Placeholder for publisher logo

Journal of XYZ, 2021, 1–14

doi: xx.xxxx/xxxx
Manuscript in Preparation
Research

RESEARCH

Pathogenic potential prediction for novel fungal DNA based on a curated fungi-hosts data collection

Jakub M. Bartoszewicz[1,2,3,*,†], Ferdous Nasri[1,2,3,†], Melania Nowicka[1,2,3] and Bernhard Y. Renard[1,2,*]

[1]Data Analytics and Computational Statistics, Hasso Plattner Institute for Digital Engineering, 14482 Potsdam, Germany and [2]Digital Engineering Faculty, University of Potsdam, 14482 Potsdam, Germany and [3]Department of Mathematics and Computer Science, Free University of Berlin, 14195 Berlin, Germany.

[*]jakub.bartoszewicz@hpi.de, bernhard.renard@hpi.de
[†]Contributed equally.

Abstract

Background: Emerging pathogens are a growing threat, but large data collections and approaches for predicting the risk associated with novel agents are limited to bacteria and viruses. Pathogenic fungi, which also pose a constant threat to public health, remain understudied. Relevant, curated data remains comparatively scarce and scattered among many different sources, hindering the development of sequencing-based detection workflows for novel fungal pathogens. No prediction method working for agents across all three groups is available, even though the cause of an infection is often difficult to identify from symptoms alone.
Results: We present a curated collection of fungal host range data, comprising records on human, animal and plant pathogens, as well as other plant-associated fungi, linked to publicly available genomes. We show that the resulting database can be used to predict the pathogenic potential of novel fungal species directly from DNA sequences with either sequence homology or deep learning. We develop learned, numerical representations of the collected genomes and show that the human pathogens are separable from non-human pathogens. Finally, we train multi-class models predicting if next-generation sequencing reads originate from novel fungal, bacterial or viral threats.
Conclusions: The presented data collection enables accurate detection of novel pathogens from sequencing data. It is also a comprehensive resource that can find use beyond this particular task. This can include possible applications in proteomics and genomics, employing both machine learning and direct sequence comparison.
Availability: The database and models are hosted at https://zenodo.org/record/5711852 and https://zenodo.org/record/5711877. Source code is available at https://gitlab.com/dacs-hpi/deepac.

Key words: Fungi-host interactions; data; pathogens; next-generation sequencing; machine learning

Introduction and Background

Many species of fungi are dangerous plant, animal, or human pathogens. Importantly, even usually harmless opportunists can be deadly in susceptible populations. For example, *Candida albicans* causes common, relatively benign infections like thrush and vulvo-vaginal candidosis, affecting up to 75% of women at least once in their lifetime and often re-occurring multiple times [1]. It is also frequently found in healthy humans without leading to any disease, and has been reported to be capable of stable colonization [2]. However, invasive *Candida* infections, especially bloodstream infections, can reach mortality rates of up to 75%, rivaling those of bacterial and viral sepsis [3]. A related species, *Candida auris*, has been first recognized in a human patient in 2009 [4] and quickly became one of the most urgent threats among the drug-resistant pathogens [5], reaching mortality rates of up to 60% [6]. It might have originally been a plant saprophyte which has adapted to avian, and then also mammalian hosts, possibly prompted by climate change [7]. Strik-

Compiled on: November 30, 2021.
Draft manuscript prepared by the author.

ingly, it seems to have emerged in three different clonal populations on three continents at the same time, for reasons that currently remain unexplained [8].

Despite their importance, research on fungal pathogens is consistently neglected and underfunded [9]. Even though fungal infections are estimated to kill 1.6 million people a year, they remain understudied and underreported [9–11]. Estimates suggest that between 1.5 million [12] and 5.1 million [13], or even 6 million [14] different species of fungi exist, but only a small fraction of them has been sequenced.

This poses a major challenge especially for pathogen detection workflows based on next-generation sequencing (NGS). Standard methods are based on recognition of known taxonomic units by homology detection, using either sequence alignment [15–23] or k-mer based approaches [24–26]. This in turn requires curated databases of fungal, as well as bacterial, viral, and other species labelled with information regarding the corresponding pathogenic phenotype or host information. Limited host information is available in the NCBI Genome browser [27], Database of Virulence Factors in Fungal Pathogens [28] and the U.S. National Fungus Collections Fungus-Host Database [29]. Those resources are partially complementary, and none of them encompasses all the available data. What is more, multiple literature sources describe fungal pathogens and their hosts without referring to the corresponding genomes, even if they are indeed available in databases such as GenBank [30] or FungiDB [31], which store genomic data without clear-cut host annotation. The ENHanCEd Infectious Diseases Database (EID2) [32] aims to detect all 'carrier'-'cargo' relationships, not limited to fungi or pathogens specifically, although it does contain fungal pathogens as well. It relies on automatically mining the 'host' field in NCBI Taxonomy [33] and finding co-occurrences of species names in articles indexed by PubMed [27], providing links to the associated nucleotide sequences. This method is efficient and scalable, but automated processing based on a concise set of simplifying assumptions may sometimes lead to spurious results. Many 'cargo' and 'carrier' species can be mentioned in the same paper even though one is not really a host of the other. This is often the case in literature reviews, articles discussing phylogenetic classification or taxonomy updates, and holds also for this work. The 'host' field in a database as large as NCBI Taxonomy may also contain outdated, inaccurate or incomplete information. For example, *Pneumocystis jirovecii*, the causative agent of deadly pneumocystis pneumonia, was previously called *Pneumocystis carinii*. While the latter name is now reserved for a species infecting exclusively rats and not humans [34], records in NCBI Taxonomy (and, possibly by consequence, EID2) still list humans as the hosts of *P. carinii* at the time of writing. What is more, many sequences included in EID2 are not genome assemblies, but single genes, which are not enough for open-view fungal pathogen detection based on shotgun sequencing. For this and similar applications, a new resource is needed. It should be based on three complementary sources: manually confirmed labels mined from the large, general-purpose databases, specialized, fungal databases accumulating biologically relevant evidence for each label, and literature.

In this work, we compiled a collection of metadata on a comprehensive selection of fungal species, annotated according to their reported pathogenicity towards humans, non-human animals, and plants. We also include plant-associated fungi without a clear pathogen annotation. We store the metadata in a flat-file database and link them to the corresponding representative or reference genomes, if available. To showcase the possible applications of the database, we model a scenario of novel fungal pathogen detection. While to our knowledge, this is a first systematic evaluation of feasibility for the novel fungal pathogen detection task, we note that it mirrors similar problems in bacterial and viral genomics [35–46]. We expect new agents to emerge due to environmental changes, host-switching events and growing human exposition to the unexplored diversity of potentially harmful fungi, as shown by the example of *C. auris*. What is more, advances in recombinant DNA assembly allow construction of whole synthetic fungal chromosomes and genomes [47–49], as well as optimization of yeast strains with new phenotypes [50, 51]. Genetic modifications of filamentous fungi also grow in popularity [52, 53]. As some of them are dangerous pathogens [10], this may lead to the development of new dual-use applications and potential bioterrorism risks. Biosecurity and biosafety regulations are supported by computational screening of ordered, potentially novel, sequences at DNA sequencing facilities, using adaptations of methods developed originally for pathogen detection from sequencing data [54–56].

Therefore, we evaluate if detecting homology between previously unseen species and their known relatives accurately predicts if a novel fungus is capable of colonizing and infecting humans. BLAST [18, 19] represents the gold standard in pathogen detection via taxonomic assignment to the closest relative. Although read mappers or k-mer based taxonomic classifiers are more computationally efficient on large NGS datasets [57–59], BLAST has been shown to be more accurate in similar tasks of detecting novel bacterial and viral pathogens [35, 46]. However, convolutional neural networks of the DeePaC package have been proven to outperform BLAST in both those scenarios [45, 46] for both isolated NGS reads and full genomes, and a recently presented variant of residual neural networks (ResNets) outperforms all alternatives on short NGS reads and their fragments [56]. Therefore, we also train a similar ResNet to predict if a novel fungal DNA sequence originates from a potential human pathogen. We then develop trained numerical representations of all genomes in the database and show the separation of potential human pathogens and non-human pathogens. Finally, as previously presented models were limited exclusively to bacteria or viruses, we extend our neural networks to a multi-class setting, which we envision to be applicable in open-view, clinical scenarios with only minimal assumptions regarding a novel, previously undiscovered causative agent. Following the convention introduced previously for bacteria [35] and viruses [46], we focus on the pathogenic potential of the analyzed species, rather than the pathogenicity itself. This distinction captures the fact that the pathogenic phenotype may or may not be manifested depending on specific host-pathogen interactions; whether this happens cannot be fully predicted outside the context of a particular host. We can only predict if a given sequence originates from a species that could cause disease under some circumstances (i.e. its pathogenic potential), not whether an invasive infection actually occurred.

Data description

Database construction

The database presented here is designed to facilitate sequence analysis and machine-learning approaches to pathogen detection. Therefore, we use the concept of a 'positive' and a 'negative' class, where the former corresponds to the species pathogenic to humans, and the latter can be defined in many ways depending on the downstream task. In our showcase of example applications of the database, we focus on pathogens (and hence our negative class consists of non-human animal and plant pathogens), but we include more records in the database for future use. More specifically, we collect metadata on species infecting humans, animals or plants, supplemented with information on other plant-associated species. To this end, we integrate multiple literature and database sources, relying on manual curation, but also including the automatically extracted data for future reference. Table 1 summarizes the data we collected for each species.

First, we accessed the Database of Virulence Factors in Fungal Pathogens (DFVF) [28] on October 9, 2021. The database contains records on virulence factors of a wide selection of fungal species together with their NCBI TaxIDs and information on their hosts

Table 1. Key information on each species included in the database.

Name	Species name, GenBank organism name and infraspecific name for the representative genome.
NCBI TaxID	Taxonomic identifiers of the species and the representative genome in the NCBI Taxonomy database.
Assembly accession & name	GenBank assembly accession number and assembly name for the representative genome.
Pathogenicity & host group	Manually curated information regarding pathogenicity and the relevant host groups.
Putative host group	Automatically extracted information suggesting a host group. Includes non-pathogenic associations.
Label sources	References to resources used to label the species for all host groups and evidence levels.
Source name and TaxID	Species name and TaxID as mentioned in the first resource used to label it.
Assembly Level	Assembly level of the representative genome.
Sequence release date	Sequence release date as reported in GenBank.
FTP path	GenBank FTP path to the representative genome.

and the disease caused. It also includes separate parts for animal and plant pathogens. We searched the records for all proteins in the database; if the "Disease" or "Disease host" fields contained the word "human", we added the corresponding species to the list of human pathogens. We added remaining species from the animal part of the database to the list of animal pathogens. Finally, we put all species mentioned in the plant part of the database on the list of plant pathogens, irrespective of whether humans were also mentioned as a possible host. For all those species, we also extracted the TaxID. As some of those TaxIDs corresponded to taxa below the species level, we linked them to their species-level ancestors in the NCBI Taxonomy tree. This resulted in a list of unique, labelled species names and species-level TaxIDs. Note that as DFVF was built using manual curation of text-mining results, we treat the obtained records as manually curated as well.

