{Comment}

Improving the usability and comprehensiveness of microbial databases

Caitlin Loeffler[1†], Aaron Karlsberg[1†], Lana S. Martin[1], Eleazar Eskin[1], David Koslicki[2,3,4], Serghei Mangul[5*]

[1] Department of Computer Science, University of California Los Angeles, 404 Westwood Plaza, Los Angeles, CA 90095, USA
[2] School of Computer Science and Engineering, The Pennsylvania State University, 207 Electrical Engineering West University Park, PA 16802, USA
[3] Department of Biology, The Pennsylvania State University, 208 Curtin Rd, State College, PA 16801, USA
[4] The Huck Institutes of the Life Sciences, The Pennsylvania State University, 101 Huck Life Sciences Building, University Park, PA 16802, USA
[5] Titus Department of Clinical Pharmacy, University of Southern California School of Pharmacy, 1985 Zonal Ave, Los Angeles, CA 90089, USA

[†] These authors contributed equally to this work.

*Correspondence: mangul@usc.edu

Abstract

Metagenomics studies leverage genomic reference databases for discoveries in basic science and translational research. We analyze existing databases and discuss guidelines for development of a master reference database that promises to improve the quality and quantity of omics research.

Text

High-throughput sequencing has revolutionized microbiome research by enabling the detection of thousands of microbial genomes directly from their host environments[1]. This approach, known as metagenomics, is capable of capturing the complex interactions that take place between thousands of different microbial organisms in their natural habitats. Metagenomic methods rely on comparison of a sample genome to multiple reference genomes[2]. Metagenomics is more expensive to perform than traditional, culture-based taxonomic identification techniques, but today's methods can produce a more comprehensive reconstruction of microbial genomes[3]. Emerging technologies for identifying and analysing microbial genomes can provide valuable insights into the interactions between human microbiomes and medicines. However, the current ad hoc practice of storing reference genomes in multiple, disparate reference databases challenges the accuracy and comprehensiveness of future microbial metagenomics studies.

Metagenomic studies isolate DNA found in a sample of the environment (e.g., soil, sea water, fecal matter), compare the sampled genomes to verified reference genomes,

and identify the organism from which the DNA came. Ideally, a metagenomic study uses a reference database that contains all known genomic references. Today's researcher can choose from many different genomic reference databases that contain verified reference genomes, but these databases lack a universal standard of specimen inclusion, data preparation, and accessibility.

Several limitations in genomic sequencing present unique challenges to assembling reference databases. Notably, reference genomes can exist in various stages of completion: typically short DNA fragments (e.g., Illumina reads) are reassembled to form larger sequences referred to as contigs; a computationally intensive task. Contigs represent an unordered subset of contiguous portions of chromosomes, and so a reference genome comprised of contigs is considered incomplete. Given sufficient data, contigs can be further assembled into scaffolds that represent larger portions of individual chromosomes but gaps can remain (consisting of a possibly unknown number of unknown nucleotides). In contrast, an idealized "complete" genome contains no gaps and no unknown nucleotides with practical definitions varying for when a reference genome is considered complete. Most reference genomes are in different stages of completeness, with portions of even the human genome remaining unknown (in particular, the centromere and telomere regions).

In addition, the location of possibly incomplete reference genomes on taxonomic or phylogenetic trees can be contentious. Metagenomics researchers must take into account discrepancies in the types of taxa included in each reference genome database, as well as differences in how the genomes are constructed, identified, and made available for distribution.

The future of metagenomics research would benefit from a standardized, comprehensive approach to reference genome database development. To begin assembling a set of recommendations for reference genome database construction, we assessed the concordance and usability of available reference databases for fungal genomics. Our study considered the concordance of fungal species and genera across four large, generalized reference databases: Ensembl[4], RefSeq[5], JGI's 1000 fungal genomes project (JGI 1K)[6], and FungiDB[7]. We compared the fungal taxa in each of the four databases using universal taxonomic identifiers (hereafter called taxIDs) at the taxonomic ranks of strain (i.e., intraspecific), species, and genus.

