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].

The amount of public sequencing data available for *C. auris* is relatively modest (Table 1).

Table 1: Summary of *C. auris* sequencing data in NCBI SRA (December 2025)

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**	

RNA sequencing has become major methodology for studying *C. auris* biology and pathogenesis.

A literature survey combining NCBI GEO database mining and PubMed/Europe PMC searches identified 32 published RNA-seq studies spanning 2018-2025 (Table 2).

Table 2: Summary of *C. auris* RNA-seq literature survey (32 studies, 2018-2025)

Category	Finding
Research Focus	Drug resistance (34.4%), Stress response (18.8%), Biofilm (12.5%), Host-pathogen (12.5%)
Peak Year	2021 (11 papers, 34.4% of total)
Reference Genome	B8441/GCA_002759435 (75% of studies)
Alignment Tool	HISAT2 (62.5%), STAR (25%), Bowtie2 (12.5%)
Quantification	HTSeq/featureCounts (50%), Cufflinks (18.8%)
DE Analysis	DESeq2 (68.8%), edgeR (12.5%), Cuffdiff (9.4%)

Analysis of NCBI Sequence Read Archive (SRA) shows that 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 differences: 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 bioinformatics pipeline has emerged: HISAT2 (62.5%), HTSeq/featureCounts (50%), and DESeq2 (68.8%). Given that RNA-seq accounts for over one-quarter of *C. auris* research projects and is primary approach for understanding pathogen biology, standardizing RNA-seq analysis is a critical priority.

Despite this growth, analysis approaches remain inconsistent regarding reference genomes and tool versions. Survey of 32 studies revealed that while 75% use B8441 reference genome (GCA_002759435 family), annotation versions vary widely—some cite only "B8441" without version, others specify precise assemblies (GCA_002759435.2, GCA_002759435.3, or variants like "s01-m01-r11") (Figure 1). This creates challenges for cross-study comparison and reproducibility. Gene identifiers differ between annotation versions: older annotations use 6-digit suffixes (e.g., B9J08_001458) while newer versions use 5-digit suffixes (e.g., B9J08_03708). Tool versions are frequently unspecified—while the field converges on HISAT2->HTSeq->DESeq2, exact versions and parameters vary. These findings underscore need for standardized platforms specifying precise genome versions, tool versions, and parameters—exactly what BRC-Analytics provides.

