THE UNIVERSITY OF CHICAGO

INVESTIGATING CELLULAR VARIABILITY IN FUNGAL PATHOGENS BY DEVELOPING MDROP-SEQ, A HIGH THROUGHPUT, SINGLE CELL RNASEQ TECHNOLOGY FOR YEAST SPECIES

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE BIOLOGICAL SCIENCES
AND THE PRITZKER SCHOOL OF MEDICINE
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

COMMITTEE ON GENETICS, GENOMICS, AND SYSTEMS BIOLOGY

BY

RYAN PATRICK DOHN

CHICAGO, ILLINOIS

JUNE 2023

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Acknowledgements

Many of the members of the Basu lab as well as external collaborators were involved in the intellectual and physical work represented by this dissertation. Above all, I would like to thank my advisor, Dr. Anindita Basu, for going above and beyond in both guidance and support during my time in graduate school. Her mentorship was invaluable for the scientific work presented by this dissertation, as well as projects and work that is not represented here. Perhaps more importantly, her guidance and support were key in the personal growth, both as a scientist and as a person, that I achieved during my time at the University of Chicago. I think of myself as immensely fortunate to have had Dr. Basu in such a position of influence on my life and am immeasurably grateful for all that she has done. Additionally, I would like to thank the members of my thesis committee: Dr. Yoav Gilad, Dr. Luis Barreiro, and Dr. Bana Jabri. Their advice, mentorship, patience, and support were instrumental in the creation of this dissertation.

I am immensely thankful for the many current and former members of the Basu lab, whom working with will have been one of the greatest joys of my graduate school career. I would like to thank Dr. Bingqing Xie, Dr. Susan Olalekan, Dr. Katelyn Mika, Dr. Pretty Bajwa, Dr. Ran Zhou, Dr. Alan Selewa, Dr. Andres Moya-Rodriguez, Heather Eckart, Rebecca Back, Trevor Wood, Dylan Cook, Allison Hohreiter, and Jianqiao Liu for their support and help over the years. In addition, I would like to give a special thanks to both Heather Eckart and Rebecca Back for their significant contributions to this work and years of assistance and support, both as scientists, but also as friends. Much like how I found myself fortunate to work underneath Dr. Basu, I am forever grateful to have had the privilege and opportunity to work with such a brilliant and wonderful group of people. Their influence will be felt in whatever endeavors, scientific or otherwise, come next.

Furthermore, I would like to thank our collaborator, Dr. Reeta Prusty Rao, for her part in the analysis and interpretation of the work presented here. Her insight into the biology of yeast, and specifically *C. albicans*, was greatly appreciated. I regret not having the opportunity to learn more from Dr. Rao had the world not conspired against us with a global pandemic. I would like to thank Dr. Sebastian Pott for his input on data analysis during some of the early stages of yeast scRNA-seq.

In my time here, I received a significant amount of administrative support both from the Genetics, Genomics, and Systems Biology program and from the Section of Genetic Medicine. I would like to thank Susan Levison, who has been a pillar of support for my entire graduate career, both administratively and personally. I would also like to thank Tamiko Charley for her aid and support. Finally, I would like to thank Sandra Dantzler for her administrative support in the final stretch of my time here.

Finally, I would like to share my gratitude towards the people I have met throughout my time at the University of Chicago including my cohort and many friends. I am forever grateful to have met so many people that I am proud to have been able to share this experience with. For my cohort, our diverse backgrounds and perspectives allowed for learning and growth beyond my expectations, and I look forward to the future to see what we can achieve. For my friends, thank you for being part of my life, both now during this journey and beyond. I have found people here that have changed my life for the better and I cannot express how much their support and friendship have meant to me.

Abstract:

The rise of high throughput single-cell RNA sequencing increased our understanding of cellular population dynamics and the heterogeneity and stochasticity between individual cells. These gains have thus far been lost on microbial cells due to complicating factors that rendered microbes incompatible with technologies developed on mammalian cells. However, not all these drawbacks are present within Eukaryotic yeast cells, making them an ideal target microbe for technological development. In this dissertation, the development of mDrop-seq, a high throughput scRNA-seq for yeast species, is displayed through the processing of thousands of cells of two yeast species. In the first chapter, we use the model organism *S. cerevisiae* for initial development, testing, and profiling of 35,109 total yeast cells. In doing so, we test appropriate lysis conditions and time to allow for droplet microfluidic compatible cell lysis. *S. cerevisiae* cells are subjected to a 42°C heat shock in order to determine mDrop-seq capability to detect a large scale stress response at single cell resolution. Analytical pipelines for single-cell analysis that were developed for mammalian data are shown to work with yeast libraries, allowing for differential gene expression (DGE), clustering analysis, cell cycle assignment, and pseudo-time trajectory analysis. In the second chapter, we described further modifying and testing mDrop-seq on the clinically relevant species *Candida albicans*. Despite challenges such as thicker cell walls, we display mDrop-seq's ability to process *C. albicans* cells using exposure to the antifungal drug Fluconazole. The final chapter of this dissertation uses mDrop-seq to search for sources of variation and batch effects within our data. We show that the activation of stress response pathways causes a reduction in transcriptomic variation between *C. albicans* cells. In total, the chapters of this dissertation show mDrop-seq's value as a low cost, scalable scRNA-seq technology for yeast species.

Chapter 1: Introduction:

An individual cell, whether alone as a single-celled organism or part of a multicellular whole, serves as the base functional unit of life. A singular cell is capable of managing its own metabolism, reproducing through cellular division, and responding to the environment around it to survive. The instructions to do all this are carried within a full set of genetic material, DNA. DNA contains sequences that include protein coding genes which encode for proteins which carry out functions, but also a vast and sophisticated array of regulatory regions that control which genes are expressed and by how much. Through this regulatory control, only subsets of genes may be actively being transcribed into mRNA. This makes the full transcriptome a powerful proxy for understanding the current state of any given set of cells. The ability to profile gene expression activity is a powerful tool to probe cellular identity and state, genomic functionality, and cell response to stimuli.

In order to take advantage of such a proxy, RNA-seq was developed almost two decades ago [1], [2] to gather the total amount of mRNA within cells and create a readout for what genes are active and by how much. Since, it has become a common tool that has vastly shaped out understanding of the genome. The general workflow for RNA-seq has changed very little since its inception: extract RNA, enrich for mRNA, reverse transcribe to cDNA, and prepare a library for next-generation sequencing. Among the most common uses of RNA-seq is for Differential Gene Expression (DGE), which aims to see how gene expression changes between two samples. This can include resolving a specific cell type in both a healthy and diseased state or before and after an exposure to environmental stimuli. Through determining which genes have statistically different expression levels, DGE studies provide significant understanding into the various mechanisms that allow for phenotypic differences between cells.

While proven to be a powerful tool for DGE studies with a high species compatibility [3]–[6], bulk RNA-seq techniques have some drawbacks. Crucially, the technique cannot resolve specific cell types easily, causing the potential loss of either rarer or formally unknown cell types. Additionally, any spatial information of multicellular systems is lost as well due to the mixing of mRNA. To address this, single cell RNA-seq (scRNA-seq) protocols have been developed to look into the transcriptome of individual cells, thereby conserving much of the information lost during bulk RNA-seq. The first scRNA-seq method was published in 2009 worked on mouse oocytes and blastomeres isolated in microfuge tubes [7]. Since then, a variety of different methods have been established for scRNA-seq. Initial techniques were considered low throughput and involved sorting cells into plates with lysis buffer [8]. Later higher throughput techniques were created using a variety of different cell isolation methods including droplet microfluidics [9], [10], nanowell loading [11], in situ sequence barcodes [12], and split-pool ligation [13].

The rise of high-throughput scRNA-seq has led to a greater understanding of the functional and phenotypic heterogeneity present in our body on a cellular level. Primarily developed for mammalian cells [9], scRNA-seq uses the transcriptome of a single cell to analyze cell type [9], cell state [14], and cell response [15]. The application of scRNA-seq technology has allowed for the discovery of new cell types within complex multicellular tissues that were previously lost in bulk experiments [16]. While variation between different cell types (in a multicellular organism) or cells of different species may be expected, scRNA-seq techniques have shown that there is significant cell-to-cell heterogeneity even between otherwise identical cells [17], [18]. High-throughput techniques can examine thousands of cells at once, adding statistical power to determine variability between cells [9], [19].

The success of single cell technologies on the mammalian systems they were developed on have allowed for a more complete picture into the function and relationship of higher Eukaryotes. However, technological challenges, such as the tough cell walls, small size, and concomitantly smaller amounts of transcripts per cell [20] have prevented similar applications in unicellular microbial organisms [21]. Microbial studies have thus far had to be done at the population level, both physically through observations on growth and transcriptomically through the bulk sequencing of potentially millions of cells. Such methods, like with bulk RNA-seq on mammalian cells, do not capture how individual microbes contribute to the population as a whole. Studies that have observed microbial heterogeneity revealed different behaviors in growth speed and metabolism in otherwise genetically identical cells [22]–[25]. The development of a scRNA-seq method compatible with single-cellular microbial organisms is essential to truly understanding the full dynamics of microbial populations.

One of the most consistent issues facing the application of current scRNA-seq technologies to microbial populations is the microbial cell wall. High-throughput technologies require the isolation of individual cells, whether by droplet or microwell plate, prior to cellular lysis and mRNA capture. However, the addition of a hard to lyse cell wall often means that lysis techniques require mechanical perturbation that is incompatible with isolation compartments. Non-mechanical methods of lysis may cause issues with microfluidic components that make them difficult to use. The strength and rigidity of the microbial cell walls composed of diverse components, e.g., peptidoglycans in bacteria and chitin and β-glucan layers in yeasts [26]–[28], make them resistant to most lysis agents used for scRNA-seq. Split-pool barcoding based techniques such as SPLiT-seq [13] and microSPLiT [29] may not be as susceptible to the lysis issue as bulk lysis is possible after barcoding. Microbes also have significantly less mRNA

compared to animal cells, with estimates ranging up to two orders of magnitude less for bacterial cells [30], [31]. This issue is further complicated by the capture rates of current scRNA-seq technologies, where only a fraction of the mRNA in any given cell is successfully reverse transcribed [32].

Despite the challenges, achieving high-throughput microbial scRNA-seq is essential to understanding the heterogeneity and complex interactions between cells in a population. Yeasts, such as the model organism *S. cerevisiae*, offer a few advantages as microbes that make them an ideal candidate for development. As Eukaryotic fungi, yeasts have polyadenylated mRNA similar to that of mammalian cells, which is often used as the method of capture to prevent dilution from rRNA and other molecules [33]. Yeasts are also much larger than a typical bacterial cell, lowering the impact of low mRNA counts within an individual cell. *S. cerevisiae* is a model organism that has an exceptionally well annotated genome, and its biology is higher studied, making it an ideal candidate. Indeed, others who have attempted to accomplish a similar technique chose similarly. A few recent studies have profiled *S. cerevisiae* at high throughput with single-cell resolution using multiple cells from clonal populations to increase signal [34], the 10X platform [35], [36], as well as the Drop-seq platform [37].

However, clinically relevant yeast species, such as *C. albicans*, have yet to be characterized at single-cell resolution and high-throughput. Plate-based scRNA-seq in Muñoz *et al.* [38] was used for isolated macrophage - *C. albicans* pairs and numbered in the low hundreds of cells. The pathogenicity of *Candida* species requires higher levels of precautions, cells are more resistant to lysis, and their overall biological understanding is not as far progressed. Single-cell RNA-seq on fungal pathogens such as *C. albicans* and *Candida auris* can help understand the commensal-to-pathogenic switch that leads to opportunistic infections [18,19]. As one of the

most common hospital-acquired infections, *C. albicans* can cause both superficial infections in humans and severe systemic infections in immunocompromised individuals [41]. Understanding the heterogeneity in how individual yeast cells respond to changes in the hosts' immune system or their shifting microbiome can help treat and prevent such infections. Additionally, the emergence of drug-resistant microbes is an urgent human health crisis [42]. Molecular mechanisms that confer protection to a select few cells when the parent population is decimated by anti-microbial agents can help understand the rise of drug-resistant strains [43].

In this dissertation, we present microbial Drop-seq, or mDrop-seq as a method of high-throughput scRNA-seq of different yeast species [37]. With modifications to the Drop-seq platform [9], we were able to accomplish microfluidic compartmentalization of single cells, in-droplet lysis and cellular barcoding of two species of yeast, *S. cerevisiae* and *C. albicans* for scRNA-seq at scale. We quantified the transcriptional heterogeneity within clonal populations of yeast cells and profiled their response at single-cell resolution to environmental stresses, such as heat shock and exposure to fluconazole, a common anti-fungal agent. Under heat shock at 42 °C, we observe differential expression of key stress response genes, including *HSP12*, *HSP26*, and *HSP42* in *S. cerevisiae* cells. When exposed to fluconazole, *C. albicans* cells show differential upregulation of key drug target pathways such as ergosterol biosynthesis and ribosome activity and an overall increase in histone activity. Importantly, both *S. cerevisiae* and *C. albicans* show disruption in their cell cycle patterns under heat shock and fluconazole exposure, respectively. Taken together, we posit that the cell cycle state of the yeast cell in a population of continuously cycling cells is related to the variability in the cell's response to stress. Finally, we display that several sources of variation may have an impact on batch effects between yeast replicates and

that these batch effects appear to be alleviated in stressed samples due to the convergence of the transcriptome to the stress response.

mDrop-seq's ability to simultaneously profile a mix of *S. cerevisiae* and *C. albicans* cells demonstrate that different fungal species are amenable to simultaneous single-cell transcriptomic profiling. mDrop-seq will help decipher transcriptional variability and the use of alternate stress response pathways in yeasts in response to external stresses such as drug treatment and find immediate use in developing new classes of drugs or vaccines for emerging drug-resistant populations.

Chapter 2: Development of mDrop-seq using the Heat Shock response of the model organism *S. cerevisiae*

2.1 Introduction:

The first task of designing a scRNA-seq technique compatible with various yeast species is to start with one of the most basic, widely used yeasts in science: the baker's yeast *S. cerevisiae*. *S. cerevisiae* has been widely used as a model organism for a variety of molecular biology applications. Model organisms are among the best places to start when designing and optimizing new methodology, as their various biological processes are highly studied, often easier to handle and manipulate, and are capable of serving as proxies for hard-to-use samples [44]. In this case, budding yeast has been a staple model in the study of genomics applications such as the function of genes and the consequences of gene silencing. Methods of genetic perturbation and mutagenesis for *S. cerevisiae* have been available and widely used for decades. Additionally, yeasts provide a fast growing, Eukaryotic model that is available in both haploid and diploid forms, allowing for a diverse range of molecular biology applications.

The *S. cerevisiae* genome was first fully sequenced and released in 1996, becoming the first Eukaryotic organism to have its full genome sequenced [45]. Upon the release if this genome, genomics tools were created including genome wide deletion and overexpression panels [46]–[48], GFP fluorescent tagged and labeled strains [49], and many others. The creation of these tools allowed for many discoveries into the function of the 6,275 known genes within the *S. cerevisiae* genome. The purpose and function of many of the genes was recorded through the use of knockouts and expression studies, and eventually a systems biology approach could be applied to start exploring how genes effect full pathways beyond their initial functions [50]–[52]. The

vast amount of knowledge we have gained through studying budding yeast is why it remains one of the most commonly used Eukaryotic models.

The wide degree of knowledge and understanding of *S. cerevisiae* biology makes it an excellent candidate for technique development. The process of adapting budding yeast to a single-cell RNA-seq platform comes with many technical difficulties, including lysis compatible with high throughput isolation methods and downstream data analysis. It is for these reasons that much of the progress that has been made on microbial single cell, both within this dissertation and within published works within recent years, has started with the widely used model organism *S. cerevisiae*. mDrop-seq makes use of the open source Drop-seq protocol, adapting it to the special requirements of yeast lysis. Urbonaite *et al.* also makes use of the Drop-seq platform,[37]. Jariani [35] and Jackson [36] both use 10X Genomics technology and have a similar approach to achieving within-droplet lysis on yeast cells.

One of the first issues that must be overcome in any attempt at microbial single cell RNA-seq is the issue of cell lysis. Microbial lysis techniques often take advantage of physical perturbations such as crushing in bead mills, freeze-thaw cycles, or sonication to break apart microbial cell walls [53]. However, these types of perturbations are often incompatible with many common forms of high throughput single-cell protocols. Creating a droplet compatible lysis buffer for *S. cerevisiae* proved to be much easier than it was for the *Candida* species (discussed in later chapters) due to easier lysis of the species' relatively thinner cell wall (~120 nm [54]). While the methods presented here include slight modifications to the Drop-seq lysis buffer in addition to adding the lytic enzyme Zymolyase, Urbonaite *et al* displays that Zymolyase combined with the standard Drop-seq lysis buffer is adequate when followed by a

heated incubation to allow for lysis [9], [37]. We determined that a stronger lysis buffer and enzyme combination allows for complete lysis while still being droplet compatible.

The creation of mDrop-seq allows for the probing of the variability we captured between simple, unstimulated cells within a single library. However, it also allows us to observe how cells respond to ever changing environments on a single cell basis, granting a greater understanding of how microbial populations change and survive. Single-celled organisms are constantly at the mercy of their environment. Environmental changes can come in the form of temperature swings, nutrient availability, osmotic stresses, and many more [55]. Even in the absence of lethal environmental changes, subtle reactions to small perturbations may confer advantages within a population. The ability to see these subtle differences can be a powerful tool in further understanding microbial adaptions.

A cell's response to environmental stresses often includes a transcriptional upregulation of stress pathways and a complete reconfiguration of the transcriptome to protect itself. *S. cerevisiae* contains an Environmental Stress Response (ESR), controlled by transcription activators MSN2/MSN4 which upregulate ~200 genes in response to environmental stresses such as heat shock, osmotic shock, oxidative stress, glucose depletion, and many others [56], [57]. In specific stress environments, such as a sudden heat shock, more specific heat shock responses controlled by HSF proteins are activated alongside the ESR, creating a further transcriptional cascade of various gene sets including protein chaperones and cell wall integrity genes [55]. As the response to stresses such as heat shock involve rapid, large scale transcriptional changes within a cell, it is an ideal way to test the ability of our technology to detect such changes.

In this chapter, we will showcase the ability of mDrop-seq using *S. cerevisiae*. With modifications to the Drop-seq platform [9], we were able to accomplish microfluidic

compartmentalization of single cells, in-droplet lysis and cellular barcoding of *S. cerevisiae* scRNA-seq at scale. We quantified the transcriptional heterogeneity within clonal populations of yeast cells and profiled their response at single-cell resolution. We took advantage of the well annotated genome and highly studied heat shock response to be able to determine our technique's ability to capture unstimulated cell to cell variability, such as cell cycle, as well as its ability to capture large transcriptional changes that we know to occur during the heat shock stress response. In doing so, we can determine how the unique data provided by microbial cells such as yeast interacts with scRNA-seq protocols that have thus far been designed solely for higher eukaryotic cells.

Chapter 2 is based on and mostly a reprint of the paper: R. Dohn *et al.*, "mDrop-Seq: Massively Parallel Single-Cell RNA-Seq of Saccharomyces cerevisiae and Candida albicans," *Vaccines*, vol. 10, no. 1, 2022, doi: 10.3390/vaccines10010030.

2.2 Methods:

Yeast Cell Culture

Saccharomyces cerevisiae (strain BY4741, Open Biosystems) cells were grown as a dense culture (>1x10^8 cells/mL) in YEPD (MP Biochemicals, #MP114001022) at 27 °C overnight. *The S. cerevisiae* culture was heavily diluted (1:20) in fresh YEPD medium and grown for 4 h at 27 °C, following which the cells were placed on ice and chilled.

Heat Shock Stimulation of S. cerevisiae

After *S. cerevisiae* was grown in YEPD for 4 h post-dilution, cells were counted in YEPD using Neubauer Improved (NI) hemocytometer (InCyto, #DHC-N01-2). A total of 700,000 cells were aliquoted into a 1.7 mL microfuge tube and heat shock stimulation was

applied by placing the tube in an Eppendorf F1.5 Thermomixer set to 42 °C and 500 rpm for 20 min. At the end of the heat shock incubation, the microfuge tube containing the *S. cerevisiae* cells were placed on ice for 5 min. The cells were then washed once with ice-cold 1X PBS (Teknova, #P0195) and 0.01% BSA (NEB, #B9000Sm), henceforth referred to as PBS-BSA, quickly recounted, and brought to a concentration of 700,000 cells/mL in PBS-BSA. A total of 10 µL of RNase Inhibitor was added to 1 mL of yeast cells in PBS-BSA and mDrop-seq was performed as described below. The emulsion droplets were collected on ice to preserve the heat shock signal during the droplet encapsulation period.

mDrop-seq Cell Preparation and Co-Encapsulation in Droplets

Yeast cells were centrifuged separately at 1000 xg for 3 min in a swinging bucket centrifuge at 4 °C. The cells were washed twice with ice-cold PBS-BSA. Following the washes, 10 µL of cells were sampled and counted using a NI hemocytometer (InCyto, #DHC-N01-2). ~1 mL of cells at 700,000 cells/mL suspended in PBS-BSA was placed in a 2.5 mL syringe (BD, #309657). A total of 10 µL of RNase Inhibitor (Lucigen, #F83923) was added per 1 mL suspension immediately before microfluidic encapsulation. A 75 µm DroNc-seq device fabricated in the cleanroom using published design and protocol [58] was used for droplet generation. Cells and beads were co-flowed into the microfluidic device at 1.5 mL/h. Cells at 700,000 cells/mL and 4,500,000 droplets/mL gives a Poisson loading distribution with $\lambda = 0.15$.

Barcoded beads (ChemGenes, #Macosko-2011-10(V+)) were suspended in Droplet Yeast Lysis Buffer (DYLB, see below) at 350,000 beads/mL and kept in suspension by constant stirring with a magnetic tumble stirrer and flea-magnet setup (V&P Scientific, #VP 710, #772DP-N42-5-2); the flea magnet is placed in the syringe containing the barcode beads suspended in lysis buffer and the stirrer kept in the vicinity of the syringe during droplet

generation. Cells and beads in lysis buffer were co-encapsulated in drops using a surfactant-oil mix (BioRad, #1864006) flowed at 8 mL/h in a 10 mL syringe (BD, #302995) as the outer carrier oil phase. Droplets were collected at ~3750 droplets/sec for 30 min in 50 mL tubes (Genesee Scientific, #28-106).

Droplet Yeast Lysis Buffer: Barcoded beads were suspended in Droplet Yeast Lysis Buffer or DYLB, consisting of 6.7% (w/v) Ficoll PM-400 (GE Healthcare, #17-0300-05), 225 mM Tris pH 7.5 (Teknova, #T2075), 22 mM EDTA (Fisher, #BP2482-500), 0.67% sarkosyl (Teknova, #S3377), 55 mM KH_3PO_4 (Sigma, #P5629-25G), 1.3 mM DTT (Teknova, #D9750), 0.1% (v/v) β-mercaptoethanol (Sigma, #M6250-10ML), and 450 units/mL zymolyase (Zymo Research, #E1005); this mix was optimized for *S. cerevisiae* lysis in drops.

S. cerevisiae and *C. albicans* species mixing: Each yeast cell population, *S. cerevisiae* or *C. albicans* were processed separately as follows: Each yeast species was centrifuged at 1000 xg for 3 min in a 4 °C swinging bucket centrifuge and washed twice with ice-cold PBS-BSA. Following the washes, 10 µL of each cell aliquot was sampled from each species and counted using a NI hemocytometer.

Two experiments at two different Poisson loading concentrations were performed to calculate doublet rates at these loading conditions: 175,000 cells from each species were combined and the final volume adjusted to 1 mL of PBS-BSA for $\lambda \approx 0.077$; 350,000 cells from each species were suspended in a total volume of 1 mL PBS-BSA for $\lambda \approx 0.15$. Due to the presence of *C. albicans*, the stronger *C. albicans* lysis buffer (see 3.2 Methods) was used for both experiments.

Cell Lysis, Reverse Transcription, cDNA Amplification and Next era Library Generation for mDrop-seq

After droplet collection, the 50 mL tubes were transferred to a 37 °C water bath for zymolyase digestion and lysis for ~20 min; different lysis incubation times ranging from 10–25 min were tested. After the incubation, the Drop-seq protocol was followed for breaking droplets, collecting barcode beads with mRNA hybridized onto them and washing them in 6x Saline-Sodium Citrate (Teknova, #S0282) [9]. Reverse transcription was performed in 1.5 mL microfuge tubes under end-over-end rotation using a modified Reverse Transcription mix (1x Maxima H- RT buffer, 4% Ficoll PM-400 (GE Healthcare, #17-0300-05), 3 mM $MgCl_2$ (Sigma, #M1028), 1 M Betaine (Sigma, #14300), 1 mM dNTPs (Clontech, #639125), 1 U/μL Rnase Inhibitor (Lucigen, #F83923), 2.5 μM Template-Switching Oligo primer, (Integrated DNA Technologies, AAGCAGTGGTATCAACGCAGAGTGAATrGrGrG), and 10 U/μL Maxima H-RT enzyme (ThermoScientific, #EP0751) with a 30 min incubation at room temperature, followed by a 90 min incubation at 50 °C. This generated barcoded cDNA affixed to the barcoded beads referred to as Single Transcriptome Attached to MicroParticles or STAMPs. The beads were then washed once in 0.5% SDS (Teknova, #S0288), twice in 0.02% Tween 20 (Teknova, #T0710) both prepared in TE buffer (Teknova, #T0228) and treated with Exonuclease I digestion (Fisher, #M0293L). The total number of STAMPS collected was counted manually under the microscope. cDNA amplification was performed on RNA-DNA conjugates attached to ~120,000 barcode beads in a 96-well plate (Genesee Scientific, #24-302) loaded at 10,000 STAMPs per well. The STAMPS were amplified for 17 PCR cycles, using Kapa Hifi Hotstart 2x Mastermix (Fisher, #NC0465187) and Drop-seq PCR primer (Integrated DNA Technologies, AAGCAGTGGTATCAACGCAGAGT) [9]. Post-PCR cleanup was performed by removing the

STAMPs and pooling the supernatant from the wells together into a single 1.7 mL low-retention tube (Genesee Scientific, #22-281LR) along with 0.6X Ampure XP beads (Beckman Coulter, #A63880) [9]. After adding the Ampure beads to the PCR product, the tube was incubated at room temperature for 2 min on a thermomixer (Eppendorf Thermomixer C, #5382000023) set to 1250 rpm, and for another 2 min on bench for stationary incubation. Next, the tube was placed on a magnet and washed 4X times using 1 mL ethanol (Sigma, #E7023) at 80% concentration in each wash. cDNA was eluted in ultra-pure water (Life Tech, #10977-023) at 2.5 µL/well and the concentration and library size were measured using Qubit 3 fluorometer (Thermo Fisher) and BioAnalyzer High Sensitivity Chip (Agilent, #5067-4626).

500 pg of the cDNA library was used in Nextera library (Illumina, #FC-131-1096) preparation, following the original Drop-seq protocol, with a 3 min 72 °C incubation step added at the beginning of the thermocycling program to yield Nextera libraries averaging 500–700 bp [1,23].

