UNIVERSITY OF CALIFORNIA, MERCED

Predicting novel transcription factor-target gene interactions in the *Candida albicans* biofilm network using machine learning

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Quantitative and Systems Biology

by

Akshay Deepak Paropkari

Committee in charge:

Professor Aaron Hernday, Chair of Advisory Committee

Professor Juris Grasis

Dr. Susannah Tringe

Professor Clarissa J. Nobile, Supervisor

Professor Suzanne S. Sindi, Supervisor

2021

The dissertation of Akshay Deepak Paropkari, titled, “Predicting novel transcription factor-target gene interactions in the *Candida albicans* biofilm network using machine learning”, is approved, and is acceptable ion quality and form for publication on microfilm and electronically:

______________________________ Date ______________

Dr. Susannah Tringe

______________________________ Date ______________

Professor Juris Grasis

Supervisor ______________________________ Date ______________

Professor Clarissa J. Nobile

Supervisor ______________________________ Date ______________

Professor Suzanne S. Sindi

Chair ______________________________ Date ______________

Professor Aaron Hernday

University of California, Merced

2021

Dedications

I dedicate my dissertation to my lovely wife, Shubhra, and my family.

Table of Contents

I. List of Abbreviations viii

II. List of Figures ix

III. List of Tables xii

IV. Acknowledgments xiii

V. Curriculum Vita xiv

VI. Abstract xvii

Chapter 1 Predicting novel transcription factor-target gene interactions in the *Candida albicans* biofilm network using machine learning 1

1.1 Graphical Abstract 1

1.2 Abstract 2

1.3 Introduction 2

1.4 Results 4

1.4.1 Background TFBS sequences differ from foreground TFBS sequences 4

1.4.2 Sequence similarity and DNA shape significantly contribute to SVM classification of the TFBSs of the master *C. albicans* biofilm TFs 6

1.4.3 Sequence-based and DNA shape-based features contribute to TFBS classification 7

1.4.4 Signal recovery of the SVM classifier to obtain high confidence novel TFBSs 9

1.4.5 Predicting life cycle stage-specific biofilm transcriptional regulatory networks 11

1.4.6 Gene Ontology (GO) analysis of predicted target genes for each master *C. albicans* biofilm TF 15

1.4.7 Network centrality provides additional insights into potentially influential target genes 25

1.5 Discussion 26

1.6 Future directions 29

1.7 Methods 29

1.7.1 Computational transcription factor binding site identification 29

1.7.2 Generating background (negative) datasets 31

1.8 References 34

Chapter 2 A CUT&RUN method and data analysis workflow for genome-wide analysis of transcription factor-DNA interactions in *Candida albicans* 41

2.1 Abstract 41

2.2 Introduction 41
2.3 Protocol...42
2.3.1 Results 62
2.4 Discussion and Summary 65
2.5 References 68
Chapter 3 A computational workflow for the analysis of 3' Tag-Seq data 70
3.1 Abstract...70
3.2 Introduction 70
3.3 Basic Protocol 1 72
3.3.1 Downloading RNAseq workflow on local machine 73
3.3.2 Loading the Conda virtual environment 74
3.3.3 Creating an input data folder...74
3.3.4 Running the RNAseq pipeline...75
3.3.5 Output files 75
3.3.6 Visualizing overlaps in multiple experimental conditions 76
3.4 Support Protocol 1 77
3.5 Summary 78
3.6 References 78

I. List of Abbreviations

Abb.	Description
ATAC-seq	Assay for transposase-accessible chromatin followed by high-throughput sequencing
ChIP-chip	Chromatin immunoprecipitation followed by microarray
ChIP-seq	Chromatin immunoprecipitation followed by high-throughput sequencing
ConA	Concanavalin A
cPCR	Colony polymerase chain reaction
CUT&RUN	Cleavage Under Targets and Release Using Nuclease
dDNA	Donor DNA
eGFP	Enhanced green fluorescent protein
EP	Electrostatic potential
GFP	Green fluorescent protein
GO	Gene ontology
gRNA	Guide ribonucleic acid
HelT	Helix twist
IgG	Immunoglobulin G antibody
MACS2	Model-based Analysis for ChIP-Seq, version 2
MGW	Minor groove width
polyA	Polyadenylated RNA
ProT	Propeller twist
RBF	Radial basis function
RNA-seq	Ribonucleic acid isolation followed by high-throughput sequencing
ROC	Receiver operating characteristic
SVM	Support vector machine
TF	Transcription factor
TFBS	Transcription factor binding site

II. List of Figures

Figure 1.1: Graphical abstract of computational method. The top box shows generation of training and testing data. The middle box illustrates support vector machine (SVM) classification process. The bottom box represents genome-wide output of the model. ... 1

Figure 1.2: The potential for machine learning to predict novel TFBSs for the master *C. albicans* biofilm TFs. Principal component analysis of foreground and background training datasets for the six *C. albicans* master biofilm TFs. Each data point on the plots represents one sample. Black colored circles are true foreground sequences obtained from Nobile et al. (2012), while blue cross marks represent simulated background data. 5

Figure 1.3: Top ten features of TFBSs for each master *C. albicans* biofilm TF. For each of the six master *C. albicans* biofilm TFs, the contributions of the top ten influential features in descending order on the Y-axis and their mean contribution towards model decision making on the X-axis are shown. The contribution score was calculated by permuting one feature at a time and measuring the reduction in classification accuracy scores compared to the baseline (non-permuted input data) accuracy scores. 6

Figure 1.4: High precision and recall rates were observed for the SVM model for all six master *C. albicans* biofilm TFs during model training. Precision-recall curves for the SVM model for all six master *C. albicans* biofilm TFs are shown. The plot title depicts the model precision and accuracy scores for the indicated TF. 8

Figure 1.5: An “ROC-like” curve comparing known target genes to novel target genes. The number of known target genes (true positives) is plotted on the Y-axis, while the number of novel target gene predictions is plotted on the X-axis. The black filled circle in each plot indicates the threshold for identifying high confidence target genes. At this threshold, the ratio of known target genes to novel target genes is displayed in the title of each plot...10

Figure 1.6: Comprehensive transcriptional regulatory network controlling *C. albicans* biofilm development. Transcriptional regulatory network controlling biofilm formation in *C. albicans* encompassing all stages of the biofilm life cycle. The six master biofilm TFs are represented by magenta circles (nodes). Smaller blue (upregulated in biofilms) and yellow (downregulated in biofilms) nodes are target genes. The grey lines (edges) represent high confidence model predicted TF-target gene interactions at all stages of the biofilm life cycle. 13

Figure 1.7: Transcriptional regulatory networks controlling *C. albicans* biofilm development at each stage of the biofilm life cycle. Transcriptional regulatory networks controlling biofilm formation in *C. albicans* at each specific stage of the four stages (adherence, initiation, maturation, dispersal) of the biofilm life cycle. The six master biofilm TFs are represented by magenta circles (nodes). Smaller blue (upregulated in biofilms) and yellow (downregulated in biofilms) nodes are target genes. The grey lines (edges) represent model predicted high confidence TF-target gene interactions during each stage of biofilm development. Magenta nodes are sized by their degree counts in each of the four networks. 14

Figure 1.8: Life cycle stage-specific enriched cellular components during *C. albicans* biofilm development. GO terms are depicted on the Y-axis and colored horizontal grouped bars indicate enrichment of predicted target genes by one or more of the six master biofilm TFs for each respective GO term. The presence or absence of a colored bar indicates the TF’s influence over a specific GO term. 17

Figure 1.9: Life cycle stage-specific enriched molecular functions during *C. albicans* biofilm development. GO terms are depicted on the Y-axis and colored horizontal grouped bars indicate enrichment of predicted target genes by one or more of the six master biofilm TFs for each respective GO term. The presence or absence of a colored bar indicates the TF's influence over a specific GO term. 18

Figure 1.10: Life cycle stage-specific enriched biological processes during *C. albicans* biofilm development. GO terms are depicted on the Y-axis and colored horizontal grouped bars indicate enrichment of predicted target genes by one or more of the six master biofilm TFs for each respective GO term. The presence or absence of a colored bar indicates the TF's influence over a specific GO term. 19

Figure 1.11: Methods overview. The top panel (A) denotes the training process for the support machine learning classifier. The bottom panel (B) provides the process to classify genome-wide TF binding sites identified through BLASTn. The red outlined boxes denote novel implementations to generate negative data to create an informative classifier. Whereas in the bottom B panel, BLASTn was utilized to identify all regions matching the consensus TF motif for all master regulators. 33

Figure 2.1: Schematic of the CUT&RUN protocol. First, *C. albicans* cells are permeabilized to isolate intact nuclei. ConA beads are activated, and the intact nuclei are then bound to the activated ConA beads. Antibody of interest is added to the bead-bound nuclei and incubated at 4 °C. Next, pAG-MNase is added and allowed to bind to the target antibody. Only after the addition of $CaCl_2$, pAG-MNase is activated and targeted chromatin digestion proceeds until the addition of the chelating regent to inactivate pAG-MNase. The pAG-MNase bound antibody complex is allowed to diffuse out of the permeabilized nuclei and the resulting DNA extracted and cleaned-up. Sequencing ready libraries are prepared from the CUT&RUN enriched DNA fragments. The resulting libraries are then run on a 10% PAGE gel to separate and remove contaminating adaptor-dimers from the CUT&RUN enriched fragments which are now ligated with sequencing adaptor. 63

Figure 2.2: Schematic of CUT&RUN sequence data analysis. The workflow starts by performing quality check on raw FASTQ files using FASTQC, followed by trimming to remove sequencing adapters. The trimmed reads are then aligned to reference genome and the aligned reads are filtered based on their size to enrich for transcription factor-sized binding signals (20 bp ≤aligned read ≤120 bp). Size selected reads are then calibrated against spike-in *Escherichia coli* reads, and lastly, calibrated reads are used to call peaks using MACS2. 64

Figure 2.3: Quality control steps critical for successful CUT&RUN experiments. (A) Cells were stained with calcofluor white and syto 13 green fluorescent nucleic acid stain before and after nuclei isolation and visualized using EVOS fluorescent microscope. (B) CUT&RUN libraries are analyzed using Bioanalyzer. Successful CUT&RUN TF libraries (indicated by the green checkmark) are enriched for short-fragments smaller than 200 bp and is the best indicator of success. (C) 48 CUT&RUN libraries pooled together and analyzed using Bioanalyzer. High quality pooled CUT&RUN libraries (indicated by the green checkmark) are free of adaptor-dimers (lane 1) while low-quality pooled libraries (indicated to the red 'X') retain small amounts of adaptor-dimers (lane 2). (D) Representative IGV tracks from CUT&RUN datasets showing significant enrichment for Efg1 and Ndt80 binding at the intergenic regions flanking Tec1. 65

Figure 2.4: Evaluation of Ndt80 and Efg1 enriched peaks identified using our CUT&RUN protocol and data analysis pipeline on *C. albicans* biofilm cells. The Venn diagrams in the top row illustrate the degree of overlap between Ndt80 and Efg1 binding sites identified via CUT&RUN with previously published ChIP-chip data. The bottom row highlights CUT&RUN signals for all binding events for Ndt80 and Efg1 as colored heatmaps (red = high peak signal, blue = low/no peak signal) and the signal intensity as a profile plot above the heatmaps. 1000 bp region upstream (-1 kb) and downstream (+1 kb) are also shown in the heatmap..66

Figure 3.1: RNA-seq analysis workflow. (i) Adapters added to raw RNA-seq reads are trimmed using BBDuk tool. (ii) A quality control (QC) report is generated for trimmed reads using FastQC. (iii) Reads passing the QC check are aligned to the reference genome using STAR and a gene count table is created. (iv) The gene count table is used to run differential gene expression analysis using DESeq2, and the DESeq2 output is saved as a table to a file for downstream usage. 72

Figure 3.2: Example UpSet plot output. Three transcription factors (TFs) – Nrg1, Rfg1 and Zcf8 – knockout (KO) RNAseq experimental data is shown. The bar plot on the top represents the overlap of upregulated genes in each of the three experimental samples as well as control WT sample. The horizontal bar on the left highlights the number of significantly upregulated genes in each of the four experimental conditions. The black circles and lines in the bottom show the categories for which overlap is calculated and shown by the bar chart on the top. E.g., the fourth bar on the top represents 199 significantly upregulated genes observed in both WT and Nrg1-KO conditions. The blue bar is highlighting overlaps for the three experimental conditions. 77

III. List of Tables

Table 1.1: Predicted target gene changes observed in each biofilm developmental stage for the six master biofilm TFs. Frequency of new target genes gained, and previous target genes lost, was measured by comparing target genes of each biofilm stage downstream of adherence (stage 1) to the preceding and succeeding developmental stages (for novel targets). 12

Table 1.2: Shared GO terms significantly enriched by the predicted target genes of all six of the master biofilm TFs during each stage of the *C. albicans* biofilm life cycle. NA indicates no shared significantly enriched GO terms. 16

Table 1.3: Cellular functions enriched by the predicted target genes of all six of the master biofilm TFs at each stage of the biofilm life cycle. 20

Table 1.4: Molecular functions enriched by the predicted target genes of all six of the master biofilm TFs at each stage of the biofilm life cycle. 22

Table 1.5: Biological processes enriched by the predicted target genes of all six of the master biofilm TFs at each stage of the biofilm life cycle. 23

Table 1.6: Significant target genes based on information flow centrality metric. Z-test, FDR corrected p-values < 0.05. 25

Table 3.1: Sample metadata file. Users can use this file as a template to generate their metadata file. 74

IV. Acknowledgments

I am grateful to have been supported by many people during my time as a PhD student. I am fortunate to be part of the University of California, Merced family who were with me during my ups and downs of graduate school.

First and foremost, I would like to thank my wife, Shubhra. You celebrated my joyful moments, you motivated me when I was doubtful, and imbued energy when I felt dreary. I could not have completed this journey without you. I am because you are. I love you.

I am very grateful my PhD advisors, Dr. Clarissa Nobile and Dr. Suzanne Sindi. They have been very supportive of my unique graduate journey and have always encouraged me to follow my passion. They encouraged my professional growth from every direction, even when I was interning for three summers away from lab work. They created an open and safe environment for me to express my ideas and patiently guided me to pursue my goals.

I would like to thank my committee members Dr. Aaron Hernday, Dr. Juris Grasis and Dr. Susannah Tringe. Everyone on my committee generously submitted their time and inputs to help guide me in refining my research interests. Their questions inspired to critically think about my research topics.

