Brazilian Journal of Biology®

ISSN 1519-6984 (Print)
ISSN 1678-4375 (Online)

Original Article

# Sketched reference databases for genome-based taxonomy and comparative genomics

## Bases de dados de referência esboçados para taxonomia baseada em genoma e genômica comparativa

A. Sánchez-Reyes[a]* and M. G. Fernández-López[b]

[a]Consejo Nacional de Ciencia y Tecnología – CONACyT, Universidad Nacional Autónoma de México – UNAM, Instituto de Biotecnología, Cuernavaca, Morelos, México

[b]Universidad Autónoma del Estado de Morelos – UAEM, Centro de Investigación en Dinámica Celular, Instituto de Investigaciones Básicas y Aplicadas, Cuernavaca, Morelos, México

**Abstract**

The analysis of curated genomic, metagenomic and proteomic data is of paramount importance in the fields of biology, medicine, education, and bioinformatics. Although this type of data is usually hosted in raw format on free international repositories, the full access requires lots of computing power and large storage disk space for the domestic user. The purpose of the study is to offer a comprehensive set of microbial genomic and proteomic reference databases in an accessible and easy-to-use form to the scientific community and demonstrate its advantages and usefulness. Also, we present a case study on the applicability of the sketched data, for the determination of overall genomic coherence between two members of the *Brucellacea* family, which suggests they belong to the same genomospecies that remain as discrete ecotypes. A representative set of genomes, proteomes (from type material), and metagenomes were directly collected from the NCBI Assembly database and Genome Taxonomy Database (GTDB), associated with the major groups of Bacteria, Archaea, Virus, and Fungi. Sketched databases were subsequently created and stored on handy reduced representations by using the MinHash algorithm implemented in Mash software. The obtained dataset contains more than 133 GB of space disk reduced to 883.25 MB and represents 125,110 genomics/proteomic records from eight informative contexts, which have been prefiltered to make them accessible, usable, and user-friendly with limited computational resources. Potential uses of these sketched databases are discussed, including but not limited to microbial species delimitation, estimation of genomic distances and genomic novelties, paired comparisons between proteomes, genomes, and metagenomes; phylogenetic neighbor's exploration and selection, among others.

**Keywords:** microbial Mash database, genomic distance, genome containment, type material, microbial taxonomy.

**Resumo**

A análise de dados genômicos, metagenômicos e proteômicos com curadoria é de suma importância nos campos da biologia, medicina, educação e bioinformática. Embora esse tipo de dados geralmente seja hospedado em formato bruto em repositórios internacionais gratuitos, o acesso total requer muita capacidade de computação e grande espaço em disco de armazenamento para o usuário doméstico. Os objetivos do estudo são oferecer um conjunto abrangente de bancos de dados de referência genômica e proteômica microbiana de forma acessível e fácil de usar para a comunidade científica e demonstrar suas vantagens e utilidade. Além disso, apresentamos um estudo de caso sobre a aplicabilidade dos dados esboçados para a determinação da coerência genômica geral entre dois membros da família Brucellacea, o que sugere que eles pertencem às mesmas genomoespécies que permanecem como ecótipos discretos. Um conjunto representativo de genomas, proteomas (de material tipo) e metagenomas foi coletado diretamente do banco de dados NCBI Assembly e do banco de dados de taxonomia do genoma (GTDB), associada aos principais grupos de bactérias, Archaea, vírus e fungos. Bancos de dados esboçados foram subsequentemente criados e armazenados em representações reduzidas práticas usando o algoritmo MinHash implementado no software Mash. O conjunto de dados obtido contém mais de 133 GB de espaço em disco reduzido para 883,25 MB e representa 125,110 registros genômicos/proteômicos de oito contextos informativos, que foram pré-filtrados para torná-los acessíveis, utilizáveis e amigáveis com recursos computacionais limitados. Os usos potenciais desses bancos de dados esboçados são discutidos, incluindo, mas não se limitando, a delimitação de espécies microbianas, estimativa de distâncias genômicas e novidades genômicas, comparações emparelhadas entre proteomas, genomas e metagenomas, exploração e seleção filogenética de vizinhos, entre outros.

**Palavras-chave:** banco de dados de Mash microbiano, distância genômica, contenção de genoma, material de tipo, taxonomia microbiana.

