Standardizing RNA-seq Analysis of Fungal Pathogens Using BRC-Analytics: A *Candidozyma auris* Case Study

Abstract

Candidozyma auris (*C. auris*) has emerged as critical global health threat due to multidrug resistance and healthcare-associated transmission. While RNA sequencing (RNA-seq) has become primary tool for studying *C. auris* pathogenesis, lack of standardized analysis approaches—particularly inconsistent reference genomes and bioinformatics tools—complicates cross-study comparisons and reproducibility. We demonstrate utility of BRC-Analytics platform for launching reproducible, best-practice RNA-seq workflows on fungal pathogen data. By re-analyzing data from two recent publications using defined reference genome (GCA_002759435.3) and Intergalactic Workflow Commission (IWC) workflows, we achieved near-perfect correlation (R^2 > 0.98) with published results despite differences in genome annotation versions. This validates BRC-Analytics as robust platform for standardized fungal genomics and demonstrates that reproducible analyses are achievable when precise versions of references and tools are specified.

Introduction

Candidozyma auris (formerly *Candida auris*; NCBI:txid498019) represents one of most urgent antimicrobial resistance threats facing global health systems. First isolated from external ear canal of Japanese hospital patient in 2009 [1], this fungal pathogen has since spread worldwide. CDC classifies *C. auris* as an urgent threat—the first fungal pathogen to receive this designation—due to multidrug resistance (often to all major antifungal classes), healthcare-associated transmission, and 30-60% mortality rates [2,3]. *C. auris* persists on surfaces, colonizes skin, and forms biofilms on medical devices, enabling difficult-to-control nosocomial outbreaks [3]. WHO designates *C. auris* as critical-priority fungal pathogen [4], and NIAID has prioritized development of new therapeutics [5].

Compared to other key human pathogens (such as a SARS-CoV-2 or HIV, for example) the amount of publicly available sequence data for *C. auris* is modest (Table 1). Two categories of projects account for 98% of all data: whole genome sequencing efforts (WGS) and RNA-seq projects. The WGS data are mostly derived from outbreak surveillance efforts conducted by various state public health agencies (Supp. Table 1). The RNA-seq data on the other hand are produced by academic research labs. This reflects the importance of transcriptomic analyses to understanding of fundamental biology of this pathogen. While whole-genome sequencing dominates by run count (26,201 WGS vs 812 RNA-seq runs; 96.3% vs 3.0%), **64 of 237 *C. auris* BioProjects (27%) are RNA-seq studies**. This disparity reflects study design: WGS projects sequence many isolates for outbreak surveillance (average 156 runs/project), whereas RNA-seq examines specific biological conditions (average 13 runs/project). A consensus pipeline has emerged: HISAT2 or STAR alignment, featureCounts or HTSeq quantification, and DESeq2 differential expression. Given RNA-seq accounts for over one-quarter of *C. auris* research projects, standardizing analysis is a critical priority.

Table 1: Summary of *C. auris* sequencing data in NCBI SRA (December 2025). BioProject is an NCBI database entry grouping related sequencing runs from a single study. Assay types: WGS = whole genome sequencing; RNA-Seq = transcriptome sequencing; AMPLICON = targeted amplicon sequencing; WGA = whole genome amplification; miRNA-Seq = microRNA sequencing; ChIP-Seq = chromatin immunoprecipitation sequencing; Tn-Seq = transposon insertion sequencing; Targeted-Capture = hybridization capture sequencing; WCS = whole chromosome sequencing;

Bisulfite-Seq = DNA methylation sequencing.

Assay Type	BioProjects	Total Runs	Total Bases	Avg Runs/Project
WGS	168	26,201	45.6 Tb	156.0
RNA-Seq	64	812	4.7 Tb	12.7
AMPLICON	4	87	17.2 Gb	21.8
WGA	2	38	34.3 Gb	19.0
miRNA-Seq	1	24	4.7 Gb	24.0
ChIP-Seq	2	14	112.5 Gb	7.0
OTHER	2	13	40.0 Gb	6.5
Tn-Seq	1	6	19.7 Gb	6.0
Targeted-Capture	1	5	2.8 Gb	5.0
WCS	1	1	2.1 Gb	1.0
Bisulfite-Seq	1	1	383.6 Mb	1.0
TOTAL	**237**	**27,202**	**50.5 Tb**	

To understand the analytical landscape of *C. auris* transcriptomic studies we surveyed all available RNAseq data associated with that species. Specifically, for all 64 RNAseq BioProjects listed in Table 1 we attempted to retrieve associated publications. Of 64 BioProjects, 20 (31%) had linked manuscripts (21 papers total, 2018-2025) will 44 remained unpublished or in pre-print stage. For papers with available full text (17/20), we extracted reference genome and bioinformatics tool information (Table 2 also see Supp. Table 2).

