Comparative Proteomic Analysis of *Saccharomyces cerevisiae* Strains and Mutants under Varied Growth Conditions

Ignacy Makowski

June 25, 2025

Abstract

This report describes a study on *Saccharomyces cerevisiae* proteins. Data from W303 and BY4742 yeast strains were reanalyzed and new data from maf1Δ and rpc128-1007 mutant strains were included. In both data sources mass spectrometry was used to measure protein levels, and then the data was analyzed with statistical methods and hierarchical clustering. A key step in this analysis was the use of Combat [1] to correct for differences between experiments, which was vital for combining data from multiple sources. The results show that different yeast strains and conditions (such as growth phase, carbon source, and temperature) lead to distinct protein patterns. Differentially abundant proteins between W303 and BY4742 strains and maf1Δ and rpc128-1007 mutants were found in different growth phases.

1 Introduction

1.1 Background on *Saccharomyces cerevisiae* Strains

Saccharomyces cerevisiae is a key model organism in biology due to its easy genetic manipulation and conserved cellular processes. In this study, two common lab strains, W303 and BY4742, and two mutants, maf1Δ and rpc128-1007, were examined to understand how changes in genes and the environment affect their proteomes. W303 (MATa, ade2-1, trp1-1, leu2-3,112, his3-11, 15, ura3-1, ssd1, can1-100, psy+) is often used for cell cycle studies because it can be easily synchronized. BY4742 (MATa, his3Δ1, leu2Δ0, lys2Δ0, ura3Δ0) is a standard strain with many molecular tools available [2]. Comparing their proteins helps to understand how their genetic differences lead to different traits. The analysis was extended by adding two additional mutants maf1Δ and rpc128-1007. The maf1Δ mutant lacks the *MAF1* gene, which is a negative regulator of RNA Polymerase III (RNAP III), an enzyme that synthesizes tRNAs and 5S rRNAs. Without *MAF1*, RNAP III activity increases, which has been shown to be associated with a decrease in TCA cycle activity, causing mitochondrial dysfunction and affecting basic cell processes. This leads to poor growth on non-fermentable carbon sources (such as glycerol) at 30°C and death at higher temperatures. Maf1 also affects fat metabolism, cell growth, and lifespan. The rpc128-1007 mutant has a specific change in its *RPC128* gene, which encodes a subunit of RNAP III. This mutation impairs the enzyme's ability to increase tRNA synthesis, leading to reduced overall tRNA levels, harming RNAP III assembly, and stopping cell division. Interestingly, rpc128-1007 can fix the growth problem of maf1Δ on non-fermentable carbon sources and exhibits increased activity of the TCA cycle under these conditions. It has been previously shown that glycolytic flux in S. cerevisiae is dependent on RNAP III and its regulator Maf1. Both mutants also exhibit different preferences in their use of glucose [3]. By studying these mutants under different food sources (glucose vs. glycerol) and temperatures (30°C vs. 37°C), their adaptation to stress and changes in metabolism can be observed.

1.2 Study Objectives

This project aimed to broadly study yeast proteomes, building on existing research by adding new genetic variations. In Rossio et al. (2023), proteins in W303 and BY4742 yeast strains during both rapid growth (exponential phase) and slower growth (stationary phase) were compared. The plan to recreate and expand this analysis involved the following:

1. **Recreation:** To repeat the main analyses of the original study (PCA, hierarchical clustering, volcano plots) using existing raw data. This serves to confirm that the earlier results are reliable.

2. **Expansion:** To add new comparisons with maf1Δ and rpc128-1007 mutants, under diverse conditions. This allows new biological questions to be addressed.

3. **Data Integration:** To use Combat for batch correction, which fixes technical differences between experiments.

2 Materials and Methods

2.1 Yeast Strains and Growth Conditions

The yeast strains W303, BY4742, maf1Δ, and rpc128-1007 were used. Their details are provided in Table 1.

Table 1: Yeast Strains and Growth Conditions

Strain Name	Genotype	Growth Medium	Carbon Source/Temperature	Growth Phase/Relevance
W303	MATa, ade2-1, trp1-1, leu2-3,112, his3-11, 15, ura3-1, ssd1, can1-100, psy+	YPD , 50 mg/L adenine	30°C	Exponential, Stationary (Cell cycle studies)
BY4742	MATa, his3Δ1, leu2Δ0, lys2Δ0, ura3Δ0	YPD , 50 mg/L adenine	30°C	Exponential, Stationary (Genomic/molecular tools)
maf1Δ	Deletion of MAF1 gene (RNAP III negative regulator)	YPD, YPGly	Glucose, Glycerol (30°C, 37°C)	Increased RNAP III activity, altered metabolism
rpc128-1007	Point mutation in RPC128 (RNAP III subunit)	YPD, YPGly	Glucose, Glycerol (30°C, 37°C)	Reduced tRNA synthesis, suppresses maf1Δ defect

W303 and BY4742 cells were grown in two phases: exponential (fast growth) and stationary (slower growth due to limited nutrients). For the maf1Δ and rpc128-1007 mutants and their controls, conditions YPD Glucose, YPGly 30°C, YPGly 37°C were used.

2.2 Proteomic Data Acquisition

2.2.1 maf1Δ and rpc128-1007 Mutant Strains

For the *maf1Δ* and *rpc128-1007* mutant strains, proteomic measurements were performed on a TripleTof instrument (Sciex) configured for SWATH (Sequential Window Acquisition of all Theoretical Mass Spectra). Data was analyzed using Sciex's proprietary software against a "YEAST" database dated 8/15/2017. UniProt accession numbers for identified proteins are available in supplementary Excel files [2].

2.2.2 W303 and BY4742 Mass Spectrometry and Data Processing

Mass spectrometric data for W303 and BY4742 strains were acquired using an Orbitrap Fusion Lumos mass spectrometer with a Proxeon NanoLC-1200 UHPLC and a FAIMSpro interface. An in-house manufactured 100 μm capillary column, packed with 35 cm of C18 beads, was used with a 90 min gradient. An RTS-MS3 method was employed. MS1 scans were acquired in the Orbitrap (60,000 resolution). MS2 analysis involved CID in the ion trap (35% NCE) following a Real Time Search (RTS) against an *S. cerevisiae* UniProt database. Subsequent MS3 spectra of matched peptides were acquired in the Orbitrap (50,000 resolution) using HCD (55% NCE). Gas-phase fractionation utilized two sets of FAIMS compensation voltages (CVs): -40/-60/-80 V and -30/-50/-70 V. Raw files were converted to mzXML using MSconvert. Database searching was performed with Comet against the *S. cerevisiae* UniProt database (plus a decoy version) with a 50 ppm precursor and 0.9 Da product ion tolerance. Peptide-spectrum matches (PSMs) and proteins were filtered to a 1% FDR. Proteins were quantified by summing the signal-to-noise

(S/N) of reporter ions, which were then column-normalized and represented as a relative percentage across all channels [2].

