Data Descriptor

Mash Sketched Reference Dataset for Genome-Based Taxonomy and Comparative Genomics

Ayixon Sánchez-Reyes [1]*, Maikel Gilberto Fernández-López[2]

[1] Cátedras Conacyt-Instituto de Biotecnología, Universidad Nacional Autónoma de México.
Avenida Universidad 2001, Chamilpa, 62210 Cuernavaca, Morelos. México; ayixon.sanchez@mail.ibt.unam.mx
[2] Centro de Investigación en Dinámica Celular-Instituto de Investigaciones Básicas y Aplicadas, Universidad Autónoma del Estado de Morelos, Cuernavaca 62209, México; fmaikel44@gmail.com
* Correspondence: ayixon.sanchez@mail.ibt.unam.mx; Tel.: + 52 7771021529

Abstract: The analysis of curated genomic, metagenomic, and proteomic data are of paramount importance in the fields of biology, medicine, education, and bioinformatics. Although this type of data is usually hosted in raw form in free international repositories, its access requires plenty of computing, storage, and processing capacities for the domestic user. The purpose of the study is to offer a comprehensive set of genomic and proteomic reference data, in an accessible and easy-to-use form to the scientific community. A representative type material set of genomes, proteomes and metagenomes were directly downloaded from the site: https://www.ncbi.nlm.nih.gov/assembly/ and from Genome Taxonomy Database, associated with the major groups of Bacteria, Archaea, Virus, and Fungi. Sketched databases were subsequently created and stored on handy raw reduced representations, by using Mash software. Our dataset contains near to 100 GB of space disk reduced to 585.78 MB and represents 87,476 genomics/proteomic records from eight informative contexts, which have been prefiltered to make them accessible, usable, and user-friendly with computational resources. Potential uses of this dataset include but are not limited to, microbial species delimitation, estimation of genomic distances, genomic novelties, paired comparisons between proteomes, genomes, and metagenomes.

Dataset: https://doi.org/10.6084/m9.figshare.14408801.v3.

Dataset License: CC BY 4.0

Keywords: Microbial Mash database, Mash distance, Genome containment, Type material, Microbial taxonomy

1. Summary

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 index and genome-to-genome sequence comparison methods, constitute standard procedures in genome-based taxonomy and comparative genomics [1]. These methods depend on reference databases that contain numerous records classified as genomes, proteomes, transcriptomes, or even metagenomes. Currently, the National Center for Biotechnology Information (NCBI) contains more than 900 thousand bacterial assemblies, almost seven thousand Archaean genomes, and near to 800 fungal genomes [2]. Storing such amount of information is difficult to reach for the common and personal user due to storage, broadband internet access, or computing limitations. Other representative databases such as the NCBI type material for prokaryotes use 20 GB of space in compressed format [3], and the Genome Taxonomy Database (GTDB) -release 95- contains

194,600 genomes and requires almost 40 GB of compressed space [4]. 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 definition 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. Finally, 31,910 genomes from GTDB -release 89- (GTDB r89) were downloaded as an external updated taxonomic resource. With the downloaded elements, representative datasets were built using the powerful MinHashing dimensional reduction technique over genomics data [5]. In all cases, the Mash sketch function included in the Mash program was used to obtain reduced representations of the DNA or amino acid alphabets. These datasets contain only a small fraction of the original size of the raw data; for comparison, the sketched GTDB r89 just contains 250 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 for assigning taxonomic categories, and genomic resemblances. The sketched *.*msh* dataset from this work can be quickly downloaded from figshare.com and executed on a modest desktop computer with as little as 4GB of ram and 1 GB of free space.

2. Data Description

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 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 [5] to obtain small sketches of massive sequence data, could be systematically applied to generate accessible 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 [5]; that information just represents the RefSeq genomes Release 70 for bacteria [6], which came out in May 2015 (available on https://mash.readthedocs.io/en/latest/data.html).

This paper presents a set of representative data from eight informative contexts, which have been filtered by the mash tool to make them accessible, usable, and user-friendly with computational resources (Table 1). Altogether, this dataset contains near to 100 GB of space disk reduced to 585.78 MB and represents 87,476 genomics/proteomic records.

Table 1. Mash sketched reference database for genome-based taxonomy or comparative genomics

Dataset	Size raw (GB)	Size processed (MB)	Entries downloaded	File Type	Data collection date
Bacteria_Archaea_type_assembly_set.msh	19.50	128	16,304	sketched genome_fasta	April, 2021
Bacteria_Archaea_type_proteome_set.msh	9.86	3.13	12,767	sketched prot_fasta	April, 2021
GTDB_r89_assembly_set.msh	32.40	250	31,910	sketched genome_fasta	March, 2021
Fungi_type_assembly_set.msh	5.50	5.96	753	sketched genome_fasta	April, 2021
Fungi_type_proteome_set.msh	0.70	0.088	248	sketched prot_fasta	April, 2021
Virus_ Latest GenBank_assembly_set.msh	0.62	318	40,708	sketched genome_fasta	April, 2021
Soil_Metgenome_assembly_set.msh	31.10	3.78	479	sketched genome_fasta	September, 2020
Freshwater_Metagenome_assembly_set.msh	18.20	4.82	611	sketched genome_fasta	April, 2021

