined are then

$$[t_0, t_1], (t_1, t_2], (t_2, t_3], \ldots, (t_m, t_{m+1}]$$

and are called percentage (subintervals). Consequently, $s(G_i, G_j)$ falls into exactly one percentage. A Life Identification Number (LIN) is an $(m + 1)$-tuple $\ell = (n_0, n_1, \ldots, n_m)$ of non-negative integers, while the p LIN prefix of ℓ is $\ell^p = (n_0, n_1, \ldots, n_p)$, where $0 \leq p \leq m$. We call the location of n_i in ℓ position i, or we simply call n_i position i. The goals of the LIN concept are twofold. First, each unique genome sequence is assigned a unique LIN; in particular, we assign ℓ_i to genome G_i. As there are an infinite number of LINs, this goal can certainly be met. Second, and more importantly, the LINs are chosen in such a manner that the LINs ℓ_i and ℓ_j of G_i and G_j provide evidence of the similarity of G_i and G_j according to similarity measure s. More specifically, LINs identify position $p + 1$ as the leftmost position of ℓ_i and ℓ_j where they differ. This implies that $\ell_i^p = \ell_j^p$; they share their p LIN prefix. Then, the goal is that $s(G_i, G_j)$ occurs in the interval $(t_p, t_{p+1}]$.

To achieve the two goals of the LIN concept, we provide simplified pseudocode for the core LIN algorithm in Figure 1.

```
(1) Assign genome $G_1$ the LIN $(1, 0, 0, \ldots, 0)$.
(2) For each subsequent genome $G_{i+1}$, find the genome $G_j$, $1 \leq j \leq i$, which minimizes $s(G_i, G_j)$.
(3) Find the leftmost position where they differ and make sure the LIN assigned to $G_{i+1}$ differs from all other LINs and differs from $\ell_j$ at that position first.
(4) Continue with the next genome (Step 2) until all genomes have been assigned LINs.
```

Figure 1: Pseudocode for our original, naive implementation of a prototype of the LIN concept. Note that it is straightforward to make this an online algorithm that accepts new genomes as they are sequenced and assembled.

Depending on the characteristics of the similarity measure, the core algorithm is not guaranteed to achieve the second goal in all cases. However, we have successfully employed a 20-position scheme in our LINbase database [20] using ANI as the similarity measure and have demonstrated that the second goal is indeed achieved in practice [12, 23, 25, 26]. The original naive implementation of the core algorithm is quite slow, given the time complexity of algorithms for computing ANI, but we have sped up the algorithm immensely by employing sourmash [15], a tool for rapid searching of a database of sequences using sets of k-mer based signatures, called sketches, in our LINflow implementation [21].

Once some portion of the universe U of genomes has been assigned LINs, all the genomes that share a LIN prefix are called a LINgroup. A LINgroup with a percentage of about 95% often contains exactly the genomes from a recognized prokaryotic species. The power of LINgroups, however, goes far beyond describing species, as the percentage can be any one present in the current scheme. Moreover, as additional genomes are added, they will automatically fall into the proper LINgroups.

We now identify a number of characteristics and advantages of the LIN concept and of LINgroups. As demonstrated in the formal description above, the LIN concept is highly flexible in several ways: the similarity measure utilized can be selected to achieve the desired ends of a particular implementation and to be as efficient as possible; new sequenced genomes can be incorporated into a LIN database in a natural, online fashion; and the implied taxonomic ranks of the LIN scheme are not fixed by the LIN concept, in contrast to existing taxonomies. The resolution of LINs can extend well below that of the traditional species rank, making fine distinctions a natural implication of using LINs. LIN assignment can be quite rapid computationally, leading to near instantaneous feedback to the group that provided the sequenced genome in the first place. Multiple similarity measures can be employed to span a large range of percentage similarities. One or more LINgroups together can represent a known taxonomic rank, especially one that was not originally characterized by genomic sequence similarities; we will informally call such a collection of LINgroups a *cluster*. The LIN concept organizes genomes into a hierarchy, much as traditional taxonomy employs hierarchy as its organizational principle. Finally, any given LIN scheme and associated database is stable in the sense that no recomputation is required in the event of the acquisition of additional data. In one of our prior works [21] we have shown implementation details, computational speed and memory usage hence, now we focus on other aspects of the approach.