Population-Level RNA-seq Library Preparation of S. cerevisiae

We performed the same standardized procedure to prepare bulk RNA-seq libraries for *S. cerevisiae* as follows: We lysed ~8,000,000 cells each using DYLB and incubation at 37 ºC for 20 min. Total RNA was extracted from each lysate using the Direct-zol RNA Miniprep Plus kit (Zymo Research, #R2071), assessed RNA quality using Qubit HS RNA Assay (Invitrogen, #Q32852) and diluted to 25 pg/µL. The total RNA library was annealed to 5′-AAGCAGTGGTATCAACGCAGAGTACTTTTTTTTTTTTTTTTTTTTTTTTTTTN-3′ primer (Integrated DNA Technologies) that allow polyA selection of mRNA and template switching, similar to Drop-seq. Briefly, 11 µL of total RNA library was mixed with 11 µL of 10 µM primer above, 11 µL of 10 mM dNTP (Takara, #639125), 13.75 µL of ultra-pure water and

2.75 µL of RNase Inhibitor (Fisher, #NC1081844), and incubated at 75 °C for 3 min on a PCR thermocycler. A reverse transcription master-mix consisting of 11 µL of 5X Maxima RT buffer, 2.2 µL of H_2O, 11 µL of 5 M Betaine (Sigma, #14300-500G), 1.65 µL of 100 mM $MgCl_2$ (Sigma, #M1028), 2.2 µL of 50 µM Drop-seq TSO primer (Integrated DNA Technologies, AAGCAGTGGTATCAACGCAGAGTGAATrGrGrG), 1.1 µL of RNase inhibitor (Fisher, #NC1081844) and 2.75 µL of Maxima H-RTase enzyme (Fisher, #FEREP0753) was added immediately after the annealing step. The final RT reaction volume of 45 µL was pipetted several times to mix, centrifuged briefly to spin down the contents and incubated on a PCR thermocycler using the following program: 42 °C for 90 min; 5 cycles*(42 °C for 2 min, 50 °C for 2 min); 70 °C for 15 min, to perform reverse transcription of the polyadenylated mRNA selectively annealed to the primer above. The cDNA was amplified for 12 cycles using Drop-seq PCR primer (Integrated DNA Technologies, AAGCAGTGGTATCAACGCAGAGT) and KAPA HiFi HotStart ReadyMix PCR Kit (Fisher, #NC0465187K). Amplified cDNA was quantified using Qubit and BioAnalyzer, followed by Nextera library (Illumina, #FC-131-1096) generation.

Sequencing

Nextera libraries of samples, including bulk RNA-seq, were loaded at ~ 15 pM concentration and sequenced on an Illumina NextSeq 500 using the 75 cycle v3 kits for paired-end sequencing. A total of 20 bp were sequenced for Read 1, 60 bp for Read 2 using Custom Read 1 primer, GCCTGTCCGCGGAAGCAGTGGTATCAACGCAGAGTAC (Integrated DNA Technologies), according to protocol. Due to the low complexity of yeast cDNA libraries, Illumina PhiX Control v3 Library (Illumina, #FC-110-3001) was added at 5–10% of the total loading concentration for all sequencing runs. Samples for each experiment were loaded at 7–15% of a NextSeq 500 lane and yielded 10–40 million reads for each sample. Some samples

were sequenced twice, depending on library complexity. For these samples, the Fastq files were concatenated using the UNIX zcat function before running the *dropRunner* pipeline. Data are available at: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE154515

mDrop-seq Data Preprocessing, Alignment and Quality Control

There were ~5000–10,000 cells processed per sample and each library was sequenced at ~40–90 million reads. We developed a Snakemake [59] protocol called dropRunner that takes paired-end reads in FASTQ format as input and produces an expression matrix corresponding to the UMI of each gene in each cell. The protocol initially performs FastQC [60] to obtain a report of read quality. The reports are later inspected manually to ensure high-quality reads were generated. Next, it creates a whitelist of cell barcodes using umi_tools [61] 0.5.3, which is a list of the top 10,000 valid cell barcodes in terms of number of reads. Next, reads were aligned to the respective yeast genomes using STAR [62] 2.7.0a. STAR 2.7 introduced STARsolo, a turnkey solution for processing droplet single-cell RNA-seq data built directly into the STAR aligner. The whitelist and paired-end reads are used as input for STARsolo, which performs alignment, gene UMI counting, and cell-barcode-filtering in one step. STARsolo uses a heuristic approach for filtering cells with low or noise-level UMI counts. It does so by constructing a UMI count rank plot for each cell (a knee-plot) and picks a cut-off based on the knee of the curve. The pipeline can be found at GitHub (aselewa/dropseqrunner). The filtered digital expression matrices from STARsolo were loaded in Seurat (v3.1.1), an R package for downstream single-cell transcriptome analyses.

All data from *S. cerevisiae* were aligned to the *Saccharomyces cerevisiae* reference genome, version sacCer3_s288c (https://www.yeastgenome.org/strain/S288C) obtained from the *Saccharomyces* Genome Database (SGD).

Bulk RNA-seq Data Processing

Bulk RNA-seq data obtained from *S. cerevisiae* was assessed for read quality using *FastQC*, mapped to the respective genomes described above using *STAR* 2.7.0a aligner [62], and RNA counts were generated from the bam files using *FeatureCounts* [63]. Read lengths were down-sampled during alignment using the *STAR* aligner, "-clip3pNbases" and "-clip5pNbases" parameters. Count matrices were compared to the mDrop-seq datasets using the Seurat package.

Clustering Cells and Generating UMAP

We followed the analysis pipeline recommended by Seurat. Data were normalized and scaled using default commands provided by the Seurat package in R. Seurat was used to calculate the gene dispersion and mean expression to find highly variable genes. This reduces the computational time of PCA compared to using the full set of genes. Highly variable genes were used to calculate the PCs for the yeast mDrop-seq data. An elbow plot displaying the variation explained by each PC was used to determine the number of PCs needed to explain the majority (>90%) of the variation. The top PCs determined in this way were used to perform clustering which was visualized with Uniform Manifold Approximation and Projection (UMAP) [64].

Calculating Doublet Rates from Species Mixing Experiments

An mDrop-seq dataset containing a mix of *S. cerevisiae* and *C. albicans* cells was aligned once to the *S. cerevisiae* genome and again separately, to the *C. albicans* genome (see 3.2 Methods for *C. albicans*). Cells were removed based on data quality. The 7% Poisson loading experiment had gene cutoffs of 50 for both experiments, while the 15% Poisson loading experiment had cutoffs of 275 and 400 for *S. cerevisiae* and *C. albicans,* respectively. We identify cell barcodes that capture genes from both *S. cerevisiae* and *C. albicans* genomes. Due

to the similarities between the yeast genomes caused by shared ancestry, there are genes that will map to both species. This was checked by taking a *C. albicans* dataset and mapping it to the *S. cerevisiae* genome, and vice versa, to identify the genes common to both species. After removing these common genes from the mixed-species dataset, we identified the cell barcodes that show significant mapping (>50 genes) to both genomes as true doublets.

Batch Correction

Dataset merging and batch correction were performed using the *Anchor Integration* function in the Seurat package. Datasets were independently normalized and highly variable genes were detected using gene dispersion and mean expression. The datasets were scaled before running the *Canonical Correlation Analysis* (CCA) function in Seurat to determine dataset anchors and merge the objects. A new integrated dataset was then created using the detected anchors. This integrated dataset was used for dimensional reduction and clustering analyses.

Cell Cycle Analysis

Lists of genes that serve as cell cycle phase markers for G1, S, and G2M phases were obtained from Spellman,[65] for *S. cerevisiae*. Cell cycle assignment was made based on the G1, S, and G2M markers for cell cycle phases. The *AddModuleScore* function implemented in Seurat was used to calculate cell cycle scores for each phase. This function sampled random genes as a control set; the number of the genes in the control set was determined by the number of markers in the cell cycle gene lists. Since there were three such lists of markers, the minimum size of cell cycle marker lists was used. For each cell, the largest module score was selected among the three phases and the corresponding cell cycle phase for the selected module score was assigned to the cell. A threshold on the selected module score was applied to ensure the cell cycle assignment

was robust. If all module scores for a cell were below zero, the phase for the cell was left undecided and counted as “Not Assigned” or NA in Supplementary Table 2.2.

This module score list was used with Seurat to create a column in the object metadata containing the assigned cell cycle phase. Cell cycle phase metadata was used to calculate PCs instead of highly variable genes. Cell cycle variation was regressed out during data scaling and centering. The expression percentage and level for the G1, S, and G2M marker genes were visualized using dot plots where the size of the dots indicates the percentage of cells expressing a marker and the color intensity reflects normalized expression.

Hierarchical Clustering of Cells

To investigate variations in single-cell expression that are not related to cell division and proliferation, the cell cycle effects were removed when scaling the gene expression per cell. The module scores for each of the G1, S, and G2M phases were regressed out against each gene. PCs and Shared Nearest Neighbor (SNN) graph [66] were constructed from the scaled gene expression matrix.

Trajectory Analysis on Single-Cell Data

To trace the lineage or process of temporal activation in yeast cells in response to heat shock, trajectory analysis [67] was performed on the single-cell data from control and stimulated cells. Datasets from different conditions/time points were integrated using anchor-based integration described above. The R-based pipeline, Monocle 2 (v 2.10) was used to process the data and construct the trajectory.

The genes used to order cells along the pseudo-time trajectory, or ordering genes were set based on differentially expressed (DE) genes obtained from unsupervised clustering in Seurat.

DE genes with q-value <0.01 were selected as the ordering genes in Monocle. The count matrix was log-transformed after adding one to the counts to eliminate logarithms of zero values. PCA was performed on the normalized count matrix using the ordering genes and the top variable PCs were selected based on the scree plot of variance explained per component. The PCs were reduced into a tree structure by Discriminative Dimensionality Reduction with Trees (DDRTree) [68]. The backbone of the tree branches formed the cell trajectory. The root of the trajectory was set as the tip of the tree branch that contained the largest number of cells from the control sample, and each cell was ordered in pseudo-time based on the root. During the PCA and DDRTree dimension reduction phases, we removed any cell cycle effects by specifying the G1, S, and G2M module scores obtained from Seurat as variables to be linearly subtracted from the data to look for changes in gene expression that are independent of cell cycle effects as response of external stimuli.

2.3 Results:

Optimizing Single Yeast Cell Lysis in Droplets

In order to develop mDrop-seq as a protocol for the scRNA-seq of yeast, we first had to make our microfluidic setup compatible with yeast. mDrop-seq requires cell lysis after single cells are encapsulated in droplets. We developed targeted in-drop lysis protocols for each yeast species that consist of cocktails of zymolyase and sarkosyl with an incubation at 37 ºC. Sarkosyl is a strong detergent used for cell lysis [69]; zymolyase is an enzyme mixture with optimal activity at 37 ºC, commonly used to digest yeast cell wall [70]. We note that zymolyase or sarkosyl alone were insufficient to obtain total lysis of yeast cells in drops. Because droplet stability is influenced by constituent fluids and flow parameters, we optimized the Droplet Yeast Lysis Buffer (DYLB) composition with oil-surfactant mix, droplet size and flow rates for each

yeast species. Monodisperse droplets of 75 μm diameter that are stable under flow and thermal incubation were used for mDrop-seq experiments. The scRNA-seq workflow for yeasts in microfluidic drops is shown in Figure 2.1A. To establish a high-throughput scRNA-seq method for yeast cells such as *S. cerevisiae* and *C. albicans*, we adapted the emulsion droplet and bead-based DNA barcoding scheme used in Drop-seq [9] and DroNc-seq [58]. To determine the duration of droplet incubation for optimal zymolyase activity needed for each species, a single collection of mDrop-seq droplets on *S. cerevisiae* was split into three different pools and each pool was incubated in a 37 °C water bath for 10, 15, and 20 min, guided by bulk lysis experiments. The droplets were inspected by optical microscopy following incubation to ensure droplet stability and cell lysis.

The 10 min lysis incubation yielded a lower quantity of cDNA, indicating incomplete lysis and was excluded from further analysis. The 15- and 20-min incubations generated two libraries, indicated as SC_15min_Rep1 and SC_20min_Rep1. There were ~5000–10,000 cells processed per sample. Figure 2.1B shows the number of features (left), as a proxy for genes, and Unique Molecular Identifiers or UMI (right), as a proxy for a number of mRNA molecules captured per cell barcode. The UMI identifies individual mRNA molecules detected, allowing correction for PCR replicates to prevent PCR bias. Supplementary Table 2.1 summarizes the total and average number of reads from each sample, number of cells, mean number of genes and UMI obtained from a single cell, and the total number of unique genes obtained per experiment. We filtered out cell barcodes with less than 200 or more than 2000 genes detected, as they likely represent poor quality cells or multiple cells loaded in a single droplet, respectively. Across all experiments, we saw only 30–35% more UMI compared to the number of genes detected [71]. This experiment was performed twice.

The two mDrop-seq datasets from the lysis trial, SC_15min_Rep1, SC_20min_Rep1 overlap in principal component (PC) space (Figures 2.1C), with a Pearson correlation >0.99 between the datasets for UMI counts. This allowed us to combine the SC_15min_Rep1, SC_20min_Rep1 datasets into a single dataset of 12,012 cells, which we used to systematically explore differential expression (DE) in *S. cerevisiae* genes. Figure 2.1D shows the combined dataset in Uniform Manifold Approximation and Projection (UMAP) space [64]. The genes with the highest expression were involved in glycolysis (*TDH3*, *ENO2*, *FBA1*), stress response (*HSC82*), or ribosome biogenesis (*RPP2B*) (Figure 2.1E). As these cells were grown in a 2% glucose medium and lysed in log phase, the presence of glycolysis genes is expected. We also saw the expression of general stress response genes common to heat shock, along with DNA replication stress and oxidative stress that we attribute to the yeast cells responding to stresses (e.g., enzymatic lysis, heat) during cell lysis. We also observed a stress-induced batch effect within our biological replicates.

To compare mDrop-seq with bulk RNA-seq, we also performed population-level RNA-seq on *S. cerevisiae*. A pseudo-bulk [9] comparison of the mDrop-seq dataset showed a Pearson's correlation of 0.85 with our bulk RNA-seq, and an overall good correlation (~0.8, on average), with public RNA-seq datasets [72]. The correlation in transcript counts between our bulk RNA-seq and public datasets was 0.85, comparable to mDrop-seq.

Single-Cell Specificity of mDrop-seq Confirmed by Species Mixing Experiments

To establish that mDrop-seq has single-cell specificity, species-mixing experiments [9] were performed using a mix of *S. cerevisiae* and *C. albicans* cells. Species-mixing experiments allow checking for "doublets", or cell barcodes that capture two cells, assuming that a fraction of such drops will capture both *S. cerevisiae* and *C. albicans* cells and the corresponding cell

barcodes will align to both *S. cerevisiae* and *C. albicans* genomes. The probability of finding one or more cells in a single drop is estimated assuming a uniform concentration of cells in the loading medium following a Poisson distribution. The Poisson parameter (λ) governs the cell distribution in drops and may be used to calculate the fraction of cell doublets. in the experiment. Two different cell-loading conditions were tested at 350,000 and 700,000 cells/mL, representing Poisson loading parameter, $\lambda = 0.077$ and 0.15, respectively. Across 4,120 cells detected at $\lambda = 0.077$ or 7.7% Poisson loading, the majority of the barcodes map to one genome only, with only 47 barcodes mapping to both *S. cerevisiae* and *C. albicans* genomes (Figure 2.1F, left). Assuming that cross-species doublets make up half of all doublets (a similar number of barcodes would contain two *S. cerevisiae* cells or two *C. albicans* cells), we estimate the doublet rate to be ~2.3% of all barcodes detected. Figure 2.1F, right, shows a similar species-mixing experiment

Figure 2.1: mDrop-seq of *Saccharomyces cerevisiae* cells

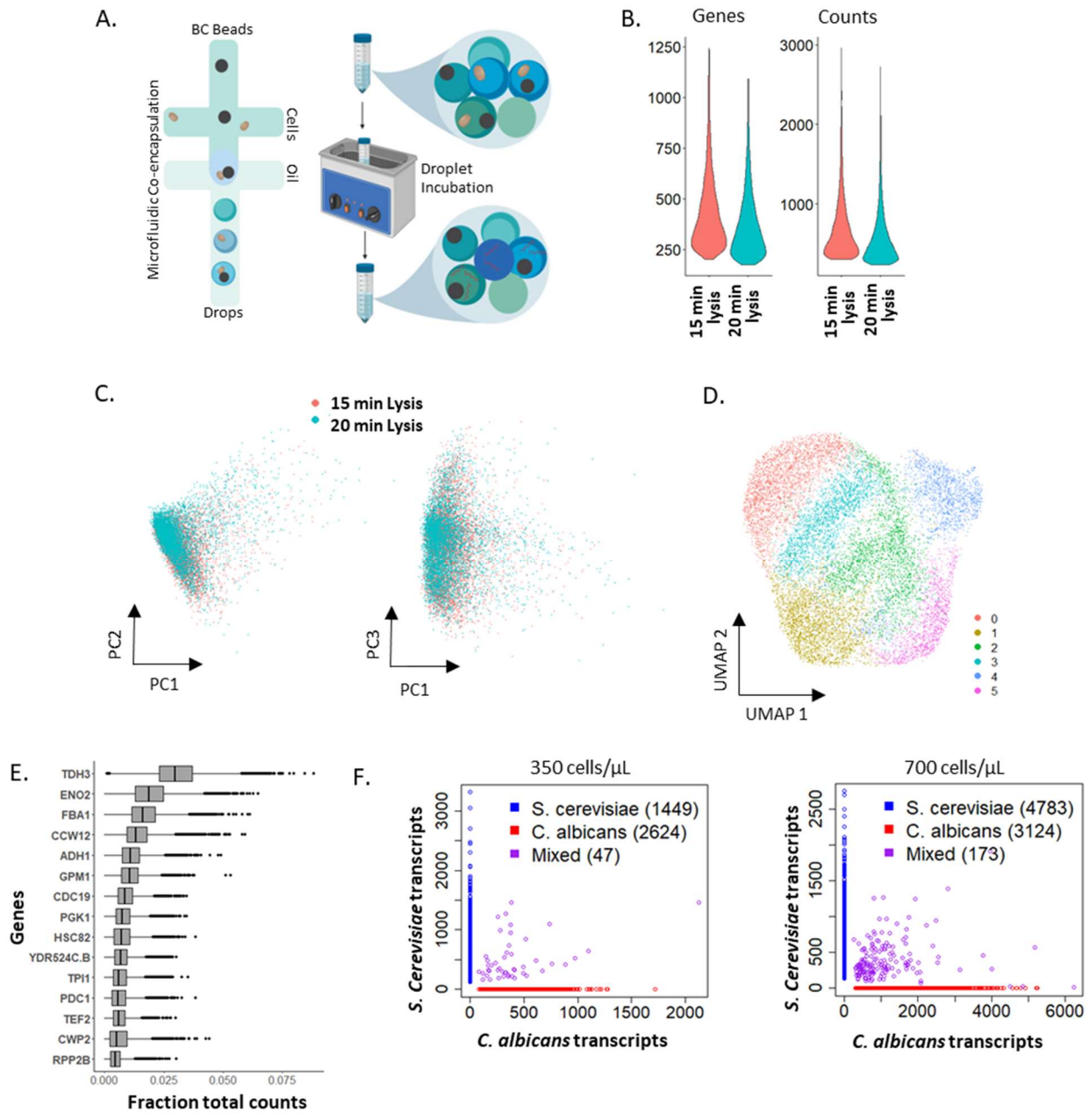


Figure 2.1: Legend

A. mDrop-seq experimental schematic.
B. Number of genes and UMI detected for each cell at two cell lysis times: 15 and 20 minutes.
C. Plots of PCs 1-3 displaying the overlapping datasets.
D. UMAP visualizing the clustering on 12,012 *S. cerevisiae* cells.

Figure 2.1: Legend continued

E. Boxplot displaying the top 15 genes expressed by fraction of total counts for 15 and 20 min lysis times.
F. Species-mixing plots where each dot depicts a unique cellular barcode that align to *S. cerevisiae* (blue), *C. albicans* (red), or both genomes (purple). Two Poisson loading concentrations of 350 cells/μL (left; $\lambda = 0.08$) and 700 cells/μL (right; $\lambda = 0.15$) are tested.

at $\lambda = 0.15$ or 15% Poisson loading; across 8080 cells detected, only 173 cells map to both genomes, giving a final doublet rate of ~4.2%. A Poisson loading of 15% was used for all *S. cerevisiae* and *C. albicans* experiments. These species mixing experiments also demonstrate that mDrop-seq is capable of simultaneous single-cell profiling of a mix of different yeast species. This feature may be useful in characterizing a fungal microbiome composed of multiple yeast species without the need to sort cells by species.

Cell Cycle Assignment and Analysis of *S. cerevisiae* cells

In an unstimulated dataset sampling continuously cycling cells, one may expect to find variation attributable to cycling cells. Using cell cycle gene lists provided in by Spellman *et al.* [65], we assigned a unique G1, S, or G2/M stage of the cell cycle to each *S. cerevisiae* cell that passed our quality filter. Due to low transcript counts inherent in yeasts, we were able to assign G1, S and G2M phases to only a subset of cells in each experiment (Supplementary Table 2.2). The remaining cells were designated as 'not assigned' or NA. Cells in the same stage of cell cycle largely grouped together in UMAP space (Figure 2.2A). Cell cycle genes in the S and G2M phases appear enriched in the upper left clusters (clusters 0 and 3 in Figure 2.1D, respectively) while G1 and additional G2M cells are spread across the lower clusters (clusters 1, 2, and 5 in Figure 2.1D). All cell cycle phases are represented in the upper right cluster (corresponding to cluster 4 in Figure 2.1D). Figure 2.2B shows six histone genes enriched in cells assigned to S

phase, as expected. After regressing out cell cycle effects, cluster 2 in Figure 2.2C shows increased expression of stress response genes, *DDR2, NCE103, HSP12, TMA10* (shown as feature plots in Figure 2.2D), likely induced during cell lysis. Figure 2.2E shows expression of genes *AIM44* and *PIR1* associated with GPI-anchored cell wall proteins that cluster together, independently of the cell cycle.

Figure 2.2: Cell cycle analysis of 12,012 *Saccharomyces cerevisiae* cells, Replicate 1

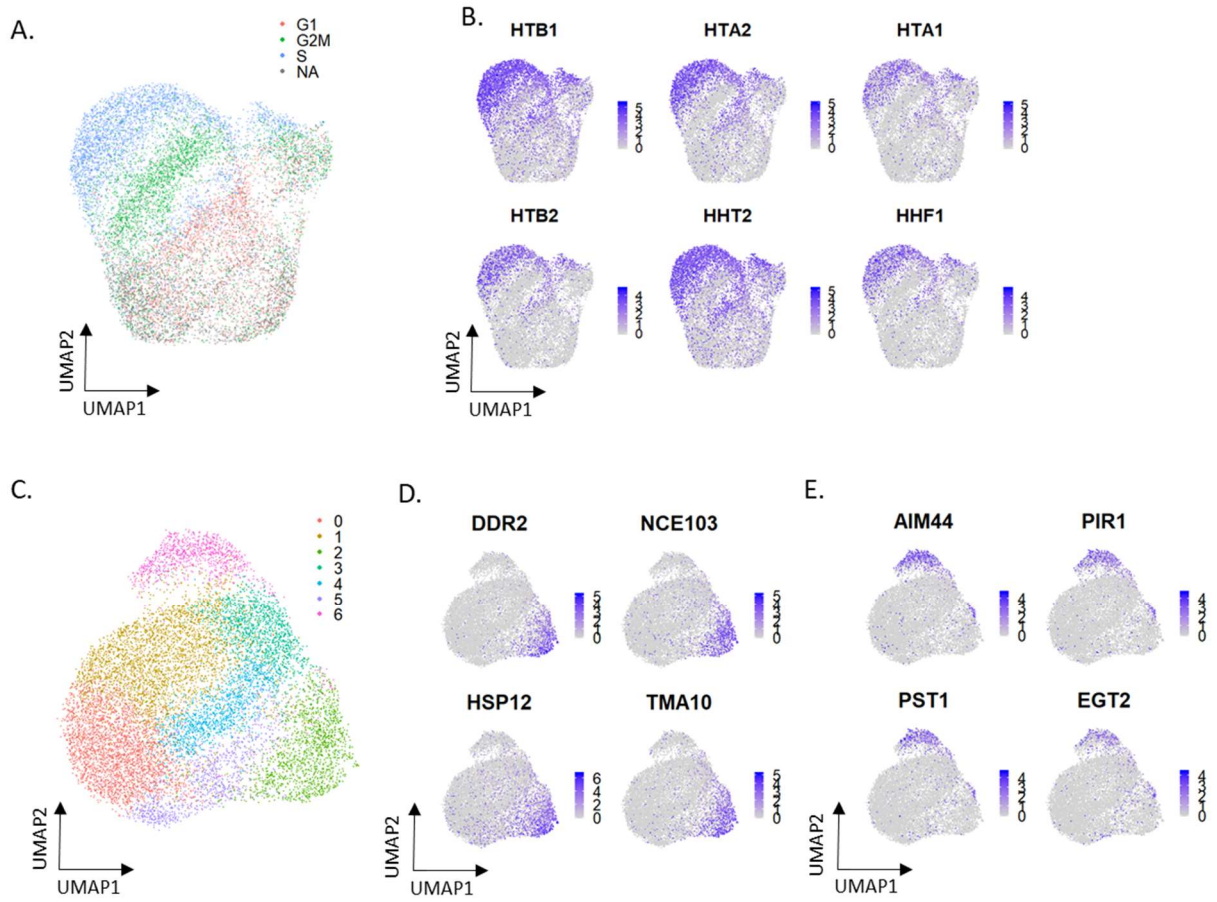


Figure 2.2: Legend

A. UMAP of *S. cerevisiae* labeled by cell cycle stages. Cells that could not be assigned to a cell cycle phase are marked as NA in D, E.
B. Feature plots of histone genes enriched in S phase cells.

Figure 2.2: Legend continued

C. UMAP plots of cells after cell cycle regression, labeled by clusters obtained by unsupervised clustering.
D. *Stress response genes DDR2, NCE103, HSP12, and TMA10 forming a separate cluster (cluster 2) after cell cycle regression.*
E. *Feature plots of AIM44, PIR1, PST1, and EGT2 for cell wall and budding marker expression that cluster separately (cluster 5) after cell cycle regression.*

A replicate of the lysis experiment shows signs of a stress response, requiring batch correction

The control experiment was repeated with 15 and 20 min incubation times for lysis, labeled SC_15min_Rep2 and SC_20min_Rep2, respectively. For both datasets, we obtained slightly lower numbers of cells, and comparable numbers of genes and UMI per cell within the datasets (Supplementary Table 2.1 and Figure 2.3A). Overall, the two lysis times in replicate 2 yielded comparable data with Pearson correlation of 0.98. The two replicates, SC_XXmin_Rep1 and SC_XXmin_rep2 (Pearson correlation = 0.86) show overlapping but slightly different sets of differentially expressed genes and some differences in their expression profiles. Several genes that are more highly expressed in replicate 2 included *HSP12*, *HSP42*, *DDR2*, and *NCE103* (Figure 2.3B), many of which involve stress response. Figure 2.3C shows the 15 highest expressed genes in replicate 2, with *HSP82*, *HSP12*, and *HOR7* being among the top 4 genes expressed. This, along with SC_20min_Rep2 having slightly higher genes and UMI detected compared to SC_15min_Rep2, may indicate that this replicate underwent longer lysis incubation or otherwise experienced stress.

