




RESEARCH ARTICLE

# Finding *Candida auris* in public metagenomic repositories

Jorge E. Mario-Vasquez<sup>1</sup> , Ujwal R. Bagal<sup>2</sup>, Elijah Lowe<sup>3</sup>, Aleksandr Morgulis<sup>4</sup>, John Phan<sup>3</sup>, D. Joseph Sexton<sup>1</sup> , Sergey Shirayev<sup>4</sup>, Rytis Slatkevičius<sup>5</sup>, Rory Welsh<sup>1</sup>, Anastasia P. Litvintseva<sup>1</sup>, Matthew Blumberg<sup>5</sup>, Richa Agarwala<sup>4</sup>, Nancy A. Chow<sup>1</sup> \*

**1** Mycotic Diseases Branch, Centers for Disease Control and Prevention, Atlanta, Georgia, United States of America, **2** ASRT Inc., Atlanta, Georgia, United States of America, **3** General Dynamics Information Technology Inc., Atlanta, Georgia, United States of America, **4** National Center for Biotechnology Information, Bethesda, Maryland, United States of America, **5** GridRepublic, Cambridge, Massachusetts, United States of America

\* [yln3@cdc.gov](mailto:yln3@cdc.gov)



## OPEN ACCESS

**Citation:** Mario-Vasquez JE, Bagal UR, Lowe E, Morgulis A, Phan J, Sexton DJ, et al. (2024) Finding *Candida auris* in public metagenomic repositories. PLoS ONE 19(1): e0291406. <https://doi.org/10.1371/journal.pone.0291406>

**Editor:** Ricardo Santos, Universidade Lisboa, Instituto superior Técnico, PORTUGAL

**Received:** August 28, 2023

**Accepted:** January 4, 2024

**Published:** January 19, 2024

**Peer Review History:** PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: <https://doi.org/10.1371/journal.pone.0291406>

**Copyright:** This is an open access article, free of all copyright, and may be freely reproduced, distributed, transmitted, modified, built upon, or otherwise used by anyone for any lawful purpose. The work is made available under the [Creative Commons CC0](https://creativecommons.org/licenses/by/4.0/) public domain dedication.

**Data Availability Statement:** SRPRISM is available at <https://github.com/ncbi/SRPRISM>. Software for rank calculation, sequence data for reference genomes, and a sample run have been deposited at Zenodo.org and are available at <https://doi.org/10.5281/zenodo.10000000>.

## Abstract

*Candida auris* is a newly emerged multidrug-resistant fungus capable of causing invasive infections with high mortality. Despite intense efforts to understand how this pathogen rapidly emerged and spread worldwide, its environmental reservoirs are poorly understood. Here, we present a collaborative effort between the U.S. Centers for Disease Control and Prevention, the National Center for Biotechnology Information, and GridRepublic (a volunteer computing platform) to identify *C. auris* sequences in publicly available metagenomic datasets. We developed the MetaNISH pipeline that uses SRPRISM to align sequences to a set of reference genomes and computes a score for each reference genome. We used MetaNISH to scan ~300,000 SRA metagenomic runs from 2010 onwards and identified five datasets containing *C. auris* reads. Finally, GridRepublic has implemented a prospective *C. auris* molecular monitoring system using MetaNISH and volunteer computing.

## Introduction

*Candida auris* is an emerging and often multidrug-resistant yeast that can cause invasive candidiasis, a life-threatening disease with high mortality [1]. The World Health Organization (WHO) classified *C. auris* as a critical priority pathogen due to its high outbreak potential, resistance to most available antifungal medicines, and ability to persist in the healthcare environment despite intensive infection prevention strategies [2].

Although the pathogen was first described in Japan in 2009 [3], the earliest known *C. auris* isolates were retrospectively identified and date back to 1996 in South Korea [4]. Whole-genome sequencing (WGS) of *C. auris* isolates from four world regions revealed four phylogenetically distinct clades of this fungal pathogen wherein isolates clustered geographically (Clade I, South Asia; Clade II, East Asia; Clade III, Africa; and Clade IV, South America). This finding supported the hypothesis that *C. auris* emerged independently and simultaneously in geographically separated human populations [5]. WGS of a recently identified isolate from

5281/zenodo.10214980. The benchmark dataset is available at NCBI under BioProject PRJNA631031. All additional data analyzed in the manuscript is already publicly available in SRA at NCBI (<https://www.ncbi.nlm.nih.gov/sra>).

**Funding:** The authors received no specific funding for this work.

**Competing interests:** The authors have declared that no competing interests exist.

Iran showed the existence of a fifth major clade, with hundreds of thousands of single nucleotide polymorphisms (SNPs) separating this isolate from the four known clades [6,7]. The five major clades are separated by tens to hundreds of thousands of SNPs. Within each clade, isolates from varying countries are typically separated by hundreds to thousands of SNPs [8].

Despite intense efforts to understand how this pathogen emerged and spread to healthcare facilities worldwide, the natural reservoirs of *C. auris* are poorly understood. Two alternative hypotheses have been proposed to explain the origin of *C. auris*. One suggests that *C. auris* existed in the environment before clinical recognition and emerged as a human pathogen due to thermal adaptation in response to environmental changes [9]. Several biological properties of *C. auris*, such as thermotolerance and halotolerance, that allow this fungus to survive in hypersaline environments provide indirect evidence supporting this theory [10]. The other hypothesis is based on the *C. auris* unique propensity to colonize human skin [11,12] and suggests that *C. auris* might have existed as a minor skin commensal colonizing poorly studied sites on the human body, such as the external ear canal, in isolated human populations and emerged globally in response to the increased use of antifungals in medical and agricultural practices [10,13]. This hypothesis is indirectly supported by the results of molecular dating, which showed that the emergence of outbreak causing strains in three different lineages (Clade I, Clade III, and Clade IV) coincided with the introduction of azoles into clinics and agriculture [13]. Of course, other potential explanations are also possible, and more research is needed to better understand the environmental and human reservoirs of this pathogen. Two recent publications reported the isolation of *C. auris* from a salt marsh and sandy beach on the Andaman Islands in India and Colombia's coastal estuaries [14,15]. These findings suggest a need for further environmental and human microbiome evaluations.

To conduct extensive environmental evaluations in a financially and logistically feasible manner, investigators have utilized metagenomic data in public repositories [16] like the Sequence Read Archive (SRA), the largest global sequence repository [17]. In this study, the U. S. Centers for Disease Control and Prevention (CDC), the National Center for Biotechnology Information (NCBI), and GridRepublic partnered to develop MetaNISH (**Metagenomic Needles In Sequence Hay**) and pipelines that utilize it. With these pipelines, we retrospectively screened ~300,000 shotgun metagenomic SRA runs from 2010 to 2022 to identify and describe datasets containing *C. auris*. In addition, we started prospectively screening datasets for this fungal pathogen daily in April 2023.

## Materials and methods

### MetaNISH design

NCBI developed the MetaNISH pipeline to screen metagenomic read sets for the presence of each genome in the given set of reference genomes (see benchmark development section). The pipeline consists of two steps: (I) the alignment step using SRPRISM [18] that aligns reads to all reference genomes as a single database, and (II) the score computation step, which increases the score for samples with reads aligned across the genome compared to samples with reads aligned to a small section of the genome.