*e-mail: ayixon.sanchez@mail.ibt.unam.mx

Received: September 23, 2021 – Accepted: June 15, 2022

## 1. Introduction

The analysis of the growing genomic universe derived from massive sequencing is within the limits of binary computation and represents a major challenge for current biology and computer sciences. Microbial taxonomy has been one of the most favored by genomic effervescence, through the development of accurate and quasi-universal circumscription concepts and methods for the prokaryotic domain. Among them, overall genome relatedness indexes (OGRI) and genome-to-genome sequence comparison methods constitute standard procedures in genome-based taxonomy and comparative genomics (Chun and Rainey, 2014). These methods depend on reference databases that contain numerous records classified as genomes, proteomes, transcriptomes, and metagenomes, among others. Currently, the National Center for Biotechnology Information (NCBI) contains more than 900 thousand bacterial assemblies, more than eight thousand Archaean genomes, and near to 800 fungal genomes (NCBI Resource Coordinators, 2018). Storing such amount of information is difficult to reach for the common and personal users due to storage, broadband internet access, or computing limitations. Other representative databases such as the NCBI type material for prokaryotes (Federhen, 2015) use ~20 GB of space in compressed format, and the Genome Taxonomy Database (GTDB) requires near to 46 GB of compressed space (Parks et al., 2018, 2020). To facilitate access to this type of information for taxonomists, microbiologists, biotechnologists, or any other domestic user, we have collected an up-to-date representation of several genomic, proteomic, and metagenomic databases, useful for systematic classification of new genomes -cultivable or not- and the circumscription of species-specific contexts, the comparison by similarity of genomic objects, as well as a representation of more than 1000 soil and freshwater metagenomes (close to 50 GB) for clustering and comparisons of metagenomics dataset. Also, 47,894 representative species clusters genomes from GTDB -release 202- (GTDB r202) were obtained as an external and updated taxonomic resource. With this raw information, representative datasets were built using the powerful MinHashing dimensional reduction technique over genomics or proteomics data (Ondov et al., 2016). In all cases, the *mash sketch* function included in the Mash program (Ondov et al., 2016) was used to obtain reduced representations of the DNA or amino acid alphabet. These datasets contain only a small fraction of the original size of the raw data; for example, the sketched GTDB r202 just contains 377 MB in size while conserving the signatures for the totality of the original genomic assemblies. The potential uses of this dataset include but are not limited to, estimation of genomic distances, genomic novelties, and paired comparisons between proteomes, genomes, and metagenomes. Each dataset is a curated database that can be used essentially to assess taxonomic categories and overall genomic resemblances. The sketched **.msh* files from this work can be quickly accessed from figshare.com (Sánchez-Reyes and Fernández-López, 2021a) and the analysis could be executed on a modest desktop computer with as little as 4GB of ram and 1 GB of free space. An early version of this work has been deposited as a preprint on Preprints.org site (Sánchez-Reyes and Fernández-López, 2021b).

## 2. Material and Methods

### *2.1. Acquisition of raw genomic, proteomic and metagenomic data*

Genomic data acquisition is a key goal to infer accurate relationships among the members of the universal tree of life. As genomes remain our most comprehensive signature, their importance in reconstructing the evolutionary history through comparative genomics is increasingly recognized. In particular, the genome-based taxonomy is an up-to-date criterion to delineate microbial genomospecies by overall genome relatedness index comparison. Genome assemblies, proteomic and metagenomic data were directly downloaded from the NCBI's Assembly resource site (Kitts et al., 2016) between April and September 2021 (except for the "soil metagenome set" that was acquired in September 2020). The search details were as follow: The prokaryotic genomic fraction with relation to type material on NCBI was extracted from (NCBI, 2021a), with the following search details: ("Bacteria"[Organism] OR "Archaea"[Organism]) AND ("latest genbank"[filter] AND (all[filter] NOT partial[filter] AND all[filter] NOT anomalous[filter]) AND "from type"[Properties]). We used "NOT partial" and "NOT anomalous" filters to avoid low coverage, contaminated or chimeric sequences. Prokaryotic protein Fasta files were downloaded for all assemblies with annotation status as: "has annotation".

The genomes from Genome Taxonomy Database release 202, (April-2021) were obtained remotely from the GTDB public repository (GTDB, 2021a) with the command *wget* (GTDB, 2021b).

The fungal fraction with relation to type material on NCBI was extracted from https://www.ncbi.nlm.nih.gov/assembly/?term=Fungi, with the criteria: ("Fungi"[Organism] OR Fungi[All Fields]) AND ("latest genbank"[filter] AND (all[filter] NOT partial[filter] AND all[filter] NOT anomalous[filter]) AND "from type"[Properties]). Genomic and protein Fasta files were downloaded for all assemblies with annotation status as: "has annotation".

Viral assemblies were downloaded from (NCBI, 2021b), with search details: ("Viruses"[Organism] OR Viruses[All Fields]) AND ("latest genbank"[filter] AND (all[filter] NOT partial[filter] AND all[filter] NOT anomalous[filter])).

Finally, soil and freshwater metagenomes were obtained with the query title: "freshwater metagenome" or "soil metagenome" on assembly database (Kitts et al., 2016), with options: AND (all[filter] NOT partial[filter] AND all[filter] NOT anomalous[filter]) Sort by: ORGN.

