Comparative Analysis of Authenticated Genomic and Phenotypic Data for *Yarrowia lipolytica* Reference Strains

ATCC®
Credible leads to Incredible®

James Crill, MSc;[1] Anthony Muhle, MSc;[2] John Bagnoli, BS;[2] Briana Benton, PMP;[2] Ana Fernandes, BS;[2] Nikhita Puthuveetil, MSc;[2] Corina Tabron, MSc;[2] Scott V. Nguyen, PhD;[2] Shahin Ali, PhD;[2] Jonathan Jacobs, PhD[2] | [1]Syracuse University, Syracuse, New York 13244; [2]ATCC, Manassas, VA 20110

Abstract

Yarrowia lipolytica stands as a pivotal microbial platform in the realm of bioindustrial applications. This presentation delves into a comparative analysis encompassing genomic and phenotypic data from over 30 haploid and diploid strains of *Y. lipolytica* within the ATCC® reference collection. Employing long-read (Oxford Nanopore Technologies®) and short-read (Illumina®) sequencing technologies, we pursued whole-genome sequencing and *de novo* hybrid assembly for each strain, all conducted under ISO 9000 quality standards at ATCC®.

Phylogenetic scrutiny revealed five distinct sub-clades among *Y. lipolytica* reference strains, unveiling conserved alleles within shared metabolic pathways. Subsequently, representative strains from each sub-clade underwent comprehensive phenotypic profiling on the OmniLog® platform (Biolog®). Discrepancies in metabolic responsiveness to diverse growth conditions correlated with allelic distinctions within each subclade. This multidimensional approach identifies strains boasting optimal traits for augmented biofuel and biomaterials synthesis.

Here, we provide an overview of foundational reference data, which is accessible through the **ATCC® Genome Portal (https://genomes.atcc.org)**. By *optimizing Y. lipolytica* strains, our study contributes to the sustainable evolution of bioindustrial applications, propelling advancements in biofuel and biomaterials production.

Strains in the ATCC® Collection

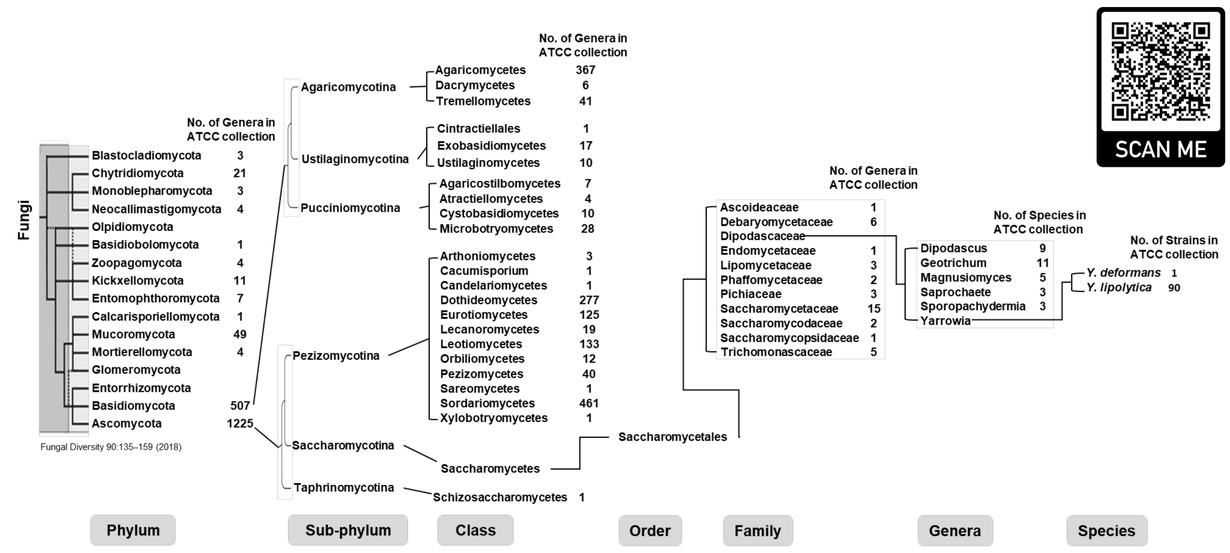
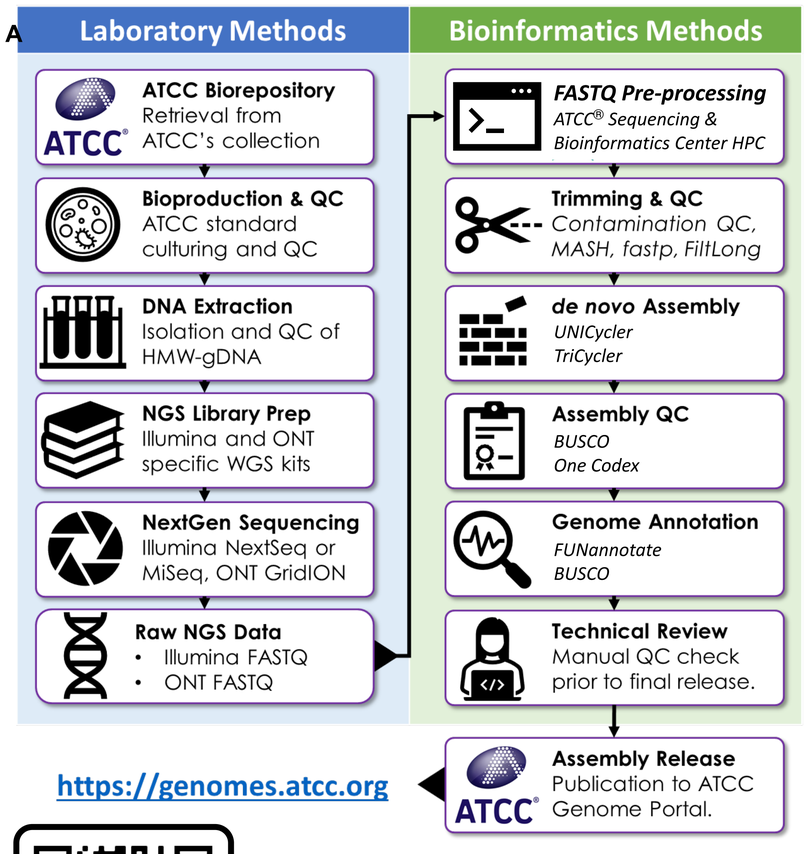