A PC plot combining the two replicates with two lysis incubation times each into a single dataset of 20,548 cells (Figure 2.3D) shows clear separation between SC_XXmin_Rep1 and SC_XXmin_Rep2. These replicates were performed on different days and may have sampled *S. cerevisiae* cells from slightly different growth phases during culture. Nevertheless, we were able

Figure 2.3: Comparison of replicates for different *S. cerevisiae* lysis incubation time experiments

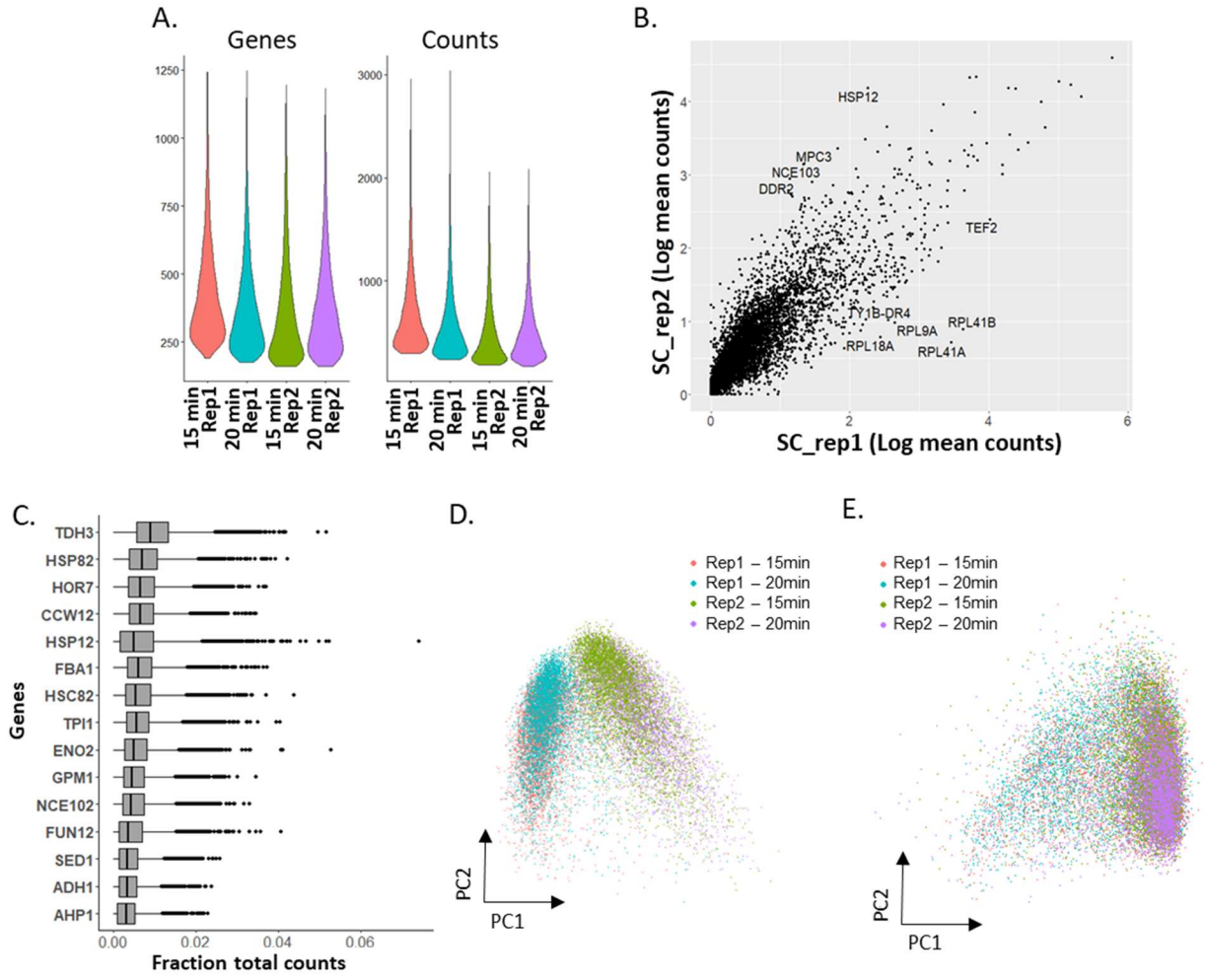


Figure 2.3: Legend

A. The number of genes and UMI detected in 4 datasets.
B. Plot of the average expression of genes in 22,491 *S. cerevisiae* cells from the control replicates.
C. Top 15 genes expressed in the second replicate dataset.
D. Plots of PC 1 and 2 displaying the four datasets without (D) and with (E) batch correction

to adjust for these differences computationally (Figure 2.3E) using anchor based integration in Seurat [73]. Because of the noted increase in stress response in the second replicate, we compared the second "stressed" control replicate as experiencing intermediate levels of heat shock related stress between the *S. cerevisiae* replicate 1 and the heat shock experiments. Indeed,

when comparing expression levels, Supplemental Figure 2.1C shows increased overall expression of *HSP12*, *HSP26*, and *SSA4* in replicate 2 or the intermediate "stressed" control, though lower than expression in the heat shock data (logFC = 1.99, 2.04, 2.29, respectively). Housekeeping genes show similar expression across control (SC_XX_Rep1), intermediate (SC_XX_Rep2) and heat shock data (Supplemental Figure 2.1D).

Activation of Stress Response in *S. cerevisiae* under Heat Shock

To test the efficacy of mDrop-seq in detecting transcriptional changes in yeast cells to stimuli, *S. cerevisiae* cells were subjected to heat shock prior to running mDrop-seq. Heat shock is a widely conserved response in cells that involves the expression of protein chaperones [74]. The cells underwent a 20 min heat shock at 42 °C and then were chilled on ice before running mDrop-seq. The *S. cerevisiae* dataset in Figure 2.1D was used as the control for the heat shock experiments.

Two replicate mDrop-seq experiments on heat shock stimulation of *S. cerevisiae* cells showed significant similarity in expression (Pearson correlation of 0.96). The second replicate showed a slight increase in genes and UMI detected (587 and 1053, respectively) in Figure 2.4A. The heat shock replicate experiments were performed on different days, using independent cultures grown from the same stock. Compared to the control, we see the upregulation of several stress response genes (the mean expression of the control and heat shock data showed a Pearson correlation of 0.74 in Figure 2.4B). We ordered the genes in the control and heat shock datasets by descending log fold change (logFC). For many of these genes, the p-values are small (Wilcoxon Rank Sum test, adj-p < 1e-10, Bonferroni correction) with many adj-p < 2.225e-308 (lowest value reported in Seurat). Among the genes induced under heat shock, we see several genes involved in heat shock related stress response, such as *HSP12* (logFC = 1.78) as well as

other heat shock protein (HSP) family genes associated with other types of stress response (Figure 2.4C, Supplemental Figure 2.1C). We also see significant differential expression in genes marked for protein transport, such as *KAR2* (logFC = 2.49; Supplemental Figure 2.1A) and DNA replication stress, such as *RTC3*, *NCE103* (logFC = 2.68, 2.51, respectively; Supplemental Figure 2.1B) under heat shock. Housekeeping genes, such as *ACT1*, histone gene *HTB1*, ribosomal genes, and glycolysis genes (Figure 2.4D, Supplemental Figure 2.1D) are present in both datasets, with slightly higher expression (logFC = 0.6, 1.1, and 1.5 for *HTB1, ACT1,* and *RPP2B,* respectively) in the control dataset. These results suggest that mDrop-seq has the power to detect cellular responses to stimuli on a single-cell level.

To investigate the sequence of activation in stress response genes in *S. cerevisiae* under heat shock, we applied trajectory analysis [67] on a subset of control and heat shocked *S. cerevisiae* data. Three datasets, SC_15min_Rep2, SC_20min_Rep2 and SC_HeatShock_Rep2 were integrated (Supplemental Figures 2.2A, B) and cell cycle module scores were assigned to each cell. We then used Monocle v2 [67] to infer the expression changes during heat shock in pseudo-time (Figures 2.4E–G, with Supplemental Figure 2.2C showing the corresponding pseudo-time ordering of cells). Figure 2.4E shows the trajectory where the control samples, SC_15min_Rep2 and SC_20min_Rep2 overlap with each other, while the heat shock sample SC_HeatShock_Rep2 diverge into two separate branches indicative of two distinct pseudo-states. When marked by each cell's cell cycle phase (Figure 2.4F), we note that both branches of the trajectory taken by the heat shocked sample are dominated by S phase cells. In contrast, the G1, S and G2M phases largely overlap for the control samples. Next, we compared the cell-type clustering results with trajectory analysis. Figure 2.4G shows the pseudo-time trajectory where each cell is colored according to the unsupervised cluster it belongs to, shown in Supplemental Figure 2.2A. The

Figure 2.4: Heat shock treatment of 26,019 *Saccharomyces cerevisiae* cells profiled using mDrop-seq

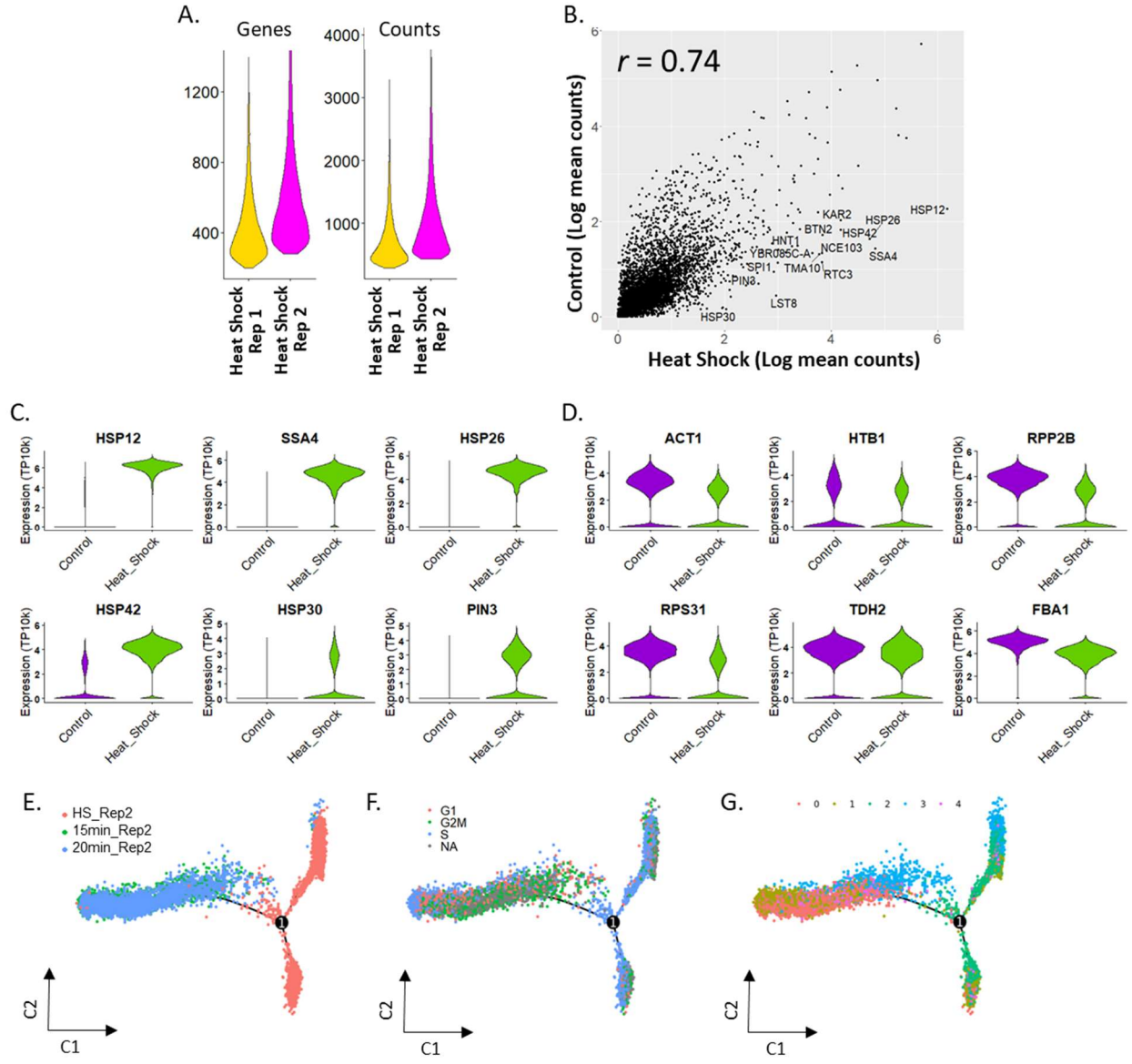


Figure 2.4: Legend

- **A.** Violin plots displaying the number of genes and UMI for heat-shock replicates.
- **B.** Correlation between average gene expression values for the control and heat-shocked *S. cerevisiae* datasets.
- **C.** Violin plots displaying expression differences in control and heat-shocked datasets for heat-shock related genes
- **D.** Violin plots displaying expression differences in control and heat-shocked datasets "house-keeping" (actin, histones, ribosomal, and glycolysis) genes.

Figure 2.4: Legend continued

E. Pseudo-time trajectory of gene expression inferred from the combined control and heat shocked *S. cerevisiae* dataset shown in this figure. Colors indicate E) experimental time points, F) Cell cycle stages (cells that could not be assigned to a cell cycle stage are marked NA), and G) cell-type clusters shown in Fig. S4A.

UMAPs in Supplemental Figure 2.2B show the contribution of each sample to the overall clustering: we note that cells in cluster 0 come mostly from the control samples (left, middle) and cluster 2 is composed of cells primarily from the heat shock sample (right). These results are consistent with Supplemental Figure 2.2D where cells in cluster 0 predominantly occur in control samples (left, middle) and cluster 2 in the heat shock sample (right). DE analysis was performed on the two branches of the heat shock sample using the Wilcoxon Rank Sum test. The expression level in the logarithmic scale was visualized on the trajectory tree plot (Supplemental Figure 2.2E). Heat shock genes, *HSC82*, *HSP12* and *HSP82* (Supplemental Figure 2.2E) are differentially expressed between the two branches of the trajectory indicating differential response in *S. cerevisiae* cells to heat shock.

2.4 Discussion:

As noted earlier, single-cell genomic analyses of microbial species have been difficult due to challenges in single-cell lysis and low input material in microbial cells. Yeasts and other fungi have poly-adenylated tails on the 3′ end of their mRNA, allowing selective mRNA capture using poly-dT oligonucleotides that is not possible in bacterial cells making fungi more experimentally tractable among microbial species.

We established the feasibility of mDrop-seq to profile transcriptional heterogeneity in fungal species at single-cell resolution and at scale by performing mDrop-seq on a total of 35,109 single cells of *S. cerevisiae* across multiple replicates, experimental conditions, and

environmental stimuli in the form of heat shock. Based on Drop-seq and DroNc-seq [9] used to profile gene expression in mammalian cells, mDrop-seq leverages existing single-cell experimental and computational tools and allows for the lower barrier of entry and easy adaption of single-cell RNA-seq on fungal species for labs that are set up for Drop-seq or similar workflows.

Modification of the Drop-seq protocol allows for microfluidic droplet stability and efficient lysis of *S. cerevisiae* cells

To implement mDrop-seq, we needed to overcome the challenge of microbial cell lysis in emulsion drops, while maintaining droplet stability and RNA integrity for downstream molecular biology reactions. This was accomplished through a combination of zymolyase and sarkosyl activity in drops, along with thermal incubation. The enzyme, zymolyase targets a common component of fungal cell walls and requires thermal activity; the detergent sarkosyl is a strong lytic agent that works ubiquitously on mammalian, zebrafish, *C. elegans* and fruit fly cells [45–47]. While we modified the Drop-seq lysis buffer beyond just zymolyase with additional sarkosyl, potassium phosphate, and β-mercaptoethanol for more efficient lysis (see 2.2 Methods), it has been shown by other similar methods that the Drop-seq lysis buffer along with just Zymolyase would be sufficient for *S. cerevisiae.*

The stability of emulsion drops and the efficacy of downstream reactions are affected by droplet contents. However, the DYLB designed for *S. cerevisiae* does not appear to have had large scale negative effects and microfluidics were performed at expected flow rates. Since reverse transcription in mDrop-seq was performed outside microfluidic drops after the emulsion was broken following single-cell lysis and mRNA capture on barcode beads in drops [9], [58], the compatibility of lysis buffer and reverse transcriptase was not an issue. The DYLB used in

our experiments were optimized for the microfluidic device [58], oil-surfactant mix, and flow parameters used here. While stable droplets may be generated at higher flow rates with other surfactants, we prefer the oil-surfactant mix used here due to its relatively low cost, long shelf-life, and easy availability.

S. cerevisiae libraries are compatible with current next generation sequencing

We used Illumina Paired End (PE) sequencing for mDrop-seq. Due to the relatively lower complexity of the yeast mDrop-seq libraries in general, or GC content bias (~41% in *S. cerevisiae*, reported by *FastQC*), we found it beneficial to sequence these libraries multiplexed with more complex libraries such as human (~45%, from *FastQC*) or use higher Illumina PhiX concentration to improve the overall quality of sequencing runs.

The 20 bp long Read 1 sequence was used to de-multiplex the cell barcode and UMI while the 60 bp Read 2 was used to identify the 3′ end of transcripts. While longer read lengths can help reduce multimapping in complex genomes such as the human at 3.1 Bb [78], yeast genomes are typically much smaller, e.g., 12 Mb for *S. Cerevisiae* [79] and, for later chapters, 14.7 Mb for *C. albicans* [80]; so transcripts with shorter read lengths (~30 bp) can be uniquely mapped to them. Figure S12A shows the percent of uniquely mapped reads (left) and reads mapping to multiple loci (right) as functions of Read2 fragment lengths for *S. cerevisiae* and human genomes. We also compared the effect of clipping the transcript fragments on the 3′ vs. 5′ ends on STAR [62] aligner and saw no noticeable difference in mapping rates between the two.

UMI identifies individual mRNA molecules, allowing us to collapse PCR replicates and prevent PCR bias. Across all mDrop-seq experiments, we saw ~30–120% more UMI compared to the number of genes detected, with higher percent UMIs detected in the heat shock stimulated

S cerevisiae, compared to controls. The majority of genes detected had only a single count of transcripts attributed to them. This is expected because most yeast genes are expressed as single copies of mRNA at any time [71]. This may make it difficult to differentiate between true variation and drop-outs in the data; a problem in scRNA-seq that is exacerbated in yeasts.

Analysis of cellular response to heat shock displays cellular transcriptomic response

When comparing the control and heat shock sample in *S. cerevisiae*, we see a very clear heat shock response and notable separation in PC and UMAP space. This separation is primarily driven by upregulation of known heat shock genes along with genes involved in DNA replication stress (e.g., *KAR2, LST8, ERO1)* and protein transport (e.g., *RTC3, NCE103, TMA10*). Using a pseudo-time trajectory analysis, we clearly see the progression from a non-stimulated control to the heat-stimulated cells, with a branch point indicating two separate heat shock responses. The most significant difference between these two branches is the differential expression of ribosomal structure (e.g., *RPL8B, RPL25, RPL36B*) vs. oxidative-reduction energetic processes (e.g., *PIG2, PCL8, GDB1*).

The *S. cerevisiae* data were normalized and batch-corrected to allow comparisons between experimental conditions, and categorized into three sub-groups: control, "stressed" control, and heat shock, for comparison. We find the "stressed" control group to be intermediate between the control and heat shock in that it showed an elevated stress-response signature for *HSP12*, *HSP26* and *HSP42* compared to the control set, but lower than the heat shock data. In addition, the heat shock data showed expression of genes related to heat shock stress, e.g., *HSP30*, and *PIN3*, DNA replication stress, e.g., *LST8*, *BTN2*, *ERO1*, and protein transport, e.g., *RTC3*, *TMA10* and *SPI1* that were absent in the control and "stressed" control samples. We saw similar levels of transcription for housekeeping genes, e.g., *ACT1, HTB1*, *TDH2*, and *FBA1*

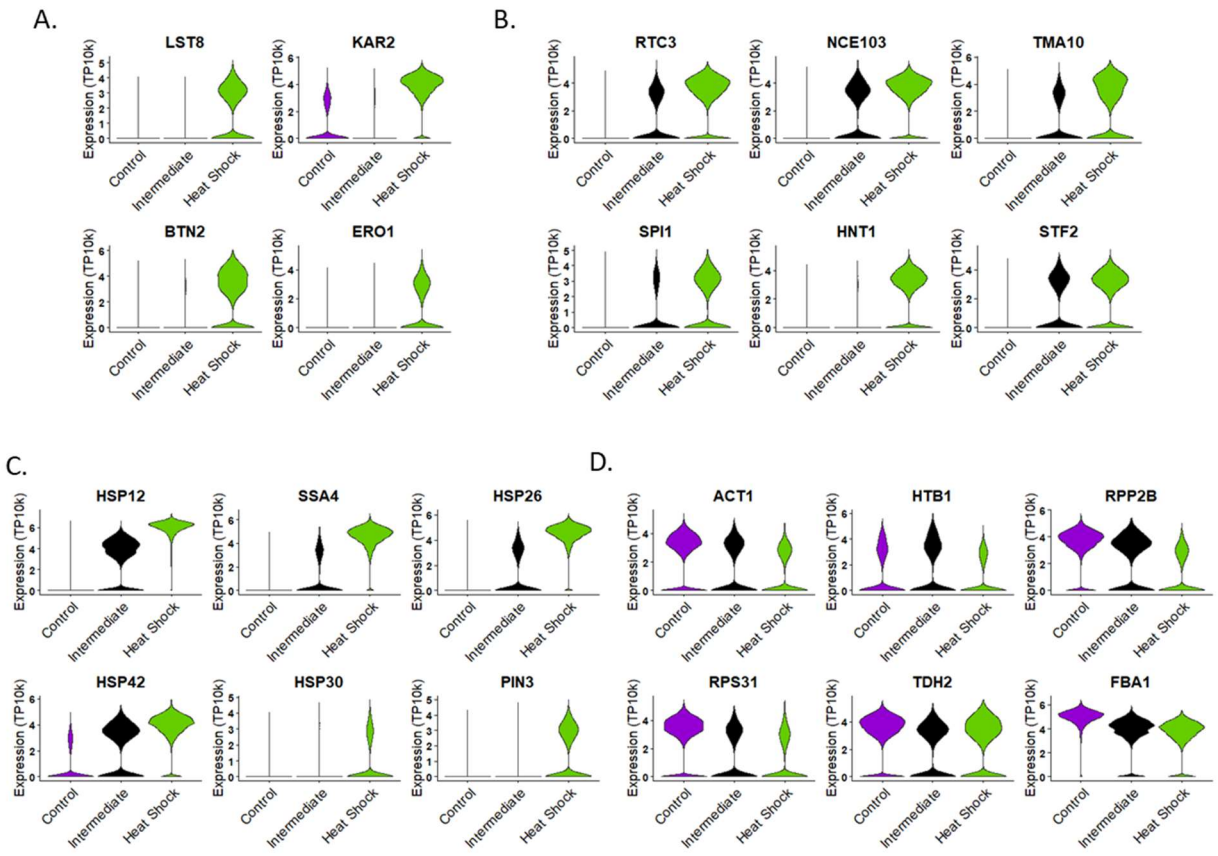
across all groups. Others of interest are genes such as *AIM44*, *PIR1*, *PST1*, and *EGT2*, associated with cell wall stability and cell budding that were expressed in a subset of cells clustering together in the control data.

Using gene lists specific to the G1, S and G2M phases to the cell cycle, we scored and assigned each cell to a unique cell cycle phase for both *S. cerevisiae*. Cells that could not be unambiguously assigned to any particular cell cycle phase were marked as NA. In heat shock, we see a significant decrease in the number of G2M phase cells (Supplementary Table 2.2) and an increase in cell numbers assigned to the S phase. Since fluconazole treatment may be expected to elicit a stress response in yeasts, we propose that yeast cells are possibly getting arrested in the S phase under stress. This would go against conventional understanding of the yeast cell cycle arresting during G1 phase during stress and may require more work to determine the cause of our consistent rise in S-phase cells.

2.5 Acknowledgement of Work Performed:

I would like to acknowledge those who contributed to the work presented here. Rebecca Back aided in the design and testing of the mDrop-seq protocol as well as the bulk RNA-seq of *S. cerevisiae*. Heather Eckart assisted in Nextera Library construction and sequencing submission. Project design originated with Dr. Anindita Basu. Dr. Alan Selewa and Dr. Bingping Xie assisted with alignment, QC, and data analysis. All other experimentation and analysis present in this chapter were performed by Ryan Dohn.

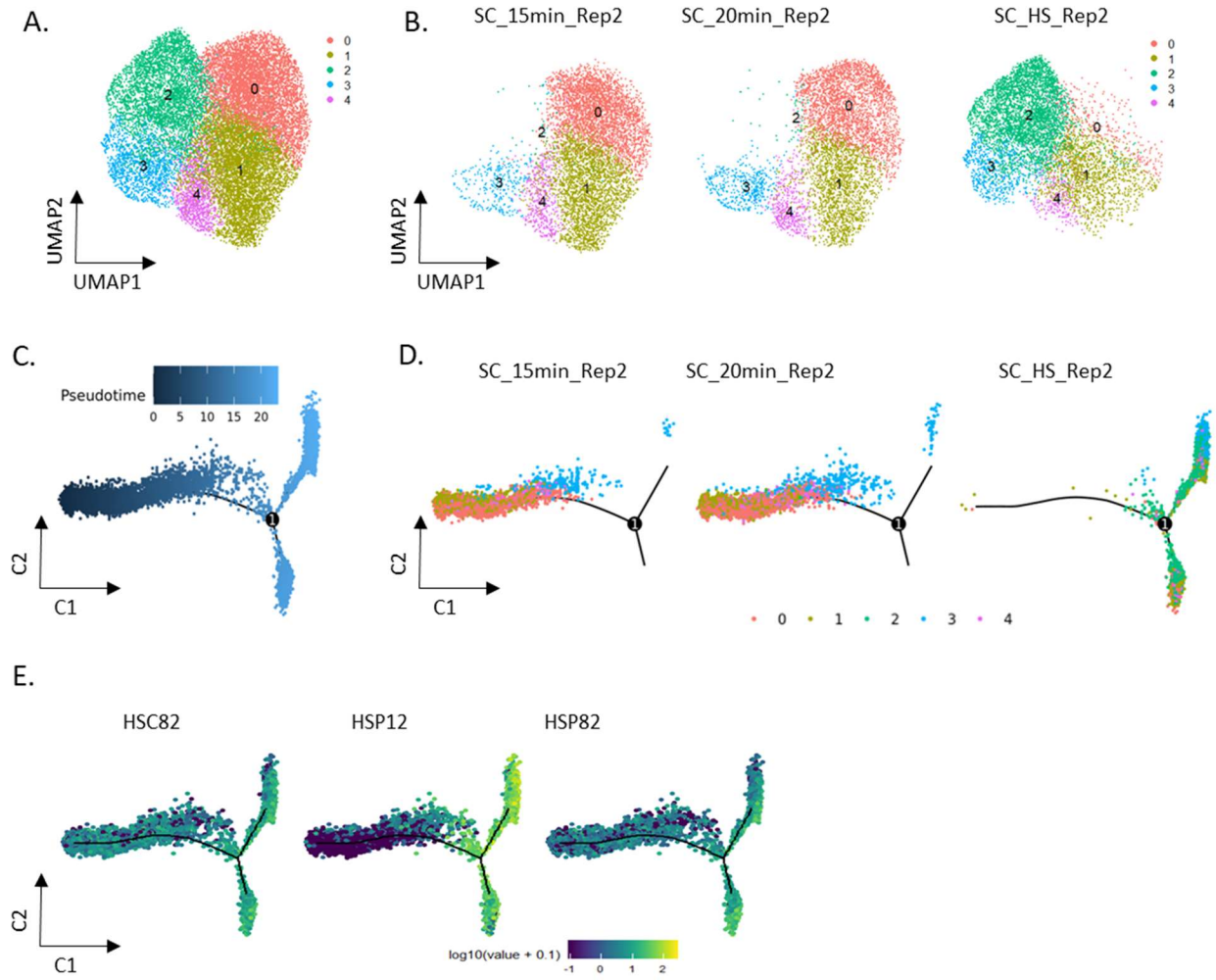
Supplemental Figure 2.1: Analysis of 35,109 *Saccharomyces cerevisiae* cells including 14,007 cells after heat shock and 21,102 cells as control



Supplemental Figure 2.1: Legend

A. Four genes for DNA replication stress show significant upregulation during heat shock when compared to the replicate 1 control.
B. Several genes involved in protein transport through the ER and Golgi show elevated expression during heat shock.
C. Violin plots displaying heat-shock genes that are progressively elevated in the second control replicate and heat-shock experiments.
D. Violin plots displaying housekeeping genes (actin, histones, ribosomes, and glycolysis) that appear in both heat shock and control replicate 2 data.