Because LINs have many more thresholds of similarity (20 in our current implementation) than the two taxonomic ranks genus and species, LINgroups can be used to precisely circumscribe (or define or delineate) genera and species of different breadth from different taxonomies (such as GTDB and NCBI) and thus allow to integrate and compare taxonomies with each other. Our current intervals of similarity (LIN scheme) described in Section 3.2 cover genus and species at the lower and middle thresholds. For example, 95% is a useful standard ANI for determining species. But, as the thresholds extend to 99.999%, the LIN scheme allows for far higher taxonomic resolution than either the NCBI or GTDB hierarchies. Moreover, a taxon in either of the other hierarchies corresponds to one or more LINgroups in this LIN scheme. Hence, the LIN concept successfully spans and connects multiple taxonomies in a neutral setting depending only on the available sequenced genomes. Moreover, with additional similarity resolution, higher taxonomic ranks can be incorporated as well.

3 MATERIALS AND METHODS

3.1 Data Sets

We have selected a set of 1207 genomes close to the Agrobacterium genus with their corresponding NCBI and GTDB (version R07-RS207) taxonomic lineages as our data set. Using the computed LINs as described in Section 2, we can perform a three-way comparison among the NCBI taxonomy, GTDB, and the computed LINs. Figure 2 shows an overview of the number of distinct lineages (taxonomic clusters) considering each method. All genomes in the data set have identical taxonomic ranks from kingdom to order and are mainly positioned within two and three families in NCBI and GTDB respectively. Hence, we have combined results for the ranks kingdom through order, favoring result simplicity. Consequently, we focus on taxonomies at the family rank and below, which provide more variation allowing us to better compare the three taxonomic methods. When comparing taxonomies at each rank, we are only able to compare them when the ranks are defined and ordered. Below are the major taxonomic ranks that we considered, when available, in order of consideration: (1) kingdom, (2) phylum, (3) class, (4) order, (5) family, (6) genus, (7) species, and (8) subspecies.

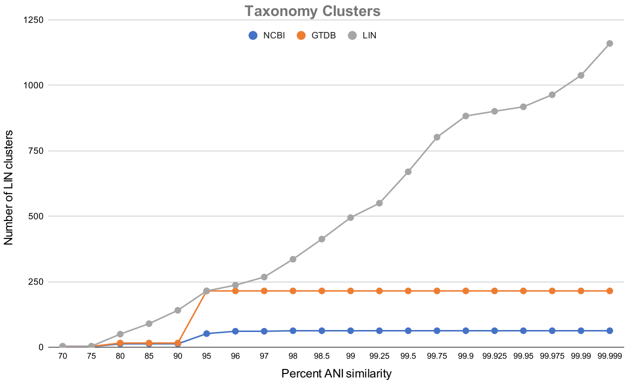

Figure 2: The number of potential taxonomic ranks currently segmenting the Agrobacterium data set. Family, Genus, Species, and Phenotype ranks, when available, were assigned to 70, 80, 95, and 96% ANI similarity respectively.

3.2 Experiments

Figure 3 provides a complete overview of our LIN assignment process. Our LINflow [21] implementation utilizes a combination of Jaccard similarity computations using sourmash and ANI comparisons using pyani to reduce the number of comparisons done by ANI, which in turn reduces the overall runtime. First, with the assumption of genomes of similar species having about 95% ANI, we create an initial measurement layer to identify species representatives using Jaccard (Species separator scheme). At this step, we decide whether to create a new species cluster (case A in Figure 3), when Jaccard similarity is low to existing species representative genomes, or we find the closest species cluster to our genome. Next, we compare genomes within the species cluster using Jaccard and choose the top three genomes as references to compare the new genome against using ANI. The genome with the top ANI similarity and at least 20% genome coverage is considered the closest genome. In certain cases, Jaccard might be computed a second time with a higher stringency (case B in Figure 3) if genomes are highly similar.