### *2.2. Data analysis and filtering*

All data gathered from the former section were decompressed with ZIP Extractor Pro version 2.0.1. Subsequently, we created sketched files from both genomic assemblies and protein data with Mash tool software version: 2.2.2 (Ondov et al., 2016), option *mash sketch *.gz*. Created sketched files in the format **.msh* can be used for fast trait distance estimations using nucleotide or protein alphabet. The functionality of each file generated was tested using the *mash dist* command. For processing data, we used the Ubuntu 18.04 LTS bash terminal for Windows 10, in a LENOVO workstation (MT_11D2_BU_Think_FM_ThinkCentre M90s) with Intel (R) Core (TM) i3-10300 CPU @ 3.70GHz, 3696 Mhz, with 8 logic processors, and total physical memory of 128 GB.

### *2.3. Determination of overall genomic coherence with the sketched data*

The genome of the type species for *Brucella ciceri* (GCA_012103155.1) was compared against the Database *Bacteria_Archaea_type_assembly_set.msh* (containing a representation of all prokaryotic genomes corresponding to type material) in order to explore its genomic coherence with other contexts within the genus *Brucella*. This genome belongs to a taxonomic group with historical importance in clinical and environmental microbiology (the *Brucellacea* family), and its relevance as a separate species is not entirely clear. The comparison was made through the command *mash dist GCA_012103155.1 Bacteria_Archaea_type_assembly_set.msh > out.txt*. The output was ordered from lowest to highest according to the Mash genomic distance parameter (D). The contexts with $D \leq 0.1$ were selected for the discussion of this work.

### *2.4. Repositories and data availability*

A GitHub repository (Sánchez-Reyes, 2021) ) has been implemented with usage and examples. Also in this repository, we will be communicating about the periodic updates and additions of the data. Individual databases could be reached at (Sánchez-Reyes and Fernández-López, 2021a)

## 3. Results

### *3.1. Up-to-date prokaryotic catalog with standing in nomenclature, content, and potential uses*

All prokaryotic genomes in the NCBI corresponding to the type material, updated to September 2021, were collected and processed to obtain representatives **.msh* sketched files. We have entitled this set as Bacteria_Archaea_type_assembly_set.msh, which comprises 17,442 genome assemblies with standing in nomenclature according to the List of Prokaryotic Names with Standing in Nomenclature (LPSN) (Parte et al., 2020) (Table 1). The raw data consume about 20 GB of storage (approximate download time > 1 hour), but after sketching it is reduced to only 137 MB (approximate download time: < 5 minutes, average transfer rate of 8 MB/s). Access to the material is found on (Sánchez-Reyes and Fernández-López, 2021c). As one of the purposes of this work is to offer curated references for taxonomic surveys, this set of genomes excludes anomalous assemblies; that is misassembled or chimeric, contaminated, or coming from unverified

**Table 1.** Mash sketched databases for genome-based taxonomy or comparative genomics.

| Dataset | Size raw (GB) | Size processed (MB) | Assemblies composition | File Type | Data collection |
|---|---|---|---|---|---|
| Bacteria_Archaea_type_assembly_set.msh | 20.80 | 137.46 | 17,442 | sketched genome_ fasta | September, 2021 |
| Bacteria_Archaea_type_proteome_set.msh | 9.86 | 3.13 | 12,767 | sketched prot_fasta | April, 2021 |
| GTDB_r202_assembly_set.msh | 46.30 | 377 | 47,894 | sketched genome_ fasta | September, 2021 |
| Fungi_type_assembly_set.msh | 5.90 | 6.32 | 801 | sketched genome_ fasta | September, 2021 |
| Fungi_type_proteome_set.msh | 0.70 | 0.088 | 248 | sketched prot_fasta | April, 2021 |
| Virus_Sept21_GenBank_assembly_set.msh | 0.66 | 351 | 44,916 | sketched genome_ fasta | September, 2021 |
| Soil_Metgenome_assembly_set.msh | 31.10 | 3.79 | 479 | sketched genome_ fasta | September, 2020 |
| Freshwater_Metagenome_assembly_set.msh | 18.20 | 4.83 | 611 | sketched genome_ fasta | April, 2021 |

source organisms, among others. The recommended use involves downloading the dataset and comparing the user genomic query through the Mash tool (Ondov et al., 2016) to obtain the reference genome(s) with the smallest mutational distance to the query. These genomes ideally will be the closest phylogenetic neighbors. The reduced list of neighbors can be promptly downloaded from the NCBI for phylogenomic analysis and calculation of genome relatedness index such as average nucleotide identity (ANI) among others (Konstantinidis and Tiedje, 2005). At this point, important clues can be obtained, if the D ≤ 0.05 a known species-specific context is likely present. This should be corroborated with the ANI calculation in which case it is expected to be ≥ 95%. If the Mash distance is ≥ 0.1 the genome query can be a totally new species (ANI must be <95%).

### *3.2. The GTDB assemblies set*

We have summarized the most updated genome representation corresponding to the Genome Taxonomy Database release 202 (GTDB r202) in a comprehensive and light sketched database (GTDBr202_genomic.fna.msh). It is one of the most complete databases for prokaryotic taxonomic purposes to date, the sketched version that we present contains 47,894 genomes, its raw space would be 47.00 GB nonetheless, after the reduction in **.msh* format it only uses 377 MB (Table 1). Access to the material is found on (Sánchez-Reyes and Fernández-López, 2021d). Not all his material has standing in nomenclature, but his proposals for cryptic taxa possess high phylogenetic support. Recommended use follows the same guidelines as the Bacteria_Archaea_type_assembly_set.msh set.