Our comparison of four major fungal reference genomes revealed substantial discrepancies across databases in the presence of microbial species at taxonomic levels below the family rank. In other words, a researcher's selection of one particular fungal genome reference database could substantially impact the number and types of unique fungal taxa identified in a study.

Calculating the coverage of each reference genome database shows that a researcher using the largest—and most comprehensive—reference database would only find identification for 63% of all known fungal genera (Figure 1A) and 66% of all known fungal species (Figure 1B).

Calculating the percentage of fungal species per reference database reveals that a relatively small percentage of species are represented as complete genomes. Only 23.5% of species are represented as complete genomes in RefSeq, 1.4% in JGI 1K,

18.5% in Ensemble, and 13.4% in FungiDB. Conversely, our study shows that the percentage of species represented as contigs are relatively high: 69.6% in RefSeq, 92.2% in JGI 1K, 76.2% in Ensembl, and 79.9% in FungiDB. Only a minority of species in the reference databases studied are represented by both contigs and complete chromosomes: 6.8% in RefSeq, 6.3% in JGI 1K, 5.3% in Ensembl, and 6.7% in FungiDB (Figure 1D). Remaining genomes are comprised of contigs or a mixture of chromosomes and contigs.

The completeness of a reference database is always subject to limitations imposed by funding or scope. As one example of the latter, the JGI 1K database contains many novel and previously unpublished genomes. The introductory text of the JGI database indicates that, for this reason, it is not designed to be used in metagenomics studies[7]. However, such a large database of novel references may be a top choice for metagenomics researchers who want to learn as much as they can about their genomic fungal samples.

Even taking these database-specific limitations into account, researchers would benefit from a universal approach to constructing fungal genomic reference databases. Since the ideal reference database containing all the reference genomes for all known samples does not yet exist, researchers are potentially failing to identify key organisms within their samples.

The first consideration of a master reference database would be a standardized approach to assembling and presenting data from existing reference databases, which promises to help improve fungal coverage in newly developed metagenomic analysis tools.

One approach to developing a comprehensive database of the most compete known genomic references is to combine all existing fungal reference databases into one master set—a complex, time-consuming task. References unique to one database could simply be added to a master set. However, a reference that is found in more than one database presents several problems. Multiple references may be assigned the same taxID, yet these references may contain differing genomic information. For example, references comprised of contigs could cover different segments of a given gene. Selecting both unedited contig-based reference genomes would unnecessarily extend the run time of a comparison algorithm utilizing the master set. On the other hand, eliminating one reference would ignore entire segments of the genome represented in the discarded contigs. In such cases, the database developer needs a consistent method for selecting one of the references to include in the master set.

An alternative approach would be to develop an open source computational method that continuously merges any number of disjointed microbial reference databases as new sequences become available. The sequencing and storing of microbial species in multiple repositories present an opportunity to improve sequence quality through an approach based on alignment and consensus. An open source format would encourage computational developers to contribute to the reference database by engineering support for the integration of other, lesser known, reference sequence repositories.

A second consideration of a master reference database is usability. Bioinformatics is an interdisciplinary field comprised of researchers with varied backgrounds—from

computer science to biology. In order to maximize potential use by both skilled and novice computational users, this complete database would need an intuitive user interface.

The four fungal databases analysed in this study presented challenges to data access and manipulation. For example, JGI 1K asks the user to select the genomes of interest from a picture of the fungal tree of life, which can be unintuitive to many researchers. Adequate user support would also increase the usability of a comprehensive reference database; at the time of our study, Ensembl did not publish any contact information on their webpage.