Figure 1: 89 strains of *Yarrowia lipolytica* are publicly available in ATCC®'s general collection. Taxonomy lineage highlighting these strains and the number of additional strains available across our mycology collection.

Yarrowia lipolytica Genome Assemblies



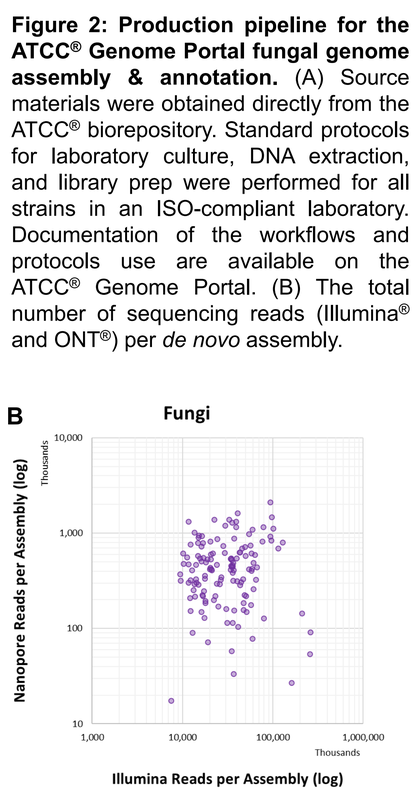


Figure 2: Production pipeline for the ATCC® Genome Portal fungal genome assembly & annotation. (A) Source materials were obtained directly from the ATCC® biorepository. Standard protocols for laboratory culture, DNA extraction, and library prep were performed for all strains in an ISO-compliant laboratory. Documentation of the workflows and protocols use are available on the ATCC® Genome Portal. (B) The total number of sequencing reads (Illumina® and ONT®) per *de novo* assembly.

Yarrowia lipolytica reference genome assemblies on the ATCC® Genome Portal

Nguyen SV, et al. "***The ATCC Genome Portal: 3,938 Authenticated Microbial Reference Genomes.***" Microbiology Resource Announcements (2024) e01045-23.