### *3.3. Fungal assemblies related to type material*

To offer valuable genomic records from the Fungi kingdom, we have collected type material assemblies updated to September 2021. The Fungi_type_assembly_set.msh is the sketched file for all Fungi assemblies in the NCBI corresponding to the type material. This data set contains 801 records (5.9 GB raw space) and after sketching it only has 6.32 MB (Table 1). This data set is also intended as a highly curated reference for taxonomic comparisons or phylogenetic purposes. Access to the material is found on (Sánchez-Reyes and Fernández-López, 2021e). Altogether, the former three sets of microbial data make up a very complete group of genomic resources especially useful in taxogenomics, due to their relationship either with the type material or precise phylogenetic evidence from the GTDB.

### *3.4. A microbial predicted-proteomes collection*

We also collected an up-to-date group of predicted proteomes from Bacteria, Archaea, and Fungi groups pertaining to type material. The prokaryotic data set represents protein sequence in Fasta format, containing 12,767 predicted proteomes (Bacteria_Archaea_type_proteome_set.msh), and after sketching it only has 3.13 MB. The equivalent file for all fungal annotated assemblies in the NCBI corresponding to the type material, updated to June 2021 (Fungi_type_proteome_set.msh), represents protein sequence in Fasta format, contains 248 predicted proteomes, sketched to only 88 KB (Table 1). The recommended use for this dataset is the comparison of predicted proteomes with the Mash tool (Ondov et al., 2016) in a similar way to the previously described but using an amino acid alphabet. Any genome relatedness index calculated with data from fungi-type datasets should be taken only as indicators of genomic coherence. Taxospecific assignments should be complemented with phenotypic methods. Access to both materials is found on (Sánchez-Reyes and Fernández-López, 2021f); (Sánchez-Reyes and Fernández-López, 2021g).

### *3.5. Viral and metagenomic representations*

As a resource for accessing viral assemblies, we offer a sketched collection for the Viruses group pertaining to the NCBI GenBank. It is the most populated dataset with 44,916 viral assemblies and uses 351 MB of disk space after sketching (Virus_Sept21_GenBank_assembly_set.msh accessible on (Sánchez-Reyes and Fernández-López, 2021h). The assemblies exclude partial or anomalous records to ensure their quality. This material is especially useful for rapid surveys on the containment of viral composites in experimental sequenced data (Ondov et al., 2019). Metagenomic representations of two important biomes were also made, freshwater and soil metagenomes. Soil metagenomes reported in this work contain 479 representative microbiomes from NCBI, obtained in September 2020 (Soil_Metgenome_assembly_set.msh accessible on (Sánchez-Reyes and Fernández-López, 2021i). The total disk usage for all raw sequences can be close to 31.1 GB, which is reduced to 3.78 MB after sketching. Freshwater biome collection contains 611 freshwater metagenome assemblies from NCBI downloaded on April 2021 (Freshwater_Metagenome_assembly_set.msh accessible on (Sánchez-Reyes and Fernández-López, 2021j). The disk storage for raw sequences would be 18.20 GB, after sketching the disk space is reduced to 4.82 MB (See Table 1 for metadata information). Both metagenomic datasets are intended to estimate overall nucleotide-level genomic similarity between the coding regions of compared metagenomes through the Mash distance or ANI calculation; however, due to the mixed composition of these data, there is no species-specific interpretation of these metrics.

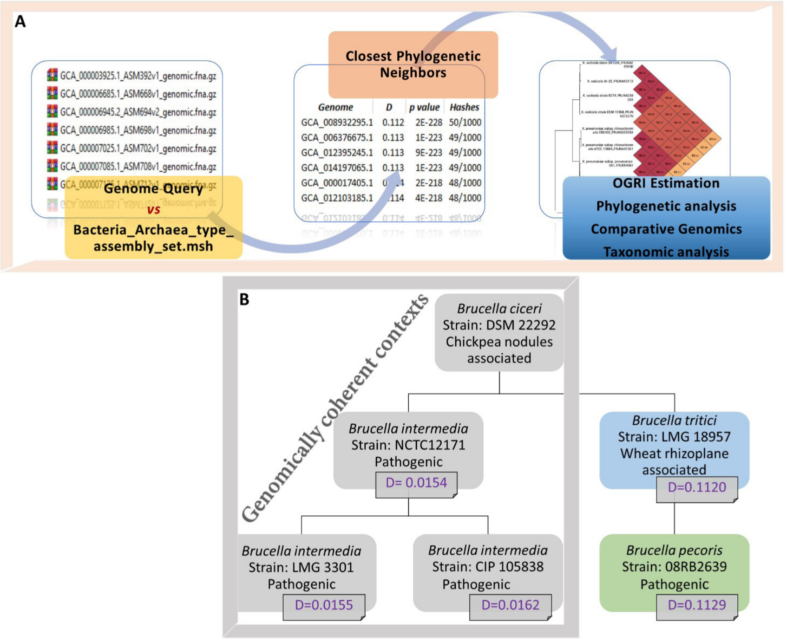
### *3.6. Brucella ciceri and Brucella intermedia, a case study using overall genome relatedness indexes*