I am grateful to my lab members Jonathan Anzules, Ashley Valle Arevalo, Priyanka Bapat, Jordan Collignon, Craig Ennis, Ali Heydari, Deepika Gunasekaran, Mikahl Banwarth-Kuhn, Mohammad Qasim, Alex Quijano, Fabián Santiago, Thaddeus Seher, Michael Stobb, and Tyrome Sweet for our friendship. Your support and encouragement have been invaluable during my years in graduate school. I also thank Sarvani Chadalapaka and Matthias Bussonnier for supporting me in expanding my horizons with regards to computing resources and infrastructure.

I thank the National Institute of General Medical Sciences (NIGMS) award R35GM124594 as well as the National Science Foundation Grant ACI-1429783 for providing funding for my research endeavors. Furthermore, my work was partially sponsored by a Pew Biomedical Scholar Award from the Pew Charitable Trusts, and by the Kamangar family in the form of an endowed chair to Dr. Clarissa Nobile.

I am thankful to all my mentors who were my biggest advocates during my time in each of my summer internships. Dr. Jessica Jarett (DOE Joint Genome Institute), Dr. Thomas Sandmann (Denali Therapeutics) and Dr. Javier Carerra (Zymergen) put faith in me by giving me the opportunity to learn and grow as a scientist. Your faith in my skills taught me to believe in myself.

I thank Abdirahim Abdi, Dr. Purnima Kumar, Dr. Sukirth Ganesan, Dr. Shareef Dabdoub and Niharika Gupta for igniting the love of science and research. Your support, encouragement and friendship continue to motivate me in my scientific journey.

Lastly, I thank my family for their unconditional encouragement, support, and love.

V. Curriculum Vita

EDUCATION

PhD in Quantitative and Systems Biology ***Aug 2017 – Dec. 2021***
University of California, Merced, CA

Master of Science in Quantitative and Systems Biology ***Aug 2017 – May 2020***
University of California, Merced, CA

Master of Science in Electrical and Computer Engineering ***Aug 2012 - May 2014***
The Ohio State University, Columbus, OH

Bachelor of Science in Electrical and Computer Engineering ***Sept 2008 – Jun 2012***
The Ohio State University, Columbus, OH

TECHNICAL SKILLS

- Programming languages: Python, R, Bash
- Tools: Jupyter, Pandas, NumPy, SciPy, scikit-learn, matplotlib, tidyverse, ggplot, Inkscape, Gephi
- GitHub: akshayparopkari

WORK EXPERIENCE

Data Science Intern**, Zymergen Inc.* ***May 2021 – Aug 2021

- Designed, implemented, and refined scikit-learn based predictive ML workflow for analyzing hybrid datasets – gene expression and fermentation KPI – to improve productivity of engineered microbial strains
- Enhanced ML workflow with the addition of (1) automated hyperparameter optimization to improve model performance - using one-sigma theory and (2) input data augmentation to expose the model to diverse input data - via employing DESeq2, documented on company GitHub repo
- Authored an internal company-wide LIMS Wiki post reporting my findings from my summer project

Cyberinfrastructure Graduate Associate**, UC Merced* ***May 2020 – Aug 2020

- Collaborated with UC Merced Research HPC team in addressing user requests and IT incidents
- Designed, scripted, and documented high-performance cluster – MERCED – system monitoring tools using Bash and Python
- Coordinated JupyterHub setup on two high-performance clusters to assist in teaching and research needs
- Documented system monitoring tools and submitted summary report to MERCED leadership

Computational Biology Intern**, Denali Therapeutics Inc.* ***Jun 2019 – Aug 2019

- Utilized GWAS data from UK BioBank and academic publications to understand heritability from summary statistics and cross-correlated results with non-neurodegenerative disease phenotypes
- Coded and documented LDSC and enrichment analysis workflow in R, Python and Bash for convenient reuse and reproduction of results

Genomics Grad Intern**, DOE Joint Genome Institute* ***Jun 2018 – Aug 2018

- Designed novel computational method in Python and R to detect all types of transposable elements (plasmids, prophages, virus, etc.) in single-cell genomes and metagenomes
- Enhanced capability to detect transposable elements in environmental samples using CRISPR spacers
- Reported software code, data analysis and workflow using reproducible Jupyter notebook

Staff Bioinformatician**, The Ohio State University* ***Sept 2014 – May 2017

- Designed a novel workflow for processing 16S data in Python using flash, fastx toolkit and skewer
- Devised research strategy plan to analyze taxonomic and genetic biological networks from 16S and metagenomics for NIH grant
- Enhanced PhyloToAST by adding error handling capabilities, introducing feature selection and extraction, hypothesis testing, data visualization, and unit–testing in R for QIIME/PhyloToAST pipeline analysis
- Created Bash scripts for pre-processing of microbial metagenomics data using MG-RAST web tool

RESEARCH EXPERIENCE

Graduate Research**, University of California, Merced* ***Aug 2017 – Dec 2021

- Project 1 – Modeling gene regulatory networks associated with biofilm formation
  - Integrate transcription factor binding site (TFBS) data and temporal gene expression data to predict novel TFBS during *Candida albicans* biofilm development
  - Characterize true vs. false positives using a SVM classifier using scikit-learn Python library
  - Utilize predicted TFBS to simulate perturbation in gene regulatory network using next-generation machine learning models
- Project 2 – Implemented CUT&RUN analysis pipeline
  - Designed and implemented computational workflow to analyze CUT&RUN data to validate predictions from Project 1 to expand our understanding of biofilm specific TF-target gene interactions

- Project 3 – Built custom RNA-seq processing pipeline and deployed on UC Merced's high performance computing resource - MERCED
    - Engineered an automated pipeline, for quality control, genome alignment and differential expression analysis for 3' Tag-seq datasets using fastx toolkit, STAR and DESeq2

***Graduate Statistics Tutoring Assistant**, UC Merced* **_Aug 2019 – May 2020_**

- Taught statistical concepts ranging from power analysis, hypothesis testing, experimental design, etc.
- Presented workshops to student groups on utilization of high-performance computing resources
- Assessed Python, R and Stata code for resolving logical and syntactical issues
- Resolved student queries over emails and through online Zoom meeting and maintained a log of inquiry topics and number of students assisted

PUBLICATIONS

- **Akshay D. Paropkari**, Suzanne S. Sindi and Clarissa J. Nobile "*Identifying novel transcription factor-target gene interactions in the Candida albicans biofilm network using machine learning*" (pending)
- Lopez-Oliva, Isabel & **Paropkari, Akshay** et al. "Dysbiotic subgingival microbial communities in periodontally healthy patients with rheumatoid arthritis." Arthritis & Rheumatology 70.7 (2018): 1008-1013.
- **Paropkari** et al. "Smoking, pregnancy and the subgingival microbiome." Scientific reports 6.1 (2016): 1-9.
- Dabdoub, Shareef… **Paropkari, Akshay**… et al. "PhyloToAST: Bioinformatics tools for species-level analysis and visualization of complex microbial datasets." Scientific reports 6.1 (2016): 1-9.

CONFERENCES PRESENTATIONS

- Society for Mathematical Biology (SMB) Virtual Conference ***Jun 2021***
- CSHL Biological Data Science Virtual Conference ***Nov 2020***
- SciPy 2020 Virtual Conference ***Jul 2020***
- ASM Microbe 2019, San Francisco, CA ***Jun 2019***
- Fifth Recent Advances in Microbial Control (RAMC) in Clearwater, FL ***Nov 2018***

HONORS AND AWARDS

- Fred and Mitzie Ruiz Fellowship ***Sept 2020***
- 2020-21 UC President's Lindau Nobel Laureate Meetings Fellow ***Jun 2020***
- CSHL Conference Helmsley Charitable Trust Fellowship ***Mar 2019***
- Santa Fe Institute James S. McDonnell Foundation Travel Grant ***Oct 2018***
- Interdisciplinary Computational Graduate Education Fellowship ***Jan 2018***

VI. Abstract

Predicting novel transcription factor-target gene interactions in the *Candida albicans* biofilm network using machine learning

Doctor of Philosophy

in

Quantitative and Systems Biology

by

Akshay Deepak Paropkari

University of California, Merced

2021

Chair of Advisory Committee: Professor Aaron Hernday

Transcription is a complex process underlying many cellular functions. DNA structure was discovered by Rosalind Franklin over 50 years ago. Since then, we have been steadily dissecting our understanding of the biological logic governing all life on Earth. Ultimately, Francis Crick discovered the central dogma of molecular biology in 1970. Early studies on *Escherichia coli* began formulating the idea of gene regulation as the basis of information flow in biological systems. Relatively soon, due to advancements in computing power, computer aided analyses of DNA and RNA were introduced to understand regulatory sequences. Since then, many improvements in experimental and computation protocols have accelerated our understanding of the control of biological information flow.

My thesis work, which uses the fungal species *Candida albicans* as a model, is the latest advancement in our understanding of gene regulation. In chapter one, I present my work on identifying the gene regulatory networks controlling biofilm formation in *C. albicans* across the biofilm life cycle. I explain my novel workflow that utilizes sequence-based as well as DNA-shape based features to predict transcription factor binding sites (TFBSs) genome-wide in *C. albicans*. In chapter two, I present CUT&RUN sequencing data implemented to assess binding events for specific TFs in *C. albicans*. I also present my novel CUT&RUN computational pipeline to analyze CUT&RUN sequencing data. In chapter three, I present a computational workflow to analyze 3’ Tag-Seq data. In this 3’ Tag-Seq, the 3’ end of the transcript is selected and amplified to yield one copy of cDNA from each transcript in a biological sample.

Chapter 1

Predicting novel transcription factor-target gene interactions in the *Candida albicans* biofilm network using machine learning

1.1 Graphical Abstract

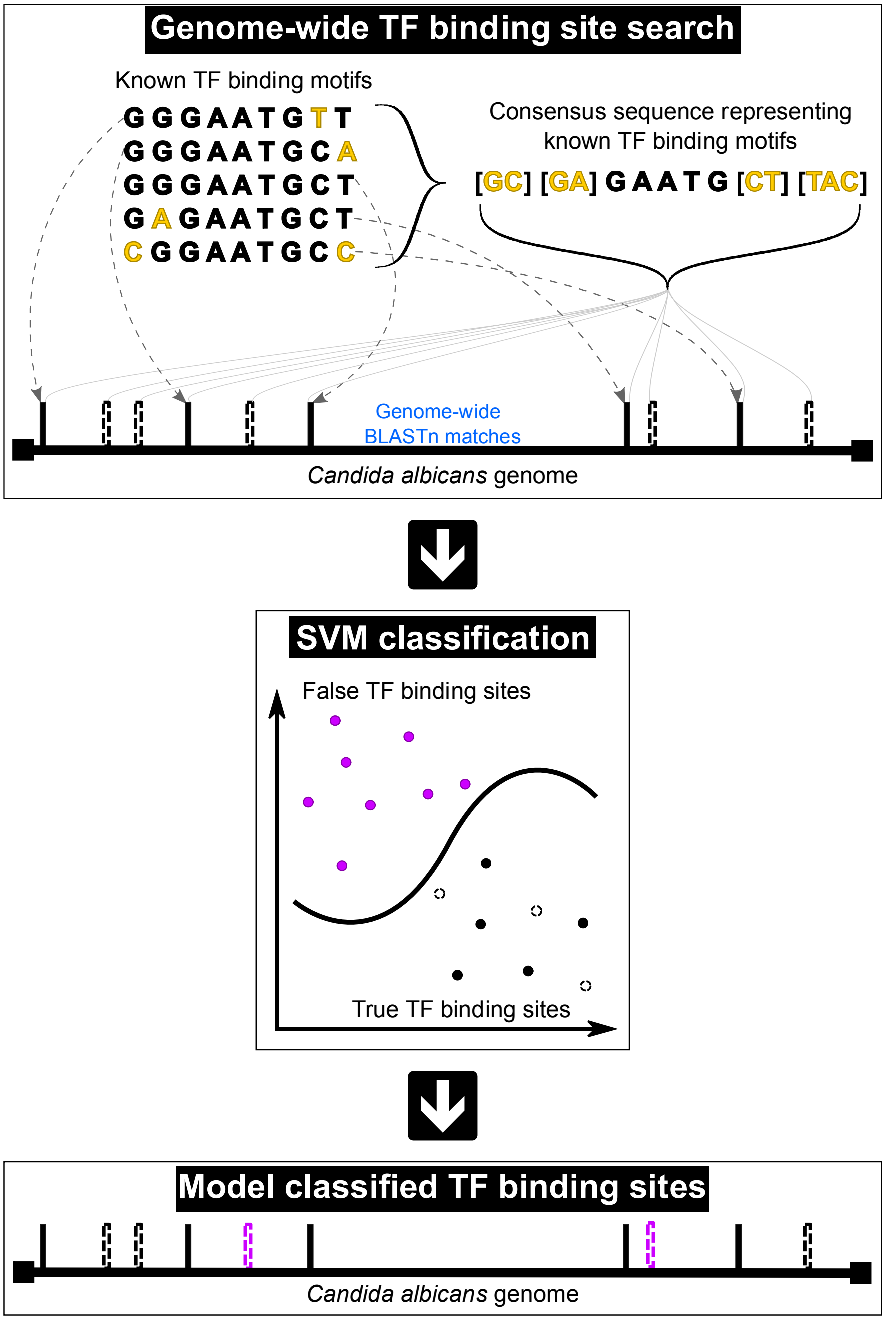

Figure 1.1: Graphical abstract of computational method. The top box shows generation of training and testing data. The middle box illustrates support vector machine (SVM) classification process. The bottom box represents genome-wide output of the model.

1.2 Abstract

Biofilms are surface-adhered communities of microbial cells that can serve as reservoirs of infection. *Candida albicans* is a common human fungal pathogen, capable of forming biofilms on biotic and abiotic surfaces. Transcription factors (TFs), defined as sequence specific DNA binding proteins, are important players in regulating transcription during complex developmental processes, such as biofilm formation. The transcriptional network controlling biofilm formation in *C. albicans*, consisting of six “master” TFs, Bcr1, Brg1, Efg1, Ndt80, Rob1 and Tec1, and 1,061 downstream “target” genes, has been previously elucidated for mature *C. albicans* biofilms. However, the roles of these TFs in controlling target gene expression at different stages of biofilm development have yet to be determined.

In this study, we use a supervised support vector machine (SVM) classifier and a validated set of TF binding sites (TFBSs), to predict novel TF-target gene interactions for each biofilm master TF temporally over the course of *C. albicans* biofilm formation. First, target sequences were created using previously identified transcription factor binding site (TFBS) consensus sequences that represent potential binding sites. The number of TFBS consensus sequences for each TF depended on both the number of validated sites as well as the fidelity of the motifs. Second, a feature matrix was built to capture the DNA shape and sequence qualities of each *Candida*te TFBS motif. Next, a positive/true set of potential TFBSs was predicted for each TF using a trained SVM classifier based on the feature matrix. The sequence similarity score was the top contributing feature to classify novel TFBSs. Finally, active TF-target gene interactions were identified by correlating TF binding activity with previously reported time-series gene expression data of target genes. Interestingly, Ndt80 and Efg1 are predicted to control the greatest number of target genes at any given stage of biofilm development. Overall, by coupling TFBS sequence and DNA shape information, we predict novel TFBSs, TF-target gene interactions, and ultimately, transcriptional regulatory networks controlling each stage of the *C. albicans* biofilm life cycle.