Supplemental Figure 2.2 Integration and pseudo-time trajectory analysis of *Saccharomyces cerevisiae* cells



Supplemental Figure 2.2: Legend

- **A.** UMAP plot of unsupervised clustering on integrated combined control (15 & 12 min lysis) and heat shocked *S. cerevisiae*, Replicate 2 data after the cell cycle effects are regressed out. Colors represent cell clusters also shown in Fig. 2H.
- **B.** UMAP plots displaying the membership of cells from the controls (15 min-left; 20 min-middle) and heat-shocked (right) *S. cerevisiae* cells, Replicate 2.
- **C.** Inferred pseudo-time trajectory of the combined dataset. Color bar indicates pseudo-time.
- **D.** Inferred trajectories split by control (15 min-left; 20 min- middle) and heat-shocked (right) datasets shown in Fig. 2H. Colors represent cell clusters shown in Figs. 2H.
- **E.** Expression of heat shock genes HSC82, HSP12 and HSP82 along the pseudo-time trajectory. Color bar indicates scaled expression level.

Supplementary Table 2.1: Summary of *S. cerevisiae* datasets

A table showing average read counts, genes, and UMI detected per cell for *S. cerevisiae* mDrop-seq experiments. The average reads and counts for the species-mixing experiments are reported after filtering out misaligned genes.

Library	Raw Counts	Cells Detected	Average Counts per cell	Average Genes per cell	Median Genes per cell	Average UMI per cell	Median UMI per cell
SC_15min_Rep1	227,973,901	6,620	34,437	474	417	781	639
SC_20min_Rep1	137,063,719	5,392	25,419	402	347	626	502
SC_15min_Rep2	108,347,980	4,350	24,907	380	334	503	424
SC_20min_Rep2	58,362,030	4,740	12,312	404	365	529	460
SC_Heat Shock_Rep1	65,393,958	7,483	8,739	434	377	721	590
SC_Heat Shock_Rep2	54,662,686	6,524	8,740	587	515	1,053	866
Species Mix- SC - 7%	88,687,370	1,495	21,288	544	487	1,108	911
Species Mix- CA - 7%	88,687,370	2,671	21,288	193	137	422	348
Species Mix - SC - 15%	88,953,431	4,901	10,962	578	529	1,173	981
Species Mix - CA - 15%	88,953,431	3,214	10,961	797	717	1,565	1,276

Supplementary Table 2.2: Summary of Cell Cycle in *S. cerevisiae*

The number and percentage of cells in different cell cycle phases across the different *S. cerevisiae* experimental replicates.

Dataset	G1 Cells	G1 (%)	S Cells	S (%)	G2M Cells	G2M (%)	NA Cells	NA (%)
SC_Rep1	1,194	15.8	3,435	45.4	2,938	38.8	1,523	16.7
SC_Rep2	2,453	27.5	3,615	40.5	2,862	32	3,082	25.7
SC_Heat Shock _Rep1	1,016	18.4	3,770	68.2	738	13.4	1,369	19.9
SC_Heat Shock _Rep2	831	15.3	3,636	66.8	979	18	1,078	16.5

Chapter 3: Expanding mDrop-seq to the pathogenic species *C. albicans* and examining the response to the antifungal drug fluconazole

3.1 Introduction:

While the use of a model organism was ideal for establishing the mDrop-seq protocol for yeast scRNA-seq, other species of yeast come with their own requirements and difficulties. *Candida* species are of clinical interest to study as they are among the most significant sources of fungal infections and one of the most common hospital-acquired infections [39], [81]. While many *Candida* species are able to cause infections, the majority of infections are caused by 5 species: *Candida albicans*, *Candida glabrata*, *Candida tropicalis*, *Candida parapsilosis* and *Candida krusei,* with another species *Candida auris* rising to become a significant health threat in certain parts of the world [82]. Typically, *Candida* species are commensal, asymptomatically inhabiting the oral, genital, and gastrointestinal tracts in a large proportion of populations [83]. However, disruptions to the environment these yeasts live in can cause opportunistic infections as pathogenic pathways activate. These fungal infections can take the form of mucosal infections or far more severe systemic infections, particularly in immune-compromised individuals [82].

Of the various *Candida* species, *C. albicans* is the most common cause of infection and the is most extensively studied *Candida* species to date [82]. While superficial *C. albicans* infections can be common, they are nonlethal. However, should a systemic infection occur, the crude mortality rate can be high even with the application of antifungal drugs and therapies [39], [81]. These severe infections can result from damage or trauma to the gastrointestinal tract through injury or surgery, application of catheters which can give direct access to the bloodstream, and disruption of the microbiome bacteria which may give room for fungal blooms [41], [82], [84].

The transition from commensalism to opportunism can be caused by multiple factors and occur across a variety of environments, indicating that *C. albicans* must have a wide array of different tools to use in opportune situations. *C. albicans* has different morphologies including a yeast form, pseudohyphae, and hyphae and can switch between these morphotypes as it needs. Morphological switching has been determined to be an important phenotype of virulence, as *C. albicans* cells that are locked in either yeast form or hyphal form show an overall decrease in virulence in mice [85], [86]. The hyphal form of yeast is believed to be the more invasive form, while the yeast form allows for easier dissemination [87]. Additionally, there is evidence that morphological switching allows for immune evasion due to changing the composition of the outer cell wall and exposure of different surface components for immune recognition [88]–[90].

Beyond morphological switching, *C. albicans* also has an array of adhesins and invasins at its disposal to aid in evading and attacking the host during an infection. Adhesins, such as agglutinin-like sequence (ALS) proteins, allow for the attachment to both abiotic surfaces and host cells and plays a role in the formation of biofilms [41], [91]. Additionally, *C. albicans* has two separate ways to enter host cells for evasion. The use of invasins allows for the induction of endocytosis, causing the uptake of the invading yeast cell [92]. Invading yeast cells may also go a more active route by directly penetration host cells through extension of hyphae [92]. Other adaptations that can be made include biofilm formation [93], [94], pH regulation [95], and vast metabolic flexibility allowing for survival in a wide range of environments [41], [96] among many others. *C. albicans* different mechanisms for survival are diverse and can explain how opportunistic infections can so quickly become difficult to deal with properly.

When *Candida* infections occur, there are four classes of anti-fungal drugs that are used to treat systemic cases [97]. The polyene class of anti-fungal drugs include amphotericin B

(AMB) which work through binding to ergosterol on the cell membrane, creating holes in the membrane eventually leading to cell death [98]. This class of anti-fungal has lower reported levels of drug resistance found and is not typically used except for extreme cases due to negative side effects. The second class of anti-fungal drug, azoles, include a commonly used drugs to treat candidiasis: Fluconazole [99]. Azoles are only fungistatic to *C. albicans* by disrupting ergosterol biosynthesis and preventing the generation of new membrane for budding reproduction [100]. Echinocandins, such as Caspofungin, functions through blocking the synthesis of β-d-glucans, a component of the fungal cell wall [101]. Finally, the last class of anti-fungal drug is flucytosine (5-FC) which negatively affects the biosynthesis of proteins and DNA [97].

As is becoming increasingly common, fungi have been developing resistances to all four classes. Resistance to polyene drugs can be achieved through mutations that lower ergosterol levels within the cell membrane often achieved through alterations in the 25 known ERG enzymes [102]. Thicker cell walls and more intense stress responses have been noted to increase resistance to echinocandins [103], [104]. Resistance to 5-FC develops rapidly such that this drug is only ever used in combination with another class [105]. Finally, there are two mechanisms known that confer a resistance to azoles. As the azoles target the ergosterol biosynthesis pathway, overexpression of the various genes in this pathway does overcome the anti-fungal drug's effects [106]. A second mechanism of resistance also originates with pathway upregulation: the ABC transporters. Adenosine triphosphate binding cassette (ABC) transporters such as Cdr1, Cdr2, and Mdr1, alongside other transporters that may have an effect on drug efflux, giving a resistance to azoles by pumping them out of the cell [107]. Strains that are not considered resistant to any of the above classes may still develop a temporary contact resistance upon exposure as the cell responds to the drugs presence. As is true with other pathogens, drug

resistance is a threat to the ability to treat infections. A deeper understanding of how resistance forms and how it can be prevented is required as more and more drug resistant strains and species are discovered.

The vast ability of *C. albicans* to respond to its environment, whether it be its various metabolic and morphological states, ability to hide from a host immune system, or adapting to and resisting anti-fungal agents, makes it all to more important to build a compatible scRNA-seq technique. Thus far, published yeast scRNA-seq protocols such as yeastDrop-seq [37] and adaptations to the 10X Genomics platform [35], [36], [108] only have been tested on *S. cerevisiae*. These techniques would require further testing and validation to work on *C. albicans* and may not be compatible with across these yeast species. *C. albicans* features a thicker cell wall (~150 nm [54]), a problem that can be made worse due to the formation of hyphae in response to stressful environments. In order to achieve in-droplet lysis of *C. albicans*, a stronger lysis buffer is required, even compared to the DYLB used in the previous chapter. *C. albicans* also does not share the benefits of a well annotated genome and high levels of genetic understanding that *S. cerevisiae* does, potentially making data analysis more difficult.

After developing our technique on *S cerevisiae*, we modified mDrop-seq further to allow for compatibility with *C. albicans* and likely many other yeast species. We tested our lysis incubation to optimize our protocol for the highest gene and UMI capture rate. Once we had achieved *C. albicans* single cell libraries, we investigated interesting structures within the data, including cell cycle, virulence factors, and stress response. We further tested our ability to measure wide scale transcriptomic responses with our technique by stimulating cells with fluconazole over a time course study. Fluconazole, as noted above, is a commonly used antifungal drug that is fungistatic, not fungicidal, making it ideal for use in the transcriptomic

study. In doing so, we can measure the response of individual *C. albicans* cells as they respond to the presence of fluconazole by increasing ergosterol biosynthesis and attempting to resist the negative reproductive effects. This work provides the proof of concept that our technique works on the pathogenic yeast *C. albicans*.

Chapter 3 is based on and mostly a reprint of the paper: R. Dohn *et al.*, "mDrop-Seq: Massively Parallel Single-Cell RNA-Seq of Saccharomyces cerevisiae and Candida albicans," *Vaccines*, vol. 10, no. 1, 2022, doi: 10.3390/vaccines10010030.

3.2 Methods:

Yeast Cell Culture

Candida albicans (strain SC5314, ATCC) were grown in YEPD at 27 °C after heavy dilution (~1:375). After 20 hr of cell culture, *Candida albicans* cells were chilled on ice prior to processing using mDrop-seq.

Fluconazole Stimulation of C. albicans

After *C. albicans* was grown overnight in YEPD, five million cells were counted using a Neubauer Improved (NI) hemocytometer (InCyto, #DHC-N01-2) and diluted into 2 mL of fresh YEPD. A total of one million cells were removed from this pool and put on ice as the control population and processed using mDrop-seq. Fluconazole (Sigma, #F8929-100MG) was freshly diluted to 100 µg/mL in fresh YEPD and added to the remaining four million *C. albicans* to a final concentration of 15 µg/mL. The *C. albicans* were then incubated in fluconazole under end-over-end rotation at room temperature for 1.5 and 3 hr, removed and put on ice prior to running mDrop-seq.

mDrop-seq Cell Preparation and Co-Encapsulation in Droplets

Yeast cells were centrifuged separately at 1000 xg for 3 min in a swinging bucket centrifuge at 4 °C. The cells were washed twice with ice-cold PBS-BSA. Following the washes, 10 µL of cells were sampled and counted using a NI hemocytometer (InCyto, #DHC-N01-2). ~1 mL of cells at 700,000 cells/mL suspended in PBS-BSA was placed in a 2.5 mL syringe (BD, #309657). A total of 10 µL of RNase Inhibitor (Lucigen, #F83923) was added per 1 mL suspension immediately before microfluidic encapsulation. A 75 µm DroNc-seq device fabricated in the cleanroom using published design and protocol [58] was used for droplet generation. Cells and beads were co-flowed into the microfluidic device at 1 mL/h. Cells at 700,000 cells/mL and 4,500,000 droplets/mL gives a Poisson loading distribution with $\lambda = 0.15$.

Barcoded beads (ChemGenes, #Macosko-2011-10(V+)) were suspended in *C. albicans* lysis buffer (see below) at 350,000 beads/mL and kept in suspension by constant stirring with a magnetic tumble stirrer and flea-magnet setup (V&P Scientific, #VP 710, #772DP-N42-5-2); the flea magnet is placed in the syringe containing the barcode beads suspended in lysis buffer and the stirrer kept in the vicinity of the syringe during droplet generation. Cells and beads in lysis buffer were co-encapsulated in drops using a surfactant-oil mix (BioRad, #1864006) flowed at 8 mL/hr in a 10 mL syringe (BD, #302995) as the outer carrier oil phase. Droplets were collected at ~3750 droplets/sec for 30 min in 50 mL tubes (Genesee Scientific, #28-106).

C. albicans Lysis Buffer: Barcode beads were suspended in *C. albicans* lysis buffer containing 6.7% Ficoll PM-400 (GE Healthcare, #17-0300-05), 225 mM Tris pH 7.5 (Teknova, #T2075), 22 mM EDTA (Fisher BP2482-500), 3.3% (w/v) sarkosyl (Teknova, #S3377), 55 mM KH3PO4 (Sigma, #P5629-25G), 1.3 mM DTT (Teknova, #D9750), 0.1% (v/v) β-

mercaptoethanol (Sigma, #M6250-10ML), 650 units/mL zymolyase (Zymo Research, #E1005), that was optimized for lysing *C. albicans* in drops

Cell Lysis, Reverse Transcription, cDNA Amplification and Next era Library Generation for mDrop-seq

See section "Cell Lysis, Reverse Transcription, cDNA Amplification and Nextera Library Generation for mDrop-seq" in Chapter 2 methods.

Population-Level RNA-seq Library Preparation of C. albicans

We performed the same standardized procedure to prepare bulk RNA-seq libraries for *C. albicans* as follows: We lysed ~8,000,000 cells each using *C. albicans* Lysis Buffer and incubation at 37 ºC for 20 min. Total RNA was extracted from each lysate using the Direct-zol RNA Miniprep Plus kit (Zymo Research, #R2071), assessed RNA quality using Qubit HS RNA Assay (Invitrogen, #Q32852) and diluted to 25 pg/µL. The total RNA library was annealed to 5′-AAGCAGTGGTATCAACGCAGAGTACTTTTTTTTTTTTTTTTTTTTTTTTTTTN-3′ primer (Integrated DNA Technologies) that allow polyA selection of mRNA and template switching, similar to Drop-seq. Briefly, 11 µL of total RNA library was mixed with 11 µL of 10 µM primer above, 11 µL of 10 mM dNTP (Takara, #639125), 13.75 µL of ultra-pure water and 2.75 µL of RNase Inhibitor (Fisher, #NC1081844), and incubated at 75 °C for 3 min on a PCR thermocycler. A reverse transcription master-mix consisting of 11 µL of 5X Maxima RT buffer, 2.2 µL of H_2O, 11 µL of 5 M Betaine (Sigma, #14300-500G), 1.65 µL of 100 mM $MgCl_2$ (Sigma, #M1028), 2.2 µL of 50 µM Drop-seq TSO primer (Integrated DNA Technologies, AAGCAGTGGTATCAACGCAGAGTGAATrGrGrG), 1.1 µL of RNase inhibitor (Fisher, #NC1081844) and 2.75 µL of Maxima H-RTase enzyme (Fisher, #FEREP0753) was added

immediately after the annealing step. The final RT reaction volume of 45 μL was pipetted several times to mix, centrifuged briefly to spin down the contents and incubated on a PCR thermocycler using the following program: 42 °C for 90 min; 5 cycles*(42 °C for 2 min, 50 °C for 2 min); 70 °C for 15 min, to perform reverse transcription of the polyadenylated mRNA selectively annealed to the primer above. The cDNA was amplified for 12 cycles using Drop-seq PCR primer (Integrated DNA Technologies, AAGCAGTGGTATCAACGCAGAGT) and KAPA HiFi HotStart ReadyMix PCR Kit (Fisher, #NC0465187K). Amplified cDNA was quantified using Qubit and BioAnalyzer, followed by Nextera library (Illumina, #FC-131-1096) generation.

Sequencing

See section "Sequencing" in Chapter 2 methods.

mDrop-seq Data Preprocessing, Alignment and Quality Control

See section "mDrop-seq Data Preprocessing, Alignment and Quality Control" in Chapter 2 methods.

C. albicans data aligned to *Candida albicans* SC5314 reference genome, version A21-s02-m09-r10 (http://www.candidagenome.org/download/gff/C_albicans_SC5314/Assembly21/).

Bulk RNA-seq Data Processing

Bulk RNA-seq data obtained from *C. albicans* was assessed for read quality using *FastQC*, mapped to the respective genomes described above using *STAR* 2.7.0a aligner [62], and RNA counts were generated from the bam files using *FeatureCounts* [63]. Read lengths were down-sampled during alignment using the *STAR* aligner, "-clip3pNbases" and "-clip5pNbases" parameters. Count matrices were compared to the mDrop-seq datasets using the Seurat package.

Clustering Cells and Generating UMAP

See section “Clustering Cells and Generating UMAP” in Chapter 2 methods.

Batch Correction

See section “Batch Correction” in Chapter 2 methods.

Cell Cycle Analysis

See section “Cell Cycle Analysis” in Chapter 2 methods. Lists of genes that serve as cell cycle phase markers for G1, S, and G2M phases were obtained from the Candida Genome Database for C. albicans [109].

Hierarchical Clustering of Cells

See section “Hierarchical Clustering of Cells” in Chapter 2 methods.

Trajectory Analysis on Single-Cell Data

See section “Trajectory Analysis on Single-Cell Data” in Chapter 2 methods.

3.3 Results:

To demonstrate the utility of mDrop-seq in profiling different yeast species, we applied mDrop-seq to *Candida albicans*, a common hospital-acquired infection that can become life-threatening [110]. This yeast has several features that make it challenging for droplet single-cell profiling. Thicker cell walls in *C. albicans* compared to *S. cerevisiae* (~150 nm and ~120 nm, respectively [54]) make its lysis more difficult. This species also has a hyphal phenotype that can clog microfluidic channels and disrupt flow and droplet generation. The Droplet Yeast Lysis Buffer or DYLB (see Methods) used for *S. cerevisiae* cells failed to lyse *C. albicans* cells (as

assessed under a microscope). The *C. albicans* Lysis Buffer used in our experiments is similar to DYLB but with higher concentrations of both detergent and enzyme (see Methods). We performed two replicates of lysis experiments (15, 20, and 25 min) to establish the ideal incubation time for lysis and RNA capture of *C. albicans* in drops. A total of 14,320 *C. albicans* cells were detected across two replicate datasets. Lysis time of 20 min outperformed the other lysis times in both replicates in terms of genes and UMI captured with an average of 735 genes and 1,356 UMI detected.

Investigating Patterns of Variation within the separate biological replicates of *C. albicans*

We combined the 15- and 20-min incubation datasets from replicate 1 (CA_20min_Rep1 and CA_25min_Rep1) to construct a combined dataset with 4006 cells (Figure S5), with lower and upper cutoffs of 220 and 1600 genes per cell. Average counts varied between the five libraries (Supplementary Table 3.1), consistent with lysis experiments performed in bulk. Lysis replicates of *C. albicans,* CA_20min_Rep1 and CA_25min_Rep1, showed heterogeneity both between and within experimental libraries. In this replicate, the 15 min lysis also produced a usable library, but was not included due to differences in total counts per cell with the other two time-points (Supplementary Table 3.1). We combined the 20 and 25 min incubation datasets, CA_20min_Rep1 and CA_25min_Rep1 to construct a combined dataset with lower and upper cutoffs of 250 and 2,000 genes per cell (Pearson correlation = 0.95). The top expressed genes in this dataset are shown in Figure 3.1A. Much like CA_15min_Rep2 and CA_20min_Rep2 in Figure 3C, integration [73] was performed prior to combining these datasets (Figure 3.1B).

We saw several distinct clusters (clusters 6-10) in Figure 3.1B that separated out from the rest of the cells. Cluster 6 shows differential expression of genes associated with cell wall stability, e.g., *FGR41*, *PGA38*, and *SCW11* (p_adj<1e-292; Figure S5C). *CHT3*, a chitinase gene,

was also expressed in cluster 6 (Figure 3.1C). Transcription factors *STP4* and *ADR1,* encoding for zinc finger proteins and implicated in *C. albicans* virulence, [38], [111] and *GNP1*, a transmembrane transporter of amino acids, were moderately expressed across clusters (Figure 3.1D). *OLE1*, involved in filamentation, exhibited high expression across all cells, shown in Figure S5D. Cluster 10 represents cells producing histones, with many histone genes serving as the most significant markers for this cluster ($p\_adj<5.11e-33$; Figure S5E). The heatmap in Figure 3.1F displays the expression profiles of the genes featured in these clusters.

Figure 3.1: mDrop-seq of 4,006 *Candida albicans* cells, replicate 1

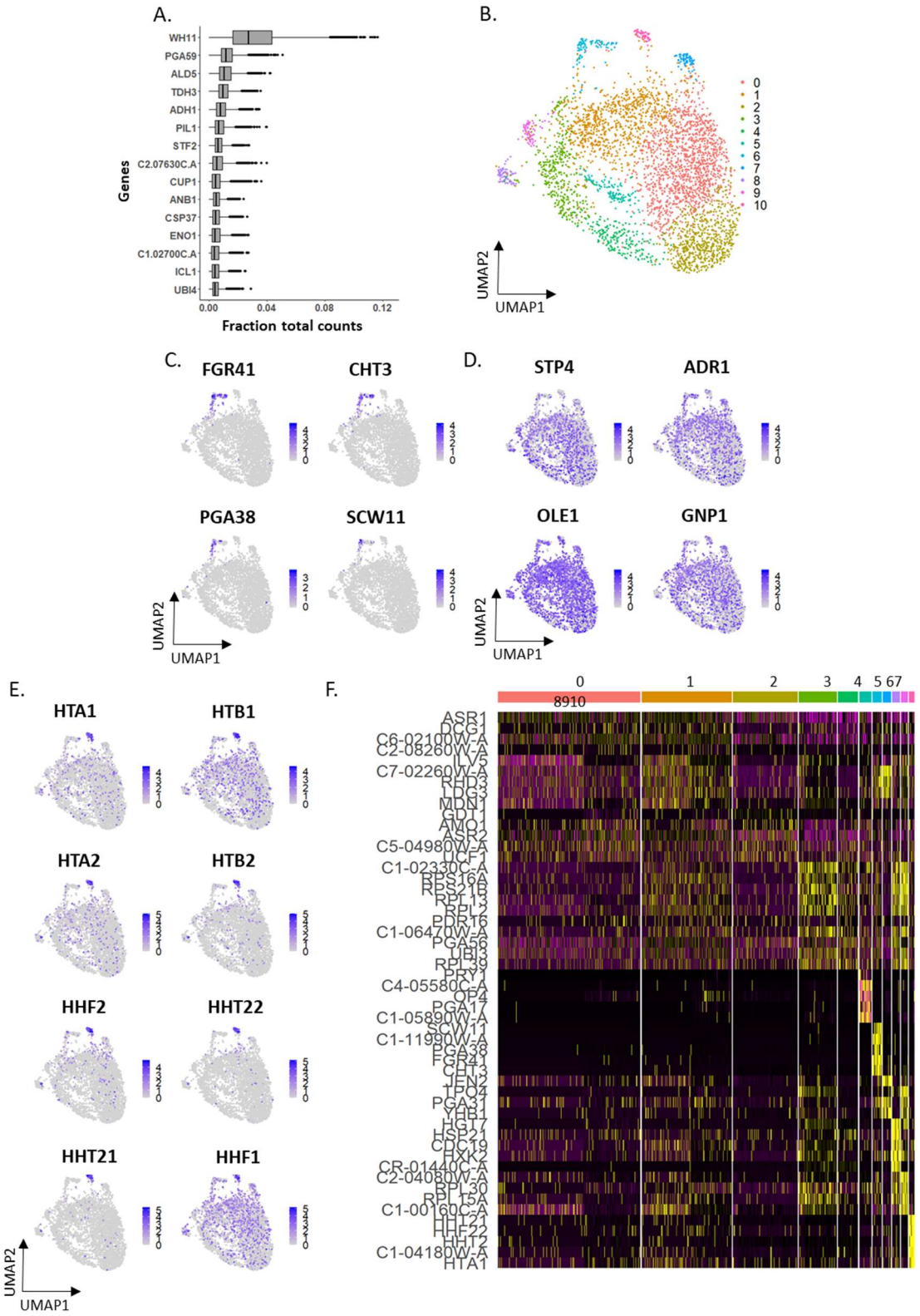


Figure 3.1: Legend

A. Boxplot of the 15 highest expression genes by the fraction of total counts in Candida data.
B. UMAP of clustering analysis of Candida cells after batch correction.
C. Feature plots displaying cell wall genes that represent markers for cluster 6.
D. Feature plots of transcription factors, fatty acid biosynthesis, and hyphal formation genes involved with C. albicans virulence.
E. Feature plots displaying histone tail genes that mark cluster 10.
F. Heatmap displaying expressions of the top marker genes for each cluster shown in B.