Further, we searched the Taxonomy database [33] using the query '"host human"[Properties] AND Fungi [orgn]', obtaining a list of species with a putative human host and their TaxIDs. We then used a comprehensive selection of literature reviews and comparative studies [60–76] to collect additional evidence and extend the collected list of human pathogens. We manually extracted the species names and searched the NCBI Taxonomy (including synonyms) to link them to their names as reported in Taxonomy and the respective species-level TaxIDs. In both cases (searching with extracted TaxIDs or species names), we handled the possibility of multiple, ambiguous matches by preferring exact name matches to any of the synonymous names recorded in Taxonomy. If this was not possible (e.g. because of additional annotations in the "name" field in Taxonomy), we considered matches where the first two words exactly match the query name, and varieties of a query species (i.e. matches containing the abbreviation 'var.'). If no exact match was found, we selected the first record containing the query species name. We ignored species hybrids and *formae speciales*, excluded the organism names explicitly expressing taxonomic ambiguity (marked with 'sp.', 'cf.' or 'aff.'), and filtered out fungal viruses. If the species name was not mentioned in Taxonomy at all, we still included it in the database along its associated host group label. This will enable easy extension of the database in the future, as more fungal sequences become available.

Similarly, we consulted the literature on animal [64, 66, 75–93] and plant pathogens [85, 94–104] and linked the results to the respective species TaxIDs wherever possible. Further, we accessed the Fungal Databases of the Germplasm Resources Information Network (GRIN). On October 9, 2021, we searched for all records with 'Homo sapiens' as a host [29]. We also added the species listed in the Nomenclature Fact Sheets for plant-associated fungi with quarantine importance [29] and the Fungal Diagnostic Fact Sheets of invasive and emerging fungal pathogens [105] to the list of plant-pathogenic species, and searched NCBI Taxonomy as described earlier.

We then linked the resulting lists of species TaxIDs with their representative of reference genomes in Genbank, accessed on October 9, 2021. Further, we selected all species with available genomes, but without a label assigned by any of the sources used. On the same day, we manually searched the GRIN Fungal Database for each of those species, allowing for synonyms. We checked the "Host, "Disease" and "Notes" fields; if the disease of a plant or animal host was clearly described, we added the query species to the appropriate list. If no disease was mentioned for a confirmed plant host, we added the species to a list of plant-associated fungi.

Next, we used the EID2 database [32] to extract species with putative human, animal, or plant hosts. Note that those labels were collected automatically and are therefore prone to errors. Moreover, they do not consist of pathogens only, and may include commensals or symbionts. We nevertheless added them to our database for completeness, albeit noting that they represent only a putative, automatically extracted carrier-cargo relationship, rather than a manually confirmed, true pathogenic potential of a species. The same, 'putative' category includes human-hosted species found in NCBI Taxonomy and GRIN, unless confirmed in other literature sources as well. We then selected all species with a putative human host and manually searched the Atlas of Clinical Fungi [106] with their names and synonyms noted in NCBI Taxonomy, considering also additional species referred to in the search results. If the Atlas confirmed a species to be a pathogen, we added it to the appropriate category. Note that only three of the human-hosted species (all belonging to the genus *Malassezia*) were clearly described as human commensals without any reports of causing disease (and hence, we excluded them from our list of pathogens). In cases when the Atlas mentioned two names to be synonyms, even though NCBI Taxonomy listed them as two separate species, we followed the nomenclature suggested by the Atlas. We retained both records to explicitly reflect this in the database, keeping one name unlabelled and linking it to the main record with an appropriate annotation regarding the synonym TaxID and name. Finally, we linked all labelled species to their GenBank genomes, if available.

The database contains 14,502 records in total. In the following parts of the manuscript, we will focus on what we will call the core database, comprising metadata on genomes of 954 manually confirmed pathogens (including 332 species reported to cause disease in humans). This forms a collection of species most relevant to the pathogen detection task, belonging to 6 phyla, 37 classes, 82 orders and 182 families. We also include records on 474 plant-associated fungi. The supplementary part of the database contains information on 465 putatively labelled genomes, 1,099 unlabelled species with available genomes, 6 synonyms derived from the Atlas, 886 labelled species without genomes (including 284 species without TaxIDs) and 10,618 putatively labelled species without genomes (including 9 without TaxIDs). This subset will enable easy updating of the database in the future, as more genomes of already labelled species are sequenced. It also serves as a record of all screened genomes and species to ensure reproducibility and facilitate future extensions (e.g. adding new data or sources of evidence). Figure S1 presents the logical relations between the genomes with manually confirmed labels and genomes for which putative labels could be found in EID2.

Methods

Training, validation and test sets

While we envision a wide range of possible applications of the database, we present an example usecase allowing to take advantage of the wealth of collected data – detection of novel fungal pathogens from NGS data. The core of the database contains 332 genomes of human pathogens (including opportunists), forming the positive class. The negative class comprises 622 species not reported to infect humans; this includes 565 plant pathogens and 58 non-human animal pathogens. To evaluate the performance of the selected pathogenic potential prediction methods, we divided the corresponding genomes into non-overlapping training, validation and test sets. In this setup, the training set is used as a reference database for the methods based on sequence homology and to train the neural networks, while the performance metrics are calculated on the held-out test set. The validation set is used for hyperparameter tuning and to select the best training epoch. Note that while evaluating performance on genuinely unknown species is by definition impossible, by testing the accuracy of BLAST and ResNets on sequences removed from the database, we effectively model the 'novel species' scenario.

We manually placed two clinically important pathogens – *Candida auris* (a recently emerged threat [4, 5]) and *Aspergillus fumigatus* (a causative agent of aspergillosis, reaching up to 95% mortality rates [3]) – in the test set. We also added two non-human pathogens: *Pyricularia oryzae* (syn. *Magnaporthe oryzae*; voted the most important fungal pathogen by a panel of almost 500 fungal pathologists [107] as it causes up to 30% of global rice production losses [108]) and *Batrachochytrium dendrobatidis* (blamed for recent decline of 500 amphibian species and complete extinction of 90 of them [109]). The rest of the species were assigned randomly. Overall, we placed 10% of all species in test and validation sets each, while the remaining 80% formed the training set.

To model the task of predicting pathogenic potentials from NGS samples, we simulated training, validation and test reads using Mason [110]. First, we generated 10 million single reads, 1.25 million single reads and 1.25 million paired reads per class for the training, validation and test sets respectively, keeping the number of reads per genome proportional to each genome's length. This mirrors the protocol used for bacteria [45] and viruses [46], resulting in equal mean coverage for all genomes. A side effect of such an approach is that reads originating from longer genomes could be over-represented compared to those from shorter or incomplete genomes, possibly causing generalization problems for trained machine learning models. To tackle this issue, we also simulated a second version of the training set, where the number of reads per genome is proportional to a logarithm of a given genome's length. We call this version of the training set the 'logarithmic-size' set to differentiate it from the previous ('linear-size') approach. Note that while the 'logarithmic-size' training set may help balance the resulting classifiers' performance, in a real sample, we expect the coverage to be approximately equal for equally abundant species. Therefore, we only used the 'linear-size' versions of validation and test sets. In both cases, the datasets maintain the 8-1-1 proportions on both read and genome level, with a total count of 25 million reads per class.

A crucial difference compared to previous pathogenic potential prediction studies is that fungal genomes are orders of magnitude longer than bacterial or viral ones. While the procedure mentioned above produced a mean coverage of 1.82 for the bacterial dataset presented in [45], it results in a mean coverage of only 0.15 on our data. We expected this could cause training problems, as the machine learning models would only have access to a small fraction of the overall sequence diversity of the training set. Therefore, we also simulated 'high-coverage' versions of both the 'linear-size' and 'logarithmic-size' training sets. In this setup, we increased the total number of training reads to 240 million, compared to 20 million for the 'low-coverage' versions described before. The 'high-coverage' versions keep the mean coverage of 1.82. As the reads were simulated randomly, we expected the 'low-coverage' validation and test sets to be representative, and correctly model a common case of low abundance of the target pathogen compared to other DNA sources in the sample (e.g. the host). We therefore used only 'low-coverage' validation and test sets. In summary, we generated four versions of the training read set and one version of the validation and test set each. The datasets contain the same number of reads for each of the classes and can be easily reused for future machine learning and benchmarking applications.

Pathogenic phenotype prediction

Next, we evaluated the feasibility of pathogenic potential prediction for novel fungal species. We used a ResNet architecture implemented in the DeePaC package, previously shown to outperform alternatives based on deep learning, traditional machine learning and sequence homology in the context of novel bacteria and viruses [56]. Briefly, the architecture consists of 17 convolutional layers of between 64 and 512 filters (with filter size of 7 for the first layer and 5 for the following layers) using skip connections, followed by a global average pooling layer and a single-neuron output layer with sigmoid activation. It returns an output score between 0 and 1, where a threshold of 0.5 is used by default as a boundary between 'positive' and 'negative' predictions. The network uses batch normalization and input dropout, which can be understood as randomly switching a predefined fraction of input nucleotides in training samples to *N*s. It also guarantees identical predictions for any given sequence and its reverse-complement via parameter sharing. For more details, we refer the reader to the corresponding publication [56].

As previous research reported that the input dropout rate is an important hyperparameter [45, 46, 56], we investigated two different values – the default 0.25 and 0 (no dropout). To this end, we trained the networks using the 'linear-size' and 'logarithmic-size' versions of the low-coverage training set for the default maximum of 30 epochs with an early stopping patience of 10 epochs, selecting the epoch and input dropout rate resulting in the highest accuracy on the validation set. We then retrained the networks using the selected input dropout rates and both versions of the high-coverage training set.

As shown in [45], predictions for two reads in a read pair can be averaged, improving prediction accuracy; training the networks on single reads allows to use them for both single reads and read pairs. Further, averaging over predictions for all reads originating from the same organism (e.g. for classification of whole genomes or single-species samples) has been demonstrated to yield accurate species-level predictions [35, 45]. We adapted those steps to our models as well. However, since the networks are explicitly trained for read-based classification only, the default classification threshold may be suboptimal for the genome-wise classification scenario. Therefore, we hypothesized that it should be retuned for optimal performance according to a preselected metric, e.g. balanced accuracy. To test this assumption on a fully independent dataset first, we used the 'novel viral species' dataset from [46] and their convolutional neural network (CNN) without any retraining. We selected the threshold optimizing the balanced accuracy on the appropriate validation set, rounded to two decimal places. We compared the original and retuned models and applied the procedure resulting in the better test set performance to our fungal models.