Alignment is performed with SRPRISM as it guarantees the reporting of all equally good alignments (max 255) across all sequences in the database. Additionally, it supports specifying the region on the reads that must align and a maximum number of errors (mismatches, insertions, or deletions) in the reported alignments. For this study, we required SRPRISM to align the first 100 bases of the reads and specified a maximum of 15 errors for the reported alignments. For reads shorter than 100 bases, the full length of the read is aligned. The design of SRPRISM guarantees that the first 100 bases will have at most 5 of the 15 errors allowed. We

chose the first 100 bases as the region that must align as the read quality drops beyond that in many Illumina runs, and 100 bp is also long enough to avoid spurious matches. An example of such a read is SRR11734778.40769.1, which is a paired read with each mate of length 251 bases. Alignments for this read are included in the data released at Zenodo.org (<https://doi.org/10.5281/zenodo.10214980>). It was shown that the first 163 bases of the read were an exact match to *C. auris* genomes. However, the remaining portion of the read (specifically, the substring from 164 to 251) seems of inferior quality as it did not report a match to anything in the non-redundant (nr) nucleotide database at NCBI as of November 2023.

Score computation was devised for metagenomic read sets where the depth of coverage by reads could vary considerably over the genome. To determine the extent to which reads were aligned with the genome, regardless of coverage at aligned regions, a padding of up to 100 Kb was added to either end of the genome region aligned by the read. A scaling factor was applied to adjust the padding length for each read and genome based on the number of alignments, so this ensures that multiple mappings of a read to a genome do not exceed 100 Kb for each location. The score was the percentage of the genome covered by padded alignments. For example, read SRR11734778.40769.1 aligned to four *C. auris* genomes with only one location in each genome. Therefore, full padding of 100 Kb was used for each of the four alignments. However, read SRR11734778.1429600.1 aligned to four contigs on three genomes, as shown in Fig 1. Fig 1A and 1B show four alignments on two genomes, which reduced the padding to 25 Kb. Fig 1C shows two alignments on the third genome, which reduced the padding to 50 Kb. Padded coverage cannot extend beyond contig boundaries.

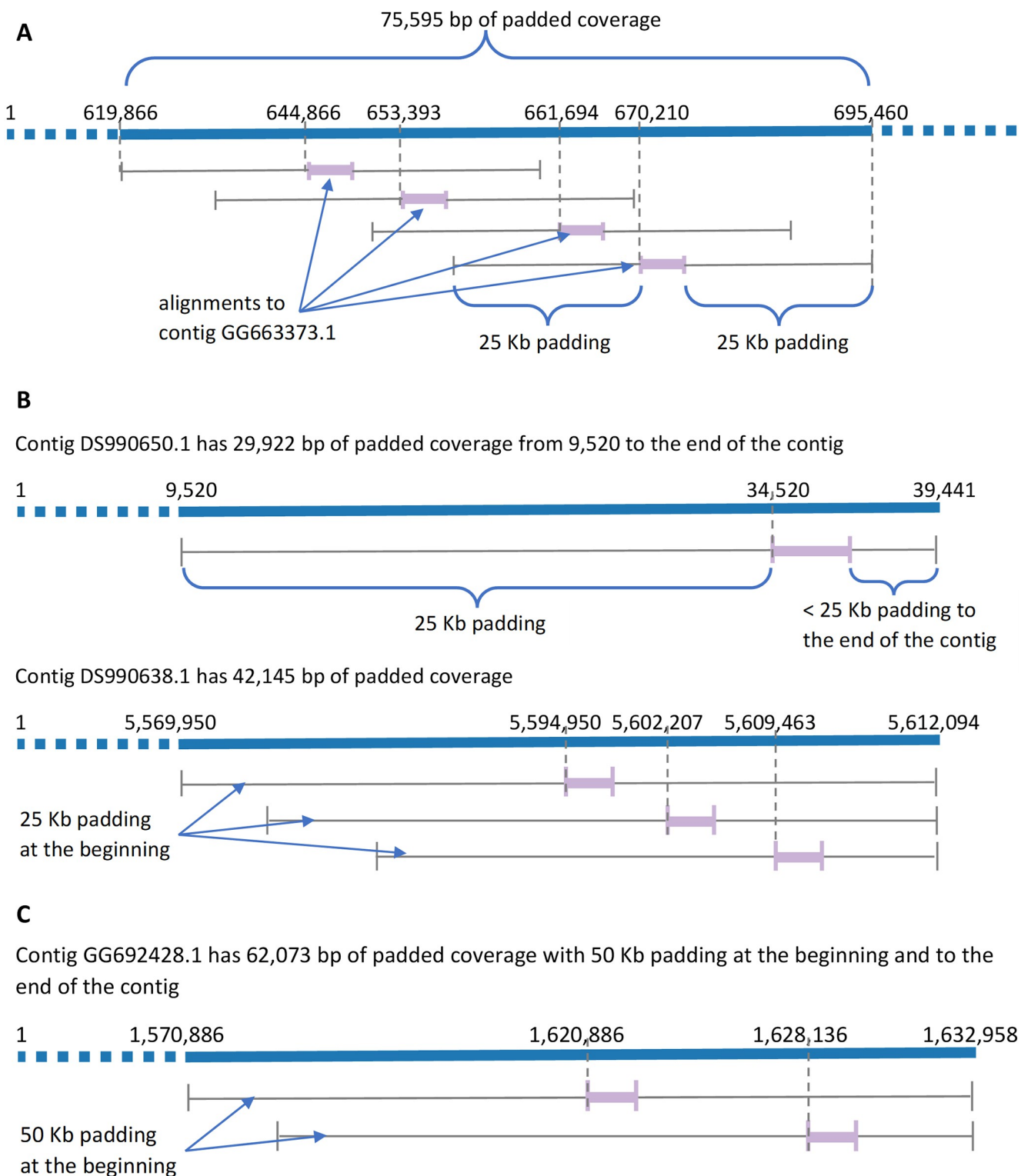
Using the reference genomes provided and empirical data from a set of 4,000 SRA runs, we proposed padding of 100 Kb and a score of at least 75 to indicate the presence of the corresponding genome in the read set. The choice is conservative, where it can potentially flag a few runs as scoring at least 75 when the genome is not present (false positives) but is unlikely to miss any (false negatives). At the same time, the parameters are not too conservative to make the false positive rate unacceptable. An example that illustrates the importance of padding regardless of the number of reads aligned at any genome location is SRR9016983. The alignment of reads from SRR9016983 to the reference genome B12037.1 is 2.8% across 1,975 locations throughout the genome. Among these reads, 50.4% have 1X coverage, and 38.7% have 2X coverage. Adding 100 Kb padding to these adjacent alignments allows the coverage to reach 100%, thus increasing the likelihood of detection.

The CPU time for the 4,000 runs used for determining the parameters varied from 25 seconds (for SRR7125652) to 17 hours 14 minutes (for SRR8550535) with a median time of 25 minutes. SRR7125652 has 86,554 paired reads, while SRR8550535 has over 418 million paired reads. We noted that SRPRISM, which takes almost all the time in the MetaNISH pipeline (as score calculation takes only a couple of seconds), can be run in multi-threaded mode with good scaling till eight threads, but we did not use that option for results reported here.

