Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Manuscript Category

Subject Section

Detecting DNA of novel fungal pathogens using ResNets and a curated fungi-hosts data collection

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

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

*To whom correspondence should be addressed. †Contributed equally.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

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 neural networks trained using the presented data collection enable accurate detection of novel pathogens from DNA inputs. A curated set of over 1,400 fungal genomes with host and pathogenicity annotation supports development of machine learning models and direct sequence comparison, not limited to the pathogen detection task.
Availability: The data, models and code are hosted at `https://zenodo.org/record/5846345`, `https://zenodo.org/record/5711877` and `https://gitlab.com/dacs-hpi/deepac`.

1 Introduction

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 vulvovaginal candidosis, affecting up to 75% of women at least once in their lifetime and often re-occurring multiple times (Sobel, 2007). It is also frequently found in healthy humans without leading to any disease, and has been reported to be capable of stable colonization (Raimondi *et al.*, 2019). However, invasive *Candida* infections, especially bloodstream infections, can reach mortality rates of up to 75%, rivaling those of bacterial and viral sepsis (Brown *et al.*, 2012). A related species, *Candida auris*, has been first recognized in a human patient in 2009 (Satoh *et al.*, 2009) and quickly became one of the most urgent threats among the drug-resistant pathogens (CDC, 2019), reaching mortality rates of up to 60% (Spivak and Hanson, 2018). It might have originally been a plant saprophyte which has adapted to avian, and then also mammalian hosts, possibly prompted by climate change (Casadevall *et al.*, 2019). Strikingly, it seems to have emerged in three different clonal populations on three continents at the same time, for reasons that currently remain unexplained (Lockhart *et al.*, 2017).

Despite their importance, research on fungal pathogens is consistently neglected and underfunded (noa, 2017). Even though fungal infections are estimated to kill 1.6 million people a year, they remain understudied and underreported (Chowdhary *et al.*, 2016; Huseyin *et al.*, 2017; noa, 2017). Estimates suggest that between 1.5 million (Hawksworth, 2001) and 5.1 million (Blackwell, 2011), or even 6 million (Taylor *et al.*, 2014) 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

© The Author 2021.

are based on recognition of known taxonomic units by homology detection, using either sequence alignment (Li and Durbin, 2010; Langmead and Salzberg, 2012; Li, 2018; Altschul *et al.*, 1990; Camacho *et al.*, 2009; Hong *et al.*, 2014; Naccache *et al.*, 2014; Ahn *et al.*, 2015; Andrusch *et al.*, 2018), k-mer based approaches (Wood *et al.*, 2019; Piro *et al.*, 2020; Breitwieser *et al.*, 2018) or combinations thereof (Piro *et al.*, 2017). 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 (Sayers *et al.*, 2021a), Database of Virulence Factors in Fungal Pathogens (Lu *et al.*, 2012) and the U.S. National Fungus Collections Fungus-Host Database (Farr and Rossman, 2021). 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 (Sayers *et al.*, 2021b) or FungiDB (Basenko *et al.*, 2018), which store genomic data without clear-cut host annotation. The ENHanCEd Infectious Diseases Database (EID2) (Wardeh *et al.*, 2015) 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 (Schoch *et al.*, 2020) and finding co-occurrences of species names in articles indexed by PubMed (Sayers *et al.*, 2021a), 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 *P. carinii*. While the latter name is now reserved for a species infecting exclusively rats and not humans (Stringer *et al.*, 2002), records in 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 (Deneke *et al.*, 2017; Mock *et al.*, 2020; Zhang *et al.*, 2019; Barash *et al.*, 2018; Guo *et al.*, 2021; Gałan *et al.*, 2019; Wardeh *et al.*, 2021; Brierley and Fowler, 2021; Bergner *et al.*, 2021; Tang *et al.*, 2015; Bartoszewicz *et al.*, 2020, 2021b). 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 (Richardson *et al.*, 2017; Burgess, 2017; Dai *et al.*, 2020), as well as optimization of yeast strains with new phenotypes (Szymanski and Calvert, 2018; Luo *et al.*, 2018). Genetic modifications of filamentous fungi also grow in popularity (Amores *et al.*, 2016; Martins-Santana *et al.*, 2018). As some of them are dangerous pathogens (Chowdhary *et al.*, 2016), 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 (Diggans and Leproust, 2019; Balaji *et al.*, 2021; Bartoszewicz *et al.*, 2021a).

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 (Altschul *et al.*, 1990; Camacho *et al.*, 2009) 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 (Breitwieser *et al.*, 2017; Ye *et al.*, 2019; Alser *et al.*, 2021), BLAST has been shown to be more accurate in similar tasks of detecting novel bacterial and viral pathogens (Deneke *et al.*, 2017; Bartoszewicz *et al.*, 2021b). However, convolutional neural networks of the DeePaC package have been proven to outperform BLAST in both those scenarios (Bartoszewicz *et al.*, 2020, 2021b) 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 (Bartoszewicz *et al.*, 2021a). 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 (Deneke *et al.*, 2017) and viruses (Bartoszewicz *et al.*, 2021b), 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.

