Title: Analysis of multiple fungal sequence repositories highlights shortcomings in microbial databases

Caitlin Loeffler[1]*, Aaron Karlsberg[1]*, Eleazar Eskin[1], David Koslicki[2], Serghei Mangul[1,3]

[1] Department of Computer Science, University of California Los Angeles, USA

[2] Department of Mathematics, Oregon State University, USA

[3] Institute for Quantitative and Computational Biosciences, University of California Los Angeles, USA

*- These authors contributed equally to this work

Abstract

Reference genomes are essential for metagenomics studies, which require comparing short metagenomic reads with available reference genomes to identify organisms within a sample. Current efforts promise to extend genomic representation of microbial organisms across the tree of life. However, new efforts are not always integrated into existing databases. Lack of comprehensive microbial databases complicate the choice of which database to use and potentially results in failing to classify portions of a sample due to missing database organisms. To illustrate, we have considered fungal reference genomes to estimate their consistency and comprehensiveness across various databases. The overlap of genera and species in the databases analyzed was alarmingly small. In other words, using only a single reference database for analysis of metagenomic samples possibly results in the failure to identify some organisms in a sample. We have identified an emerging need to integrate and disambiguate such

databases. The current state of microbial databases is hampering metagenomic research and it is time to establish an effective dialogue between all the parties involved in creating microbial databases, where any new database will incorporate the information from the previous efforts in a consistent manner to avoid discrepancies between the databases and hence aid in metagenomic tool development.

Introduction

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 superior when compared to traditional, culture-based techniques and is incapable of capturing the complex interactions that take place between thousands of different microbial organisms in their natural habitats, though at the cost of increased expense. Reference genomes are essential for metagenomics studies, which require comparing short metagenomic reads with available reference genomes to identify species within a sample[2].

Recent advances in metagenomics promise to extend the reference database of microbial organisms across the tree of life[3,4]. However, newly constructed reference databases often do not contain previously discovered reference genomes. Lack of comprehensive

databases, which include all discovered microbial genomes, complicate the choice of which database to use in a metagenomic analysis. Unstandardized databases may result in a failure to classify portions of the sample due to missing references within the chosen database[5].

In order to assess the potential usability of fungal reference databases in a metagenomic analysis, we performed a comprehensive analysis of four fungal genome reference databases. Our focus was on the organization and accessibility of nucleotides sequences only. We considered fungal reference databases as a case study to estimate concordance across various reference databases. In particular, we estimated the fraction of fungal genomes present in only one of the databases. Our results show an alarmingly low concordance of species and genera across these reference databases. We observed that a majority of the species and genera were present in only one database, making the use of a single database insufficient for metagenomics research. Using multiple databases at once could improve coverage, but combining databases requires disambiguating and deduplicating species and genera. There are currently no methods capable of merging disjoint reference databases.

Results

We considered fungal species and genera across four reference databases: Ensembl[6], RefSeq[7], JGI's 1000 fungal genomes project (JGI 1K)[8], and FungiDB[9]. We used universal taxonomic IDs from NCBI to match the species and genera across the databases. The taxonomic rank of provided universal taxonomic ID's varied across individual references in JGI 1K and

Ensembl. Either species or strain level taxonomic ID's were used with no clear indication of which was which. FungiDB used almost exclusively strain level taxonomic IDs, except in 11 cases where no ID was given (**Figure S2**). We found species and genus level IDs for all references, which were used for analysis (See Methods for more detail). We identified that 63% of the genera were present only in a single database (**Figure 1a**). On the species level, we observed a larger discrepancy across databases with 66% of species being present only in a single database (**Figure 1b**). The Methods Section describes the details of this analysis along with links to accompanying code.

