# Synergy2: 04/10/2025 Output Analysis

**Author:** Hubert Kicinski

**Affiliation:** Dr. Bin Z He Lab @ The University of Iowa

**Date:** April 2025

**Contact:** hkicinski@uiowa.edu

## 1. Introduction

This report presents the results of an orthology analysis using SYNERGY2, performed on **04/10/2025** to detect orthologous relationships between four yeast species:

- *Candida albicans*
- *Candida glabrata*
- *Kluyveromyces lactis*
- *Saccharomyces cerevisiae*

To ensure that the predictions obtained via the SYNERGY2 run are robust, these predictions were compared against reference orthology data from YGOB (Yeast Gene Order Browser) and CGOB (Candida Gene Order Browser) databases to evaluate performance.

## 2. Analysis Pipeline

The analysis workflow consisted of three main components:

1. **Reference Data Preparation** (`parse_pillars_for_comp.py`)
    - Processed both YGOB and CGOB phylogenetic pillar files
    - Extracted ortholgo pairs for all six species pairs
    - Used *S. cerevisiae* as a "bridge" to connect the isolated *C. albicans* that is found on the CGOB dataset
    - Created standardized reference files.

The reference files obtained were stored in the `Homology_data` subdirectory. These reference files include the following:

- `C_albicans_C_glabrata_pairs`
- `C_albicans_K_lactis_pairs`
- `C_albicans_S_cerevisiae_pairs`
- `C_glabrata_K_lactis_pairs`
- `C_glabrata_S_cerevisiae_pairs`
- `K_lactis_S_cerevisiae_pairs`

These 6 pairwise combinations were obtained by considering all unordered combinations.

$$\binom{4}{2} = 6$$

2. **SYNERGY2 Ortholog Extraction** (`extract_synergy_orthologs.py`)
    - Extracted ortholog pairs from SYNERGY2 clustering results (from 04/10/2025 execution)
    - Identified species and gene for each homology cluster based on identifier patterns (e.g., "orf19." for *C. albicans*)
    - Generated ortholog pair files for all six species combinations
3. Orthology Comparison (`compare_orthologs.py`)
    - Standardized gene identifiers to enable accurate comparison
    - Compared SYNERGY2 predictions against reference orthologs
    - Calculated precision, recall, and F1 scores
    - Generated visualizations and reports

## 3. Performance Metrics

The analysis used standard evaluation metrics for orthology prediction. The motivation of this set analysis comes from this bioRxiv pre-print that compared algorithmic detection of homologous sequences (via RNAcMap2) in RNA to a known database (Rfam). The similarity in context--comparing standard database results to an algorithmic result--motivated the use of an $F_1$ Score Analysis for these obtained results. The evaluation metrics used in this analysis are as follows:

- **Precision**: Fraction of SYNERGY2 ortholog predictions that match the reference data (CGOB+YGOB)

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

- **Recall**: Fraction of reference orthologs correctly identified by SYNERGY2

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

- **$F_1$ Score**: Harmonic mean of precision and recall

$$F_1 ;=; \frac{2 \times (\text{Precision} \times \text{Recall})}{\text{Precision} + \text{Recall}}$$

## 4. Results

### 4.1 Overall Performance

These results are published in the `orthology_comparison_report.md` that has been published via GitHub-Hkicinski.

| Species Pair | Precision | Recall | $F_1$ Score | SYNERGY2 Pairs | Reference Pairs | True Positives |
|---|---|---|---|---|---|---|
| C. albicans – S. cerevisiae | 0.8342 | 0.9033 | 0.8674 | 4,343 | 4,011 | 3,623 |
| K. lactis – S. cerevisiae | 0.8596 | 0.8850 | 0.8721 | 4,700 | 4,565 | 4,040 |
| C. albicans – K. lactis | 0.8147 | 0.9012 | 0.8558 | 3,988 | 3,605 | 3,249 |

| Species Pair | Precision | Recall | $F_1$ Score | SYNERGY2 Pairs | Reference Pairs | True Positives |
|---|---|---|---|---|---|---|
| C. glabrata – K. lactis | 0.8672 | 0.3358 | 0.4841 | 1,740 | 4,494 | 1,509 |
| C. albicans – C. glabrata | 0.7815 | 0.3345 | 0.4685 | 1,542 | 3,602 | 1,205 |
| C. glabrata – S. cerevisiae | 0.7070 | 0.3367 | 0.4561 | 2,191 | 4,601 | 1,549 |

### 4.2 Species-Specific Patterns

- Strong performance (F1 > 0.85) for three species pairs:
    - C. albicans - S. cerevisiae
    - K. lactis - S. cerevisiae
    - C. albicans - K. lactis
- Moderate performance (F1 ~ 0.45-0.49) for all C. glabrata comparisons:
    - C. glabrata - K. lactis
    - C. albicans - C. glabrata
    - C. glabrata - S. cerevisiae
- C. glabrata pattern: Consistent in relatively high precision (0.71-0.87) but low recall (0.33-0.34)

## 5. Validation

A sanity check was performed to investigate why C. glabrata comparisons showed lower recall. We examined a random sample of "false negative" ortholog pairs (present in reference data but missed by SYNERGY2) using `sanity-cgla-synvygob.py`.

### 5.1 Sanity Check Results

For all sampled gene pairs, the following trends were observed:

- Both genes existed in the SYNERGY2 dataset
- SYNERGY2 placed them in different orthology clusters
- Each gene was assigned orthologs from other species

Examples:

- CAGL0C01463g (Cluster3571) vs. YPL180W (Cluster10714)
- CAGL0H08910g (Cluster1428) vs. YGL138C (Cluster4373)
- CAGL0H03091g (Cluster4054) vs. YGL091C (Cluster3316)

Though there appears to be a misalignment in the orthologous gene clustering mechanism between SYNERGY2 and the synteny-aided genome order browser databases, additional analysis must be performed to identify the potential root of this misalignment in homologous gene clustering.

## 7. Preliminary Conclusions

1. SYNERGY2 demonstrates strong orthology prediction performance for most yeast species pairs, with excellent concordance with reference databases for S. cerevisiae, K. lactis, and C. albicans relationships.
2. The lower recall for C. glabrata comparisons reflects biological differences and algorithm choices rather than methodological errors.
3. The analysis pipeline successfully integrated YGOB and CGOB reference data, allowing comprehensive orthology validation across all four yeast species.
4. SYNERGY2 orthology predictions are particularly reliable for C. albicans - S. cerevisiae, K. lactis - S. cerevisiae, and C. albicans - K. lactis relationships.
5. For C. glabrata comparisons, SYNERGY2 makes accurate predictions (high precision) but identifies a different subset of orthologs than reference databases (lower recall).

Nevertheless, the discrepancy of *C. glabrata's* orthogroup clustering will be further explored in prevailing README documents.

## References

1. Wapinski I, Pfeffer A, Friedman N, Regev A. “Natural history and evolutionary principles of gene duplication in fungi.” *Nature*. 2007;449(7158):54–61.

2. Byrne KP, Wolfe KH. “The Yeast Gene Order Browser: combining curated homology and syntenic context reveals gene fate in polyploid species.” *Genome Research*. 2005;15(10):1456–1461.

3. Fitzpatrick DA, O'Gaora P, Byrne KP, Butler G. “Analysis of gene evolution and metabolic pathways using the Candida Gene Order Browser.” *BMC Genomics*. 2010;11:290.

4. Singh J, Paliwal K, Singh J, Litfin T, Zhou Y. “Improved RNA homology detection and alignment by automatic iterative search in an expanded database.” *bioRxiv*. 2022.