The 15- and 20-min libraries (CA_15min_Rep2 and CA_20min_Rep2) in replicate 2 were combined into a dataset containing 10,314 cells. The 20-min library had higher detected genes and UMI across the dataset as mentioned earlier, shown in Figure 3.2A. The gene with the highest expression, on average, is a white-phase yeast transcript *WH11* (Figure 3.2B). *C. albicans* in the white phase is the generic, round yeast form, and expected when grown in rich medium [112], as opposed to the elongated, mating-competent opaque phase [113]. These morphological forms favor growth, unlike the *C. albicans'* filamentous hyphal form [112]. *TDH3*, a gene involved in glycolysis (that also showed the highest expression in *S. cerevisiae*), is the fourth highest expressed gene in the *C. albicans* (in both replicates, also see Figure 3.1A). Potentially due to the difference in genes and UMI detected, CA_15min_Rep2 and CA_20min_Rep2 were separated in PC space (Pearson correlation = 0.96). The data were integrated to find common structures in the data without the interference of batch variation. After integration [73], we see several distinct clusters (clusters 2, 3, and 5) in this dataset (Figure 3.2C) that separate out from the remaining cells. Cluster 3 is marked by the expression of GPI-anchored cell wall genes such as *FGR41* and *PGA38* (Figure 3.2D). Clusters 2 and 5 are heavily represented by histone tail genes as markers (Figure 3.2E), indicating that these clusters may represent variation caused by the cell cycle. Transcription factors *STP4* and *ADR1* encoding zinc

finger proteins and implicated in *C. albicans* virulence, [38], [111] and *GNP1*, a transmembrane transporter of amino acids are moderately expressed across the entire population (Figure 3.2F). *OLE1*, involved in filamentation, also shows high expression across all cells (Figure 3.2F). Figure 3.2G shows a heatmap of the most significantly expressed genes in each cluster.

Our population-level RNA-seq data showed a modest Pearson correlation of 0.70 with pseudo-bulk [9] data constructed from mDrop-seq and 0.64 with a publicly available dataset of *C. albicans* [114]. The correlation between our population-level data and the public dataset was 0.84, by comparison.

Figure 3.2: mDrop-seq of 10,314 *Candida albicans* cells, replicate 2

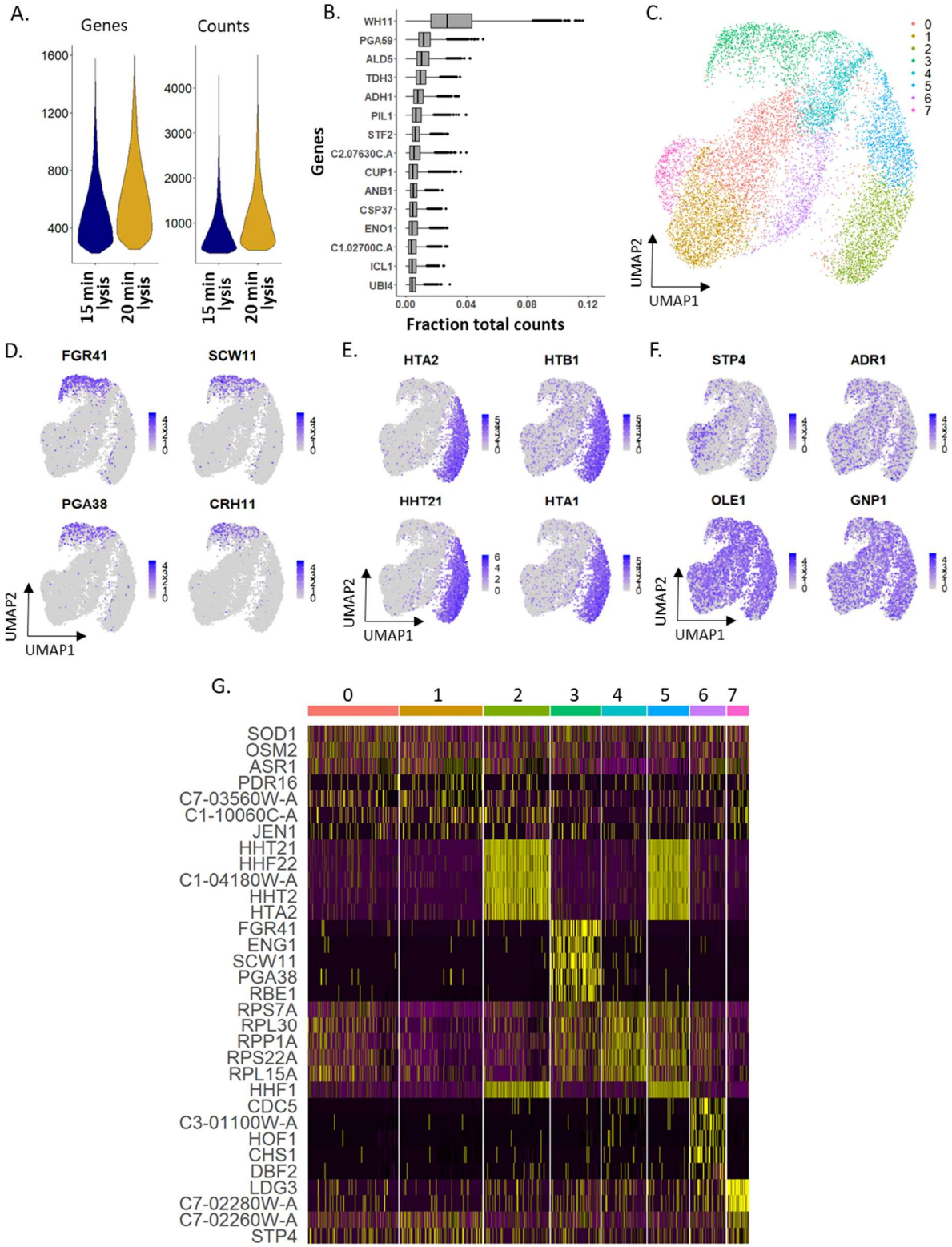


Figure 3.2: Legend

A. Violin plots of genes and UMI counts per cell.
B. Boxplot of 15 genes with highest expression in *C. albicans*, plotted by fraction of total counts.
C. UMAP displaying the clustering analysis of 10,314 *C. albicans* cells after batch correction.
D. Feature plots displaying GPI-anchored cell wall proteins that represent markers for cluster 3.
E. Feature plots displaying histone tail genes that represent markers for clusters 2 and 5.
F. Feature plots displaying genes for transcription factors, fatty acid biosynthesis, and hyphal formation involved with *C. albicans* virulence.
G. Heatmap displaying expressions of the top marker genes for each cluster.

Analysis of cell cycle phases in *Candida albicans* lysis replicates

Cell cycle marker genes for *C. albicans,* determined through microarray analysis of synchronized cells, were obtained from the Candida Cell Cycle Database [109]. Cells from the combined *C. albicans* lysis datasets were scored for G1, S and G2M markers and assigned to the cell cycle phase with the highest score (Figure 3.3A for replicate 1, Figure 3.3B for replicate 2). As seen before in *S. cerevisiae* cells, we were able to assign cell cycle phases unambiguously to only a subset of *C. albicans* cells. The remaining cells were designated as NA. The number and percentage of cells in the different cell cycle phases in each dataset are summarized in Supplementary Table 3.2. We note that the cells marked as NA (gray) are more or less uniformly distributed across the UMAP plot. Figures 3.3C, D show the *C. albicans* replicates 1 and 2 after the variation from cell cycle genes were regressed out. Regressing out any variation in expression due to the cycling genes can be useful when comparing experimental conditions where the effects may be small (e.g., environmental stimuli) which will be useful in later analysis.

Figure 3.3: Analysis of cell cycle genes in *Candida albicans* lysis replicates

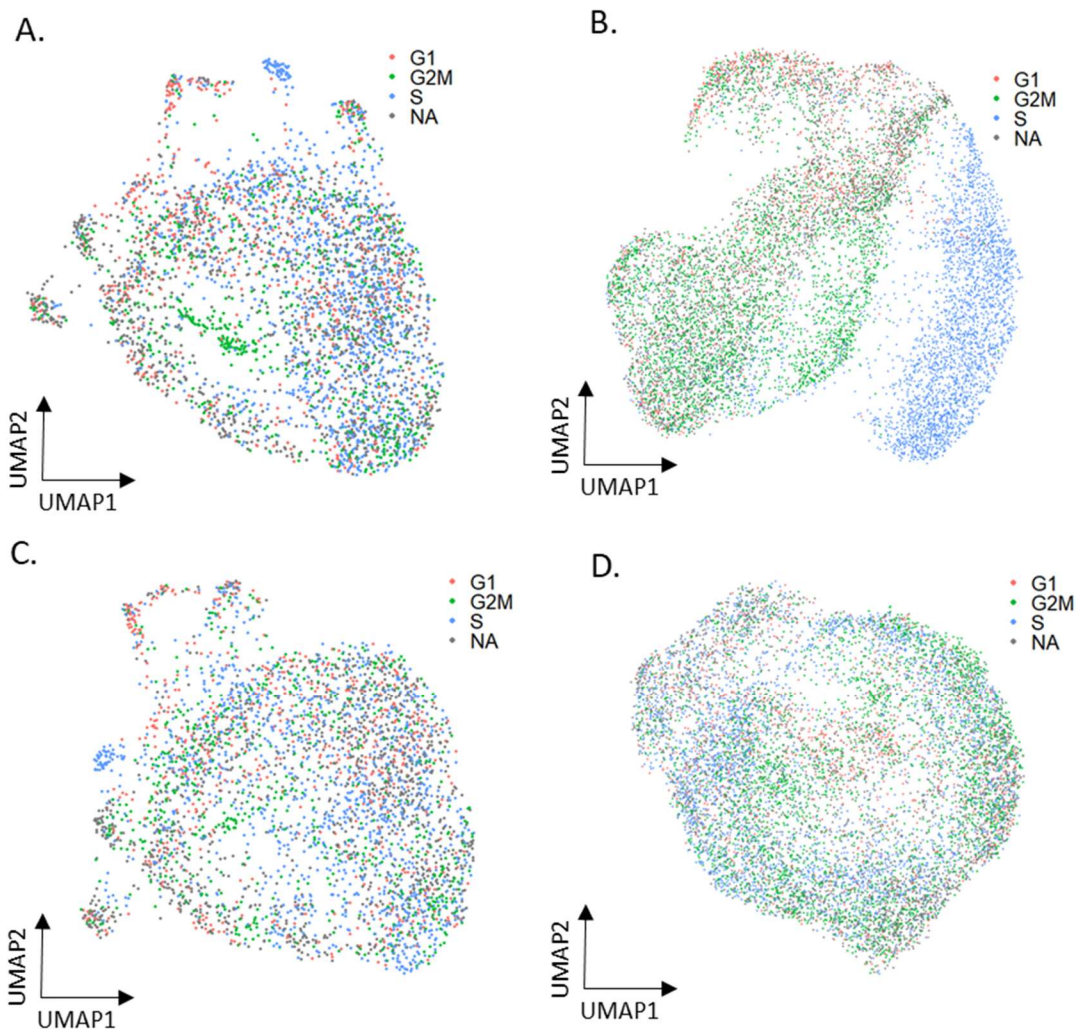


Figure 3.3: Legend

A. UMAP plot of Replicate 1 (Fig. S6B) with cells marked for the G1, G2M and S phases of the cell cycle.
B. UMAP plot of replicate 2 (Fig. 3C) with cells marked for the G1, G2M and S phases of the cell cycle.
C. UMAP plot of replicate 1 showing the integrated dataset after regressing out the variations from cell cycle
D. UMAP plot of replicate 2 after regressing out the variations attributed to cell cycle.

Differential Expression in *Candida albicans* in Response to Fluconazole Exposure

Fluconazole is an anti-fungal agent commonly used to treat infections from various *Candida* species. *C. albicans* cells were exposed to fluconazole over the course of 3 hr, with samples taken for running mDrop-seq prior to exposure, at 1.5 hr and 3 hr. Fluconazole is a bis-triazole antifungal agent that binds to cytochrome P-450 to disrupt the conversion of lanosterol to ergosterol [100]. Previous experiments showed that *C. albicans* responds to the presence of fluconazole in a variety of ways, such as increasing expression of the drug target genes, increasing drug efflux, and finding compensatory pathways for ergosterol biosynthesis [115]. We sampled a population of *C. albicans* cells before (as control) and after (1.5 and 3 hr) exposure to 15 μg/mL fluconazole, which is slightly higher than the C_{max} dose of 400 mg [116]. These experiments were performed twice. *C. albicans* strain SC5314 was shown to be susceptible to fluconazole with a tested MIC50 of 0.156 μg/mL and an MIC80 of 1.25 μg/mL [117]. MIC values were determined via broth microdilution assay in accordance with CSLI standards [117].

We saw an increase in UMI and genes detected per cell when exposed to fluconazole, across replicates (Figures 3.4A). Figure 3.4B shows a UMAP plot of the integrated dataset (control, 1.5 and 3 hr) from replicate 1 with a slight separation between the control and fluconazole-exposed samples. In contrast, there was very little separation between the samples at 1.5 and 3 hr fluconazole exposures. The mean gene expression of the control library yielded a Pearson correlation of 0.91 for the 1.5 hr and 0.88 with the 3 hr time-points (Figure 3.4C).

When comparing the 1.5 and 3 hr time-points of fluconazole treatment to the control, we saw significant upregulation of several ergosterol biosynthesis pathway genes that alter *C. albicans* susceptibility to different classes of antifungal drugs such as azoles and allylamines [118]. Figure 3.4D shows six of these genes, with *ERG11* being the main drug target of

fluconazole and *ERG1,* associated with terbinafine resistance [119] (LogFC = 1.26, 1.71, 1.02, 1.53, 1.34, and 2.00 for *ERG11*, *ERG252*, *ERG1*, *ERG13*, *ERG10*, and *ERG6*, respectively). ABC transporters used for drug efflux were not detected, likely due to the short duration of our fluconazole treatment [115].

Using DE analysis on the combined dataset, we identified the following genes of interest: DE genes that show higher expression in the fluconazole treated datasets, e.g., *CHT2*, *INO1*, *POL30, TNA1, RHD3, HXK2* (Figure 3.4E) that included several antigenic genes and genes upregulated during a host immune response; DE genes that show increased expression in the control data that decreased with time under fluconazole treatment, e.g., *ASR1*, *ASR2*, *WH11*, *HSP70*, *AHP1* (Figure 3.4F) associated with core and heat shock specific stress responses; and DE genes that show the highest expression transiently in the 1.5 hr fluconazole treated sample, e.g., *MRV5, ADH2*, *SOD5*, *CAR1* (Figure 3.4G) associated with acid, osmotic and alkaline stress responses. Violin plots of housekeeping genes *ACT1*, *PDA1*, *TDH3*, and *PGK1* are shown in Figure 3.4H for comparison.

Since the fluconazole treatment led to steady changes in gene expression along the 3 hr time course, we attempted to capture the temporal changes in gene expression by constructing pseudo-time trajectories for the *C. albicans* stimulation using Monocle, an R package [67]. We assumed the control (untreated) sample as time t = 0 h in the fluconazole treatment for this analysis. Figures 3.4J, K show the pseudo-temporal trajectories of CA_Fluconazole_Rep1 response to fluconazole, marked by experimental time-point and cell-type clusters (identified by unsupervised clustering and shown in Figure 3.5C), respectively. Based on prior knowledge, the tip (bottom left) in Figure 3.4J was set as the starting point for pseudo-time construction. The occurrence of the untreated control (blue) on the left of trajectory in Figure 3.4J and the

fluconazole treated samples (coral–1.5 hr; green–3 hr) to the right are consistent with the inferred pseudo-time progression (Supplemental Figure 3.1A). Supplemental Figure 3.1B shows the increasing expression of *ERG10* and *ERG11* genes that mediate resistance to fluconazole and other antifungal agents along the trajectory [115].

The contributions of each experimental time-point to the different branches of the pseudo-time trajectory are shown in Supplemental Figures 3.1C, D, marked by cell-type clusters and cell cycle stages, respectively, while broken down by control (left), 1.5 hr (middle) and 3 hr (right) time-points. The branches to the left were predominantly composed of cells from the control sample. S phase assignment dominated the fluconazole treated cells (Supplemental Figure 3.1D, middle, right), as seen from cell clustering.

In summary, pseudo-time analysis of *C. albicans* that were exposed to fluconazole shows that cell activation trajectory in pseudo-time can be used to infer the temporal sequence of gene expression in yeast cells under external stimuli.

Figure 3.4: Fluconazole treatment and analysis of 15,503 Candida albicans cells

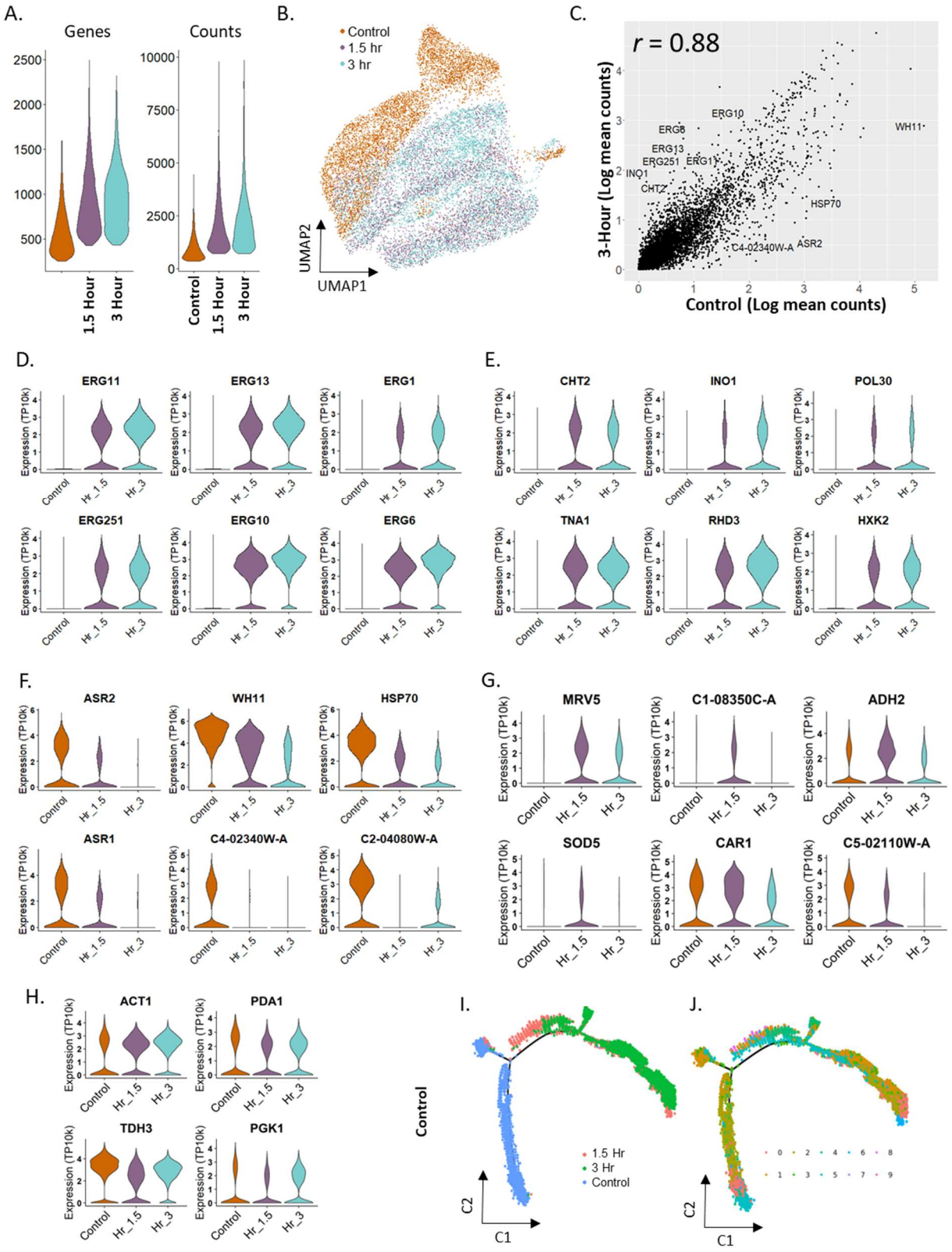


Figure 3.4 Legend:

A. Number of genes and UMI for each fluconazole treatment times.
B. UMAP displaying the clustering of the combined control and fluconazole data.
C. Correlation between the control and 3-Hour fluconazole exposed yeast for gene expression.
D. Violin plots displaying expression in control and fluconazole datasets for several ergosterol biosynthesis genes.
E. Violin plots displaying the expression differences of genes detected at significantly higher expression in fluconazole treated data.
F. Violin plots of genes that show significantly higher expression in the control data.
G. Violin plots of DEGs for the 1.5 hour fluconazole exposed data.
H. Expression levels of housekeeping genes ACT1, PDA1, TDH3 andPGK1 in CA, replicate 1.
I. Pseudo-time trajectory of the combined data inferred using Monocle. Colors indicate experimental time points.
J. Pseudo-time trajectory of the combined data inferred using Monocle. Colors indicate cell-type clusters shown in I.

Cell Cycle Analysis in Fluconazole Stimulated Cells

Next, we performed cell cycle analysis on the combined dataset (CA_Fluconazole-ctrl, 1.5, 3hr_Rep1) from the control, 1.5 and 3 hr time-points of the fluconazole treatment. When colored by the cell cycle phase, the UMAP of the combined dataset (Figure 3.5A) showed some clustering by the cell cycle. We also saw a significant increase in the number of cells assigned to the S phase under fluconazole treatment (3.2x and 1.8x for 1.5 and 3 hr, respectively; Supplementary Table 3.2) compared to the control dataset. Cells under stress tend to go into cell cycle arrest [120]. Increased expression of ERG genes was also associated with slow growth in yeasts [121]. Since many histone genes (Figure 3.5B) occur in the list of marker genes for the S phase, we verified that high histone activity alone in the fluconazole treated cells was not skewing our cell cycle assignment towards the S phase. Note that histone genes in yeasts have polyadenylated tails, unlike in humans [122]. For the experiments where *C. albicans* cells were treated with fluconazole, we noted significant increase in the fraction of cells assigned to the S phase and decrease in the fraction of cells assigned to the G2M phase, compared to their

controls. The fraction of cells in the G1 phase did not change much, by comparison. These observations are consistent for both replicates of fluconazole treatment. This is also qualitatively similar to *S. cerevisiae* cells that underwent heat shock treatment.

To verify that cell assignment to the S phase was not biased by high expression of histone genes, the expression levels for S phase markers were examined with and without the contribution of histone genes. Figures 3.5D, E show dot plots of expression levels for S phase marker genes for all assigned phases of the cell cycle (G1, S, G2M and NA) in the fluconazole datasets, CA_Fluconazole_Rep1 and CA_Fluconazole_Rep2. The average expression levels of all S phase genes in the cells assigned to the S phase were higher than the expression in the remaining cells. Even when the histone genes were excluded from the S phase marker list (Figures 3.5F, G), cells assigned to the S phase showed higher expression than G1 and G2M phases. This indicates that high S phase assignment among the *C. albicans* cells was not driven by the high expression of histone genes alone.

Unsupervised clustering of the combined CA_Fluconazole_Rep1 data, after cell cycle effects were regressed out (Figure 3.5C), showed clusters of cells exhibiting histone activity (cluster 4), ribosome activity (clusters 4, 6), synthesis of ribonucleoproteins and Hap43 induced proteins, along with a reduction in iron metabolism (cluster 5; violin plots of expression for some genes in this cluster are shown in Figure S10A), stress response (cluster 7), synthesis of the cell wall and vacuolar proteins (cluster 8), and nucleolar activity (cluster 9). When cells from each time-point were plotted separately (Figure 3.5H), we saw that some cell clusters were present predominantly in either control or fluconazole-treated time-points, e.g., clusters 5 and 7 in control, cluster 4 at 1.5 hr, and cluster 6 at 3 hr. The number of cells in clusters 2 and 3 decreased monotonically between control, 1.5 and 3 hr time-points. These results show interesting

variability in transcriptomic response between cells to fluconazole, potentially providing insight into differences in resistance between cells. Biological replicates were analyzed separately due to batch effects between the replicates potentially interfering with biological variation within the data.

Figure 3.5: Cell cycle analysis of 15,503 control and fluconazole treated C. albicans cells, replicate 1

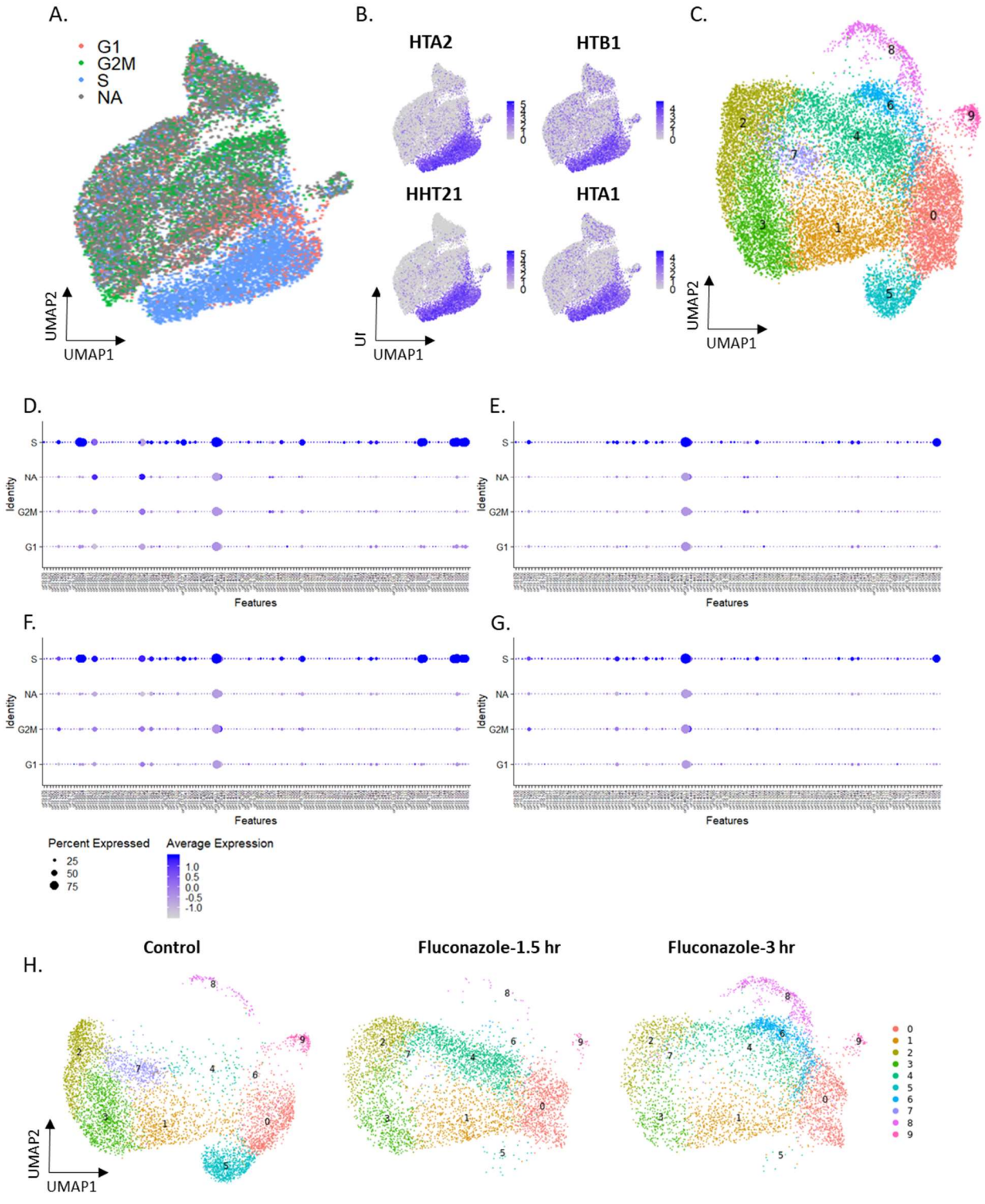


Figure 3.5: Legend

A. Feature plots displaying four histone tail genes across control and fluconazole exposed cells.
B. Cell Cycle assignment on the integrated control and fluconazole data.
C. UMAP of the integrated data after cell cycle regression.
D. Dot plots of S phase markers D) with histone genes on replicate 1; E) without histone genes on replicate 1; F) with histone genes on replicate 2; G) without histone genes on replicate 2.