The design of MetaNISH can be used for tracking any pathogen. It requires developing a reference set with representative genomes from all clades for the pathogen to be tracked and for nearby species. Doing so allows SRPRISM to find the best matches for each read among the genomes where a match can be expected. Then, empirical analysis is needed to find suitable parameters for padding and score threshold. If read properties change substantially over time, the alignment method and parameters may need revisited.

## Benchmark development

CDC collated a set of 100 reference genomes representing priority pathogens for fungal molecular surveillance, including a representative subset of genomes for *C. auris* [19,20]. Specifically,



**Fig 1. Padding in alignments of read SRR11734778.1429600.1 on different assemblies.** (A) Assembly GCA\_000150115.1 (B) Assembly GCA\_000151005.2 (C) Assembly GCA\_000151035.1.

<https://doi.org/10.1371/journal.pone.0291406.g001>

14 genomes were of *C. auris*, 44 were of other *Candida* species, and 42 were of other fungal genera (i.e., *Ajellomyces*, *Blastomyces*, *Clavispora*, *Coccidioides*, *Cryptococcus*, *Emergomycetes*, *Emmonsia*, *Paracoccidioides*, *Pichia*, *Pneumocystis*, *Saitozyma*, *Sporothrix*, and *Talaromyces*). Detailed information regarding assemblies used are found in [S1 Table](#). The data released for this paper also includes sequences for all 100 reference genomes.

For the benchmark dataset, sequencing data for 20 metagenomic runs were generated by sequencing clinical specimens with *C. auris* spiked in at various concentrations. Briefly, residual material from *C. auris* colonization screening swabs collected from the anterior nares were used as a benchmarking dataset. The qualitative presence of *C. auris* was first confirmed by enrichment broth culture [21]. The concentration of *C. auris* cells was then assessed through a quantitative Sybr Green qPCR as previously described [22]. Cell concentrations were interpolated from a standard curve built using samples spiked with *C. auris* AR 0385 at serial dilutions ranging between  $10^7$  CFU/mL to  $10^3$  CFU/mL. Concentrations in the standard curve were confirmed by CFU counts and tested with three biological replicates at each concentration. The melt curve was referenced for both standard curve and benchmark samples to confirm that a strong melt peak was present in positive samples at  $\sim 83$ – $84^\circ\text{C}$ , the signature temperature indicative of *C. auris*. No unspecific amplification was observed in the standard curve or benchmarking samples. Five "no template controls" were included in the run. As expected, there was no amplification in these samples. Sequence data were deposited in PRJNA631031.

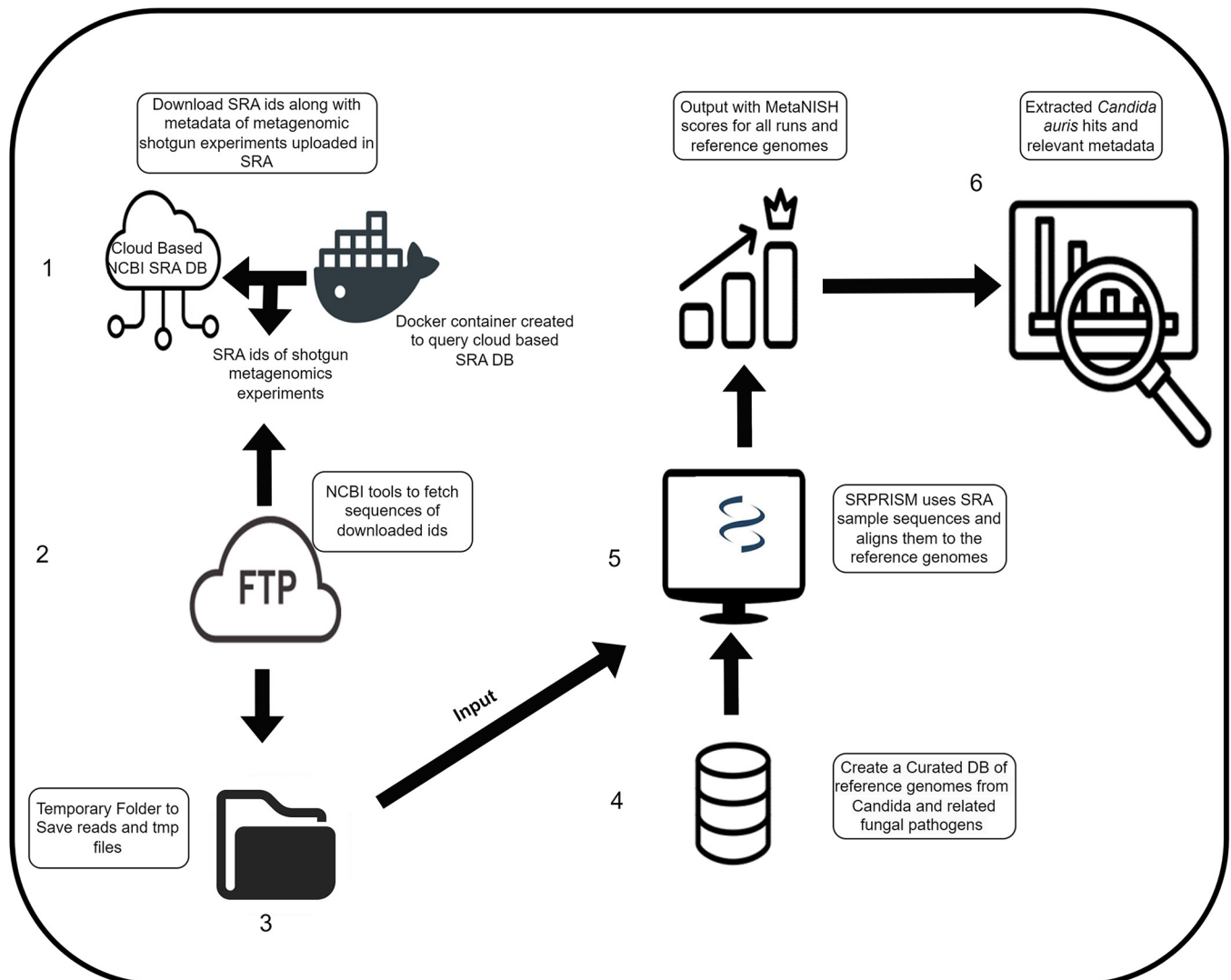
## MetaNISH implementation

The bash pipeline used by CDC for searching NCBI's SRA database integrated with MetaNISH is depicted in [Fig 2](#). Following filtering criteria were applied (*platform*: Illumina; *library\_source*: metagenomic; *consent*: public; *assay\_type*: WGS; *library\_selection*: random) to download only whole-genome sequence metagenomic datasets temporarily with additional metadata (accession ID, biosample, bioproject, release date, library layout, mbases, and organism). The alignment and scoring were done as per the MetaNISH design described earlier. The scores for all 100 reference genomes for each SRA ID are reported by the pipeline.

