UNDERSTANDING HOW THE RELATIVE ABUNDANCE OF CANDIDA SPECIES IMPACTS TRANSCRIPTIONAL REGULATION IN COCULTURE BIOFILMS

Diksha Kool

Follow this and additional works at: https://scholarworks.sjsu.edu/etd_projects

Part of the Bioinformatics Commons

Recommended Citation
https://scholarworks.sjsu.edu/etd_projects/1093

This Master's Project is brought to you for free and open access by the Master's Theses and Graduate Research at SJSU ScholarWorks. It has been accepted for inclusion in Master's Projects by an authorized administrator of SJSU ScholarWorks. For more information, please contact scholarworks@sjsu.edu.
UNDERSTANDING HOW THE RELATIVE ABUNDANCE OF CANDIDA SPECIES IMPACTS TRANSCRIPTIONAL REGULATION IN COCULTURE BIOFILMS

A Project

Presented to

The Faculty of the Department of Computer Science
San José State University

In Partial Fulfillment
Of the Requirements for the
Degree Master of Science in Bioinformatics

By
Diksha Kool
May 2022
Candida albicans and Candida glabrata are common fungal species that can change from commensal to pathogen due to their ability to form robust biofilms. Candida species are the leading cause of life-threatening conditions like Candidemia, and the existing treatments for biofilm-related infections are suboptimal. Research shows that the relative abundance of the two Candida species promotes biofilm formation, enhances pathogenicity, and increases antibiotic resistance. Thus, focusing on the importance of coculture, this paper utilizes RNA sequencing to investigate the gene expression leading to biofilm development in coculture through a time-series study.

Index Terms—RNA-Seq, Biofilm, Candida Coculture, Antibiotic Resistance, Candidemia
# Table of Contents

1 Introduction .............................................................................................................................. 5
   1.1 Biological Background ............................................................................................... 5
   1.2 Technical Background .............................................................................................. 8

2 Methods ................................................................................................................................... 9
   2.1 Overview of the RNA-Seq Pipeline ................................................................................. 9
   2.2 Upstream Analysis Steps and Tools Used ..................................................................... 10
      2.2.1 Quality Check using FASTQC and Trimming using Trimmomatic ..................... 10
      2.2.2 Align Reads to the Reference Genome using STAR aligner ................................. 11
      2.2.3 Alignment Quality Check and Removal of Duplicates using Picard ..................... 12
      2.2.4 Expression Quantification using htsseq-count..................................................... 13
   2.3 Downstream Analysis Steps and Tools Used ............................................................... 13
      2.3.1 Differential Gene Expression (DGE) Analysis using DESeq2 .............................. 13
      2.3.2 Pairwise Comparison Analysis using Wald Test .................................................. 14
      2.3.3 Time-series Analysis using Likelihood Ratio Test (LRT)....................................... 16
      2.3.4 Gene Set Enrichment Analysis ............................................................................ 17

3 Results .................................................................................................................................... 19
   3.1 Upstream Analysis Results .......................................................................................... 19
      3.1.1 Quality Check and Trimming .............................................................................. 19
      3.1.2 Align Reads to the Reference Genome ................................................................. 20
      3.1.3 Alignment Quality Check and Removal of Duplicates ........................................... 22
   3.2 Downstream Analysis Results ...................................................................................... 24
      3.2.1 Differential Gene Expression Analysis (DGE) ...................................................... 24
      3.2.2 Pairwise Comparison Analysis ............................................................................ 25

4 Discussion .............................................................................................................................. 34
   4.1 Significant Ca Genes and Biological Pathways for Ca:Cg Cocultures Compared to Ca Monoculture ................................................................. 35
   4.2 Significant Cg Genes and Biological Pathways for Ca:Cg Cocultures Compared to Cg Monoculture ................................................................. 38
   4.3 Significant Ca and Cg Genes and Biological Pathways for Ca:Cg Coculture Comparisons ......................................................................................... 40
   4.4 Significant Time Influenced Ca and Cg Genes and Biological Pathways for Time Series Analysis ........................................................................... 41

References ...................................................................................................................................... 43
LISTS OF FIGURES

Figure 1. Life cycle of C. albicans biofilm, adapted from [7] .......................................................... 6
Figure 2. High level overview of RNA-Seq pipeline ........................................................................ 9
Figure 3. Different experimental conditions schematic for (A) pairwise comparison analysis cases and (B) time-course analysis cases .......................................................... 14
Figure 4. Models of LRT .................................................................................................................. 16
Figure 5. MultiQC generated FastQC mean quality score plots for (A) reads before trimming and (B) reads after trimming .......................................................... 19
Figure 6. MultiQC generated FastQC adaptor content plots for (A) reads before trimming and (B) reads after trimming ................................................................................. 20
Figure 7. MultiQC generated STAR mapping plots for (A) Ca aligned samples and .......... 21
Figure 8. Top BLAST hits of the unmapped too short reads generated using STAR ................. 22
Figure 9. Plot showing PCR read duplicates in STAR generated bam files for (A) Ca and (B) Cg samples ...................................................................................................................... 23
Figure 10. PCA for all Ca samples and all coculture samples at three time points .......... 24
Figure 11. GSEA enrichment map showing enriched GO biological processes for coculture Ca: Cg ratio 1:3 vs. Ca monoculture at 24h ........................................................................ 29
Figure 12. GSEA heat map showing significant genes for the enriched GO biological processes along with their expression levels for case1- coculture Ca: Cg ratio 1:3 vs. Ca monoculture at 24h .......................................................... 30
Figure 13. GSEA enrichment plot showing filamentous growth linked leading edge genes and ES for Ca:Cg 1:3 vs. Ca monoculture at 24h .............................................................................. 30
Figure 14. GSEA enrichment map showing enriched GO biological processes for case4- time-series analysis for cocultures vs Ca monoculture ........................................................................... 32
Figure 15. GSEA heat map showing significant genes for the enriched GO biological processes along with their expression levels for case4- time-series analysis for cocultures vs Ca monoculture .......................................................................................................................... 33
Figure 16. GSEA enrichment plot showing filamentous growth linked leading edge genes and ES for case4- time-series analysis ............................................................................................. 33
Figure 17. GSEA enrichment map showing enriched GO biological processes for case1- Ca:Cg 1:3 vs. Ca monoculture at (A) 6h and (B) 12h ........................................................................ 36
Figure 18. GSEA heat map showing significant genes for the enriched GO biological processes along with their expression levels for Case1- Ca:Cg 1:3 vs. Ca monoculture at (A) 6h and (B) 12h .......................................................................................................................... 37
Figure 19. GSEA heat map showing significant genes for the enriched GO biological processes along with their expression levels for case2- cocultures vs Cg monoculture at 24h for (A) Ca:Cg 3:1, (B) Ca:Cg 1:1, and (C) Ca:Cg 1:3 ......................................................................................................................... 39
Figure 20. GSEA enrichment map showing upregulated GO biological processes for case3- Ca:Cg 1:3 vs. Ca:Cg 3:1 at 24h ..................................................................................................................... 40
Figure 21. GSEA heat map showing significant genes for the enriched GO biological processes along with their expression levels for case3- Ca:Cg 1:3 vs. Ca:Cg 3:1 at 24h ........................................................................ 41
Figure 22. GSEA generated heat map for time-series analysis for cell adhesion signal .......... 42
LISTS OF TABLES

Table 1. Description of all Pairwise comparison cases ......................................................................15
Table 2. Description of the comparisons for all cocultures with Ca and Cg monocultures across all time-points .......................................................................................................................17
Table 3. Summary of significant genes generated by DESeq2 for case1- cocultures vs. Ca monoculture ...........................................................................................................................................25
Table 4. Summary of significant genes generated by DESeq2 for case2- cocultures vs. Cg monoculture ...........................................................................................................................................26
Table 5. Summary of significant genes generated by DESeq2 for case3- coculture vs. coculture 27
Table 6. Summary of significant genes generated by DESeq2 for case4- cocultures vs Ca monoculture and cocultures vs Cg monoculture ..........................................................................................28
Table 7. Top 5 biological process with (+ve ES) for case4- time-series analysis ...............................31
Table 8. Top 5 biological process with (-ve ES) for case4- time-series analysis .................................32
1 INTRODUCTION