2 Data description

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 (Taylor *et al.*, 2001; Chowdhary *et al.*, 2014; Bombassaro *et al.*, 2020; Brunner-Mendoza *et al.*, 2019; Franzen and Müller, 2001; Colombo *et al.*, 2011; Han and Weiss, 2017; Hu *et al.*, 2015; Jančič *et al.*, 2015; Kjærbølling *et al.*, 2018; Köhler *et al.*, 2015;

Kwon-Chung *et al.*, 2017; Paulussen *et al.*, 2017; Prakash *et al.*, 2017; Dellière *et al.*, 2020; Teixeira *et al.*, 2017; Seyedmousavi *et al.*, 2018; Smith, 2006; Bałazy *et al.*, 2008; Becnel and Andreadis, 2014; Franzen and Müller, 2001; Cissé *et al.*, 2021; Chandler *et al.*, 2000; Evans *et al.*, 2011; Gerson *et al.*, 2008; Han and Weiss, 2017; Schulenburg and Félix, 2017; Leonhardt *et al.*, 2018; Lovett and St. Leger, 2017; Rehner and Buckley, 2005; Seyedmousavi *et al.*, 2015; Shang *et al.*, 2015; Wu *et al.*, 2021; St. Leger and Wang, 2020; Sung *et al.*, 2007; Teixeira *et al.*, 2017; van der Geest *et al.*, 2000; Vacher *et al.*, 2008; Doehlemann *et al.*, 2017; Barbara and Clewes, 2003; Costa *et al.*, 2021; Coutinho *et al.*, 2017; Gurung *et al.*, 2015; Jones and Baker, 2007; Kim *et al.*, 2012, 2019; Leonhardt *et al.*, 2018; Schardl *et al.*, 2013; Stukenbrock *et al.*, 2012; de Hoog *et al.*, 2020) and database (Lu *et al.*, 2012; Schoch *et al.*, 2020; Farr and Rossman, 2021; Wardeh *et al.*, 2015) sources, relying on manual curation, but also including the automatically extracted data for future reference. Table S1 summarizes the data we collected for each species. We describe the curation procedure in detail in Supplementary Note 1. The database contains 14,555 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), available on October 9, 2021. This forms a collection of species most relevant to the pathogen detection task, belonging to 6 phyla, 37 classes, 82 orders and 182 families. A 'temporal benchmark' subset contains 15 further pathogens (including one infecting humans), collected in a database update on January 2, 2022. We also include records on 486 plant-associated fungi. The supplementary part of the database contains information on 481 putatively labelled genomes, 1,147 unlabelled species with available genomes, 8 synonyms (with 6 alternative genomes) derived from the Atlas of Clinical Fungi (de Hoog *et al.*, 2020), 885 labelled species without genomes (including 284 species without TaxIDs) and 10,579 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 (Wardeh *et al.*, 2015).

3 Methods

3.1 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. While evaluating performance on genuinely unknown species is by definition impossible, we effectively model the 'novel species' scenario by testing on a wide range of sequences removed from the database, including selected important species (Satoh *et al.*, 2009; Brown *et al.*, 2012; Dean *et al.*, 2012; Skamnioti and Gurr, 2009; Scheele *et al.*, 2019). We simulated low- and high-coverage read sets using a protocol outlined in Supplementary Note 2 (Bartoszewicz *et al.*, 2020; Holtgrewe, 2010). We considered two methods of balancing the number of samples between taxa: setting the read number per species to be proportional to the respective genome's length ('linear-size') or its logarithm ('logarithmic-size').

3.2 Phenotype prediction and genome representations

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 (Bartoszewicz *et al.*, 2021a). We also adapted a BLAST-based pipeline used by Bartoszewicz *et al.* (2020) for benchmarking. Details of the pipeline, the ResNet architecture and the hyperparameter tuning procedure are described in Supplementary Note 3.

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 (Bartoszewicz *et al.*, 2020). 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. As classifier outputs based on average activations indeed approximate classifier outputs averaged over all reads originating from a given species (see Supplementary Note 4 and 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 (McInnes *et al.*, 2020) 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.

3.3 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 (Bartoszewicz *et al.*, 2020, 2021b,a). 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 (Bartoszewicz *et al.*, 2021a) only by the output layer, which has as many units as the number of considered classes and uses a softmax activation (see Supplementary Note 3). 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 (Zielezinski *et al.*, 2021b,a) 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 (Loka *et al.*, 2018; Ahmed *et al.*, 2021; Wood *et al.*, 2019). Therefore, we fused the previously published datasets used in DeePaC (Bartoszewicz *et al.*, 2021a) for bacteria (pathogens vs. commensals) and viruses (human vs. non-human) with the 'positive' (human-pathogenic) class of our database (Supplementary Note 2). 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. 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 (Bartoszewicz *et al.*, 2021a), 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 'lost' by using a more open-view classifier. Note that in the case of the purely 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 (Bartoszewicz *et al.*, 2020), viral for viruses (Bartoszewicz *et al.*, 2021b)). 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.