Several reference databases highlight features that would be helpful to implement in an ideal master reference database. FungiDB, which had the most intuitive interface, simply asks the user to select data as though shopping online. To download all organisms, one only had to hover over "About FungiDB", click "Organisms" under "------ Data in FungiDB", click "add to basket". Once all the organisms are in the basket, it is possible to customize an annotation table containing download links for all references within the basket. While downloading data from NCBI can be challenging, once the user knows to select "Assembly" in the dropdown menu on the home page and type "Fungi" into the search bar, the filtering process becomes more intuitive.

A third consideration of a master reference database is maintenance support and archival stability. Maintaining a master reference sequence database would carry a substantial cost in terms of computational power and storage. An open source, continuous assembly approach would depend on support from an institution, governing body, or a global consortium.

Other areas of biological and genomic study currently benefit from comprehensive reference databases. For example, the Pathosystems Resource Integration Center (PATRIC) online bacterial reference database can be used as a gold standard for database website design. In PATRIC, all genomes for the selected organism taxa are present, and the filtering is intuitive. One drawback to the PATRIC website is the current protocol for downloading genomes; the best way to transfer data between servers is to download data locally, then manually transfer the data over to the server. A more efficient and usable method for downloading reference data would be to develop a dedicated FTP site (for researchers using the command line) or provide an option to utilize a data transfer service (such as Globus).

Our study indicates that the current approach to developing genomic reference databases for fungal species is not meeting the needs of metagenomics research. As metagenomic data becomes increasingly high-resolution, researchers need tools that enable a similarly high precision of taxonomic identification of DNA derived from samples. We believe that a systematic approach to developing a centralized master reference database will increase coverage and dramatically improve the quality and quantity of -omics research.

Declarations

Acknowledgements
Not Applicable

Funding
Not Applicable

Availability of Data and Materials
The data supporting the conclusions of this article, including the species and genera names, are available at: https://github.com/Mangul-Lab-USC/db.microbiome.

Competing interests
The authors declare that they have no competing interests.

Author contributions
C.L. and A.K. performed the analysis. C.L., L.M., D.K., and S.M. wrote the manuscript. S.M. conceived of the presented idea. E.E. supervised the project. All authors have read and agree to the content.

References
There should be no more than 10 references in a Comment

[1] Sharpton TJ. An introduction to the analysis of shotgun metagenomic data. Frontiers in plant science. 2014 Jun 16;5:209.

[2] Nasko DJ, Koren S, Phillippy AM, Treangen TJ. RefSeq database growth influences the accuracy of k-mer-based lowest common ancestor species identification. Genome biology. 2018 Dec;19(1):165.

[3] Hilton SK, Castro-Nallar E, Pérez-Losada M, Toma I, McCaffrey TA, Hoffman EP, Siegel MO, Simon GL, Johnson WE, Crandall KA. Metataxonomic and metagenomic approaches vs. culture-based techniques for clinical pathology. Frontiers in microbiology. 2016 Apr 7;7:484.

[4] Kersey PJ, Allen JE, Allot A, Barba M, Boddu S, Bolt BJ, Carvalho-Silva D, Christensen M, Davis P, Grabmueller C, Kumar N. Ensembl Genomes 2018: an integrated omics infrastructure for non-vertebrate species. Nucleic acids research. 2017 Oct 30;46(D1):D802-8.

[5] National Center for Biotechnology Information. https://www.ncbi.nlm.nih.gov. Accessed: 27 November 2018.

[6] Grigoriev IV, Nordberg H, Shabalov I, Aerts A, Cantor M, Goodstein D, Kuo A, Minovitsky S, Nikitin R, Ohm RA, Otillar R. The genome portal of the department of energy joint genome institute. Nucleic acids research. 2011 Nov 22;40(D1):D26-32.

[7] Basenko E, Pulman J, Shanmugasundram A, Harb O, Crouch K, Starns D, Warrenfeltz S, Aurrecoechea C, Stoeckert C, Kissinger J, Roos D. FungiDB: an integrated bioinformatic resource for fungi and oomycetes. Journal of Fungi. 2018 Mar;4(1):39.