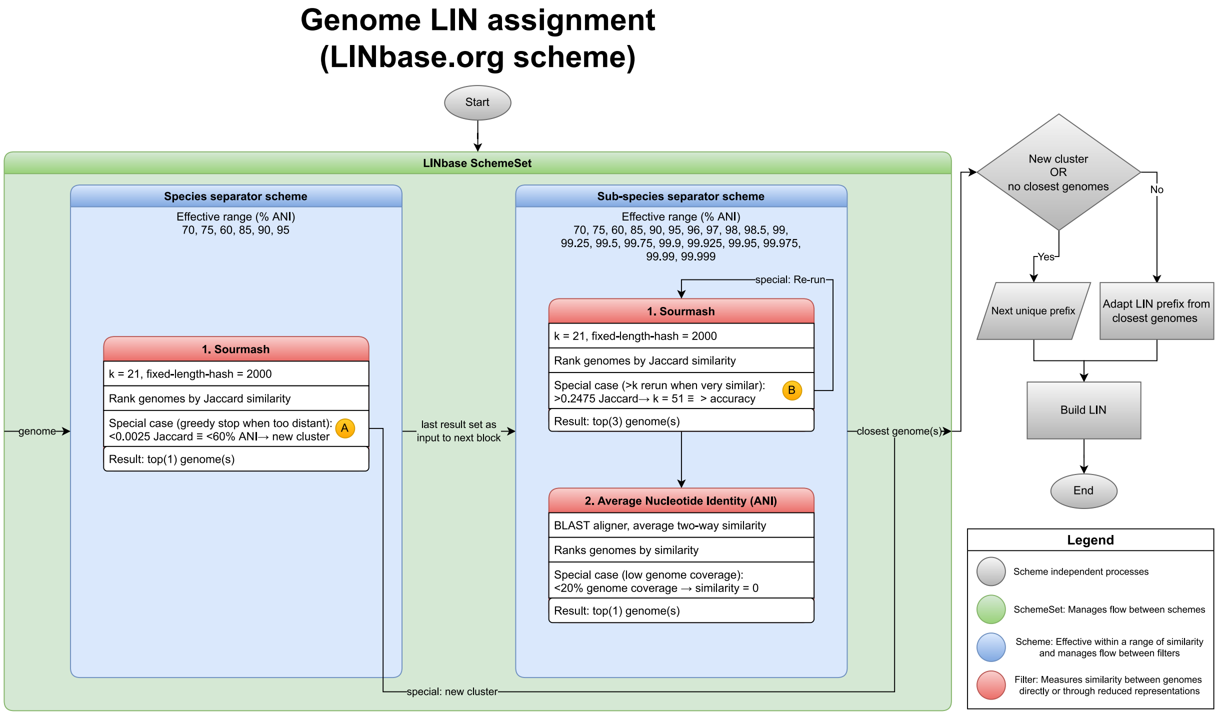
Genome LIN assignment
(LINbase.org scheme)
Start
LINbase SchemeSet
Species separator scheme
Effective range (% ANI)
70, 75, 60, 85, 90, 95
1. Sourmash
k = 21, fixed-length-hash = 2000
Rank genomes by Jaccard similarity
Special case (greedy stop when too distant):
<0.0025 Jaccard ≡ <60% ANI→ new cluster
A
Result: top(1) genome(s)
genome
last result set as
input to next block
Sub-species separator scheme
Effective range (% ANI)
70, 75, 60, 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
special: Re-run
1. Sourmash
k = 21, fixed-length-hash = 2000
Rank genomes by Jaccard similarity
Special case (>k rerun when very similar):
>0.2475 Jaccard→ k = 51 ≡ > accuracy
B
Result: top(3) genome(s)
2. Average Nucleotide Identity (ANI)
BLAST aligner, average two-way similarity
Ranks genomes by similarity
Special case (low genome coverage):
<20% genome coverage → similarity = 0
Result: top(1) genome(s)
closest genome(s)
special: new cluster
New cluster
OR
no closest genomes
Yes
No
Next unique prefix
Adapt LIN prefix from
closest genomes
Build LIN
End
Legend
Scheme independent processes
SchemeSet: Manages flow between schemes
Scheme: Effective within a range of similarity
and manages flow between filters
Filter: Measures similarity between genomes
directly or through reduced representations

Figure 3: Process of assigning LIN to a single genome sequence using the current LINbase.org LIN scheme. All scheme and measure related criteria can be modified at will to best fit one's needs.

Finally, given the closest genome, or the lack thereof where similarity is set to zero, we use the algorithm in Figure 1 to assign a LIN to the genome. We are using a set of threshold similarities (LIN scheme) with 20 percentage ANI thresholds creating clusters at 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, and 99.999. The LIN of the closest genome is assigned as a prefix of the new LIN up to the highest scheme threshold, given the similarity between the genomes is less than or equal to that threshold. For example if the similarity is 96.1% the two genomes will share identical LINs up to the 96% threshold and differ at the 97% threshold.