4 Results

4.1 Fungal pathogenic potential prediction

The best network, trained on the high-coverage, 'logarithmic-size' dataset without dropout, required 8 days of training on four Tesla V100 GPUs 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 S2), 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 suboptimal for both BLAST and the ResNet, probably reflecting the extreme difficulty of the task (Supplementary Note 5, Table S3). The error estimates based on the held-out test dataset are consistent with the results of the temporal benchmark (Table S4). Balanced accuracy is much higher on full genomes (88.4–90.3), suggesting that the main performance bottleneck is the total amount of information (total sequence length) available as input. This is consistent with previous observations (Deneke *et al.*, 2017; Bartoszewicz *et al.*, 2020, 2021b), although more drastic than in the case of bacteria or viruses. All approaches correctly classify the single human pathogen in the temporal benchmark dataset, but the ResNet achieves higher specificity (92.9) than read-based (78.6) and contig-based BLAST (85.7).

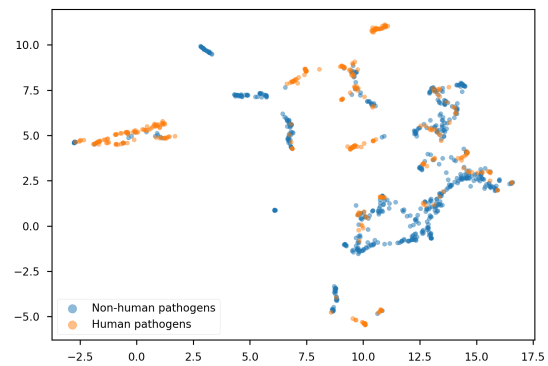

Fig. 1: UMAP embeddings of the learned genome representations for the core database, enlarged in Figure S5. Each point represents a genome of a single species, coloured by its ground truth label. The learned representations offer a way of visualizing the core database along relevant labels for each genome. The ResNet correctly classifies most of the genomes (Figure S6). The clusters are related, but not fully reducible to the taxonomic classification of the analysed species (Supplemetary Note 6, Figures S7–9).

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 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 (Supplementary Note 5, Table S3).

4.2 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. 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 S6). Classification errors seem to originate from an 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

Table 1. Performance of the multi-class classifiers on non-fungal datasets, read pairs.

		Acc.	Prec.	Rec.	Spec.	AUPR
Bacteria	4-class ensemble (ours)	85.1	**84.1**	87.2	**83.0**	88.3
	2-class (Bartoszewicz *et al.*, 2021a)	**87.3**	83.6	**92.7**	81.8	**89.0**
	BLAST	70.0	**84.1**	86.6	53.5	-
Viruses	4-class ensemble, reassigned (ours)	89.6	95.3	**89.3**	90.0	**98.1**
	2-class (Bartoszewicz *et al.*, 2021a)	**90.3**	94.7	85.5	**95.2**	97.2
	BLAST	80.6	**98.4**	79.1	82.2	-

Table 1. The 4-class classifier has been trained to detect non-pathogenic bacteria and non-human viruses, bacterial pathogens, human-infecting viruses, and human-pathogenic fungi. The 2-class classifiers are the original DeePaC ResNets (Bartoszewicz et al., 2021a) for bacteria and viruses accordingly. We use BLAST as shown in (Bartoszewicz et al., 2020, 2021b); since this only returns binary labels, AUPR is not defined. The multi-class models do not assume a purely bacterial or viral sample, so they are more flexible than the 2-class networks – a single classifier can be used for both datasets without retraining, although a minor cost in performance should be expected of a more general approach. For the viral dataset, assignments to the bacterial pathogens class may be assumed to actually reflect bacteriophages infecting those bacteria due to high bacteriophage-host sequence similarity, and reassigned accordingly. The multi-class models and the specialized binary classifiers perform comparably well, even though the latter can only be used in very specific situations. BLAST, using an appropriate reference database and representing the estimated upper bound on performance of homology-based approaches, is outperformed by a significant margin. Acc. – accuracy, Prec. – precision; Rec. – recall, Spec. – specificity, AUPR – area under the PR curve (calculated for the respective positive class).

Table 2. Performance on the multi-class dataset, read pairs.