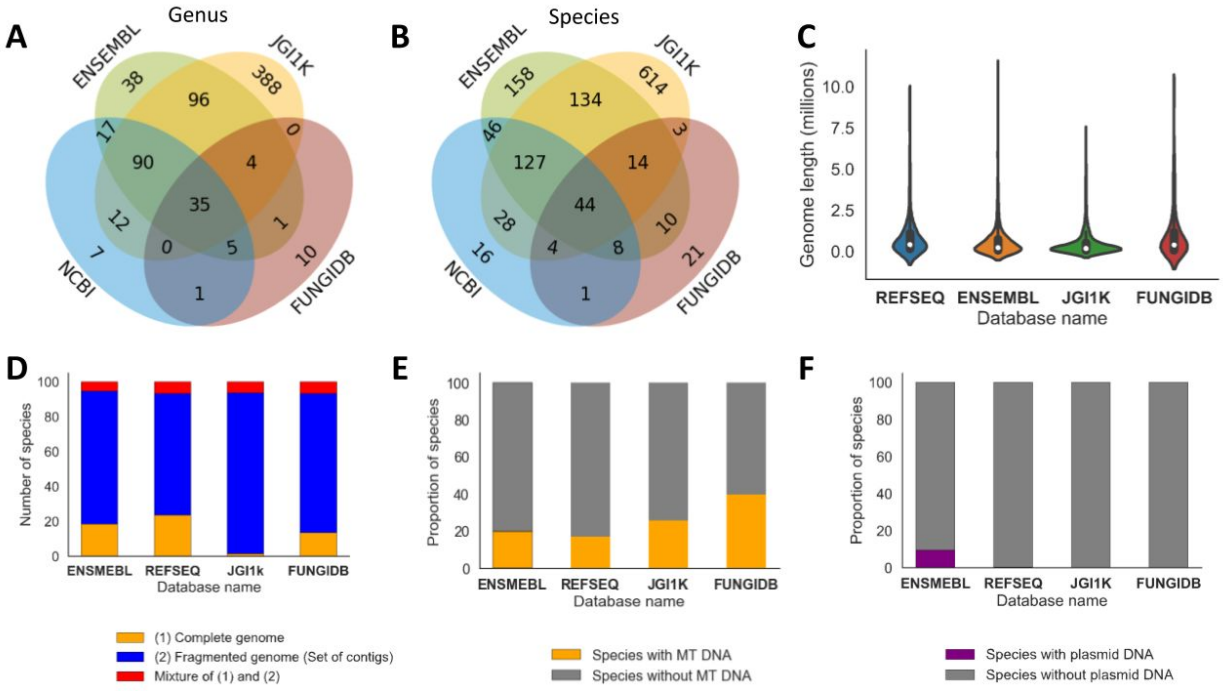


Figure 1. Consensus of fungal genome representation across multiple reference databases. **(a)** In total, there are 704 unique genera represented across four databases. Of this, 443 genera were found in only one of the databases, 127 genera in two databases, and 99 in three. Only 35 genera were identified 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.

We investigated the lengths of the fungal genomes across the four databases considered in this study (**Figure 1c**). We observed a shorter length of genomes in the JGI 1K database. The shorter overall length of genomes in the JGI 1K database can be attributed to the limitation of current assembly algorithm often not able to assemble the full length microbial genomes. We also separately investigated the length of complete and incomplete genomes (represented as a set of contigs). As expected, we observed a greater length of complete genomes compared to incomplete ones consistently across all four databases (**Figure S1**). The percent of species represented as complete genomes are 23.5% (RefSeq), 1.4% (JGI 1K), 18.5% (Ensemble), and 13.4% (FungiDB). The percentage of species represented as contigs are 69.6%

(RefSeq), 92.2% (JGI 1K), 76.2% (Ensembl), and 79.9% (FungiDB). The percentage of species containing both contigs and complete chromosomes are 6.8% (RefSeq), 6.3% (JGI 1K), 5.3% (Ensembl), and 6.7% (FungiDB) (**Figure 1d**). Analysis of all four databases revealed that some species were represented as a mixture of complete and incomplete genomes (**Figure 1d**). Additionally, for the same species, RefSeq, Ensembl, and FungiDB had mitochondrial reference genomes in the same files as the reference genome (**Figure 1e**). At the same time, none of the complete and incomplete genomes in the JGI 1K database were annotated as mitochondrial reference genomes. JGI 1K did, however, have mitochondrial sequences represented in 271 separate files meaning that mitochondrial sequences were available for 25.5% of references. Finally, 9.5% of references in Ensembl contained plasmid genomes as did 0.36% of references in RefSeq. JGI 1K had one separate file containing plasmid genomes, and FungiDB did not contain any plasmid genomes (**Figure 1f**).