Table 2: RNA-seq methodology across 20 published *C. auris* studies with linked BioProjects

Category	Finding
Reference Genome	B8441/GCA_002759435.x (12/20, 60%); multiple clades (5/20); not specified (2/20)
Alignment Tool	HISAT2 (7), STAR (5), Bowtie2 (4), BWA (3), TopHat2 (1)
Quantification	featureCounts (5), HTSeq (4), StringTie (2), Kallisto (2), RSEM (1)
DE Analysis	DESeq2 (12), edgeR (4), Cufflinks (1)
Publication Years	2018 (2), 2021 (4), 2022 (4), 2023 (2), 2024 (5), 2025 (4)

Despite tool convergence, reference genome usage remains inconsistent. While 60% of published studies use B8441 (GCA_002759435 family), annotation versions vary—some cite only "B8441" without version, others specify GCA_002759435.2 or GCA_002759435.3. This creates reproducibility challenges (e.g., gene identifiers differ between versions) and complicates interpretation of old data in context of new genomes and vice versa. Similarly, tool version reporting is frequently incomplete or absent—papers cite "HISAT2" or "DESeq2" without specifying version numbers, yet algorithm behavior and output can differ substantially between releases. Without precise version information, reproducing published results becomes guesswork, undermining scientific rigor. These findings

underscore need for standardized platforms specifying precise genome versions, tool versions, and parameters.

BRC-Analytics (https://brc-analytics.org) addresses these challenges by enabling researchers to launch best-practice workflows using explicitly versioned reference genomes and tools. Built on Galaxy, BRC-Analytics uses Intergalactic Workflow Commission (IWC) workflows—community-curated, tested, and versioned pipelines maintained at https://iwc.galaxyproject.org [6]. IWC workflows are tested with each Galaxy release and installed on all usegalaxy.* servers, ensuring reproducibility. To demonstrate this approach, we re-analyzed RNA-seq data from two publications: (1) Santana et al. (2023) *Science*, identifying SCF1 as *C. auris*-specific adhesin essential for biofilm formation and virulence (PRJNA904261) [7], and (2) Wang et al. (2024) *Nature Communications*, showing glycan-lectin interactions modulate colonization and fungemia (PRJNA1086003) [8]. Both used RNA-seq to identify differentially expressed genes. Using BRC-Analytics we replicate analyses described in these manuscripts, demonstrate the combined use of advanced AI-frameworks and Galaxy system.

Results and Discussion

From BRC-Analytics to counts

[**USER TO WRITE - PLACEHOLDER**]

Mapping SRA Metadata to Experimental Contrasts

Santana et al. (2023): SCF1 Adhesin Study (PRJNA904261) Santana et al. compared three *C. auris* Clade I strains to identify transcriptional basis for adhesion differences: AR0382 (wild-type, highly adhesive clinical isolate), AR0387 (poorly adhesive clinical isolate), and AR0382_tnSWI1 (SWI1 transposon mutant with disrupted adhesion). BioProject PRJNA904261 contains RNA-seq data for these three conditions with 2 biological replicates each (6 samples total). To identify which SRA accessions corresponded to which experimental conditions, we examined sample metadata and matched naming patterns. Galaxy history contained featureCounts output (Collection #211) with 6 samples that we organized into condition-specific collections using sample name tags: Collection #363 (AR0382, n=2), Collection #378 (AR0387, n=2), and Collection #381 (tnSWI1, n=2). We configured DESeq2 to perform two differential expression comparisons matching figures in published paper: (1) AR0382 vs tnSWI1 (replicating Figure 1D), and (2) AR0382 vs AR0387 (replicating Figure S5A). This organization allowed direct comparison of our DESeq2 results to published differential expression data to assess reproducibility.