		Acc.	F1	Prec.	Rec.	AUPR
All classes	4-class ens. (ours)	**87.6**	**87.7**	87.7	**87.6**	93.4
	BLAST	78.3	84.0	**90.6**	78.3	-
Non-pathogens	4-class ens. (ours)	**77.4**	**78.7**	**80.1**	**77.4**	86.7
	BLAST	66.5	71.6	77.5	66.5	-
Path. bacteria	4-class ens. (ours)	**87.2**	85.1	83.2	**87.2**	90.4
	BLAST	83.8	**87.5**	**91.6**	83.8	-
Human viruses	4-class ens. (ours)	**90.9**	**93.7**	96.7	**90.9**	98.4
	BLAST	78.9	87.9	**99.2**	78.9	-
Fungi	4-class ens. (ours)	**95.0**	**92.9**	90.9	**95.0**	97.9
	BLAST	84.1	88.9	**94.2**	84.1	-

Table 2. 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, recall is equal to accuracy for each class. We use the F1 score as an additional measure. As expected, BLAST's predictions are very precise, since when it finds a match, it is usually a relevant one. This does not hold for the non-pathogen class, which could indicate confusion between bacteriophage and bacterial pathogen reads. Our classifier significantly outperforms BLAST in terms of recall and prediction accuracy for all classes. BLAST, representing the estimated upper bound on performance of homology-based approaches, yields no predictions for 12.5% of all read pairs. Acc. – accuracy, F1 – F1 score, Prec. – precision, Rec. – recall, AUPR – area under the PR curve.

cluster 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, although this is importantly not a simple one-to-one mapping (Supplementary Note 6, Figure S7).

4.3 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 then 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 (Bartoszewicz *et al.*, 2021a) 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 S5, 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. They also outperform BLAST, which represents the estimated upper limit on the accuracy of pathogen detection by finding the closest taxonomic matches, by a wide margin (Table 1). 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 S6). 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, cutting the average error rate by over 40% compared to BLAST (Table 2). 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 (Bartoszewicz *et al.*, 2021a).

5 Discussion

Fungal pathogens have been under-studied compared to human-infecting bacteria and viruses, leading to repeated calls for more research in this area (Huseyin *et al.*, 2017; noa, 2017). 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 respective 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 (Garcia-Solache and Casadevall, 2010). 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 (Piro *et al.*, 2020) 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 1,455 labelled genomes to over 14,500 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 (Wardeh *et al.*, 2015) 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.

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 (Deneke *et al.*, 2017; Bartoszewicz *et al.*, 2020, 2021b,a). 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 possibly 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 (Bartoszewicz *et al.*, 2021a). Although we only evaluate them on simulated reads, this has been shown to approximate performance on real datasets very well (Bartoszewicz *et al.*, 2020, 2021b,a). Further work could extend the presented multi-class setup to Nanopore reads, as shown for bacterial and viral models (Bartoszewicz *et al.*, 2021a), enabling selective sequencing of mixed-pathogen samples.

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 4). 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 (Zielezinski *et al.*, 2017), including being used as input features for machine learning approaches as in Deneke *et al.* (2017). 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 (Zhou *et al.*, 2021). 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.

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 (Bartoszewicz *et al.*, 2021b). 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 a 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 (Renard *et al.*, 2012; Schiebenhoefer *et al.*, 2019, 2020).

6 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, hosted at `https://zenodo.org/record/5846345`, can be reused for future research on fungal pathogenicity based on both machine learning and other methods of computational biology. The models, read sets and code are available at `https://zenodo.org/record/5711877`, `https://zenodo.org/record/5846397`, and `https://github.com/dacs-hpi/deepac`.

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).

Conflict of interest: None declared.

Acknowledgements

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

References

(2017). Stop neglecting fungi. *Nature Microbiology*, **2**(8), 1–2.

Ahmed, O. *et al.* (2021). Pan-genomic matching statistics for targeted nanopore sequencing. *iScience*, **24**(6), 102696.

Ahn, T.-H. *et al.* (2015). Sigma: Strain-level inference of genomes from metagenomic analysis for biosurveillance. *Bioinformatics*, **31**(2), 170–177.

Alser, M. *et al.* (2021). Technology dictates algorithms: recent developments in read alignment. *Genome Biology*, **22**(1), 249.

Altschul, S. F. *et al.* (1990). Basic local alignment search tool. *Journal of Molecular Biology*, **215**(3), 403–410.

Amores, G. R. *et al.* (2016). Recent Progress on Systems and Synthetic Biology Approaches to Engineer Fungi As Microbial Cell Factories. *Current Genomics*, **17**(2), 85–98.

Andrusch, A. *et al.* (2018). PAIPline: pathogen identification in metagenomic and clinical next generation sequencing samples. *Bioinformatics*, **34**(17), i715–i721.

Balaji, A. *et al.* (2021). SeqScreen: Accurate and Sensitive Functional Screening of Pathogenic Sequences via Ensemble Learning. *bioRxiv*, page 2021.05.02.442344. Publisher: Cold Spring Harbor Laboratory Section: New Results.

Barash, E. *et al.* (2018). BacPaCS—Bacterial Pathogenicity Classification via Sparse-SVM. *Bioinformatics*.