1.1 Biological Background

The human microflora consists of a wide variety of microorganisms, and the interactions occurring among the host cells and the normal flora can be advantageous or harmful to the human host [1]. Candida is a common fungal microbe that is a part of the human normal microbial flora, and it often colonizes the gastrointestinal (GI) tract, oral cavity, and skin of healthy humans. However, the fungus can opportunistically become virulent if the host is immuno-compromised [2], [3]. Candida albicans (Ca) and Candida glabrata (Cg) are the widely known opportunistic fungal pathogens causing superficial or systemic infections [4], [5]. While the superficial infections mostly affect the mucous membranes and are treatable, Candida systemic infections can affect the blood, brain, heart, bones, or other parts of the body and often leads to life-threatening conditions like Invasive Candidiasis and Candidemia [4]. Invasive Candidiasis is a serious Candida infection where Candida species spreads into the blood stream or internal organs. Candidemia is an extremely frequent form of Invasive Candidiasis resulting in Candida bloodstream infections with a mortality rate of 40% and is pushing 60% in some studies [6], [7].

A major factor linked with Candida pathogenesis and increased virulence resulting in fatal diseases is its ability to form biofilms in the polymicrobial environment [1]. The biofilm cycle in Candida starts with the rounded yeast cell adherence to a stable surface, the next step involves cell growth and filament formation to form a basal layer attached to the solid surface, and this results in the initiation of the biofilm formation [8].
The initiation step is followed by the biofilm maturation step during which an extracellular polymeric substance (EPS) is produced by the organism, and there is formation of hyphal and pseudohyphal cells and all the cells are covered in an extracellular matrix, providing the biofilm a rigid structure. The final stage of biofilm formation is called the dispersal phase, where a few rounded yeast cells scatter from the biofilm site to plant to new locations. The life cycle of Ca biofilm is shown in Figure 1. It also shows the polymorphic nature of Ca and its capability to grow as yeast, pseudohyphae, and hyphae. In contrast, Cg primarily develops as budding yeast and has demonstrated its attachment onto the hyphae of Ca for tissue invasion which triggers an infection [9]. Ca is considered as the most common cause of most of the yeast infections because of its polymorphic nature, on the other hand, Cg is less pathogenic but often known for its resistant to the antifungals.

Figure 1. Life cycle of C. albicans biofilm, adapted from [7]

These biofilms are known to protect the fungus from host immune defense and antifungal drugs, thus making the existing treatments for life-threatening Candida
infections limited and suboptimal [1], [10]. Biofilms also pose a threat as they are observed in medical equipments, such as urinary catheters and cardiac devices, making Candida a leading cause of nosocomial infections [4], [11]. The Candida biofilms also provide antimicrobial resistance to other disease-causing bacteria like Staphylococcus aureus [12].

It has been previously demonstrated that Ca and Cg are frequently isolated collectively from the sites of infection and the co-isolation is linked with increased pathogenesis, indicating the significance of Candida coculture biofilms [1], [13]. A clinical study on Ca and Cg co-infection of Oral Candidiasis revealed that the co-isolation of Candida species is associated with enhanced pathogenesis as the incidence of co-isolation was identified in about 80% of patients with serious Candida infection [14]. This signaled for a possible synergistic interaction between Ca and Cg in pathogenesis. In the laboratory of Dr. Katy Kao at San José State University, Candida coculture biofilms were studied, and it was reported that biofilm development of the Candida coculture depends on the relative ratio of starting Ca and Cg cell concentrations [1]. Results of the research indicated that Cg in monocultures produces a weaker biofilm than Ca and coculture biofilms of Ca and Cg have shown a higher biomass than monoculture biofilms. This study demonstrated a ~6.5-fold boost in biofilm biomass when Ca:Cg ratio was 1:3, and a ~2.5-fold increase with Ca:Cg ratio of 1:1, observed relative to Ca monoculture. This suggested interspecies interaction between the two species in cocultures, specifically when the relative ratio of Ca to Cg was 1:3 which greatly increased the extent of biofilm formation. With more biofilm biomass, the two Candida species in coculture are known to enhance pathogenesis and antifungal resistance compared to monoculture biofilms and
other relative ratios. This indicates the importance of *Candida* coculture and despite the collaborative relationship between the two *Candida* species, there is limited information to explain the molecular basis of interspecies interaction. Thus, there is clinical relevance to utilize RNA sequencing (RNA-Seq) to identify the gene expression leading to biofilm development in coculture across different time frames. In this work, we have established an RNA-Seq pipeline which starts from the raw reads and ends with several differentially expressed genes and significant pathways in between two experimental conditions. This pipeline is used to study the dynamic transcriptomic changes in Ca and Cg cocultures when compared to Ca and Cg monocultures. Finally, this pipeline can also be used to identify significant Gene Ontology (GO) biological processes with enriched genes that can help us understand the underlying genetic basis of Ca and Cg interspecies interactions.

### 1.2 Technical Background

The sequencing data was generated in the laboratory of Dr. Katy Kao at San José State University, Department of Chemical and Materials Engineering. Twelve hours before experiments, single colonies of Ca and Cg strains were isolated and cultured in a growth medium and were grown overnight at 30°C [1], [15]. Biofilms were grown on coverslips and RNA was obtained from the samples by employing the protocol mentioned on the Illustra RNAspin mini-isolation kit. The quality of RNA was determined using a Nanodrop, and its concentration was estimated using Qubit. TruSeqRNA kit was used for RNA library preparation, and single-end sequencing was accomplished using an Illumina HiSeq 2500 v4 High Output instrument. This RNA sequencing project includes a total of 30 samples collected at five different biofilm
conditions – i. Ca monoculture, ii. Cg monoculture, iii. Ca:Cg 3:1, iv. Ca: Cg 1:1, and v. Ca: Cg 1:3. All the samples were collected at three-time points – i. 6h, ii. 12h, and iii. 24h with two biological replicates. Detailed information about sample names, and their conditions can be found in the appendix.

2 Methods

2.1 Overview of the RNA-Seq Pipeline

This RNA-Seq pipeline analyzes the transcriptome at the gene level during the course of biofilm formation in mono and cocultures, it starts from the raw reads, and ends with a number of differentially expressed genes and pathways. An overview of the RNA-Seq pipeline workflow of this project is shown in Figure 2.

![Figure 2. High level overview of RNA-Seq pipeline](image)

Black arrows indicate the workflow, blue arrows specify the input files, and green arrows specify the output files. Bold text in the boxes indicate different tools used with the specific function of each tool mentioned below the box.
The RNA-Seq data analysis workflow consists of upstream and downstream analysis. The upstream analysis includes raw reads quality check, quality trimming, aligning reads to the reference genome, removal of duplicates, and mapped transcript quantification while the downstream analysis consists of Differential Gene Expression (DGE) analysis, and Gene Set Enrichment Analysis (GSEA).

After obtaining the raw read fastq files, quality control for raw reads was performed using FastQC [16]. Low-quality bases with Phred quality score of less than 33 and adaptors were removed using Trimmomatic [17]. Trimmed reads were aligned to Ca reference genome and the unmapped reads, and six Cg monocultures were mapped to Cg reference genome using Spliced Transcripts Alignment to a Reference (STAR) alignment tool [18]. The quality of the alignment was checked using the Broad Institute Picard CollectAlignmentSummaryMetrics tool and the duplicates were removed using Picard MarkDuplicates [19]. The aligned reads with duplicates removed were utilized by htseq-count for expression quantification [20]. The Bioconductor DESeq2 package [21] was used for differential expression analysis and to analyze the upregulated and downregulated genes. The last step of the pipeline was enrichment analysis using GSEA which provided information about the enriched pathways and significant genes [22].