## Data analysis

The samples in the benchmark set were spiked using *C. auris* AR 0385 (Biosample SAMN05379620 as per CDC's AR isolate bank; strain B11244) that has reads in SRA under accession SRR3883465 but no published assembly. Therefore, we used the assembly in our reference set of 100 genomes that is closest to the spike in strain for presenting the data analysis. We found the closest assembly in the following manner: The reads in SRR3883465 were assembled using SKESA [23], resulting in an assembly with a length of 12.21 Mb and N50 of 22 Kb. The assembly was then aligned to all reference genomes using BLAST, retaining only the best e-value alignments, and coverage on the reference genomes was determined using the retained alignments. The analysis revealed that 12.19 Mb of the assembly had alignments to reference genomes, with the maximum coverage for the reference assembly B12342.1 at 11.52 Mb aligned. The second-best coverage was for B11245, but it had only 1.4 Mb aligned. All alignments to B12342.1 had a percent identity of at least 99.6%, of which all except 8,465 bp aligned at a percent identity of at least 99.9%. Hence, MetaNISH scores for the benchmark dataset were presented using reference genome B12342.1. These scores were compared to KrakenUniq [24], a method for metagenomic classification that provides a quantitative measure of genome coverage. KrakenUniq was run with defaults, except no information was printed for unclassified sequences using parameter—only-classified-output.





**Fig 2. Pipeline for *C. auris* sequence-based monitoring using MetaNISH.** Steps 1–4 comprise collecting the required input data (samples sequence reads and reference database) for MetaNISH (step 5), whose output is a file with the scores for all references for each sample processed. Finally (step 6), this stack of files is processed and analyzed to obtain the samples with positive hits (score  $\geq 75$ ) of *C. auris*.

<https://doi.org/10.1371/journal.pone.0291406.g002>

Heatmaps were generated using alignments to reference genome B12342.1, contigs in the reference assembly were split into consecutive intervals of size 200 Kb and 2 Kb to represent padding of 100 Kb and 1 Kb on both ends of the alignments for reads, respectively. For each alignment, the starting position of the alignment on the contig was used to determine the bin where the alignment contributes to the count and to increase the count for that bin by one. The counts were plotted in MATLAB (version R2020a, Update 2) using the *imagesc* function to produce the heatmaps.

SRA reads were aligned to the set of reference genomes, and a score for each reference genome using padded coverage was obtained for SRA ids from January 2010 to November 2022, retrospectively. Using the output from MetaNISH, we scanned the scores for all *C. auris* reference assemblies using a MetaNISH score  $\geq 50$  up to the maximum possible score of 100 to obtain the number of SRA runs with at least a hit on any of the *C. auris* assemblies. With the

suggested score of  $\geq 75$  as the threshold for positive pathogen identification, samples with *C. auris* positive hits were described using the metadata collected.

## Results

### Benchmarking the reference dataset in the monitoring tool

The benchmark dataset is further described in Table 1. The presence of *C. auris* spike-in, cell concentration, scores computed by MetaNISH using different padding lengths, and the assembly coverage reported by KrakenUniq (Table 1 and Fig 3) are indicated for each metagenomic run. Using a score of 75 with 100 Kb padding, MetaNISH was able to detect all true positive as well as one false positive sample, while a score of 80 was able to separate all positive and negative samples. Significant variation was observed in scores with padding of less than 100 Kb, no padding, and KrakenUniq. For example, SRR11734778 compared to SRR11734781 had similar cell concentrations ( $7.1 \times 10^4$  CFU/mL and  $6.7 \times 10^4$  CFU/mL, respectively) and the same score (100) using 100 Kb padding; however, SRR11734778 compared to SRR11734781 had substantially different scores with 1 Kb padding (39.97 and 89.09, respectively), no padding (7.42 and 24.38, respectively) and KrakenUniq (3.73 and 13.52, respectively). Increasing padding to even just 10 Kb brings the padded coverage to over 98 for SRR11734778 and SRR11734781. As reflected in Table 1 and Fig 3, a padding length of 100 Kb was found to be effective in differentiating positive samples like SRR11734782 with a score of 92.74 from negative samples, while MetaNISH, with lesser than 100 Kb or no padding and KrakenUniq coverage values were not effective. Fig 3 depicted that the score is not affected by the number of

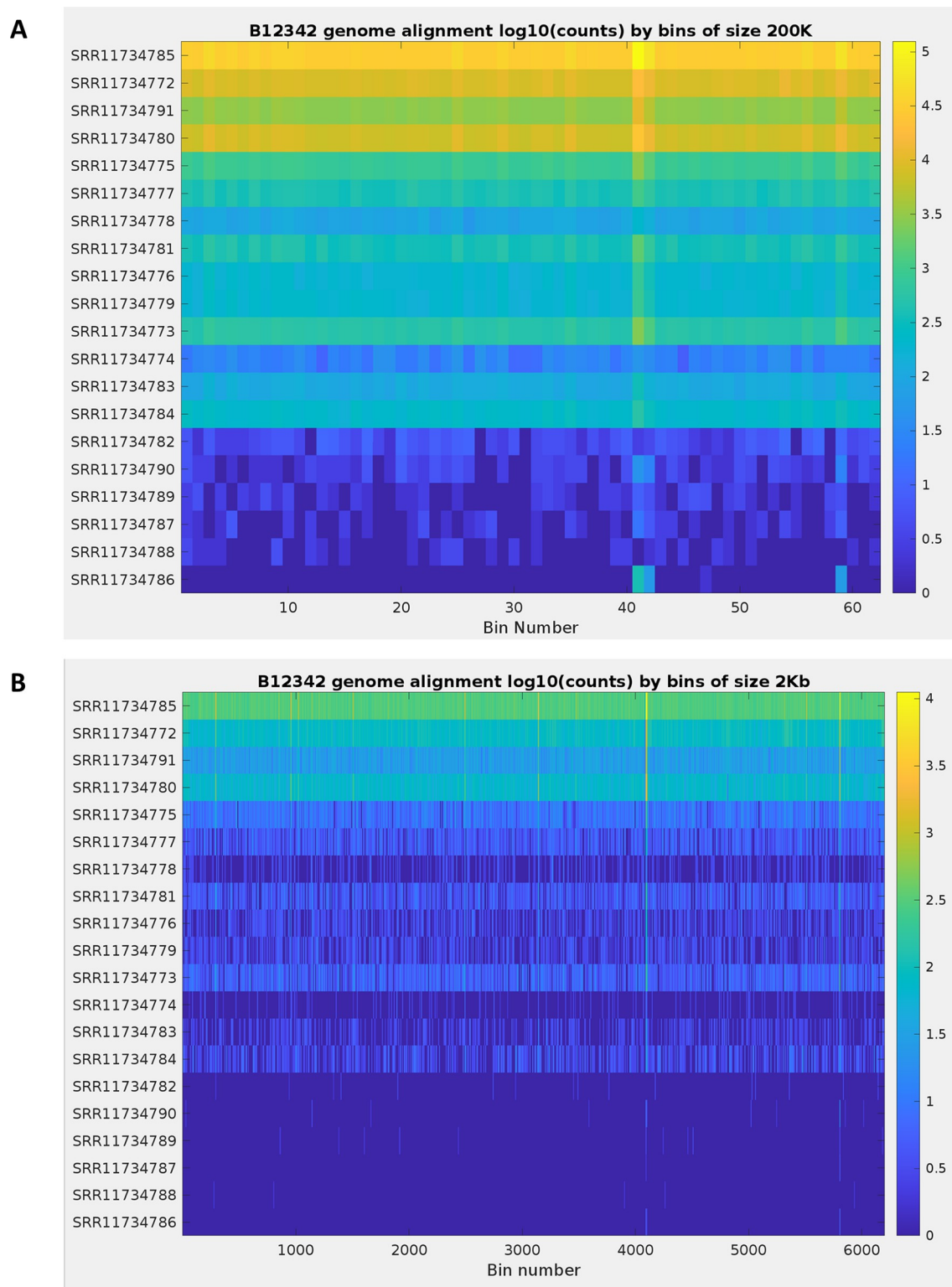
Table 1. Benchmark results using B12342 reference genome.