To classify novel fungal sequences based on their closest taxonomic matches, we used BLAST [18, 19] as described for bacterial data in [45]. The reference database was created from the training set genomes. We used discontiguous megablast with an E-value cutoff of 10 and default parameters, selecting the label of the top hit ('pathogenic to humans' or not) as a predicted label for each

query sequence. Note that if no hits are found, no prediction can be returned for a given read. This lowers the true positive and true negative rates, defined as the ratios of correct positive or negative predictions to all reads in the positive or negative class, respectively. For read pairs, a single match is enough to assign a label to a pair, but conflicting matches result in no prediction. Similarly to [46], we considered two approaches for generating species-level predictions. First, we found BLAST matches of contigs from the test set, using a majority vote over all contigs belonging to the same species to assign a corresponding label. Second, we used the majority vote over predictions for all reads originating from a given species. The latter case is especially relevant for a use-case of a sequencing sample containing a single-species, and is more directly comparable to the genome-level evaluation of the ResNet.

Genome representations and dataset structure

To visualize the structure of the dataset as learned by the trained classifier, we developed numerical representations for the collected genomes. This poses a challenge, as the networks are trained on reads rather than full genomes. However, we observe that final outputs of the network can be averaged over all reads originating from a single genome to generate a prediction for the genome in question [45]. Analogously, we can average the activations of the intermediate layers to construct vector representations for whole genomes based on the corresponding reads. Note that averaging the activations of the penultimate layer is approximately equivalent to using a full genome as input (assuming full coverage), as our architecture uses global average pooling just before the output layer.

More formally, by averaging the activations of the penultimate layer for a set of reads belonging to the same genome and passing the output to the final layer, we obtain the value $s(E[z])$, where $s(x) = 1/(1 + e^{-x})$ is the sigmoid activation and $z = \mathbf{w}^\top \mathbf{h} + b$, for the penultimate activation vector $\mathbf{h}$, weights of the last layer $\mathbf{w}$ and its bias b. Note that this is *not* mathematically equivalent to $E[s(z)]$, or the expected output of the network for reads originating from a given genome, which was used for the genome-level predictions in previously published literature [45, 46], as well as in this study. However, we suspected that although usually $s(E[z])$ and $E[s(z)]$ are not identical, those terms could have similar values in practice, as they both express a notion of aggregating information contained in individual reads to yield a prediction for a full genome. If this is the case, mean activations for a set of reads originating from the same genome can be used as if they were the internal genome representations of our species-level classifiers, even though the models rely on aggregating read-level predictions. To investigate the relationship between $s(E[z])$ and $E[s(z)]$, we performed additional experiments described in Supplementary Note 1.

Assuming that $s(E[z]) \approx E[s(z)]$ (see Figure S2-S3), we extracted penultimate activations for all simulated reads in the low-coverage, 'linear-size' training, validation and test sets. We then used the averaged activation vectors for each species to map the distances between them as learned by our networks. The last convolutional layer has 512 filters, but we removed the filters which activations were equal to zero for all samples in the dataset and used UMAP [111] to embed the resulting representations in a 2-dimensional space for visualization. We used the Euclidean distance metric with a minimum possible distance of 0.1 and the neighbourhood size of 15. The inputs had been randomly shuffled beforehand to avoid artifacts that can appear if an embedding is learned based on representations ordered by class.

Multi-class evaluation

Finally, we investigated an application requiring merging the 'positive' subset of our database with previously available resources for pathogenic potential prediction in bacteria and viruses [45, 46, 56]. We aimed to integrate the separate classifiers for fungi, bacteria and viruses into a single, multi-class model capable of predicting whether unassembled NGS reads originate from (possibly novel) pathogens present in a human-derived sample. To this end, we extended the DeePaC package adding the multi-class classification functionality. The resulting architectures differ from the previously described ResNets [56] only by the output layer, which has as many units as the number of considered classes and uses a softmax activation. They use the default input dropout rate, as required by the addition of viral and bacterial data. We train the models for 50 epochs due to the increased difficulty of the task.

We note that in practice, only human-hosted fungi are expected to be found in clinical samples. In this context, a slightly constrained view is admissible: we assume that only human-pathogenic fungi, human-hosted bacteria (pathogenic or commensal), human viruses and non-human viruses (mainly bacteriophages) will be present in the sample. Further, bacteriophage sequences tend to be very similar to the sequences of their bacterial hosts [112, 113] and difficult to differentiate, but both commensal bacteria and non-human viruses can be viewed here as a joint 'negative' (i.e. harmless) class. Hence, learning a precise decision boundary between them can be omitted. Human reads can be ignored, as they can be relatively easily filtered out with traditional methods based on read mapping or k-mers [24, 114, 115].

Therefore, we fused the previously published datasets used in DeePaC [56] for bacteria (pathogens vs. commensals) and viruses (human vs. non-human) with the 'positive' (human-pathogenic) class of our database. As the original datasets contain only 12.5 million reads per class (with the same 8-1-1 proportion of the training, validation and test sets, stratified by bacterial species or virus), we used the low-coverage versions of our 'positive' read sets, matching this number. We also merged the original negative classes (commensal bacteria and non-human viruses) into a joint non-pathogen class, downsampling accordingly. Further, the original datasets contain reads of between 25 and 250bp instead of 250bp only, which makes the problem more challenging, but increases robustness on very short sequences [56]. To make our fungal reads compatible with this setup, we randomly shortened the reads in our validation and low-coverage training sets to 25-250bp. We only used the test sets of 250bp in order to highlight the upper performance limit of the resulting classifier.

The final result is a dataset divided into 4 classes: nonpathogenic bacteria and non-human viruses, bacterial pathogens, human-infecting viruses and human-pathogenic fungi, in either the 'linear-size' or the 'logarithmic-size' variant. Each of the classes contains 12.5 million reads divided in the training, validation and test subsets while keeping separate species or viruses in each of the subsets. This is intended to model a scenario of analysing a clinical sample containing a mixture of previously unseen pathogens and non-pathogens, while keeping the task feasible by resigning from differentiating between bacteriophages and commensal bacteria, as well as fungi with human and non-human hosts. Note that even in this case, the 'negative' part of the presented database is useful, allowing us to constrain our view to a curated set of clinically relevant fungi only.

Using this dataset, we trained two models including all four classes (using the 'linear-size' or the 'logarithmic-size' variant of the fungal training set). We further evaluated the one resulting in higher validation accuracy and a simple ensemble averaging the predictions of both models. Then, we trained a 3-class model including only the bacterial and viral classes. This allows us to measure the 'difficulty' of integrating the fungal dataset with the others within a single network in terms of resulting differences in prediction accuracy on the original DeePaC datasets. By comparing the performance of our models to the performance of the original binary classifiers [56], we can disentangle the 'difficulty' of adding the fungal class from the 'difficulty' of integrating the bacterial and viral classes, and assess how much performance is

Table 2. Classification performance comparison between BLAST and the ResNet model on single reads and paired reads from the test data-set.

		BAcc.	Prec.	Rec.	Spec.	AUC	AUPR	Pred.	Time (CPU)	Time (GPU)
Single reads	ResNet	64.7	65.8	**61.3**	68.1	70.9	72.0	**100.0**	**25 min.**	**3 min.**
	BLAST	**66.2**	**92.6**	61.1	**71.3**	-	-	77.6	181 min.	-
Read pairs	ResNet	68.5	69.8	**65.1**	71.9	75.7	76.5	**100.0**	**50 min.**	**6 min.**
	BLAST	**69.7**	**94.5**	62.3	**77.2**	-	-	79.8	362 min.	-
Genomes	ResNet	88.4	77.5	**91.2**	85.7	95.1	90.1	**100.0**	**25-50 min.**	**3-6 min.**
	BLAST (reads)	**90.3**	**90.6**	85.3	**95.2**	-	-	**100.0**	181-362 min.	-
	BLAST (contigs)	89.5	87.8	85.3	93.7	-	-	**100.0**	2190 min.	-

The best performance value for each performance metric (row) and setting is displayed in bold. Read datasets are balanced, and the differences in performance for the first and second mate were negligible for all metrics. In genome datasets, negative class is the majority class. We use BLAST as shown in [45]; since this only returns binary labels, AUC and AUPR are not defined. Overall, BLAST is much more precise, but the ResNet yields only slightly less accurate predictions for all reads, even those impossible to match with BLAST, in a fraction of the time. AUC and AUPR are marginally higher (95.2 and 90.2) for the genome-level ResNet if only first mates are used compared to using both mates (95.1 and 90.1). Other metrics are equal, and the computation time is 50% lower. BAcc. – balanced accuracy (equivalent to accuracy on read sets), Prec. – precision; Rec. – recall, Spec. – specificity, AUC – area under the ROC curve, AUPR – area under the PR curve, Pred. – prediction rate; Time (CPU) – prediction time on 2x AMD EPYC 7742 (256 threads); Time (GPU) – prediction time on 1x Tesla V100 GPU (not possible for BLAST).

'lost' by using a more open-view classifier. Note that in the case of the viral dataset, spurious assignments to the bacterial pathogens class may be treated as detection of bacteriophages infecting the bacterial species of this class, and hence reassigned into predictions for the non-pathogen class by adding the predicted probabilities for both classes. This effectively merges the non-pathogen and bacterial pathogen classes at test time when appropriate, but still keeps the possibility to use the trained networks in a fully open-view setting (with all classes) without the need for retraining. We performed an additional comparison to BLAST with a pre-selected training database (bacterial for bacteria [45], viral for viruses [46]). This resulted in an estimated upper bound on the performance of non-machine learning approaches on those datasets (as extending the training database with irrelevant reference genomes can only lower BLAST's performance). Finally, we also evaluated the neural networks on the full dataset of all four classes, reporting performance measures for each class separately and the 'macro' average performance over all of them.