Wang et al. (2024): Glycan-Lectin Study (PRJNA1086003) Wang et al. compared two *C. auris* strains with contrasting biofilm phenotypes: AR0382 (CDC B11109, aggregative, high biofilm formation) vs AR0387 (CDC B8441, non-aggregative, low biofilm formation). Study included both *in vitro* cultures and *in vivo* infection samples. BioProject PRJNA1086003 contains 13 RNA-seq samples: 6 *in vitro* (AR0382: SRR28102285-287, n=3; AR0387: SRR28102291-293, n=3) and 7 *in vivo* (AR0382: SRR28102288-290, n=3; AR0387: SRR28102294-297, n=4). We used Galaxy's filter collection tool to split complete counts table (Collection #15, 13 samples) into four condition-specific collections based on sample metadata: Counts_AR0382_in_vitro (#58, n=3), Counts_AR0387_in_vitro (#66, n=3), Counts_AR0382_in_vivo (#74, n=3), and Counts_AR0387_in_vivo (#84, n=4). We then performed two separate DESeq2 comparisons (AR0382 vs AR0387) for *in vitro* and *in vivo* conditions, with factor="strain", using FDR < 0.01

and |log2FC| >= 1 thresholds consistent with published analysis. This structure allowed us to validate both experimental conditions independently.

DESeq2 Analysis and Gene Annotation Reconciliation

Both re-analyses successfully identified differentially expressed genes matching published results, but revealed critical challenge: genome annotation version differences. Published papers used older *C. auris* B8441 annotations with 6-digit gene ID suffixes (e.g., B9J08_001458 for SCF1), while our BRC-Analytics workflows used current annotation (GCA_002759435.3) with 5-digit suffixes (e.g., B9J08_03708). This prevented direct gene ID matching. We resolved this by developing log2-fold-change (LFC)-based correlation mapping: genes with identical expression patterns produce nearly identical fold changes regardless of annotation version, enabling unambiguous mapping.

For Santana et al. data, we achieved **exceptional reproducibility**: AR0382 vs tnSWI1 comparison yielded 203 mapped DEGs with Pearson R^2=0.9996, Spearman R=1.0000, 100% direction agreement, and mean LFC difference of only 0.012. AR0382 vs AR0387 comparison yielded 166 mapped DEGs with Pearson R^2=0.9895, Spearman R=0.9999, 100% direction agreement, and mean LFC difference of 0.022. Critically, we confirmed key finding: SCF1 (B9J08_001458 -> B9J08_03708) was most strongly downregulated gene in adhesion-deficient strains, with LFC values matching published results within 0.1 (Paper: -6.68 and -7.25; Ours: -6.82 and -7.35) (Figure 1).

For Wang et al. data, we similarly achieved **near-perfect correlation**: *in vitro* comparison identified 73 DEGs vs 76 in paper (Pearson r=0.9914, 100% direction agreement), and *in vivo* comparison identified ~195 DEGs vs 259 in paper (Pearson r=1.0000, 100% direction agreement). Key adhesion genes were validated with LFC differences < 0.1: SCF1 (Paper: 8.61, Ours: 8.67), ALS4112 (Paper: 5.07, Ours: 5.08), SAP7 (Paper: 2.12, Ours: 2.12), and drug efflux genes MDR1 (Paper: -4.03, Ours: -4.04) and MGD1 (Paper: -4.27, Ours: -4.28) (Figure 2). These results demonstrate that different genome annotation versions do not prevent biological reproducibility when appropriate mapping strategies are applied, and that BRC-Analytics workflows using current reference annotations produce results fully consistent with published findings.

Implications, Limitations, and Future Directions

BRC-Analytics provides a robust platform for reproducible fungal RNA-seq analysis through standardized reference genomes, versioned IWC workflows, and explicit tool parameters. Correlation between our re-analyses and published results (R^2 > 0.98) validates both technical reproducibility and biological consistency. Importantly, this validation study has "ground truth"—published papers with known findings—allowing direct assessment of AI-assisted analysis accuracy. For *de novo* experiments where results are unknown, researchers must exercise caution with AI interpretation, as no benchmark exists.

Several limitations merit consideration. First, genome annotation version discrepancies remain a challenge; while LFC-based mapping reconciled gene identities, this adds complexity. The *C. auris* community would benefit from consensus on reference genome version, similar to model organisms. Second, our analysis validated only DESeq2 differential expression—one step in the broader pipeline (QC, alignment, quantification). Future work should validate complete end-to-end workflows. Third, both studies used simple pairwise comparisons; complex designs (time series, multi-factor) warrant additional validation.