Benchmark design			MetaNISH scores with the specified padding						KrakenUniq
Run	Status	Concentration <sup>a</sup>	150 Kb	100 Kb	50 Kb	10 Kb	1 Kb	None	coverage*100
SRR11734785	pos	$5.8 \times 10^5$	100	100	100	100	99.98	99.9	82.9
SRR11734772	pos	$3.5 \times 10^5$	100	100	100	100	99.97	99.57	76.98
SRR11734791	pos	$2.0 \times 10^5$	100	100	100	100	99.96	85.92	56.38
SRR11734780	pos	$1.6 \times 10^5$	100	100	100	100	99.93	97.88	71.1
SRR11734775	pos	$1.2 \times 10^5$	100	100	100	100	99.32	50.33	30.6
SRR11734777	pos	$8.6 \times 10^4$	100	100	100	99.98	89.8	30.07	17.36
SRR11734778	pos	$7.1 \times 10^4$	100	100	100	98.8	39.97	7.42	3.73
SRR11734781	pos	$6.7 \times 10^4$	100	100	100	100	89.09	24.38	13.52
SRR11734776	pos	$4.3 \times 10^4$	100	100	100	99.98	66.17	13.89	9.003
SRR11734779	pos	$2.7 \times 10^4$	100	100	100	99.99	68.22	13.6	7.575
SRR11734773	pos	$1.1 \times 10^4$	100	100	100	99.99	93.81	29.67	16.4
SRR11734774	pos	$1.1 \times 10^4$	100	100	99.8	71.75	14.28	2.36	1.756
SRR11734783	pos	$1.0 \times 10^4$	100	100	100	99.74	48.4	8	4.61
SRR11734784	pos	$2.9 \times 10^3$	100	100	100	99.99	67.56	13.65	8.118
SRR11734782	pos	$1.9 \times 10^3$	96.72	92.74	74.01	23.71	3.04	0.43	0.8356
SRR11734790	neg	NA	85.08	79.43	53.74	14.74	1.84	0.25	0.9267
SRR11734789	neg	NA	65.16	57.18	34.14	8.43	0.99	0.13	0.4599
SRR11734787	neg	NA	50.28	41.88	22.96	5.29	0.68	0.11	0.1616
SRR11734788	neg	NA	46.98	39.93	22.75	5.4	0.63	0.08	0.8494
SRR11734786	mock	NA	3.56	2.88	1.47	0.57	0.14	0.04	0.4101

Pos: Positive for *C. auris*; neg: Negative for *C. auris*; mock: Pooled skin swab samples negative for *C. auris*.

<sup>a</sup> Units in CFU/mL.

<https://doi.org/10.1371/journal.pone.0291406.t001>



**Fig 3. Distribution of the number of reads aligned in the benchmark set to the B12342 reference genome.** The reference genome is binned by 200 Kb to reflect padding of 100 Kb on both ends of read alignments (A) and similarly by 2 Kb to reflect the padding of 1 Kb (B). The read sets are sorted by decreasing concentration levels of *C. auris*, with the topmost run SRR11734785 having the highest concentration.

<https://doi.org/10.1371/journal.pone.0291406.g003>



Table 2. Number of SRA records scanned.

Year	Records Scanned	Cumulative
2010	1	1
2011	41	42
2012	551	593
2013	2570	3163
2014	7023	10186
2015	17402	27588
2016	11529	39117
2017	21458	60575
2018	26897	87472
2019	51395	138867
2020	47766	186633
2021	46952	233585
2022	57756	291341

<https://doi.org/10.1371/journal.pone.0291406.t002>

reads aligning in a specific bin, such as high coverage for SRR11734785 (all yellow) and relatively low but well distributed throughout the genome coverage for SRR11734774 (primarily light blue).

### Detection of *Candida auris* in metagenomic datasets

Using MetaNISH (Fig 2), the number of samples per year that met the filtering criteria increased from one sample in 2010 to 57,756 in 2022 (Table 2). As of December 2022, 291,341 SRA samples were analyzed (Table 2) to produce an output of *C. auris* hits with varying genome coverage (Table 3). *C. auris* was identified in five sample datasets: PRJNA488992 (2 SRA runs), PRJNA657014 (4 SRA runs), PRJNA475330 (1 SRA run), PRJNA631031 (15 SRA

Table 3. Binned padded genome coverage (score) of SRA runs (samples) with *Candida auris* hits.

Year	Score Ranges				
	[50–75)	[75–85)	[85–95)	[95–100)	100*
2010	0	0	0	0	0
2011	0	0	0	0	0
2012	0	0	0	0	0
2013	0	0	0	0	0
2014	0	0	0	0	0
2015	0	0	0	0	0
2016	5	0	0	0	0
2017	0	0	0	0	0
2018	0	0	0	0	0
2019	1	0	2	4	6
2020	6	3	5	9	14
2021	0	0	0	0	0
2022	1	0	0	0	0
Total	13	3	7	13	20

Numbers in bold correspond to the samples where *C. auris* was identified, and its metadata is shown in Table 4.

\*This is not an interval; it equals the number of runs with a score of 100. For all other columns, the interval is closed at the beginning and open at the end.

<https://doi.org/10.1371/journal.pone.0291406.t003>

Table 4. Bioproject metadata for samples with WGS data at SRA with *C. auris* positive hits.