After the LIN assignment process is complete, we look into the correspondence between the LINs and their corresponding lineages within different methods. Based on our prior work [12, 20–26], we are aware that different lineages at the same taxonomic rank do not necessarily diverge at the same ANI similarity threshold. Therefore, we consider all unique taxonomic lineages at each rank (taxonomic cluster) to all unique LIN clusters at each threshold. For example, we analyze the uniqueness of the species rank both when considering the first LIN threshold and all 20 LIN thresholds. Figure 4 illustrates the full process of comparing all the available taxonomy pairings and the eventual results that manifest from the analysis.

3.3 Evaluation

Evaluating the LIN-based taxonomy can be done by observing how taxonomic lineages such as ones defined by NCBI and GTDB can also be defined by LINs reliably. Simply put, we can measure how the clusters made by LINs conform with the lineages. One simple method to measure conformance is to compute how unique the LINs are when compared to one unique lineage. This is done for every combination of lineage and every LIN length. For instance, if we consider all the genomes with the lineage Agrobacterium at the genus rank, for a fixed LIN length, we can place the LINs in three groups.

The first group consists of *Unique LINs*. These LIN(s) are unique to the lineage, at the genus rank and do not identify any genomes with a different lineage. Furthermore, if all LINs are unique to the lineage, we can rely on the LIN(s) to reliably identify/separate this lineage within GTDB from the others.

The second group consists of *Dominant LINs*. These LIN(s) are not unique to the lineage however, 90% of the genomes within each LIN are members of the lineage. This allows for a more relaxed identification of lineages while keeping taxonomy conformance above a threshold. Depending on how strict this conformity needs to be the threshold can be increased where 100% is the same as unique LINs. Higher thresholds mainly affect taxonomy conformance on shorter LIN lengths since they are shared by more genomes. On longer length LINs though the chance of conformity tends to zero the higher the threshold and the lower the number of genomes sharing that LIN.

The third group consists of *Non-unique LINs*. Any LIN(s) not unique (including dominants) are part of this group. Any lineages corresponding to LINs within these groups cannot be reliably identified/separated from the other lineages within this group. The

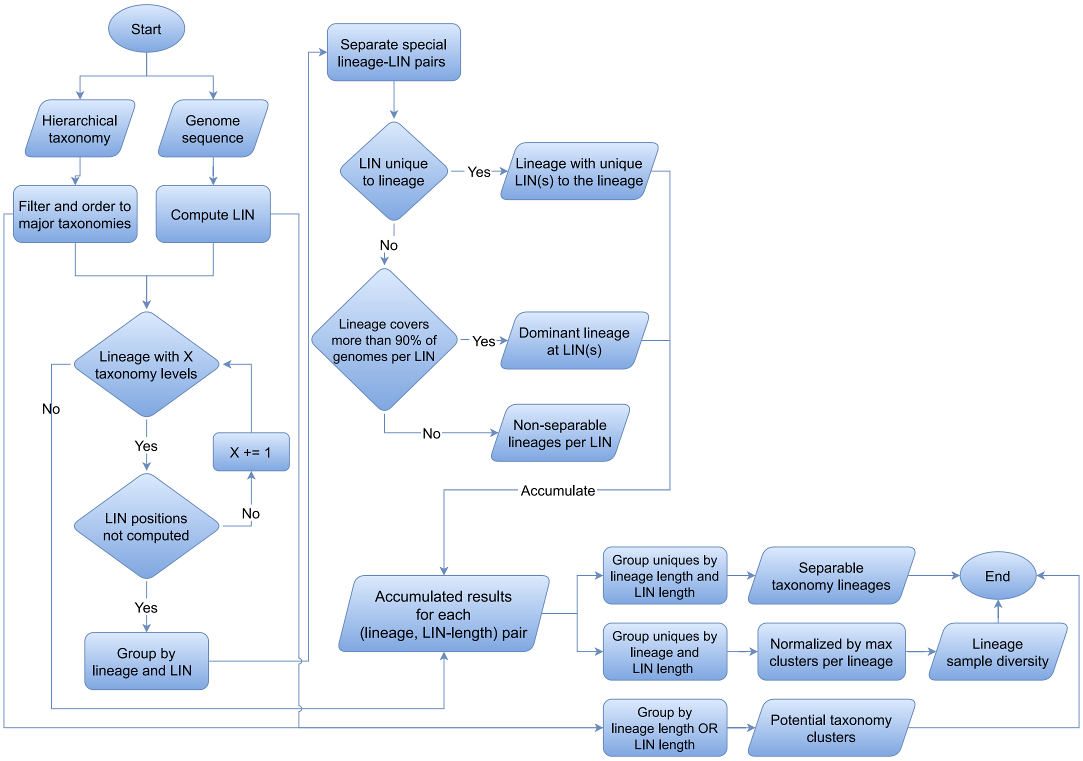
Start
Hierarchical taxonomy
Genome sequence
Filter and order to major taxonomies
Compute LIN
Lineage with X taxonomy levels
No
Yes
X += 1
LIN positions not computed
No
Yes
Group by lineage and LIN
Separate special lineage-LIN pairs
LIN unique to lineage
Yes
Lineage with unique LIN(s) to the lineage
No
Lineage covers more than 90% of genomes per LIN
Yes
Dominant lineage at LIN(s)
No
Non-separable lineages per LIN
Accumulate
Accumulated results for each (lineage, LIN-length) pair
Group uniques by lineage length and LIN length
Separable taxonomy lineages
End
Group uniques by lineage and LIN length
Normalized by max clusters per lineage
Lineage sample diversity
Group by lineage length OR LIN length
Potential taxonomy clusters