In addition to the discrepancies between fungal reference databases, we identified numerous issues that limit the usability of the databases. Namely the lack of an easy-to-use interface to download the genome references. Also, some databases provided only limited, difficult to find, documentation. For example, obtaining the universal taxonomic ID's for the JGI 1K fungal genomes was a non-intuitive process involving six steps. Overall, the lack of user-friendly interfaces and inconsistent use of unique identifiers in reference databases requires substantial time and effort from the user.

Discussion

This study is the first to systematically investigate the consistency of fungal databases. We determined that discrepancies between the fungal reference databases are alarmingly large. In the best case scenario, a researcher only using one database will be missing 21% of the reference fungal species. We recognize that the JGI 1K database contains a number of previously unpublished, novel genomes, and the publication introducing JGI 1K indicates that it was not originally designed to be used as a metagenomics reference database[8]. However, the existence of such novel genomes in JGI 1K motivates a comparison to NCBI, Ensembl, and FungiDB databases and benefits metagenomics researchers who wish to build or take advantage of complete reference databases. Additionally, we understand that other databases may be limited in scope or funding and cannot create a complete database. This unfortunate state of fungal databases perhaps explains the general lack of fungal organisms in many metagenomic analysis tools, the absence of which stalls metagenomic discoveries centering around the Fungal Kingdom. Furthermore, since the fungal reference genomes are from databases that also contain reference genomes for bacteria and other organisms, it is likely that these issues extend to general microbial databases as well. Establishing, between all parties involved, an effective dialogue centered on systematic creation of microbial databases promises to accelerate metagenomics discoveries. In order to optimize metagenomic tool development, any centralized database should consistently incorporate information from previous efforts to avoid introducing discrepancies between the databases. Emergent long read technologies can assist in assembling longer contigs and eventually may obtain full-length genomes[10]. Implementing systematic reference databases today will improve the outcome of

these efforts. It is important to impose stringent standards on the way reference microbial databases are organized and distributed, as has been successfully initiated for vertebrate genomes[11].

Methods

Downloading the databases

We 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 of these had a separate process for downloading the fungal reference genomes:

- **JGI 1K Fungal Genomes Database**. On the download page, we downloaded only the assembled masked fungal reference database. This appeared as a zip file, which was downloaded locally. When unzipped, the file yielded 1063 directories, each representing one species (in some cases where strain information was available, each strain was represented within its own directory), inside of each was a zipped FASTA file (in 2 directories there were 2 such files) which contained the genetic reference information. Plasmid and mitochondrial sequences can be found in separate files.
- **Ensembl.** There was no efficient GUI with which to download all 811 fungal reference files on the site. Each DNA FASTA link led to a FTP page with multiple gzip downloads. Only the files that ended in “dna.toplevel.fa.gz” were selected. A wget command was

called on each of the links that were selected for the 811 available fungal references in Ensembl.

- **RefSeq.** First, we have downloaded the table of available fungal reference genomes from: ftp://ftp.ncbi.nlm.nih.gov/genomes/refseq/fungi/assembly_summary.txt. We extracted the URLs from the table, and used wget to download the corresponding FASTA files with the reference genomes.
- **FungiDB.** Once all the reference genomes were placed into the online basket, we downloaded the list of FASTA download links and created a bash file that called "wget" on each link.

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

Standardize the names of the species across the fungal reference databases

In order to standardize the names of the species across all four fungal reference databases, universal taxonomic IDs were used in place of scientific names. The taxonomic IDs provided by the databases were either strain level or species level identifiers. The species taxonomic IDs were used to identify which species were present in one or more databases. The Ete3 module[12] was used to find species level taxonomic ID when strain level was given, and get genus level IDs for all references. In finding genus level IDs, we found that four references have not been placed into a genus, thus have no genus level ID. Instead they have a taxonomic ID indicated as 'no rank' where the genus level ID should be. This unranked ID was used as the

genus ID in these cases. Only species and genus level taxonomic IDs were used to quantify the consensus of fungal genome representation across the four databases. Strain level IDs themselves were not analyzed.