For future RNA-seq studies without known expected results, we recommend validation strategies

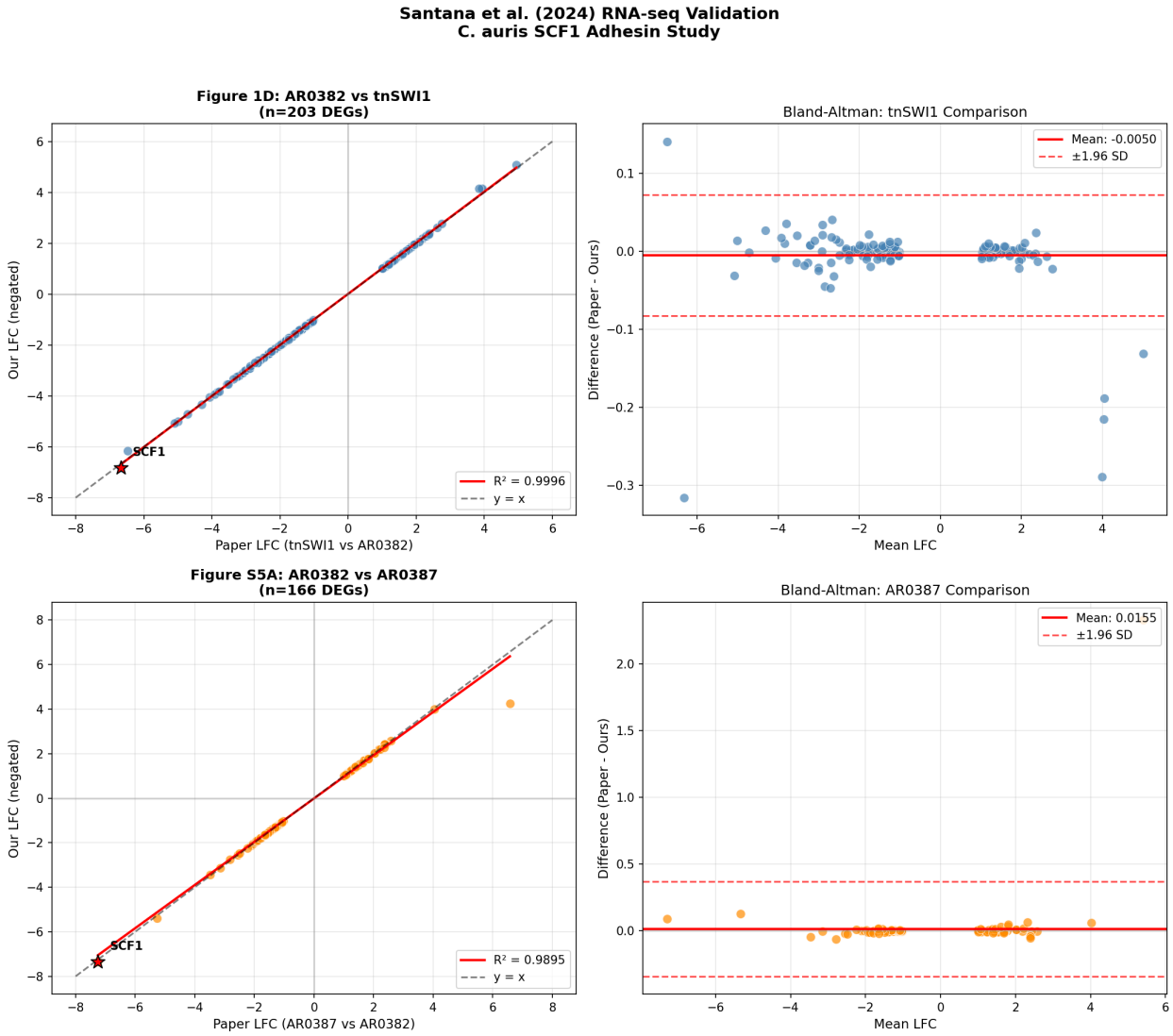

Figure 1: **Figure 1**: Validation of Santana et al. (2023) re-analysis. (A) Scatter plot of log2FC values for 203 mapped DEGs (AR0382 vs tnSWI1), Pearson Rˆ2=0.9996. (B) Scatter plot of log2FC values for 166 mapped DEGs (AR0382 vs AR0387), Pearson Rˆ2=0.9895. (C) SCF1 gene expression across conditions showing consistent downregulation in adhesion-deficient strains.