Figure 4: Process of pairing lineage and LIN(s) and how grouping (aggregating) by different parameters results in different views of relation between LIN clusters and the taxonomic lineage. The last three nodes before the end are the three figures included in Section 4.

number of genomes within this group can be decreased by increasing the length of the LIN, decreasing the dominance threshold, or considering a higher-level lineage.

Now given the unique LINs only or their union with dominant LINs, we can analyze how well LINs can identify lineage clusters already defined by the reference taxonomies and whether the lineage clusters can potentially be further broken down into smaller clusters when considering the LIN clusters. In Section 4, we mainly focus on unique LINs when analyzing our results so as to present the most conservative results when considering lineage conformance.

4 RESULTS

Looking at the number of clusters at each taxonomic rank that has varying values in our data set, NCBI having four such taxonomic ranks and GTDB having three, we can look at the available clusters. Figure 2 illustrates the number of clusters at each rank. Although the taxonomic ranks family and genus do not correspond to set ANI thresholds, in Figure 5 we equated family with 70% ANI and genus with 75% ANI to best fit the number of LIN clusters to their corresponding rank. We used the established 95% ANI threshold for species and 96% for phenotype (when this rank was used by NCBI). We can easily see the number of LIN clusters and how the two taxonomic methods diverge at the 95% ANI (species) rank while, interestingly, the number of LIN and GTDB clusters match perfectly. Another interesting trend is the roughly linear increase in LIN clusters despite the nonlinear ANI ranks. Normally, with a linear ANI scheme one would see an exponential increase in clusters at the latter ranks, assigning many LIN clusters to subspecies ranks. This continuous increase of LIN clusters is an indication that the selected scheme captures the range of genomic distances in our data set. Simply put, we can define subspecies ranks simply through a few short (taxonomically speaking) LIN clusters, rather than many long LIN clusters.

Just focusing on cluster counts, however, does not illustrate how well the LIN clusters match current taxonomic ranks. Figure 5 illustrates how well genomes can be grouped into their unique taxonomic cluster at different taxonomic ranks with only a limited number of thresholds used in the LIN scheme. It is expected that higher-level ranks will be easier to identify than lower ranks, which holds mostly true, with the exception of NCBI's genus and species ranks at scheme ranks 6-10 (95-98.5). Furthermore, identification of GTDB ranks is more easily achieved through fewer LIN ranks compared to NCBI. This indicates GTDB ranks are more aligned to sequence similarities, which is expected based on how GTDB species clusters are defined [14]. Additionally the LIN ranks 2, 5, and 9 improve identification significantly for genus and species in both taxonomies; hence, they are meaningful ranks for this set of genomes.

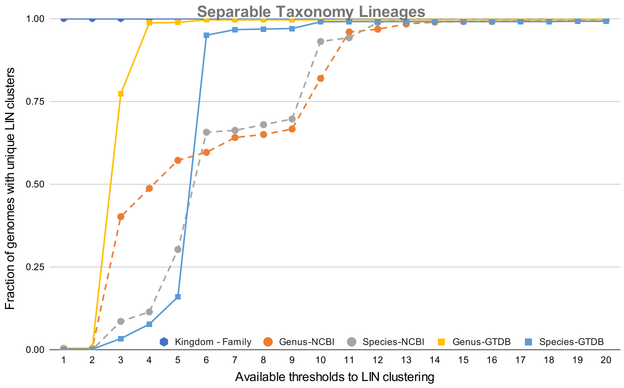

Figure 5: Fraction of genomes that can be uniquely identified using their LIN cluster(s) given how many thresholds can be used for clustering. For both NCBI and GTDB, taxonomy levels Kingdom through Family can all be uniquely identified since our genomes all are part of one Order. The GTDB taxonomic ranks on the Genus and Species levels are more easily separable than the NCBI ranks using the ANI-based LIN clusters.