2.3 Statistical Analysis

2.3.1 General Statistical Analysis

Data visualization and analysis included volcano plots, hierarchical clustering and Principal Component Analysis (PCA). To correct for significant technical variations between experiments, which accounted for 98.9% of the initial data variance, Combat batch correction was applied. Volcano plots were used to visualize changes in abundance based on fold change and statistical significance. Hierarchical clustering with Euclidean distance and Ward's method was also used to group similar samples and proteins.

Table 2: Key Statistical Methods and In-Depth Parameter Analysis

Method/Function	Library	Argument = Assigned Value / Description	Script
`rowMeans`	dplyr	`na.rm = TRUE` (Calculates mean protein abundance from replicates, ignoring missing values)	Both
`pheatmap`	pheatmap	`cluster_rows = TRUE`, `cluster_cols = TRUE` (Performs hierarchical clustering on both proteins and samples)	Both
`prcomp`	stats	`scale. = TRUE` (Performs PCA on scaled data to prevent high-abundance proteins from dominating the analysis)	Both
`t.test`	stats	`t.test(x, y)` (Performs Welch's t-test on replicates to calculate p-values for differential abundance)	'protein_analysis2.R'
`p.adjust`	stats	`method = "BH"` (Adjusts p-values for multiple comparisons using the Benjamini-Hochberg method)	'protein_analysis2.R'
`case_when`	dplyr	`sig = case_when(log2FC > 1 & adj_p_value < 0.05 ...)` (Categorizes proteins as significantly up/down-regulated based on fold-change and adjusted p-value thresholds)	'protein_analysis2.R'
`ComBat`	sva	`batch = batch_vector`, `mod = model.matrix( 1, ...)` (Corrects for batch effects between datasets while preserving biological variance)	'protein_analysis_combat.R'

Explanation of Parameters

ComBat This function was used to correct for non-biological variability between the two datasets ('Table_S1' and 'Poland_Yeast') [1].

- **`batch = batch_vector`**: The `batch_vector` variable contained the identity of each sample, explicitly labeling them as belonging to either Table_S1 or Poland_Yeast. This allowed ComBat to identify and remove the systematic technical differences between the two experiments.

prcomp This function performs Principal Component Analysis to visualize the major sources of variation in the data.

- **`scale. = TRUE`**: It normalizes the intensity values for every protein to have unit variance before analysis.

t.test This function was used to perform a formal statistical comparison of protein abundance between experimental groups for volcano plot visualization.

- A Welch's two-sample t-test was applied to the replicate intensity values for each protein.
- The test calculates a p-value for each protein. To account for multiple comparisons, these p-values were adjusted using the Benjamini-Hochberg method.

3 Results

3.1 Recreation of Original Proteomic Analysis (W303 vs BY4742)

First, a reanalysis of proteins from yeast strains W303 and BY4742 was performed. A total of 4480 proteins were measured across 12 samples (W303 stationary replication 1-3, BY4742 stationary replication 1-3, W303 exponential replication 1-3, BY4742 exponential replication 1-3). To assess the consistency of the protein measurements, the Coefficient of Variation (CV) was calculated for each condition. The original W303/BY4742 dataset showed low CVs, with median values ranging from 3.1% to 3.5% across the different conditions (Figure 1). This indicates high reproducibility within the experimental replicates for this dataset.

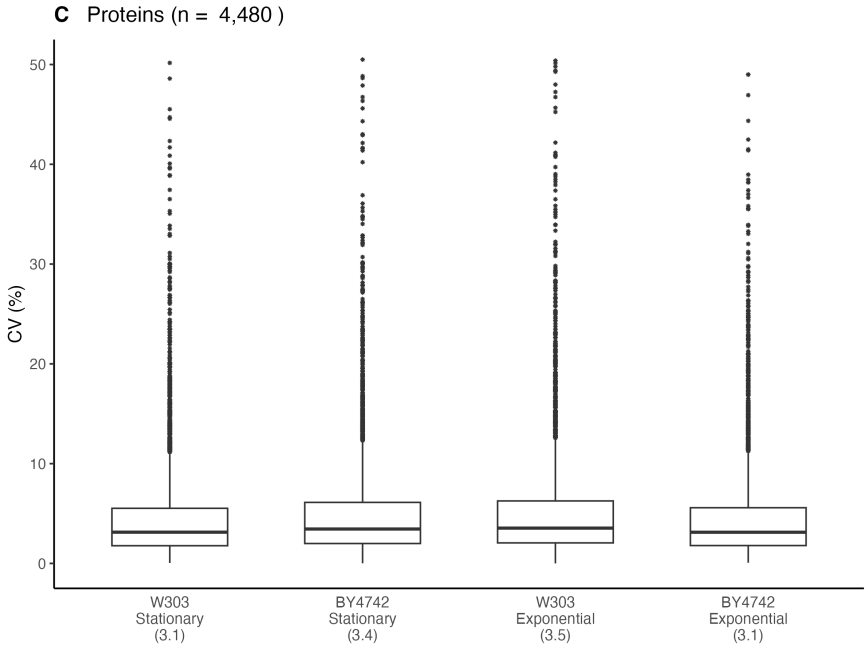

Figure 1: Distribution of Coefficient of Variation (CV) for Original Dataset Proteins. Box plots show CVs for proteins in the W303 and BY4742 dataset, separated by strain and growth phase. Median CVs are shown at the bottom of each box.

3.1.1 Hierarchical Clustering Analysis

The heatmap (Figure 2) shows clear separation between samples, with the growth phase being a decisive clustering factor.