Phylogeny

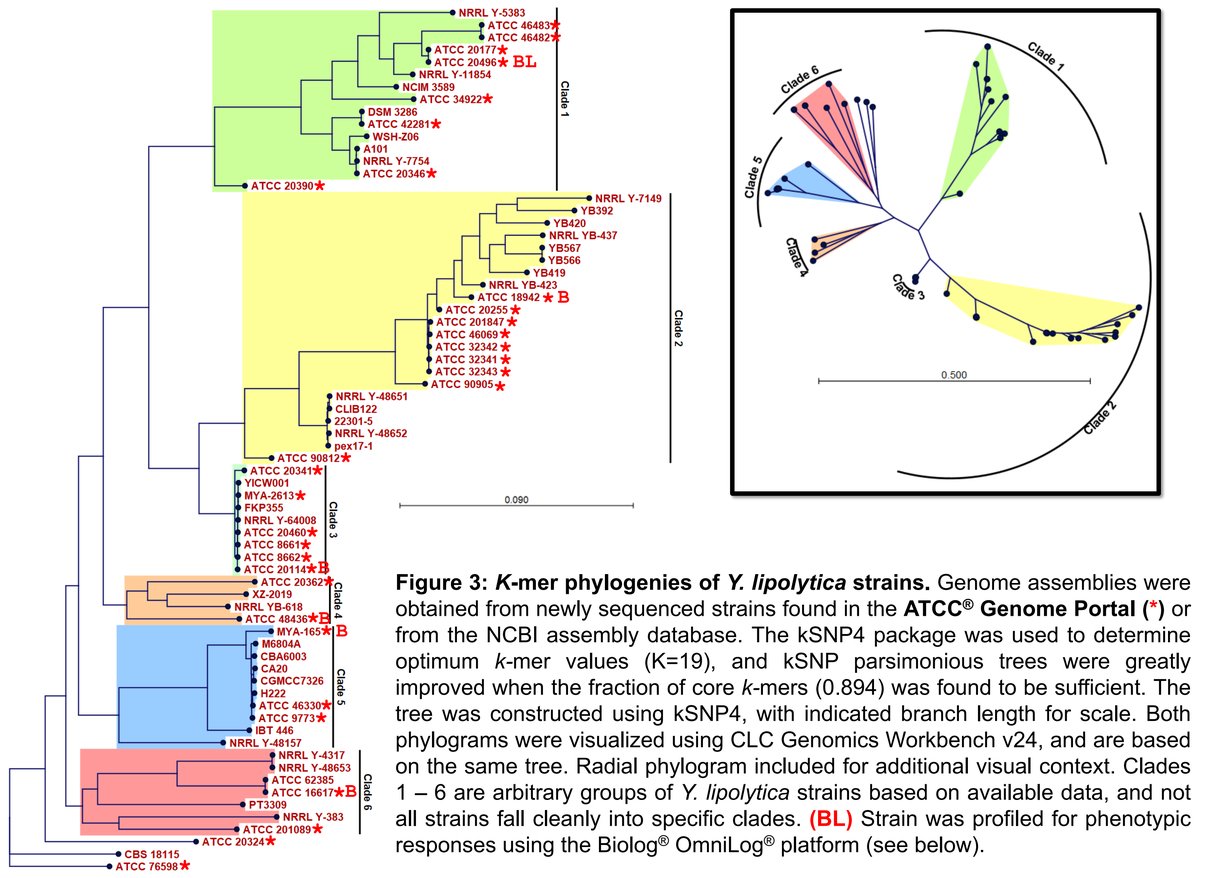
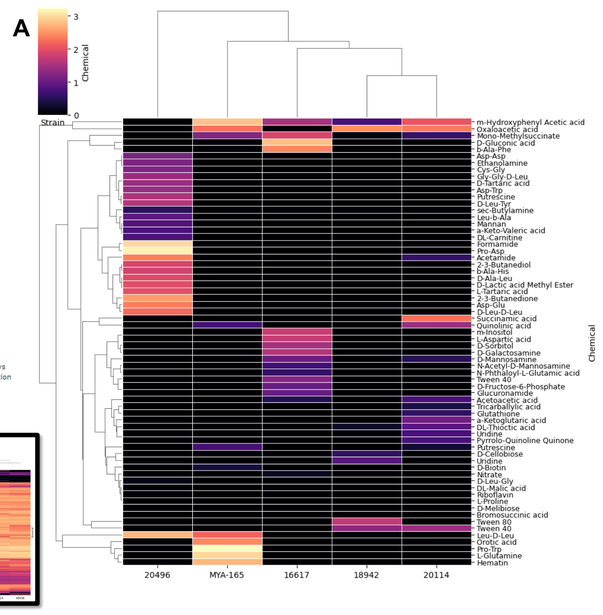