[8] Nordberg H, Cantor M, Dusheyko S, Hua S, Poliakov A, Shabalov I, Smirnova T, Grigoriev IV, Dubchak I. The genome portal of the Department of Energy Joint Genome Institute: 2014 updates. Nucleic acids research. 2013 Nov 12;42(D1):D26-31.

{Figure Legends}

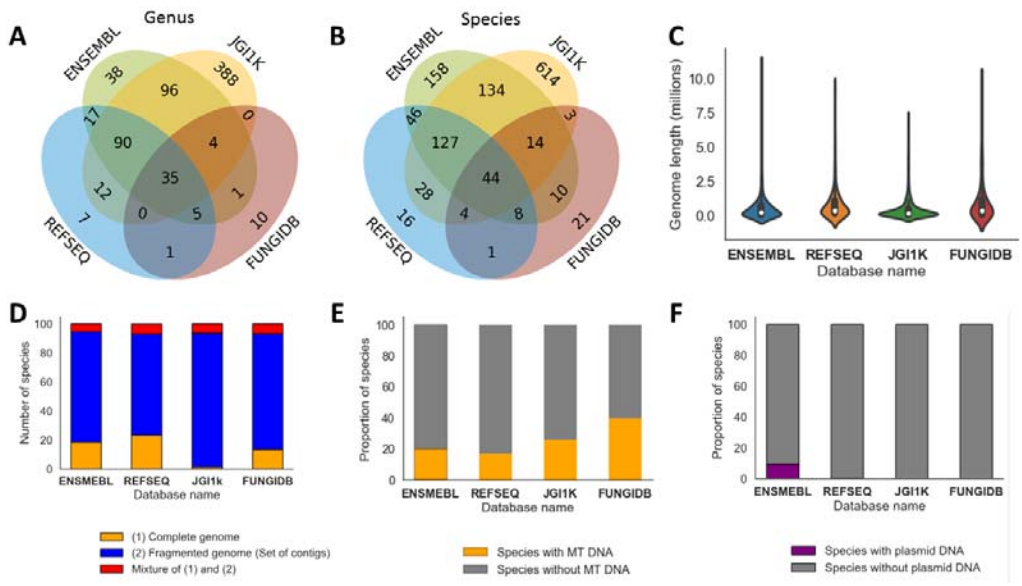
Figure 1.

Consensus of fungal genome representation across multiple reference databases.

(a) In total, there are 704 unique genera represented across four databases. Of this, 443 genera were found in only one of the databases, 127 genera in two databases, and 99 in three. Only 35 genera were identified in all databases. (b) In total there are 1228 unique species represented across four databases. Of this, 809 species can be found in only one of the four databases, 222 species in two databases, and 153 found across three databases. Only 44 species are represented in all four databases. (c) Length distribution of the fungal genomes (both chromosomes and contigs) across the databases. The combined chromosome and contig mean lengths are 843K (RefSeq), 464K (JGI 1K), 598K (Ensembl), and 947K (FungiDB). (d) Percentage of species per database available as complete genomes (orange), fragmented genomes (i.e., set of contigs) (blue), and a mixture of full chromosomes and contigs (red). (e) Percentage of species per database containing mitochondrial DNA (mtDNA) (orange). The percent of species that contained mitochondrial sequences are 17% (RefSeq), 25.5% (JGI 1K), 19.5% (Ensembl), and 39.6% (FungiDB). (f) Percentage of species per database containing plasmid DNA (violet). These percentages are 0.36% (RefSeq), 0.09% (JGI 1K), 9.5% (Ensembl), and 0.0% FungiDB.