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 with reference genome

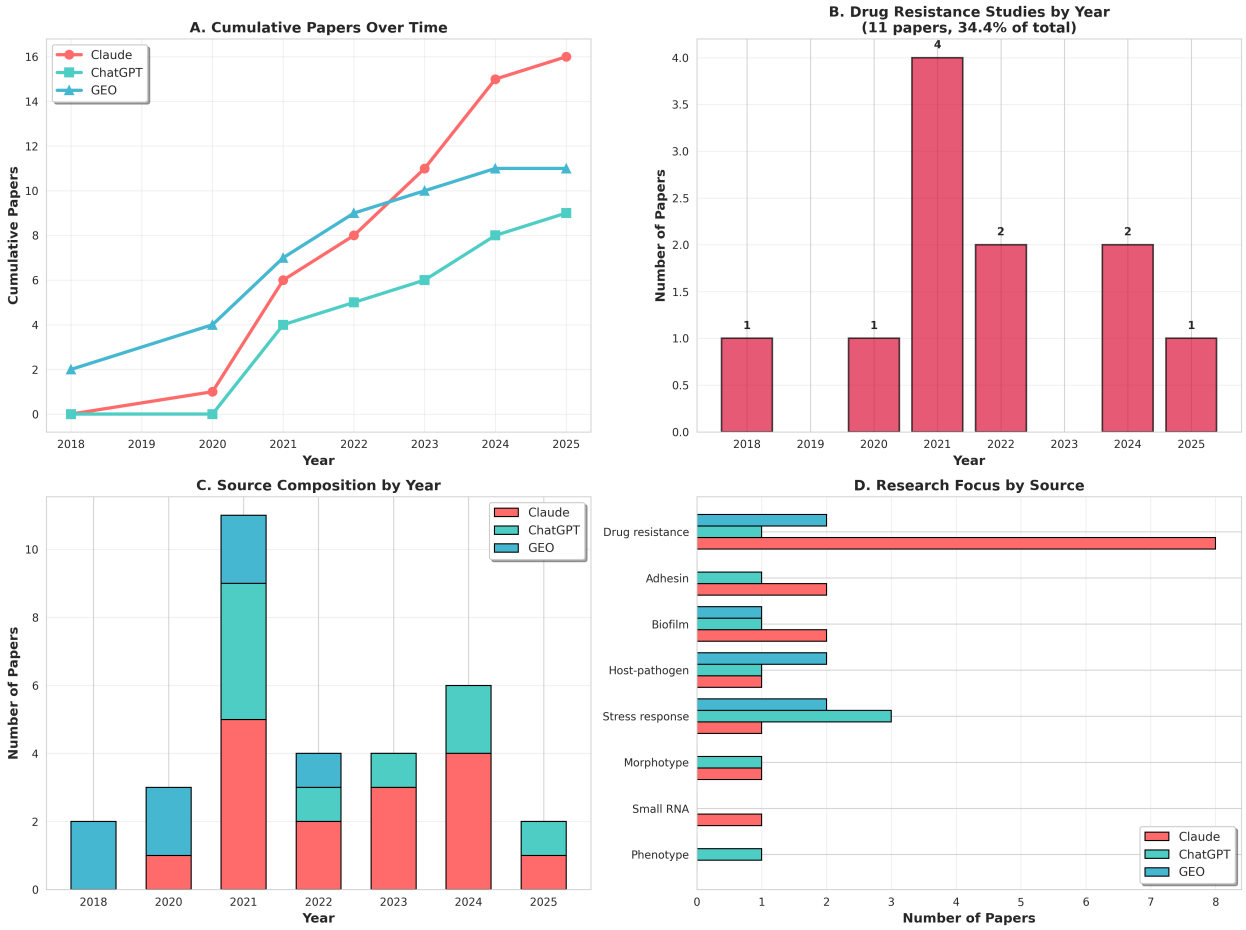

Figure 1: **Figure 1**: Challenges in *C. auris* RNA-seq standardization. (A) Genome annotation version variability across 32 studies, (B) Gene ID format differences between annotation versions requiring reconciliation, (C) Tool version reporting inconsistency.

GCA_002759435.3 and IWC workflows, our results correlate with published findings (R^2 > 0.98), validating the platform's reproducibility.

Results and Discussion

Obtaining Data from BRC-Analytics

[**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 2).