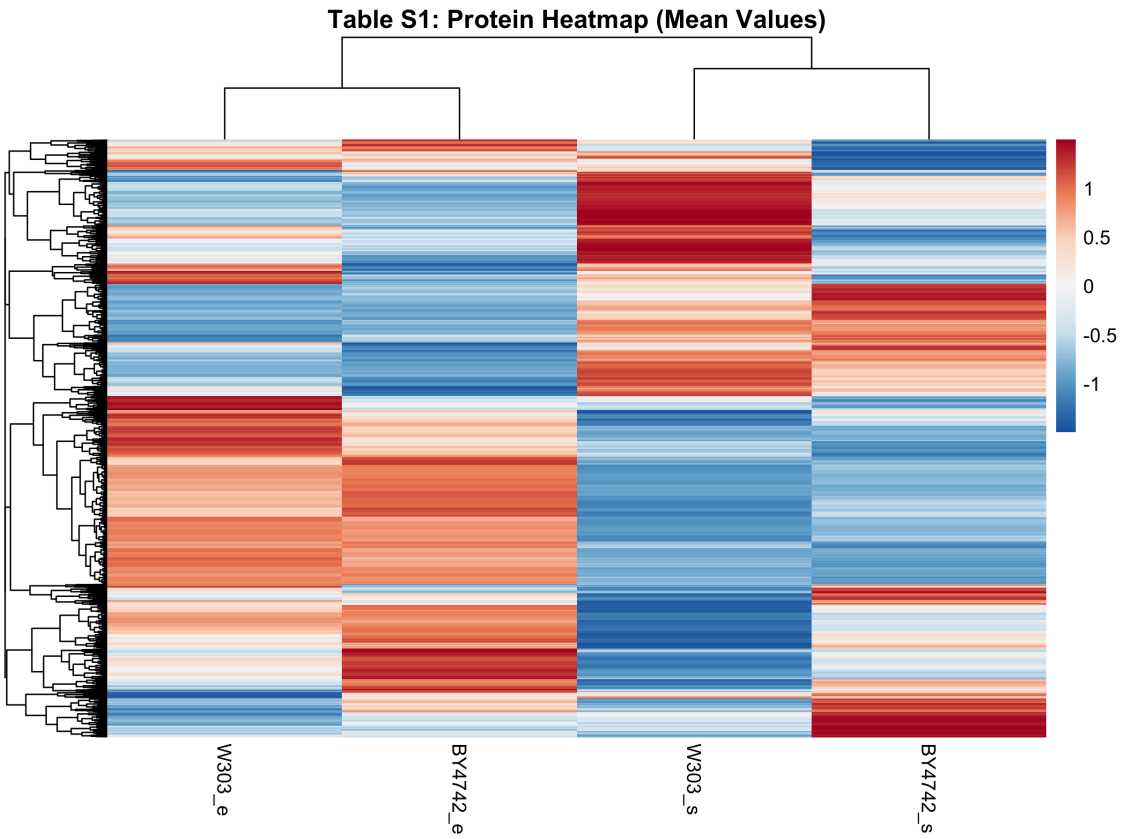

Figure 2: Hierarchical Clustering of W303 and BY4742 Proteomes. Heatmap showing protein abundance patterns across W303 and BY4742 strains in exponential (e) and stationary (s) growth phases. Red indicates higher protein abundance, blue indicates lower abundance.

3.1.2 Principal Component Analysis

PCA revealed the major sources of variation in the W303 and BY4742 proteome dataset. The first two principal components (PC1 and PC2) explained 64.5% and 25.8% of the total variance, respectively (Figure 3). Samples clustered clearly by both strain and growth phase.

3.1.3 Differential Protein Abundance Analysis

Volcano plot analysis identified significant differences in protein abundance between W303 and BY4742 strains in both growth phases. In the exponential phase comparison (Figure 4A), 231 proteins showed significant differential abundance, with 57 proteins more abundant in W303 and 174 more abundant in BY4742. The stationary phase comparison (Figure 4B) revealed more extensive changes, with 522 differentially abundant proteins: 218 higher in W303 and 304 higher in BY4742. Despite these substantial differences, nearly 90% of the proteins did not change significantly between strains.

3.2 Expansion of Proteomic Analysis to New Strains (maf1Δ, rpc128-1007)

The study was expanded to compare maf1Δ and rpc128-1007 mutants with wild type yeast under various conditions. The consistency of protein measurements in this new dataset, comprising 1,660 proteins, was assessed using the Coefficient of Variation (CV), as shown in Figure 5. The median CVs for this set ranged from 12.4% to 27.9%, indicating more variability compared to the original W303/BY4742 dataset (median CVs of 3.1% to 3.5%). This difference in reproducibility is an important consideration for data interpretation.

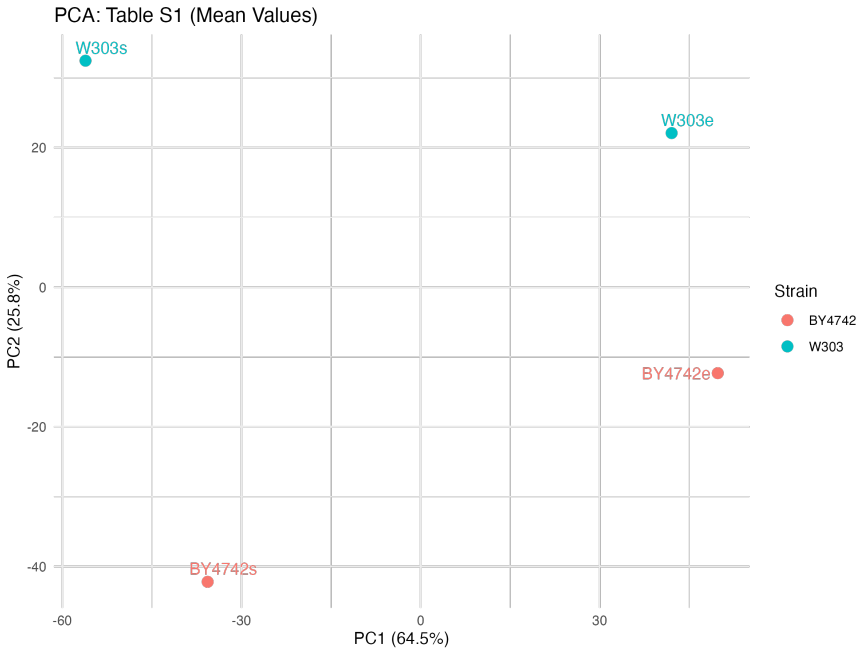

Figure 3: Global Proteome Differences (PCA of W303 and BY4742). This PCA plot shows how W303 and BY4742 yeast strains cluster in exponential (e) and stationary (s) growth phases. PC1 and PC2 explain 64.5% and 25.8% of the variation, respectively, clearly separating samples by strain and growth condition.