Barbara, D. J. and Clewes, E. (2003). Plant pathogenic Verticillium species: how many of them are there? *Molecular Plant Pathology*, **4**(4), 297–305.

Bartoszewicz, J. M. *et al.* (2020). DeePaC: predicting pathogenic potential of novel DNA with reverse-complement neural networks. *Bioinformatics*, **36**(1), 81–89.

Bartoszewicz, J. M. *et al.* (2021a). Deep learning-based real-time detection of novel pathogens during sequencing. *Briefings in Bioinformatics*, (bbab269).

Bartoszewicz, J. M. *et al.* (2021b). Interpretable detection of novel human viruses from genome sequencing data. *NAR Genomics and Bioinformatics*, **3**(1).

Basenko, E. Y. *et al.* (2018). FungiDB: An Integrated Bioinformatic Resource for Fungi and Oomycetes. *Journal of Fungi*, **4**(1), 39.

Bałazy, S. *et al.* (2008). Diversity of acaropathogenic fungi in Poland and other European countries. *Experimental and Applied Acarology*, **46**(1), 53–70.

Becnel, J. J. and Andreadis, T. G. (2014). Microsporidia in Insects. In *Microsporidia*, pages 521–570. John Wiley & Sons, Ltd.

Bergner, L. M. *et al.* (2021). Characterizing and Evaluating the Zoonotic Potential of Novel Viruses Discovered in Vampire Bats. *Viruses*, **13**(2), 252.

Blackwell, M. (2011). The fungi: 1, 2, 3 … 5.1 million species? *American Journal of Botany*, **98**(3), 426–438.

Bombassaro, A. *et al.* (2020). Genomics and Virulence of Fonsecaea pugnacius, Agent of Disseminated Chromoblastomycosis. *Frontiers in Genetics*, **11**, 822.

Breitwieser, F. P. *et al.* (2017). A review of methods and databases for metagenomic classification and assembly. *Briefings in Bioinformatics*, **20**(4), 1125–1136.

Breitwieser, F. P. *et al.* (2018). KrakenUniq: confident and fast metagenomics classification using unique k-mer counts. *Genome Biology*, **19**(1), 198.

Brierley, L. and Fowler, A. (2021). Predicting the animal hosts of coronaviruses from compositional biases of spike protein and whole genome sequences through machine learning. *PLOS Pathogens*, **17**(4), e1009149.

Brown, G. D. *et al.* (2012). Hidden killers: Human fungal infections. *Science Translational Medicine*, **4**(165).

Brunner-Mendoza, C. *et al.* (2019). 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*, **29**(1), 83–102.

Burgess, D. J. (2017). Synthetic Biology: Building a custom eukaryotic genome de novo. *Nature Reviews Genetics*, **18**(5), 274–274.

Camacho, C. *et al.* (2009). BLAST+: architecture and applications. *BMC Bioinformatics*, **10**(1), 421.

Casadevall, A. *et al.* (2019). On the emergence of candida auris: climate change, azoles, swamps, and birds. *MBio*, **10**(4), e01397–19.

CDC (2019). *Antibiotic resistance threats in the United States, 2019*. U.S. Department of Health and Human Services, CDC.

Chandler, D. *et al.* (2000). Fungal Biocontrol of Acari. *Biocontrol Science and Technology*, **10**(4), 357–384.

Chowdhary, A. *et al.* (2014). Recognizing filamentous basidiomycetes as agents of human disease: A review. *Medical Mycology*, **52**(8), 782–797.

Chowdhary, A. *et al.* (2016). Filamentous Fungi in Respiratory Infections. What Lies Beyond Aspergillosis and Mucormycosis? *PLOS Pathogens*, **12**(4), e1005491.

Cissé, O. H. *et al.* (2021). Genomic insights into the host specific adaptation of the Pneumocystis genus. *Communications Biology*, **4**(1), 1–14.

Colombo, A. L. *et al.* (2011). Current knowledge of Trichosporon spp. and Trichosporonosis. *Clinical Microbiology Reviews*, **24**(4), 682–700.

Costa, M. M. *et al.* (2021). Colletotrichum falcatum and Fusarium species induce symptoms of red rot in sugarcane in Brazil. *Plant Pathology*, **70**(8), 1807–1818.

Coutinho, I. B. L. *et al.* (2017). Diversity of genus Lasiodiplodia associated with perennial tropical fruit plants in northeastern Brazil. *Plant Pathology*, **66**(1), 90–104.

Dai, J. *et al.* (2020). Sc3.0: revamping and minimizing the yeast genome. *Genome Biology*, **21**(1), 205.

de Hoog, G. *et al.* (2020). *Atlas of Clinical Fungi, 4th edition*. Hilversum.

Dean, R. *et al.* (2012). The Top 10 fungal pathogens in molecular plant pathology. *Molecular Plant Pathology*, **13**(4), 414–430.