To illustrate the relevance of the proposed databases, we have carried out a case study with the prokaryotic species *Brucella ciceri* and *Brucella intermedia* of the *Brucelaceae* family, an important group of microorganisms with medical, agricultural, and livestock importance. *Brucella intermedia* (Velasco et al., 1998) was described as a medically significant species often associated with human nosocomial infections, although the distribution of the species is not only limited to clinical locations (Aujoulat et al., 2014). *Brucella ciceri* (Imran et al., 2010) was described as associated with root nodules of chickpea *Cicer arietinum*. However, an environmental survey comparing all available genomic composites of the genus *Brucella* (unpublished data) revealed that the type strains of *B. intermedia* and *B. ciceri* share a significant genomic coherence: more than 95% of ANI, Mash genomic distance <0.05, and DNA-DNA hybridization rate of 95% (Sánchez-Reyes, 2020). These data constituted the first body of evidence to hypothesize that possibly *B. intermedia* and *B. ciceri* constituted the same genomospecies. Under the currently accepted concept to define a prokaryotic species, it is necessary phylogenetic evidence that indicates monophyletic relationships between both taxa compared to their closest neighbors. A quick and easy way to obtain the closest phylogenetic neighbors is by using a database such as *Bacteria_Archaea_type_assembly_set.msh* and accurately estimating the Mash genomic distance (D) (see Figure 1A). *B. ciceri* strain DSM 22292 shares distances around 0.01 with *B. intermedia* type strains NCTC12171, LMG 3301, and CIP 105838. All other 17,442 comparisons take less than two minutes and fall likely within the continuum of diversity, with mutational distance > 0.1 and clear species segregation from other neighbors like *Brucella tritici* strain LMG 18957 and *Brucella pecoris* strain 08RB2639 (see Figure 1B).

## 4. Discussion

The advent of massive DNA sequencing has rendered more than 1.5 billion sequences hosted in public databases. The handling of massive genomic information has shortcomings such as storage, internet access, or computing restrictions. To facilitate the access and use of this type of information for the scientific community with no access to a high-performance computational cluster or with storage limitations, it is necessary to obtain reduced representations of massive genomic databases with the least loss of information possible. The minhashing technique implemented by Ondov et al. (2016, 2019) to obtain small sketches of massive sequence data, could be systematically applied to generate accessible genomic reference data, useful in a wide variety of analyses of the properties of genomes, metagenomes, or predicted proteomes (as it was originally proposed). The publicly existing sketched databases came from the pioneering work of Ondov et al. (2016); that information just represents the RefSeq genomes Release 70 for bacteria (O'Leary et al., 2016), which came out in May 2015. This is one of the reasons that motivated us to present a set of updated and representative data from eight informative contexts, which have been filtered to make them accessible, usable, and user-friendly with limited computational resources (Table 1). Altogether, this dataset contains near to 133 GB of space disk reduced to 883.25 MB and represents 125,110 genomics/proteomic records.

These datasets offer an advantage over other curated datasets like MG-RAST (Meyer et al., 2008), MGnify platform (Mitchell et al., 2020), or EzBioCloud Genome Database (Yoon et al., 2017). MG-RAST is a database that we consider to be at the forefront of hosting genomic data; however, its main drawback is support and updating. Since July 2020 the site has not been updated, although remains functional. The MGnify platform contains highly curated metagenomic data and plentiful metadata. Both platforms are not user-friendly and raw data in large volumes is prohibitive for the common user. EzBioCloud Genome Database provides a highly curated genome set of Bacteria and Archaea for taxonomic purposes in raw format. Although this database


A
GCA_000003925.1_ASM392v1_genomic.fna.gz
GCA_000006685.1_ASM668v1_genomic.fna.gz
GCA_000006945.2_ASM694v2_genomic.fna.gz
GCA_000006985.1_ASM698v1_genomic.fna.gz
GCA_000007025.1_ASM702v1_genomic.fna.gz
GCA_000007085.1_ASM708v1_genomic.fna.gz
Genome Query
vs
Bacteria_Archaea_type_assembly_set.msh
Closest Phylogenetic Neighbors
OGRI Estimation
Phylogenetic analysis
Comparative Genomics
Taxonomic analysis
B
Genomically coherent contexts
Brucella ciceri
Strain: DSM 22292
Chickpea nodules associated
Brucella intermedia
Strain: NCTC12171
Pathogenic
D= 0.0154
Brucella tritici
Strain: LMG 18957
Wheat rhizoplane associated
D=0.1120
Brucella intermedia
Strain: LMG 3301
Pathogenic
D=0.0155
Brucella intermedia
Strain: CIP 105838
Pathogenic
D=0.0162
Brucella pecoris
Strain: 08RB2639
Pathogenic
D=0.1129