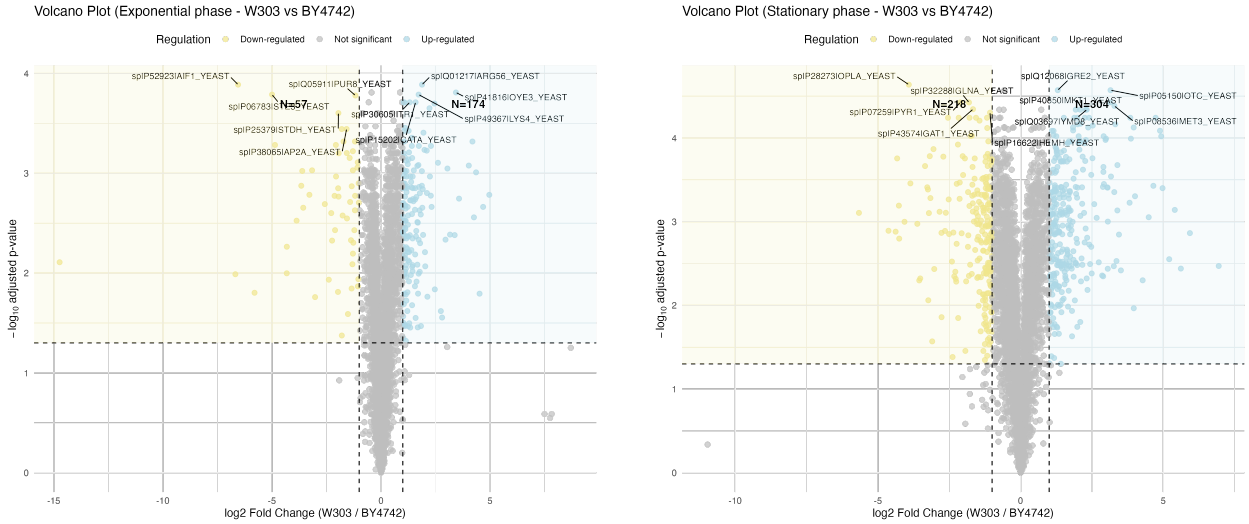

Figure 4: Volcano Plots of Differential Protein Abundance (Mean Values). (A) Exponential phase (W303 vs. BY4742): Shows 58 proteins higher in W303 (left, yellow) and 177 higher in BY4742 (right, blue). (B) Stationary phase (W303 vs. BY4742): Shows 218 proteins higher in W303 and 302 higher in BY4742. These plots display log2 Fold Change (x-axis) versus statistical significance (y-axis, $-log_{10}$ p-value).

A heatmap (Figure 6) showed how protein levels for 1654 proteins clustered across different strains and conditions. Samples were grouped mainly by carbon source (YPD vs. YPGly), then by temperature and strain within the YPGly conditions. Red colors indicate higher protein levels, while blue means lower.
PCA of the new dataset (Figure 7) showed that the carbon source was the biggest factor separating samples (YPD Glucose vs. YPGly conditions).

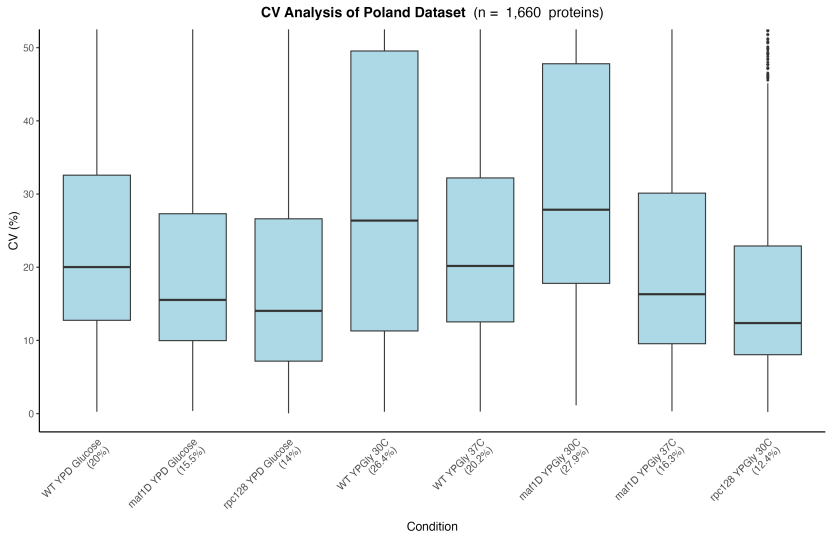

Figure 5: Distribution of Coefficient of Variation (CV) for the Poland Dataset. Box plots show CVs for 1,660 proteins across different strains and conditions. The median CV for each condition is noted in parentheses on the x-axis.