2.2 Upstream Analysis Steps and Tools Used

2.2.1 Quality Check using FASTQC and Trimming using Trimmomatic

FastQC takes fastq files as the input and provides an instant view on the quality of the raw reads on various metrics ranging from the sequence quality, GC content, and adapter contamination in the form of html files. FastQC [16] version 0.11.9 was used
twice to generate quality metrics; the first quality check was performed on the raw reads and the second FastQC check was performed on the trimmed reads. To aggregate the output of FastQC into one report, MultiQC [23] was used on the FastQC output files. Detailed FastQC and MultiQC results of the raw and trimmed reads are available in the appendix. Trimmomatic is a flexible read trimming tool that works on raw fastq files, and it removes any low-quality bases and trims adaptors from the raw reads. In Trimmomatic [17] version 0.39, TruSeq3-SE Illumina adapter sequences were used, and selected parameters were MINLEN (20) which drops reads below a specified length, LEADING (3) to cut off bases from the start, TRAILING (3) to cut off bases from the end. Detailed Trimmomatic results for all the samples, including the input reads, output reads, dropped reads and their percentage, and percentage of reads remaining for the raw and trimmed sample files are available in the appendix.

2.2.2 *Align Reads to the Reference Genome using STAR aligner*

STAR is a fast read aligner and it also provides support for splice-junction and fusion read detection. STAR is shown to have high precision and surpasses other aligners by more than a factor of 50 in aligning speed, but it is memory demanding [18]. It achieves this highly efficient mapping by performing seed searching which focuses on finding the Maximal Mappable Prefix (MMP) hits between reads and the genome. It performs local alignment, and if a good alignment is not found, it soft clips the ends of reads with high mismatches. The workflow for the STAR aligner consists of two steps – generating indexes files and the actual mapping step. To generate index files, reference genome sequence (FASTA file), and annotations file (GTF file) were provided as an
input to the aligner. For the alignment step, STAR generated index files and trimmed fastq files are supplied as an input.

STAR version 2.7.5 was used to align the trimmed reads to the Ca reference genome. The trimmed reads of Ca monocultures, Ca:Cg 3:1, Ca:Cg 1:1, Ca:Cg 1:3, and Cg monocultures were mapped to Ca reference genome and --outReadsUnmapped Fastx parameter of STAR was used to generate unmapped reads files for all the samples. Next, the unmapped reads for all the samples from the previous alignment were mapped to Cg reference genome using STAR. The aligner creates several output files - alignments files such as Sequence Alignment Map (SAM) or Binary Alignment Map (BAM), alignment summary statistics including uniquely mapped reads, reads mapped to multiple loci, and unmapped too short reads (Log files), and splice junctions (Tab files). As the alignment result indicated unmapped too short reads, these reads were also examined for any possible source of contamination using the Basic Local Alignment Search Tool (BLAST) [24]. MultiQC was performed on all the STAR alignment metrics files to generate a summary report for the trimmed reads of all the samples.

2.2.3 Alignment Quality Check and Removal of Duplicates using Picard

Picard is a java application and consists of various metrics that can provide valuable information about the quality of the mapped reads. CollectAlignmentSummaryMetrics of Picard [19] version 2.21.8 was used to obtain a detailed summary of STAR alignment from the BAM files. It provides statistics to evaluate the alignment, such as the number of inputs reads and the percentage of reads that are mapped to the reference. Duplicates can be present in the sample files as a result of library preparation step performed before sequencing using PCR. Duplicate reads can
also be produced during the actual sequencing process. Picard MarkDuplicates was used to locate and remove all the duplicate reads in the STAR aligner generated BAM file.

2.2.4 Expression Quantification using htseq-count

After the reads are mapped to both reference genomes, the subsequent step was to count the number of reads aligned to each gene. From the HTSeq tool suite, htseq-count uses the mapped reads in SAM/BAM format and the GTF file as its input and counts the number of reads mapped to certain genomic features such as transcripts or genes. The htseq-count [20] version 0.11.3 was utilized for expression quantification and the count files that were generated as its output were used for DGE analysis. To compare all cocultures at same time points, the reads count for each coculture samples were concatenated.

2.3 Downstream Analysis Steps and Tools Used

2.3.1 Differential Gene Expression (DGE) Analysis using DESeq2

To analyze the differences in abundance of genes within a transcriptome between two experimental conditions, DGE analysis was performed using the Bioconductor package DESeq2 version 1.28.1 in R [21]. The main steps of the analysis were making DESeq data set object, size factor and gene dispersion estimation, hypothesis testing, creating DEG table and data visualization. Wald test and Likelihood Ratio Test (LRT) were the two different hypothesis tests performed for DGE analysis. The Wald test is like a t-test, and it calculates the p-value which is the probability of each gene to be significantly differentially expressed in one of the sample groups compared to a null hypothesis which assumes no differential expression across the two given sample groups.
This hypothesis test is commonly used for comparing two groups which is referred to as pairwise comparison.

The other hypothesis test which is the LRT helps in finding significant genes that react in a condition-specific manner over time, compared to a group of baseline samples. It’s an alternative when evaluating expression change across more than two groups which includes Ca, Cg monoculture, cocultures and three timepoints. To analyze the significantly perturbed genes in between different conditions, four different analyses will be considered for DGE which are described in Figure 3.

2.3.2 Pairwise Comparison Analysis using Wald Test

To investigate the significant genes during the biofilm growth, three different pairwise comparison cases were studied - Case1: compare all cocultures with Ca monocultures at same time points, Case2: compare all cocultures with Cg monocultures at same time points, and Case3: compare all cocultures among each other at same time.
points. As the test is used to compare two groups, conditions (monoculture and coculture) and time points were combined into a factor in R. The two conditions for pairwise comparisons for each case is summarized in Table 1.