Figures

Figure 1



Tables

Not Applicable

Supplementary Materials

Part I: Downloading the databases

Our study considered fungal species and genera across four reference databases:

- JGI 1000 Fungal Genomes Database (JGI 1K), https://genome.jgi.doe.gov/programs/fungi/index.jsf
- Ensembl, http://fungi.ensembl.org/index.html
- RefSeq, https://www.ncbi.nlm.nih.gov/
- FungiDB http://fungidb.org/fungidb/

Each reference database had a different process for downloading fungal reference genomes:

- **JGI 1K Fungal Genomes Database.** We locally downloaded the assembled masked fungal reference database, which appeared in the "download" section of the website as a zipped file. The unzipped file yielded 1063 directories. Each directory represented one species or, when strain information was available, one strain. The contents of each directory included a zipped FASTA file (in two directories there were two such files) that contained the genetic

reference information. Plasmid and mitochondrial sequences were available in separate files.

- **Ensembl.** We could not identify an efficient GUI with which to download all 811 fungal reference files available on the website. Each DNA FASTA link led to an FTP page with multiple gzip files. We only selected the files that ended in "dna.toplevel.fa.gz". For each of the links that were selected for the 811 available fungal references in Ensembl, a wget command was called.
- **RefSeq.** First, we downloaded the table of available fungal reference genomes from the following FTP address: ftp://ftp.ncbi.nlm.nih.gov/genomes/refseq/fungi/assembly_summary.txt. We then extracted the URLs from the table, and we used wget to download the corresponding FASTA files that contained the reference genomes.
- **FungiDB.** Once all the fungal reference genomes were placed into the online basket, we downloaded the links corresponding to FASTA files. We then created a bash file that called "wget" on each link.

Scripts and commands used to download the reference databases are freely available at https://github.com/smangul1/db.microbiome.

Part II: Standardizing the names of the species across the fungal reference databases

In order to standardize the names of species across all four fungal reference databases, universal taxIDs were used in place of scientific names. The taxIDs provided by the databases indicate the strain or species of an identified fungal specimen. Only species- and genus-level taxIDs were used to quantify the consensus of fungal genome representation across the four databases. We did not analyze the consensus of strain-level IDs. We used the Ete3 module to assign a species-level taxID, when the database provided a strain-level taxonomic identification, and to obtain a genus-level ID for all reference genomes. We encountered four reference genomes that had not been assigned a genus and lacked a genus-level ID; in the data, such reference genomes indicated 'no rank' where the genus-level ID would have appeared. In such cases, we used the unranked ID as the genus ID.

As with the procedures for downloading reference genomes, the processes for obtaining corresponding taxIDs for each file were different for each of the four databases.

- **JGI 1K Fungal Genomes Database.** We followed the six-step process for obtaining a Microsoft Excel document with the taxIDs. First, we created an advanced search that produced a "reports" button where the user can download the taxID information via an Excel spreadsheet. Second, we prepared a Python script to automatically match filenames with corresponding taxIDs. The Python script is available at https://github.com/cloeffler745/Database_source_code/tree/changing_compare_files/Compare/Fungus/fix_onek_csv
- **Ensembl.** The FTP server (ftp://ftp.ensemblgenomes.org/pub/fungi/release-40/) offers a file named "species_EnsemblFungi.txt," a mapping file that shows which Ensembl files match to corresponding taxonomic IDs. We ran a Python script that used the mapping file and a list of Ensembl file names to assign taxIDs (https://github.com/cloeffler745/Database_source_code/blob/changing_compare_files/Compare/Fungus/ensmebl_prep/new_make_ensembl_taxid_table.py).
- **RefSeq.** The FASTA headers within each reference genome file contained the species names. We isolated these names into a text file, which we used to obtain taxIDs via the same NCBI taxonomy browser (https://www.ncbi.nlm.nih.gov/Taxonomy/TaxIdentifier/tax_identifier.cgi).
- **FungiDB.** After creating an account on the website and downloading all the reference species, a csv file could be custom generated by clicking the "download" link, selecting "choose columns", and checking the "NCBI taxon ID" box under "taxonomy".