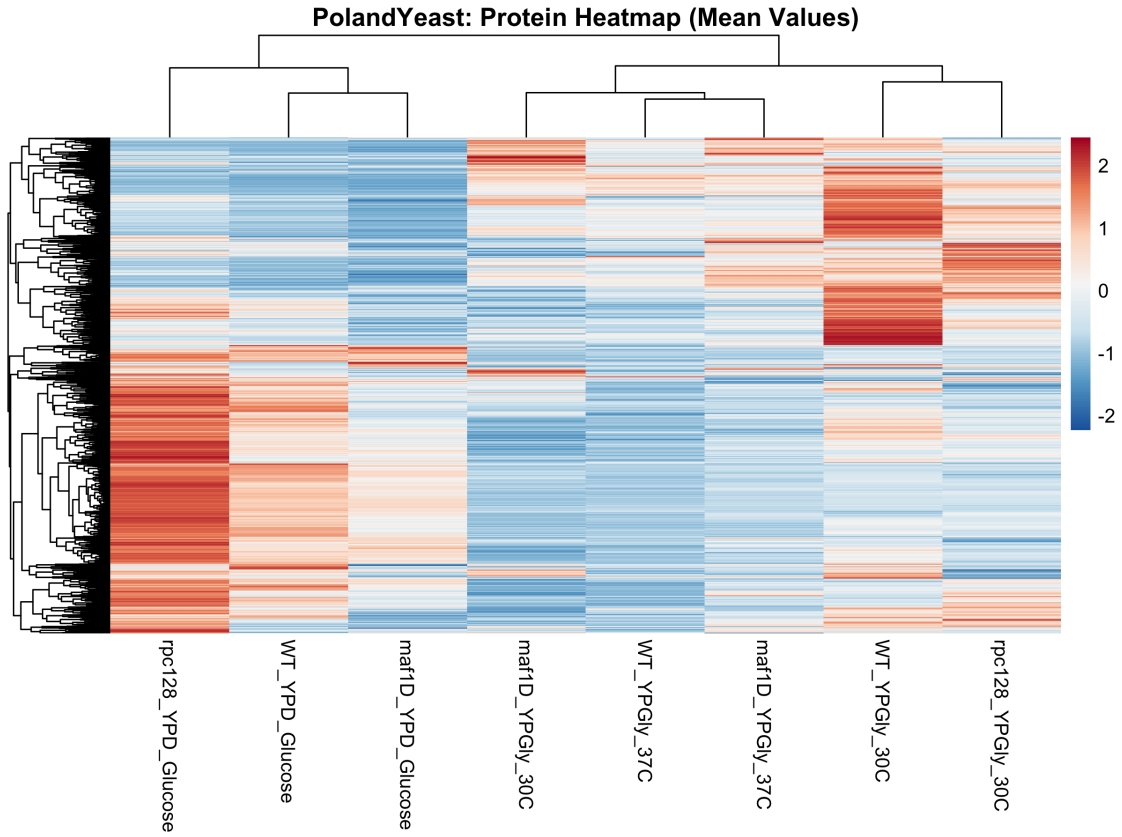

Figure 6: Hierarchical Clustering of Proteins (1654 proteins). Heatmap showing protein clusters across yeast strains and growth conditions. Colors show log2 fold change (red for high, blue for low). Samples group by protein profiles.

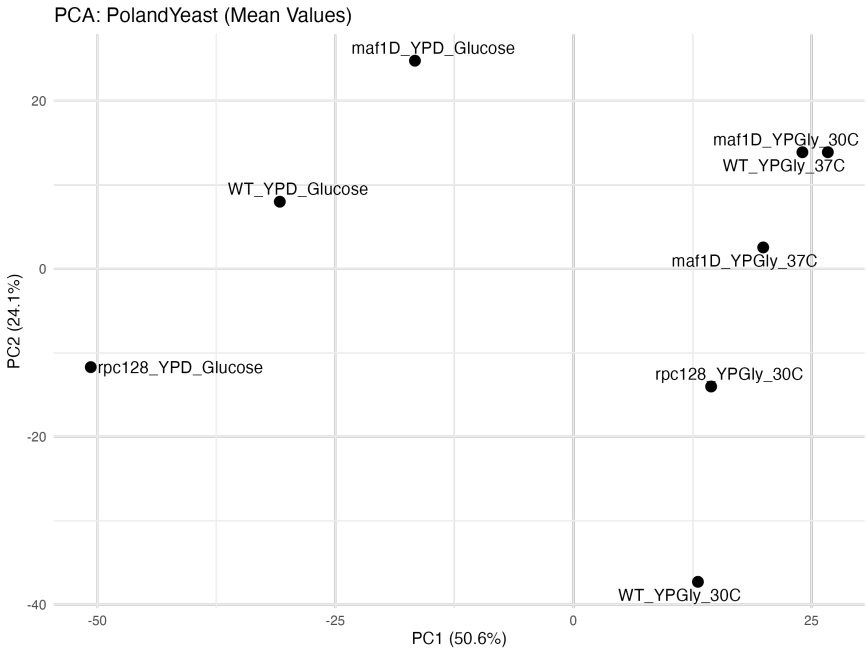

Figure 7: PCA of PolandYeast (Mean Values). This PCA plot for the expanded dataset shows samples clustering mostly by carbon source, then by temperature and mutant type. PC1 and PC2 account for the most variation.

Volcano plots (Figure 8) showed many protein changes in the mutant strains:

- When comparing maf1Δ to normal yeast in YPGly at 30°C (Figure 8A), 232 proteins were higher in maf1Δ, but only 5 were higher in normal yeast.
- On the other hand some comparisons (Figure 8B-C), 60 proteins were higher in maf1Δ have not revealed differncially abundant proteins.
- Other comparisons (maf1Δ vs WT in YPGly 37°C, rpc128-1007 vs WT in YPD Glucose, rpc128-1007 vs WT in YPGly 30°C) also showed distinct protein changes, as seen in Figure 8D-E.

Table 3 summarizes the number of changing proteins.

Table 3: Summary of Differentially Abundant Proteins

Comparison	**Upregulated in Cond. 1 (Count)**	**Upregulated in Cond. 2 (Count)**
W303 vs BY4742 (Exponential)	57 (W303)	174 (BY4742)
W303 vs BY4742 (Stationary)	218 (W303)	304 (BY4742)
WT YPGly 30C vs maf1Δ YPGly 30C	5 (WT)	232 (maf1Δ)
WT YPD vs maf1Δ YPD	4 (WT)	0 (maf1Δ)
WT YPGly 30C vs rpc128-1007 YPGly 30C	0 (WT)	8 (rpc128-1007)

3.3 Combined Proteomic Analysis

Protein data from two different experiments (named Poland Yeast and Table S1 [2]) were combined for a broader analysis.

3.3.1 Combined Analysis without Batch Correction

Before correction, the protein intensity distribution (Figure 9) showed clear differences between datasets, indicating technical variations. The initial combined heatmap also suggested these batch effects.