Results

Fungal pathogenic potential prediction

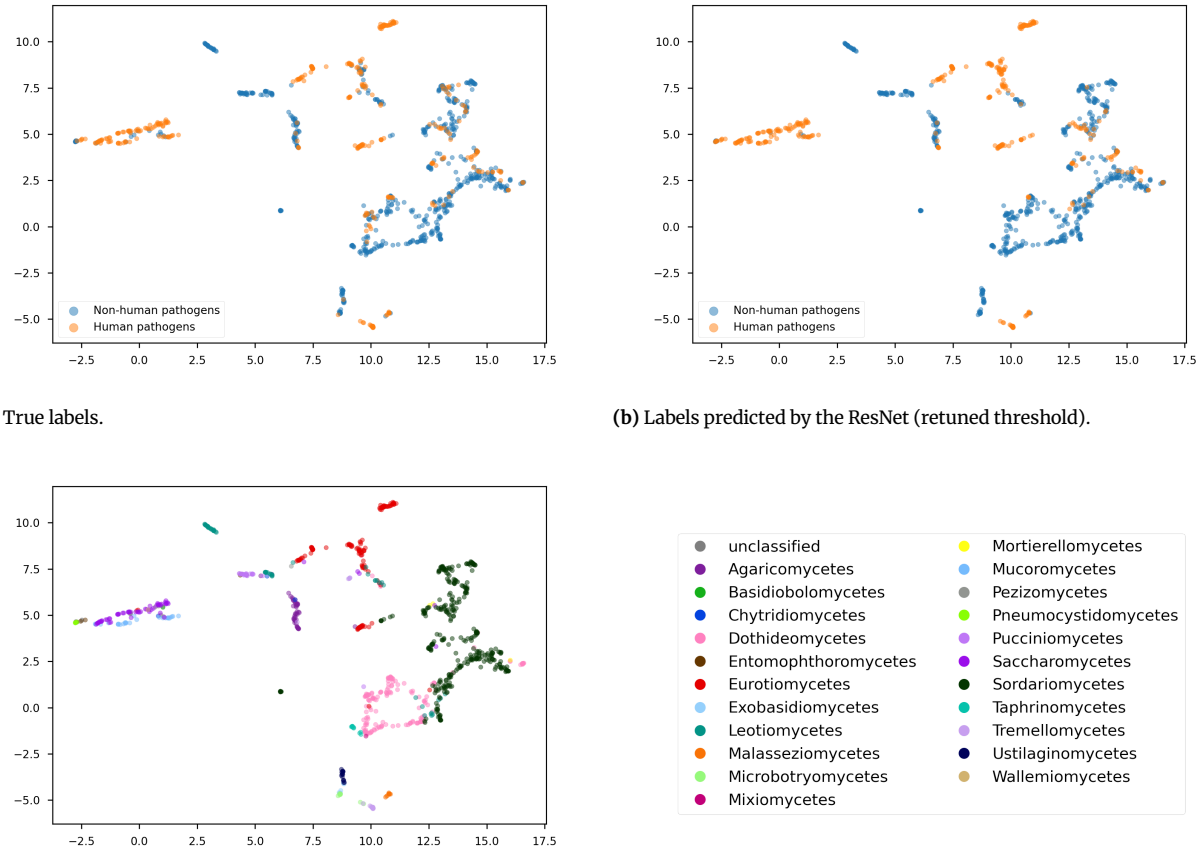
The networks trained without input dropout achieved higher validation accuracy compared to those trained with dropout, and the 'logarithmic-size' training set resulted in higher accuracy than the 'linear-size' set in all cases. As expected, the 'high-coverage' training set also improved performance, although the training time of around 8h45' per epoch on four Tesla V100 GPUs was much higher than around 45 minutes per epoch for the 'low-coverage' variant on the same hardware. The best network, trained on the high-coverage, 'logarithmic-size' dataset without dropout required 22 epochs (approx. a week) of training and was selected for further evaluation. Proper retuning of the classification threshold for species-level predictions appears to be a necessary step for an independent, viral dataset (Table S1), so we also retuned the threshold (0.46 instead of the default 0.5) for the respective fungal ResNet setup.

Overall, prediction accuracy for reads and read pairs is low for both BLAST and the ResNet, probably reflecting the extreme difficulty of the task (Table 2). The difference in performance for the first and second mate of the read pairs is negligible; we present the mean values. Interestingly, while BLAST's accuracy on the test set is higher, its accuracy on the validation set is lower by a similar margin (61.6%, compared to 63.3% for the ResNet). While this comparison should not be overinterpreted, as the ResNet epoch maximising validation accuracy was chosen, those results suggest that the performance differences between the two approaches are small and could be dependent on a particular composition of species in the test set. Although this could be explicitly tested using a nested cross-validation scheme, such a setup would be computationally prohibitive.

We do not observe much overfitting (the single read training accuracy for the selected epoch is 68.9%), which suggests that the unsatisfactory performance is a form of underfitting. However, we speculate that even those predictions could be useful in some circumstances. High precision of BLAST proves that its positive predictions are trustworthy. On the other hand, the ResNet offers only slightly worse accuracy, but for all reads in the sample (BLAST finds no matches for over 20% of reads even if read pairs are used). What is more, the ResNet can process 1.25 million reads in just 3 minutes on a single GPU, while BLAST needs over three hours and 256 threads on a high-performance computing node for the same task. The ResNet is slower when used on CPUs, but still outperforms BLAST more than sevenfold. This is very important in practice, as screening of large NGS datasets must be performed quickly to remain feasible. For this reason, BLAST is usually too slow for NGS analysis. Note that in this work, it represents the upper-bound on accuracy of homology-based approaches; faster alternatives like NGS mappers or k-mer apporaches have been shown to underperform in the pathogenic potential tasks for other pathogen groups [35, 46].

Both methods perform much better on full genomes, suggesting that the main performance bottleneck is the total amount of information (total sequence length) available as input. This is consistent with previous observations [35, 45, 46], although more drastic than in the case of bacteria or viruses. Interestingly, BLAST's precision is actually a bit lower for full genomes than for reads, while the ResNet becomes more precise when more information (more reads) is considered. Contrary to what was reported for viruses, using raw, unassembled reads improves BLAST's performance compared to relying on assembled contigs. A probable reason is that when a majority vote over all contigs is taken, all contigs have the same voting weight, which means that the ratio of voting weight to sequence length can be much higher for short contigs of potentially lower quality than for the very long ones. In contrast, when reads are used, all regions of the genome weight the same. Using contigs also markedly increases the processing time. This slowdown is probably best explained by the computational difficulty of the local alignment of long contigs compared to a low-coverage set of representative reads.

Despite the low coverage, the test set seems to be indeed representative. The genome-wide predictions are equally accurate if all test reads are used, and when only a half of the dataset (corresponding to either the first or the second mate) is analyzed. Therefore, computations can be sped up by considering first mates only, as they are enough to deliver an accurate prediction. Strikingly, this



(a) True labels.

(b) Labels predicted by the ResNet (retuned threshold).

(c) Taxonomic rank: class.

Figure 1. UMAP embeddings of the learned genome representations for the core database. Enlarged in Figures S4, S5 and S7. Each point represents a genome of a single species, coloured by its ground truth label, predicted label, or its taxonomic classification. The learned representations offer a way of visualizing the core database along relevant labels for each genome. The ResNet correctly classifies most of the genomes. Its predictions within clusters are more homogeneous than the ground truth, but many 'positive' members of otherwise 'negative' clusters are detected. The clusters are related, but not fully reducible to the taxonomic classification of the analysed species. In general, related species are close to each other in this space, but some taxonomic units are distributed among more than one cluster (e.g. *Eurotiomycetes*, in red), and some clusters contain members of distant taxa (e.g. the leftmost cluster).

is the case even though they correspond to a mean coverage below 0.08. As a result, the read-based ResNet yields only slightly less accurate predictions than contig-based BLAST, but requires 700-fold less time if a GPU is used.

Both BLAST and the ResNet correctly predict *C. auris* and *A. fumigatus* to be pathogens, even though they were kept in the held-out test dataset unused during training. *P. oryzae* is also recognized as a a member of the negative class. The ResNet (but not BLAST) incorrectly classifies *B. dendrobatidis*, a deadly amphibian pathogen, as able to infect humans. This prediction could be in principle indicative of a hidden zoonotic potential that could be realized under some yet unreported circumstances (e.g. in susceptible individuals or tissues). However, without further evidence, we assume this is simply an error of the model – perfect generalization is rarely possible, even though as noted in Table 2, the prediction performance of the ResNet on full genomes is still high. It is possible that the erroneous prediction for *B. dendrobatidis* is caused by underrepresentation of related species in the training set, as the core database contains only two other species of the same order, *Chytridiomycota*, hindering the classifier from correctly modelling their pathogenic potential.

The landscape of fungal pathogenicity

Good overall accuracy of the ResNet is reflected in the visualisation of learned genome representations for the entirety of the core database. Figure 1 presents UMAP embeddings of the extracted representations for all labelled genomes, i.e. a sum of the training, validation and test datasets. Although some noise is present, the positive and the negative class are mostly separated. As shown in Figure 1a (enlarged in Figure S5), several clusters of human pathogens and non-human pathogens are present. The ResNet correctly recovers most of the labels, including many of the 'positive' members of the otherwise 'negative' clusters (Figure 1b). Classification errors seem to originate from a interpolation based on neighbouring data points – within clusters, the predicted labels are more homogeneous than the ground truth annotations. This is expected, as the clusters represent similarity in the space of learned representations. The network should in general assign similar labels to inputs similar in this space. In contrast, BLAST works analogously to a k-nearest neighbours classifier in the input sequence space (finding the single closest match for each query). The ResNet, interpolating between multiple data points, may be less efficient in modelling situations where a small set of 'negative' data points is embedded within a larger 'positive' cluster of similar species, or vice versa. This hypothesis is supported by the visualization of BLAST-predicted labels in the learned representation space (Figure S6). BLAST recovers mixed, contrasting labels within cluster

Table 3. Performance of the multi-class classifiers on the multi-class dataset, read pairs.

		Prec.	Rec.	AUPR
All classes	4-class ensemble	87.7	87.6	93.4
	4-class	86.8	86.6	92.8
	3-class	85.8	85.5	92.4
Non-pathogens	4-class ensemble	80.1	**77.4**	86.7
	4-class	81.5	72.5	85.3
	3-class	**81.8**	75.5	**87.6**
Path. bacteria	4-class ensemble	**83.2**	87.2	90.4
	4-class	79.1	89.0	90.4
	3-class	79.3	**89.7**	**91.1**
Human viruses	4-class ensemble	**96.7**	90.9	98.4
	4-class	95.7	**91.2**	98.1
	3-class	96.3	91.1	**98.5**
Fungi	4-class ensemble	**90.9**	**95.0**	**97.9**
	4-class	90.7	93.7	97.4

The 4-class classifier includes the fungi class along the three viral and bacterial classes included in the 3-class classifier. The best performance for each class is marked in bold. In this setting, the true positive rate corresponds to the rate of correct assignments within a given class. Hence, 'macro'-average recall is equivalent to 'macro'-average accuracy. Integrating fungi into the classifiers does significantly impact the overall accuracy for any of the non-fungal classes, and the 4-class ensemble offers a more balanced performance on putative non-pathogen reads than the single-network alternative. Prec. – precision, Rec. – recall, AUPR – area under the PR curve.

more accurately, and its errors seem to be more evenly distributed across the space. At the same time, its slightly lower sensitivity is especially visible within the diverse *Sordariomycetes* class placed in the rightmost cluster.