Dellière, S. *et al.* (2020). Emerging mould infections: Get prepared to meet unexpected fungi in your patient. *Medical Mycology*, **58**(2), 156–162.

Deneke, C. *et al.* (2017). PaPrBaG: A machine learning approach for the detection of novel pathogens from NGS data. *Scientific Reports*, **7**, 39194.

Diggans, J. and Leproust, E. (2019). Next Steps for Access to Safe, Secure DNA Synthesis. *Frontiers in Bioengineering and Biotechnology*, **7**.

Doehlemann, G. *et al.* (2017). Plant Pathogenic Fungi. *Microbiology Spectrum*, **5**(1).

Evans, H. C. *et al.* (2011). Hidden Diversity Behind the Zombie-Ant Fungus Ophiocordyceps unilateralis: Four New Species Described from Carpenter Ants in Minas Gerais, Brazil. *PLOS ONE*, **6**(3), e17024.

Farr, D. F. and Rossman, A. Y. (2021). Fungal databases. Retrieved October 9, 2021 and January 2, 2022.

Franzen, C. and Müller, A. (2001). Microsporidiosis: human diseases and diagnosis. *Microbes and Infection*, **3**(5), 389–400.

Garcia-Solache, M. A. and Casadevall, A. (2010). Global warming will bring new fungal diseases for mammals. *mBio*, **1**(1), e00061–10.

Gałan, W. *et al.* (2019). Host Taxon Predictor - A Tool for Predicting Taxon of the Host of a Newly Discovered Virus. *Scientific Reports*, **9**(1), 3436.

Gerson, U. *et al.* (2008). A tale of three acaropathogenic fungi in Israel: Hirsutella, Meira and Acaromyces. *Experimental and Applied Acarology*, **46**(1), 183–194.

Guo, Q. *et al.* (2021). Predicting Hosts Based on Early SARS-CoV-2 Samples and Analyzing Later World-wide Pandemic in 2020. *bioRxiv*, page 2021.03.21.436312.

Gurung, S. *et al.* (2015). Host Range of Verticillium isaacii and Verticillium klebahnii from Artichoke, Spinach, and Lettuce. *Plant Disease*, **99**(7), 933–938.

Han, B. and Weiss, L. M. (2017). Microsporidia: Obligate Intracellular Pathogens Within the Fungal Kingdom. *Microbiology Spectrum*, **5**(2).

Hawksworth, D. L. (2001). The magnitude of fungal diversity: the 1.5 million species estimate revisited. *Mycological Research*, **105**(12), 1422–1432.

Holtgrewe, M. (2010). Mason – a read simulator for second generation sequencing data. *Technical Report FU Berlin*.

Hong, C. *et al.* (2014). PathoScope 2.0: a complete computational framework for strain identification in environmental or clinical sequencing samples. *Microbiome*, **2**(1), 33.

Hu, W. *et al.* (2015). Alternaria arborescens Infection in a Healthy Individual and Literature Review of Cutaneous Alternariosis. *Mycopathologia*, **179**(1), 147–152.

Huseyin, C. E. *et al.* (2017). Forgotten fungi—the gut mycobiome in human health and disease. *FEMS Microbiology Reviews*, **41**(4), 479–511.

Jančič, S. *et al.* (2015). A Taxonomic Revision of the Wallemia sebi Species Complex. *PLOS ONE*, **10**(5), e0125933.

Jones, D. R. and Baker, R. H. A. (2007). Introductions of non-native plant pathogens into Great Britain, 1970–2004. *Plant Pathology*, **56**(5), 891–910.

Kim, S. W. *et al.* (2012). Antifungal Effects of Silver Nanoparticles (AgNPs) against Various Plant Pathogenic Fungi. *Mycobiology*, **40**(1), 53–58.

Kim, W. *et al.* (2019). Comparative Genomics and Transcriptomics During Sexual Development Gives Insight Into the Life History of the Cosmopolitan Fungus Fusarium neocosmosporiellum. *Frontiers in Microbiology*, **10**, 1247.

Kjærbølling, I. *et al.* (2018). Linking secondary metabolites to gene clusters through genome sequencing of six diverse Aspergillus species. *Proceedings of the National Academy of Sciences*, **115**(4), E753–E761.

Kwon-Chung, K. J. *et al.* (2017). The Case for Adopting the "Species Complex" Nomenclature for the Etiologic Agents of Cryptococcosis. *mSphere*, **2**(1), e00357–16.

Köhler, J. R. *et al.* (2015). The Spectrum of Fungi That Infects Humans. *Cold Spring Harbor Perspectives in Medicine*, **5**(1), a019273.

Langmead, B. and Salzberg, S. L. (2012). Fast gapped-read alignment with bowtie 2. *Nature methods*, **9**(4), 357–359.