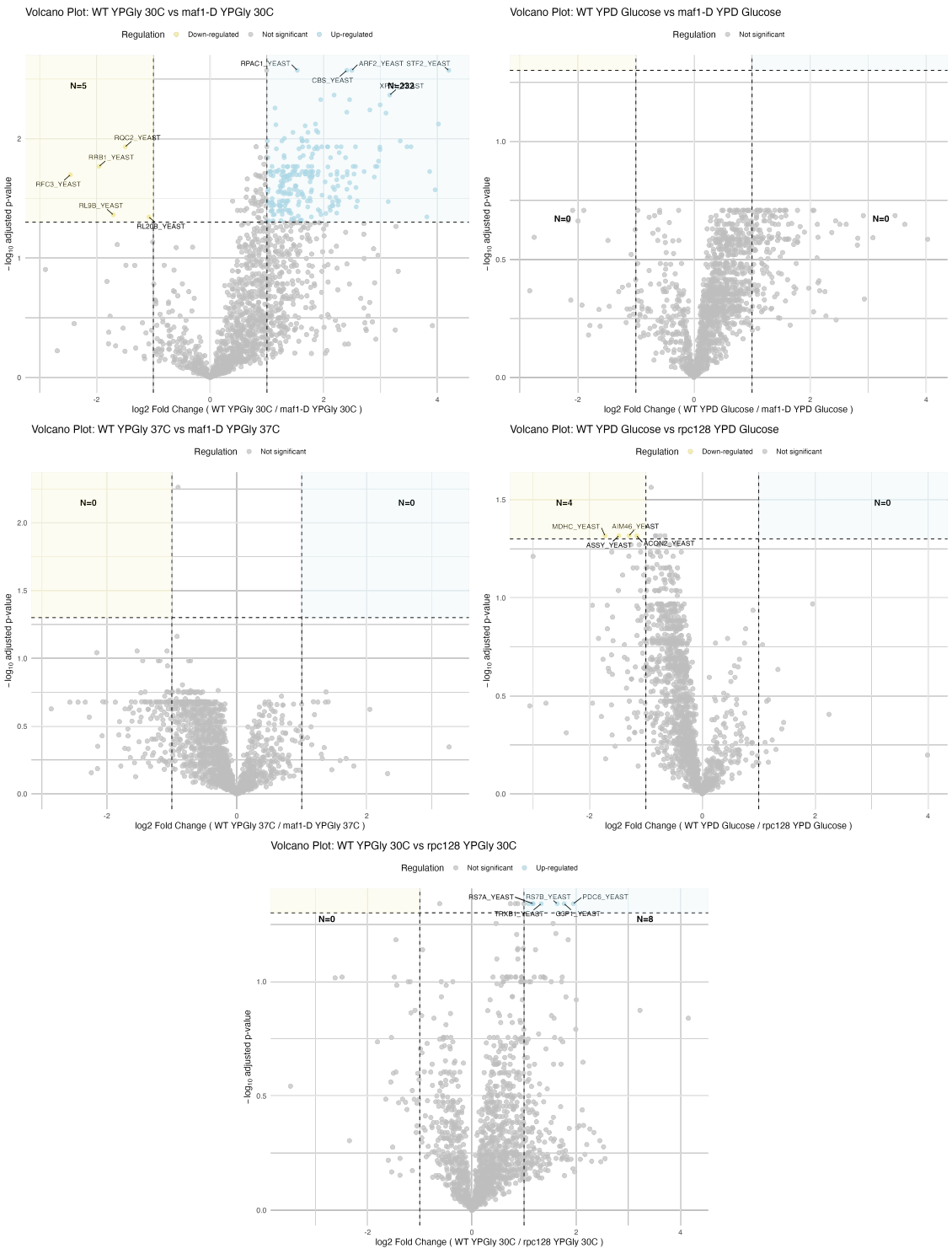

Figure 8: Volcano Plots of Differential Protein Abundance for Mutants. These plots show protein changes (log2 Fold Change vs. negative log10 p-value) for maf1Δ and rpc128-1007 mutants. (A) WT YPGly 30C vs. maf1-D YPGly 30C. (B) WT YPD Glucose vs. maf1-D YPD Glucose. (C) WT YPGly 37C vs. maf1-D YPGly 37C. (D) WT YPD Glucose vs. rpc128 YPD Glucose. (E) WT YPGly 30C vs. rpc128 YPGly 30C.

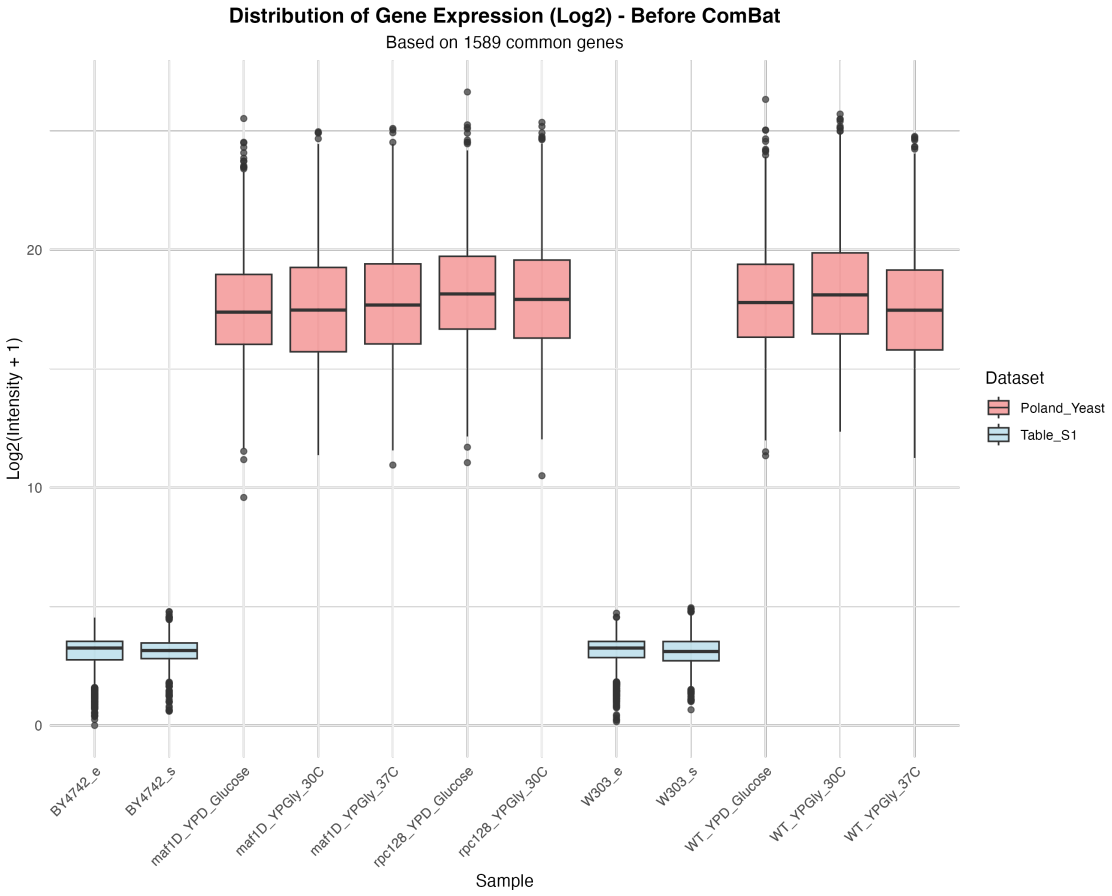

Figure 9: Distribution of Log2 Intensities Before Batch Correction. Box plots show protein intensity distributions across all samples before correction. Different intensity ranges across datasets highlight potential batch effects.