As with downloading, the process for initially obtaining corresponding taxonomic IDs for each file was different for each of the four databases.

- **JGI 1K Fungal Genomes Database** There was a six-step process necessary to obtain a Microsoft Excel document that contained taxonomic ID information that involved making an advanced search which reveals a "reports" button where the user can download an excel spreadsheet containing the taxonomic ID information. We have prepared a python script to match filenames to taxonomic IDs. The script is available at https://github.com/cloeffler745/Database_source_code/tree/changing_compare_files/Compare/Fungus/fix_onek_csv
- **Ensembl** On the FTP server (ftp://ftp.ensemblgenomes.org/pub/fungi/release-40/) is a file named "species_EnsemblFungi.txt" which is a mapping file that shows what Ensembl files match to which taxonomic IDs. A list of Ensembl file names and the mapping file were used to match taxonomic IDs to file names using Python (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 file contained the species names. These names were isolated into a text file which was used to get taxonomic IDs on the same NCBI

taxonomy browser (https://www.ncbi.nlm.nih.gov/Taxonomy/TaxIdentifier/tax_identifier.cgi).

- **FungiDB** Once an account had been created and all the reference species had been placed into the users basket, a csv file could be custom generated by clicking the "download" link, clicking "choose columns", and checking the "NCBI taxon ID" box under "taxonomy".

Classify reference genome as complete or incomplete

In order to determine the presence or absence of genetic reference types (scaffolds, contigs, fully assembled chromosomes) and extra genetic references (mitochondrial and plasmid sequences), the text of the reference files was searched for predetermined patterns and words. The key words "chromosome" and "chr" were used to identify sequences that were marked as complete genomes. The key words "contig", "scaffold", and "sca" were used to identify sequences marked as incomplete.

Compare the species and genera across the fungal reference databases

To generate statistical data for cross-database species comparison, individual sequence attributes were extracted from each fungus file and stored in a structured query language relational database management system (SQL RDBM). The attributes extracted from each fungal reference sequence included database name, species level taxonomic ID, genus level taxonomic ID, species name, genus name, a flag indicating if the species is composed of chromosomes, contigs or mixture of both, a flag indicating if the species contains mitochondrial

and plasmid DNAs. We have also recorded the length of contigs and chromosomes for each of the species. Individual files could have more than one sequence classification depending on the contents of the DNA sequences within. The data for sequence composition contained the number of sequences for a given sequence classification that existed within each file. Furthermore, the average, minimum, and maximum sequence lengths for each sequence classification within each file were also stored. Due to the variation in file formatting and naming conventions within each database, several flags were implemented to determine sequence classifications. In particular, the keywords “scaffold” and “contig” were used to catch instances of contig classified sequences. Variations of the keyword “chromosome” such as “chrom” and “chr” were used to catch instances of chromosome classified sequences. Variations of the keyword “mitochondria” such as “mitochondrion”, were used to catch instances of mitochondria classified sequences. The keyword “plasmid” was used to catch instances of plasmid classified sequences. A link to the SQL database can be found here:

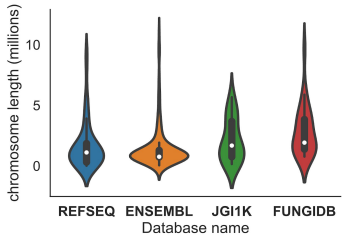
https://github.com/smangul1/db.microbiome/blob/master/Fungi/data/refSeqFungiStats.db

Data Location

The data used in this study, including the species and genera names, are available here:
https://github.com/smangul1/db.microbiome/tree/master/Fungi/data

Supplementary figures

(a)



(b)

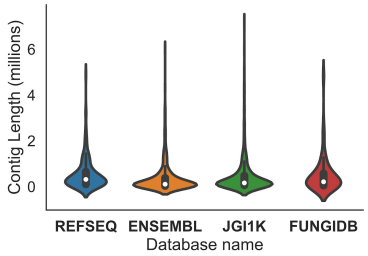


Figure S1. Length distribution of the fungal genomes (both chromosomes and contigs) across the databases. **(a)** The mean lengths of chromosomes are 1.6M (RefSeq), 2.1M (JGI 1K), 1.4M (Ensembl) and 2.8M (FungiDB). **(b)** The mean lengths of the contigs are 535K (RefSeq), 433K (JGI 1K), 348K (Ensembl), and 505K (FungiDB).