Run Record	Score	Release Year	Bioproject	SRA study	Title	Environment or isolation source
SRR8584355	100%	2019	PRJNA488992	SRP159446	Metagenomics of wastewater drains and river samples from Delhi, India	Wastewater drain
SRR8584356	100%					Urban river
SRR9016982	100%	2019	PRJNA657014	SRP277451	Sequencing data from point prevalence study associated with <i>C. auris</i> Raw sequence reads	Combined axilla and inguinal crease (groin) and anterior nares (Human skin metagenome)
SRR9016983	100%					
SRR9016984	100%					
SRR9016985	100%					
SRR10237756	>90%	2019	PRJNA475330	SRP161559	Metagenomic assembly of the iron-reducing, 1-methylnaphthalene-degrading enrichment culture (1MN)	Sulfur-oxidizing nitrate-reducing enrichment culture
SRR11734772	100%	2020	PRJNA631031	SRP260772	Study of microbial diversity of anterior nares swabs from patients colonized by the pathogen <i>Candida auris</i>	Human nasopharyngeal metagenome
SRR11734773	100%					
SRR11734774	100%					
SRR11734775	100%					
SRR11734776	100%					
SRR11734777	100%					
SRR11734778	100%					
SRR11734779	100%					
SRR11734780	100%					
SRR11734781	100%					
SRR11734783	100%					
SRR11734784	100%					
SRR11734785	100%					
SRR11734791	100%					
SRR11734782	>90%					
SRR10680803	>90%	2020	PRJNA557323	SRP237407	Human gut metagenomes from Hong Kong populations	Stool samples (Human gut metagenome)
SRR10680804	>90%					

<https://doi.org/10.1371/journal.pone.0291406.t004>

runs), and PRJNA557323 with two SRA runs (Table 4). Sequence reads from PRJNA488992 were collected from wastewater drains and river samples from Delhi, India. Reads from PRJNA657014 and PRJNA631031 datasets were from skin swabs of the residents of healthcare facilities in the United States where *C. auris* had been identified [25] and our benchmark dataset, respectively. Sequence reads from PRJNA557323 were collected from human stool samples from Hong Kong. Finally, PRJNA475330 samples were collected from Germany's sulfur-oxidizing nitrate-reducing enrichment culture of a groundwater sample (Table 4).

### Prospective monitoring

GridRepublic has implemented a molecular monitoring system using MetaNISH on a volunteer computing platform (i.e., a distributed computing platform comprised of resources volunteered by the general public). This system successfully screens all new metagenomic data submitted to SRA daily for *C. auris* (averaging 925 runs per day). These results are available on the web at [www.gridrepublic.org/biosurveillance](http://www.gridrepublic.org/biosurveillance).

### Discussion

In a collaborative effort between CDC, NCBI, and GridRepublic, we developed and benchmarked bioinformatics tools for the prospective monitoring of metagenomic datasets for the detection of *C. auris* and examined ~300,000 SRA runs released between 2010 and 2022 to

identify this pathogen. Using benchmarking samples generated by spiking human skin microbiome with known concentrations of *C. auris*, we found that the MetaNISH pipeline with the following parameters was successful in identifying samples with *C. auris*: (i) alignment of the first 100 bases to the target using SRPRISM, (ii) padding length of 100 Kb, and (iii) score threshold of at least 75. Increasing a cutoff score to 80 was able to separate all positive and negative samples; however, using a cutoff of 75 may increase the chances of identifying samples with a low prevalence of *C. auris*. The proposed parameters were especially beneficial for the detection of benchmarking samples with lower concentrations of *C. auris* reads, such as SRR11734774 and SRR11734782, which showed low base pair coverage of 2.36% and 0.43% by SRPRISM, and 1.8% and 0.8%, by KrakenUniq but generated scores of 100 and 92.4 with MetaNISH and 100 Kb padding (Table 1). The alignment scores above 90 indicate that the alignments were well-distributed throughout the genome, increasing confidence in the results.

Our study identified five metagenomic datasets containing *C. auris* sequences in the public SRA repository. The first was the benchmark dataset used to test our pipeline. The second was from skin swabs of patients colonized with *C. auris* (24). Detection of *C. auris* in these samples was not surprising, although important for providing an independent validation of the developed method. The third dataset was from a study of stool microbiome of healthy individual in Hong Kong [26], which was novel and unexpected. Although *C. auris* has previously been isolated from the gastrointestinal tract [27,28], it is generally accepted that this fungus is primarily a skin colonizer [12]. Several publications show that *Candida* spp. can survive to passage through the gut in healthy adults and possibly generate further spread via wastewater [29,30]. Our observation raises the question of whether patients colonized with *C. auris* on the skin are also colonized in the gut and whether some human communities may harbor the previously unknown reservoirs of *C. auris* [31,32]. More studies of healthy people are needed to understand the prevalence of *C. auris* in the community [33].

The presence of *C. auris* in the fourth set, laboratory culture of the iron-reducing bacteria most likely indicates contamination [34], although it suggests its ability to survive under such iron-reducing conditions. The detection of *C. auris* in aquatic biome samples from Delhi is consistent with the recent report showing isolation of *C. auris* from the coastal waters in India and Colombia [14,15], which provided support to the hypothesis of an environmental origin of this pathogen [9,10]. However, it is also equally likely that *C. auris* might have been spread into aquatic environment from contaminated wastewater after being excreted from the gastrointestinal tract or washed off the skin of a colonized people [35]. These findings point to its most likely mode of spread (any aquatic stream or aqueous medium) between the human populations and environment and vice versa [36].

A limitation of this study is that only shotgun metagenomic data were analyzed, and amplicon sequencing data were excluded. Relatively high costs and the need for more advanced bioinformatics have limited the use of shotgun metagenomics for microbiome analysis on a large scale [37–39]. In contrast, the amplicon sequencing approach is the most widely used method for analyzing microbial communities due to its cost-effectiveness, established data analysis pipelines, and availability of an extensive archive of reference data [12,25,36]. Thus, building and validating a search option for amplicon datasets into MetaNISH by generating benchmark amplicon datasets and *C. auris* reference databases will complement the existing metagenomic search function. The other limitation of the study is that for most SRA submissions, only limited metadata on the specimens is available. A follow-up with the submitters is often needed to identify additional details of the study and to determine whether a finding of *C. auris* sequences in the sample is indeed a reflection of the sequenced community and not an artifact of laboratory practices, in which *C. auris* might have been used as a loading control or occurred as a contaminant. It is also important to point out that the identification of reads of

*C. auris* in certain samples may not necessarily indicate that these samples represent the ecological niche for this fungus. As described above with an example of finding *C. auris* in coastal waters, the directionality of *C. auris* transition between the human skin and coastal waters is not clear. It is equally likely that the fungus might have emerged in coastal habitats and later transitioned into human population, or in contrast, that it has emerged elsewhere and was introduced into the coastal waters from colonized persons.

Because, in many cases, there is a significant time lag between the collection of a sample and the submission/publication of its sequence reads, the detections that can be made (even daily) do not imply an active outbreak response but are valuable *post hoc* information that allows tracking trends of spread and are encompassed in the data integration of a One Health surveillance system [40].

The findings presented in this study using MetaNISH on public metagenomic data support the results of previous work on *C. auris* in natural environments. This work also lays the foundation for the prospective monitoring system for *C. auris* because the modular design of MetaNISH makes it suitable for this daily job, which in addition to addressing scientific questions about the origin of *C. auris*, provides a necessary public health monitoring tool for investigating the spread of *C. auris* into the new areas. GridRepublic has implemented the pipeline developed and evaluated in our study on a distributed computing network, adapting it into a real-time monitoring system. Future efforts can adapt this tool to monitor other emerging pathogens and public health threats.

## Supporting information

**S1 Table. Reference genomes.**  
(DOCX)

## Acknowledgments

This research work was supported in part by the National Center for Biotechnology Information of the National Library of Medicine (NLM), National Institutes of Health. This work was also made possible through support from the CDC Office of Advanced Molecular Detection (OAMD). The contents of this publication are solely the responsibility of the authors and do not necessarily represent the official views of the Centers for Disease Control and Prevention.