Table 1. Description of all pairwise comparison cases

<table>
<thead>
<tr>
<th>Case</th>
<th>Condition 1</th>
<th>Condition 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1a (Coculture 3:1 vs. Ca Monoculture)</td>
<td>Ca:Cg 3:1_6h</td>
<td>Ca Mono_6h</td>
</tr>
<tr>
<td></td>
<td>Ca:Cg 3:1_12h</td>
<td>Ca Mono_12h</td>
</tr>
<tr>
<td></td>
<td>Ca:Cg 3:1_24h</td>
<td>Ca Mono_24h</td>
</tr>
<tr>
<td>1b (Coculture 1:1 vs. Ca Monoculture)</td>
<td>Ca:Cg 1:1_6h</td>
<td>Ca Mono_6h</td>
</tr>
<tr>
<td></td>
<td>Ca:Cg 1:1_12h</td>
<td>Ca Mono_12h</td>
</tr>
<tr>
<td></td>
<td>Ca:Cg 1:1_24h</td>
<td>Ca Mono_24h</td>
</tr>
<tr>
<td>1c (Coculture 1:3 vs. Ca Monoculture)</td>
<td>Ca:Cg 1:3_6h</td>
<td>Ca Mono_6h</td>
</tr>
<tr>
<td></td>
<td>Ca:Cg 1:3_12h</td>
<td>Ca Mono_12h</td>
</tr>
<tr>
<td></td>
<td>Ca:Cg 1:3_24h</td>
<td>Ca Mono_24h</td>
</tr>
<tr>
<td>2a (Coculture 3:1 vs. Cg Monoculture)</td>
<td>Ca:Cg 3:1_6h</td>
<td>Cg Mono_6h</td>
</tr>
<tr>
<td></td>
<td>Ca:Cg 3:1_12h</td>
<td>Cg Mono_12h</td>
</tr>
<tr>
<td></td>
<td>Ca:Cg 3:1_24h</td>
<td>Cg Mono_24h</td>
</tr>
<tr>
<td>2b (Coculture 1:1 vs. Cg Monoculture)</td>
<td>Ca:Cg 1:1_6h</td>
<td>Cg Mono_6h</td>
</tr>
<tr>
<td></td>
<td>Ca:Cg 1:1_12h</td>
<td>Cg Mono_12h</td>
</tr>
<tr>
<td></td>
<td>Ca:Cg 1:1_24h</td>
<td>Cg Mono_24h</td>
</tr>
<tr>
<td>2c (Coculture 1:3 vs. Cg Monoculture)</td>
<td>Ca:Cg 1:3_6h</td>
<td>Cg Mono_6h</td>
</tr>
<tr>
<td></td>
<td>Ca:Cg 1:3_12h</td>
<td>Cg Mono_12h</td>
</tr>
<tr>
<td></td>
<td>Ca:Cg 1:3_24h</td>
<td>Cg Mono_24h</td>
</tr>
<tr>
<td>3a (Coculture 1:3 vs. Coculture 3:1)</td>
<td>Ca:Cg 1:3_6h</td>
<td>Ca:Cg 3:1_6h</td>
</tr>
<tr>
<td></td>
<td>Ca:Cg 1:3_12h</td>
<td>Ca:Cg 3:1_12h</td>
</tr>
<tr>
<td></td>
<td>Ca:Cg 1:3_24h</td>
<td>Ca:Cg 3:1_24h</td>
</tr>
<tr>
<td>3b (Coculture 1:3 vs. Coculture 1:1)</td>
<td>Ca:Cg 1:3_6h</td>
<td>Ca:Cg 1:1_6h</td>
</tr>
<tr>
<td></td>
<td>Ca:Cg 1:3_12h</td>
<td>Ca:Cg 1:1_12h</td>
</tr>
<tr>
<td></td>
<td>Ca:Cg 1:3_24h</td>
<td>Ca:Cg 1:1_24h</td>
</tr>
<tr>
<td>3c (Coculture 3:1 vs. Coculture 1:1)</td>
<td>Ca:Cg 3:1_6h</td>
<td>Ca:Cg 1:1_6h</td>
</tr>
<tr>
<td></td>
<td>Ca:Cg 3:1_12h</td>
<td>Ca:Cg 1:1_12h</td>
</tr>
<tr>
<td></td>
<td>Ca:Cg 3:1_24h</td>
<td>Ca:Cg 1:1_24h</td>
</tr>
</tbody>
</table>
2.3.3 Time-series Analysis using Likelihood Ratio Test (LRT)

LRT analysis all the factors at once and are especially designed for time course analysis to reflect the dynamic nature of the significant ally expressed genes. To perform the LRT test, two models are required. It compares the models to generate a set of differentially expressed genes that are perturbed in a condition-specific manner over time compared to a group of baseline samples. The first model is referred as a full model and it comprises of all the experimental conditions as well as the interaction condition (all cocultures vs. monoculture interaction with time) while the second model which is called a reduced model consists only experimental conditions (coculture vs. monoculture). The conditions for both models are described in Figure 4.

\[
\text{full\_model} \leftarrow \sim \text{condition} + \text{time} + \text{condition:time} \\
\text{reduced\_model} \leftarrow \sim \text{condition} + \text{time}
\]

*Figure 4. Models of LRT*

In order to identify significant genes between the conditions across the time points, the two models are compared and LRT tests whether the increase in the log likelihood from the additional coefficients would be expected if those coefficients were equal to zero. If the adjusted p-value is small, the additional coefficient in full and not in reduced model increases the log likelihood more than that would be estimated if their actual value was zero. The LRT was performed twice with Ca and Cg monoculture samples as the baseline sample and were compared to all the coculture samples. The different tests that are computed as a part of LRT are summarized in the Table 2.
Table 2. Description of the comparisons for all cocultures with Ca and Cg monocultures across all time-points

<table>
<thead>
<tr>
<th>Tests comparing conditions</th>
<th>Coculture</th>
<th>Monoculture</th>
</tr>
</thead>
<tbody>
<tr>
<td>3:1</td>
<td></td>
<td>Ca and Cg</td>
</tr>
<tr>
<td>1:1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1:3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tests comparing all time periods after T1 = 6h</td>
<td>T2 VS. T1</td>
<td>T3 VS. T1</td>
</tr>
<tr>
<td>12h VS. 6h</td>
<td>24h VS. 6h</td>
<td></td>
</tr>
<tr>
<td>Tests for Interaction</td>
<td>Ca:Cg 3:1, Ca:Cg 1:1, Ca:Cg 1:3 at T2 = 12h</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ca:Cg 3:1, Ca:Cg 1:1, Ca:Cg 1:3 at T3 = 24h</td>
<td></td>
</tr>
</tbody>
</table>

Pairwise comparison and LRT both generated a table of significant genes, with p-values, p-adj values, and log2 fold change, log fold change standard error, base mean, and some related statistics. A threshold of a p-adj value of 0.01 was applied to extract genes which were not statistically meaningful. The notable genes were ordered by p-adj value in ascending order and filtered into a .csv file for all the cases. These were used as input to for GSEA.

2.3.4 Gene Set Enrichment Analysis

GSEA is a computational technique that establishes whether an a priori defined group of genes are statistically significant and shows agreeable differences between two biological states [22]. It is a desktop application that analyzes ranked gene lists using a permutation-based test. Two types of enrichment analysis were performed – GSEAPreranked for Wald test (Case1 – Case3) and GSEA for LRT (Case4). Preparing
input files for enrichment analysis is an important step and for GSEAPreranked, two
input files were prepared - Gene Matrix Transposed file format (GMT) file that describes
gene sets and Ranked list file format (RNK) file that contains a single, rank ordered genes
arranged in ascending order based upon log2fold change metric. *Candida* Genome
Database [25] GO Term Finder and GO Slim Mapper were used to generate statistically
significant Gene Ontology and for broad categorization of gene sets. In case of the LRT,
GSEA was used, and apart from GMT, two other files were prepared - Gene Cluster Text
file format (GCT) file that included the normalized gene counts, and Continuous Label
format (CLS) file with the phenotypic information for each sample and was used to
analyze a time series experiments. Ranking of the genes was done using the Pearson
Correlation as the ranking metric. GSEA examines each gene in the ranked list and
increments a running-sum when a ranked gene fit into the gene set in the GMT file and
decrements it when the gene is not a part of the gene set. The Enrichment Score (ES) is
 calculated by GSEA to identify significant genes and pathways and its significance is
 estimated using a permutation test.

Both GSEA and GSEAPreranked generated a list of top positively (upregulated)
and negatively (downregulated) enriched genes and the significant GO biological
processes. These genes are referred to as the leading-edge gene as these genes contribute
the most to the specific gene set’s enrichment score. GSEA was also used to generate
heat maps, and enrichment plots which provides a graphical view of the ES which is the
peak of the plot for a specific gene set. It indicates the degree to which the ranked genes
which are also a present in a gene set are overrepresented at the top or bottom of the
complete ranked list of genes. An enriched gene set shows leading edge genes at the top
or bottom of the ranked list. Cytoscape was used to visualize the results from GSEA pathway enrichment analysis and for generating enrichment maps [26].

3 RESULTS

3.1 Upstream Analysis Results

3.1.1 Quality Check and Trimming

Two FASTQC check were performed – the first check was performed on the raw reads and the second was performed on the trimmed reads. The mean quality scores generated using MultiQC shown in Figure 5(A) indicates that the raw reads have an overall good quality score. Following the trimming step, the quality improved slightly towards the end of the reads as shown in Figure 5(B).