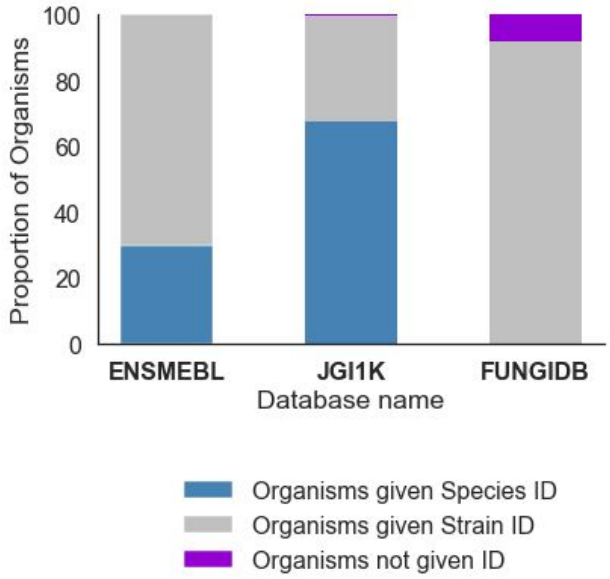


Figure S2. Proportion of references that were identified by species level taxonomic ID, strain level taxonomic ID, or were missing a taxonomic ID across three of the four databases. Species level taxonomic IDs were given to 29.6% of references in Ensembl, 67.5% in JGI 1K, and 0.0% in FungiDB. Strain 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 curated manually, and are therefore not included.

References

1. Sharpton, T. J. An introduction to the analysis of shotgun metagenomic data. - PubMed - NCBI. Available at: https://www.ncbi.nlm.nih.gov/pubmed/24982662. (Accessed: 15th February 2019)
2. Nasko, D. J., Koren, S., Phillippy, A. M. & Treangen, T. J. RefSeq database growth influences the accuracy of k-mer-based lowest common ancestor species identification. *Genome Biol.* **19**, 165 (2018).
3. Parks, D. H. *et al.* Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nat Microbiol* **2**, 1533–1542 (2017).
4. Hug, L. A. *et al.* A new view of the tree of life. *Nature Microbiology* **1**, 16048 (2016).
5. Sczyrba, A. *et al.* Critical Assessment of Metagenome Interpretation—a benchmark of metagenomics software. *Nat. Methods* **14**, 1063 (2017).
6. Kersey, P. J. *et al.* Ensembl Genomes 2018: an integrated omics infrastructure for non-vertebrate species. *Nucleic Acids Res.* **46**, D802–D808 (2018).
7. National Center for Biotechnology Information. Available at: https://www.ncbi.nlm.nih.gov/. (Accessed: 27th November 2018)
8. Grigoriev, I. V. *et al.* The Genome Portal of the Department of Energy Joint Genome Institute. *Nucleic Acids Res.* **40**, D26–32 (2012).
9. Basenko, E. Y. *et al.* FungiDB: An Integrated Bioinformatic Resource for Fungi and Oomycetes. *J Fungi (Basel)* **4**, (2018).

10. Diaz-Viraque, F. *et al.* Nanopore sequencing significantly improves genome assembly of the eukaryotic protozoan parasite Trypanosoma cruzi. (2018). doi:10.1101/489534

11. A reference standard for genome biology. *Nat. Biotechnol.* **36**, 1121 (2018).

12. Huerta-Cepas, J., Serra, F. & Bork, P. ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. *Mol. Biol. Evol.* **33**, 1635–1638 (2016).