Part III: Classify reference genome as complete or incomplete

In order to determine the presence or absence of genetic reference types (e.g., scaffolds, contigs, fully assembled chromosomes) and extra genetic references (e.g., mitochondrial and plasmid sequences), we searched the text of the reference files for predetermined patterns and words. We used the key words "chromosome" and "chr" to identify sequences that had been marked as complete genomes. The key words

“contig”, “scaffold”, and “sca” were used to identify sequences that had been marked as incomplete genomes.

Part IV: Compare the species and genera across the fungal reference databases
In order to generate statistical data for cross-database species comparison, we extracted individual sequence attributes from each fungus file. We stored these attributes in a structured query language relational database management system (SQL RDBM). Attributes extracted from each fungal reference sequence included database name; species-level taxID; genus-level taxID; species name; genus name; a flag indicating species composition (e.g., chromosomes, contigs or mixture of both); a flag indicating if the species contains mitochondrial and plasmid DNAs.

For each species, we also recorded the length of contigs and chromosomes. Individual files could have more than one sequence classification depending on the contents of the DNA sequences. The data for sequence composition contained the number of sequences for a given sequence classification that existed within each file. We also stored within each file the average, minimum, and maximum sequence lengths for each sequence classification.

Due to variation in file formatting and naming conventions inherent to each database, we implemented several flags to determine sequence classifications. For example, the keywords “scaffold” and “contig” were assigned to catch instances of contig classified sequences. Variations of the keyword “chromosome” such as “chrom” and “chr” were used to identify instances of chromosome classified sequences. Variations of the keyword “mitochondria,” such as “mitochondrion,” were searched to tag instances of mitochondria classified sequences. We included the keyword “plasmid” to identify instances of plasmid classified sequences. A link to the SQL database used in this study can be found here:
https://github.com/smangul1/db.microbiome/blob/master/Fungi/data/refSeqFungiStats.db

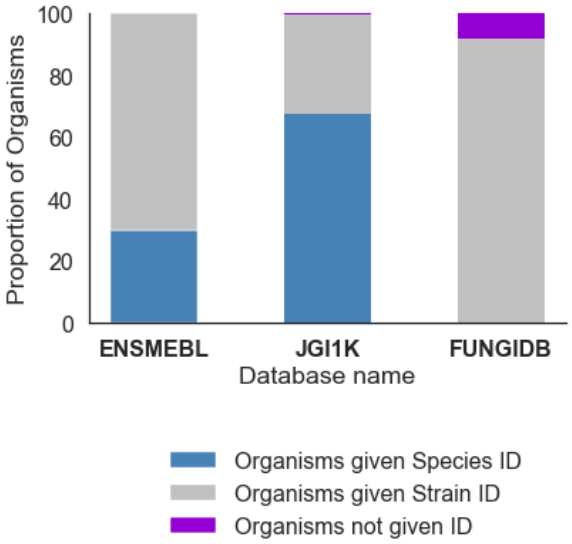


Figure S1. Proportion of references that were identified by species-level or strain-level taxonomic ID, or were missing a taxonomic ID across three of the four databases. Species-level taxonomic IDs were given to 29.6% of references in Ensembl, 67.5% in JGI 1K, and 0.0% in FungiDB. Strain-level taxonomic IDs were given to 70.4% of references in Ensembl, 32.1% in JGI 1K, and 91.8% in FungiDB. Taxonomic IDs were invalid or missing from 0.5% of references in JGI 1K and 8.2% in FungiDB. RefSeq taxonomic IDs were discarded because they were manually curated.