3.4 Discussion:

Much like the previous chapter, we have established mDrop-seq to be able to profile *C. albicans* at single-cell resolution. In total, 39,705 *C. albicans* cells were detected across multiple experiments and replicates. The use of fluconazole as an environmental stimulus shows that we can detect and analyze the cellular response to outside stimuli within *C. albicans* despite the enhanced difficulties of the species. We note many of the same trends seen in *S. cerevisiae,* such as UMI increases in stimulated libraries, held true for *C. albicans*.

Further Modification of the Lysis Buffer Affected Droplet Quality

In Chapter 2, we developed the DYLB for in-droplet lysis of *S. cerevisiae* through a combination of zymolyase and sarkosyl in drops followed by a thermal incubation. When tested on *C. albicans*, the DYLB proved insufficient to achieve the same level of lysis, with many cells surviving intact beyond 30 minutes (data not shown). As such, the *C. albicans* Lysis Buffer was further optimized, containing higher concentrations of both zymolyase and sarkosyl. We note that a high concentration of detergent in the lysis buffers, e.g., 3.3% sarkosyl in the *C. albicans* lysis buffer, is detrimental to stable droplet formation, necessitating lower flow rates on the microfluidic device. Due to lower flow rates, experiments run using the *C. albicans* Lysis Buffer are likely to have lower expected cell counts for the same collection time.

We posit that similar cocktails consisting of zymolyase and sarkosyl will prove effective on a broad class of fungal species that share similar cell wall properties to *S. cerevisiae* and *C. albicans*, including clinically relevant species such as *C. auris*. Highly concentrated lysis buffers such as the *C. albicans* Lysis Buffer may be universally used, or buffers can be modified to be appropriate for the species of interest. Anti-fungal peptides [123] that target specific components of the fungal cell wall may also be added to the lysis cocktail of zymolyase and sarkosyl.

Analysis of Cellular Response to Anti-Fungal Agent Fluconazole

mDrop-seq on *C. albicans* cells exposed to the anti-fungal agent fluconazole showed significant increases in the number of counts and genes detected, compared to control data. Much of these increases appeared to be driven by increased expression of ribosomal structure genes (e.g., *RPS27, RPS6A, RPS12*). Across both replicates, we noted significant upregulation of several genes belonging to the Ergosterol Biosynthesis pathway, including *ERG11* that produces the cytochrome P-450 target of fluconazole. The upregulation of this pathway is a known mechanism for resisting the membrane disruptive effects of fluconazole. We did not detect any upregulation of ABC transporter genes as mechanisms of resistance. During the 3 hr exposure time, we also noted several genes that increased their expression transiently, increasing quickly in expression in 1.5 hr before decreasing by the end of the 3 hr period. These genes included some stress-induced genes such as *SOD5* and *ADH2*, as well as *C1-08350C-A, C5-02110W-A.* After integrating the control and fluconazole treated data for each replicate, followed by cell cycle regression and clustering analysis, we noted that the signatures of ribosomal and histone expression differences persisted within the clusters. A cluster marked by GPI-anchored proteins was also seen in the integrated data, as well as a cluster involving nucleolar and pre-ribosomal genes.

In particular, we noted increased histone activity in *C. albicans* under fluconazole exposure. Since many cell cycle genes for the S phase in *C. albicans* are histone-related, we confirmed that the signal for the S phase assignment was preserved, compared to the G1 and G2M phases even when the histone genes were excluded from the S phase marker list (Supplementary Material). Since chromatin accessibility is needed for increased transcription under stress response, this may drive up histone expression under heat-shock in *S. cerevisiae* or fluconazole treatment in *C. albicans*.

Due to the use of an antifungal drug, there are some limitations in how the data in our study can be interpreted. While the strain of C. albicans used in our experiments is susceptible to fluconazole, antifungal drug exposures may cause contact or acquired resistance that can be reversible. The adaptability of C. albicans can be confounding due to the many pathways it can use to increase resistance in response to different classes of antifungal drugs [53,54]. Our experimental design simply demonstrates C. albicans response to fluconazole over our time-course study but cannot distinguish if these responses are intrinsic, temporary, or permanent. Further experiments will be needed to resolve the issue that is beyond the scope of this study.

Scalability as a Technology

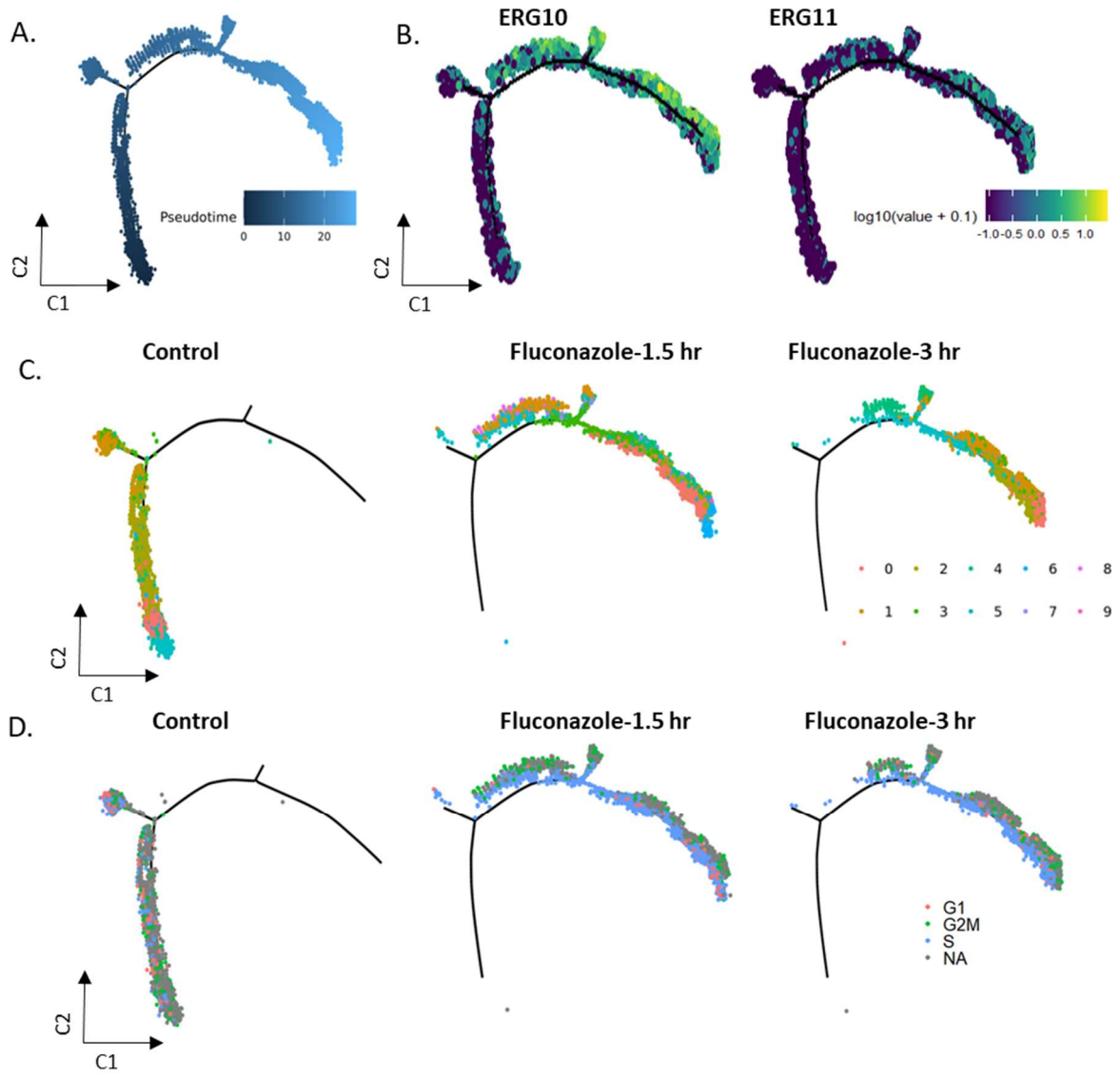
With two separate species have been profiled at high throughput, we can show that mDrop-seq works as a cheap transcriptomic profiling solution for unicellular fungi and may be easily adapted in laboratories that use Drop-seq or similar techniques. Existing bioinformatics and statistical tools for single-cell analyses may be effectively leveraged to analyze fungal species at single-cell resolution across multiple species. According to our estimate, the cost of single-cell library preparation using mDrop-seq is ~USD250 per sample (~5,000 cells/sample).

At ~50 million reads per sample, we estimate sequencing to cost an additional ~USD190 per experiment.

3.5: Acknowledgement of Work Performed

I would like to acknowledge those who contributed to the work presented here. Rebecca Back aided in the implementation of the mDrop-seq protocol on *C. albicans* as well we the bulk RNA-seq. Heather Eckart assisted in Nextera Library construction and sequencing submission. Project and experiment design was guided by Dr. Anindita Basu. Dr. Alan Selewa and Dr. Bingping Xie assisted with alignment, QC, and data analysis. All other experimentation and analysis present in this chapter were performed by Ryan Dohn.

Supplemental Figure 3.1: Integration and cell cycle regression analysis of 15,503 *C. albicans* cells, replicate 1



Supplemental Figure 3.1: Legend

A. UMAP plots displaying the membership of cells from the control (left) and fluconazole (1.5 hr- middle; 3 hr- right) samples in the combined dataset shown in Fig 3.4I.
B. Inferred trajectory of gene expression in CA, replicate1 in response to fluconazole, with color bar indicating pseudo-time.
C. Expression levels of ergosterol synthesis genes, ERG10 and ERG11 increase with pseudo-time.
D. Inferred trajectories split by control (left) and fluconazole (1.5 hr- middle; 3 hr- right) datasets shown in Fig.4K. Colors represent E) cell clusters shown in Fig 4I, F) Cell cycle phases, G1, S, G2M and not assigned (NA).

Supplementary Table 3.1: Summary of *C. albicans* datasets

A table showing average read counts, genes, and UMI detected per cell for *C. albicans* mDrop-seq experiments. The average reads and counts for the species-mixing experiments are reported after filtering out misaligned genes.

Library	Raw Counts	Cells Detected	Average Counts per cell	Average Genes per cell	Median Genes per cell	Average UMI per cell	Median UMI per cell
CA_15min_Rep1	30,400,506	1,215	25,021	451	376	685	504
CA_20min_Rep1	93,129,841	1,693	55,008	837	706	1,660	1,223
CA_25min_Rep1	128,016,172	2,364	54,152	573	494	1,029	783
CA_15min_Rep2	96,778,937	6,452	7,132	508	549	836	610
CA_20min_Rep2	129,649,186	3,862	15,021	611	561	1,132	976
CA_Fluconazole-ctrl_Rep1	154,646,838	5,952	25,982	563	510	944	784
CA_Fluconazole-1.5hr_Rep1	122,638,009	4,924	24,906	893	816	1,948	1586
CA_Fluconazole-3hr_Rep1	111,642,039	4,627	24,128	953	922	2,130	1,882
CA_Fluconazole-ctrl_Rep2	95,442,703	4,227	10,918	351	324	527	456
CA_Fluconazole-1.5hr_Rep2	115,687,150	3,687	12,978	967	918	2,152	1,863
CA_Fluconazole-3hr_Rep2	96,686,038	3,028	11,989	935	896	1,909	1,689

Supplementary Table 3.2: Summary of Cell Cycle in *C. albicans*

The number and percentage of cells in different cell cycle phases across the different *C. albicans* experimental replicates.

Dataset	G1 Cells	G1 (%)	S Cells	S (%)	G2M Cells	G2M (%)	NA Cells	NA (%)
CA_Rep1	268	14.6	862	47.1	702	38.3	2,174	54.3
CA_Rep2	671	11.4	3,282	55.8	1,932	32.8	4,429	42.9
CA_Fluconazole-ctrl_Rep1	585	25.3	702	30.3	1,029	44.4	3,636	61.1
CA_Fluconazole-1.5hr_Rep1	370	11.7	2,257	71.7	522	16.6	1,775	36
CA_Fluconazole-3hr_Rep1	603	22.2	1,325	48.9	782	28.9	1,917	41.4
CA_Fluconazole-ctrl_Rep2	351	17.9	756	38.6	854	43.5	2,266	53.6
CA_Fluconazole-1.5hr_Rep2	233	10.2	1,344	58.8	708	31	1,402	38
CA_Fluconazole-3hr_Rep2	273	16.6	927	56.3	447	27.1	1,381	45.6

Chapter 4: Quantifying transcriptional variation between yeast scRNA-seq experiments and reduced heterogeneity in samples under stress response

4.1 Introduction:

There are a multitude of variables that can contribute to unwanted variations or 'batch effects' within single cell data. Whether they be on mammalian or microbial cells, scRNA-seq studies often have to deal with experiments being run at different times, with different reagent lots, or handled by different personnel among other issues [126], [127]. Studies with logistical issues that need to be processed in subsets due to sample size or equipment limitations have even greater potential for bias [128]. These issues can be compounded with variations in gene capture rates and dropouts which often occur in single cell data [129]. A variety of techniques to computationally handle batch effects have been created, but identifying the best technique may not always be obvious as some variation may be complex and non-linear [130]. However, using an integration technique that is not well equipped to handle the biases within the data may also cause the loss of real biological variation.

These issues, like others discussed in this dissertation, are also relevant for microbial organisms. Microbes are incredibly dynamic, with a wide range of phenotypes for responding to minor changes in environment [131]. The dynamic nature of microbes makes them even more susceptible to biases introduced by differences in handling and experimental conditions[132]. Confounding factors like dropouts are expected to be far worse as well due to low levels of mRNA (including low copy numbers of any given gene) and higher amplification requirements increasing bias [71]. As such, scRNA-seq studies must be carefully designed reduce these effects as much as possible to prevent batch effects from confounding real biological insight.

The best way to tackle variation between libraries is good study design. Identifying sources of variation and planning an experiment around those sources can reduce batch effects without the need for integration. In the previous chapters of this dissertation, we noticed transcriptomic differences within many of our libraries, despite the cells being similar in cell type and treatment. These batch effects necessitated a redesign of our study in order to find additional sources of bias. The use of YEPD in continuous culture for yeast availability during technique development likely led to increased heterogeneity within the data presented in this dissertation. YEPD is known to be inconsistent due to the use of dead cells in its formulation. To reduce levels of variation from continuous culture, we adapted many of the later experiments to be consistently planned for a set amount of time post opening from frozen stocks. Some variability however cannot be helped, as there are inconsistencies within microfluidic fabrication and handling of samples that are difficult to overcome. It is difficult to de-couple the variation from experimental conditions from the intrinsic variability in transcription within or between microbial populations [131].

In this chapter, we investigate possible sources of patterns of transcriptional variability within our earlier data and discuss we can improve the study design in of single cell genomics in yeasts. We also observed that yeast stimulation experiments preformed on different dates displayed lower variability in transcription, compared to their unstimulated controls. We determined that the stress response in our experiments displayed reduced transcriptional variation potentially due to the transcriptome converging on the set of stress response genes. In addition, we also determined ways in which our cell culture conditions could be systematically controlled to reduce transcriptional heterogeneity to our data, thus helping guide the experimental design for future experiments.

4.2 Methods:

Yeast Cell Culture

Candida albicans (strain SC5314, ATCC) were grown in YEPD at 27 °C for 3 weeks and propagated into fresh YEPD twice weekly. When multiple populations were necessary, cells were opened and propagated simultaneously. 100 µL samples were taken for processing both Day 1 and Week 3 of growth after opening from a frozen stock. Cells were opened from the frozen stock (Day 1) or propagated into fresh media (Week 3) 17 hours prior to mDrop-seq. 100 µL samples were taken for processing both Day 1 and Week 3 of growth after opening from a frozen stock, following which the cell samples were placed on ice and chilled.

Fluconazole Stimulation of C. albicans

After *C. albicans* was grown overnight in YEPD at either the 1 Day or 3 Week time point, five million cells were counted using a Neubauer Improved (NI) hemocytometer (InCyto, #DHC-N01-2) and diluted into 2 mL of fresh YEPD for each population. A total of two million cells were removed from this pool and put on ice as the unstimulated control and processed using mDrop-seq for each population. Fluconazole (Sigma, #F8929-100MG) was freshly diluted to 100 µg/mL in fresh YEPD and added to the remaining three million *C. albicans* to a final concentration of 15 µg/mL. The *C. albicans* were then incubated in fluconazole for 2 h in a cell culture tube (Fisher, # 14-956-1J) in an incubated shaker set to 27 °C and 450 rpm, removed, and put on ice prior to running mDrop-seq. Experiments requiring multiple populations were processed side-by-side simultaneously.

mDrop-seq Cell Preparation and Co-Encapsulation in Droplets

See section “mDrop-seq Cell Preparation and Co-Encapsulation in Droplets” in Chapter 3 methods.

Experiments requiring biological or technical replicates be processed simultaneously were performed on different Drop-seq apparatuses at the same time.

Cell Lysis, Reverse Transcription, cDNA Amplification and Next era Library Generation for mDrop-seq

See section “Cell Lysis, Reverse Transcription, cDNA Amplification and Next era Library Generation for mDrop-seq” in Chapter 2 methods.

Sequencing

See section “Sequencing” in Chapter 2 methods.

mDrop-seq Data Preprocessing, Alignment and Quality Control

See section “mDrop-seq Data Preprocessing, Alignment and Quality Control” in Chapter 2 methods.

C. albicans data aligned to *Candida albicans* SC5314 reference genome, version A21-s02-m09-r10 (http://www.candidagenome.org/download/gff/C_albicans_SC5314/Assembly21/).

Clustering Cells and Generating UMAP

See section “Clustering Cells and Generating UMAP” in Chapter 2 methods.

Permutational Multivariate Analysis of Variance

To perform Permutational Multivariate Analysis of Variance or PERMANOVA [133], relevant datasets were combined into a single object and clustered together. After clustering, libraries that were not part of any given measurement of variance were subset out. UMAP coordinates were extracted from the object and used to compute the pairwise Euclidean distance matrix. The adonis2() command from the *vegan* R package [134] was used to generate the PERMANOVA results.

4.3 Results:

Identifying differences in clustering patterns within stimulated and unstimulated data

In the previous chapters, experimental data was analyzed within individual replicates to prevent significant batch effects from effecting the data (see Figures 2.1, 3.1, 3.2, and 3.4). In some cases, CCA integration was applied to remove these batch effects, however it was still noted that differences between libraries of what would otherwise be cells grown in the same conditions may have effects on analysis. While data integration does appear to be a suitable way to remove many of these batch effects, we noted a pattern emerged within the clustering analysis of many of the stimulated data sets. When comparing across biological replicates of stimulation experiments, we noticed replicates of unstimulated controls cluster separately but stimulated libraries cluster together, regardless of yeast species or stimulation conditions. This observation is summarized in Figure 4.1, showing A) the combined datasets from different replicates of heat-shocked *S. cerevisiae* (Chapter 2) and B) replicates of *C. albicans* under fluconazole stimulation (Chapter 3). We hypothesize that this observation may be the result of individual cells, which

otherwise have heterogeneous gene expressions, now show a convergence in their transcriptomes, as specific stress-response pathways get activated.

In Figure 4.1A, there are four total libraries, two unstimulated and two undergoing heat shock (at 42°C for 20 min). The heat shocked libraries, while still somewhat clustering apart, appear closer in UMAP space than the unstimulated controls. It should be noted that one of the controls used in this comparison is referred to as the "Intermediate" heat-shock response (as discussed in Chapter 2), such that we may not expect similar clustering between intermediate heat shock and control libraries. However, this trend persisted in the fluconazole data (Figure 4.1B). Both the 1.5 and 3 hour fluconazole exposures cluster together across two biological replicates, while their respective controls cluster apart. Figure 4.1C shows Pearson correlations in average gene expression between the various libraries represented in Figure 4.1B. These correlations indicate how similar each of these libraries are, both within and between biological replicates. Higher correlations between fluconazole-treated libraries across biological replicates (black box in Figure 4.1C) compared to biological replicates of unstimulated controls may indicate that fluconazole exposure caused the transcriptomes of single *C. albicans* cells to converge under stress response, despite any batch effect that may exist between biological replicates.

In order to develop a generalized framework of comparison for overall heterogeneity in single-cell gene expression *between* and *across* experimental conditions, we used permutational multivariate analysis of variance, or PERMANOVA [133]. PERMANOVA creates a pseudo-F ratio as a measure of variance within and between comparison groups in a multivariate dataset without needing to make the assumption of multivariate normality, and tests significance through random permutations of the objects against this F ratio [133]. The magnitude of this pseudo-F

ratio can give us an estimate of the amount of variation between different comparison groups within a complex dataset, with the caveat that this statistic may also be affected by high levels of dispersion within each dataset. The R^2 value produced by this test indicates the percentage of variation explained by the model. We used the UMAP distance matrices for various datasets as input for PERMANOVA tests, as a proxy for variance in our gene expression data to lower computational costs. To make the pseudo-F ratios comparable, each cell embedding was calculated for all comparison datasets integrated in a single UMAP space, prior to subsetting the conditions of interest. A total of 99 permutations were performed to generate each pseudo-F ratio, putting a lower limit on p-values to 0.01. The pseudo-F ratios for all tests reported in this chapter can be found in Table 4.1.

We performed PERMANOVA test on the *S. cerevisiae* cells (shown in Figure 4.1A) to estimate the similarity between the two unstimulated libraries (marked 'Control' and 'Intermediate') and again between the stimulated libraries. We obtained pseudo-F ratios of 62,825 for the libraries of unstimulated cells and 17,370 for the heat-shock stimulated libraries (Table 4.1). It is important to note that PERMANOVA is best used for balanced experimental designs [133]; therefore, the lack of biological controls for the heat shock experiments, along with previously noted differences in the unstimulated libraries, may confound the comparison. Despite this, we note that the replicates of heat-shocked cells yield lower pseudo-F ratios than the unstimulated cells.

We performed a similar analysis on the combined dataset of the fluconazole stimulation experiments in Chapter 3 (Figure 4.1B). This experiment was performed with control experiments drawing from the same population of cell culture and processed alongside the stimulation experiments. Two such experiments were performed on different days, as biological

replicates. Under these conditions, we obtain a pseudo-F ratio of 31,576 for the unstimulated controls and a pseudo-F ratio of 2,923 for the fluconazole-stimulated cells (Table 4.1). Taken together, we see that stimulated cells of different species and under different stimulation/stress conditions, displayed lower variation in transcription compared to unstimulated cells.

Stimulation with fluconazole reduces variation between replicates

We further investigated the sources of transcriptional variation within single cells under stimulation and control conditions with the aim to reduce heterogeneity between datasets obtained from similar experimental conditions. To do so, we designed an experiment in which *C. albicans* cells were opened from frozen stocks into three identical culture flasks and grown in the same shaker under identical conditions (and considered as 3 identical populations or '3-pop'). A cell sample was taken from each of these three populations and profiled using mDrop-seq, with and without a 2-hour fluconazole stimulation (at a concentration of 15 µg/mL fluconazole); This set of experiments was performed on the next day from starting cell culture (Day 1) and after three weeks of continuous growth (Day 21). Unfortunately, mDrop-seq of the unstimulated cells performed on the Day 1 (as control) did not produce usable libraries and had to be discarded; these experiments will not be discussed further. Supplemental Figure 4.1A displays this experimental design (Day 1 has been excluded from the figure). The 3-week time point produced six (three unstimulated and three stimulated) mDrop-seq libraries which were analyzed to compare transcriptional variation between cell cultures after several weeks of growth under culture identical conditions. Unlike the experiments performed in previous chapters in which biological replicates were performed on different days, growing and running biological replicates side-by-side had an impact on clustering behavior within the unstimulated samples. The UMAP in Figure 4.1D displays these libraries, colored by stimulation status. We do not see the

unstimulated samples clustering separately, as was previously noted in previous chapters. PERMANOVA was also used to determine levels of variation in unstimulated and stimulated conditions. The three unstimulated populations yielded a pseudo-F ratio of 877, compared to the pseudo-F ratio of 133 in the three stimulated populations, confirming that stimulated cells indeed display lower levels of transcriptomic variation, compared to unstimulated cells.

To confirm this observation in *C. albicans* stimulation experiments, another experiment was performed with the same experimental design with the exception that only 2 culture flasks were grown together under identical conditions (and considered as 2 identical populations or '2-pop', shown in Supplemental Figure 4.1B). As such, we wanted to compare variation between stimulated and unstimulated cells with some populations grown together (3-pop, 2-pop) or grown separately (3-pop vs. 2-pop). In Figures 4.1E, F, we see the clustering results when these experiments are combined without integration. In Figure 4.1E, the unstimulated (control) cells of each experiment cluster apart in UMAP, as seen in previous chapters. However, when the cells in these experiments were exposed to 2 hours of fluconazole stimulation at 15 μg/mL concentration, we see that each replicate clusters together, perhaps indicating a transcriptomic convergence of the cellular stress response, as seen before. The unstimulated cells, when comparing the 2-pop and 3-pop experiments, have a pseudo-F ratio 15,263 while the stimulated cells have a ratio of just 6.67 (Table 4.1). The pseudo F-ratios also support that stimulation and stress response may cause a convergence of the transcriptome, allowing for less cellular variation within a population. It also seems that variations in single cell transcription noted in previous chapters likely arises from cells grown at different times, even if the conditions are the same. GO analysis of genes retrieved from the PC loadings of PCs with the most pronounced batch variation did not indicate any consistent sources of variation in unstimulated cells.

