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[1,4]

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

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

[4] 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. We analyzed the current state of fungal reference databases to assess their usability as reference databases for metagenomic studies. 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 the sample. Communication between database developers needs to be established to create a set of standards for the way reference databases are organized and distributed.

Introduction

High-throughput sequencing has revolutionized microbiome research by enabling the detection of thousands of microbial genomes directly from their host environments. This approach, known as metagenomics, is superior when compared to traditional, culture-based techniques which are incapable of capturing the complex interactions that take place between the thousands of different microbial organisms in their natural habitats. Reference genomes are essential for metagenomics studies, which require comparing short metagenomic reads with available reference genomes to identify species within a sample[1].

Recent advances in metagenomics promise to extend the reference database of microbial organisms across the tree of life[2]. 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.

In order to assess the potential usability of fungal reference databases in a metagenomic analysis, we performed a comprehensive analysis of three fungal genome reference databases. 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 database into a single one.

Results

We considered fungal species and genera across three reference databases: Ensembl[3], RefSeq[4], and JGI's 1000 fungal genomes project (1K)[5]. We used universal taxonomic IDs from NCBI to match the species and genera across the databases. We identified 64% of genera that were present only in a single database (**Figure 1a**). On the species level, we observed an even larger discrepancy across databases with 77% of species being present only in a single database (**Figure 1b**). The Methods Section describes the details of this analysis along with accompanying code.

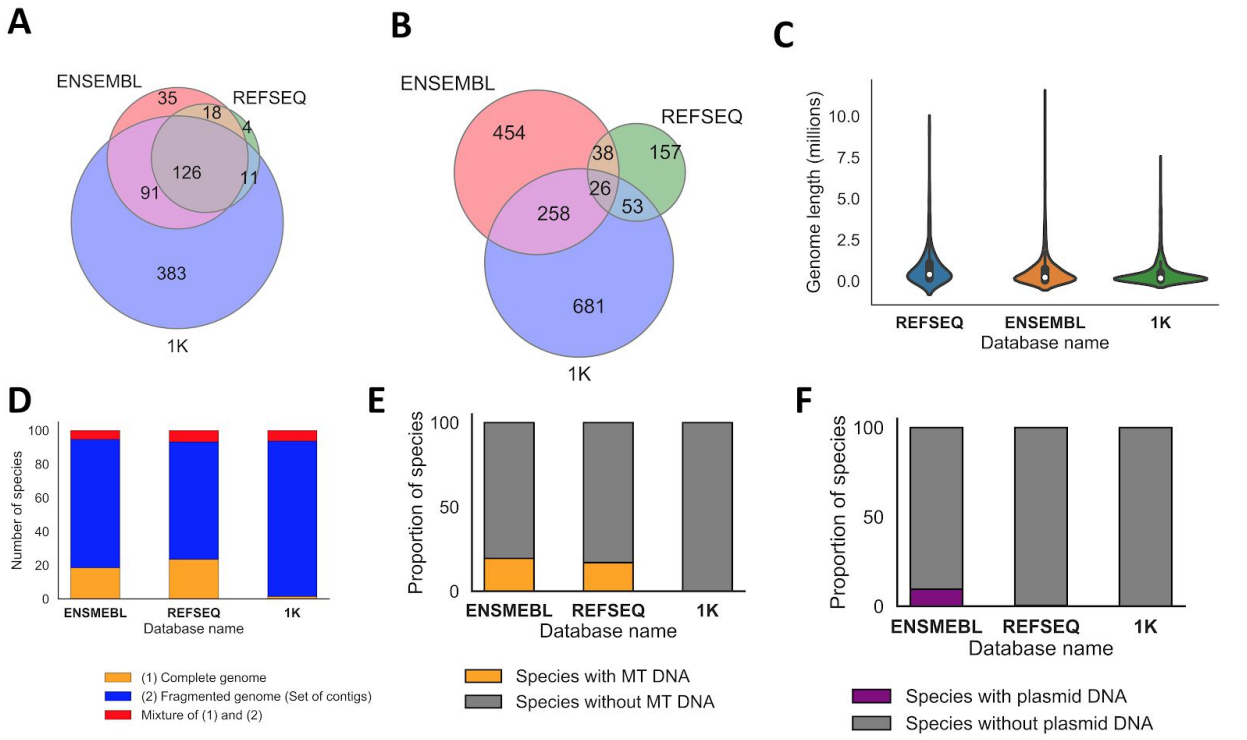


Figure 1. Consensus of fungal genome representation across multiple reference databases. **(a)** In total, there are 668 unique genera represented across three databases. Of this, 422genera were found in only one of the databases and 120genera in two databases. Only 126 genera were identified in all databases. **(b)** In total there are 1667 unique species represented across three databases. Of this, 1292 species can be found in only one of the three databases and 349 species in two databases. Only 26 species are represented in all three databases. **(c)** Length distribution of the fungal genomes (both chromosomes and contigs) across the databases. **(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), 0.0% (JGI), and

19.5% (Ensembl). **(f)** Percentage of species per database containing plasmid DNA (violet). These percentages are 0.36% (RefSeq), 0.0% (JGI), and 9.5% (Ensembl).