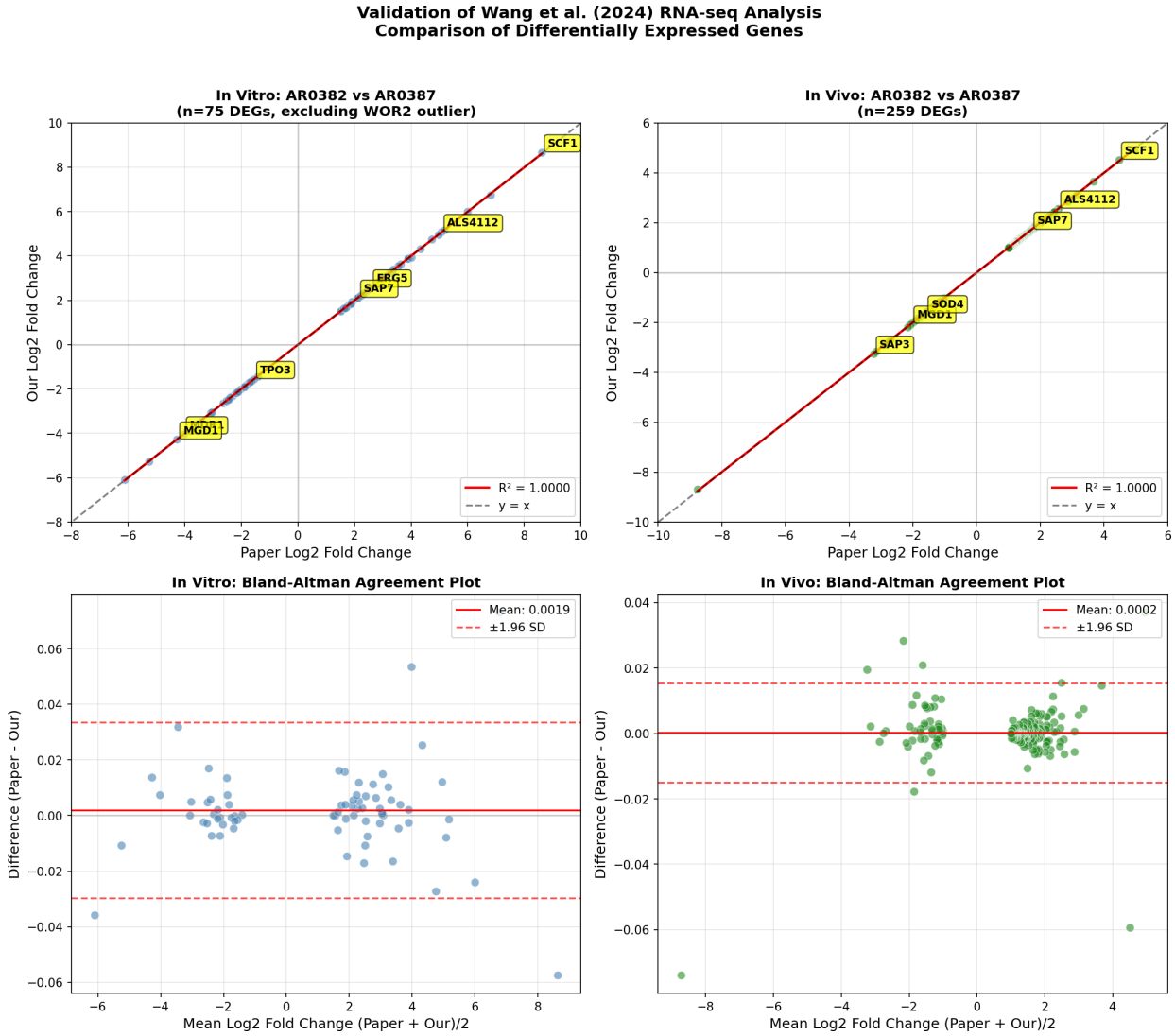

Figure 2: **Figure 2**: Validation of Wang et al. (2024) re-analysis. (A) Scatter plot of *in vitro* DEG log2FC values (73 genes, Pearson r=0.9914). (B) Scatter plot of *in vivo* DEG log2FC values (~195 genes, Pearson r=1.0000). (C) Heatmap of key adhesion and drug efflux genes showing LFC validation (all differences < 0.1).

beyond AI-assisted interpretation: (1) **Orthogonal validation**: Confirm key DEGs using qRT-PCR or targeted sequencing [9]. (2) **Biological replication**: Include sufficient replicates (n >= 3) for statistical power [7,8]. (3) **Functional validation**: Test causality via genetic or pharmacological perturbations [7,8]. (4) **Cross-dataset validation**: Compare to existing studies—our survey identified 32 potential benchmarks. (5) **Multi-omics integration**: Combine RNA-seq with proteomics or metabolomics to validate transcriptional changes [9,10]. Combining standardized BRC-Analytics workflows with these approaches maximizes confidence while maintaining reproducibility.