Breaking down the ranks and comparing the lineages at each rank, we can analyze the LIN clustering behavior. We have specifically selected genera that have more than 10 genome sequences associated with them to allow flexibility in cluster formations in Figure 6. Note that not all genera at the final rank (20) have LIN clusters equivalent to the number of genome sequences ($y < 1$). This means that some genome sequences share their full LIN with others. This indicates that, within these genera, there exist sequences that have ANI similarities above the maximum threshold (99.999%). This could be easily mitigated by increasing the maximum threshold of the LIN scheme.

Furthermore, the different rates of cluster formation between genera shows more insight into each lineage. The slope of the lines at certain LIN thresholds is a good indicator of internal clusters, specific to the lineage. When studying that lineage exclusively, we need to focus on these threshold ranges to further see the internal clustering of the genomes. We could similarly decrease focus on ranges with low or zero slope. This can be seen in Sinorhizobium, where it would be best to increase focus on thresholds 12-13 since the number of clusters almost doubles from 56 to 109 having a total of 231 genomes assigned to the lineage.

Given the same information, we also hypothesize that lineages that have fewer clusters at the first few thresholds than others could be lacking sample diversity. Simply put, highly similar (less diverse) genomes lead to many clusters on the latter LIN ranks, while less similar (highly diverse) genomes do the opposite. Hence, the lack of diversity is apparent when comparing the genus Esnifer with Neorhizobium or Pararhizobium despite their roughly equivalent genome count. Genera with high genome counts in the data set, namely Rhizobium, Agrobacterium, and Sinorhizobium, also show a lack of sample diversity, which stems from their low diversity with respect to genome count.

5 DISCUSSION

Taxonomy has been an indispensable tool for biologists to structure and simplify the abundance of variation among organisms. However, with the emergence of inexpensive DNA sequencing and the tremendous increase in genome sequence availability, the traditional methods of taxonomic identification are benefitting from data-driven methods. In this paper, we have proposed our concept of Life Identification Numbers (LINs) to integrate and compare multiple taxonomies based on genome similarity. LINs provide a powerful system to structure the similarity relationships among genomes and can be flexibly implemented utilizing any set of similarity measures. For example, a LIN implementation may use the most precise measures (for example, a single nucleotide polymorphism (SNP) based similarity measure) for contexts where the genomes are known to be highly similar, since such measures are more sensitive to single mutations occurring between almost identical genomes. Even though these measures are sometimes computationally expensive, they are needed for only a fraction of the comparisons, which has little effect on the overall runtime.

Basing a LIN implementation on ANI, we created clusters that were not only able to uniquely identify the current taxonomic lineages up to the species rank, but also identify clusters at the subspecies level. Using the variation of these clusters, we also compared the sample diversity among genera and identified thresholds of importance for studying their subspecies groupings.

LINs provide a conceptually simple approach to organize sequenced genomes into a hierarchy based strictly on sequence similarity. The resulting taxonomy can serve as a starting point for automated placement of new genomes as well as a common coordinate system for conversion between other taxonomies. The increased granularity offered by LINs over the more traditional seven-point system of kingdom and phylum through species also provides opportunities to more finely resolve relationships as new genome sequences are introduced. Finally, the availability of well-defined ranks beyond those of species will be a critically important tool for the rapid analysis and characterization of many closely related genomes during a pathogenic outbreak. The overall framework provided by the LIN concept is further enhanced by a relatively simple set of algorithmic tools for building LINs.

ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Grant No. DBI-2018522 to BAV, CTB, and LSH. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

[1] Timothy G Barraclough, Kevin J Balbi, and Richard J Ellis. 2012. Evolving concepts of bacterial species. *Evol. Biol.* 39, 2 (2012), 148–157.

[2] Pierre-Alain Chaumeil, Aaron J Mussig, Philip Hugenholtz, and Donovan H Parks. 2020. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics* 36 (Nov. 2020), 3 pages.

[3] Rob DeSalle and Margaret Riley. 2020. Should Networks Supplant Tree Building? *Microorganisms* 8, 8 (Aug. 2020), 14 pages.

[4] J Peter Gogarten, W Ford Doolittle, and Jeffrey G Lawrence. 2002. Prokaryotic evolution in light of gene transfer. *Mol. Biol. Evol.* 19, 12 (Dec. 2002), 2226–2238.