We investigated the lengths of the fungal genomes across the three databases considered in this study (**Figure 1c**). We observed a shorter length of genomes in the 1K database. The shorter overall length of genomes in the 1K database can be attributed to the lower amount of complete genomes in the 1K database compared to other databases. 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 three databases (**Figure S1**). The percent of species represented as complete genomes are 23.5% (RefSeq), 1.4% (JGI), and 18.5% (Ensemble). The percentage of species represented as contigs are 69.6% (RefSeq), 92.2% (JGI), and 76.2% (Ensembl). The percentage of species containing both contigs and complete chromosomes are 6.8% (RefSeq), 6.3% (JGI), and 5.3% (Ensembl) (**Figure 1d**). Analysis of all three databases revealed that some species were represented as a mixture of complete and incomplete genomes (**Figure 1d**). Additionally, for the same species, RefSeq and Ensembl had mitochondrial reference genomes (**Figure 1e**). At the same time, none of the complete and incomplete genomes in the 1K database were annotated as mitochondrial reference genomes. Finally, Ensembl and RefSeq contained plasmid references while 1K did not (**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 limited documentation for the references that were difficult to find. 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

Our 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 38% of the reference fungal species. 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 new database should consistently incorporate information from previous efforts to avoid introducing discrepancies between the databases. Current emergent long read technologies promise to promise to deliver reads helping to assemble

longer contigs and eventually obtaining full-length genomes[7]. 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[8].

Methods

Downloading the databases

We considered fungal species and genera across three reference databases:

- JGI 1000 Fungal Genomes Database, https://genome.jgi.doe.gov/programs/fungi/index.jsf
- Ensembl, ftp://ftp.ensemblgenomes.org/pub/fungi/release-40/
- RefSeq, https://www.ncbi.nlm.nih.gov/

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.
- **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 .gz downloads. Only the file that ended in dna.toplevel.fa.gz was selected. The wget command was called on each of the links that were selected for the 811 available fungal references on Ensembl.

- **RefSeq.** First, we have downloaded the table of available fungal reference genomes from here ftp://ftp.ncbi.nlm.nih.gov/genomes/refseq/fungi/assembly_summary.txt. We have extracted the URLs from the table, and use wget to download corresponding FASTA files with the reference genomes.

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 three fungal reference databases, universal taxonomic IDs were used. Once the taxonomic ID for each file was determined, the species name was found by converting the taxonomic ID to a species name on NCBI (https://www.ncbi.nlm.nih.gov/Taxonomy/TaxIdentifier/tax_identifier.cgi). The taxonomic IDs provided species-level information. As with downloading, the process for which taxonomic IDs needed to be determined for each file was different for each of the three 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 site that was used to get species names from taxonomic IDs.

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, TAXID, 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:

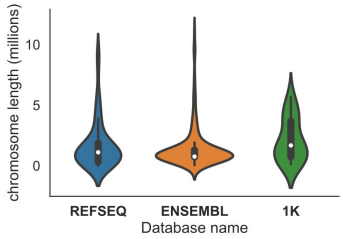
https://github.com/aaronkarlsberg/db.microbiome/blob/master/Fungi/data/refSeqFungiStats.db

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)

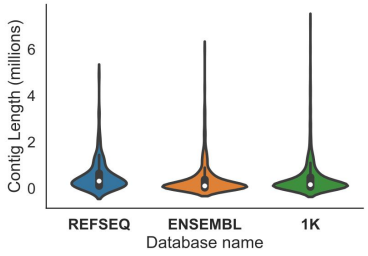


Figure S1. Length distribution of the fungal genomes (both chromosomes and contigs) across the databases. The mean lengths of chromosomes are 1.6M (RefSeq), 2.1M (JGI), 1.4M (Ensembl). The mean lengths of the contigs are 535K (RefSeq), 433K (JGI), 348K (Ensembl). The combined chromosome and contig mean lengths are 843K (RefSeq), 464K (JGI), 598K (Ensembl).

References

1. 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).
2. 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).
3. Website. Available at: http://fungi.ensembl.org/index.html). (Accessed: 27th November 2018)
4. National Center for Biotechnology Information. Available at: https://www.ncbi.nlm.nih.gov/. (Accessed: 27th November 2018)
5. JGI Fungi Portal - Home. Available at: https://genome.jgi.doe.gov/programs/fungi/index.jsf. (Accessed: 27th November 2018)
6. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25,** 3389–3402 (1997).
7. Diaz-Viraque, F. *et al.* Nanopore sequencing significantly improves genome assembly of the eukaryotic protozoan parasite Trypanosoma cruzi. (2018). doi:10.1101/489534
8. A reference standard for genome biology. *Nat. Biotechnol.* **36,** 1121 (2018).