![Figure 5. MultiQC generated FastQC mean quality score plots for (A) reads before trimming and (B) reads after trimming](image-url)
The adaptor content plot generated using MultiQC shown in Figure 6(A) indicates the presence of Illumina Universal adapters which were trimmed using Trimmomatic as shown in Figure 6(B). A small amount of residue adapter sequences that were not trimmed were expected to be soft clipped during the read alignment process. Less than 1% of the reads were dropped following trimming. Detailed FastQC and Trimmomatic results are available in the appendix.

![MultiQC generated FastQC adaptor content plots](image)

*Figure 6. MultiQC generated FastQC adaptor content plots for (A) reads before trimming and (B) reads after trimming*

### 3.1.2 Align Reads to the Reference Genome

The trimmed reads of Ca monoculture and coculture samples were mapped to Ca and Cg reference genomes using the STAR aligner. MultiQC summarized the mapping percentage results for Ca and Cg samples as shown in Figure. 7(A) and (B), respectively.
Understanding How the Relative Abundance of *Candida* Species Impacts Transcriptional Regulation in Coculture Biofilms

Figure 7. MultiQC generated STAR mapping plots for (A) Ca aligned samples and (B) Cg aligned samples
The percentage of uniquely mapped reads varies for most of the samples. There are also a high percentage of unmapped too short reads that were left due the short read length resulting from the trimming process or possible contamination. The BLAST results indicated various *E. coli* sequences to be present in the unmapped reads and were considered as a possible source of contamination as shown in Figure 8. On analyzing the BWA-MEM [27] mapping results using the SAMtools flagstat [28], 81.57% of the unmapped reads were successfully aligned to the *E. coli* reference genome. This confirmed *E. coli* as a possible contaminant in the samples. The Ca monoculture at 24h appeared to contain a high level of *E. coli* contamination compared to the different ratios of Ca:Cg coculture. Detailed BWA-MEM alignment results are available in the appendix.

![Figure 8. Top BLAST hits of the unmapped too short reads generated using STAR](image)

### 3.1.3 Alignment Quality Check and Removal of Duplicates

To analyze the quality of the alignment of the reads, Binary Alignment Map (BAM) files were then processed by the CollectAlignmentSummaryMetrics of the Picard tool. The results of the analysis indicated a good mapping quality for all the samples.
which warranted analyzing the duplicates in the mapped reads. Picard’s MarkDuplicates was used to find PCR duplicates from library preparation or optical duplicates from Illumina sequencing. The results indicated that both Ca and Cg reads contained a high level of the PCR duplicates as shown in Figure 9(A) and (B). The number of reads remained after removing the PCR duplicates can be found in the appendix.

Figure 9. Plot showing PCR read duplicates in STAR generated bam files for (A) Ca and (B) Cg samples

3.1.4 Expression Quantification

After removing the PCR duplicates from the BAM files, htseq-count was used for quantifying the gene expression level by counting the number of reads aligned to the genic regions of the genome and generated count files for each sample. As one of the aspects of the project was to compare Ca and Cg cocultures among each other, the count files generated from aligning the trimmed reads of coculture samples to Ca reference genome and mapping the unmapped reads to Cg reference genome were concatenated.
3.2 Downstream Analysis Results

3.2.1 Differential Gene Expression (DGE) Analysis

The count files generated from the htseq-count was imported into DESeq2 in R to perform DGE analysis. To determine the similarity of samples in each group, Principal Component Analysis (PCA) plot was generated and each sample group with same time points clustered well together as shown in Figure 10.

![PCA for all Ca samples and all coculture samples at three time points](image)

*Figure 10. PCA for all Ca samples and all coculture samples at three time points*

PC1 and PC2 were identified by variance stabilizing transformation of normalized counts in DESeq2.

DESeq2 produced a table of genes with their p-values, p-adj values, and log2 fold change and the total number of upregulated and downregulated genes. A threshold of a p-adj value of 0.01 was used to filter the differentially expressed genes for each case.
3.2.2 Pairwise Comparison Analysis

To analyze the significantly perturbed genes during the biofilm growth, three different pairwise comparisons were performed – Case1: cocultures vs. Ca monoculture, Case2: cocultures vs. Cg monoculture, and Case3: cocultures vs. cocultures. The number and percentage of upregulated (positive log2 fold change) and downregulated (negative log2 fold change) genes for each case is summarized in Table 3 – 5.

Table 3. Summary of significant genes generated by DESeq2 for case1- cocultures vs. Ca monoculture

<table>
<thead>
<tr>
<th>Condition</th>
<th>Up</th>
<th>Down</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ca:Cg 3:1 6h vs. Ca Monoculture 6h</td>
<td>3(6%)</td>
<td>46(94%)</td>
<td>49(100%)</td>
</tr>
<tr>
<td>Ca:Cg 3:1 12h vs. Ca Monoculture 12h</td>
<td>16(25%)</td>
<td>49(75%)</td>
<td>65(100%)</td>
</tr>
<tr>
<td>Ca:Cg 3:1 24h vs. Ca Monoculture 24h</td>
<td>621(65%)</td>
<td>338(35%)</td>
<td>959(100%)</td>
</tr>
<tr>
<td>Ca:Cg 1:1 6h vs. Ca Monoculture 6h</td>
<td>50(23%)</td>
<td>169(77%)</td>
<td>219(100%)</td>
</tr>
<tr>
<td>Ca:Cg 1:1 12h vs. Ca Monoculture 12h</td>
<td>81(46%)</td>
<td>95(54%)</td>
<td>176(100%)</td>
</tr>
<tr>
<td>Ca:Cg 1:1 24h vs. Ca Monoculture 24h</td>
<td>722(53%)</td>
<td>630(47%)</td>
<td>1352(100%)</td>
</tr>
<tr>
<td>Ca:Cg 1:3 6h vs. Ca Monoculture 6h</td>
<td>553(48%)</td>
<td>600(52%)</td>
<td>1153(100%)</td>
</tr>
<tr>
<td>Ca:Cg 1:3 12h vs. Ca Monoculture 12h</td>
<td>408(52%)</td>
<td>382(48%)</td>
<td>790(100%)</td>
</tr>
<tr>
<td>Ca:Cg 1:3 24h vs. Ca Monoculture 24h</td>
<td>845(52%)</td>
<td>772(48%)</td>
<td>1617(100%)</td>
</tr>
</tbody>
</table>

For Case1, where all cocultures were compared with Ca monoculture at same time points, DESeq2 results indicated a very high number of differentially expressed genes for coculture Ca:Cg 1:3 vs. Ca monoculture at all three time point compared to other cocultures. A total of 1153 DEG were reported for Ca:Cg 1:3 vs. Ca at 6h, 790
DEG at with upregulated and downregulated at 12h, and 1617 DEGs were reported with 845(52%) upregulated and 772(48%) downregulated genes at 24h.