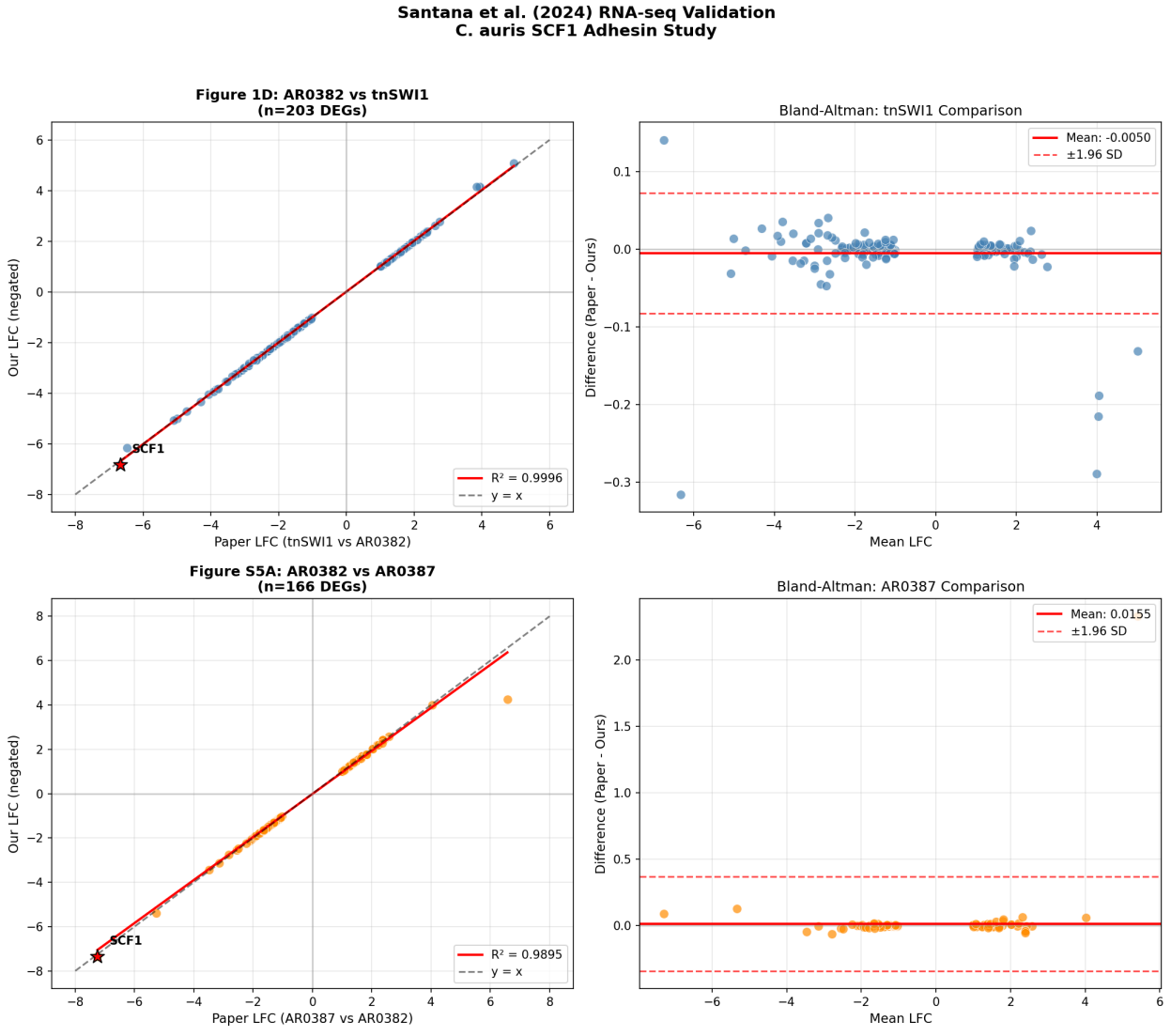

Figure 2: **Figure 2**: Validation of Santana et al. (2023) re-analysis. (A) Scatter plot of log2FC values for 203 mapped DEGs (AR0382 vs tnSWI1), Pearson R^2=0.9996. (B) Scatter plot of log2FC values for 166 mapped DEGs (AR0382 vs AR0387), Pearson R^2=0.9895. (C) SCF1 gene expression across conditions showing consistent downregulation in adhesion-deficient strains.

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 3). 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.