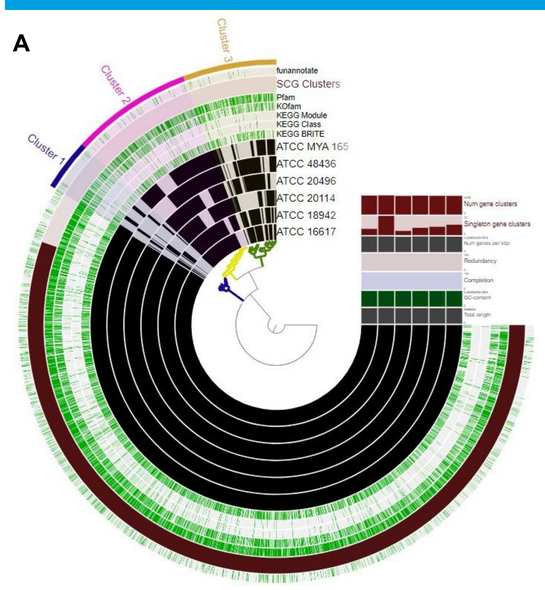
Figure 3: *K*-mer phylogenies of *Y. lipolytica* strains. Genome assemblies were obtained from newly sequenced strains found in the **ATCC® Genome Portal (*)** or from the NCBI assembly database. The kSNP4 package was used to determine optimum *k*-mer values (K=19), and kSNP parsimonious trees were greatly improved when the fraction of core *k*-mers (0.894) was found to be sufficient. The tree was constructed using kSNP4, with indicated branch length for scale. Both phylograms were visualized using CLC Genomics Workbench v24, and are based on the same tree. Radial phylogram included for additional visual context. Clades 1 – 6 are arbitrary groups of *Y. lipolytica* strains based on available data, and not all strains fall cleanly into specific clades. **(BL)** Strain was profiled for phenotypic responses using the Biolog® OmniLog® platform (see below).

Phenotypic Profiling

Figure 4: Phenotypic profiling of representative strains (A) Hierarchical clustering of outliers for log-transformed values of growth in presence of various chemicals from Biolog® OmniLog®. (B) Principal Component Analysis (PCA) of chemicals across strains, color-coded by assay type. Each data point represents a chemical, and its position reflects its similarity to other chemicals based on their expression patterns across strains. (INSET) Data for all 950 chemical exposure conditions tests on the Biolog® OmniLog® platform.



Comparative Genomics



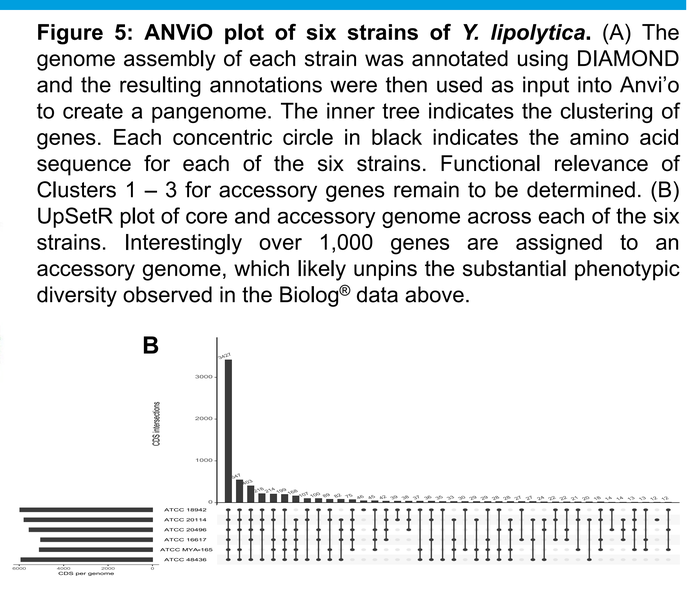


Figure 5: ANViO plot of six strains of *Y. lipolytica*. (A) The genome assembly of each strain was annotated using DIAMOND and the resulting annotations were then used as input into Anvi'o to create a pangenome. The inner tree indicates the clustering of genes. Each concentric circle in black indicates the amino acid sequence for each of the six strains. Functional relevance of Clusters 1 – 3 for accessory genes remain to be determined. (B) UpSetR plot of core and accessory genome across each of the six strains. Interestingly over 1,000 genes are assigned to an accessory genome, which likely unpins the substantial phenotypic diversity observed in the Biolog® data above.

Summary

Here, we explored the connection between genomic and phenotypic diversity across *Yarrowia lipolytica* strains held in ATCC®'s mycology collection.

- The *Y. lipolytica* pangenome consists of 6,462 genes, which collectively includes an estimated 3,427 genes (100% identity) forming a core genome for the species.
- Three distinct clusters of genes were found across the six strains, which may represent the phenotypic differences observed. Additional research is needed to understand these genes and the potential role they play in each strain.

ATCC 10801 University Boulevard, Manassas, Virginia 20110-2209 **Phone:** 800.638.6597 **Email:** sales@atcc.org **Web:** www.atcc.org