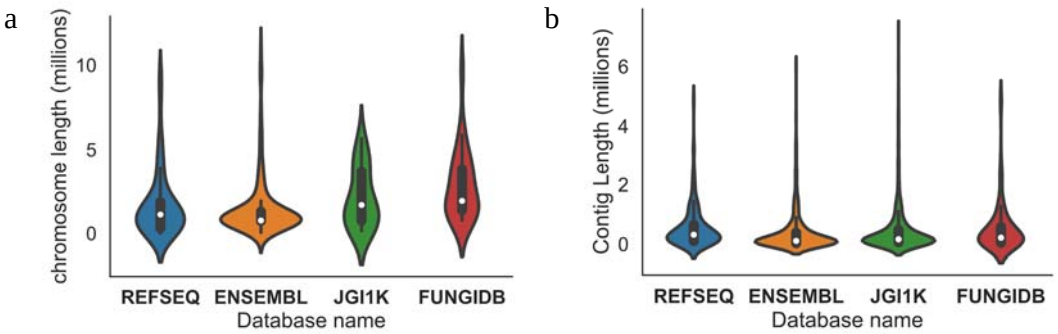
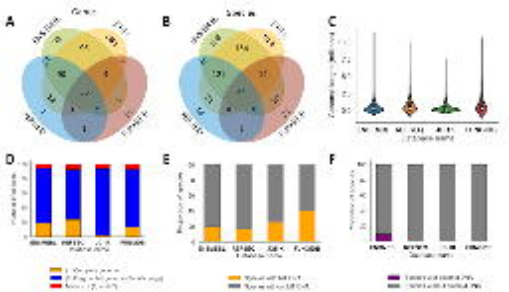
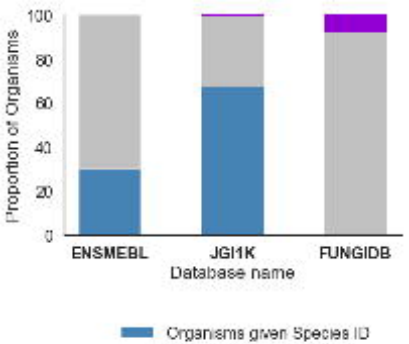
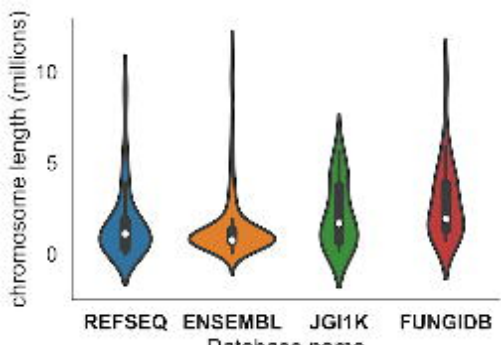
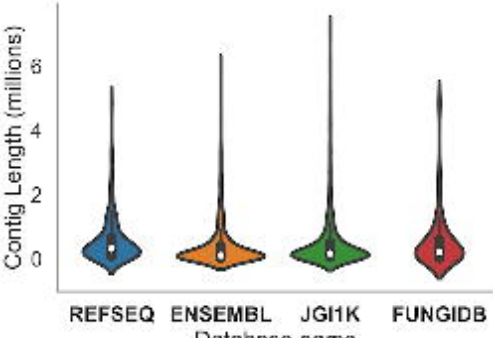


Figure S2. Length distribution of the fungal genomes (both chromosomes and contigs) across the databases. (a) The mean lengths of chromosomes are 1.6M (RefSeq), 2.1M (JGI 1K), 1.4M (Ensembl) and 2.8M (FungiDB). (b) The mean lengths of the contigs are 535K (RefSeq), 433K (JGI 1K), 348K (Ensembl), and 505K (FungiDB).


A
B
C
D
E
F


100
80
60
40
20
0
Proportion of Organisms
ENSMEBL
JGI1K
FUNGIDB
Database name
Organisms given Species ID
Organisms given Strain ID
Organisms not given ID


chromosome length (millions)
10
5
0
REFSEQ
ENSEMBL
JGI1K
FUNGIDB
Database name


Contig Length (millions)
0
2
4
6
REFSEQ
ENSEMBL
JGI1K
FUNGIDB
Database name