## Author Contributions

**Conceptualization:** Richa Agarwala, Nancy A. Chow.

**Data curation:** Jorge E. Mario-Vasquez, Ujwal R. Bagal, Richa Agarwala.

**Formal analysis:** Jorge E. Mario-Vasquez, Ujwal R. Bagal, Richa Agarwala.

**Funding acquisition:** Rory Welsh, Nancy A. Chow.

**Investigation:** Jorge E. Mario-Vasquez, Ujwal R. Bagal, Richa Agarwala.

**Methodology:** Jorge E. Mario-Vasquez, Ujwal R. Bagal, Elijah Lowe, Aleksandr Morgulis, John Phan, Sergey Shirayev, Rytis Slatkevičius, Rory Welsh, Matthew Blumberg, Richa Agarwala.

**Project administration:** Nancy A. Chow.

**Resources:** D. Joseph Sexton, Matthew Blumberg, Richa Agarwala, Nancy A. Chow.

**Software:** Jorge E. Mario-Vasquez, Ujwal R. Bagal, Elijah Lowe, Aleksandr Morgulis, John Phan, Sergey Shirayev, Rytis Slatkevičius, Matthew Blumberg, Richa Agarwala.

**Supervision:** Richa Agarwala, Nancy A. Chow.

**Validation:** Jorge E. Mario-Vasquez, Ujwal R. Bagal, Richa Agarwala.

**Visualization:** Jorge E. Mario-Vasquez, Ujwal R. Bagal, Richa Agarwala.

**Writing – original draft:** Jorge E. Mario-Vasquez.

**Writing – review & editing:** Jorge E. Mario-Vasquez, Ujwal R. Bagal, D. Joseph Sexton, Rory Welsh, Anastasia P. Litvintseva, Matthew Blumberg, Richa Agarwala, Nancy A. Chow.

## References

1. Watkins RR, Gowen R, Lionakis MS, Ghannoum M. Update on the Pathogenesis, Virulence, and Treatment of *Candida auris*. *Pathog Immun*. 2022; 7(2):46–65. <https://doi.org/10.20411/pai.v7i2.535> PMID: 36329818
2. WHO. WHO fungal priority pathogens list to guide research, development and public health action: World Health Organization; 2022 [Available from: <https://www.who.int/publications/item/9789240060241>.
3. 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):41–4. <https://doi.org/10.1111/j.1348-0421.2008.00083.x> PMID: 19161556
4. Lee WG, Shin JH, Uh Y, Kang MG, Kim SH, Park KH, et al. First Three Reported Cases of Nosocomial Fungemia Caused by *Candida auris*. *Journal of Clinical Microbiology*. 2011; 49(9):3139–42. <https://doi.org/10.1128/JCM.00319-11> PMID: 21715586
5. 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. *Clin Infect Dis*. 2017; 64(2):134–40. <https://doi.org/10.1093/cid/ciw691> PMID: 27988485
6. Chow NA, De Groot T, Badali H, Abastabar M, Chiller TM, Meis JF. Potential Fifth Clade of *Candida auris*, Iran, 2018. *Emerging Infectious Diseases*. 2019; 25(9):1780–1. <https://doi.org/10.3201/eid2509.190686> PMID: 31310230
7. Spruijtenburg B, Badali H, Abastabar M, Mirhendi H, Khodavaisy S, Sharifisooraki J, et al. Confirmation of fifth *Candida auris* clade by whole genome sequencing. *Emerging Microbes & Infections*. 2022; 11(1):2405–11. <https://doi.org/10.1080/22221751.2022.2125349> PMID: 36154919
8. Chow NA, Gade L, Tsay SV, Forsberg K, Greenko JA, Southwick KL, et al. Multiple introductions and subsequent transmission of multidrug-resistant *Candida auris* in the USA: a molecular epidemiological survey. *The Lancet Infectious Diseases*. 2018; 18(12):1377–84. [https://doi.org/10.1016/S1473-3099\(18\)30597-8](https://doi.org/10.1016/S1473-3099(18)30597-8) PMID: 30293877
9. Casadevall A, Kontoyiannis DP, Robert V. Environmental *Candida auris* and the Global Warming Emergence Hypothesis. *mBio*. 2021; 12(2). <https://doi.org/10.1128/mBio.00360-21> PMID: 33727350
10. Jackson BR, Chow N, Forsberg K, Litvintseva AP, Lockhart SR, Welsh R, et al. On the Origins of a Species: What Might Explain the Rise of *Candida auris*? *Journal of Fungi*. 2019; 5(3):58. <https://doi.org/10.3390/jof5030058> PMID: 31284576
11. Huang X, Hurabielle C, Drummond RA, Bouladoux N, Desai JV, Sim CK, et al. Murine model of colonization with fungal pathogen *Candida auris* to explore skin tropism, host risk factors and therapeutic strategies. *Cell Host & Microbe*. 2021; 29(2):210–21.e6. <https://doi.org/10.1016/j.chom.2020.12.002> PMID: 33385336
12. Proctor DM, Dangana T, Sexton DJ, Fukuda C, Yelin RD, Stanley M, et al. Integrated genomic, epidemiologic investigation of *Candida auris* skin colonization in a skilled nursing facility. *Nature Medicine*. 2021; 27(8):1401–9. <https://doi.org/10.1038/s41591-021-01383-w> PMID: 34155414
13. Chow NA, Muñoz JF, Gade L, Berkow EL, Li X, Welsh RM, et al. Tracing the Evolutionary History and Global Expansion of *Candida auris* Using Population Genomic Analyses. *mBio*. 2020; 11(2). <https://doi.org/10.1128/mBio.03364-19> PMID: 32345637
14. Arora P, Singh P, Wang Y, Yadav A, Pawar K, Singh A, et al. Environmental Isolation of *Candida auris* from the Coastal Wetlands of Andaman Islands, India. *mBio*. 2021; 12(2). <https://doi.org/10.1128/mBio.03181-20> PMID: 33727354