1.3 Introduction

Candida albicans is a diploid polymorphic commensal fungus commonly found in the oral cavity, gastrointestinal tract, genitourinary tract, and skin of healthy humans (Lohse et al., 2018; Nobile et al., 2012; Nobile and Johnson, 2015). *C. albicans* is also an opportunistic pathogen that predominantly causes severe disease in immunocompromised individuals, such as those with HIV/AIDS and patients undergoing chemotherapy or organ transplantation (Lohse et al., 2018; Nobile et al., 2012; Nobile and Johnson, 2015) but can also cause superficial mucosal and cutaneous infections in healthy individuals, such as vulvovaginal candidiasis in women and diaper rash in babies (Mayer et al., 2013; Soll and Daniels, 2016). An important virulence trait of *C. albicans* is its ability to form biofilms – complex communities of cells that form on biotic surfaces (e.g., mucosal epithelial layers) and abiotic surfaces (e.g., catheters, heart valves, and prosthetic devices) (Gulati and Nobile, 2016; Lohse et al., 2018; Mayer et al., 2013; Nobile and Johnson, 2015). *C. albicans* biofilms are often resistant and/or tolerant to antifungal drugs, making biofilm infections notoriously difficult to treat (Gulati and Nobile, 2016; Lohse et al., 2018; Nobile and Johnson, 2015). The *C. albicans* biofilm life cycle occurs in four stages: adherence, initiation, maturation, and dispersal (Lohse et al., 2018; Nobile and Johnson, 2015; Uppuluri et al., 2010). In the adherence stage, planktonic yeast-form cells attach to a surface. In the initiation stage, the yeast-form cells proliferate to form an anchoring cell layer and begin to differentiate into hyphal and pseudohyphal cells. In the maturation stage, the hyphal cells elongate, and a protective extracellular matrix composed of proteins, carbohydrates, nucleic acids, and lipids, encases the biofilm cells. Finally, in the dispersal stage, yeast-form cells are released from the biofilm, where they repeat the biofilm life cycle by adhering to and forming biofilms at new surfaces or grow planktonically. Understanding the genetic and molecular mechanisms regulating the *C. albicans* biofilm life cycle is fundamental to the development of effective therapeutics for biofilm infections caused by this important human fungal pathogen.

Transcription factors (TFs) or transcriptional regulators, defined here as sequence specific DNA-binding proteins, are proteins that regulate the expression of downstream "target" genes to ultimately control important cellular functions. TFs bind to specific cis-regulatory DNA sequences in the genome upstream of the transcription start sites of their target genes to mediate gene expression. The transcriptional network controlling biofilm formation in *C. albicans* was initially described in 2012 for mature (48-hour) in vitro biofilms (Nobile et al., 2012). In this study, using genome-wide binding approaches combined with genome-wide transcriptional profiling approaches, six "master" biofilm TFs (Bcr1, Tec1, Efg1, Ndt80, Rob1, and Brg1) were identified along with 1,061 downstream "target" genes. Generally, the six master biofilm TFs were found to bind to their own upstream intergenic regions and to positively regulate the expression of one other, forming a closely knit biofilm circuit comprised of many feed-forward loops (Nobile et al., 2012). In 2015, this biofilm circuit was expanded to also include Gal4, Rfx2 and Flo8 as additional TFs in the core circuit (Fox et al., 2015). Since experimentally determined cis-regulatory sequences, which we refer to as TF binding sites (TFBSs), were not established for Gal4, Rfx2 and Flo8 under biofilm conditions, we did not consider these three additional TFs in our supervised learning model in the present study.

Three genome-wide transcriptional profiling studies compared *C. albicans* biofilms to planktonic cells temporally over the course of biofilm development in vitro (Fox et al., 2015; García-Sánchez et al., 2004; Yeater et al., 2007) and one genome-wide transcriptional profiling study compared *C. albicans* biofilms to planktonic cells over the course of biofilm development in vivo in a rat central venous catheter biofilm model (Nett et al., 2009). Together, all four of these transcriptional profiling studies identified many genes differentially expressed over time during *C. albicans* biofilm development, highlighting the coordinated changes in gene expression that occur in biofilm formation. Based on the fact that the temporal transcriptional profiling datasets from the Fox et al. (2015) study (Fox et al., 2015) used the same conditions as the binding datasets for mature biofilms in the Nobile et al. (2012) study (Nobile et al., 2012), we used the datasets from these comprehensive genome-wide studies to predict the regulatory relationships between the six master biofilm TFs and their downstream target genes throughout the four major stages (adherence, initiation, maturation, and dispersal) of the *C. albicans* biofilm life cycle.

Many computational approaches have been proposed to date to predict TFBSs based on genome-wide binding experimental datasets, such as those from chromatin immunoprecipitation followed by sequencing (ChIP-seq) and assay for transposase-accessible chromatin followed by sequencing (ATAC-seq) experiments (Berman et al., 2002; Rogers and Bulyk, 2018; Sinha et al., 2003). Early methods were based on scanning genomes to calculate sequence similarities to known binding site sequences to identify novel binding specificities (Berman et al., 2002). More recent experimental and modeling studies have shown that information such as GC content of regions flanking binding sites, multiple binding specificities, and 3D DNA structure or DNA shape upstream of binding sites, enhance the prediction of transcription factor binding site (TFBS) specificities (Cuellar-Partida et al., 2012). Consequently, certain modern methods now incorporate such information to more accurately predict TFBSs (Cuellar-Partida et al., 2012; Inukai et al., 2017; Stormo, 2013). In addition, with the recent advancements in machine learning, scientists are beginning to exploit computational power to study larger and more complex TFBS datasets (Chiu et al., 2016; Kantorovitz et al., 2009; Mathelier et al., 2016; Mathelier and Wasserman, 2013; Zhou et al., 2015). For example, Mathelier et al. (2016) combined DNA sequence information with DNA shape information as additional metrics for predicting TFBSs (Mathelier et al., 2016).

In this study, we use previously characterized sets of TFBS interactions obtained for mature *C. albicans* biofilms to predict new interactions over the course of biofilm development using a supervised learning model. More specifically, we develop a support vector machine (SVM) that considers the full space of potential binding sites, based on known binding motifs, to computationally predict novel TFBSs, TF-target gene interactions, and ultimately, entire gene regulatory networks, temporally over the course of the four known stages of the *C. albicans* biofilm life cycle. These novel predictions will set the framework for us – and the larger community – to further explore the roles of transcriptional regulation in biofilm dynamics.

1.4 Results

1.4.1 Background TFBS sequences differ from foreground TFBS sequences

A critical challenge in using experimentally derived genome-wide binding data, such as ChIP-seq data, to develop computational models is that there is only a selection of known or positive examples of binding. To address this challenge, we developed a framework for generating theoretically negative binding data that we utilize as training data (see Methods). The input data to our machine learning framework is an equally balanced set of positive/true (foreground) and presumptive negative/false (background) binding sites characterized by two broad feature categories: (a) sequence similarity (MinHash, Poisson additive and product similarity score) and (b) DNA features (TFBS GC proportion and DNA shape values, i.e., DNA roll, helix twist (HelT), minor groove width (MGW), propeller twist (ProT) and electrostatic potential).

To evaluate if the simulated background sequences represent a negative dataset, we conducted principal component analysis (PCA) on foreground and background sequences to assess the performance of machine learning on predicting novel TFBSs, as was implemented in Mathelier et al. (2016) (Mathelier et al., 2016). Ideally, the background dataset should cluster separately from the foreground sequences. For all six master *C. albicans* biofilm TFs, the true/foreground dataset was separated from the simulated background dataset (Figure 1). Some overlap between foreground and background data points was observed since the negative data originated from foreground sequences. The balance between intersecting and exclusive foreground and background data points that we observed in the PCA plots allowed for a better trained model, avoiding the creation of overly simplistic background data points as well as background data that was too similar to foreground data. The frequency of foreground sequence counts influenced the negative/background binding site sequence counts. From the Nobile et. al. (2012) dataset, we started with 94, 137, 482, 612, 43 and 94 unique foreground sequences for Bcr1, Brg1, Efg1, Ndt80, Rob1 and Tec1, respectively (Nobile et al., 2012). Using a dual strategy for background sequence generation, we obtained 188, 274, 964, 1224, 86 and 188 distinct background sequences for Bcr1, Brg1, Efg1, Ndt80, Rob1 and Tec1, respectively. The background/negative TFBS sequences were generated using dual strategies (see Methods); thus, we have twice as many background/negative TFBS sequences as foreground/true TFBS sequences. These differences in data points are reflected in the density of the point clouds in the PCA scatter plots in **Figure 1.2**.

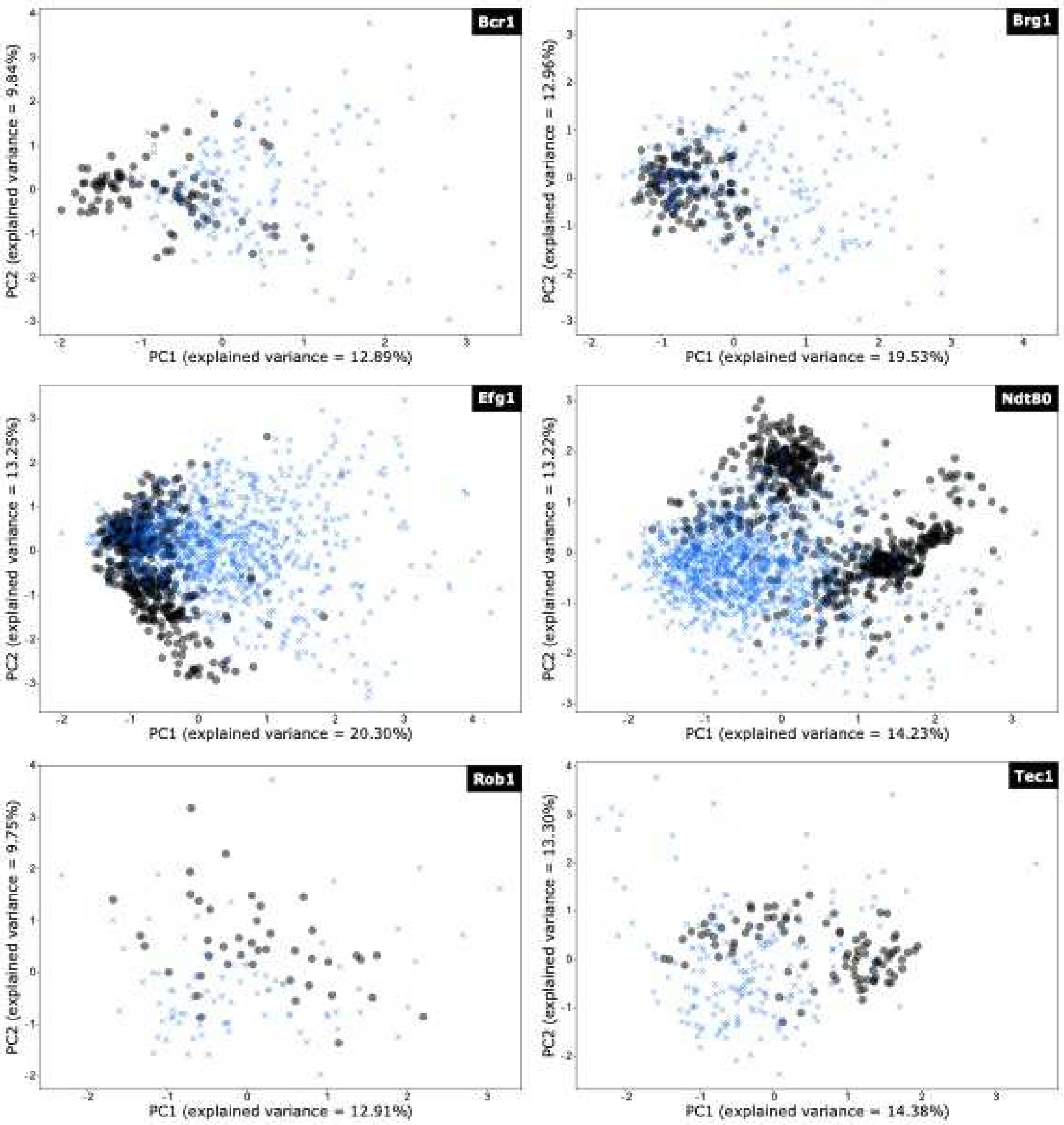

Figure 1.2: The potential for machine learning to predict novel TFBSs for the master *C. albicans* biofilm TFs. Principal component analysis of foreground and background training datasets for the six *C. albicans* master biofilm TFs. Each data point on the plots represents one sample. Black colored circles are true foreground sequences obtained from Nobile et al. (2012), while blue cross marks represent simulated background data.

1.4.2 Sequence similarity and DNA shape significantly contribute to SVM classification of the TFBSs of the master *C. albicans* biofilm TFs

A support vector machine (SVM) model using a radial basis function (RBF) kernel was created for each *C. albicans* master biofilm TF. A balanced training dataset was derived from foreground and background data. The SVM model was trained on 80% of the data and its performance was tested on the remaining 20% of the data. A tenfold cross validation method was used to evaluate the performance of the model.

To determine the most important features used by SVM to classify TFBS sequences, we measured the influence of each feature by permuting the feature column for all samples in the training data and calculating the decrease in the prediction precision score compared to the original training data. Contribution scores for each feature are shown in **Figure 1.3**. For each feature, the contribution score represents its influence, and therefore importance, in facilitating the SVM model to classify a sample as a TFBS or not.