Figure 4.1: Clustering variation between replicates differs in stimulated and unstimulated libraries

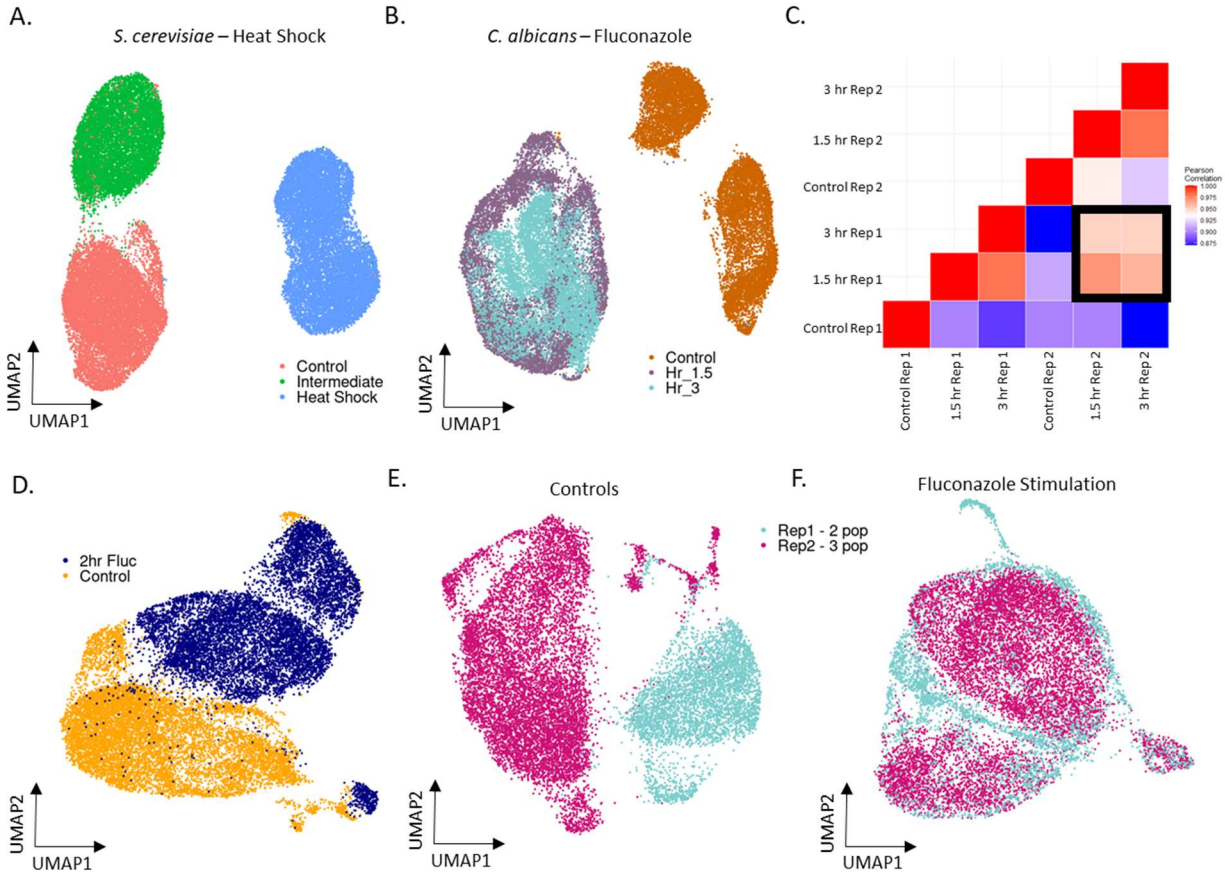


Figure 4.1: Legend

A. UMAP plot of 2 biological replicates of *S. cerevisiae* heat shock stimulation experiments (see Figure 2.4).
B. UMAP plot of 2 biological replicates of *C. albicans* fluconazole stimulation experiments (see Figure 3.4).
C. Pearson Correlation plot comparing average expression data between libraries displayed in 4.1B.
D. UMAP plot of 3 populations of *C. albicans* cells grown together prior to fluconazole exposure.
E. UMAP plot featuring 5 total unstimulated *C. albicans* populations across 2 total replicates displayed by the replicate each population belong to.
F. UMAP plot featuring 5 total fluconazole stimulated *C. albicans* populations across 2 total replicates displayed by the replicate each population belong to.

Cells grown under identical conditions but at different times show higher levels of variation

In order to test if cells that are grown at different times is a strong source of variation for unstimulated cells, we performed a series of experiments over separate three-week intervals. In the first round, a single population of *C. albicans* was opened from the frozen stock, grown overnight before two samples of this single population (serving as replicates) were processed using mDrop-seq side-by-side. The population was then allowed to grow for three weeks at which point two replicates were again processed side-by-side (Supplemental Figure 4.1C). After the completion of the first round of experiments, the previous experiments were repeated for a second round, where a new population was opened from the frozen stock and subjected to the same series of experiments, allowing it to grow in separate 3-week window, same as the first round of experiments. Through these set of experiments (first and second rounds), we can compare levels of variation between unstimulated cells that were grown at different times: one day after thawing, and three weeks later.

The above experiment yielded eight libraries in total, four from Day 1 (from starting cell culture from freezer stock), and four from Week 3. Within each time point, there are two replicates (e.g. Day 1_Rep X.Y where X is the biological replicate (first or second round) and Y is the technical replicate). In Figure 4.2A (top), we show a UMAP of the four 'Day 1' libraries. While differences between the two rounds of 'Day 1' replicates are evident, the clustering appears closer together. Pearson correlations for these libraries are greater than 0.92. Day 1_Rep 2.2 had lower data quality (in terms of average number of genes and UMI captured per cell) than the remaining libraries, which may potentially influence the correlations. Using PERMANOVA, we obtain a pseudo-F ratio of 4,444 between the replicates when cells have only been grown for 1 day. Week 3 data look qualitatively different from the Day 1 data. In Figure 4.2B (top), we see

two distinct clusters representing the two rounds of experiments performed at different times. The replicates from each time point and processed in parallel continue to cluster together, but cells grown in a different three-week time point, ‘Week 3’ cluster apart. This difference is also reflected in the Pearson correlations in the data, which fall below 0.9 (Figure 4.2B, (bottom)). The higher correlations in the replicates performed on the same-day (Figure 4.2B, bottom) are consistent with previous observations. These ‘Week 3’ libraries when compared across the 2 rounds of experiments, yield a pseudo-F ratio of 40,490, which much higher than the ‘Day 1’ libraries. It appears that cells from the same frozen stock remain somewhat similar, but when grown at different times will diverge in their scRNA-seq profiles and create transcription heterogeneity, that may confound experimental outcomes.

Figure 4.2: Growing cells at different times under identical protocols causes heterogeneity

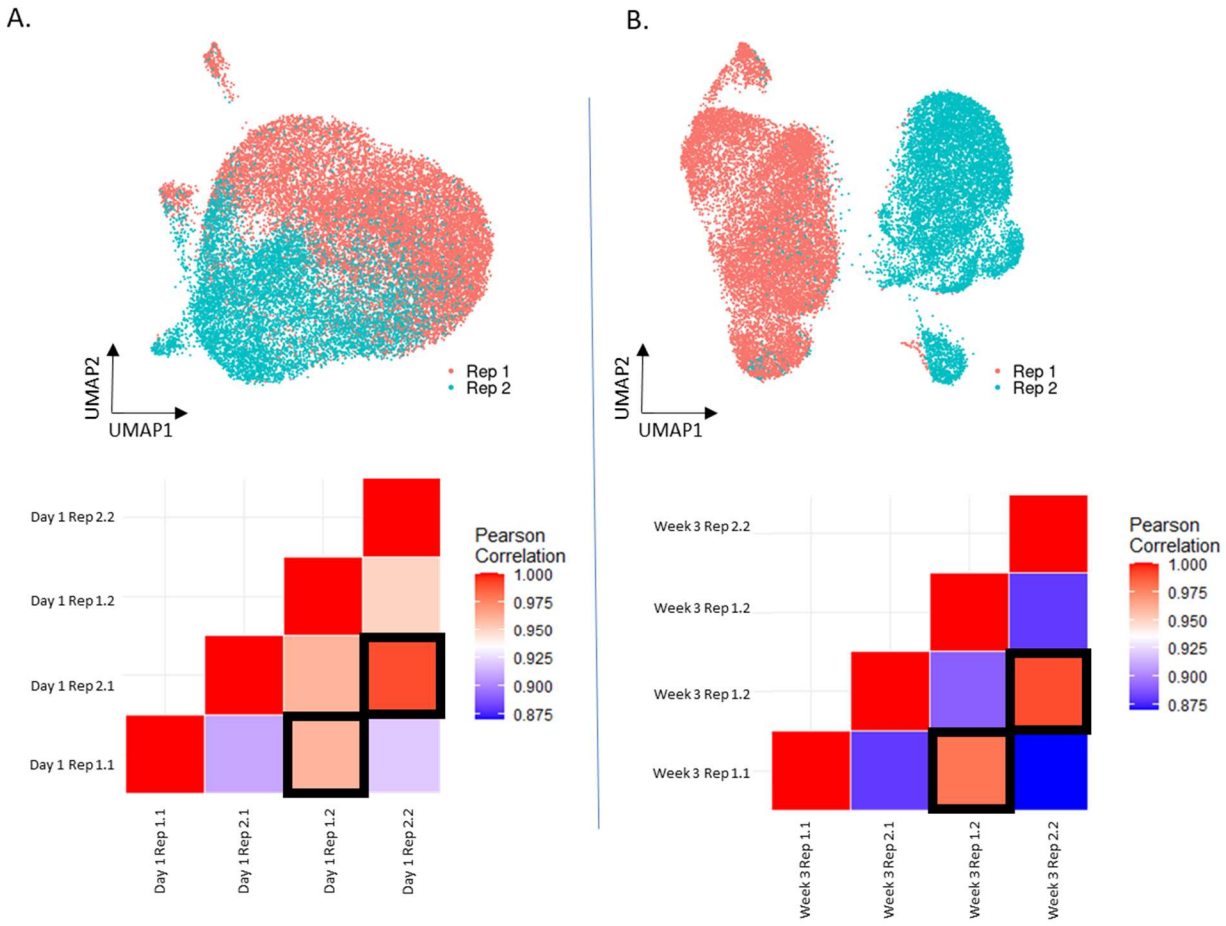


Figure 4.2: Legend

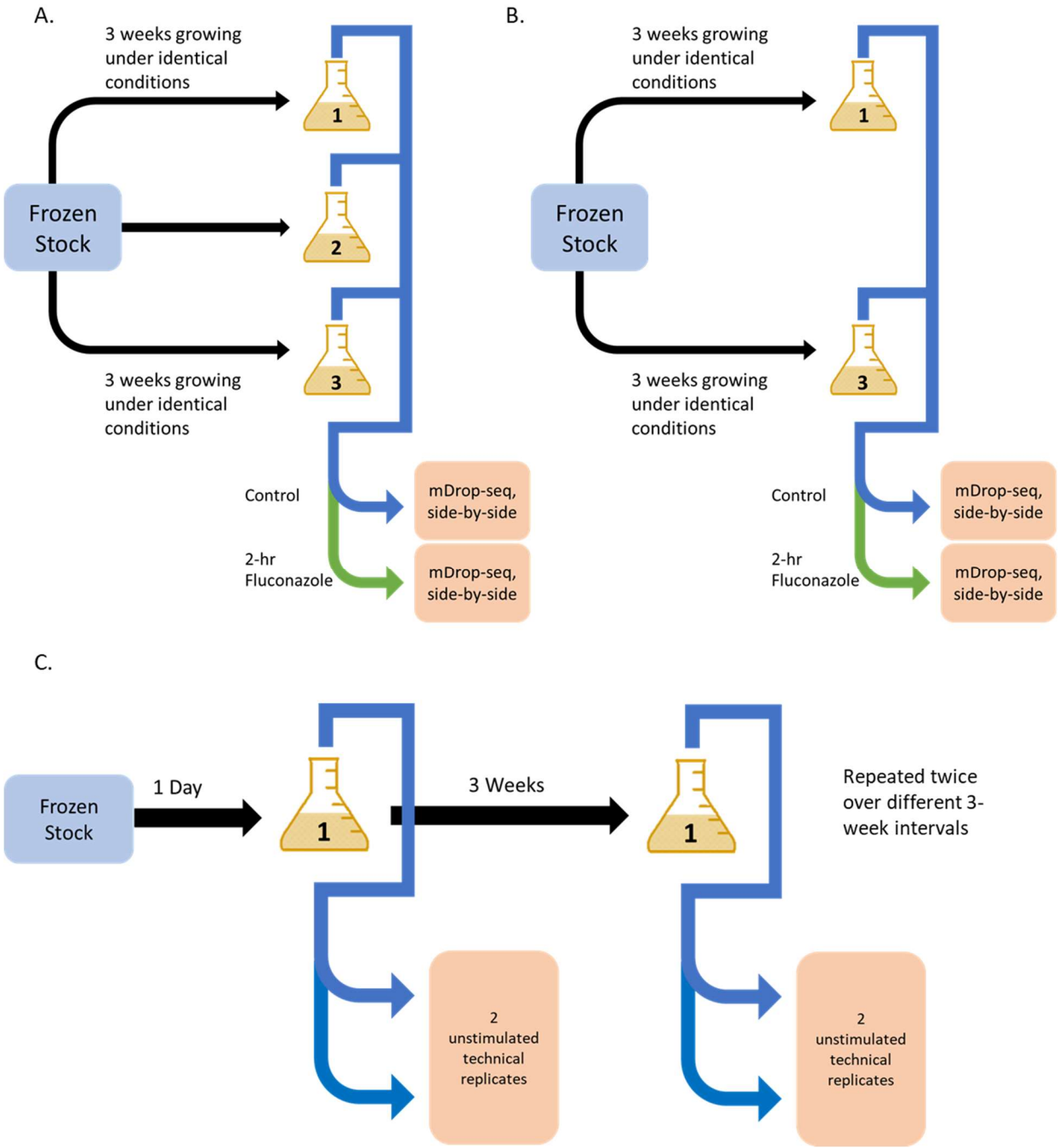
A. Top: UMAP plot all 'Day 1' replicates from the two rounds of experiments. Rep 1 and 2 refer to Biological Replicates from the 2 rounds of experiments performed at different times; each contain 2 technical replicates. Bottom: Pearson correlation between four 'Day 1' replicates, with boxes highlighting comparisons in same-day technical replicates.
B. Top: UMAP plot all 'Week 3' replicates from the two rounds of experiments. Rep 1 and 2 refer to Biological Replicates from the two rounds; each contain 2 technical replicates. Bottom: Pearson correlation between the four 'Week 3' replicates, with highlight boxes around the comparisons for same-day technical replicates.

Table 4.1: Pseudo-F Ratios between various datasets show levels of variation differences based on stimulation

The R^2, pseudo-F ratio, and p-value of various statistical tests. The data each test is referencing can be matched with an associated UMAP from which the UMAP distances were pulled.

Data Reference	Comparison	R2	F	Pr(>F)
4.1A	Replicates - All *S. cerevisiae*	0.882	129069.4	0.01
	Replicates - Controls/Intermediate	0.7485	62825.47	0.01
	Replicates - Heat Shock Stim	0.5642	17370.27	0.01
4.1B	Replicates - All *C. albicans*	0.1141	3397.8	0.01
	Replicates - Controls	0.7569	31576.17	0.01
	Replicates - Fluconazole Stim	0.1525	2923.58	0.01
4.1D	All Populations - Control and Stim	0.498	3953.05	0.01
	Control Populations 1-3	0.1462	877.991	0.01
	Fluconazole Populations 1-3	0.027	133.53	0.01
4.1E	2population-3population Controls	0.5103	15263.28	0.01
4.1F	2population-3population Fluconazole Stim	0.0005	6.67	0.01
4.2 A	Biological Replicates - Day 1	0.153	4444.94	0.01
4.2 B	Biological Replicates - Week 3	0.63	40490.4	0.01

Supplemental Figure 4.1: Experimental design diagrams for variation comparison



Supplemental Figure 4.1: Legend

A. Diagram of the experiment design for the '3-pop' experiment displayed in Figure 4.1D.
B. Diagram of the experiment design for the '2-pop' experiment displayed in Figure 4.1E,F.
C. Diagram of the experiment design to determine differences in variation between Day 1 and Week 3 yeast growth shown in Figure 4.2.

4.4 Discussion:

We sought to determine some of the sources of technical variations or batch effects that appeared within our datasets. In the previous chapters, we often analyzed our data as separate experiments, one replicate at a time, due to the differences between replicates in some libraries. In some instances, we integrated different libraries together in order to find common structures within the data that may otherwise have been lost in the variation between replicates. Batch correction and sample integration are expected in mammalian cells, for which scRNA-seq analysis was first developed. However, it was important to determine the best way to design studies in yeast cells to reduce the variation between replicates, whether they are performed on the same day or at different times or days. Microbial cells such as yeast already face significant challenges when it comes to scRNA-seq, due to higher dropout rates and higher cDNA amplification requirements; being able to control batch effects through careful study design can help mitigate additional variations from culture conditions.

Cell stress lowers transcriptomic variation between replicates

While we often worked with individual replicates, we noted early on that stressed cells, whether by heat shock or fluconazole, had higher correlations between libraries, even across replicates. This makes sense as we might expect the single-cell transcriptomes to converge when similar genetic pathways are upregulated as a stress response. The experiments performed here were designed to determine if behavior was consistent under fluconazole exposure. Through growing three separate rounds and populations of *C. albicans*, we were able to show that cells exposed to fluconazole were transcriptomically more similar and thus clustered closer together, more than unstimulated cells, even when the cells are grown side by side, under identical conditions. This is also supported by the lower pseudo-F ratio in fluconazole exposed

populations, compared to unstimulated cells (F= 134 vs. 878, Table 4.1, Data reference 4.1D).

This opens an interesting possibility in ways to lower cell variation during experiments. It can be beneficial to synchronize the cells to a similar state prior to experimentation, to get more consistent behavior across a population. A common example of this is cell cycle synchronization, in which a drug or genetic modification is used to temporarily stall the cells' cycling in a certain stage, allowing all cells within a population to synchronize [135]. However, much like how stressing cells has an obvious effect on the transcriptome, these synchronization techniques may also affect the cells' molecular pathways and gene expression. Our data here simply indicates that stress response is a potential source of transcriptional variation and needs to be considered when designing experiments.

Cells grown together remain similar compared to cells grown at different times under identical conditions

We were able to show that fluconazole-exposed cells are transcriptomically more similar and cluster closer together than their unstimulated counterparts. The hypothesis behind the experiment design in this chapter was to test if replicates of unstimulated *C. albicans* cells in Figure 4.1D would cluster far apart (similar to Figure 4.1B) or closer together under carefully controlled culture conditions. We posited earlier that cells grown in continuous culture was a likely source of variation in our mDrop-seq data. Previous experiments had indicated that three weeks of continuous growth was enough time for the cells' gene expression to change, causing this variation to increase. Yet, when three separate populations of *C. albicans* cells were grown in parallel under identical conditions, they remained more similar than we had previously seen (pseudo-F ratio = 878 vs. 31,576). This suggests that simply growing cells independently in different flasks is not enough for transcriptome divergence to occur. When we compared this

data to a previous experiment that involved two independent populations that were cultured in parallel under identical conditions, we recognized the previous variations and clustering patterns to persist. In this experiment, the two unstimulated libraries yielded pseudo-F ratios that agree with our findings that cells grown independently under the same conditions at the same time remain somewhat similar. When we combine all *C. albicans* datasets, the original observation that unstimulated libraries cluster apart and stimulated libraries cluster together is also observed.

A different set of experiments on unstimulated cells agrees with these findings. The final set of experiments was designed to show how cells grown in continuous culture for multiple weeks diverge (3 weeks, in this case), and that they diverge in different ways when grown at different times; we also compared this variation with cells grown for one day (starting from the freezer stock) also show these differences. Indeed, when cells are grown for 3 weeks vs. 1 day, we see larger differences in clustering (Figure 4.2) and pseudo-F ratio (40,490 vs. 4,445). This effect seems to be far less extreme in cells that were only given one day to grow from the freezer stocks. This indicates that a potential way to reduce variation between experiments run at different times may be to establish a cell culture from freezer stock and process cells within days of opening.

The findings of the experiments discussed in this chapter may reduce the need for integration of datasets using computational approaches. While data integration is a powerful tool that is commonly used in the context of mammalian cells, it is not yet clear whether these integration techniques are equally applicable to single cell RNA-seq data from yeasts due to issues such as high dropout rates and low copy numbers of certain genes. Even when integration techniques may be applicable, it may still be important to design studies that reduce batch effects

in order to provide the most power in elucidating complex biological processes in single-cell data.real biology within complex single-cell data.

4.6: Acknowledgement of Work Performed

I acknowledge those who contributed to the work presented here. Trevor Wood, Dylan Cook, Allison Hohreiter, Dr. Ran Zhou, and Dr. Katie Mika contributed to the processing of multi-population experiments. Project and experiment design was guided by Dr. Anindita Basu. Dr. Bingping Xie assisted with data analysis. All other experimentation and analysis present in this chapter were performed by Ryan Dohn.

Chapter 5: Future Directions:

Expanding mDrop-seq to further yeast species

In this work, we have demonstrated that mDrop-seq is capable of processing *S. cerevisiae,* a model yeast organism and *C. albicans,* a pathogen of global health concern. We believe that mDrop-seq is applicable to most yeast species for single-cell genomics study at low cost and easy adaptation. This may include common lab species such as *S. pombe*, or clinically relevant species such as *C. auris*. Notably, we are interested in *C. auris* due to its emergence as a new pathogen with widespread resistance to many classes of antifungal drugs. Despite being discovered only in 2009, this species of yeast is quickly growing to be a global health concern due to several properties advantageous to its survival. *C. auris* has shown itself to be commonly drug resistant, with several of the strains detected showing resistance across the major antifungal classes including a diverse array of ERG mutations to protect against azoles like fluconazole [136]. Additionally, this species has been noted to spread from person to person, capability to quickly colonize the human skin, as well as survive on abiotic surfaces for extended periods of time [137]. These adaptations make *C. auris* a tempting target for scRNA-seq. Understanding how adaptations under stress may be crucial to responding to this new threat. It will take work beyond the scope of mDrop-seq to achieve high-quality, informative single cell data, however. For example, there is still work to be done to achieve better genomic annotation for this species and its different strains isolated from the clinical setting. The lack of a highly annotated genome will be a significant roadblock for further development and application.

Investigating the development of drug resistance within yeast species

Much like the multidrug resistance in *C. auris*, drug resistance in other pathogenic yeast species is of great concern, as fungal infections are on the rise [138]. In this work, we observed the effects of fluconazole on *C. albicans* cells. However, there are other classes of antifungal drugs to investigate the cellular response and adaption in yeasts with single-cell resolution. Resistance in some form to each of the four major antifungal drug classes discussed earlier in this dissertation have been observed. Cytotoxic drugs will be harder to study than fungistatic drugs like fluconazole, meaning that experiments and exposure amounts must be carefully planned to achieve usable single-cell genomics data. The CDC has assembled panels of drug-resistant yeast species including various *Candida* species and a few *S. cerevisiae* strains. We have obtained a full panel of drug-resistant yeast species, primarily consisting of *C. auris* strains. It is likely that probing drug-resistant yeast species and expanding mDrop-seq to *C. auris* strains would occur together.

Being able to measure transcriptomic differences between drug resistant and susceptible strains can inform how drug resistance develops. While there are strains that contain specific mutations that confer permanent resistance, yeasts are also capable of developing contact resistance to antifungal drug exposure. Data presented in chapter 3 displays contact resistance on some level as *C. albicans* cells upregulate the ERG pathway to try and overcome fluconazole interference. The next step is to use mDrop-seq to see similar responses to other antifungal drugs: Caspofungin (an echinocandin) and Amphotericin B (a polyene). Much like with fluconazole, we expect to see changes in genetic pathways that these drugs target. It is likely that these drug exposures will need to occur below the MIC for any given yeast strain to prevent cell death prior to droplet encapsulation and processing.

Interaction between *C. albicans* and host macrophages during infection

While drug resistance is one avenue of research to fight fungal infections such as Candidiasis, understanding the role of transcriptional heterogeneity in hosts and pathogens that may lead to different infection outcomes is also open to investigation. Using a low throughput system, Muñoz *et al* [38] managed to capture single-cell data for mouse macrophages and *C. albicans* cells that were endocytosed by the macrophages through the use of a fluorescent reporter system. In order to infect macrophages, yeast cells were simply added to the plate at an average Multiplicity of Infection (MOI). This methodology did not control MOI at single cell level; individual cells across the plate may experience different numbers of yeast in their vicinity. With the advent of high-throughput yeast scRNA-seq techniques, we can also characterize the transcriptomes of matched host-pathogen pairs. Using droplet microfluidics, we can create chambers to isolate and infect host-pathogen pairs at controlled MOI.

Bibliography

[1] M. N. Bainbridge *et al.*, "Analysis of the prostate cancer cell line LNCaP transcriptome using a sequencing-by-synthesis approach.," *BMC Genomics*, vol. 7, p. 246, Sep. 2006.

[2] A. P. M. Weber, K. L. Weber, K. Carr, C. Wilkerson, and J. B. Ohlrogge, "Sampling the Arabidopsis transcriptome with massively parallel pyrosequencing.," *Plant Physiol.*, vol. 144, no. 1, pp. 32–42, May 2007.

[3] R. Lister *et al.*, "Highly integrated single-base resolution maps of the epigenome in Arabidopsis.," *Cell*, vol. 133, no. 3, pp. 523–536, May 2008.

[4] U. Nagalakshmi *et al.*, "The transcriptional landscape of the yeast genome defined by RNA sequencing.," *Science*, vol. 320, no. 5881, pp. 1344–1349, Jun. 2008.

[5] A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, and B. Wold, "Mapping and quantifying mammalian transcriptomes by RNA-Seq.," *Nat. Methods*, vol. 5, no. 7, pp. 621–628, Jul. 2008.

[6] J. C. Marioni, C. E. Mason, S. M. Mane, M. Stephens, and Y. Gilad, "RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays.," *Genome Res.*, vol. 18, no. 9, pp. 1509–1517, Sep. 2008.

[7] F. Tang *et al.*, "mRNA-Seq whole-transcriptome analysis of a single cell.," *Nat. Methods*, vol. 6, no. 5, pp. 377–382, May 2009.

[8] S. Islam *et al.*, "Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq.," *Genome Res.*, vol. 21, no. 7, pp. 1160–1167, Jul. 2011.

[9] E. Z. Macosko *et al.*, "Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets," *Cell*, vol. 161, no. 5, pp. 1202–1214, 2015.

[10] A. M. Klein *et al.*, "Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells.," *Cell*, vol. 161, no. 5, pp. 1187–1201, May 2015.

[11] L. D. Goldstein *et al.*, "Massively parallel nanowell-based single-cell gene expression profiling.," *BMC Genomics*, vol. 18, no. 1, p. 519, Jul. 2017.

[12] J. Cao *et al.*, "Comprehensive single-cell transcriptional profiling of a multicellular organism.," *Science*, vol. 357, no. 6352, pp. 661–667, Aug. 2017.

[13] A. B. Rosenberg *et al.*, "Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding.," *Science*, vol. 360, no. 6385, pp. 176–182, Apr. 2018.

[14] T. S. Andrews and M. Hemberg, "Identifying cell populations with scRNASeq," *Mol. Aspects Med.*, vol. 59, pp. 114–122, 2018.

[15] P. A. Szabo *et al.*, "Single-cell transcriptomics of human T cells reveals tissue and activation signatures in health and disease," *Nat. Commun.*, vol. 10, no. 1, 2019.

[16] D. T. Montoro *et al.*, "A revised airway epithelial hierarchy includes CFTR-expressing ionocytes.," *Nature*, vol. 560, no. 7718, pp. 319–324, Aug. 2018.

[17] A. Raj and A. van Oudenaarden, “Nature, Nurture, or Chance: Stochastic Gene Expression and Its Consequences,” *Cell*, vol. 135, no. 2, pp. 216–226, Oct. 2008.