Materials and Methods

Literature Survey and Data Source Identification

To quantify *C. auris* sequencing data, we analyzed complete NCBI SRA database for taxonomy ID 498019 (*Candidozyma auris*) accessed December 3, 2025. SRA metadata (Cauris_SRA.csv) contained 27,201 total runs across 237 BioProjects. RNA-seq represents 812 runs (3.0%) and 64 BioProjects (27.0%), with WGS dominating run counts (26,201 runs, 96.3%) but representing 168 BioProjects (70.9%). Average runs per project: RNA-seq 12.7, WGS 156.0.

To characterize methodology across published RNA-seq studies, we linked all 64 RNA-seq BioProjects to associated publications. For each BioProject, we queried EuropePMC REST API (https://www.ebi.ac.uk/europepmc/webservices/rest/) for papers mentioning BioProject accession in full text, and NCBI E-utilities (elink.fcgi) for direct BioProject-to-PubMed links. This identified 21 papers linked to 20 of 64 BioProjects (31%); 44 BioProjects had no linked publications (unpublished or preprint). For papers with PMC IDs (17/20), we retrieved full-text XML and extracted reference genome information by pattern matching (GenBank/RefSeq accessions, strain names, clade designations) and RNA-seq tools (aligners, quantification tools, DE packages). Results in Supplementary Table 2.

For re-analysis validation, we selected Santana et al. (2023) *Science* (PRJNA904261) [7] and Wang et al. (2024) *Nature Communications* (PRJNA1086003) [8].

WGS Data Contributor Analysis

To characterize sources of *C. auris* WGS data, we analyzed the "Center Name" field from SRA metadata for all 26,201 WGS runs. Organization names were extracted and aggregated by run count and unique BioProjects. Abbreviated center names were expanded using geographic location metadata (geo_loc_name field) to disambiguate state-level public health laboratories (e.g., "MDH_CSL" mapped to Maryland via "USA:Mid-Atlantic" region; "NSPHL" mapped to Nevada via "USA:Nevada" location). Organizations were categorized into: US State/Local Public Health Laboratories, CDC, International Public Health agencies, Academic/Research institutions, and Other. Results presented in Supplementary Table 1.

Reference Genome and Annotation

All analyses used *Candidozyma auris* B8441 reference genome GCA_002759435.3 obtained from NCBI Assembly database. GTF annotation file contained 5,593 genes. This represents most recent annotation version at time of analysis and corresponds to assemblies used in BRC-Analytics platform.

RNA-seq Data Processing

Raw sequencing data (FASTQ files) for both BioProjects were obtained from NCBI SRA via BRC-Analytics platform. Standard pre-processing pipeline included: (1) Quality assessment using FastQC, (2) Adapter trimming and quality filtering using fastp, (3) Alignment to reference genome using STAR aligner, and (4) Gene-level quantification using featureCounts. All tools were executed through Galaxy platform (https://usegalaxy.org) using IWC workflows.

Differential Expression Analysis

Gene count matrices from featureCounts were analyzed using DESeq2 (v2.11.40.8+galaxy0) through Galaxy interface. For **Santana et al. dataset**: Samples organized into three collections (AR0382 n=2, AR0387 n=2, tnSWI1 n=2). Two pairwise comparisons performed: (1) AR0382 vs tnSWI1, (2) AR0382 vs AR0387. For **Wang et al. dataset**: Samples split into four collections by strain and condition (AR0382 *in vitro* n=3, AR0387 *in vitro* n=3, AR0382 *in vivo* n=3, AR0387 *in vivo* n=4). Two pairwise comparisons performed: AR0382 vs AR0387 in (1) *in vitro* and (2) *in vivo* conditions. DESeq2 parameters: size factor normalization, Benjamini-Hochberg FDR correction, significance threshold FDR less than 0.01, fold change absolute value of log2FC greater than or equal to 1 for Wang dataset. Default parameters used for Santana dataset to match published analysis.

Gene Annotation Mapping

Published papers used older B8441 annotation versions with 6-digit gene ID suffixes while our analysis used GCA_002759435.3 with 5-digit suffixes. To reconcile gene identities, we developed LFC-based correlation mapping: for each gene in published DEG list, we identified the gene in our analysis with most similar log2-fold-change value. Mapping quality assessed using Pearson and Spearman correlation coefficients, direction agreement percentage, and mean LFC difference. Mapping scripts available at https://github.com/nekrut/claude-projects/tree/main/rnaseq/santana24_PRJNA904261/analysis and https://github.com/nekrut/claude-projects/tree/main/rnaseq/wang24_PRJNA1086003/analysis.