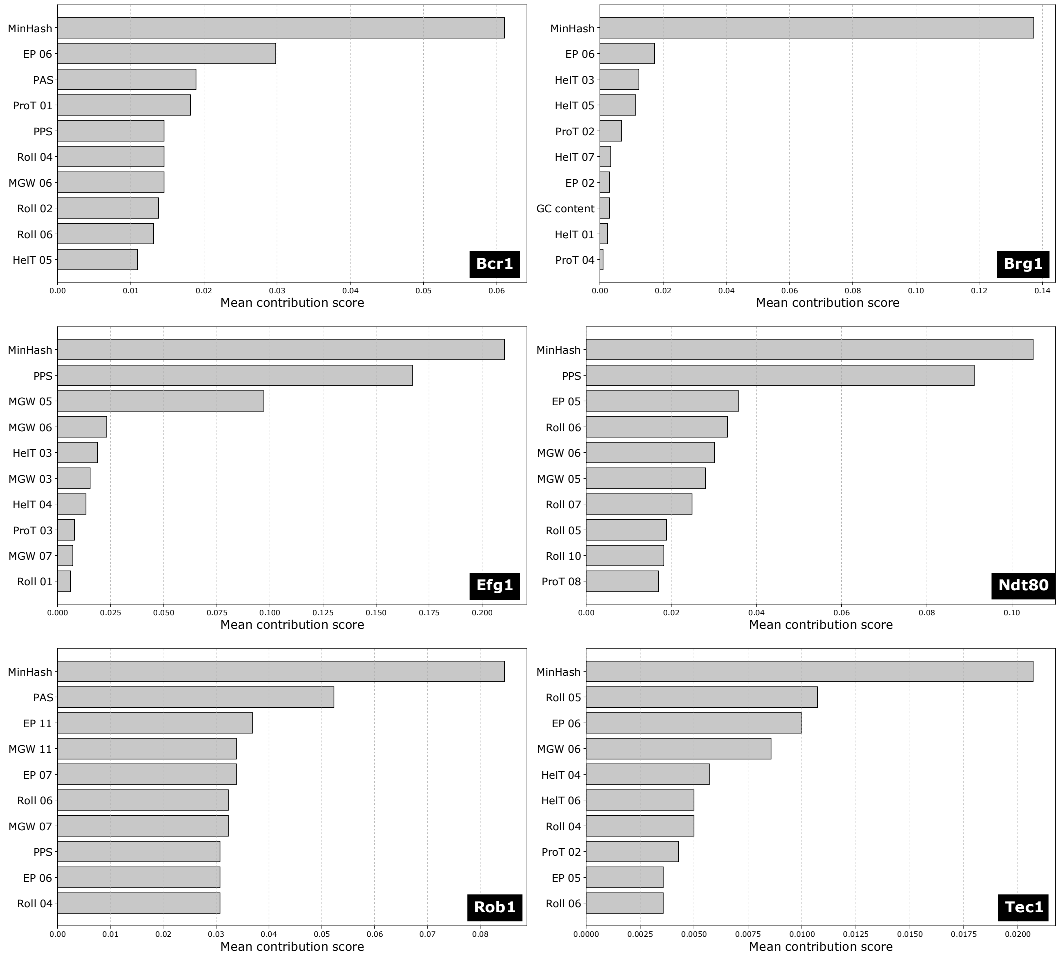

Figure 1.3: Top ten features of TFBSs for each master *C. albicans* biofilm TF. For each of the six master *C. albicans* biofilm TFs, the contributions of the top ten influential features in descending order on the Y-axis and their mean contribution towards model decision making on the X-axis are shown. The contribution score was calculated by permuting one feature at a time and measuring the reduction in classification accuracy scores compared to the baseline (non-permuted input data) accuracy scores.

High scores (along the X-axis) represent high influencing power. The sequence similarity score feature, MinHash, was the most influential feature in all six models. The second most influential feature was the electrostatic potential at position six (EP 06) for four of the master *C. albicans* biofilm TFs (Bcr1, Brg1, Rob1 and Tec1). EP at position five (EP 05) was influential for Ndt80, while the electrostatic potential was not present in the top ten influential features for Efg1. Interestingly, distinct DNA shape values at positions 5 and 6 were the most influential features for all six of the master *C. albicans* biofilm TFs.

1.4.3 Sequence-based and DNA shape-based features contribute to TFBS classification

For the classification problem of categorizing a TFBS as valid (true) or invalid (false), the model hyperparameters controlling its learning and training need to be set. SVM creates a decision boundary that separates the inputs into True or False, depending on which side of the decision boundary they lie. For SVM, the parameter C controls the sensitivity of a decision boundary that separates foreground from background data. This parameter was identified using an exhaustive grid search that runs the model through a range of values for parameter C and identifies its best value based on model precision and recall metrics. Similarly, linear and RBF kernels were supplied to the grid search method to identify the best kernel for this dataset. Overall, six separate grid searches, one for each master *C. albicans* biofilm TF, were utilized to identify the optimal model parameters. In all six models, the RBF kernel performed better than the linear kernel.

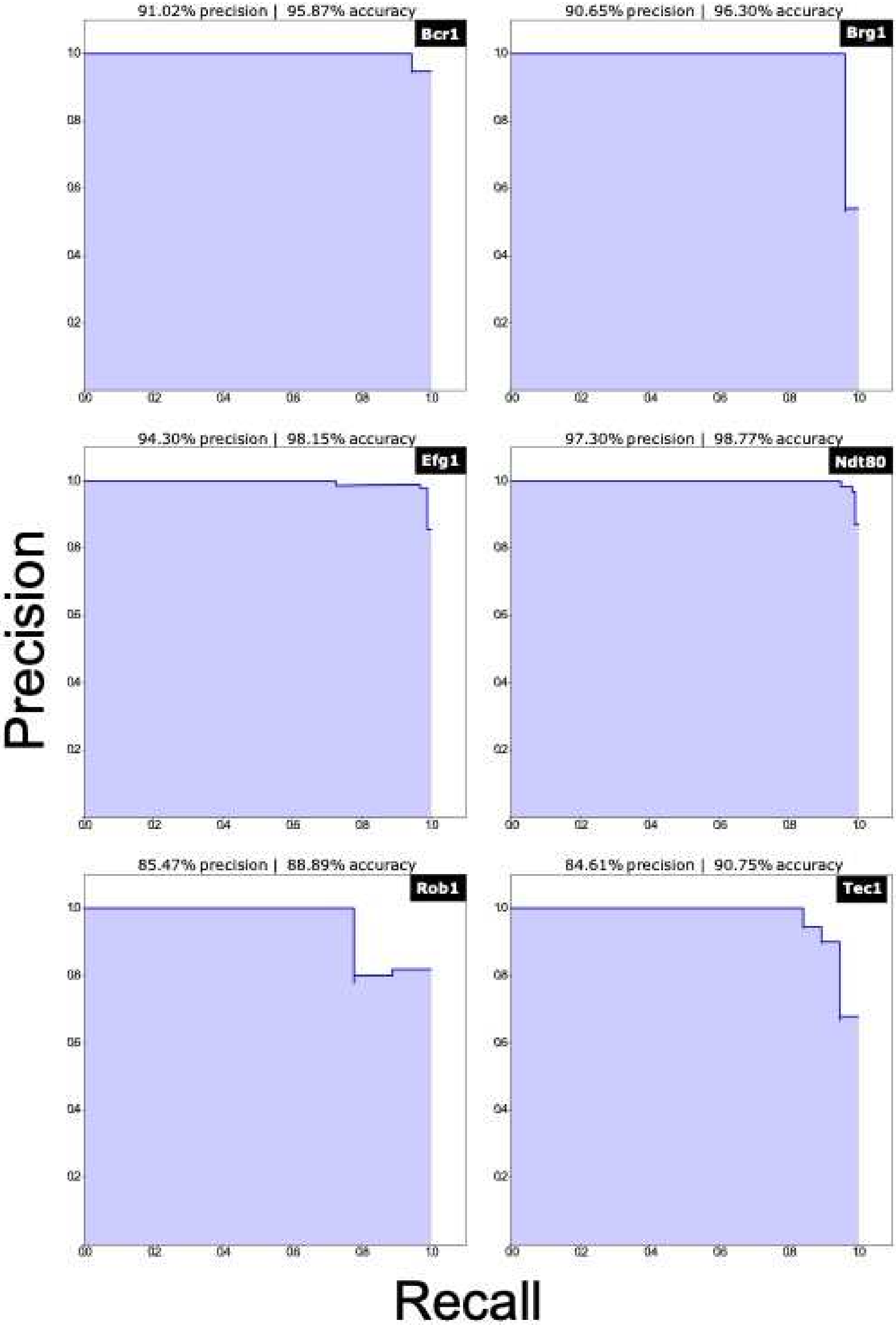

Figure 1.4: High precision and recall rates were observed for the SVM model for all six master *C. albicans* biofilm TFs during model training. Precision-recall curves for the SVM model for all six master *C. albicans* biofilm TFs are shown. The plot title depicts the model precision and accuracy scores for the indicated TF.

To assess the model accuracy, we divided the input background and foreground datasets into model training (80%) and model testing (20%) datasets. **Figure 1.4** displays precision-recall curves for each of the six SVM models – one for each master *C. albicans* biofilm TF. For all TF models, over 84% precision and 88% accuracy were achieved by parameter optimization using the grid search method. Efg1 and Ndt80 had the most data points relative to the other master biofilm TFs from Nobile et al. (2012) and their precision and accuracy scores were the top two among all six master *C. albicans* biofilm TFs, indicating a positive correlation between model performance and the size of the input training data.

1.4.4 Signal recovery of the SVM classifier to obtain high confidence novel TFBSs

We tested the performance of our SVM classifier to recover validated TFBS datasets obtained from Nobile et al. (2012) (Nobile et al., 2012). We used the distance to the decision boundary as the metric to understand how our classifier performs on known and unknown TFBS datasets. For each of the six master *C. albicans* biofilm TFs, we categorized all positive predictions as either high confidence or low confidence depending on their distance from the separating decision boundary. A low confidence TFBS prediction would be found between the separating decision boundary and the closest experimentally validated TFBS to the decision boundary. Similarly, a high confidence TFBS prediction would be found at a greater distance away from the decision boundary when compared to the distance at which the closest validated TFBS would be found. For example, if an experimentally validated TFBS is located x units away from the SVM separating decision boundary, then a novel TFBS prediction at a distance y away from the separating decision boundary is categorized using:

$$TFBS\ prediction = \begin{cases} high\ confidence, & only\ if\ y\ \geq x \\ low\ confidence, & 0 < y < x. \end{cases}$$

Target genes for all high confidence TFBSs were identified based on proximity of the open reading frame (ORF) to the binding location. After identifying the target genes, we compared novel target genes predicted by our model to previously known target genes (true positives) from Nobile et al. (2012) (Nobile et al., 2012) to evaluate the model predictions. Based on the distribution observed in the “receiver operating characteristic (ROC)-like” plots in **Figure 1.5**, the distance of previously reported Efg1 and Ndt80 target genes are more uniformly distributed away from the SVM decision boundary compared to the other four master *C. albicans* biofilm TFs.

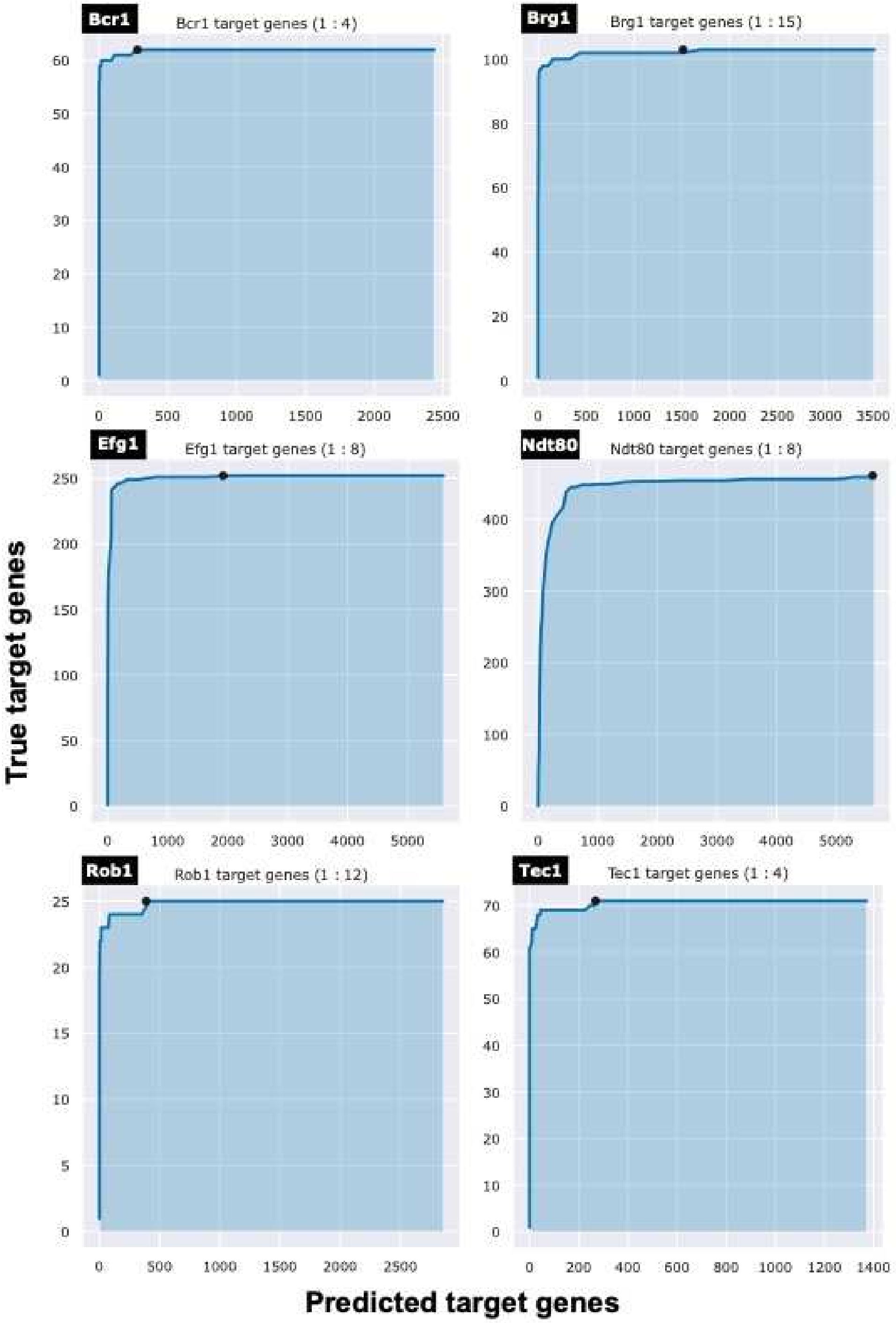

Figure 1.5: An "ROC-like" curve comparing known target genes to novel target genes. The number of known target genes (true positives) is plotted on the Y-axis, while the number of novel target gene predictions is plotted on the X-axis. The black filled circle in each plot indicates the threshold for identifying high confidence target genes. At this threshold, the ratio of known target genes to novel target genes is displayed in the title of each plot.