The clusters themselves are noticeably related to the taxonomic units represented in the database (Figure 1c, enlarged in Figure S7). The number of units at most levels of taxonomy poses a challenge for intuitive visualization of the dataset structure. Nevertheless, some correspondence between cluster membership, label and the taxonomic rank of a class is clearly visible (note that the 'class' as a taxonomic term is distinct from the concept of a positive or negative class in machine learning). This extends also to the lower and higher taxonomic ranks of order and phylum, respectively (Figure S8-S9). On the other hand, this is importantly not a simple one-to-one mapping. Some orders of one class may be belong to different clusters. For example, the orders *Chaetothyriales*, *Eurotiales* (including the genus *Aspergillus*) and *Onygenales* all belong to the class *Eurotiomycetes* (Figure 1c and S7, in red), but are members of four different clusters, with *Eurotiales* split into two (Figure S8). Even more strikingly, the leftmost cluster contains at least one member of all six phyla, with two most noticeable groups corresponding to the orders *Saccharomycetales* and *Mucorales* in phyla *Ascomycota* and *Mucormycota*, respectively (Figures 1c, S7-S9). 89% of the members of both those orders are at least opportunistic human pathogens; this includes *Candida* species in *Saccharomycetales* and causative agents of mucormycosis, known also as the 'black fungus', in *Mucorales*. Therefore, we hypothesize that the grouping represents a cluster of 'conserved' potential pathogens, relatively easy to classify with an acceptable accuracy. A common signal at the functional level or technical artifacts could be alternative explanations; they are however less likely due to large phylogenetic distance on one side, and a consistent grouping of almost all species from those well-represented orders on the other.

Multi-class models

For the final evaluation of our database, we focused on a generalized version of the pathogenic potential prediction task. Here, we aimed to develop a model capable of classifying NGS reads originating from novel viruses, bacterial and fungal species into appropriate pathogen and non-pathogen classes. We trained the multi-class ResNets on data including four classes (human-pathogenic fungi, bacterial pathogens, human viruses and non-pathogens) as described in section Multi-class evaluation. The network trained on the dataset containing the 'logarithmic-size' version of the fungal positive class achieved slightly better validation accuracy and was selected for further evaluation, but the difference was small (<0.5%). We also evaluated a simple ensemble of both 4-class ResNets.

First, we used the DeePaC datasets consisting of bacteria and viruses to compare the 4-class models to a classifier including the three non-fungal classes only, as well as the original binary ResNets [56] and BLAST. This procedure allows us to a) measure the effect of integrating the fungal dataset with the bacterial and viral data in one task, and b) disentangle the effects of adding the fungal data from the effects of merging the bacterial and viral datasets. We expected the fungal sequences to be relatively easy to differentiate from the others, but whether the ResNet architecture would be expressive enough to accurately represent all those diverse sequences was unclear.

As shown in Table S2, integrating the fungal dataset with three bacterial and viral classes indeed does not negatively influence the prediction accuracy. The main challenge in multi-class pathogenic potential prediction seems to be actually differentiating between bacteriophages and the pathogenic bacteria that they infect. The multi-class networks confuse 18-21% of viral non-pathogen reads for pathogenic bacteria. However, their accuracy is still similar to BLAST's, even though BLAST was used with a purely viral reference database, and hence would always classify ambiguous bacterial or viral sequences as viral. This problem is unrelated to the fungal database presented here, and can be circumvented by assuming that spurious assignments to bacteria should be reassigned to bacteriophages in the case of a purely viral dataset (see section Multi-class evaluation). The resulting networks achieve accuracy similar to that of the simple binary classifiers, while still being capable of multi-class predictions for more complex datasets.

The fungal dataset can be integrated with the other classes without causing any significant performance hits on the full, multi-class dataset as well (Table 3). Consistently with the other results, performance is lower on the non-pathogen class, since many bacteriophage reads can be confused with pathogenic bacteria. While this issue requires further research, we expect future solutions to remain compatible with our database. The 4-class ensemble achieves the most balanced performance on non-pathogen data, the best recall on fungal reads and is also the most accurate overall. The results show that using the data collected in our database, we can efficiently extend previously developed pathogenic potential methods to be compatible with a metagenomic setting including fungi, without assuming a purely bacterial or viral sample as it was done previously [56].

Discussion

Database of pathogenic fungi and their hosts

Fungal pathogens have been under-studied compared to human-infecting bacteria and viruses, leading to repeated calls for more research in this area [9, 11]. What is more, a large part of the research effort has been focused on plant pathogens due to their agricultural significance. A subset of them could in principle also have an unreported or undetected ability to infect a human host. An analogous problem applies also to incomplete data regarding pathogenicity towards non-human animals or plants. For this reason, we do not claim that species not listed as potential pathogens are indeed non-pathogens. In our database, we include confirmed labels alongside appropriate sources; in case of lack of evidence, we treat the re-

spective label as missing. It is therefore possible that some of the fungi currently labelled as 'non-human pathogens' would have to be reclassified as the state of science evolves. This may be especially important as it has been suggested that the ongoing climate change will lead to more frequent host-switching events, including expansion of host-range to mammals, which are usually relatively resistant to fungal infections [116]. Even though the very goal of the presented classifiers is to generalize to newly emerging species, large, comprehensive datasets are crucial – often more important than the actual analysis method used. This has been shown before for metagenomic data [25] and likely applies to the tasks analyzed here as well. Therefore, extending the database to include more species, as more genomes are sequenced in the future, could facilitate the downstream tasks. To support future extensions, we include all considered species in the database – even those without assigned TaxIDs, genomes, or labels (in the case of screened GenBank genomes). This broadens the scope of the data from 1,428 labelled genomes to over 14,000 records, enabling easy labelling of newly published genomes and minimizing the workload needed for addition of new, non-redundant records. It is also possible to link the species TaxIDs to taxa below the species level. However, it should be kept in mind that the fungal taxonomy is in constant flux – taxa previously considered variants of a single species may be reclassified into separate species in the future. While automatically curated databases like EID2 [32] are relatively easy to update and scale, we note that they may be prone to errors introduced by the automated protocol used. Manual curation is not fully error-free either, but we see it as a necessary step to maximize the quality of the collected labels. Both approaches are complementary and may be best suited for different use-cases.

Application to pathogenic potential prediction

We show that both BLAST and ResNet can accurately predict if a fungus is a human pathogen based on its genome. However, the task remains inherently difficult – a DNA sequence by itself, outside of the context of the host, can be probabilistically associated with a certain trait, but the trait itself is only realized in the biological system comprising of both the host and the colonizing microbe as a whole. Therefore, we intentionally include opportunistic pathogens in the 'positive' class of human pathogens. These considerations mirror challenges previously described for other pathogen classes [35, 45, 46, 56].

Given limited information content of a short NGS read, predicting if a fungus can infect humans directly (and only) from reads is therefore an extremely challenging task. The read-level performance is admittedly low for predicting a fungal host. More expressive neural architectures could possible help alleviate this problem, but it is also possible that prediction accuracy known for other pathogen classes is simply impossible for fungi, for example due to the higher complexity of eukaryotic genomes with abundant regions of non-coding DNA. On the other hand, multi-class networks detecting fungal, bacterial or viral pathogens yield satisfactory accuracy under some simplifying assumptions. We envision that the networks trained using the database can be applied to predict pathogenic potentials of short metagenomic reads, as previously suggested for purely bacterial and viral datasets [56]. Although we only evaluate them on simulated reads, this has been shown to approximate performance on real datasets very well [45, 46, 56]. Further work could extend the presented multi-class setup to Nanopore reads, as shown for bacterial and viral models [56], enabling selective sequencing of mixed-pathogen samples.

Data visualization and genome representations

We show that full genomes can be represented by aggregating representations of reads originating from each genome. In addition to that, we observe that coverage as low as 0.08 is enough to correctly classify a species. Taken together, those two facts warrant a view of a species genome as a distribution generating subsequences (i.e. reads) originating from it; such a distribution can also be considered in an abstract representation space (Supplementary Note 1). This concept is very similar to that of a k-mer spectrum, where an *empirical* distribution of k-mers is used as a signature of a longer sequence to enable alignment-free comparisons [117], including being used as input features for machine learning approaches as in [35]. However, k-mer spectra operate in the sequence space only.

Classifiers based on aggregated *representations* are approximately equivalent to classifiers based on aggregated *predictions*, although this relation is modulated by the standard deviation of the respective, genome-specific distribution. A somewhat related effect was reported in the context of competing design choices for neural networks equivariant to DNA reverse-complementarity – models averaging the predictions for both DNA strands were found to be approximately equivalent to models applying a sigmoid transformation to an average of logits [118]. The probabilistic view of genome representations presented here deserves deeper investigation; this could potentially lead to a development of useful embeddings also for whole, multi-species samples.

Finally, the trained representations allowed us to visualize the taxonomic diversity of the database along its phenotypic landscape. As expected, the apparent fungal host-range signal seems to be related to, but not fully reducible to the fungal taxonomy. It can be captured by finding the closest taxonomic matches, as well as by a neural network, avoiding direct taxonomic assignment for a more assumption-free view of fungal pathogenicity. This suggests that both methods should generalize well to novel species sequenced in the future. At the same time, a stark dataset bias is clearly visible, with many more genomes available for some taxonomic units than for others in the same rank. Future sequencing projects focusing on underrepresented taxa could help build a more fine-grained view of the pathogenicity landscape presented here.

Reuse potential and implications

Although in this work, we focus on using the collected data in a pathogenic potential prediction task, the database can find future applications beyond this particular problem. For example, we imagine that fungal genomes could be scanned for regions associated with their ability to colonize and infect humans, as shown previously for bacteria and viruses [46]. This approach uses sliding windows of the same size as the NGS reads used for training; high-scoring sequences tend to be clustered together and often collocalized with virulence-related genes. Even though our binary fungal host classifiers achieved relatively low per-read accuracy, we speculate that finding contiguous regions of elevated pathogenic potential could be a way to identify potentially relevant genes, also for novel and weakly annotated species. On the other hand, the multitude of genomic features present in fungal genomes, often including intron features and regions without obvious functional annotation, renders the validation of such an approach a challenging project on its own. This could be perhaps facilitated by focusing exclusively on coding regions, which should in principle carry a stronger phenotype-related signal, at the risk of omitting potentially relevant, non-coding (e.g. regulatory) elements. Building such a gene-based classifier could be a future application of the presented dataset.

We imagine that, depending on annotation quality, genomes collected here could be a valuable resource for functional and comparative genomics of fungi. As as source of curated labels, it could also support application of proteomics to fungal pathogen research. For example, computational metaproteomics and proteogenomics approaches enable analysis of microbial communities based on mass spectrometry data and can be co-opted for pathogen detection

workflows independent of DNA sequencing [119, 120].

Conclusions

We compiled a comprehensive database of fungal species linked to their host group (human, non-human animal or plant), evidence for their pathogenicity, and publicly available genomes. To showcase the potential uses of the dataset, we benchmark two most promising approaches to fungal pathogenic potential prediction: a deep neural network capable of fast inference directly from fungal DNA sequences, and the gold standard in homology-based pathogen identification – BLAST. We also extended our approach to a multi-class scenario, integrating collected fungal data with bacterial and viral datasets. The database can be reused for future research on fungal pathogenicity based on both machine learning and other methods of computational biology.