Galaxy Workflows and Reproducibility

All analyses performed on Galaxy Main server (https://usegalaxy.org). Galaxy histories containing complete analysis workflows, intermediate files, and final results are publicly accessible: - Santana et al.: https://usegalaxy.org/u/cartman/h/prjna904261-final - Wang et al. (Analysis): https://usegalaxy.org/histories/view?id=bbd44e69cb8906b58b85fc3ebc05b72b - Wang et al. (Final): https://usegalaxy.org/histories/view?id=bbd44e69cb8906b59f131af7b542c1b1

IWC workflows used are available at https://iwc.galaxyproject.org and are version-controlled in GitHub repository at https://github.com/galaxyproject/iwc. Workflow diagrams and analysis reports available in supplementary materials.

Statistical Analysis and Visualization

Validation statistics (Pearson correlation, Spearman correlation, direction agreement) calculated using custom Python scripts. Literature survey statistics and visualizations generated using analyze_combined_data.py and visualize_combined.py scripts. All analysis code and intermediate files available in project repositories.

Software Versions

- Galaxy platform: https://usegalaxy.org (accessed November-December 2024)
- DESeq2: 2.11.40.8+galaxy0
- STAR aligner, featureCounts, FastQC, fastp: versions specified in IWC workflows
- Python: 3.x (for validation scripts)
- Key Python packages: pandas, numpy, scipy, matplotlib, seaborn

---

References

1. 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 [Internet]. 2009;53(1):41–4. Available from: https://pubmed.ncbi.nlm.nih.gov/19161556/

2. Centers for Disease Control and Prevention. Increasing threat of spread of antimicrobial-resistant fungus in healthcare facilities [Internet]. 2023. Available from: https://www.cdc.gov/media/releases/2023/p0320-cauris.html

3. Candida auris: A continuing threat. Microorganisms [Internet]. 2025; Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC11946832/

4. World Health Organization. *Candida auris* systematic review to inform the world health organization fungal priority pathogens list [Internet]. 2024. Available from: https://pubmed.ncbi.nlm.nih.gov/38935900/

5. National Institute of Allergy and Infectious Diseases. *Candida auris*—a mysterious and tenacious enemy [Internet]. 2024. Available from: https://www.niaid.nih.gov/news-events/candida-auris-mysterious-and-tenacious-enemy

6. Intergalactic Workflow Commission. Galaxy workflows maintained by the intergalactic workflow commission [Internet]. 2024. Available from: https://iwc.galaxyproject.org/

7. Santana DJ, O'Meara TR, Romo JA, et al. A *candida auris*-specific adhesin, Scf1, governs surface association, colonization, and virulence. Science [Internet]. 2023;381(6665):1461–7. Available from: https://pubmed.ncbi.nlm.nih.gov/37769084/

8. Wang Y, Zou Y, Chen X, et al. Cell surface glycan-lectin interactions modulate *candida auris* colonization and fungemia. Nature Communications [Internet]. 2024;15:6490. Available from: https://pubmed.ncbi.nlm.nih.gov/38562758/

9. RNA-seq analysis best practices [Internet]. Zenodo; 2020. Available from: https://zenodo.org/records/3985047

10. Act now: The global threat of candida auris and the urgent need for effective countermeasures. 2024; Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC11221456/

Supplementary Materials

Supplementary Table 1: *C. auris* WGS data contributors by organization category and top sequencing centers.

Panel A: Summary by Organization Category

Category	Organizations	Runs	% of Total
US State/Local Public Health Labs	26	20,552	78.4%
CDC	2	2,626	10.0%
Academic/Research	46	1,365	5.2%
Other	41	1,345	5.1%
International Public Health	5	313	1.2%
TOTAL	**120**	**26,201**	**100%**

Panel B: Top 15 Contributing Organizations

Organization	Full Name	Runs	%
UPHL_ID	Utah Public Health Laboratory	4,447	17.0%
NVSPHL	Nevada State Public Health Laboratory	4,363	16.7%
CDC-NCEZID-MDB	CDC Mycotic Diseases Branch	2,406	9.2%
MDH_CSL	Maryland Dept of Health, Central Services Lab	2,309	8.8%
TXDSHS	Texas Dept of State Health Services	1,487	5.7%
MDHHS-GS	Michigan Dept of Health & Human Services	1,289	4.9%
-	Wisconsin State Laboratory of Hygiene	1,211	4.6%
RIPHL	Rhode Island Public Health Laboratory	1,197	4.6%
NSPHL	Nevada State Public Health Laboratory	1,031	3.9%
-	Wadsworth Center (New York)	705	2.7%
-	Minnesota Dept of Health	688	2.6%
OCPHL_CA	Orange County Public Health Lab (California)	659	2.5%
-	Washington State Dept of Health	583	2.2%
UNLV NPM	Univ of Nevada Las Vegas, Pathogen Monitoring	443	1.7%
-	Fudan University	264	1.0%

US public health laboratories (state/local + CDC) account for 88.4% of all C. auris WGS data, reflecting outbreak surveillance priorities. Nevada appears twice (NVSPHL + NSPHL = 5,394 runs, 20.6%), indicating major outbreak focus.