1.4.5 Predicting life cycle stage-specific biofilm transcriptional regulatory networks

To identify stage-specific TF-target gene interactions, we combined our TFBS model predictions with existing temporal genome-wide biofilm transcriptional profiling data from Fox et al. (2015) (Fox et al., 2015). Specifically, for all positively predicted TFBSs within an intergenic region, we extrapolated all TF-target gene interactions based on proximity of the positively predicted TFBSs to their target genes. Figure 1.6 portrays the comprehensive transcriptional regulatory network controlling biofilm formation in *C. albicans* encompassing all possible (experimentally validated and model predicted) TF-target gene interactions at every stage of the *C. albicans* biofilm life cycle. This highly interconnected network consists of 5,430 nodes (representing the master biofilm TFs and their target genes) and 14,599 edges (model predicted high confidence TFBSs). We next identified all high confidence TF-target gene pairs specifically throughout the four stages of the biofilm life cycle by incorporating the previously published temporal transcriptional profiling datasets from Fox et al. (2015) (Fox et al., 2015). We considered the TF-target gene interaction as “active” if the expression of the target gene was differentially regulated with a twofold or greater expression change (upregulated or downregulated) in biofilms at any of the four biofilm developmental stages compared to planktonic cells (see Dataset S1, Dataset S2, Dataset S3, Dataset S4, Dataset S5, and Dataset S6 for all model predicted target genes indicating activity at each biofilm life cycle stage for Bcr1, Brg1, Efg1, Ndt80, Rob1 and Tec1, respectively). By assigning activity to all high confidence TF-target gene pairs, we have predicted transcriptional regulatory networks for all four stages of the *C. albicans* biofilm life cycle (**Figure 1.7**). Each transcriptional regulatory network for the four biofilm life cycle stages (adherence, initiation, maturation, and dispersal) consisted of distinct numbers of nodes and edges, highlighting the dynamic transcriptional changes occurring during biofilm development. We predicted a total of 1,164 nodes with 4,562 edges during the adherence stage, 400 nodes with 1,585 edges during initiation stage, 594 nodes with 2,349 edges during the maturation stage, and 739 nodes with 2,935 edges during the dispersal stage (Table S1). Of these total nodes and edges, we predicted 1,096 nodes with 2,974 edges to be active during the adherence stage, 380 nodes with 1,055 edges to be active during the initiation stage, 543 nodes with 1,479 edges to be active during the maturation stage, and 682 nodes with 1,873 edges to be active during the dispersal stage (Table S1).

To evaluate our predicted TF-target gene interaction changes during the *C. albicans* biofilm life cycle, we compared the new TF-target gene interactions gained and previous TF-target gene interactions lost by each of the six master biofilm TFs between every preceding and succeeding biofilm developmental stage downstream of the adherence stage (**Table 1.1**). Through this analysis, we observed that each biofilm developmental stage downstream adherence (initiation, maturation, and dispersal) showed evidence of TF-target gene interaction rewiring. The percentages of novel target genes predicted to be gained and previous target genes predicted to be lost by all the master biofilm TFs combined from the initiation through to the dispersal stages of biofilm formation are reported in **Table 1.1**. During the initiation stage, the six master biofilm TFs were predicted to gain 3.79 ± 1.35% of novel target genes and to lose 46.42 ± 20.1% of previous target genes compared to the adherence stage. Similarly, during the maturation stage, the six master biofilm TFs were predicted to gain 70.67 ± 29.19% of novel target genes and to lose 38.83 ± 16.4% of previous target genes compared to the initiation stage. And finally, during the dispersal stage, the six master biofilm TFs were predicted to gain 29.85 ± 12.57% of novel target genes and to lose 13.13 ± 5.42% of previous target genes compared to the maturation stage. Overall, the maturation stage had the highest number of predicted gains of novel target genes, while the initiation stage had the highest number of predicted losses of previous target genes.

Table 1.1: Predicted target gene changes observed in each biofilm developmental stage for the six master biofilm TFs. Frequency of new target genes gained, and previous target genes lost, was measured by comparing target genes of each biofilm stage downstream of adherence (stage 1) to the preceding and succeeding developmental stages (for novel targets).

	Number of novel target genes gained during initiation (stage 2)	**Number of novel target genes gained during maturation (stage 3)**	**Number of novel target gene gained during dispersal (stage 4)**	**Number of previous target genes lost during initiation (stage 2)**	**Number of previous target genes lost during maturation (stage 3)**	**Number of previous target genes lost during dispersal (stage 4)**
Bcr1	37	224	151	420	130	59
Brg1	45	266	177	531	160	73
Efg1	60	411	257	798	226	112
Ndt80	60	424	263	819	231	117
Rob1	45	253	152	464	126	75
Tec1	18	118	64	210	59	32

To prioritize our model predicted target genes into those that could be of highest functional (biological) relevance across the different stages of the biofilm life cycle, we used the available temporal gene expression data from Fox et al. (2015) (Fox et al., 2015), to identify a discreet set of three target genes that were bound by five of the master biofilm TFs and that had significant expression changes across the four biofilm stages. Significance was calculated for all target genes of a TF separately to obtain active TF-target gene interactions. For all target genes of a master regulator, we measured the Z-test derived p-value adjusted to a 5% false discovery rate, which compared the mean log2 fold change expression values of each target gene across all four biofilm growth stages to mean log2 fold change expression values of all target genes for all four biofilm growth stages. We note that there were no target genes that fit these criteria for all six of the master biofilm TFs (see Dataset S7 for the twenty target genes that were bound by all six of the master biofilm TFs and their corresponding expression changes across the four biofilm life cycle stages). Three model predicted target genes stood out from our temporal analysis: ORF19.2870/LDG11 (bound by Bcr1, Brg1, Efg1, Ndt80, and Rob1; Z-test p-value < 0.05), ORF19.2762/AHP1 (bound by Bcr1, Brg1, Efg1, Ndt80, and Rob1; Z-test < 0.05), and ORF19.2020/HGT6 (bound by Brg1, Efg1, Ndt80, Rob1, and Tec1; Z-test p-value < 0.05) (see Dataset S8 for the complete set of prioritized target genes that were bound by five, four, three, two, and one of the biofilm master TFs, and that had significant expression changes (Z-test p-value < 0.05) averaged across the four biofilm life cycle stages).

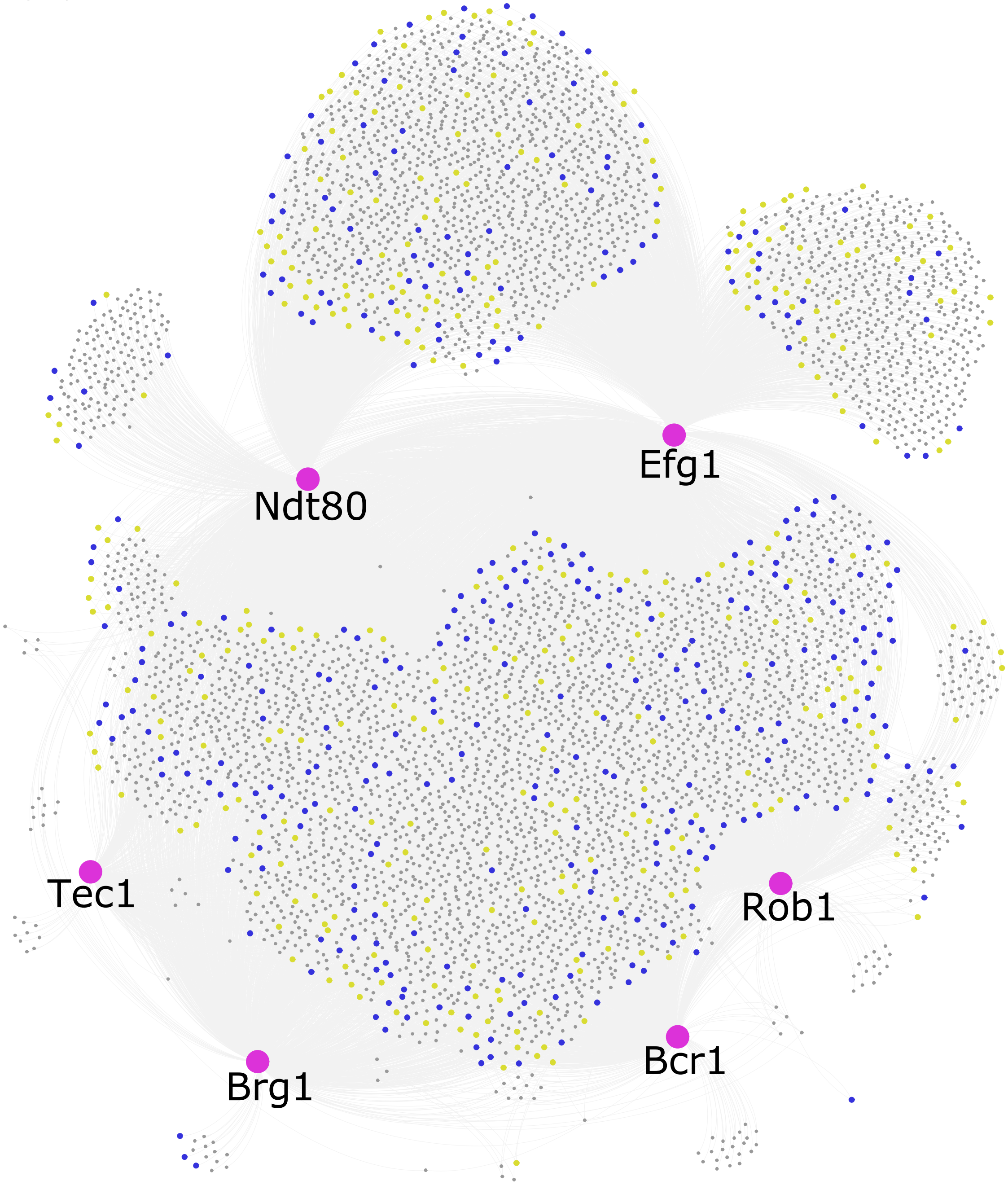

Figure 1.6: Comprehensive transcriptional regulatory network controlling *C. albicans* biofilm development. Transcriptional regulatory network controlling biofilm formation in *C. albicans* encompassing all stages of the biofilm life cycle. The six master biofilm TFs are represented by magenta circles (nodes). Smaller blue (upregulated in biofilms) and yellow (downregulated in biofilms) nodes are target genes. The grey lines (edges) represent high confidence model predicted TF-target gene interactions at all stages of the biofilm life cycle.

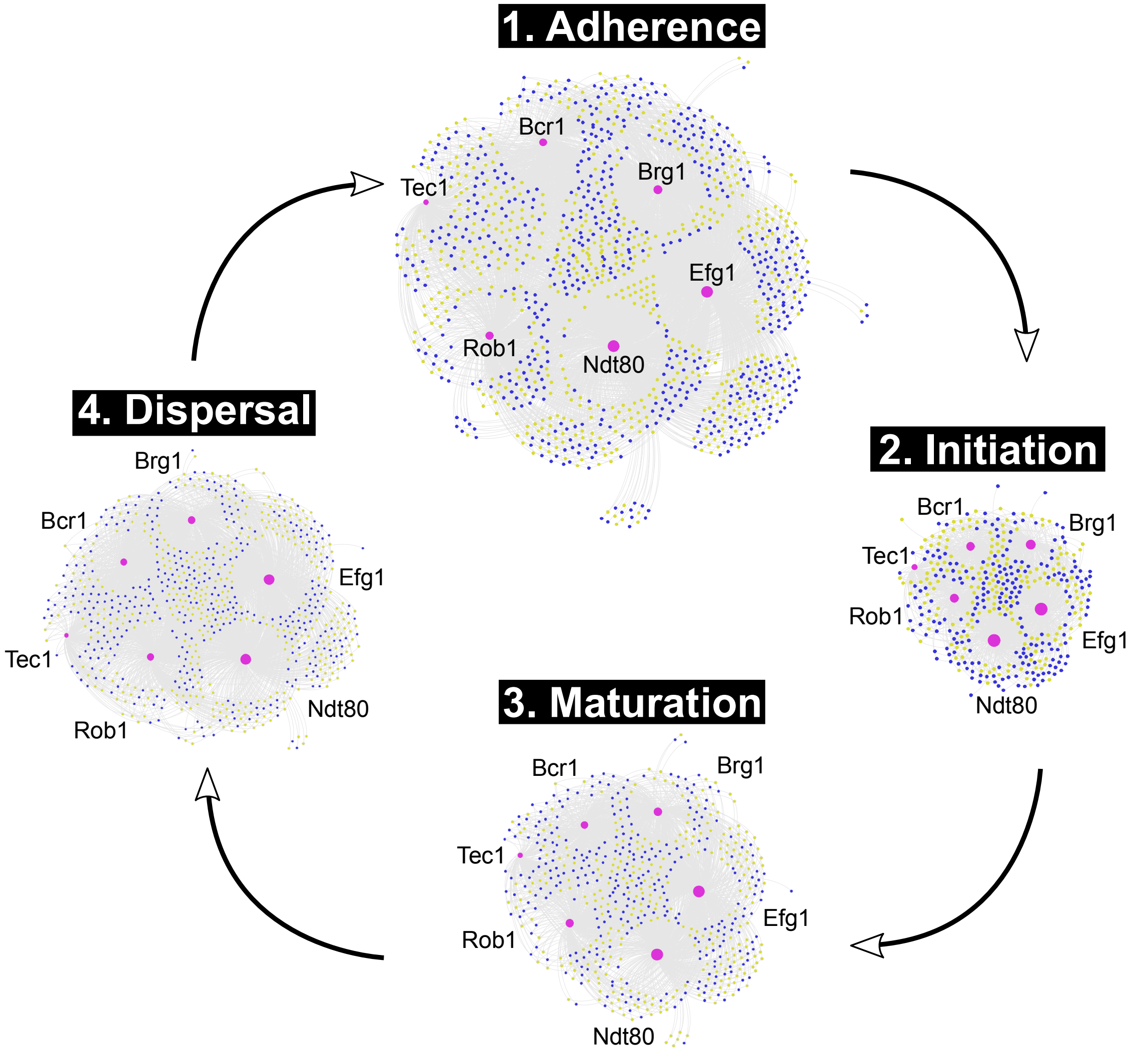

Figure 1.7: Transcriptional regulatory networks controlling *C. albicans* biofilm development at each stage of the biofilm life cycle. Transcriptional regulatory networks controlling biofilm formation in *C. albicans* at each specific stage of the four stages (adherence, initiation, maturation, dispersal) of the biofilm life cycle. The six master biofilm TFs are represented by magenta circles (nodes). Smaller blue (upregulated in biofilms) and yellow (downregulated in biofilms) nodes are target genes. The grey lines (edges) represent model predicted high confidence TF-target gene interactions during each stage of biofilm development. Magenta nodes are sized by their degree counts in each of the four networks.