| *Genome* | *D* | *p value* | *Hashes* |
|---|---|---|---|
| GCA_008932295.1 | 0.112 | 2E-228 | 50/1000 |
| GCA_006376675.1 | 0.113 | 1E-223 | 49/1000 |
| GCA_012395245.1 | 0.113 | 9E-223 | 49/1000 |
| GCA_014197065.1 | 0.113 | 1E-223 | 49/1000 |
| GCA_000017405.1 | 0.114 | 2E-218 | 48/1000 |
| GCA_012103185.1 | 0.114 | 4E-218 | 48/1000 |

**Figure 1.** Workflow using sketched database containing type genomes of prokaryotic origin and their output. Panel A illustratively shows the 3 main steps of use: comparison, output, and post-processing. Panel B shows a grouping of different species of the genus *Brucella* according to the genomic distance estimated by Mash.

meets the taxonomic precision criterion, it is only available through an identification service called TrueBacTM, which is a major limitation.

Genomic relationship indices can help us to establish species-specific contexts in microorganisms and to rethink hypotheses regarding the concept of prokaryotic species, a fundamental problem in biology. For this, access to taxonomically curated data is essential. The case of *B. ciceri* and *B. intermedia* as similar species-specific contexts has been previously discussed by Sánchez-Reyes (2020). There is genomic and phylogenetic evidence that indicates that they are the same genomospecies. This evidence comes from the analysis of highly precise relationship indices such as ANI, *in silico* DNA–DNA hybridization, and Mash D, and supports the requirement of genomic coherence between both species. Contexts with D ≤ 0.05 have been consistently shown to be strongly correlated with ANI ≥ 95%, which is a current standard for classifying prokaryotic species. Also, the phylogeny appeals on genomic analysis for the selection of phylogenetic neighbors with comparative value, showing that both species also form a monophyletic cluster separated from other members within the genus *Brucella* (See Figure 1 on (Sánchez-Reyes, 2020)). Given the wide range of distribution for *B. intermedia*, it would be important to reconsider the relevance of the diagnostic characters proposed for *B. ciceri* in the transition to a new species. In our opinion, genomic and phylogenetic coherence suggests that *B. ciceri* is actually a discrete ecotype of *B. intermedia*, which differs -as expected- in various physiological responses as a result of adaptation to local environmental conditions. These observations remained hidden for years and have come to light in the scientific discussion thanks to the analysis of complete genomes, hence the importance of genomic databases in formats accessible to the community (see Figure 1).

Finally, our data offer microbial genomic information derived from the Type Material (Federhen, 2015) in a reduced format, which ensures collective appropriation of precise taxonomic references. Since genome-based taxonomy has already surpassed traditional ribosomal gene comparison methods, the most consistent way to carry out species delineation in prokaryotes is through “taxogenomics” with whole genome information. This is only possible with suitable references that come from material with standing in nomenclature. To the best of our knowledge, there is no precedent to concentrate up-to-date prokaryotic and eukaryotic type material assemblies in a lightweight and accessible format.

## Acknowledgements

We thank to IBT-UNAM and the program “Investigadoras e Investigadores por México” from the Consejo Nacional de Ciencia y Tecnología, México, for supporting the Project 237. This research was partially funded by the Programa Presupuestario F003, under the project with code CF 2019 265222. Also we thanks to Dra. Judith Tabullo de Robles for his help with the data visualization.

## References

AUJOULAT, F., ROMANO-BERTRAND, S., MASNOU, A., MARCHANDIN, H. and JUMAS-BILAK, E., 2014. Niches, population structure and genome reduction in Ochrobactrum intermedium: clues to technology-driven emergence of pathogens. *PLoS One*, vol. 9, no. 1, p. e83376. http://dx.doi.org/10.1371/journal.pone.0083376. PMid:24465379.

CHUN, J. and RAINEY, F.A., 2014. Integrating genomics into the taxonomy and systematics of the acteria and Archaea. *International Journal of Systematic and Evolutionary Microbiology*, vol. 64, no. Pt 2, pp. 316-324. http://dx.doi.org/10.1099/ijs.0.054171-0. PMid:24505069.

FEDERHEN, S., 2015. Type material in the NCBI taxonomy database. *Nucleic Acids Research*, vol. 43, no. D1, pp. D1086-D1098. http://dx.doi.org/10.1093/nar/gku1127. PMid:25398905.

GENOME TAXONOMY DATABASE – GTDB [online], 2021a [viewed 1 April 2021]. Available from: https://data.ace.uq.edu.au/public/gtdb/data/releases/release202/202.0/genomic_files_reps.

GENOME TAXONOMY DATABASE – GTDB [online], 2021b [viewed 1 April 2021]. Available from: https://data.ace.uq.edu.au/public/gtdb/data/releases/release202/202.0/genomic_files_reps/gtdb_genomes_reps_r202.tar.gz.