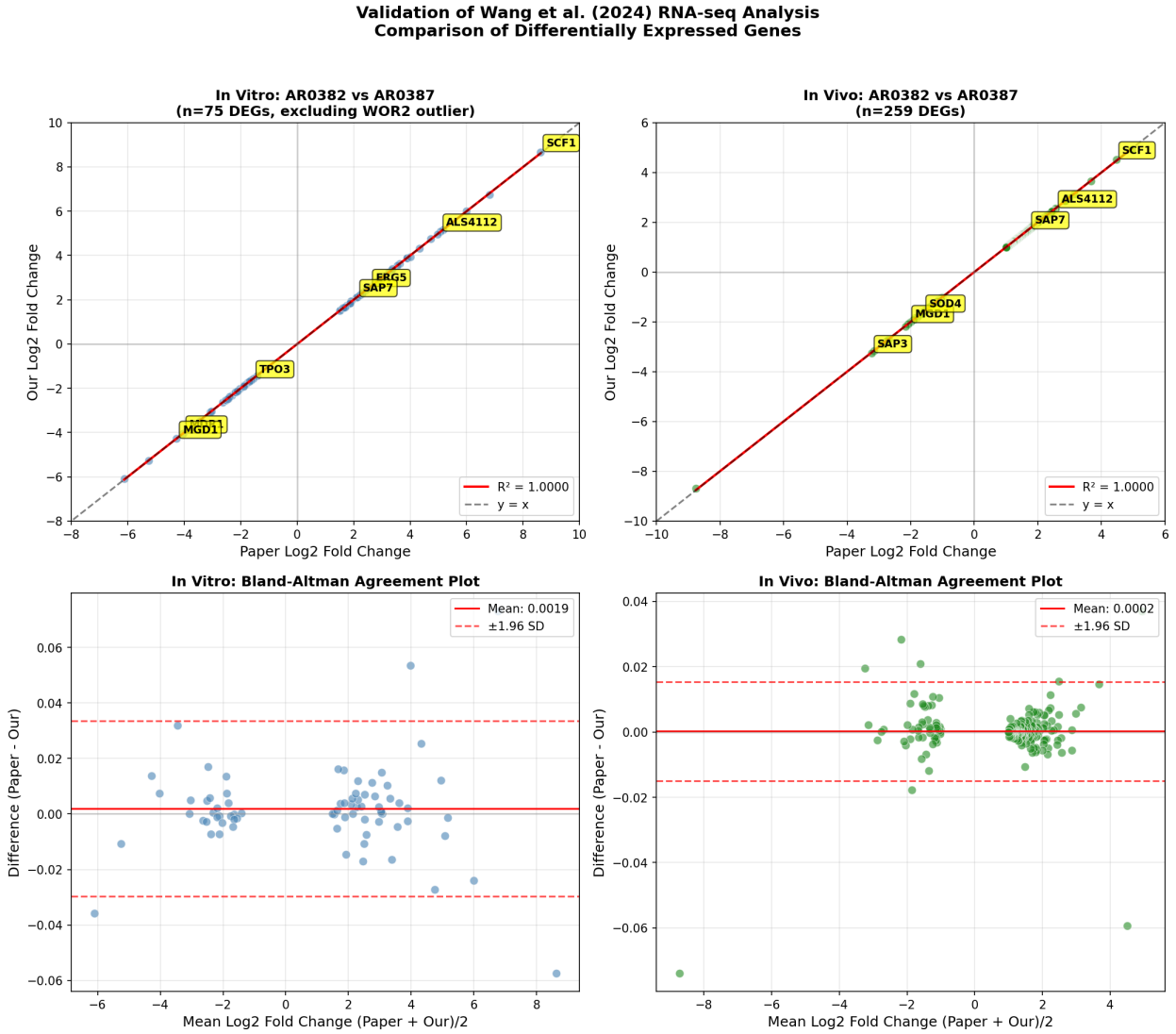

Figure 3: **Figure 3**: Validation of Wang et al. (2024) re-analysis. (A) Scatter plot of *in vitro* DEG log2FC values (73 genes, Pearson r=0.9914). (B) Scatter plot of *in vivo* DEG log2FC values (~195 genes, Pearson r=1.0000). (C) Heatmap of key adhesion and drug efflux genes showing LFC validation (all differences < 0.1).

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 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

We conducted literature survey of *C. auris* RNA-seq studies using NCBI Gene Expression Omnibus (GEO) database mining combined with PubMed/Europe PMC searches. Searches performed December 2, 2025, identifying 32 unique RNA-seq studies spanning 2018-2025.

To quantify proportion of RNA-seq data in *C. auris* research, we analyzed complete NCBI SRA database for taxonomy ID 498019 (*Candidozyma auris*) accessed December 3, 2025. Analysis of 27,201 total runs across 237 BioProjects revealed 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.

For re-analysis validation, we selected Santana et al. (2023) *Science* (PRJNA904261) [7] and Wang et al. (2024) *Nature Communications* (PRJNA1086003) [8]. Survey analysis scripts available at https://github.com/nekrut/claude-projects/tree/main/rnaseq/Cauris_rna_seq_survey.

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, others. 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, others. 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: Complete list of 32 *C. auris* RNA-seq studies with PMIDs, genome versions, tools, and research focus. *Source: https://github.com/nekrut/claude-projects/blob/main/rnaseq/Cauris_rna_seq_survey/combined/combined_data.csv*

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