15. Escandón P. Novel Environmental Niches for *Candida auris*: Isolation from a Coastal Habitat in Colombia. *Journal of Fungi*. 2022; 8(7):748. <https://doi.org/10.3390/jof8070748> PMID: 35887503
16. Abdill RJ, Adamowicz EM, Blekhman R. Public human microbiome data are dominated by highly developed countries. *PLOS Biology*. 2022; 20(2):e3001536. <https://doi.org/10.1371/journal.pbio.3001536> PMID: 35167588
17. Katz K, Shutov O, Lapoint R, Kimelman M, Brister JR, O'Sullivan C. The Sequence Read Archive: a decade more of explosive growth. *Nucleic Acids Res*. 2022; 50(D1):D387–d90. <https://doi.org/10.1093/nar/gkab1053> PMID: 34850094
18. Morgulis A, Agarwala R. SRPRISM (Single Read Paired Read Indel Substitution Minimizer): an efficient aligner for assemblies with explicit guarantees. *Gigascience*. 2020; 9(4). <https://doi.org/10.1093/gigascience/giaa023> PMID: 32315028
19. Muñoz JF, Welsh RM, Shea T, Batra D, Gade L, Howard D, et al. Clade-specific chromosomal rearrangements and loss of subtelomeric adhesins in *Candida auris*. *Genetics*. 2021; 218(1). <https://doi.org/10.1093/genetics/iyab029> PMID: 33769478
20. Welsh RM, Misas E, Forsberg K, Lyman M, Chow NA. *Candida auris* Whole-Genome Sequence Benchmark Dataset for Phylogenomic Pipelines. *J Fungi (Basel)*. 2021; 7(3). <https://doi.org/10.3390/jof7030214> PMID: 33809682
21. Welsh RM, Bentz ML, Shams A, Houston H, Lyons A, Rose LJ, et al. Survival, Persistence, and Isolation of the Emerging Multidrug-Resistant Pathogenic Yeast *Candida auris* on a Plastic Health Care Surface. *Journal of Clinical Microbiology*. 2017; 55(10):2996–3005. <https://doi.org/10.1128/JCM.00921-17> PMID: 28747370
22. Sexton DJ, Kordalewska M, Bentz ML, Welsh RM, Perlin DS, Litvintseva AP. Direct Detection of Emergent Fungal Pathogen *Candida auris* in Clinical Skin Swabs by SYBR Green-Based Quantitative PCR Assay. *Journal of Clinical Microbiology*. 2018; 56(12). <https://doi.org/10.1128/JCM.01337-18> PMID: 30232130
23. Souvorov A, Agarwala R, Lipman DJ. SKESA: strategic k-mer extension for scrupulous assemblies. *Genome Biology*. 2018; 19(1). <https://doi.org/10.1186/s13059-018-1540-z> PMID: 30286803
24. Breitwieser FP, Baker DN, Salzberg SL. KrakenUniq: confident and fast metagenomics classification using unique k-mer counts. *Genome Biology*. 2018; 19(1). <https://doi.org/10.1186/s13059-018-1568-0> PMID: 30445993
25. Huang X, Welsh RM, Deming C, Proctor DM, Thomas PJ, Gussin GM, et al. Skin Metagenomic Sequence Analysis of Early *Candida auris* Outbreaks in U.S. Nursing Homes. *mSphere*. 2021; 6(4): e0028721. <https://doi.org/10.1128/mSphere.00287-21> PMID: 34346704
26. Yeoh YK, Chen Z, Wong MCS, Hui M, Yu J, Ng SC, et al. Southern Chinese populations harbour non-nucleatum *Fusobacteria* possessing homologues of the colorectal cancer-associated FadA virulence factor. *Gut*. 2020; 69(11):1998–2007. <https://doi.org/10.1136/gutjnl-2019-319635> PMID: 32051205
27. Alam MJ, Begum K, Endres BT, McPherson J, Costa G, Miranda JM, et al. 1720. Isolation and Characterization of *Candida auris* From an Active Surveillance System in Texas. *Open Forum Infectious Diseases*. 2019; 6(Supplement\_2):S630–S.
28. Zuo T, Zhan H, Zhang F, Liu Q, Tso EYK, Lui GCY, et al. Alterations in Fecal Fungal Microbiome of Patients With COVID-19 During Time of Hospitalization until Discharge. *Gastroenterology*. 2020; 159(4):1302–10.e5. <https://doi.org/10.1053/j.gastro.2020.06.048> PMID: 32598884
29. Leonardi I, Paramsothy S, Doron I, Semon A, Kaakoush NO, Clemente JC, et al. Fungal Trans-kingdom Dynamics Linked to Responsiveness to Fecal Microbiota Transplantation (FMT) Therapy in Ulcerative Colitis. *Cell Host & Microbe*. 2020; 27(5):823–9.e3. <https://doi.org/10.1016/j.chom.2020.03.006> PMID: 32298656
30. Rossi A, Chavez J, Iverson T, Hergert J, Oakeson K, Lacross N, et al. *Candida auris* Discovery through Community Wastewater Surveillance during Healthcare Outbreak, Nevada, USA, 2022. *Emerging Infectious Diseases*. 2023; 29(2):422–5. <https://doi.org/10.3201/eid2902.221523> PMID: 36692459
31. Olsen M, Nassar R, Senok A, Moloney S, Lohning A, Jones P, et al. Mobile phones are hazardous microbial platforms warranting robust public health and biosecurity protocols. *Scientific Reports*. 2022; 12(1). <https://doi.org/10.1038/s41598-022-14118-9> PMID: 35705596
32. Tharp B, Zheng R, Bryak G, Litvintseva AP, Hayden MK, Chowdhary A, et al. Role of Microbiota in the Skin Colonization of *Candida auris*. *mSphere*. 2023; 8(1):e0062322. <https://doi.org/10.1128/msphere.00623-22> PMID: 36695588
33. Ahmad S, Asadzadeh M. Strategies to Prevent Transmission of *Candida auris* in Healthcare Settings. *Curr Fungal Infect Rep*. 2023:1–13. <https://doi.org/10.1007/s12281-023-00451-7> PMID: 36718372
34. Müller H, Marozava S, Probst AJ, Meckenstock RU. Groundwater cable bacteria conserve energy by sulfur disproportionation. *The ISME Journal*. 2020; 14(2):623–34. <https://doi.org/10.1038/s41396-019-0554-1> PMID: 31728021

35. Akinbobola AB, Kean R, Hanifi SMA, Quilliam RS. Environmental reservoirs of the drug-resistant pathogenic yeast *Candida auris*. *PLOS Pathogens*. 2023; 19(4):e1011268. <https://doi.org/10.1371/journal.ppat.1011268> PMID: 37053164
36. Irinyi L, Roper M, Malik R, Meyer W. Finding a Needle in a Haystack—In Silico Search for Environmental Traces of *Candida auris*. *Jpn J Infect Dis*. 2022; 75(5):490–5. <https://doi.org/10.7883/yoken.JJID.2022.068> PMID: 35491231
37. Hilt EE, Ferrieri P. Next Generation and Other Sequencing Technologies in Diagnostic Microbiology and Infectious Diseases. *Genes (Basel)*. 2022; 13(9). <https://doi.org/10.3390/genes13091566> PMID: 36140733
38. Ranjan R, Rani A, Metwally A, Mcgee HS, Perkins DL. Analysis of the microbiome: Advantages of whole genome shotgun versus 16S amplicon sequencing. *Biochemical and Biophysical Research Communications*. 2016; 469(4):967–77. <https://doi.org/10.1016/j.bbrc.2015.12.083> PMID: 26718401
39. Rausch P, Rühlemann M, Hermes BM, Doms S, Dagan T, Dierking K, et al. Comparative analysis of amplicon and metagenomic sequencing methods reveals key features in the evolution of animal metaorganisms. *Microbiome*. 2019; 7(1). <https://doi.org/10.1186/s40168-019-0743-1> PMID: 31521200
40. Ko KKK, Chng KR, Nagarajan N. Metagenomics-enabled microbial surveillance. *Nature Microbiology*. 2022; 7(4):486–96. <https://doi.org/10.1038/s41564-022-01089-w> PMID: 35365786