[5] Simon R Harris. 2018. SKA: Split Kmer Analysis Toolkit for Bacterial Genomic Epidemiology. (Oct. 2018), 453142 pages.

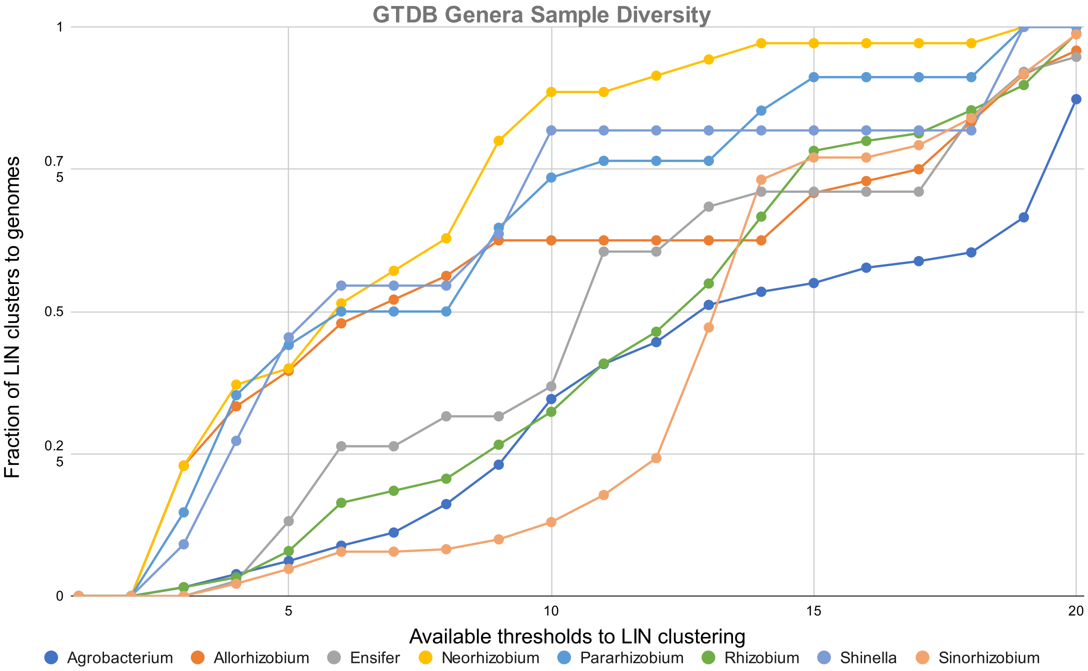

Figure 6: This figure compares all genera in our Agrobacterium data set that have more than ten genomes. The fraction of potential clusters can be taken as a good representative of genome diversity, since higher number of clusters means there exists more separable genomes based on similarity measures, and in extension more genome variation.

[6] Antonio Hernández-López, Olivier Chabrol, Manuela Royer-Carenzi, Vicky Merhej, Pierre Pontarotti, and Didier Raoult. 2013. To tree or not to tree? Genome-wide quantification of recombination and reticulate evolution during the diversification of strict intracellular bacteria. *Genome Biol. Evol.* 5, 12 (2013), 2305–2317.

[7] Jaime Iranzo, Yuri I Wolf, Eugene V Koonin, and Itamar Sela. 2019. Gene gain and loss push prokaryotes beyond the homologous recombination barrier and accelerate genome sequence divergence. *Nat. Commun.* 10, 1 (Nov. 2019), 5376.

[8] Chirag Jain, Luis M Rodriguez-R, Adam M Phillippy, Konstantinos T Konstantinidis, and Srinivas Aluru. 2018. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat. Commun.* 9, 1 (Nov. 2018), 5114.

[9] Dongwook Kim, Sein Park, and Jongsik Chun. 2021. Introducing EzAAI: a pipeline for high throughput calculations of prokaryotic average amino acid identity. *J. Microbiol.* 59, 5 (May 2021), 476–480.

[10] Konstantinos T Konstantinidis and James M Tiedje. 2005. Genomic insights that advance the species definition for prokaryotes. *Proc. Natl. Acad. Sci. U. S. A.* 102, 7 (Feb. 2005), 2567–2572.

[11] James Mallet, Nora Besansky, and Matthew W Hahn. 2016. How reticulated are species? *Bioessays* 38, 2 (Feb. 2016), 140–149.