[18] Y. H. Choi and J. K. Kim, “Dissecting cellular heterogeneity using single-cell RNA sequencing,” *Mol. Cells*, vol. 42, no. 3, pp. 189–199, 2019.

[19] G. X. Y. Zheng *et al.*, “Massively parallel digital transcriptional profiling of single cells,” *Nat. Commun.*, vol. 8, 2017.

[20] Y. Zhang, J. Gao, Y. Huang, and J. Wang, “Recent Developments in Single-Cell RNA-Seq of Microorganisms,” *Biophys. J.*, vol. 115, no. 2, pp. 173–180, 2018.

[21] A.-E. Saliba, S. C Santos, and J. Vogel, “New RNA-seq approaches for the study of bacterial pathogens.,” *Curr. Opin. Microbiol.*, vol. 35, pp. 78–87, Feb. 2017.

[22] K. Campbell, J. Vowinckel, and M. Ralser, “Cell-to-cell heterogeneity emerges as consequence of metabolic cooperation in a synthetic yeast community.,” *Biotechnol. J.*, vol. 11, no. 9, pp. 1169–1178, Sep. 2016.

[23] S. Helaine and D. W. Holden, “Heterogeneity of intracellular replication of bacterial pathogens.,” *Curr. Opin. Microbiol.*, vol. 16, no. 2, pp. 184–191, Apr. 2013.

[24] B. Cerulus, A. M. New, K. Pougach, and K. J. Verstrepen, “Noise and Epigenetic Inheritance of Single-Cell Division Times Influence Population Fitness.,” *Curr. Biol.*, vol. 26, no. 9, pp. 1138–1147, May 2016.

[25] B. Cerulus *et al.*, “Transition between fermentation and respiration determines history-dependent behavior in fluctuating carbon sources.,” *Elife*, vol. 7, Oct. 2018.

[26] N. A. R. Gow and B. Hube, “Importance of the Candida albicans cell wall during commensalism and infection,” *Curr. Opin. Microbiol.*, vol. 15, no. 4, pp. 406–412, 2012.

[27] G. Lesage and H. Bussey, “Cell wall assembly in Saccharomyces cerevisiae.,” *Microbiol. Mol. Biol. Rev.*, vol. 70, no. 2, pp. 317–343, Jun. 2006.

[28] T. J. Silhavy, D. Kahne, and S. Walker, “The bacterial cell envelope.,” *Cold Spring Harb. Perspect. Biol.*, vol. 2, no. 5, p. a000414, May 2010.

[29] A. Kuchina *et al.*, “Microbial single-cell RNA sequencing by split-pool barcoding.,” *Science*, vol. 371, no. 6531, Feb. 2021.

[30] Y. Kang, M. H. Norris, J. Zarzycki-Siek, W. C. Nierman, S. P. Donachie, and T. T. Hoang, “Transcript amplification from single bacterium for transcriptome analysis,” *Genome Res.*, vol. 21, no. 6, pp. 925–935, 2011.

[31] D. Zenklusen, D. R. Larson, and R. H. Singer, “Single-RNA counting reveals alternative modes of gene expression in yeast,” *Nat. Struct. Mol. Biol.*, vol. 15, no. 12, pp. 1263–1271, 2008.

[32] B. Hwang, J. H. Lee, and D. Bang, “Single-cell RNA sequencing technologies and bioinformatics pipelines.,” *Exp. Mol. Med.*, vol. 50, no. 8, pp. 1–14, Aug. 2018.

[33] J. Zhao, L. Hyman, and C. Moore, “Formation of mRNA 3’ ends in eukaryotes:

mechanism, regulation, and interrelationships with other steps in mRNA synthesis.," *Microbiol. Mol. Biol. Rev.*, vol. 63, no. 2, pp. 405–445, Jun. 1999.

[34] M. Nadal-ribelles, S. Islam, W. Wei, and P. Latorre, "Sensitive high-throughput single-cell RNA-Seq reveals within- clonal transcript-correlations in yeast populations," *Nat. Microbiol.*, vol. 4, no. 4, pp. 683–692, 2019.

[35] A. Jariani *et al.*, "A new protocol for single-cell RNA-seq reveals stochastic gene expression during lag phase in budding yeast," *Elife*, vol. 9, p. e55320, May 2020.

[36] C. A. Jackson, D. M. Castro, G. Saldi, R. Bonneau, and D. Gresham, "Gene regulatory network reconstruction using single-cell RNA sequencing of barcoded genotypes in diverse environments," *Elife*, vol. 9, pp. 1–34, Jan. 2020.

[37] G. Urbonaite, J. T. H. Lee, P. Liu, G. E. Parada, M. Hemberg, and M. Acar, "A yeast-optimized single-cell transcriptomics platform elucidates how mycophenolic acid and guanine alter global mRNA levels," *Commun. Biol.*, vol. 4, no. 1, p. 822, 2021.

[38] J. F. Muñoz *et al.*, "Coordinated host-pathogen transcriptional dynamics revealed using sorted subpopulations and single macrophages infected with Candida albicans," *Nat. Commun.*, vol. 10, no. 1, p. 1607, Dec. 2019.

[39] M. A. Pfaller and D. J. Diekema, "Epidemiology of invasive candidiasis: A persistent public health problem," *Clin. Microbiol. Rev.*, vol. 20, no. 1, pp. 133–163, 2007.

[40] P. Muñoz *et al.*, "Saccharomyces cerevisiae Fungemia: An Emerging Infectious Disease," *Clin. Infect. Dis.*, vol. 40, no. 11, pp. 1625–1634, Jun. 2005.

[41] F. L. Mayer, D. Wilson, and B. Hube, "Candida albicans pathogenicity mechanisms," *Virulence*, vol. 4, no. 2, pp. 119–128, 2013.

[42] "Centers for Disease Control and Prevention: Antibiotic Resistance Threats in the United States, 2019," 2019.

[43] B. R. Levin and D. E. Rozen, "Non-inherited antibiotic resistance.," *Nat. Rev. Microbiol.*, vol. 4, no. 7, pp. 556–562, Jul. 2006.

[44] H. Karathia, E. Vilaprinyo, A. Sorribas, and R. Alves, "Saccharomyces cerevisiae as a model organism: a comparative study.," *PLoS One*, vol. 6, no. 2, p. e16015, Feb. 2011.

[45] A. Goffeau *et al.*, "Life with 6000 Genes," *Science (80-.).*, vol. 274, no. 5287, pp. 546–567, 1996.

[46] G. Giaever *et al.*, "Functional profiling of the Saccharomyces cerevisiae genome.," *Nature*, vol. 418, no. 6896, pp. 387–391, Jul. 2002.

[47] E. A. Winzeler *et al.*, "Functional Characterization of the *S. cerevisiae* Genome by Gene Deletion and Parallel Analysis," *Science (80-.).*, vol. 285, no. 5429, pp. 901–906, 1999.

[48] G. M. Jones *et al.*, "A systematic library for comprehensive overexpression screens in Saccharomyces cerevisiae.," *Nat. Methods*, vol. 5, no. 3, pp. 239–241, Mar. 2008.

[49] W.-K. Huh *et al.*, "Global analysis of protein localization in budding yeast.," *Nature*, vol.

425, no. 6959, pp. 686–691, Oct. 2003.

[50] N. J. Krogan *et al.*, “Global landscape of protein complexes in the yeast Saccharomyces cerevisiae.,” *Nature*, vol. 440, no. 7084, pp. 637–643, Mar. 2006.

[51] P. Uetz *et al.*, “A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae.,” *Nature*, vol. 403, no. 6770, pp. 623–627, Feb. 2000.

[52] A. H. Y. Tong *et al.*, “Global Mapping of the Yeast Genetic Interaction Network,” *Science (80-.).*, vol. 303, no. 5659, pp. 808–813, 2004.

[53] M. Shehadul Islam, A. Aryasomayajula, and P. R. Selvaganapathy, “A Review on Macroscale and Microscale Cell Lysis Methods.,” *Micromachines*, vol. 8, no. 3. Mar-2017.

[54] F. M. Klis, C. G. de Koster, and S. Brul, “Cell wall-related bionumbers and bioestimates of Saccharomyces cerevisiae and Candida albicans.,” *Eukaryot. Cell*, vol. 13, no. 1, pp. 2–9, Jan. 2014.

[55] K. A. Morano, C. M. Grant, and W. S. Moye-Rowley, “The response to heat shock and oxidative stress in Saccharomyces cerevisiae.,” *Genetics*, vol. 190, no. 4, pp. 1157–1195, Apr. 2012.

[56] A. P. Gasch *et al.*, “Genomic Expression Programs in the Response of Yeast Cells to Environmental Changes,” *Mol. Biol. Cell*, vol. 11, no. 12, pp. 4241–4257, 2000.

[57] M. T. Martínez-Pastor, G. Marchler, C. Schüller, A. Marchler-Bauer, H. Ruis, and F. Estruch, “The Saccharomyces cerevisiae zinc finger proteins Msn2p and Msn4p are required for transcriptional induction through the stress response element (STRE).,” *EMBO J.*, vol. 15, no. 9, pp. 2227–2235, May 1996.

[58] N. Habib *et al.*, “Massively parallel single-nucleus RNA-seq with DroNc-seq,” *Nat. Methods*, vol. 14, no. 10, pp. 955–958, 2017.

[59] J. Köster and S. Rahmann, “Snakemake-a scalable bioinformatics workflow engine,” *Bioinformatics*, vol. 28, no. 19, pp. 2520–2522, 2012.

[60] S. Andrews and Babraham Bioinformatics, “FastQC: A quality control tool for high throughput sequence data,” *Manual*. 2010.

[61] T. Smith, A. Heger, and I. Sudbery, “UMI-tools: Modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy,” *Genome Res.*, 2017.

[62] A. Dobin *et al.*, “STAR: Ultrafast universal RNA-seq aligner,” *Bioinformatics*, vol. 29, no. 1, pp. 15–21, 2013.

[63] Y. Liao, G. K. Smyth, and W. Shi, “FeatureCounts: An efficient general purpose program for assigning sequence reads to genomic features,” *Bioinformatics*, vol. 30, no. 7, pp. 923–930, 2014.

[64] L. McInnes, J. Healy, N. Saul, and L. Großberger, “UMAP: Uniform Manifold Approximation and Projection,” *J. Open Source Softw.*, vol. 3, no. 29, p. 861, 2018.

[65] P. T. Spellman *et al.*, “Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization,” *Mol. Biol. Cell*, vol. 9, no. 12, pp. 3273–3297, Dec. 1998.

[66] T. Stuart and R. Satija, “Integrative single-cell analysis,” *Nat. Rev. Genet.*, vol. 20, no. 5, pp. 257–272, 2019.

[67] X. Qiu *et al.*, “Reversed graph embedding resolves complex single-cell trajectories,” *Nat. Methods*, vol. 14, no. 10, pp. 979–982, 2017.

[68] Q. Mao, L. Wang, S. Goodison, and Y. Sun, “Dimensionality Reduction Via Graph Structure Learning,” in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 10-14 August 2015; Sydney, Australia*, 2015, pp. 765–774.

[69] S. Frankel, R. Sohn, and L. Leinwand, “The use of sarkosyl in generating soluble protein after bacterial expression.,” *Proc. Natl. Acad. Sci.*, vol. 88, no. 4, pp. 1192–1196, Feb. 1991.

[70] K. Kitamura and Y. Yamamoto, “Purification and properties of an enzyme, zymolyase, which lyses viable yeast cells.,” *Arch. Biochem. Biophys.*, vol. 153, no. 1, pp. 403–406, Nov. 1972.

[71] F. Miura *et al.*, “Absolute quantification of the budding yeast transcriptome by means of competitive PCR between genomic and complementary DNAs,” *BMC Genomics*, vol. 9, no. 1, p. 574, 2008.

[72] A. Silva *et al.*, “Regulation of transcription elongation in response to osmostress,” *PLoS Genet.*, vol. 13, no. 11, pp. 1–24, 2017.

[73] T. Stuart *et al.*, “Comprehensive Integration of Single-Cell Data.,” *Cell*, vol. 177, no. 7, pp. 1888-1902.e21, Jun. 2019.

[74] K. Richter, M. Haslbeck, and J. Buchner, “The Heat Shock Response: Life on the Verge of Death,” *Mol. Cell*, vol. 40, no. 2, pp. 253–266, 2010.

[75] Y. B. Tzur, E. Winter, J. Gao, T. Hashimshony, I. Yanai, and M. P. Colaiácovo, “Spatiotemporal Gene Expression Analysis of the Caenorhabditis elegans Germline Uncovers a Syncytial Expression Switch,” *Genetics*, vol. 210, no. 2, pp. 587–605, Oct. 2018.

[76] J. A. Farrell, Y. Wang, S. J. Riesenfeld, K. Shekhar, A. Regev, and A. F. Schier, “Single-cell reconstruction of developmental trajectories during zebrafish embryogenesis,” *Science (80-.).*, vol. 360, no. 3692, 2018.

[77] M. M. Ariss, A. B. M. M. K. Islam, M. Critcher, M. P. Zappia, and M. V Frolov, “Single cell RNA-sequencing identifies a metabolic aspect of apoptosis in Rbf mutant,” *Nat. Commun.*, vol. 9, no. 5024, pp. 1–13, 2018.

[78] V. A. Schneider *et al.*, “Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly,” *Genome Res.*, vol. 27, no. 5, pp. 849–864, 2017.

[79] “Saccharomyces cerevisiae (ID 15) - Genome - NCBI.” [Online]. Available: https://www.ncbi.nlm.nih.gov/genome/?term=s+cerevisiae. [Accessed: 21-May-2020].

[80] “Candida albicans (ID 21) - Genome - NCBI.” [Online]. Available: https://www.ncbi.nlm.nih.gov/genome/?term=Candida albicans. [Accessed: 21-May-2020].

[81] M. A. Pfaller and D. J. Diekema, “Epidemiology of invasive mycoses in North America.,” *Crit. Rev. Microbiol.*, vol. 36, no. 1, pp. 1–53, 2010.

[82] P. G. Pappas, M. S. Lionakis, M. C. Arendrup, L. Ostrosky-Zeichner, and B. J. Kullberg, “Invasive candidiasis,” *Nat. Rev. Dis. Prim.*, vol. 4, no. 1, p. 18026, 2018.

[83] M.-E. Bougnoux *et al.*, “Multilocus sequence typing reveals intrafamilial transmission and microevolutions of Candida albicans isolates from the human digestive tract.,” *J. Clin. Microbiol.*, vol. 44, no. 5, pp. 1810–1820, May 2006.

[84] A. Y. Koh, J. R. Köhler, K. T. Coggshall, N. Van Rooijen, and G. B. Pier, “Mucosal damage and neutropenia are required for Candida albicans dissemination.,” *PLoS Pathog.*, vol. 4, no. 2, p. e35, Feb. 2008.

[85] H. J. Lo, J. R. Köhler, B. DiDomenico, D. Loebenberg, A. Cacciapuoti, and G. R. Fink, “Nonfilamentous C. albicans mutants are avirulent.,” *Cell*, vol. 90, no. 5, pp. 939–949, Sep. 1997.

[86] A. M. Murad *et al.*, “NRG1 represses yeast-hypha morphogenesis and hypha-specific gene expression in Candida albicans.,” *EMBO J.*, vol. 20, no. 17, pp. 4742–4752, Sep. 2001.

[87] S. P. Saville, A. L. Lazzell, C. Monteagudo, and J. L. Lopez-Ribot, “Engineered control of cell morphology in vivo reveals distinct roles for yeast and filamentous forms of Candida albicans during infection.,” *Eukaryot. Cell*, vol. 2, no. 5, pp. 1053–1060, Oct. 2003.

[88] N. A. R. Gow, F. L. van de Veerdonk, A. J. P. Brown, and M. G. Netea, “Candida albicans morphogenesis and host defence: discriminating invasion from colonization.,” *Nat. Rev. Microbiol.*, vol. 10, no. 2, pp. 112–122, Dec. 2011.

[89] R. T. Wheeler, D. Kombe, S. D. Agarwala, and G. R. Fink, “Dynamic, morphotype-specific Candida albicans beta-glucan exposure during infection and drug treatment.,” *PLoS Pathog.*, vol. 4, no. 12, p. e1000227, Dec. 2008.

[90] M. G. Netea, L. A. B. Joosten, J. W. M. van der Meer, B.-J. Kullberg, and F. L. van de Veerdonk, “Immune defence against Candida fungal infections.,” *Nat. Rev. Immunol.*, vol. 15, no. 10, pp. 630–642, Oct. 2015.

[91] C. Murciano *et al.*, “Evaluation of the role of Candida albicans agglutinin-like sequence (Als) proteins in human oral epithelial cell interactions.,” *PLoS One*, vol. 7, no. 3, p. e33362, 2012.

[92] W. Zhu and S. G. Filler, “Interactions of Candida albicans with epithelial cells.,” *Cell. Microbiol.*, vol. 12, no. 3, pp. 273–282, Mar. 2010.

[93] S. Fanning and A. P. Mitchell, “Fungal biofilms.,” *PLoS Pathog.*, vol. 8, no. 4, p. e1002585, 2012.

[94] J. S. Finkel and A. P. Mitchell, “Genetic control of Candida albicans biofilm development.,” *Nat. Rev. Microbiol.*, vol. 9, no. 2, pp. 109–118, Feb. 2011.

[95] D. A. Davis, “How human pathogenic fungi sense and adapt to pH: the link to virulence.,” *Curr. Opin. Microbiol.*, vol. 12, no. 4, pp. 365–370, Aug. 2009.

[96] M. Brock, “Fungal metabolism in host niches.,” *Curr. Opin. Microbiol.*, vol. 12, no. 4, pp. 371–376, Aug. 2009.

[97] E. M. Carmona and A. H. Limper, “Overview of Treatment Approaches for Fungal Infections.,” *Clin. Chest Med.*, vol. 38, no. 3, pp. 393–402, Sep. 2017.

[98] A. J. Coukell and R. N. Brogden, “Liposomal amphotericin B. Therapeutic use in the management of fungal infections and visceral leishmaniasis.,” *Drugs*, vol. 55, no. 4, pp. 585–612, Apr. 1998.

[99] M. Shafiei, L. Peyton, M. Hashemzadeh, and A. Foroumadi, “History of the development of antifungal azoles: A review on structures, SAR, and mechanism of action,” *Bioorg. Chem.*, vol. 104, p. 104240, 2020.

[100] J. D. Morrow, “Fluconazole: a new triazole antifungal agent.,” *Am. J. Med. Sci.*, vol. 302, no. 2, pp. 129–132, Aug. 1991.

[101] A. Patil and S. Majumdar, “Echinocandins in antifungal pharmacotherapy.,” *J. Pharm. Pharmacol.*, vol. 69, no. 12, pp. 1635–1660, Dec. 2017.

[102] C. M. Hull *et al.*, “Two clinical isolates of Candida glabrata exhibiting reduced sensitivity to amphotericin B both harbor mutations in ERG2.,” *Antimicrob. Agents Chemother.*, vol. 56, no. 12, pp. 6417–6421, Dec. 2012.

[103] A. C. Mesa-Arango *et al.*, “Cell Wall Changes in Amphotericin B-Resistant Strains from Candida tropicalis and Relationship with the Immune Responses Elicited by the Host.,” *Antimicrob. Agents Chemother.*, vol. 60, no. 4, pp. 2326–2335, Apr. 2016.

[104] W. Posch, M. Blatzer, D. Wilflingseder, and C. Lass-Flörl, “Aspergillus terreus: Novel lessons learned on amphotericin B resistance.,” *Med. Mycol.*, vol. 56, no. suppl_1, pp. 73–82, Apr. 2018.

[105] J. Houšť, J. Spížek, and V. Havlíček, “Antifungal Drugs.,” *Metabolites*, vol. 10, no. 3, Mar. 2020.

[106] P. M. Silver, B. G. Oliver, and T. C. White, “Role of Candida albicans transcription factor Upc2p in drug resistance and sterol metabolism.,” *Eukaryot. Cell*, vol. 3, no. 6, pp. 1391–1397, Dec. 2004.

[107] N. M. Revie, K. R. Iyer, N. Robbins, and L. E. Cowen, “Antifungal drug resistance: evolution, mechanisms and impact.,” *Curr. Opin. Microbiol.*, vol. 45, pp. 70–76, Oct. 2018.

[108] L. Vermeersch, A. Jariani, J. Helsen, B. M. Heineike, and K. J. Verstrepen, “Single-Cell

RNA Sequencing in Yeast Using the 10× Genomics Chromium Device.," *Methods Mol. Biol.*, vol. 2477, pp. 3–20, 2022.

[109] P. Côte, H. Herve, and M. Whiteway, "Transcriptional Analysis of the Candida albicans Cell Cycle," *Mol. Biol. Cell*, vol. 20, no. 14, pp. 3363–3373, 2009.

[110] J. Kim and P. Sudbery, "Candida albicans, a major human fungal pathogen," *J. Microbiol.*, vol. 49, no. 2, pp. 171–177, 2011.

[111] L. Issi *et al.*, "Zinc Cluster Transcription Factors Alter Virulence in Candida albicans.," *Genetics*, vol. 205, no. 2, pp. 559–576, Feb. 2017.

[112] B. Slutsky, M. Staebell, J. Anderson, L. Risen, M. Pfaller, and D. R. Soll, "'White-opaque transition': A second high-frequency switching system in Candida albicans," *J. Bacteriol.*, vol. 169, no. 1, pp. 189–197, 1987.

[113] C.-Y. Lan *et al.*, "Metabolic specialization associated with phenotypic switching in Candidaalbicans.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 99, no. 23, pp. 14907–14912, Nov. 2002.

[114] F. Cottier *et al.*, "The transcriptional response of Candida albicans to weak organic acids, carbon source, and MIG1 inactivation unveils a role for HGT16 in mediating the fungistatic effect of acetic acid," *G3 Genes, Genomes, Genet.*, vol. 7, no. 11, pp. 3597–3604, 2017.

[115] E. L. Berkow and S. R. Lockhart, "Fluconazole resistance in Candida species: a current perspective.," *Infect. Drug Resist.*, vol. 10, pp. 237–245, 2017.

[116] Pfizer, "DIFLUCAN (Fluconazole Tablets) (Fluconazole Injection - for intravenous infusion only) (Fluconazole for Oral Suspension)," *FDA drug label*, pp. 1–37, 2011.

[117] H. M. H. N. Bandara, D. L. A. Wood, I. Vanwonterghem, P. Hugenholtz, B. P. K. Cheung, and L. P. Samaranayake, "Fluconazole resistance in Candida albicans is induced by Pseudomonas aeruginosa quorum sensing," *Sci. Rep.*, vol. 10, no. 1, p. 7769, 2020.

[118] K. W. Henry, J. T. Nickels, and T. D. Edlind, "Upregulation of ERG genes in Candida species by azoles and other sterol biosynthesis inhibitors.," *Antimicrob. Agents Chemother.*, vol. 44, no. 10, pp. 2693–2700, Oct. 2000.

[119] R. Leber *et al.*, "Molecular mechanism of terbinafine resistance in Saccharomyces cerevisiae.," *Antimicrob. Agents Chemother.*, vol. 47, no. 12, pp. 3890–3900, Dec. 2003.

[120] J. I. Castrillo *et al.*, "Growth control of the eukaryote cell: a systems biology study in yeast.," *J. Biol.*, vol. 6, no. 2, p. 4, 2007.

[121] S. Bhattacharya, B. D. Esquivel, and T. C. White, "Overexpression or Deletion of Ergosterol Biosynthesis Genes Alters Doubling Time, Response to Stress Agents, and Drug Susceptibility in Saccharomyces cerevisiae.," *MBio*, vol. 9, no. 4, Jul. 2018.

[122] K. Fahrner, J. Yarger, and L. Hereford, "Yeast histone mRNA is polyadenylated," *Nucleic Acids Res.*, vol. 8, no. 23, pp. 5725–5737, Dec. 1980.

[123] A. Matejuk *et al.*, "Peptide-based Antifungal Therapies against Emerging Infections.,"

Drugs Future, vol. 35, no. 3, p. 197, Mar. 2010.

[124] L. E. Cowen and W. J. Steinbach, "Stress, drugs, and evolution: the role of cellular signaling in fungal drug resistance.," *Eukaryot. Cell*, vol. 7, no. 5, pp. 747–764, May 2008.

[125] S. Costa-de-Oliveira and A. G. Rodrigues, "Candida albicans Antifungal Resistance and Tolerance in Bloodstream Infections: The Triad Yeast-Host-Antifungal," *Microorganisms*, vol. 8, no. 2, 2020.

[126] P. Brennecke *et al.*, "Accounting for technical noise in single-cell RNA-seq experiments," *Nat. Methods*, vol. 10, no. 11, pp. 1093–1095, 2013.

[127] S. C. Hicks, F. W. Townes, M. Teng, and R. A. Irizarry, "Missing data and technical variability in single-cell RNA-sequencing experiments.," *Biostatistics*, vol. 19, no. 4, pp. 562–578, Oct. 2018.

[128] L. Haghverdi, A. T. L. Lun, M. D. Morgan, and J. C. Marioni, "Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors.," *Nat. Biotechnol.*, vol. 36, no. 5, pp. 421–427, Jun. 2018.

[129] P. V Kharchenko, L. Silberstein, and D. T. Scadden, "Bayesian approach to single-cell differential expression analysis," *Nat. Methods*, vol. 11, no. 7, pp. 740–742, 2014.

[130] H. T. N. Tran *et al.*, "A benchmark of batch-effect correction methods for single-cell RNA sequencing data," *Genome Biol.*, vol. 21, no. 1, p. 12, 2020.

[131] P. D. Schloss, "Identifying and Overcoming Threats to Reproducibility, Replicability, Robustness, and Generalizability in Microbiome Research," *MBio*, vol. 9, no. 3, pp. e00525-18, 2018.

[132] Y. Wang and K.-A. LêCao, "Managing batch effects in microbiome data," *Brief. Bioinform.*, vol. 21, no. 6, pp. 1954–1970, 2019.

[133] M. J. Anderson, "Permutational Multivariate Analysis of Variance (PERMANOVA)," in *Wiley StatsRef: Statistics Reference Online*, John Wiley & Sons, Ltd, 2017, pp. 1–15.

[134] J. Oksanen *et al.*, "vegan: Community Ecology Package." 2022.

[135] M. A. Juanes, "Methods of Synchronization of Yeast Cells for the Analysis of Cell Cycle Progression.," *Methods Mol. Biol.*, vol. 1505, pp. 19–34, 2017.

[136] S. Sarma *et al.*, "Candidemia caused by amphotericin B and fluconazole resistant Candida auris.," *Indian journal of medical microbiology*, vol. 31, no. 1. United States, pp. 90–91, 2013.

[137] R. M. Welsh *et al.*, "Survival, Persistence, and Isolation of the Emerging Multidrug-Resistant Pathogenic Yeast Candida auris on a Plastic Health Care Surface.," *J. Clin. Microbiol.*, vol. 55, no. 10, pp. 2996–3005, Oct. 2017.

[138] J. Berman and D. J. Krysan, "Drug resistance and tolerance in fungi.," *Nat. Rev. Microbiol.*, vol. 18, no. 6, pp. 319–331, Jun. 2020.