KITTS, P.A., CHURCH, D.M., THIBAUD-NISSEN, F., CHOI, J., HEM, V., SAPOJNIKOV, V., SMITH, R.G., TATUSOVA, T., XIANG, C., ZHERIKOV, A., DICUCCIO, M., MURPHY, T.D., PRUITT, K.D. and KIMCHI, A., 2016. Assembly: a resource for assembled genomes at NCBI. *Nucleic Acids Research*, vol. 44, no. D1, pp. D73-D80. http://dx.doi.org/10.1093/nar/gkv1226.

KONSTANTINIDIS, K.T. and TIEDJE, J.M., 2005. Genomic insights that advance the species definition for prokaryotes. *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 7, pp. 2567-2572. http://dx.doi.org/10.1073/pnas.0409727102. PMid:15701695.

MEYER, F., PAARMANN, D., D'SOUZA, M., OLSON, R., GLASS, E.M., KUBAL, M., PACZIAN, T., RODRIGUEZ, A., STEVENS, R., WILKE, A., WILKENING, J. and EDWARDS, R.A., 2008. The metagenomics RAST server–a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics*, vol. 9, no. 1, p. 386. http://dx.doi.org/10.1186/1471-2105-9-386. PMid:18803844.

MITCHELL, A.L., ALMEIDA, A., BERACOCHEA, M., BOLAND, M., BURGIN, J., COCHRANE, G., CRUSOE, M.R., KALE, V., POTTER, S.C., RICHARDSON, L.J., SAKHAROVA, E., SCHEREMETJEW, M., KOROBEYNIKOV, A., SHLEMOV, A., KUNYAVSKAYA, O., LAPIDUS, A. and FINN, R.D., 2020. MGnify: the microbiome analysis resource in 2020. *Nucleic Acids Research*, vol. 48, no. D1, pp. D570-D578. http://dx.doi.org/10.1093/nar/gkz1035. PMid:31696235.

NATIONAL CENTER FOR BIOTECHNOLOGY INFORMATION – NCBI, 2021a [viewed 1 September 2021]. *National Library of Medicine - National Center for Biotechnology Information* [online]. Bethesda (MD): NLM. Available from: https://www.ncbi.nlm.nih.gov/assembly/?term=Prokaryote.

NATIONAL CENTER FOR BIOTECHNOLOGY INFORMATION – NCBI, 2021b [viewed 1 September 2021]. *National Library of Medicine - National Center for Biotechnology Information* [online]. Bethesda (MD): NLM. Available from: https://www.ncbi.nlm.nih.gov/assembly/?term=Viruses.

NCBI RESOURCE COORDINATORS, 2018. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, vol. 46, no. D1, pp. D8-D13. http://dx.doi.org/10.1093/nar/gkx1095. PMid:29140470.

O'LEARY, N.A., WRIGHT, M.W., BRISTER, J.R., CIUFO, S., HADDAD, D., MCVEIGH, R., RAJPUT, B., ROBBERTSE, B., SMITH-WHITE, B., AKO-ADJEI, D., ASTASHYN, A., BADRETDIN, A., BAO, Y., BLINKOVA, O., BROVER, V., CHETVERNIN, V., CHOI, J., COX, E., ERMOLAEVA, O., FARRELL, C.M., GOLDFARB, T., GUPTA, T., HAFT, D., HATCHER, E., HLAVINA, W., JOARDAR, V.S., KODALI, V.K., LI, W., MAGLOTT, D., MASTERSON, P., MCGARVEY, K.M., MURPHY, M.R., O'NEILL, K., PUJAR, S., RANGWALA, S.H., RAUSCH, D., RIDDICK, L.D., SCHOCH, C., SHKEDA, A., STORZ, S.S., SUN, H., THIBAUD-NISSEN, F., TOLSTOY, I., TULLY, R.E., VATSAN, A.R., WALLIN, C., WEBB, D., WU, W., LANDRUM, M.J., KIMCHI, A., TATUSOVA, T., DICUCCIO, M., KITTS, P., MURPHY, T.D. and PRUITT, K.D., 2016. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research*, vol. 44, no. D1, pp. D733-D745. http://dx.doi.org/10.1093/nar/gkv1189. PMid:26553804.

ONDOV, B.D., STARRETT, G.J., SAPPINGTON, A., KOSTIC, A., KOREN, S., BUCK, C.B. and PHILLIPPY, A.M., 2019. Mash screen: high-throughput sequence containment estimation for genome discovery. *Genome Biology*, vol. 20, no. 1, p. 232. http://dx.doi.org/10.1186/s13059-019-1841-x. PMid:31690338.

ONDOV, B.D., TREANGEN, T.J., MELSTED, P., MALLONEE, A.B., BERGMAN, N.H., KOREN, S. and PHILLIPPY, A.M., 2016. Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biology*, vol. 17, no. 1, p. 132. http://dx.doi.org/10.1186/s13059-016-0997-x. PMid:27323842.