Table 4. Summary of significant genes generated by DESeq2 for case 2- cocultures vs. Cg monoculture

<table>
<thead>
<tr>
<th>Condition</th>
<th>Up</th>
<th>Down</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ca:Cg 3:1 6h vs. Cg Monoculture 6h</td>
<td>353(48%)</td>
<td>390(52%)</td>
<td>743(100%)</td>
</tr>
<tr>
<td>Ca:Cg 3:1 12h vs. Cg Monoculture 12h</td>
<td>1715(51%)</td>
<td>1660(49%)</td>
<td>3375(100%)</td>
</tr>
<tr>
<td>Ca:Cg 3:1 24h vs. Cg Monoculture 24h</td>
<td>1796(54%)</td>
<td>1506(46%)</td>
<td>3302(100%)</td>
</tr>
<tr>
<td>Ca:Cg 1:1 6h vs. Cg Monoculture 6h</td>
<td>160(41%)</td>
<td>226(59%)</td>
<td>386(100%)</td>
</tr>
<tr>
<td>Ca:Cg 1:1 12h vs. Cg Monoculture 12h</td>
<td>1716(50%)</td>
<td>1737(50%)</td>
<td>3453(100%)</td>
</tr>
<tr>
<td>Ca:Cg 1:1 24h vs. Cg Monoculture 24h</td>
<td>1899(52%)</td>
<td>1738(48%)</td>
<td>3637(100%)</td>
</tr>
<tr>
<td>Ca:Cg 1:3 6h vs. Cg Monoculture 6h</td>
<td>105(39%)</td>
<td>162(61%)</td>
<td>267(100%)</td>
</tr>
<tr>
<td>Ca:Cg 1:3 12h vs. Cg Monoculture 12h</td>
<td>1741(49%)</td>
<td>1810(51%)</td>
<td>3551(100%)</td>
</tr>
<tr>
<td>Ca:Cg 1:3 24h vs. Cg Monoculture 24h</td>
<td>1918(52%)</td>
<td>1779(48%)</td>
<td>3697(100%)</td>
</tr>
</tbody>
</table>

For Case 2, where all cocultures were compared with Cg monoculture at same time points, DESeq2 results indicated a high number of differentially expressed genes for all coculture ratios vs. Cg monoculture at 12h and 24h. A total of 3302 DEG were reported for Ca:Cg 3:1 vs. Cg at 24h, 3637 DEG for Ca:Cg 1:1 vs. Cg at 24h, and 3697 DEG for Ca:Cg 1:3 vs. Cg at 24h.
Table 5. Summary of significant genes generated by DESeq2 for case3- coculture vs. coculture

<table>
<thead>
<tr>
<th>Condition</th>
<th>Up</th>
<th>Down</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ca:Cg 1:3 6h vs. Ca:Cg 3:1 6h</td>
<td>4887(88%)</td>
<td>661(12%)</td>
<td>5548(100%)</td>
</tr>
<tr>
<td>Ca:Cg 1:3 12h vs. Ca:Cg 3:1 12h</td>
<td>827(53%)</td>
<td>732(47%)</td>
<td>1559(100%)</td>
</tr>
<tr>
<td>Ca:Cg 1:3 24h vs. Ca:Cg 3:1 24h</td>
<td>394(64%)</td>
<td>220(36%)</td>
<td>614(100%)</td>
</tr>
<tr>
<td>Ca:Cg 1:3 6h vs. Ca:Cg 1:1 6h</td>
<td>94(89%)</td>
<td>12(11%)</td>
<td>106(100%)</td>
</tr>
<tr>
<td>Ca:Cg 1:3 12h vs. Ca:Cg 1:1 12h</td>
<td>143(40%)</td>
<td>212(60%)</td>
<td>355(100%)</td>
</tr>
<tr>
<td>Ca:Cg 1:3 24h vs. Ca:Cg 1:1 24h</td>
<td>0(0%)</td>
<td>1(100%)</td>
<td>1(100%)</td>
</tr>
<tr>
<td>Ca:Cg 3:1 6h vs. Ca:Cg 1:1 6h</td>
<td>1(14%)</td>
<td>6(86%)</td>
<td>7(100%)</td>
</tr>
<tr>
<td>Ca:Cg 3:1 12h vs. Ca:Cg 1:1 12h</td>
<td>4(36%)</td>
<td>7(64%)</td>
<td>11(100%)</td>
</tr>
<tr>
<td>Ca:Cg 3:1 24h vs. Ca:Cg 1:1 24h</td>
<td>42(86%)</td>
<td>7(14%)</td>
<td>49(100%)</td>
</tr>
</tbody>
</table>

For Case3, where all cocultures were compared among each other at same time points, DESeq2 results indicated a high number of differentially expressed genes for coculture Ca:Cg 1:3 vs. Ca:Cg 3:1 6h. This comparison also indicated more DEG at 12h and 24h time points compared to Ca:Cg 1:3 vs. Ca:Cg 1:1, and Ca:Cg 3:1 vs. Ca:Cg 1:1.

3.2.3 Time Course Analysis

To analyze the time influenced genes during the biofilm growth, two different comparisons were performed - Case4: cocultures vs. Ca monoculture and cocultures vs. Cg monoculture. The number and percentage of upregulated (positive log2 fold change)
and downregulated (negative log2 fold change) genes for coculture comparison with both Ca and Cg is summarized in Table 6.

Table 6. Summary of significant genes generated by DESeq2 for case4- cocultures vs Ca monoculture and cocultures vs Cg monoculture

<table>
<thead>
<tr>
<th>Condition</th>
<th>Up</th>
<th>Down</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cocultures vs. Ca Monoculture</td>
<td>437</td>
<td>209</td>
<td>646</td>
</tr>
<tr>
<td>Cocultures vs. Cg Monoculture</td>
<td>2214</td>
<td>1987</td>
<td>4201</td>
</tr>
</tbody>
</table>

3.2.4 Gene Set Enrichment Analysis

In Case1, Ca to Cg cocultures were compared with Ca monoculture. The upregulation of genes involved in hyphal growth and biological process involved in interspecies interaction were observed in the coculture samples (Ca:Cg 3:1, Ca:Cg 1:1, Ca:Cg 1:3) compared to Ca monoculture at 24h. Cell adhesion processes were upregulated in coculture samples Ca:Cg 1:1, and Ca to Cg ratio of 1:3 compared to Ca monoculture at 6h. Filamentous growth signals were upregulated in coculture samples Ca:Cg 1:1 and Ca:Cg 1:3 compared to Ca monoculture at 24h. Figure 11 shows the gene set enrichment network generated by cytoscape for Ca: Cg ratio 1:3 compared to Ca monoculture at 24h with p-value < 0.05 and False Discovery Rate (FDR q-value) < 0.01.
Figure 11. GSEA enrichment map showing enriched GO biological processes for coculture Ca: Cg ratio 1:3 vs. Ca monoculture at 24h

The network can be interpreted as follows- nodes: gene sets, node size: number of genes in the gene set, blue lines: overlap between genes of different GO processes, and thickness of the blue line: number of genes that overlap. The significant GO biological processes relevant to this research are indicated in black boxes.

To analyze the expression of genes linked to notable biological processes, heat maps were generated. The GSEA generated heat map for the up-regulated significant GO biological processes indicated in the Enrichment Map above is shown in Figure 12.
Understanding How the Relative Abundance of *Candida* Species Impacts Transcriptional Regulation in Coculture Biofilms

Figure 12. GSEA heat map showing significant genes for the enriched GO biological processes along with their expression levels for case1- coculture Ca: Cg ratio 1:3 vs. Ca monoculture at 24h

The expression values of genes are represented as colors with red as high, pink as moderate, blue as low and dark blue as very low gene expression.

To further investigate the leading-edge genes subset, Enrichment Plots were obtained for notable biological pathways. The Enrichment Plot for the filamentous growth signal for Case1- Ca:Cg 1:3 vs. Ca Monoculture at 24h is shown in Figure 13.

Figure 13. GSEA enrichment plot showing filamentous growth linked leading edge genes and ES for Ca:Cg 1:3 vs. Ca monoculture at 24h

The topmost point of the green line indicates the ES for this gene set and shows the extent to which the gene set is overrepresented at the top or bottom of a ranked gene list. The black vertical lines indicate a total of 23 rank-based genes (leading edge gene subset) which were the part of the filamentous growth gene set and contributed the most to the positive ES.
In Case2, the comparison of all cocultures with Cg monoculture at 24h demonstrated cell adhesion and biofilm formation as significant GO biological process. Although, the DGE analysis indicated a high number of upregulated and downregulated genes for cocultures vs. Cg monoculture, the number of significant genes and pathways specific to this research were fewer than that compared to cocultures vs. Ca monoculture. When cocultures were compared in Case3, significant difference were observed only in between Ca:Cg 1:3 with Ca:Cg 3:1 at all-time points. GSEA results demonstrated upregulated filamentous growth for this comparison at all the three time points. The upregulated filamentous growth signals were detected for Cg. Case4 focused on the time-series analysis for all cocultures vs Ca monoculture as baseline and all cocultures vs Cg monoculture as the baseline using the LRT. The GSEA results for the LRT indicated upregulation of genes and pathways related to cell adhesion and biofilm formation for all cocultures vs Ca monoculture as indicated in Table 7. The downregulated pathways and significant genes generated by the LRT for this comparison are included in Table 8.