Supplementary Table 2: RNA-seq methodology across 20 published *C. auris* BioProjects with linked publications (2018-2025).

BioProject	PMID	Authors	Year	Runs	Reference Genome	RNA-seq Tools
PRJEB57846	39297640	Rhodes J et al.	2024	12	NS	WGS, RNA-seq
PRJNA1012821	40468551	Chauhan A et al.	2025	16	B8441, B11220 (CGD)	FastQC, fastp, Bowtie2, HTSeq, DESeq2
PRJNA1015296	38493178	Bing J et al.	2024	141	B8441 (GCA_002759435.2)	HiSat2, StringTie, DESeq2, BWA
PRJNA1036037	39480072	Li J et al.	2024	22	Clade IV	RNA-seq
PRJNA1086003	39455573	Wang TW et al.	2024	13	B8441 (Clade I)	HISAT2, STAR, DESeq2
PRJNA1139166	40099908	Phan-Canh T et al.	2025	15	B8441 (GCA_002759435.2)	FastQC, fastp, cutadapt, STAR, featureCounts
PRJNA1208975	40530673	Yang G et al.	2025	9	Clade I	RNA-seq
PRJNA1232830	40066990	Chauhan M et al.	2025	6	Clade I	RNA-seq
PRJNA1291775	40863525	Vidal-Montiel A et al.	2025	6	GCA_003014415.1, GCA_034640365.1 (Clades III, IV)	FastQC, Trimmomatic, STAR, featureCounts, DESeq2
PRJNA445471	30559369	Muñoz JF et al.	2018	24	B8441, B11220, B11243	Bowtie2, TopHat2, RSEM, Trinity, edgeR
PRJNA477447	29997121	Kean R et al.	2018	22	B8441 (de novo)	Trinity, HISAT2, Kallisto, DESeq2
PRJNA682185	34630944	Zamith-Miranda D et al.	2021	36	B8441 (GCA_002759435.2)	DESeq2, edgeR
PRJNA682422	34180774	Lara-Aguilar V et al.	2021	6	B8441 (GCA_002759435.2)	FastQC, Trimmomatic, fastp, STAR, featureCounts, DESeq2
PRJNA735406	34354695	Zhou W et al.	2021	6	B11221 (Clades I-V)	Trimmomatic, HISAT2, Cufflinks, HTSeq, DESeq2

BioProject	PMID	Authors	Year	Runs	Reference Genome	RNA-seq Tools
PRJNA788930	35652307	Shivarathri R et al.	2022	12	NS	RNA-seq
PRJNA792028	36913408	Bing J et al.	2023	15	GCA_002759435.2, GCF_002775015.1	HiSat2, StringTie, DESeq2, BWA
PRJNA801628	35473297	Biermann AR et al.	2022	24	B8441, B11221, B11243 (Clades I, III, IV)	HISAT2, featureCounts, edgeR
PRJNA830685	36445083	Narayanan A et al.	2022	16	B8441, CBS10913 (Clade II)	FastQC, fastp, BWA, Bowtie2, HTSeq, DESeq2
PRJNA902676	38722168	Yang B et al.	2024	40	B11220, B11221 (Clades II, III)	Kallisto, DESeq2
PRJNA904261	37769084	Santana DJ et al.	2023	6	B8441 (Clade I)	RNA-seq

NS = Not specified in available text. Data extracted from PMC full text via EuropePMC and NCBI E-utilities APIs. 44 additional BioProjects had no linked publications.

Supplementary File 1: Galaxy workflow diagrams for re-analyses. *Source: https://github.com/nekrut/claude-projects/tree/main/rnaseq/santana24_PRJNA904261/analysis and https://github.com/nekrut/claude-projects/tree/main/rnaseq/wang24_PRJNA1086003/analysis*

Supplementary File 2: Gene mapping tables with LFC correlation values. *Source: Same repositories as above*

Supplementary File 3: Analysis reports with complete methodological details. *Source: ANALYSIS_REPORT.md in each repository*

Manuscript generated with Claude Code (Anthropic) Draft version for iteration - December 3, 2025