1.4.6 Gene Ontology (GO) analysis of predicted target genes for each master *C. albicans* biofilm TF

Based on the model predictions, we wanted to identify the Gene Ontology (GO) terms enriched for the predicted target genes of each master biofilm TF at each stage of the *C. albicans* biofilm life cycle. All GO terms presented were obtained using the *Candida* Genome Database GO Term finder tool with Bonferroni corrected p-values < 0.05 (http://www.candidagenome.org/cgi-bin/GO/goTermFinder; accessed on 04/30/2021). Enrichment was calculated against all 6,473 *C. albicans* annotated genes consisting of ORFs, non-coding RNAs (ncRNAs), pseudogenes, ribosomal RNAs (rRNAs), small nuclear RNAs (snRNAs), small nucleolar RNAs (snoRNAs), and transfer RNAs (tRNAs).

A total of 1,502 GO terms were enriched by the predicted target genes of the master biofilm TFs during *C. albicans* biofilm formation (199 by Bcr1, 246 by Brg1, 408 by Efg1, 412 by Ndt80, 182 by Rob1, and 55 by Tec1). To explore the effects of all six master biofilm TFs temporally during biofilm formation, we determined common GO terms enriched for their predicted target genes at each stage of the biofilm life cycle. Shared cellular components influenced by all predicted target genes of all six master biofilm TFs are depicted in **Figure 1.8** and include hyphal cell wall, ribosomal subunit, cytosolic ribosome, cell surface, extracellular region, and ribosome for the adherence stage of the biofilm life cycle; ribosome, cytosolic ribosome, and ribosomal subunit for the initiation stage of the biofilm life cycle; and extracellular region and external encapsulating structure for both the maturation and dispersal stages of the biofilm life cycle. Shared molecular functions influenced by all predicted target genes of all six master biofilm TFs are depicted in **Figure 1.9** and include structural integrity of ribosome for the adherence and initiation stages of the biofilm life cycle; and oxidoreductase activity for the maturation stage of the biofilm life cycle. No shared molecular functions were observed during the dispersal stage of the biofilm life cycle. Shared biological processes influenced by all predicted target genes of all six master biofilm TFs are depicted in **Figure 1.10** and include peptide biosynthetic process, amide biosynthetic process, and translation for the initiation stage of the biofilm life cycle; single-species biofilm formation, biological process involved in symbiotic interaction, cell aggregation, carbohydrate metabolic process, biofilm formation, and aggregation of unicellular organisms for the maturation stage of the biofilm life cycle; and organic acid metabolic process for the dispersal stage of the biofilm life cycle. No shared biological processes were identified during the adherence stage of the biofilm life cycle. **Table 1.2** summarizes the shared enriched GO terms of the predicted target genes for the master biofilm TFs at each stage of the biofilm life cycle. Interestingly, our model predicted novel TFBSs, which included several previously unknown target genes of the six master biofilm TFs. Significantly enriched GO molecular functions for the novel target genes of the six master biofilm TFs included DNA-binding transcription factor activity for Bcr1 (5.5%, 59 out of 1074 target genes, corrected p-value = 0.02759), Brg1 (4.9%, 101 out of 2052 genes, corrected p-value = 0.00284) and Rob1 (6.2%, 55 out of 891 genes, corrected p-value = 0.00165). The novel target genes of Brg1 were specifically enriched for anion transmembrane transporter activity (2.2%, 46 out of 2052 genes, corrected p-value = 0.0454); the novel target genes of Efg1 were specifically enriched for transmembrane transporter activity (6.6%, 305 out of 4599 genes, corrected p-value = 0.04196); and the novel target genes of Ndt80 were specifically enriched for ion binding (17.5%, 900 out of 5155 genes, corrected p-value = 0.00982) and transferase activity (13%, 672 out of 5155 genes, corrected p-value = 0.03345).

Table 1.2: Shared GO terms significantly enriched by the predicted target genes of all six of the master biofilm TFs during each stage of the *C. albicans* biofilm life cycle. NA indicates no shared significantly enriched GO terms.

	Adherence	**Initiation**	**Maturation**	**Dispersal**
Enriched cellular components	hyphal cell wall, ribosomal subunit, cytosolic ribosome, cell surface, extracellular region, and ribosome	ribosome, cytosolic ribosome, and ribosomal subunit	extracellular region and external encapsulating structure	extracellular region and external encapsulating structure
Enriched molecular function	structural integrity of ribosome	structural integrity of ribosome	oxidoreductase activity	NA
Enriched biological process	NA	peptide biosynthetic process, amide biosynthetic process and translation	single-species biofilm formation, biological process involved in symbiotic interaction, cell aggregation, carbohydrate metabolic process, biofilm formation and aggregation of unicellular organisms	organic acid metabolism

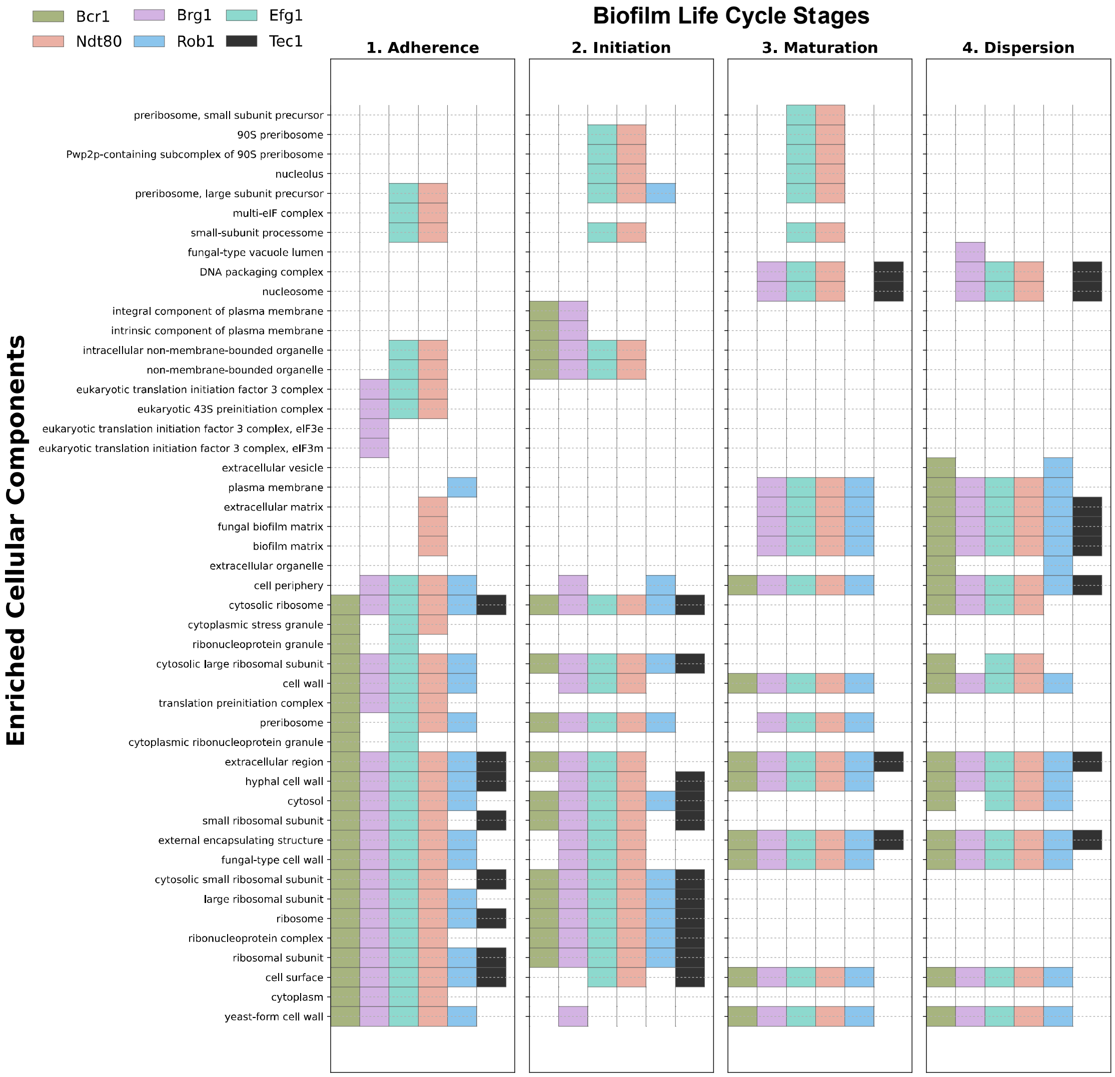

Figure 1.8: Life cycle stage-specific enriched cellular components during *C. albicans* biofilm development. GO terms are depicted on the Y-axis and colored horizontal grouped bars indicate enrichment of predicted target genes by one or more of the six master biofilm TFs for each respective GO term. The presence or absence of a colored bar indicates the TF's influence over a specific GO term.

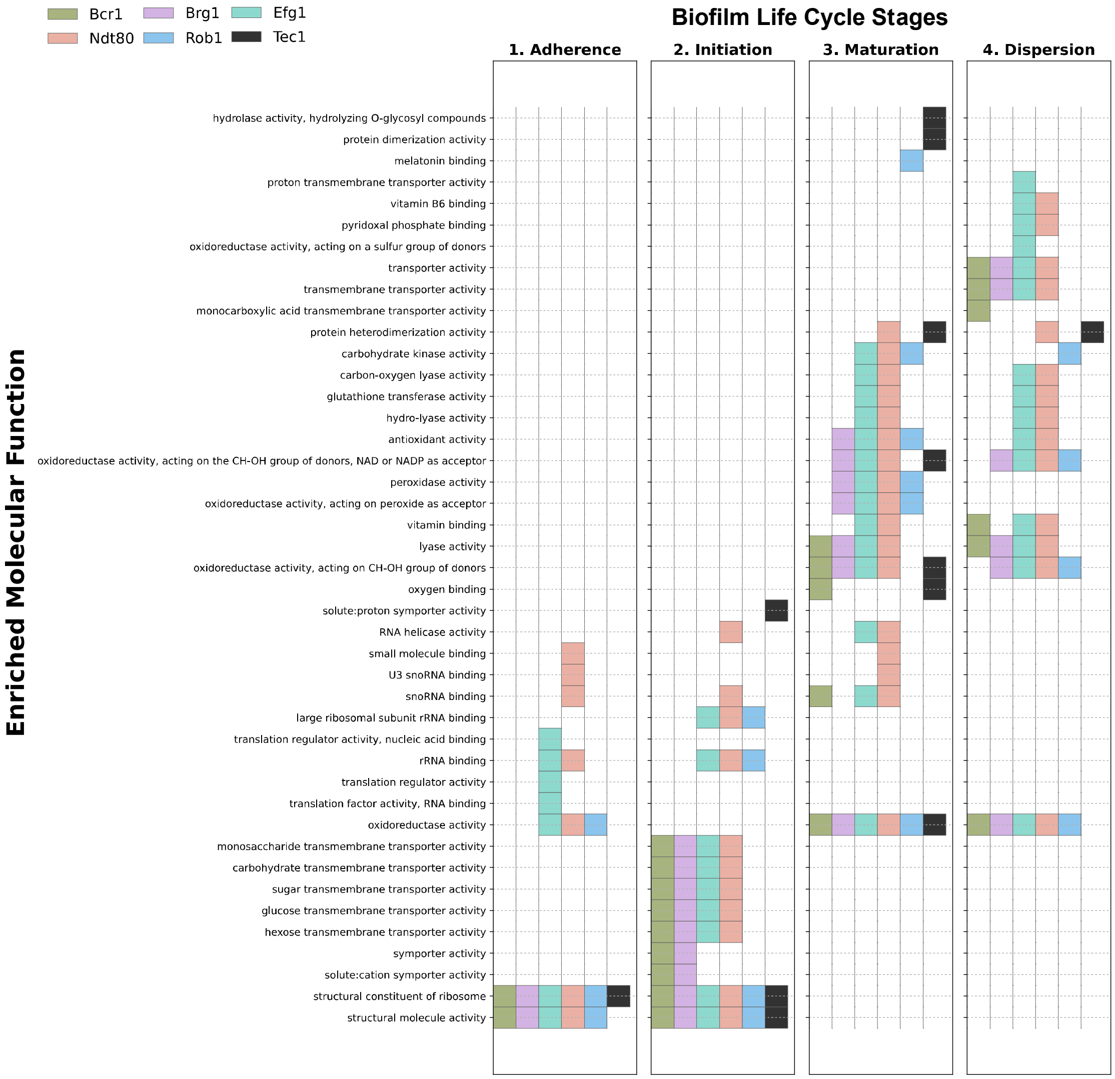

Figure 1.9: Life cycle stage-specific enriched molecular functions during *C. albicans* biofilm development. GO terms are depicted on the Y-axis and colored horizontal grouped bars indicate enrichment of predicted target genes by one or more of the six master biofilm TFs for each respective GO term. The presence or absence of a colored bar indicates the TF’s influence over a specific GO term.

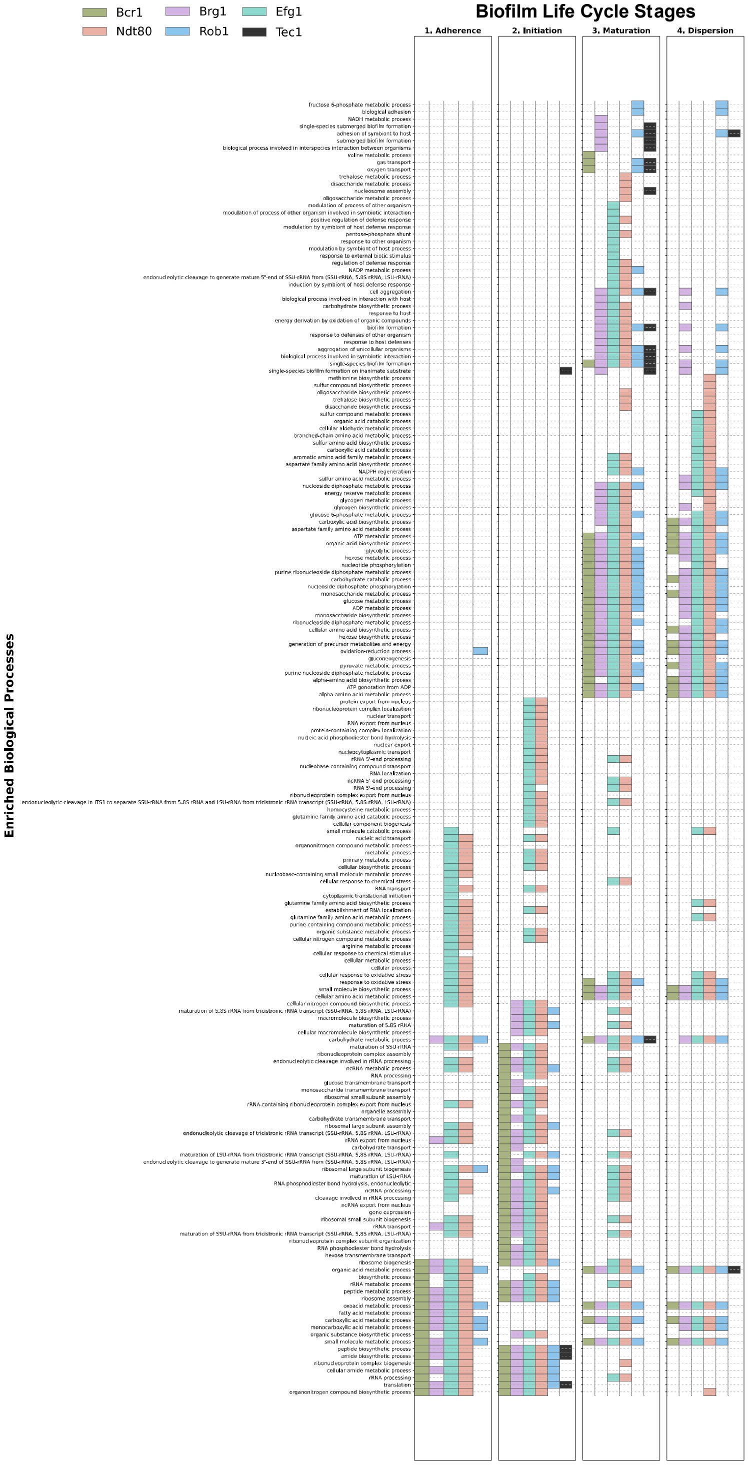

Figure 1.10: Life cycle stage-specific enriched biological processes during *C. albicans* biofilm development. GO terms are depicted on the Y-axis and colored horizontal grouped bars indicate enrichment of predicted target genes by one or more of the six master biofilm TFs for each respective GO term. The presence or absence of a colored bar indicates the TF's influence over a specific GO term.