3.3.2 Combined Analysis with Combat Batch Correction

Due to the strong batch effect, Combat batch correction was used. This greatly improved the data. The PCA after Combat correction (Figure 10) now shows samples grouping based on biological factors like strain, mutant type, and growth conditions, revealing patterns that were previously hidden. Figure 11 also shows that protein intensities are now consistent across all samples after correction.
The combined heatmap after Combat correction (Figure 12) also shows clearer biological groups and protein patterns across all samples.

4 Conclusion

This comprehensive proteomic study successfully recreated the foundational comparative analysis of W303 and BY4742 yeast strains and significantly expanded it to include maf1Δ and rpc128-1007 mutants under diverse growth conditions. Given that maf1Δ mutants are known to exhibit diminished growth on non-fermentable carbon sources at 30°C, and YPGly conditions necessitate respiratory metabolism, the extensive proteomic changes observed in maf1Δ under these conditions likely represent a significant metabolic stress response. This response is an attempt by the cell to adapt to the non-fermentable carbon source in the absence of Maf1's critical regulatory function [3]. The identification of specific proteins, such as ILV3 and ARO10 (involved in amino acid biosynthesis), among the differentially expressed set, points to particular metabolic pathways that are perturbed or rewired in the maf1Δ mutant. A significant challenge in integrating multi-dataset proteomic analyses, as observed in this study, is the presence of batch effects. This overwhelming technical variation could completely mask any true biological differences if not

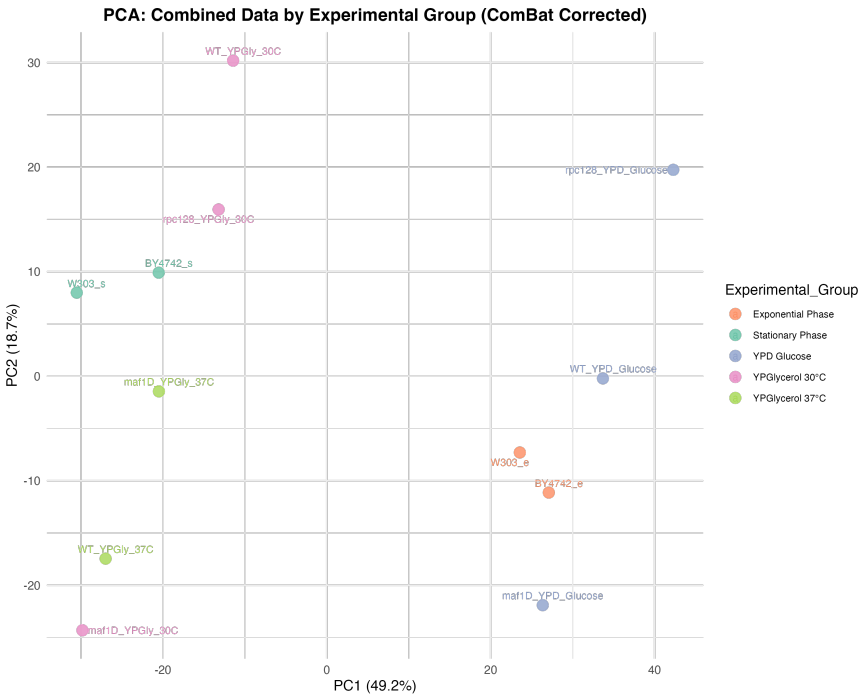

Figure 10: Combined PCA with Combat Correction. This PCA plot shows combined protein data after Combat correction. Samples now cluster by their experimental group (biological factors), showing that batch effects were successfully removed.

addressed. The application of Combat batch correction, a statistical method specifically designed to mitigate such systematic, non-biological variation, is therefore indispensable. The anticipated outcome of the Combat correction (Figures 10 and 12) is a dramatic shift in the PCA and heatmap, where biological factors such as strain, mutant status, and growth conditions become the primary drivers of sample clustering, rather than batch origin.

References

[1] W. Evan Johnson, Cheng Li, and Ariel Rabinovic. Adjusting batch effects in microarray expression data using empirical bayes methods. *Biostatistics*, 8(1):118–127, 04 2006.

[2] Valentina Rossio, Xinyue Liu, and Joao A. Paulo. Comparative proteomic analysis of two commonly used laboratory yeast strains: W303 and by4742. *Proteomes*, 11(4), 2023.

[3] Roza Szatkowska, Emil Furmanek, Andrzej M. Kierzek, Christian Ludwig, and Malgorzata Adamczyk. Mitochondrial metabolism in the spotlight: Maintaining balanced rnap iii activity ensures cellular homeostasis. *International Journal of Molecular Sciences*, 24(19), 2023.

References

[1] Google. Gemini (Version: 2.5 Pro). `https://gemini.google.com`. Large language model. 2025.