Leonhardt, S. *et al.* (2018). Draft Genome Sequence of the Sordariomycete Lecythophora (Coniochaeta) hoffmannii CBS 245.38. *Genome Announcements*, **6**(7), e01510–17.

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, **34**(18), 3094–3100.

Li, H. and Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*, **26**(5), 589–595.

Lockhart, S. R. *et al.* (2017). 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*, **64**(2), 134–140.

Loka, T. P. *et al.* (2018). PriLive: privacy-preserving real-time filtering for next-generation sequencing. *Bioinformatics (Oxford, England)*, **34**(14), 2376–2383.

Lovett, B. and St. Leger, R. J. (2017). The Insect Pathogens. *Microbiology Spectrum*, **5**(2), 5.2.01.

Lu, T. *et al.* (2012). DFVF: database of fungal virulence factors. *Database*, **2012**(bas032).

Luo, Z. *et al.* (2018). Identifying and characterizing SCRaMbLEd synthetic yeast using ReSCuES. *Nature Communications*, **9**(1), 1930.

Martins-Santana, L. *et al.* (2018). Systems and Synthetic Biology Approaches to Engineer Fungi for Fine Chemical Production. *Frontiers in Bioengineering and Biotechnology*, **6**.

McInnes, L. *et al.* (2020). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv:1802.03426 [cs, stat]*. arXiv: 1802.03426.

Mock, F. *et al.* (2020). VIDHOP, viral host prediction with Deep Learning. *Bioinformatics*.

Naccache, S. N. *et al.* (2014). A cloud-compatible bioinformatics pipeline for ultrarapid pathogen identification from next-generation sequencing of clinical samples. *Genome Research*, **24**(7), 1180–1192.

Paulussen, C. *et al.* (2017). Ecology of aspergillosis: insights into the pathogenic potency of Aspergillus fumigatus and some other Aspergillus

species. *Microbial Biotechnology*, **10**(2), 296–322.

Piro, V. C. *et al.* (2017). Metameta: integrating metagenome analysis tools to improve taxonomic profiling. *Microbiome*, **5**(1), 1–11.

Piro, V. C. *et al.* (2020). ganon: precise metagenomics classification against large and up-to-date sets of reference sequences. *Bioinformatics*, **36**(Supplement_1), i12–i20.

Prakash, P. Y. *et al.* (2017). Online Databases for Taxonomy and Identification of Pathogenic Fungi and Proposal for a Cloud-Based Dynamic Data Network Platform. *Journal of Clinical Microbiology*, **55**(4), 1011–1024.

Raimondi, S. *et al.* (2019). Longitudinal survey of fungi in the human gut: Its profiling, phenotyping, and colonization. *Frontiers in Microbiology*, **10**.

Rehner, S. A. and Buckley, E. (2005). A Beauveria phylogeny inferred from nuclear ITS and EF1-alpha sequences: evidence for cryptic diversification and links to Cordyceps teleomorphs. *Mycologia*, **97**(1), 84–98.

Renard, B. Y. *et al.* (2012). Overcoming species boundaries in peptide identification with bayesian information criterion-driven error-tolerant peptide search (biceps). *Molecular & cellular proteomics*, **11**(7), M111–014167.

Richardson, S. M. *et al.* (2017). Design of a synthetic yeast genome. *Science*, **355**(6329), 1040–1044.

Satoh, K. *et al.* (2009). Candida auris sp. nov., a novel ascomycetous yeast isolated from the external ear canal of an inpatient in a Japanese hospital. *Microbiology and Immunology*, **53**(1), 41–44.

Sayers, E. W. *et al.* (2021a). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, **49**(D1), D10–D17.

Sayers, E. W. *et al.* (2021b). GenBank. *Nucleic Acids Research*, **49**(D1), D92–D96.

Schardl, C. L. *et al.* (2013). Plant-Symbiotic Fungi as Chemical Engineers: Multi-Genome Analysis of the Clavicipitaceae Reveals Dynamics of Alkaloid Loci. *PLOS Genetics*, **9**(2), e1003323.

Scheele, B. C. *et al.* (2019). Amphibian fungal panzootic causes catastrophic and ongoing loss of biodiversity. *Science*, **363**(6434), 1459–1463.

Schiebenhoefer, H. *et al.* (2019). Challenges and promise at the interface of metaproteomics and genomics: an overview of recent progress in metaproteogenomic data analysis. *Expert Review of Proteomics*, **16**(5), 375–390.

Schiebenhoefer, H. *et al.* (2020). A complete and flexible workflow for metaproteomics data analysis based on MetaProteomeAnalyzer and Prophane. *Nature Protocols*, **15**(10), 3212–3239.

Schoch, C. L. *et al.* (2020). NCBI taxonomy: a comprehensive update on curation, resources and tools. *Database*, **2020**.

Schulenburg, H. and Félix, M.-A. (2017). The Natural Biotic Environment of Caenorhabditis elegans. *Genetics*, **206**(1), 55–86.