Enriched GO terms arising from the predicted target genes for all six master regulators are summarized in **Table 1.3** (for cellular components)**, Table 1.4** (for molecular functions)**,** and **Table 1.5** (for biological processes).

Table 1.3: Cellular functions enriched by the predicted target genes of all six of the master biofilm TFs at each stage of the biofilm life cycle.

	Adherence	**Initiation**	**Maturation**	**Dispersal**
Bcr1	hyphal and yeast-form cell wall and cytosolic small and large ribosomal subunits	plasma membrane components, cytosolic ribosomal subunits	cellular periphery including yeast-form cell wall and extracellular region	hyphal and yeast-form cell walls, plasma membrane, fungal biofilm matrix, cytosolic large ribosomal subunit, and extracellular vesicle
Brg1	cytosolic ribosomal large and small subunits, cytosol, yeast-form and hyphal cell wall, cell surface and extracellular region cellular components	preribosome, ribosome, cytosolic large and small ribosomal subunits, intracellular non-membrane-bounded organelle (includes ribosomes, the cytoskeleton, and chromosomes), integral component of plasma membrane, yeast-form and hyphal cell wall and extracellular region	yeast-form and hyphal cell wall, plasma membrane, cell surface and fungal biofilm matrix	yeast-form and hyphal cell wall, plasma membrane, cell surface, fungal biofilm matrix and cytosolic ribosome
Efg1	yeast-form and hyphal cell wall, cytosolic large and small ribosomal subunits, cytoplasmic stress granule,	90S preribosome, preribosome large subunit precursor, small subunit processome protein containing complexes as	fungal biofilm matrix, yeast-form and hyphal cell wall, plasma membrane, extracellular	cytosol, cytosolic large ribosomal subunit, extracellular region, cell surface, plasma membrane, yeast-form and hyphal cell wall and

	translation preinitiation complex, preribosome large subunit precursor and small-subunit processome	well as cell surface, extracellular region, and cytosolic large and small ribosomal subunit	region, and cell surface	fungal biofilm matrix
Ndt80	preribosome large subunit precursor, small subunit processome, cytosolic large and small ribosomal subunits, cytoplasmic stress granule, yeast-form and hyphal cell wall, cell surface, extracellular region, and fungal biofilm matrix	preribosome, cytosolic large and small ribosomal subunits, cell surface, extracellular region, and hyphal cell wall	fungal biofilm matrix, hyphal and yeast-form cell wall, plasma membrane, extracellular region, and cell surface cellular anatomical entities	cytosol, cytosolic large ribosomal subunit, extracellular region, plasma membrane, cell surface, hyphal and yeast-form cell wall, and fungal biofilm matrix
Rob1	preribosome, cytosolic large ribosomal subunit, cell surface, extracellular region, and hyphal and yeast-form cell wall	preribosome, cytosolic large and small ribosomal subunits and cellular periphery	extracellular region, cell surface, fungal type cell wall and plasma membrane	cytosol, extracellular vesicle, plasma membrane, cell surface and fungal type cell wall
Tec1	cytosolic small ribosomal subunit, hyphal cell wall, cell surface and extracellular	hyphal cell wall, ribonucleoprotein complex, cell surface and cytosolic large and small ribosomal subunits	DNA packaging with nucleosome	external encapsulating structure, extracellular region, and DN packaging with nucleosome

Table 1.4: Molecular functions enriched by the predicted target genes of all six of the master biofilm TFs at each stage of the biofilm life cycle.

	Adherence	**Initiation**	**Maturation**	**Dispersal**
Bcr1	ribosomal structural integrity	ribosomal structural integrity	oxidoreductase activity acting on diphenols, oxidizing metal ions, NAD(P)H, heme-group, and peroxide as donors	oxidoreductase, lyase and transmembrane transporter
Brg1	ribosomal structural integrity	Ribosomal structural integrity, solute:cation symporter activity and glucose transmembrane transfer activity	antioxidant and oxidoreductase activities	oxidoreductase activity acting on the CH-OH group of donors, NAD or NADP as acceptor
Efg1	rRNA binding, oxidoreductase activity and structural integrity of ribosome	structural integrity of the ribosome and large ribosomal subunit rRNA binding	peroxidate activity, oxidoreductase activity acting on the CH-OH group of donors, NAD or NADP as acceptor, and carbon-oxygen lyase activity	oxidoreductase activity acting on (1) a sulfur group of donors and (2) acting on the CH-OH group of donors, NAD or NADP as acceptor, and carbon-oxygen lyase activity
Ndt80	structural integrity of ribosome, oxidoreductase activity, small molecule and rRNA binding	RNA helicase activity, structural integrity of ribosome and large ribosomal subunit rRNA binding	oxidoreductase activity, acting on the CH-OH group of donors with NAD or NADP as acceptor, small molecule binding and carbon-oxygen lyase activity	antioxidant activity, oxidoreductase activity acting on the CH-OH group of donors with NAD or NADP as acceptor, and glutathione transferase activity
Rob1	ribosome structural integrity and	ribosomal structural integrity and	melatonin binding, oxidoreductase	oxidoreductase activity acting on the CH-OH group

	oxidoreductase activity	large ribosomal subunit rRNA binding	activity and peroxidase activity	of donors and NAD or NADP as acceptor
Tec1	structural integrity of ribosome	structural integrity of ribosome and solute:proton symporter activity	towards hydrolase activity hydrolyzing O-glycosyl compounds, oxidoreductase activity, oxygen binding and protein heterodimerization activity	protein heterodimerization activity

Table 1.5: Biological processes enriched by the predicted target genes of all six of the master biofilm TFs at each stage of the biofilm life cycle.

	Adherence	**Initiation**	**Maturation**	**Dispersal**
Bcr1	translation, fatty acid metabolism, glycolytic process, carbohydrate, glucose and vitamins and carboxylic acid metabolism	rRNA processing and messenger RNA export from nucleus	ATP, organic acid, glucose, and carboxylic acid metabolism	cellular amino acid and alpha-amino acid biosynthesis, carbohydrate metabolism and ATP production via glycolytic process
Brg1	translation, carbohydrate and fatty acid metabolism and ribosome assembly	translation, cellular macromolecule biosynthesis, ribonucleoprotein complex biogenesis and rRNA transcript maturation	defensive response to host, single-species submerged biofilm formation, NADH metabolic process, carboxylic acid metabolism, energy derivation by oxidation of organic compounds, glucose metabolism and glycolytic process	alpha amino acid and carboxylic acid biosynthesis, glycolytic process, and nucleoside diphosphate metabolic process

Efg1	nucleobase containing small molecule metabolism, cellular amino acid metabolism, ncRNA metabolism, translation, fatty acid metabolism and cellular response to chemical stimuli	endonucleolytic cleavage of tricistronic rRNA transcript, maturation of 5.8S, LSU and SSU rRNA from tricistronic rRNA transcript, ncRNA processing metabolic processes as well as rRNA-containing ribonucleoprotein complex export from nucleus	ncRNA metabolism, aspartate family amino acid biosynthesis and metabolism, carboxylic acid biosynthesis, glucose metabolism, glycolytic process and biological process involved in symbiotic interaction with host	glucose metabolism, glutamine amino acid metabolism, aspartate and sulphur amino acid metabolism, sulphur compound metabolism, carbohydrate catabolic process, and glycolytic process
Ndt80	cellular nitrogen compound biosynthesis, nucleobase-containing small molecule, cellular amino acid, ncRNA and fatty acid metabolism, translation, and cellular-level biological process	translation, ncRNA processing, maturation of SSU, LSU and 5.8S rRNA from tricistronic rRNA transcript, ribosome large subunit assembly and rRNA-containing ribonucleoprotein complex export from nucleus	symbiotic interactions, glycolysis, aspartate family amino acid biosynthesis, glucose metabolisms and ncRNA metabolism	sulphur compound, glutamine, aspartate family and methionine biosynthesis, glucose and sulphur amino acid metabolism, carbohydrate catabolism, and glycolysis
Rob1	carbohydrate, monocarboxylic acid, and small molecule metabolism. Rob1 controls ribosome structural integrity and oxidoreductase activity	ribosome biogenesis, ncRNA and rRNA processing, and translation	single-species biofilm formation, carboxylic acid, fructose 6-phosphate and glucose metabolism, NADPH regeneration and glycolytic process. Rob1's functions include melatonin binding, oxidoreductase	alpha amino acid biosynthesis and metabolism, glucose, fructose 6-phosphate sulphur amino acid and carboxylic acid metabolism and glycolytic process

			activity and peroxidase activity	
Tec1	NA	translation via amide biosynthetic process	carbohydrate metabolism, symbiotic interactions, and single-species submerged biofilm formation	adhesion of symbiont to host process

1.4.7 Network centrality provides additional insights into potentially influential target genes

We wanted to get an orthogonal view towards important or influential target genes in the biofilm regulatory network. Therefore, we implemented network theory-based "information flow centrality" metric to get a list of target genes which show control over the flow of information in the network (Brandes and Fleischer, 2005). This centrality metric treats the edges as information highway, and nodes as mediators of information flow. These target genes were identified for each master regulator and their significance was calculated using Z-test and the p-values were corrected with 5% false discovery rate. **Table 1.6** shows significant target genes of Bcr1, Brg1, Rob1 and Tec1. Ndt80 and Efg1 target genes did not show up as significant after implementing false discovery rate correction.

Table 1.6: Significant target genes based on information flow centrality metric. Z-test, FDR corrected p-values < 0.05.

Bcr1	**Brg1**	**Rob1**	**Tec1**
CaalfMp05	orf19.2108	orf19.1142	orf19.2018
orf19.1045	orf19.233.1	orf19.2228	orf19.2131
orf19.1249	orf19.2831	orf19.2301	orf19.3234.1
orf19.1308	orf19.2889	orf19.2677	orf19.3314
orf19.1492	orf19.3775	orf19.2697	orf19.5663
orf19.1633	orf19.3921	orf19.3287	orf19.6650
orf19.2190	orf19.5808	orf19.336	orf19.672
orf19.2720	orf19.5905	orf19.4234	orf19.90
orf19.4031	orf19.6115	orf19.5053	orf19.979
orf19.4306	orf19.6701	orf19.5431	
orf19.4369	orf19.701	orf19.6173	

orf19.5039	orf19.7154	orf19.6175	
orf19.5056	orf19.7436.1	orf19.7057	
orf19.519	orf19.7645	orf19.7443	
orf19.5274			
orf19.5413			
orf19.5445			
orf19.5767			
orf19.6263			
orf19.6625			
orf19.980			

1.5 Discussion

In this study, we created a computational workflow using a supervised support vector machine (SVM) classifier and multimodal genome-wide datasets, to predict novel TF-target gene interactions for each biofilm master TF temporally over the course of the *C. albicans* biofilm life cycle. Our approach is the first to model stage-specific regulatory networks controlling *C. albicans* biofilm formation. First, we created target sequences using previously identified transcription factor binding site (TFBS) consensus sequences that represent potential binding sites. Second, we built a feature matrix to capture the DNA shape and sequence qualities of each candidate TFBS motif. Third, we predicted a positive/true set of potential TFBSs for each TF using our trained SVM classifier based on a feature matrix. Lastly, we identified “active” TF-target gene interactions by correlating TF binding activity with the time-series gene expression data of target genes. Overall, by coupling TFBS sequence and DNA shape information, we successfully predicted novel TFBSs, TF-target gene interactions, and ultimately, entire transcriptional regulatory networks controlling each stage of the *C. albicans* biofilm life cycle.

The transcriptional regulatory networks controlling the four stages of the *C. albicans* biofilm life cycle that we present in this study are intricate, highly interconnected and are representative of small-world networks (Fox et al., 2015; Nobile et al., 2012). A small-world network is characterized by the presence of short paths between any two nodes within the network as well as the presence of network hubs, which act as important mediators of signal propagation within the network (Albert and Barabási, 2002; Camacho et al., 2018). The overall architecture of the *C. albicans* biofilm network is highly dynamic, with constantly changing TF-target gene interactions at each stage of the biofilm life cycle. The biofilm network appears to be structured in a way that allows *C. albicans* cells to respond to and adapt to environmental changes quickly and efficiently, yet also provides robustness to the network (Nobile et al., 2012). Our findings reinforce the idea that target genes of the six master biofilm TFs are differentially expressed at different stages of the biofilm life cycle (Fox et al., 2015) and indicate that different combinations of target genes are controlled by multiple master biofilm TFs at various stages of the biofilm life cycle.