PARKS, D.H., CHUVOCHINA, M., CHAUMEIL, P.A., RINKE, C., MUSSIG, A.J. and HUGENHOLTZ, P., 2020. A complete domain-to-species taxonomy for Bacteria and Archaea. *Nature Biotechnology*, vol. 38, no. 9, pp. 1079-1086. http://dx.doi.org/10.1038/s41587-020-0501-8. PMid:32341564.

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. *Nature Biotechnology*, vol. 36, no. 10, pp. 996-1004. http://dx.doi.org/10.1038/nbt.4229. PMid:30148503.

PARTE, A.C., CARBASSE, J.S., MEIER-KOLTHOFF, J.P., REIMER, L.C. and GÖKER, M., 2020. List of prokaryotic names with standing in nomenclature (LPSN) moves to the DSMZ. *International Journal of Systematic and Evolutionary Microbiology*, vol. 70, no. 11, pp. 5607-5612. http://dx.doi.org/10.1099/ijsem.0.004332. PMid:32701423.

SÁNCHEZ-REYES, A. and FERNÁNDEZ-LÓPEZ, M.G. (2021a): Mash sketched dataset for: mash sketched reference dataset for genome-based taxonomy and comparative genomics. *Figshare*. In press. https://doi.org/10.6084/m9.figshare.14408801.v5

SÁNCHEZ-REYES, A. and FERNÁNDEZ-LÓPEZ, M.G., 2021b. Mash sketched reference dataset for genome-based taxonomy and comparative genomics. *Preprints*, 2021060368. In press. https://doi.org/10.20944/preprints202106.0368.v1.

SÁNCHEZ-REYES, A. and FERNÁNDEZ-LÓPEZ, M.G., 2021c [viewed 1 September 2021]. Mash sketched databases for: mash sketched reference dataset for genome-based taxonomy and comparative genomics. *Figshare* [online]. Available from: https://figshare.com/ndownloader/files/30851626

SÁNCHEZ-REYES, A. and FERNÁNDEZ-LÓPEZ, M.G., 2021d [viewed 1 September 2021]. Mash sketched databases for: mash sketched reference dataset for genome-based taxonomy and comparative genomics. *Figshare* [online]. Available from: https://figshare.com/ndownloader/files/30863182.

SÁNCHEZ-REYES, A. and FERNÁNDEZ-LÓPEZ, M.G., 2021e [viewed 1 September 2021]. Mash sketched databases for: mash sketched reference dataset for genome-based taxonomy and comparative genomics. *Figshare* [online]. Available from: https://figshare.com/ndownloader/files/30871351.

SÁNCHEZ-REYES, A. and FERNÁNDEZ-LÓPEZ, M.G., 2021f [viewed 1 September 2021]. Mash sketched databases for: mash sketched reference dataset for genome-based taxonomy and comparative genomics. *Figshare* [online]. Available from: https://figshare.com/ndownloader/files/27631017.

SÁNCHEZ-REYES, A. and FERNÁNDEZ-LÓPEZ, M.G., 2021g [viewed 1 September 2021]. Mash sketched databases for: mash sketched reference dataset for genome-based taxonomy and comparative genomics. *Figshare* [online]. Available from: https://figshare.com/ndownloader/files/27631026.

SÁNCHEZ-REYES, A. and FERNÁNDEZ-LÓPEZ, M.G., 2021h [viewed 1 September 2021]. Mash sketched databases for: Mash Sketched Reference Dataset for Genome-Based Taxonomy and Comparative Genomics. *Figshare* [online]. Available from: https://figshare.com/ndownloader/files/30863599.

SÁNCHEZ-REYES, A. and FERNÁNDEZ-LÓPEZ, M.G., 2021i [viewed 1 September 2021]. Mash sketched databases for: mash sketched reference dataset for genome-based taxonomy and comparative genomics. *Figshare* [online]. Available from: https://figshare.com/ndownloader/files/27631032.

SÁNCHEZ-REYES, A. and FERNÁNDEZ-LÓPEZ, M.G., 2021j [viewed 1 September 2021]. Mash sketched databases for: mash sketched reference dataset for genome-based taxonomy and comparative genomics. *Figshare* [online]. Available from: https://figshare.com/ndownloader/files/27631020.

SÁNCHEZ-REYES, A., 2020. Reclassification of Brucella ciceri as later heterotypic synonyms of Brucella intermedia. *bioRxiv*. In press. https://doi.org/10.1101/2020.08.16.251660.

SÁNCHEZ-REYES, A., 2021 [viewed 1 September 2021]. *GitHub-ayixon/Mash-sketched-reference-databases* [online], Available from: https://github.com/ayixon/Mash-sketched-reference-databases

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. *International Journal of Systematic and Evolutionary Microbiology*, vol. 67, no. 5, pp. 1613-1617. http://dx.doi.org/10.1099/ijsem.0.001755. PMid:28005526.