Seyedmousavi, S. *et al.* (2015). Aspergillus and aspergilloses in wild and domestic animals: a global health concern with parallels to human disease. *Medical Mycology*, **53**(8), 765–797.

Seyedmousavi, S. *et al.* (2018). Fungal infections in animals: a patchwork of different situations. *Medical Mycology*, **56**(suppl_1), 165–187.

Shang, Y. *et al.* (2015). Fungi That Infect Insects: Altering Host Behavior and Beyond. *PLOS Pathogens*, **11**(8), e1005037.

Skamnioti, P. and Gurr, S. J. (2009). Against the grain: safeguarding rice from rice blast disease. *Trends in Biotechnology*, **27**(3), 141–150.

Smith, J. M. (2006). Fungal Pathogens of Nonhuman Animals. In *eLS*. American Cancer Society.

Sobel, J. D. (2007). Vulvovaginal candidosis. *The Lancet*, **369**(9577), 1961–1971.

Spivak, E. S. and Hanson, K. E. (2018). Candida auris: an emerging fungal pathogen. *Journal of clinical microbiology*, **56**(2).

St. Leger, R. J. and Wang, J. B. (2020). Metarhizium: jack of all trades, master of many. *Open Biology*, **10**(12), 200307.

Stringer, J. R. *et al.* (2002). A New Name for Pneumocystis from Humans and New Perspectives on the Host-Pathogen Relationship. *Emerging Infectious Diseases*, **8**(9), 891–896.

Stukenbrock, E. H. *et al.* (2012). Zymoseptoria ardabiliae and Z. pseudotritici, two progenitor species of the septoria tritici leaf blotch fungus Z. tritici (synonym: Mycosphaerella graminicola). *Mycologia*, **104**(6), 1397–1407.

Sung, G.-H. *et al.* (2007). Phylogenetic classification of Cordyceps and the clavicipitaceous fungi. *Studies in Mycology*, **57**, 5–59.

Szymanski, E. and Calvert, J. (2018). Designing with living systems in the synthetic yeast project. *Nature Communications*, **9**(1), 2950.

Tang, Q. *et al.* (2015). Inferring the hosts of coronavirus using dual statistical models based on nucleotide composition. *Scientific Reports*, **5**(1), 17155.

Taylor, D. L. *et al.* (2014). A first comprehensive census of fungi in soil reveals both hyperdiversity and fine-scale niche partitioning. *Ecological Monographs*, **84**(1), 3–20.

Taylor, L. H. *et al.* (2001). Risk factors for human disease emergence. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, **356**(1411), 983–989.

Teixeira, M. M. *et al.* (2017). Exploring the genomic diversity of black yeasts and relatives (Chaetothyriales, Ascomycota). *Studies in Mycology*, **86**, 1–28.

Vacher, C. *et al.* (2008). Architecture of an Antagonistic Tree/Fungus Network: The Asymmetric Influence of Past Evolutionary History. *PLOS ONE*, **3**(3), e1740.

van der Geest, L. *et al.* (2000). Diseases of Mites. *Experimental & Applied Acarology*, **24**(7), 497–560.

Wardeh, M. *et al.* (2015). Database of host-pathogen and related species interactions, and their global distribution. *Scientific Data*, **2**(1), 150049.

Wardeh, M. *et al.* (2021). Predicting mammalian hosts in which novel coronaviruses can be generated. *Nature Communications*, **12**(1), 780.

Wood, D. E. *et al.* (2019). Improved metagenomic analysis with Kraken 2. *Genome Biology*, **20**(1), 257.

Wu, S. *et al.* (2021). Identification and Virulence of Cordyceps javanica Strain wf GA17 Isolated From a Natural Fungal Population in Sweetpotato Whiteflies (Hemiptera: Aleyrodidae). *Environmental Entomology*, **50**(5), 1127–1136.

Ye, S. H. *et al.* (2019). Benchmarking Metagenomics Tools for Taxonomic Classification. *Cell*, **178**(4), 779–794.

Zhang, Z. *et al.* (2019). Rapid identification of human-infecting viruses. *Transboundary and Emerging Diseases*, **66**(6), 2517–2522.

Zhou, H. *et al.* (2021). Towards a Better Understanding of Reverse-Complement Equivariance for Deep Learning Models in Regulatory Genomics. *bioRxiv*, page 2020.11.04.368803.

Zielezinski, A. *et al.* (2017). Alignment-free sequence comparison: benefits, applications, and tools. *Genome Biology*, **18**(1), 186.

Zielezinski, A. *et al.* (2021a). PHIST: fast and accurate prediction of prokaryotic hosts from metagenomic viral sequences. *bioRxiv*.

Zielezinski, A. *et al.* (2021b). Taxonomy-aware, sequence similarity ranking reliably predicts phage–host relationships. *BMC Biology*, **19**(1), 223.