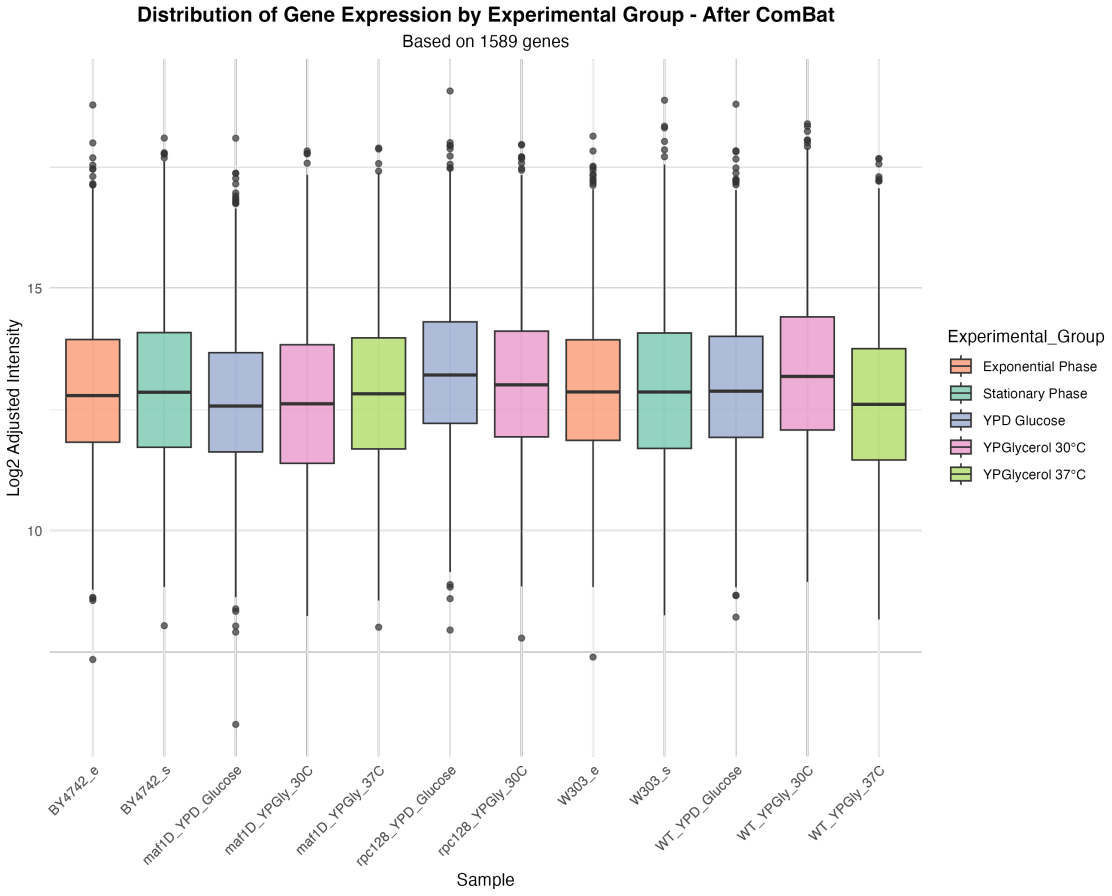

Figure 11: Distribution of Log2 Intensities After Combat Correction. Box plots show protein intensity distributions across all samples after Combat correction. The consistent intensity ranges confirm that batch effects were successfully reduced.

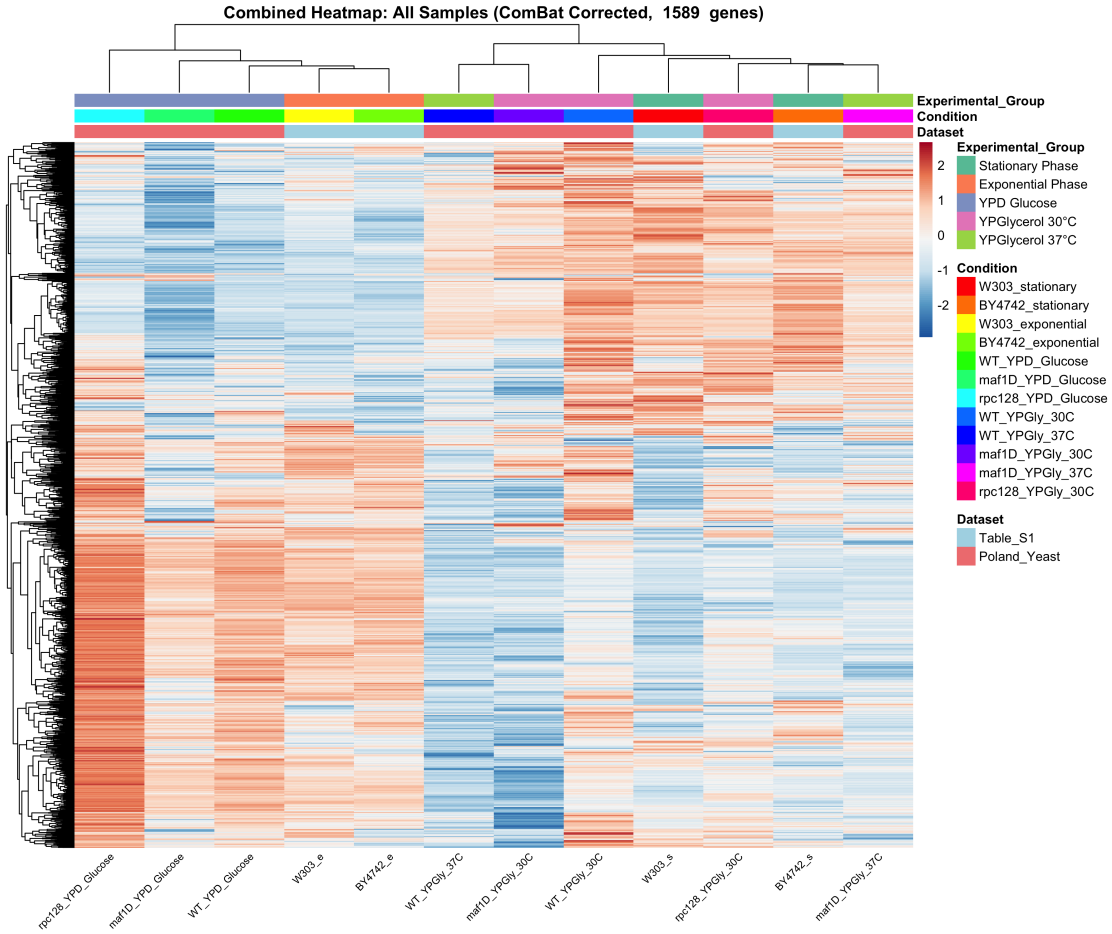

Figure 12: Combined Heatmap with Combat Correction. Heatmap of 1589 common proteins across all combined samples after Combat correction. Clearer biological clustering shows successful batch effect removal, allowing for more accurate comparisons.