Availability of source code and requirements

All R scripts used for data preparation are publicly available at https://github.com/dacs-hpi/deepac under the MIT licence alongside Python code for ResNet training, UMAP visualization and experiments on simulated genome representations. All additional package requirements and installation instructions for the DeePaC package used in this study are listed in the repository.

Availability of supporting data and materials

The database presented in this article is available in a Zenodo repository, https://zenodo.org/record/5711852. The trained models are available at https://zenodo.org/record/5711877, and the simulated reads used to train and evaluate them at https://zenodo.org/record/5713153.

Declarations

List of abbreviations

AUC: Area Under the ROC curve; AUPR: Area Under the PR (precision-recall) curve; DFVF: Database of Fungal Virulence Factors; NCBI: U.S. National Center for Biotechnology Information; NGS: Next-Generation Sequencing; ResNet: Residual Network; ROC curve: Receiver Operating Characteristic curve.

Competing Interests

The authors declare that they have no competing interests.

Funding

This work was supported by the BMBF-funded Computational Life Science initiative (project DeepPath, 031L0208, to B.Y.R.) and the BMBF-funded de.NBI Cloud within the German Network for Bioinformatics Infrastructure (de.NBI) (031A537B, 031A533A, 031A538A, 031A533B, 031A535A, 031A537C, 031A534A, 031A532B).

Author's Contributions

Conceptualization: J.M.B., F.N., B.Y.R.; Data curation: J.M.B., F.N.; Methodology, Formal Analysis, Visualization, Software and Investigation: J.M.B., F.N., M.N.; Funding acquisition and Supervision: J.M.B., B.Y.R.; Writing – original draft: J.M.B., F.N.; Writing – review & editing: J.M.B., F.N., M.N., B.Y.R.

Acknowledgements

We thank the NCBI help desk for assistance and helpful suggestions, as well as Katharina Baum (HPI) for valuable discussions and comments.

References

1. Sobel JD. Vulvovaginal candidosis. The Lancet 2007;369(9577):1961–1971. https://doi.org/10.1016/S0140-6736(07)60917-9.
2. Raimondi S, Amaretti A, Gozzoli C, Simone M, Righini L, Candeliere F, et al. Longitudinal Survey of Fungi in the Human Gut: ITS Profiling, Phenotyping, and Colonization. Frontiers in Microbiology 2019 Oct;10.
3. Brown GD, Denning DW, Gow NAR, Levitz SM, Netea MG, White TC. Hidden Killers: Human Fungal Infections. Science Translational Medicine 2012;4(165).
4. Satoh K, Makimura K, Hasumi Y, Nishiyama Y, Uchida K, Yamaguchi H. Candida auris sp. nov., a novel ascomycetous yeast isolated from the external ear canal of an inpatient in a Japanese hospital. Microbiology and Immunology 2009;53(1):41–44. https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1348-0421.2008.00083.x.
5. Centers for Disease Control and Prevention (US), National Center for Emerging Zoonotic and Infectious Diseases (US) Division of Healthcare Quality Promotion Antibiotic Resistance Coordination and Strategy Unit , editors. Antibiotic resistance threats in the United States, 2019; 2019. https://stacks.cdc.gov/view/cdc/82532.
6. Spivak ES, Hanson KE. Candida auris: an emerging fungal pathogen. Journal of clinical microbiology 2018;56(2).
7. Casadevall A, Kontoyiannis DP, Robert V. On the emergence of Candida auris: climate change, azoles, swamps, and birds. MBio 2019;10(4):e01397–19.
8. Lockhart SR, Etienne KA, Vallabhaneni S, Farooqi J, Chowdhary A, Govender NP, et al. Simultaneous Emergence of Multidrug-Resistant Candida auris on 3 Continents Confirmed by Whole-Genome Sequencing and Epidemiological Analyses. Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America 2017 Jan;64(2):134–140.
9. Stop neglecting fungi. Nature Microbiology 2017 Jul;2(8):1–2. https://www.nature.com/articles/nmicrobiol2017120.
10. Chowdhary A, Agarwal K, Meis JF. Filamentous Fungi in Respiratory Infections. What Lies Beyond Aspergillosis and Mucormycosis? PLOS Pathogens 2016 Apr;12(4):e1005491. https://journals.plos.org/plospathogens/article?id=10.1371/journal.ppat.1005491.
11. Huseyin CE, O'Toole PW, Cotter PD, Scanlan PD. Forgotten fungi—the gut mycobiome in human health and disease. FEMS Microbiology Reviews 2017 Jul;41(4):479–511. https://academic.oup.com/femsre/article/41/4/479/3738183.
12. Hawksworth DL. The magnitude of fungal diversity: the 1.5 million species estimate revisited. Mycological Research 2001 Dec;105(12):1422–1432. https://www.sciencedirect.com/science/article/pii/S0953756208620274.
13. Blackwell M. The Fungi: 1, 2, 3 … 5.1 million species? American Journal of Botany 2011;98(3):426–438.
14. Taylor DL, Hollingsworth TN, McFarland JW, Lennon NJ, Nusbaum C, Ruess RW. A first comprehensive census of fungi in soil reveals both hyperdiversity and fine-scale niche partitioning. Ecological Monographs 2014;84(1):3–20. https://esajournals.onlinelibrary.wiley.com/doi/abs/10.1890/12-1693.1.
15. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 2010 Mar;26(5):589–595.

16. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nature methods 2012;9(4):357–359. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3322381/.
17. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 2018;34(18):3094–3100. https://doi.org/10.1093/bioinformatics/bty191.
18. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. Journal of Molecular Biology 1990 Oct;215(3):403–410.
19. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinformatics 2009 Dec;10(1):421. https://doi.org/10.1186/1471-2105-10-421.
20. Hong C, Manimaran S, Shen Y, Perez-Rogers JF, Byrd AL, Castro-Nallar E, et al. PathoScope 2.0: a complete computational framework for strain identification in environmental or clinical sequencing samples. Microbiome 2014;2(1):33. https://doi.org/10.1186/2049-2618-2-33.
21. Naccache SN, Federman S, Veeraraghavan N, Zaharia M, Lee D, Samayoa E, et al. A cloud-compatible bioinformatics pipeline for ultrarapid pathogen identification from next-generation sequencing of clinical samples. Genome Research 2014 Jul;24(7):1180–1192.
22. Ahn TH, Chai J, Pan C. Sigma: Strain-level inference of genomes from metagenomic analysis for biosurveillance. Bioinformatics 2015 Jan;31(2):170–177. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4287953/.
23. Andrusch A, Dabrowski PW, Klenner J, Tausch SH, Kohl C, Osman AA, et al. PAIPline: pathogen identification in metagenomic and clinical next generation sequencing samples. Bioinformatics 2018;34(17):i715–i721. https://academic.oup.com/bioinformatics/article/34/17/i715/5093217.
24. Wood DE, Lu J, Langmead B. Improved metagenomic analysis with Kraken 2. Genome Biology 2019 Nov;20(1):257. https://doi.org/10.1186/s13059-019-1891-0.
25. Piro VC, Dadi TH, Seiler E, Reinert K, Renard BY. ganon: precise metagenomics classification against large and up-to-date sets of reference sequences. Bioinformatics 2020 Jul;36(Supplement_1):i12–i20. https://doi.org/10.1093/bioinformatics/btaa458.
26. Breitwieser FP, Baker DN, Salzberg SL. KrakenUniq: confident and fast metagenomics classification using unique k-mer counts. Genome Biology 2018 Nov;19(1):198. https://doi.org/10.1186/s13059-018-1568-0.
27. Sayers EW, Beck J, Bolton EE, Bourexis D, Brister JR, Canese K, et al. Database resources of the National Center for Biotechnology Information. Nucleic Acids Research 2021 Jan;49(D1):D10–D17.
28. Lu T, Yao B, Zhang C. DFVF: database of fungal virulence factors. Database 2012 Jan;2012(bas032). https://doi.org/10.1093/database/bas032.
29. Farr DF, Rossman AY, Fungal Databases. U.S. National Fungus Collections, ARS, USDA; 2021. https://nt.ars-grin.gov/fungaldatabases/, retrieved October 9, 2021.
30. Sayers EW, Cavanaugh M, Clark K, Pruitt KD, Schoch CL, Sherry ST, et al. GenBank. Nucleic Acids Research 2021 Jan;49(D1):D92–D96. https://doi.org/10.1093/nar/gkaa1023.
31. Basenko EY, Pulman JA, Shanmugasundram A, Harb OS, Crouch K, Starns D, et al. FungiDB: An Integrated Bioinformatic Resource for Fungi and Oomycetes. Journal of Fungi 2018 Mar;4(1):39. https://www.mdpi.com/2309-608X/4/1/39.
32. Wardeh M, Risley C, McIntyre MK, Setzkorn C, Baylis M. Database of host-pathogen and related species interactions, and their global distribution. Scientific Data 2015 Sep;2(1):150049. https://www.nature.com/articles/sdata201549.
33. Schoch CL, Ciufo S, Domrachev M, Hotton CL, Kannan S, Khovanskaya R, et al. NCBI Taxonomy: a comprehensive update on curation, resources and tools. Database 2020;2020.
34. Stringer JR, Beard CB, Miller RF, Wakefield AE. A New Name for Pneumocystis from Humans and New Perspectives on the Host-Pathogen Relationship. Emerging Infectious Diseases 2002 Sep;8(9):891–896. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2732539/.
35. Deneke C, Rentzsch R, Renard BY. PaPrBaG: A machine learning approach for the detection of novel pathogens from NGS data. Scientific Reports 2017;7:39194. https://www.nature.com/articles/srep39194.
36. Mock F, Viehweger A, Barth E, Marz M. VIDHOP, viral host prediction with Deep Learning. Bioinformatics 2020;https://doi.org/10.1093/bioinformatics/btaa705.
37. Zhang Z, Cai Z, Tan Z, Lu C, Jiang T, Zhang G, et al. Rapid identification of human-infecting viruses. Transboundary and Emerging Diseases 2019;66(6):2517–2522. https://onlinelibrary.wiley.com/doi/abs/10.1111/tbed.13314.
38. Barash E, Sal-Man N, Sabato S, Ziv-Ukelson M. BacPaCS—Bacterial Pathogenicity Classification via Sparse-SVM. Bioinformatics 2018;https://academic.oup.com/bioinformatics/advance-article/doi/10.1093/bioinformatics/bty928/5165373.
39. Guo Q, Li M, Wang C, Guo J, Jiang X, Tan J, et al. Predicting Hosts Based on Early SARS-CoV-2 Samples and Analyzing Later World-wide Pandemic in 2020. bioRxiv 2021 Mar;p. 2021.03.21.436312. https://www.biorxiv.org/content/10.1101/2021.03.21.436312v1.
40. Gałan W, Bąk M, Jakubowska M. Host Taxon Predictor - A Tool for Predicting Taxon of the Host of a Newly Discovered Virus. Scientific Reports 2019 Mar;9(1):3436. https://www.nature.com/articles/s41598-019-39847-2.
41. Wardeh M, Baylis M, Blagrove MSC. Predicting mammalian hosts in which novel coronaviruses can be generated. Nature Communications 2021 Feb;12(1):780. https://www.nature.com/articles/s41467-021-21034-5.
42. Brierley L, Fowler A. Predicting the animal hosts of coronaviruses from compositional biases of spike protein and whole genome sequences through machine learning. PLOS Pathogens 2021 Apr;17(4):e1009149. https://journals.plos.org/plospathogens/article?id=10.1371/journal.ppat.1009149.
43. Bergner LM, Mollentze N, Orton RJ, Tello C, Broos A, Biek R, et al. Characterizing and Evaluating the Zoonotic Potential of Novel Viruses Discovered in Vampire Bats. Viruses 2021 Feb;13(2):252. https://www.mdpi.com/1999-4915/13/2/252.
44. Tang Q, Song Y, Shi M, Cheng Y, Zhang W, Xia XQ. Inferring the hosts of coronavirus using dual statistical models based on nucleotide composition. Scientific Reports 2015 Nov;5(1):17155. https://www.nature.com/articles/srep17155.
45. Bartoszewicz JM, Seidel A, Rentzsch R, Renard BY. DeePaC: predicting pathogenic potential of novel DNA with reverse-complement neural networks. Bioinformatics 2020 Jan;36(1):81–89. https://doi.org/10.1093/bioinformatics/btz541.
46. Bartoszewicz JM, Seidel A, Renard BY. Interpretable detection of novel human viruses from genome sequencing data. NAR Genomics and Bioinformatics 2021 Mar;3(1). https://doi.org/10.1093/nargab/lqab004.
47. Richardson SM, Mitchell LA, Stracquadanio G, Yang K, Dymond JS, DiCarlo JE, et al. Design of a synthetic yeast genome. Science 2017 Mar;355(6329):1040–1044. http://science.sciencemag.org/content/355/6329/1040.
48. Burgess DJ. Synthetic Biology: Building a custom eukaryotic genome de novo. Nature Reviews Genetics 2017 May;18(5):274–274. http://www.nature.com/nrg/journal/v18/n5/full/nrg.2017.30.html?WT.feed_name=subjects_synthetic-biology.