[12] Haitham Marakeby, Eman Badr, Hanaa Torkey, Yuhyun Song, Scotland Leman, Caroline L Monteil, Lenwood S Heath, and Boris A Vinatzer. 2014. A System to Automatically Classify and Name Any Individual Genome-Sequenced Organism Independently of Current Biological Classification and Nomenclature. *PLoS One* 9, 2 (2014), 12 pages.

[13] Marike Palmer, Stephanus N Venter, Martin P A Coetzee, and Emma T Steenkamp. 2019. Prokaryotic species are sui generis evolutionary units. *Syst. Appl. Microbiol.* 42, 2 (March 2019), 145–158.

[14] Donovan H Parks, Maria Chuvochina, Christian Rinke, Aaron J Mussig, Pierre-Alain Chaumeil, and Philip Hugenholtz. 2022. GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic Acids Res.* 50, D1 (Jan. 2022), D785–D794.

[15] N Tessa Pierce, Luiz Irber, Taylor Reiter, Phillip Brooks, and C Titus Brown. 2019. Large-scale sequence comparisons with sourmash. *F1000Res.* 8 (July 2019), 1006.

[16] Martin F Polz, Eric J Alm, and William P Hanage. 2013. Horizontal gene transfer and the evolution of bacterial and archaeal population structure. *Trends Genet.* 29, 3 (March 2013), 170–175.

[17] Leighton Pritchard, Rachel H Glover, Sonia Humphris, John G Elphinstone, and Ian K Toth. 2016. Genomics and taxonomy in diagnostics for food security: soft-rotting enterobacterial plant pathogens. *Anal. Methods* 8, 1 (2016), 12–24.

[18] Ramon Rosselló-Móra and Rudolf Amann. 2015. Past and future species definitions for Bacteria and Archaea. *Syst. Appl. Microbiol.* 38, 4 (June 2015), 209–216.

[19] Conrad L Schoch, Stacy Ciufo, Mikhail Domrachev, Carol L Hotton, Sivakumar Kannan, Rogneda Khovanskaya, Detlef Leipe, Richard Mcveigh, Kathleen O'Neill, Barbara Robbertse, Shobha Sharma, Vladimir Soussov, John P Sullivan, Lu Sun, Seán Turner, and Ilene Karsch-Mizrachi. 2020. NCBI Taxonomy: a comprehensive update on curation, resources and tools. *Database* 2020 (Jan. 2020), 21 pages.

[20] Long Tian, Chengjie Huang, Reza Mazloom, Lenwood S Heath, and Boris A Vinatzer. 2020. LINbase: a web server for genome-based identification of prokaryotes as members of crowdsourced taxa. *Nucleic Acids Res.* 48, W1 (July 2020), W529–W537.

[21] Long Tian, Reza Mazloom, Lenwood S Heath, and Boris A Vinatzer. 2021. LINflow: a computational pipeline that combines an alignment-free with an alignment-based method to accelerate generation of similarity matrices for prokaryotic genomes. *PeerJ* 9 (March 2021), e10906.

[22] Long Tian, Y Vasebi, V Eastman, K Hirani, G Hughes, Lenwood S Heath, and Boris A Vinatzer. 2018. Genome-Based Classification and Identification of Bacteria. *Phytopathology* 108, 10 (2018), 1 pages.

[23] Boris A Vinatzer, H A Elmarakeby, A J Weisberg, C L Monteil, and Lenwood S Heath. 2015. A New Exclusively Genome-Based Species-Independent Taxonomic Framework for All Life Forms Applied to Pseudomonas syringae. *Phytopathology* 105, 11 (2015), 143.

[24] Boris A Vinatzer, L Tian, and Lenwood S Heath. 2017. A Proposal for a Portal to Make Earth'S Microbial Diversity Easily Accessible and Searchable. *Antonie Van Leeuwenhoek International Journal of General and Molecular Microbiology* 110, 10 (2017), 1271–1279.

[25] Boris A Vinatzer, Alexandra J Weisberg, Caroline L Monteil, Haitham A Elmarakeby, Samuel K Sheppard, and Lenwood S Heath. 2017. 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, 1 (2017), 18–28.

[26] Alexandra J Weisberg, Haitham A Elmarakeby, Lenwood S Heath, and Boris A Vinatzer. 2015. Similarity-Based Codes Sequentially Assigned to Ebolavirus Genomes Are Informative of Species Membership, Associated Outbreaks, and Transmission Chains. *Open Forum Infectious Diseases* 2, 1 (2015), 11 pages.