2.1. Dataset content description and potential uses

1. *The Bacteria_Archaea_type_assembly_set.msh* is the sketched file for all prokaryotic genomes in the NCBI corresponding to the type material, downloaded in April 2021. Each record of this set (16304 assemblies) has standing in nomenclature according to the LPSN (List of Prokaryotic Names with Standing in Nomenclature available at (http://www.bacterio.net) [7], which makes it ideal for taxonomic surveys. The raw data would consume about 20 GB of storage (approximate download time > 1 hour), but after sketching it is reduced to only 128 MB (approximate download time: < 5 minutes).
2. *Bacteria_Archaea_type_proteome_set.msh* is the sketched file for all prokaryotic genomes in the NCBI corresponding to the type material, downloaded in April 2021. This data set represents protein sequence in Fasta format, contains 12,767 predicted proteomes, and after sketching it only has 3.3 MB.
3. *GTDB_r89_assembly_set.msh* contains the genome representation corresponding to the Genome Taxonomy Database implemented in the MetaSanity tool [8]. It is one of the most complete databases for taxonomic purposes to date, the sketched version that we present contains 31,910 genomes, its raw space would be 32.40 GB but after the reduction, it only occupies 250 MB.
4. *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 753 records (5.5 GB raw space)

doi:10.20944/preprints202106.0368.v1

and after sketching it only has 5.96MB. This data set is also intended as a reference for taxonomic comparisons or even phylogenetic purposes.

5. *Fungi_type_proteome_set.msh* is the sketched file for all fungal annotated assemblies in the NCBI corresponding to the type material, downloaded in April 2021. This data set represents protein sequence in Fasta format, contains 248 predicted proteomes, and after sketching it only has 88 KB.

6. *Virus_ Latest GenBank_assembly_set.msh* is the sketched file for the Viruses group in the NCBI GenBank, as in the former datasets, this excludes partial or anomalous assemblies. It is the most populated dataset with 40,708 viral assemblies and occupies 318 MB of disk space.

7. *Soil_Metgenome_assembly_set.msh* is a sketched file of 479 representative soil metagenomes from NCBI downloaded on September 2020. All the raw sequences can have a disk usage close to 31.1 GB, which is reduced to 3.78 MB after sketching.

8. *Freshwater_Metagenome_assembly_set.msh* this data set contains 611 freshwater metagenome assemblies from NCBI downloaded on April 2021. The disk storage for raw sequences would be 18.20 GB, after sketching the disk space is reduced to 4.82 MM.

2.2. Value of the Data

- Databases for taxonomic and phylogenomic analysis, or metagenome comparison, comprise massive amounts of data (hundreds of GB and thousands of genomic records), difficult to store and process in domestic computers. These data, represent an updated collection of common-use genomic records by the academic community, reduced to manageable representations on a laptop or desktop computer with limited resources without sacrificing coverage.
- The dataset can be used as a reference in evolutionary studies, genome-based microbial taxonomy, as well as in comparative genomics, microbial ecology, or metagenomic surveys. The data are also useful for the exploration of microbial taxonomic profiles at the community level. Covering three major groups, the prokaryote, the fungi kingdom and viruses, microbiologists, geneticists, bioinformatics, etc., can find use to this data.

- This dataset could be used and accessed freely by the scientific community to obtain up-to-date genomics, metagenomics, and proteomic references, to explore trait similarities. The databases are particularly useful on the inference resemblance among genomes or proteomes; in the purification and selection of close phylogenetic neighbors to test phylogenomic hypotheses, and in the clustering of new metagenomic data sets obtained in environmental surveys.

- The scientific community can also benefit from this dataset for genome or metagenome containment studies. Overall data specification can be seen in Table 2

Table 2. General Data Specification

Data specification	Value
Data Field	Biological sciences
Specific field area	Omics: General; Microbiology: Microbiome; Bioinformatics; Taxonomy
Type of data	Pre-sketched genomic and proteomic archives
Data format	Filtered in **.msh* final format
Parameters for data collection	Sequence data collected here corresponds to September 2020-April 04, 2021, included in NCBI's Assembly resource (www.ncbi.nlm.nih.gov/assembly/).
Primary data source	The primary sources used are listed as follow: Assembly set Bacteria and Archaea: https://www.ncbi.nlm.nih.gov/assembly/?term=Prokaryote Assembly set Fungi: https://www.ncbi.nlm.nih.gov/assembly/?term=Fungi GTDB data repository: https://data.ace.uq.edu.au/public/gtdb/data/releases/release89/89.0 Assembly set soil metagenomes: https://www.ncbi.nlm.nih.gov/assembly Assembly set freshwater metagenomes: https://www.ncbi.nlm.nih.gov/assembly/?term=freshwater+metagenome Assembly set for viral genomes https://www.ncbi.nlm.nih.gov/assembly/?term=viruses
Repository name	Mash Sketched databases for Mash Sketched Reference Dataset for Genome-Based Taxonomy and Comparative Genomics