49. Dai J, Boeke JD, Luo Z, Jiang S, Cai Y. Sc3.0: revamping and minimizing the yeast genome. Genome Biology 2020 Aug;21(1):205. https://doi.org/10.1186/s13059-020-02130-z.
50. Szymanski E, Calvert J. Designing with living systems in the synthetic yeast project. Nature Communications 2018 Jul;9(1):2950. https://www.nature.com/articles/s41467-018-05332-z.
51. Luo Z, Wang L, Wang Y, Zhang W, Guo Y, Shen Y, et al. Identifying and characterizing SCRaMbLEd synthetic yeast using ReSCuES. Nature Communications 2018 May;9(1):1930. https://www.nature.com/articles/s41467-017-00806-y.
52. Amores GR, Guazzaroni ME, Arruda LM, Silva-Rocha R. Recent Progress on Systems and Synthetic Biology Approaches to Engineer Fungi As Microbial Cell Factories. Current Genomics 2016 Apr;17(2):85–98.
53. Martins-Santana L, Nora LC, Sanches-Medeiros A, Lovate GL, Cassiano MHA, Silva-Rocha R. Systems and Synthetic Biology Approaches to Engineer Fungi for Fine Chemical Production. Frontiers in Bioengineering and Biotechnology 2018;6. https://www.frontiersin.org/articles/10.3389/fbioe.2018.00117/full.
54. Diggans J, Leproust E. Next Steps for Access to Safe, Secure DNA Synthesis. Frontiers in Bioengineering and Biotechnology 2019;7. https://www.frontiersin.org/articles/10.3389/fbioe.2019.00086/full.
55. Balaji A, Kille B, Kappell AD, Godbold GD, Diep M, Elworth RAL, et al. SeqScreen: Accurate and Sensitive Functional Screening of Pathogenic Sequences via Ensemble Learning. bioRxiv 2021 May;p. 2021.05.02.442344. https://www.biorxiv.org/content/10.1101/2021.05.02.442344v2, publisher: Cold Spring Harbor Laboratory Section: New Results.
56. Bartoszewicz JM, Genske U, Renard BY. Deep learning-based real-time detection of novel pathogens during sequencing. Briefings in Bioinformatics 2021 Jul;(bbab269). https://doi.org/10.1093/bib/bbab269.
57. Breitwieser FP, Lu J, Salzberg SL. A review of methods and databases for metagenomic classification and assembly. Briefings in Bioinformatics 2017 Sep;20(4):1125–1136. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6781581/.
58. Ye SH, Siddle KJ, Park DJ, Sabeti PC. Benchmarking Metagenomics Tools for Taxonomic Classification. Cell 2019 Aug;178(4):779–794. https://www.cell.com/cell/abstract/S0092-8674(19)30775-5.
59. Alser M, Rotman J, Deshpande D, Taraszka K, Shi H, Baykal PI, et al. Technology dictates algorithms: recent developments in read alignment. Genome Biology 2021 Aug;22(1):249. https://doi.org/10.1186/s13059-021-02443-7.
60. Taylor LH, Latham SM, Woolhouse ME. Risk factors for human disease emergence. Philosophical Transactions of the Royal Society of London Series B, Biological Sciences 2001 Jul;356(1411):983–989.
61. Chowdhary A, Kathuria S, Agarwal K, Meis JF. Recognizing filamentous basidiomycetes as agents of human disease: A review. Medical Mycology 2014 Nov;52(8):782–797. https://doi.org/10.1093/mmy/myu047.
62. Bombassaro A, Schneider GX, Costa FF, Leão ACR, Soley BS, Medeiros F, et al. Genomics and Virulence of Fonsecaea pugnacius, Agent of Disseminated Chromoblastomycosis. Frontiers in Genetics 2020;11:822. https://www.frontiersin.org/article/10.3389/fgene.2020.00822.
63. Brunner-Mendoza C, Reyes-Montes MdR, Moonjely S, Bidochka MJ, Toriello C. A review on the genus Metarhizium as an entomopathogenic microbial biocontrol agent with emphasis on its use and utility in Mexico. Biocontrol Science and Technology 2019 Jan;29(1):83–102. https://doi.org/10.1080/09583157.2018.1531111.
64. Franzen C, Müller A. Microsporidiosis: human diseases and diagnosis. Microbes and Infection 2001 Apr;3(5):389–400.
65. Colombo AL, Padovan ACB, Chaves GM. Current knowledge of Trichosporon spp. and Trichosporonosis. Clinical Microbiology Reviews 2011 Oct;24(4):682–700.
66. Han B, Weiss LM. Microsporidia: Obligate Intracellular Pathogens Within the Fungal Kingdom. Microbiology Spectrum 2017 Apr;5(2).
67. Hu W, Ran Y, Zhuang K, Lama J, Zhang C. Alternaria arborescens Infection in a Healthy Individual and Literature Review of Cutaneous Alternariosis. Mycopathologia 2015 Feb;179(1):147–152. https://doi.org/10.1007/s11046-014-9822-9.
68. Jančič S, Nguyen HDT, Frisvad JC, Zalar P, Schroers HJ, Seifert KA, et al. A Taxonomic Revision of the Wallemia sebi Species Complex. PLOS ONE 2015 May;10(5):e0125933. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0125933.
69. Kjærbølling I, Vesth TC, Frisvad JC, Nybo JL, Theobald S, Kuo A, et al. Linking secondary metabolites to gene clusters through genome sequencing of six diverse Aspergillus species. Proceedings of the National Academy of Sciences 2018 Jan;115(4):E753–E761. https://www.pnas.org/content/115/4/E753.
70. Köhler JR, Casadevall A, Perfect J. The Spectrum of Fungi That Infects Humans. Cold Spring Harbor Perspectives in Medicine 2015 Jan;5(1):a019273. http://perspectivesinmedicine.cshlp.org/content/5/1/a019273.
71. Kwon-Chung KJ, Bennett JE, Wickes BL, Meyer W, Cuomo CA, Wollenburg KR, et al. The Case for Adopting the "Species Complex" Nomenclature for the Etiologic Agents of Cryptococcosis. mSphere 2017 Feb;2(1):e00357–16.
72. Paulussen C, Hallsworth JE, Álvarez Pérez S, Nierman WC, Hamill PG, Blain D, et al. Ecology of aspergillosis: insights into the pathogenic potency of Aspergillus fumigatus and some other Aspergillus species. Microbial Biotechnology 2017;10(2):296–322. https://onlinelibrary.wiley.com/doi/abs/10.1111/1751-7915.12367.
73. Prakash PY, Irinyi L, Halliday C, Chen S, Robert V, Meyer W. Online Databases for Taxonomy and Identification of Pathogenic Fungi and Proposal for a Cloud-Based Dynamic Data Network Platform. Journal of Clinical Microbiology 2017 Apr;55(4):1011–1024.
74. Dellière S, Rivero-Menendez O, Gautier C, Garcia-Hermoso D, Alastruey-Izquierdo A, Alanio A. Emerging mould infections: Get prepared to meet unexpected fungi in your patient. Medical Mycology 2020 Feb;58(2):156–162. https://doi.org/10.1093/mmy/myz039.
75. Teixeira MM, Moreno LF, Stielow BJ, Muszewska A, Hainaut M, Gonzaga L, et al. Exploring the genomic diversity of black yeasts and relatives (Chaetothyriales, Ascomycota). Studies in Mycology 2017 Mar;86:1–28. https://www.sciencedirect.com/science/article/pii/S0166061617300167.
76. Seyedmousavi S, Bosco SdMG, de Hoog S, Ebel F, Elad D, Gomes RR, et al. Fungal infections in animals: a patchwork of different situations. Medical Mycology 2018 Apr;56(suppl_1):165–187.
77. Smith JM. Fungal Pathogens of Nonhuman Animals. In: eLS American Cancer Society; 2006.https://onlinelibrary.wiley.com/doi/abs/10.1038/npg.els.0004235.
78. Bałazy S, Miętkiewski R, Tkaczuk C, Wegensteiner R, Wrzosek M. Diversity of acaropathogenic fungi in Poland and other European countries. Experimental and Applied Acarology 2008 Dec;46(1):53–70. https://doi.org/10.1007/s10493-008-9207-1.
79. Becnel JJ, Andreadis TG. Microsporidia in Insects. In: Microsporidia John Wiley & Sons, Ltd; 2014.p. 521–570. https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118395264.ch21.