Table 7. Top 5 biological process with (+ve ES) for case4- time-series analysis

<table>
<thead>
<tr>
<th>GO TERM</th>
<th>SIZE</th>
<th>leadingEdge</th>
<th>ES</th>
<th>NES</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 cell adhesion</td>
<td>9</td>
<td>ALS3, SSK1, SFL1, WOR4, CZF1, MRR2, PHR1, SWI1, SFU1</td>
<td>0.61</td>
<td>1.38</td>
</tr>
<tr>
<td>2 symbiont process</td>
<td>4</td>
<td>BGL2, RIM8, RFX2, SWI1</td>
<td>0.74</td>
<td>1.28</td>
</tr>
<tr>
<td>3 carbohydrate metabolic process</td>
<td>5</td>
<td>CWT1, BGL2, orf19.4031, SGA1, orf19.2638,</td>
<td>0.41</td>
<td>1.27</td>
</tr>
<tr>
<td>4 biofilm formation</td>
<td>12</td>
<td>CSA1, BGL2, CZF1, WAR1, LEU3, QDR3, ALS3 MRR2, PHR1, MFG1, HYR1, VPS1</td>
<td>0.43</td>
<td>1.22</td>
</tr>
<tr>
<td>5 vitamin metabolic process</td>
<td>4</td>
<td>BIO3, THI20, BIO4, THI13</td>
<td>0.68</td>
<td>1.22</td>
</tr>
</tbody>
</table>
Table 8. Top 5 biological process with (-ve ES) for case4- time-series analysis

<table>
<thead>
<tr>
<th>GO TERM</th>
<th>SIZE</th>
<th>leadingEdge</th>
<th>ES</th>
<th>NES</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 cytoskeleton-dependent intracellular transport</td>
<td>2</td>
<td>MYO1, ARP1</td>
<td>-0.85</td>
<td>-1.33</td>
</tr>
<tr>
<td>2 protein targeting</td>
<td>5</td>
<td>COG4, PMM1, MON2, orf19.3843, SPC3</td>
<td>0.5</td>
<td>-1.29</td>
</tr>
<tr>
<td>3 DNA metabolic process</td>
<td>3</td>
<td>POL1, CCE1, MCM3</td>
<td>0.31</td>
<td>-1.25</td>
</tr>
<tr>
<td>4 response to stress</td>
<td>4</td>
<td>SWI6, FGR41, HST3, DNA2</td>
<td>0.45</td>
<td>-1.28</td>
</tr>
<tr>
<td>5 vacuolar transport</td>
<td>5</td>
<td>COG4, MON2, orf19.4253, BUD7</td>
<td>0.63</td>
<td>-1.23</td>
</tr>
</tbody>
</table>

The GSEA Enrichment Map for Cocultures vs. Ca Monoculture with p-value < 0.05 and FDR q-value < 0.1 is shown in Figure 14 and the heat map for this case shown in Figure 15.

Figure 14. GSEA enrichment map showing enriched GO biological processes for case4-time-series analysis for cocultures vs Ca monoculture

The network can be interpreted as follows: nodes: gene sets, node size: number of genes in the gene set, blue lines: overlap between genes of different GO processes, and thickness of the blue line: number of genes that overlap. The significant GO biological processes relevant to this research are indicated in black boxes.
Understanding How the Relative Abundance of *Candida* Species Impacts Transcriptional Regulation in Coculture Biofilms

![Figure 15. GSEA heat map showing significant genes for the enriched GO biological processes along with their expression levels for case4- time-series analysis for cocultures vs Ca monoculture](image)

The expression values of genes are represented as colors with red as high, pink as moderate, blue as low and dark blue as very low gene expression.

The 19 leading edge gene subsets for the biofilm formation signal which resulted the most to the positive ES is shown in the Enrichment Plot shown in Figure 16.

![Figure 16. GSEA enrichment plot showing filamentous growth linked leading edge genes and ES for case4- time-series analysis](image)

The topmost point of the green line indicates the ES for this gene set and shows the extent to which the gene set is overrepresented at the top or bottom of a ranked gene list. The black vertical lines indicate a total of 19 rank-based genes (leading edge gene subset) which were the part of the biofilm formation gene set and contributed the most to the positive ES.
Gene set enrichment analysis provided insight into the significant biological pathways during the biofilm growth. Detailed information about the top upregulated and downregulated GO biological pathways along with their leading-edge gene subsets and enrichment plot for all the Cases can be found in the appendix.

4 DISCUSSION

Although the antifungal drugs used for treating fungal infections seem to be diverse and abundant, only a few antifungal agents are available at present to handle systemic infections with *Candida* species [10], [29]. The relative composition of Ca and Cg in coculture influences the biofilm formation resulting in life-threatening conditions like Candidemia [1]. The biofilm shields the fungus from host immune defense and antifungal agents, often leading to the failure of therapeutic options. Considering the necessity to find more efficient therapies to handle *Candida* infections, there is a need to understand the gene expression leading to biofilm development and antifungal resistance in coculture.

The upstream analysis of the RNA-Seq pipeline along with the various quality check steps is performed on SJSU High-Performance Computing (HPC) using Bash scripts for each process. FastQC indicated samples have good quality scores before trimming, and adaptors are removed using Trimmomatic. Alignment using STAR aligner show that majority of reads are either uniquely mapped or unmapped due to a possible source of contamination which was confirmed to be *E. coli*. The gene expression was quantified by counting the sequencing reads aligned to the genic regions of the reference genome using htseq-count. DESeq2 was utilized to study the differentially expressed
genes and finally, GSEA was performed to determine if there are concordant differences between the gene expression in *Candida* coculture biofilms compared to monoculture over the course of biofilm growth.

The study conducted by Dr. Katy Kao’s lab indicated a rise in antifungal drug resistance in the coculture Ca:Cg ratio of 1:3 which formed the highest biofilm [1]. This suggested interspecies interactions between Ca and Cg, which is allowing increased biofilm formation and Ca hyphal growth. The result of this RNA-Seq supported the findings as Ca:Cg ratio of 1:3 when compared with Ca monoculture at 24h time point indicated filamentous growth, hyphal growth and biological process involved in interspecies interaction as upregulated GO biological process.

4.1 Significant Ca Genes and Biological Pathways for Ca:Cg Cocultures Compared to Ca Monoculture

Ca:Cg 1:3 coculture compared to Ca monoculture at all time points showed cell adhesion, filamentous growth, and biological process involved in interspecies interaction as significantly upregulated GO pathways as shown in Figure 11 and 17. The heat maps with upregulated genes in red and downregulated genes for Ca:Cg 1:3 coculture vs Ca monoculture is shown in Figure 18. Although, Ca:Cg 1:1 and Ca:Cg 3:1 cocultures vs. Ca monoculture at 24hr indicated biofilm linked genes as upregulated, these genes were not observed at 6h and 12h time points. The result suggests the importance of interspecies interaction in coculture and a possible synergistic connection between Ca and Cg in pathogenesis, specifically when the relative ratio of Cg is higher.
Understanding How the Relative Abundance of \textit{Candida} Species Impacts Transcriptional Regulation in Coculture Biofilms

\textbf{Figure 17.} GSEA enrichment map showing enriched GO biological processes for case1-Ca:Cg 1:3 vs. Ca monoculture at (A) 6h and (B) 12h

The network can be interpreted as follows- nodes: gene sets, node size: number of genes in the gene set, blue lines: overlap between genes of different GO processes, and thickness of the blue line: number of genes that overlap.
Figure 18. GSEA heat map showing significant genes for the enriched GO biological processes along with their expression levels for Case1- Ca:Cg 1:3 vs. Ca monoculture at (A) 6h and (B) 12h

The expression values of genes are represented as colors with red as high, pink as moderate, blue as low and dark blue as very low gene expression.