3. Methods

3.1. Acquisition of raw genomic, proteomic and metagenomic data

Genome assemblies, proteomic and metagenomic data were directly downloaded from the NCBI's Assembly resource site: https://www.ncbi.nlm.nih.gov/assembly/ on April 04, 2021 (except for “soil metagenome set” that were downloaded on September 2020). The search details were as follow: The prokaryotic genomic fraction with relation to type material on NCBI was extracted from https://www.ncbi.nlm.nih.gov/assembly/?term=Prokaryote, with the criteria: ("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]). Genomic and protein Fasta files were downloaded for all assemblies with annotation status as: “Has annotation”.

The genomes from Genome Taxonomy Database release 89 were obtained remotely with the command python3 download-data.py -d gtdbtk available in the MetaSanity package

[8]. All referred genomes correspond to the fastani directory included in the downloaded folder.

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 https://www.ncbi.nlm.nih.gov/assembly/?term=Viruses, 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 https://www.ncbi.nlm.nih.gov/assembly/,with options: AND (all[filter] NOT partial[filter] AND all[filter] NOT anomalous[filter]) Sort by: ORGN

3.2. *Data analysis and filtering*

All data gathered from the former section were individually decompressed with ZIP Extractor free app. Subsequently, we created sketched files from both genomic assemblies and protein data with mash tool software version: 2.2.2 [5], 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 created 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, 8 logic processors, and total physical memory of 128 GB.

4. User Notes

A GitHub repository (https://github.com/ayixon/Mash-sketched-reference-databases) has been implemented with usage and examples.

Author Contributions: Ayixon Sánchez-Reyes: Conceptualization, Methodology, Data curation, and filtration and Drafted the original manuscript. Maikel Gilberto Fernández-López: Data usability validation, Review & editing of the original manuscript. All authors reviewed and approved the manuscript.

Funding: This research was partially funded by the program CIENCIA DE FRONTERA FORDECYT-PRONACES, grant number 265222/2020 CONACyT- México.

Data Availability Statement: The reference datasets generated in this work were deposited in the figshare repository (https://doi.org/10.6084/m9.figshare.14408801.v3) and are freely available for the scientific community.

Acknowledgments: The authors thanks UNAM-IBT, for support this study. ASR thank to the program CATEDRAS CONACYT from the Consejo Nacional de Ciencia y Tecnología, México, for supporting the Project 237.

Conflicts of Interest: The authors declare no conflict of interest.

References

1.

1 Chun, J.; Rainey, F.A.Y. 2014 Integrating Genomics into the Taxonomy and Systematics of the Bacteria and Archaea. *International Journal of Systematic and Evolutionary Microbiology 64*, 316–324, doi:10.1099/ijs.0.054171-0.

2 Database Resources of the National Center for Biotechnology Information. Nucleic Acids Res 2018, 46, D8–D13, doi:10.1093/nar/gkx1095.

3 Federhen, S. Type Material in the NCBI Taxonomy Database. Nucleic Acids Research 2015, 43, D1086–D1098, doi:10.1093/nar/gku1127.

4 Chaumeil, P.-A.; Mussig, A.J.; Hugenholtz, P.; Parks, D.H. GTDB-Tk: A Toolkit to Classify Genomes with the Genome Taxonomy Database. Bioinformatics 2020, 36, 1925–1927, doi:10.1093/bioinformatics/btz848.

5 Ondov, B.D.; Treangen, T.J.; Melsted, P.; Mallonee, A.B.; Bergman, N.H.; Koren, S.; Phillippy, A.M. Mash: Fast Genome and Metagenome Distance Estimation Using MinHash. Genome Biol 2016, 17, doi:10.1186/s13059-016-0997-x.

6 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.; et al. Reference Sequence (RefSeq) Database at NCBI: Current Status, Taxonomic Expansion, and Functional Annotation. Nucleic Acids Res 2016, 44, D733–D745, doi:10.1093/nar/gkv1189.

7 Parte, A.C.; Sardà Carbasse, J.; Meier-Kolthoff, J.P.; Reimer, L.C.; Göker, M. List of Prokaryotic Names with Standing in Nomenclature (LPSN) Moves to the DSMZ. Int J Syst Evol Microbiol 2020, 70, 5607–5612, doi:10.1099/ijsem.0.004332.

8 Neely, C.J.; Graham, E.D.; Tully, B.J. MetaSanity: An Integrated Microbial Genome Evaluation and Annotation Pipeline. Bioinformatics 2020, 36, 4341–4344, doi:10.1093/bioinformatics/btaa512.