80. Cissé OH, Ma L, Dekker JP, Khil PP, Youn JH, Brenchley JM, et al. Genomic insights into the host specific adaptation of the Pneumocystis genus. Communications Biology 2021 Mar;4(1):1–14. https://www.nature.com/articles/s42003-021-01799-7.
81. Chandler D, Davidson G, Pell JK, Ball BV, Shaw K, Sunderland KD. Fungal Biocontrol of Acari. Biocontrol Science and Technology 2000 Aug;10(4):357–384. https://doi.org/10.1080/09583150050114972.
82. Evans HC, Elliot SL, Hughes DP. Hidden Diversity Behind the Zombie-Ant Fungus Ophiocordyceps unilateralis: Four New Species Described from Carpenter Ants in Minas Gerais, Brazil. PLOS ONE 2011 Mar;6(3):e17024. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0017024.
83. Gerson U, Gafni A, Paz Z, Sztejnberg A. A tale of three acaropathogenic fungi in Israel: Hirsutella, Meira and Acaromyces. Experimental and Applied Acarology 2008 Dec;46(1):183–194. https://doi.org/10.1007/s10493-008-9202-6.
84. Schulenburg H, Félix MA. The Natural Biotic Environment of Caenorhabditis elegans. Genetics 2017 May;206(1):55–86. https://doi.org/10.1534/genetics.116.195511.
85. Leonhardt S, Büttner E, Gebauer AM, Hofrichter M, Kellner H. Draft Genome Sequence of the Sordariomycete Lecythophora (Coniochaeta) hoffmannii CBS 245.38. Genome Announcements 2018;6(7):e01510–17. https://journals.asm.org/doi/10.1128/genomeA.01510-17.
86. Lovett B, St Leger RJ. The Insect Pathogens. Microbiology Spectrum 2017 Mar;5(2):5.2.01. https://journals.asm.org/doi/full/10.1128/microbiolspec.FUNK-0001-2016.
87. Rehner SA, Buckley E. A Beauveria phylogeny inferred from nuclear ITS and EF1-alpha sequences: evidence for cryptic diversification and links to Cordyceps teleomorphs. Mycologia 2005 Feb;97(1):84–98.
88. Seyedmousavi S, Guillot J, Arné P, de Hoog GS, Mouton JW, Melchers WJG, et al. Aspergillus and aspergilloses in wild and domestic animals: a global health concern with parallels to human disease. Medical Mycology 2015 Nov;53(8):765–797. https://doi.org/10.1093/mmy/myv067.
89. Shang Y, Feng P, Wang C. Fungi That Infect Insects: Altering Host Behavior and Beyond. PLOS Pathogens 2015 Aug;11(8):e1005037. https://journals.plos.org/plospathogens/article?id=10.1371/journal.ppat.1005037.
90. Wu S, Toews MD, Castrillo LA, Barman AK, Cottrell TE, Shapiro-Ilan DI. Identification and Virulence of Cordyceps javanica Strain wf GA17 Isolated From a Natural Fungal Population in Sweetpotato Whiteflies (Hemiptera: Aleyrodidae). Environmental Entomology 2021 Oct;50(5):1127–1136. https://doi.org/10.1093/ee/nvab061.
91. St Leger RJ, Wang JB. Metarhizium: jack of all trades, master of many. Open Biology 2020;10(12):200307. https://royalsocietypublishing.org/doi/10.1098/rsob.200307.
92. Sung GH, Hywel-Jones NL, Sung JM, Luangsa-ard JJ, Shrestha B, Spatafora JW. Phylogenetic classification of Cordyceps and the clavicipitaceous fungi. Studies in Mycology 2007 Jan;57:5–59. https://www.sciencedirect.com/science/article/pii/S0166061614601305.
93. van der Geest LP, Elliot SL, Breeuwer JAJ, Beerling EAM. Diseases of Mites. Experimental & Applied Acarology 2000 Jul;24(7):497–560. https://doi.org/10.1023/A:1026518418163.
94. Vacher C, Piou D, Desprez-Loustau ML. Architecture of an Antagonistic Tree/Fungus Network: The Asymmetric Influence of Past Evolutionary History. PLOS ONE 2008 Mar;3(3):e1740. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0001740.
95. Doehlemann G, Ökmen B, Zhu W, Sharon A. Plant Pathogenic Fungi. Microbiology Spectrum 2017 Jan;5(1).
96. Barbara DJ, Clewes E. Plant pathogenic Verticillium species: how many of them are there? Molecular Plant Pathology 2003;4(4):297–305. https://onlinelibrary.wiley.com/doi/abs/10.1046/j.1364-3703.2003.00172.x.
97. Costa MM, Silva BAAS, Moreira GM, Pfenning LH. Colletotrichum falcatum and Fusarium species induce symptoms of red rot in sugarcane in Brazil. Plant Pathology 2021;70(8):1807–1818. https://onlinelibrary.wiley.com/doi/abs/10.1111/ppa.13423.
98. Coutinho IBL, Freire FCO, Lima CS, Lima JS, Gonçalves FJT, Machado AR, et al. Diversity of genus Lasiodiplodia associated with perennial tropical fruit plants in northeastern Brazil. Plant Pathology 2017;66(1):90–104. https://onlinelibrary.wiley.com/doi/abs/10.1111/ppa.12565.
99. Gurung S, Short DPG, Hu X, Sandoya GV, Hayes RJ, Koike ST, et al. Host Range of Verticillium isaacii and Verticillium klebahnii from Artichoke, Spinach, and Lettuce. Plant Disease 2015 Jul;99(7):933–938.
100. Jones DR, Baker RHA. Introductions of non-native plant pathogens into Great Britain, 1970–2004. Plant Pathology 2007;56(5):891–910. https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-3059.2007.01619.x.
101. Kim SW, Jung JH, Lamsal K, Kim YS, Min JS, Lee YS. Antifungal Effects of Silver Nanoparticles (AgNPs) against Various Plant Pathogenic Fungi. Mycobiology 2012 Mar;40(1):53–58. https://doi.org/10.5941/MYCO.2012.40.1.053.
102. Kim W, Cavinder B, Proctor RH, O'Donnell K, Townsend JP, Trail F. Comparative Genomics and Transcriptomics During Sexual Development Gives Insight Into the Life History of the Cosmopolitan Fungus Fusarium neocosmosporiellum. Frontiers in Microbiology 2019;10:1247. https://www.frontiersin.org/article/10.3389/fmicb.2019.01247.
103. Schardl CL, Young CA, Hesse U, Amyotte SG, Andreeva K, Calie PJ, et al. Plant-Symbiotic Fungi as Chemical Engineers: Multi-Genome Analysis of the Clavicipitaceae Reveals Dynamics of Alkaloid Loci. PLOS Genetics 2013 Feb;9(2):e1003323. https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1003323.
104. Stukenbrock EH, Quaedvlieg W, Javan-Nikhah M, Zala M, Crous PW, McDonald BA. Zymoseptoria ardabiliae and Z. pseudotritici, two progenitor species of the septoria tritici leaf blotch fungus Z. tritici (synonym: Mycosphaerella graminicola). Mycologia 2012 Dec;104(6):1397–1407.
105. Invasive Fungi. Systematic Mycology and Microbiology Laboratory, ARS, USDA; 2021. https://nt.ars-grin.gov/sbmlweb/fungi/diagnosticfactsheets.cfm, retrieved October 9, 2021.
106. de Hoog G, Guarro J, Gené J, Ahmed S, Al-Hatmi A, Figueras M, et al. Atlas of Clinical Fungi, 4th edition. Hilversum; 2020.
107. Dean R, Van Kan JAL, Pretorius ZA, Hammond-Kosack KE, Di PIEietro A, Spanu PD, et al. The Top 10 fungal pathogens in molecular plant pathology. Molecular Plant Pathology 2012 Apr;13(4):414–430. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6638784/.
108. Skamnioti P, Gurr SJ. Against the grain: safeguarding rice from rice blast disease. Trends in Biotechnology 2009 Mar;27(3):141–150. https://www.cell.com/trends/biotechnology/abstract/S0167-7799(09)00014-6.
109. Scheele BC, Pasmans F, Skerratt LF, Berger L, Martel A, Beukema W, et al. Amphibian fungal panzootic causes catastrophic and ongoing loss of biodiversity. Science 2019 Mar;363(6434):1459–1463. https://www.science.org/doi/10.1126/science.aav0379.
110. Holtgrewe M. Mason – A Read Simulator for Second Generation Sequencing Data. Technical Report FU Berlin 2010;http://publications.imp.fu-berlin.de/962/.
111. McInnes L, Healy J, Melville J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv:180203426 [cs, stat] 2020 Sep;http://arxiv.org/abs/1802.03426, arXiv: 1802.03426.

112. Zielezinski A, Barylski J, Karlowski WM. Taxonomy-aware, sequence similarity ranking reliably predicts phage–host relationships. BMC Biology 2021 Oct;19(1):223. https://doi.org/10.1186/s12915-021-01146-6.
113. Zielezinski A, Deorowicz S, Gudys A. PHIST: fast and accurate prediction of prokaryotic hosts from metagenomic viral sequences. bioRxiv 2021;https://www.biorxiv.org/content/early/2021/09/06/2021.09.06.459169.
114. Loka TP, Tausch SH, Dabrowski PW, Radonic A, Nitsche A, Renard BY. PriLive: privacy-preserving real-time filtering for next-generation sequencing. Bioinformatics (Oxford, England) 2018 Jul;34(14):2376–2383.
115. Ahmed O, Rossi M, Kovaka S, Schatz MC, Gagie T, Boucher C, et al. Pan-genomic matching statistics for targeted nanopore sequencing. iScience 2021 Jun;24(6):102696. https://www.sciencedirect.com/science/article/pii/S2589004221006647.
116. Garcia-Solache MA, Casadevall A. Global warming will bring new fungal diseases for mammals. mBio 2010 May;1(1):e00061–10.
117. Zielezinski A, Vinga S, Almeida J, Karlowski WM. Alignment-free sequence comparison: benefits, applications, and tools. Genome Biology 2017 Oct;18(1):186. https://doi.org/10.1186/s13059-017-1319-7.
118. Zhou H, Shrikumar A, Kundaje A. Towards a Better Understanding of Reverse-Complement Equivariance for Deep Learning Models in Regulatory Genomics. bioRxiv 2021 Feb;p. 2020.11.04.368803. https://www.biorxiv.org/content/10.1101/2020.11.04.368803v2.
119. Schiebenhoefer H, Bossche TVD, Fuchs S, Renard BY, Muth T, Martens L. Challenges and promise at the interface of metaproteomics and genomics: an overview of recent progress in metaproteogenomic data analysis. Expert Review of Proteomics 2019 May;16(5):375–390. https://doi.org/10.1080/14789450.2019.1609944.
120. Schiebenhoefer H, Schallert K, Renard BY, Trappe K, Schmid E, Benndorf D, et al. A complete and flexible workflow for metaproteomics data analysis based on MetaProteomeAnalyzer and Prophane. Nature Protocols 2020 Oct;15(10):3212–3239. https://www.nature.com/articles/s41596-020-0368-7.