For Ca:Cg 1:3 comparison with Ca monoculture, the significantly upregulated genes involved in cell adhesion were HWP1, which encodes for Hyphal wall protein 1, and ALS3 which encodes for Agglutinin-like protein 3. These genes release several adhesins that are expressed on the surface of the cell and plays an important role in adherence of Candida to the host tissues and mediating biofilm formation [30], [31]. Higher expression of adhesion genes is linked with increased virulence and increased resistance to antifungal drugs. The genes involved in filamentous growth signal, including UME6, HGC1, and ALS3 were significant at all three time points. UME6 codes a filament-specific transcriptional regulatory protein for hyphal growth in the dearth of filament-inducing conditions and the gene is linked with hyphal extension and virulence [32]. HGC1 is a hypha-specific gene that is essential for hyphal morphogenesis and encodes for virulence factors [33]. Biological process involved in interspecies interaction
Understanding How the Relative Abundance of *Candida* Species Impacts Transcriptional Regulation in Coculture Biofilms

was shown to be upregulated. The upregulation of these genes in coculture compared to Ca monoculture suggests an apparently neutral or collaborative association between the two *Candida* species.

Some of the common downregulated genes for all the cocultures vs. Ca monoculture at 24h were *NAN1*, *NOG1*, and *SPB4*. These genes belong to ribosome biogenesis gene set and known to play a crucial role in synthesis of ribosomal proteins [34]-[36].

4.2 Significant *Cg* Genes and Biological Pathways for Ca:*Cg* Cocultures Compared to *Cg* Monoculture

In Case2, where all cocultures were compared to Cg monoculture at same time points, GSEA results showed significant upregulation in cell adhesion and biofilm formation signals in Ca:Cg coculture samples compared to Cg monoculture at all time points and 24h, respectively. Similar enriched pathways were seen in case of Ca:Cg ratio of 1:1 and Ca:Cg ratio of 1:3 at 24h. This agree with some previous findings on Ca and Cg coculture biofilm studies which showed that Cg adhesion linked genes were upregulated in the presence of Ca. Heat maps with significantly up-regulated genes in red for all coculture compared to Cg monoculture at 24h are shown in Figures 19. Although, there were genes and pathways of interests that were upregulated in coculture compared to Cg monoculture, the number were fewer than that compared to Ca monoculture. Some of the Cg adhesions genes that were upregulated at all time points were *EPA1*, *EPA6*, *EPA11*, and *HSP12*. The *EPA* gene family encodes glycosylphosphatidylinositol anchored-adhesins which is a major contributor of virulence in Cg [37], [38]. The Cg
upregulated genes of the *EPA* family are also indicated to facilitate the adhesion of Ca-to-Ca hyphae.

![GSEA heat map showing significant genes for the enriched GO biological processes along with their expression levels for case2- cocultures vs Cg monoculture at 24h for (A) Ca:Cg 3:1, (B) Ca:Cg 1:1, and (C) Ca:Cg 1:3](image)

The expression values of genes are represented as colors with red as high, pink as moderate, blue as low and dark blue as very low gene expression.

Some of the downregulated genes for all the cocultures vs. CG monoculture at all time points were *DBP3, NOP8, SSB1,* and *SNF6.* Interestingly, as seen in Case1, these Cg genes were also linked to ribosome biogenesis gene set.
4.3 Significant Ca and Cg Genes and Biological Pathways for Ca:Cg Coculture Comparisons

In Case3, where cocultures were compared amongst each other at same time points, GSEA results showed upregulated filamentous growth for coculture Ca:Cg 1:3 vs. Ca:Cg 3:1 at all the three time points. Interestingly, the upregulated filamentous growth signals were identified for Cg as seen in Figure 20. *SPT3*, *SPT11*, and some unannotated Cg genes involved in filamentous growth were found to be significant as shown in the heat map in Figure 21. No gene sets were found to significantly downregulated at p-value = 0.05 and FDR < 25%.

![Figure 20. GSEA enrichment map showing upregulated GO biological processes for case3- Ca:Cg 1:3 vs. Ca:Cg 3:1 at 24h](image)

The network can be interpreted as follows- nodes: gene sets (gene sets for Ca are denoted with CA as the prefix, and gene sets for Cg are denoted with CG as a prefix), node size: number of genes in the gene set, blue lines: overlap between genes of different GO processes, and thickness of the blue line: number of genes that overlap.
Figure 21. GSEA heat map showing significant genes for the enriched GO biological processes along with their expression levels for case 3- Ca:Cg 1:3 vs. Ca:Cg 3:1 at 24h.

The expression values of genes are represented as colors with red as high, and pink as moderate gene expression.

The coculture ratio Ca:Cg 1:3 when compared to Ca:Cg 1:1 and the coculture ratio Ca:Cg 3:1 when compared to Ca:Cg 1:1 showed no significant upregulation in biofilm linked pathways or genes.

4.4 Significant Time Influenced Ca and Cg Genes and Biological Pathways for Time Series Analysis

The time-series analysis using the LRT indicated biofilm related notable pathways such as filamentous growth, biofilm formation, cell adhesion, and symbiont process as shown in Figure. In case of coculture comparison to Ca monoculture, the notable cell adhesion genes that were found to be influenced by time across all conditions were ALS3, HLS4, MRR2, PHR1, WOR4, and YVC1. The GSEA generated heatmap of the top 15 significant adhesion genes for each coculture samples vs. Ca monoculture samples as baseline is shown in Figure 22.
Understanding How the Relative Abundance of Candida Species Impacts Transcriptional Regulation in Coculture Biofilms

Figure 22. GSEA generated heat map for time-series analysis for cell adhesion signal

The expression values of the time influenced genes are represented as colors with red as high, pink as moderate, blue as low and dark blue as very low gene expression.

The downstream analysis of the RNA-Seq pipeline gave some insight into the biofilm and virulence-related gene expression in Candida. Various significant pathways and upregulated or downregulated genes were identified during the course of biofilm formation using this RNA-Seq pipeline. As the two Candida species when present together can cause more serious and lethal diseases like Candidemia, utilizing RNA-Seq to explore the transcriptomic alterations in coculture biofilms across different time frames may open avenues that could help in identifying novel drug targets. This could also help in developing new antifungals to treat lethal biofilm-related infections.
Understanding How the Relative Abundance of *Candida* Species Impacts Transcriptional Regulation in Coculture Biofilms

**REFERENCES**


APPENDIX

Sample Information

https://www.dropbox.com/s/yg04o0nrpy4x9gv/Sample_Information.xlsx?dl=0

FastQC Results (Raw Reads and Trimmed Reads)

https://www.dropbox.com/sh/tslwkb2rcrr342/AABfo-ex05JghmTCxudRyL1Ta?dl=0

Trimmomatic Results

https://www.dropbox.com/s/pe22825nd9a17bz/Trimmomatic_Result.xlsx?dl=0

Mapping Results

https://www.dropbox.com/s/dxgt4zaca43z0w6/Mapping_Results.xlsx?dl=0

Picard Results (Ca and Cg)

https://www.dropbox.com/sh/5kfz5ntd18hxpr91/AAD2p4jMSMLK_4wJfJn0CB_pa?dl=0

GSEA

https://www.dropbox.com/s/4pc3l7612e3iu2e/GSEA.xlsx?dl=0