Based on our GO analyses, we found that Tec1 influences target genes involved in hyphal formation in the first two stages of biofilm formation (adherence and initiation), while Efg1 influences target genes involved in hyphal formation specifically during adherence, maturation, and dispersal stages of biofilm formation. Interestingly, while both Efg1 and Tec1 influence target genes involved in hyphal formation, Efg1 enriched target genes are involved in the extracellular region of cellular components, while Tec1 enriched target genes are involved in the cell wall region of cellular components, suggesting that Efg1 and Tec1 contribute distinct controls over their target genes involved in hyphal formation. In the adherence and initiation stages of the biofilm life cycle, analysis of the enriched cellular components provides an entry point towards understanding the complex GO data. As is evident from our transcriptional regulatory networks, Efg1 and Ndt80 primarily exert combinatorial control over their active target genes with over 93% of target ORFs in common. Recently, Mancera et al. (2021) also found evidence of 91% overlap in the target genes between Efg1 and Ndt80 between four *C. albicans* and its closest relative C. dubliniensis (Mancera et al., 2021). Together, these two TFs influence many downstream target genes producing large scale changes to the biofilm regulatory network. Analyzing other TF pairs, we found that Bcr1 and Efg1 (2882/5664 50.9%), Brg1 and Efg1 (3392/5728 59.2%), Brg1 and Ndt80 (3479/5907 58.9%), Brg1 and Rob1 (2170/4328 50.1%), Efg1 and Rob1 (2953/5677 52%) and Ndt80 and Rob1 (2998/5898 50.8%) pairs shared 50% or more of their active target genes with Brg1, Efg1 and Rob1 participating in three of the six TF pairs indicating their high connectivity within the biofilm regulatory network. Altogether, the prevalence of TF pairs with high frequency common target genes hints at the compensatory nature of the biofilm regulatory network. Interestingly, the TFBS features of two other TF pairs also seem to follow similar patterns – Efg1 and Ndt80 (MinHash and Poisson product score) as well as Bcr1 and Brg1 (MinHash and EP at position 6) share the top two influential features indicating similar behavior.

To extract relevant translational information from our analysis of the complex biofilm transcriptional networks, we compared the average log2 fold change value of each individual target genes over all four biofilm growth stages (target gene mean expression) to mean log2 fold change of all target genes across all biofilm growth stages (global mean expression). We then identified significant target genes using Z-test, corrected with 5% false discovery rate. Based on this analysis, we came up with three target genes, encoding proteins that we hypothesize could be useful therapeutic targets: Orf19.2870/Ldg11, Orf19.2762 (Ahp1), and Orf19.2020 (Hgt6). Ldg11 is an uncharacterized protein of *C. albicans*. According to *Candida* Gene Browser Order, its orthologs are found only in closely related C. dubliniensis and C. tropicalis (Fitzpatrick et al., 2010; Maguire et al., 2013). Ldg11 has no significant human ortholog, making it a good viable candidate for antibiofilm drug target.***Ahp1 is an alkyl hydroperoxide reductase that is expressed in response to stress, including the antifungal drug fluconazole, and is repressed in response to the antifungal drugs amphotericin B and caspofungin (Bonhomme et al., 2011; Copping et al., 2005; Enjalbert et al., 2006; García-Sánchez et al., 2005; Karababa et al., 2004; Liu et al., 2005; Maglott et al., 2007; Nett et al., 2009; Seneviratne et al., 2008; Singh et al., 2011). Ahp1 belongs to a class of molecules called peroxiredoxins (Prxs), which mediate cellular responses to reactive oxygen species (ROS). Prxs are critical in the defense of pathogens to host produced ROS, and upon deletion have been reported to have significant effects on the growth rates of the fungal pathogen Cryptococcus neoformans

(Gretes et al., 2012). Interestingly, Prxs are currently being pursued as therapeutic targets for eukaryotic pathogens. For example, Prx1a in Leishmania major and Prx1a in Leishmania donovani show promise as vaccine targets against infections caused by these parasitic pathogens (Gretes et al., 2012). Although the antifungal drug amphotericin B causes oxidative damage to cells, it does not directly target Ahp1 activity in *C. albicans* (Al Balushi et al., 2018; Rybak et al., 2019). Similarly, the antifungal drug caspofungin inhibits synthesis of an essential cell wall component, β-(1,3)-D-glucan (McCormack and Perry, 2005), and thus does not target Ahp1. Thus, Ahp1 is a new potential antifungal drug target that is worthy of further exploration. Hgt6 is a putative high-affinity major facilitator superfamily (MFS) glucose transporter that is induced by fluconazole and general cell stresses (Bonhomme et al., 2011; Copping et al., 2005; Enjalbert et al., 2006; Fan et al., 2002; Fanning et al., 2012; Nobile et al., 2012). Inhibiting sugar transporters, such as Hgt6, could result in cell starvation. For example, disrupting hexose transporter genes in the parasite Plasmodium falciparum has been shown to lower intracellular ATP levels, leading to starvation (Slavic et al., 2011). In humans, glucose transporter inhibitors have been successfully targeted for use in treating type 2 diabetes in patients by lowering glucose levels independent of insulin (Lin et al., 2015).

In addition to using gene expression as the basis to identify key target genes, we implemented network theory based centrality measures to assess an orthogonal view of target genes. To identify influential target genes for each of the six master regulator, we utilized the “information flow” centrality metric (Brandes and Fleischer, 2005). Table 5 lists the influential target genes for Bcr1, Brg1, Rob1 and Tec1. Interestingly, Efg1 and Ndt80 target genes did not pass through multiple hypothesis correction threshold. There are no common or shared genes among the four master regulators. However, Bcr1 and Tec1 targets - orf19.5445 and orf19.3314 respectively – are functionally similar as these target genes are involved in endoplasmic reticulum to Golgi vesicle-mediated transport. Since transmembrane transporter genes have been identified as drug targets, we focus on these genes and list them out next. Bcr1 controls orf19.1308 (member of the drug:proton antiporter (14 spanner) (DHA2) family) and orf19.6263 (predicted MFS membrane transporter), in addition to orf19.5445. Rob1 has two transporter target genes – orf19.1142 (putative vacuolar transporter of large neutral amino acids) and orf19.2697 (regulation of dipeptide transmembrane transport by regulation of transcription from RNA polymerase II promoter and cytoplasm). Of all the target genes, only fatty acid biosynthesis gene orf19.979 (Fas1) has been shown to be susceptible to current antifungal drugs amphotericin B and caspofungin (Liu et al., 2005). Further investigation into these target genes might provide additional insights into their influencing capabilities.

SVM is categorized as a supervised machine learning method where the model utilizes the information in the feature table to classify input data. SVMs are effective for high-dimensional data such as the feature table used in this study and are versatile in their use of kernels. Due to these advantages, we implemented SVMs for identifying novel TFBS controlling *C. albicans* biofilm formation. While building the SVM classifier, we utilized PCAs to visualize the feature space of the input data – experimentally validated TFBS from Nobile et al. (2012) and simulated background data. Based on the data separation, we implemented a Grid Search approach to optimize SVM hyperparameters. During model validation, the radial basis function (RBF) kernel of the SVM consistently was the top performer based on accuracy and precision values. To control model run time and to achieve consistent model behavior, we chose RBF and linear kernel as two options for Grid Search while classifying novel genome wide TFBS. An interesting follow up to model classifications is the presence of low confidence predictions as an indicator that the classifier performance is not optimal. Low confidence predictions could originate from the experimental data used to train the SVM model. For example, the experimental TFBS data obtained from Nobile et al. (2012) for mature *C. albicans* biofilms (Nobile et al., 2012) is unlikely to contain all possible TFBSs for the TFs since some of the binding sites could be occluded due to common experimental artifacts of the chromatin immunoprecipitation procedure, such as variable crosslinking, variable chromatin fragmentation, and epitope masking (Wardle and Tan, 2015). Our SVM-based methodology provides an agnostic approach to TFBS prediction and can be generalized for use in other systems. The input to the SVM model is a list of True TF binding site sequences, which can be obtained from genome-wide binding studies or from various existing databases such as Encode, Jasper, and NCBI. The workflow contains Python scripts, which aid in negative training data generation, creating feature tables, training the SVM model and evaluating its performance and eventually predicting TFBS in test data. The test data would need to be created separately and provided to the model for during the classification stage.

Overall, our study utilizes experimentally validated data from Nobile et al. (2012) to begin to dissect the regulatory control of the *C. albicans* biofilm regulatory network (Nobile et al., 2012). Our analysis revealed the dynamic nature of TFs over each biofilm developmental stage. We acknowledge that there are limitations in the experimental aspects of the data, such as in the fact that the TFBS data from Nobile et al. (2012) was based solely on a mature biofilm, which likely precludes capturing all possible TFBSs. Additionally, chromatin immunoprecipitation-based protocols are affected by epitope masking, making it likely that some TFBS are missed. Future experimentation using more modern protocols, such as Assay for Transposase-Accessible Chromatin with high-throughput sequencing (ATAC-seq) and cleavage under targets and release using nuclease (CUT&RUN) sequencing, will certainly be beneficial in improving the model predictions (Buenrostro et al., 2015; Fox and Nobile, 2012; Skene and Henikoff, 2017).

1.6 Future directions

Given that *C. albicans* is a major human fungal pathogen, one goal of my thesis work has been to accelerate biomarker discovery for fungal infections caused by *C. albicans*. One milestone in achieving this goal is to fill in gaps in our knowledge about *C. albicans* biofilm formation, one important virulence trait of *C. albicans*. My computational approach provides predictive insights into genes, which could potentially be targeted to disrupt biofilm formation in this fungal pathogen. Based on my results, I pose the following questions as immediate next steps to my thesis work:

1. Are Ahp1 and Hgt6 potential targets for disrupting biofilm formation in *Candida albicans*?
2. What is the logic behind TF-target gene interaction changes throughout the *C. albicans* biofilm lifecycle?
3. Where do master regulators bind to DNA during each biofilm growth stage?

1.7 Methods

1.7.1 Computational transcription factor binding site identification

Approach

Transcription factors bind to DNA at specific regions in the genome (Berman et al., 2002; Fox et al., 2015; Gulati and Nobile, 2016; Inukai et al., 2017; Lohse et al., 2018; Mathelier and Wasserman, 2013; Nobile et al., 2012; Nobile and Johnson, 2015; Nobile and Mitchell, 2005; Rogers and Bulyk, 2018; Schweizer et al., 2000). TFs dynamically bind and unbind to DNA to regulate the timely expression of their target genes. *C. albicans* transcription factor behavior has been probed using chromatin immunoprecipitation followed by microarray (ChIP-chip) (Fox et al., 2015; Nobile et al., 2012). However, due to the pulse-like TF-DNA interactions and limitations of the ChIP protocol, genome-wide binding activity of biofilm regulators has been intractable and prone to false-negative results. In this paper, we use experimentally verified TFBS regions as tools to discover novel TF binding that has not yet been experimentally observed. We characterized the binding motifs by measuring their sequence and shape features (detailed below). Then, we built a supervised learning workflow to learn these motif features and classify novel sequences as true or false TFBS. A supervised machine learning classifier model uses labels assigned to input samples to learn the identity of each label, whereas an unsupervised model does not possess the sample label information. Here, we create a supervised support vector machine (SVM) classifier model by training it on a set of true (experimentally identified) TFBSs and a set of carefully curated, but false, TFBSs. We then allow the SVM to characterize all possible binding sites in the *C. albicans* genome that are TFBS candidates by sequence motif alone.

Features

Previous studies have provided evidence for how TFs recognize their binding sites (Camacho et al., 2018; Inukai et al., 2017; Kantorovitz et al., 2009; Mathelier et al., 2016; Rogers and Bulyk, 2018; Sinha et al., 2003; Zhou et al., 2015). Chiu et al. (2015) and Mathelier et al. (2016) have shown computational approaches to identifying novel TFBS using binding motifs (Chiu et al., 2016; Mathelier et al., 2016). These approaches established that TF activity is influenced by DNA sequence as well as shape (Liu et al., 2009; Rogers and Bulyk, 2018). In this paper, we implemented a DNA sequence similarity metric based on the Poisson distribution (Kantorovitz et al., 2009; Van Helden, 2004) (Kantorovitz et al., 2009; Van Helden, 2004). The Poisson distribution provides the probability of finding at most x number of common k-mers in both sequences, given a background expected count value for the k-mer of interest (Kantorovitz et al., 2009; Van Helden, 2004). Given N_i counts of kmer *i* with expected m_i count, the probability of finding N_i counts is:

$$P\ (x \geq N_i) = \begin{cases} [1 - \ Poisson(N_i - 1, m_i)]^2\ , & N_i > 0 \\ 1 & , otherwise. \end{cases}$$

The Poisson additive similarity score (PAS) and Poisson multiplicative similarity score (PPS) measure the similarities between the true binding motif and novel sequences based on k-mer counts (k-length subsequence of a motif) overlap in both sequences. For p k-mers, PAS and PPS are calculated as:

$$PAS = \frac{1}{p}\sum\nolimits_{i=1}^{p} 1 - P(x\ \geq N_i)$$

$$PPS = 1 - \sqrt[p]{\prod_{i=1}^{p} P(x \geq N_i)}.$$

PAS and PPS scores increase with higher counts of common k-mers between two sequences and if no common k-mers are found, the scores are zero (Kantorovitz et al., 2009).

Ondov et al. 2016 presented another alignment-free sequence similarity analytical method called MinHash (Ondov et al., 2016). In MinHash, k-mers are identified by sliding a k-length window across sequences to be compared. Comparing each k-mer with its reverse complement, only the canonical k-mer is chosen as the representative. Then, the nucleotide k-mers are converted to integer values using the hash function. Here, we used the mm3 Python package to generate 32-bit hashes for k-mers. Distinct k-mers will generate different 32-bit hash integers, while same k-mers will have one 32-bit hash integer. The integer representation of k-mers is compared to assess overlap and measured as a Jaccard similarity index. The Jaccard similarity index for two sequences, seqA and seqB, is calculated by:

$$Jaccard\ similarity = \frac{seqA_{kmers} \cap seqB_{kmers}}{seqA_{kmers} \cup seqB_{kmers}},$$

where the numerator represents the number of common k-mers (set intersection) found in seqA and seqB and the denominator is the total number of k-mers (set union) from seqA and seqB.

Additionally, DNA shape values are used to characterize motif sequences, and these are calculated using the DNAShapeR package in R (Chiu et al., 2016). Minor groove width (MGW), Roll, Propeller twist (ProT), Helix twist (HelT) and electrostatic potential (EP) are the five shape values provided by the package. These five shapes are measured for each base of the input motif sequence, using a pentamer window centered on base of interest (Chiu et al., 2016). Hence, for a 5 bp length sequence, the resulting shape values will be (n x 5) 25 shape values.

Classification Model: By combining the similarity score and DNA shape feature, separate feature tables were calculated for each of the six master biofilm TFs. Using principal component analysis (PCA) to reduce dimensionality and visualize the distribution of true (foreground) and curated background (negative) datasets, support vector machine (SVM) using an RBF kernel was chosen for implementing the classification model (Pedregosa et al., 2011).

1.7.2 Generating background (negative) datasets

Foreground (true positive) sequences were obtained from Step 1. Compared to foreground sequences, the background sequences were created to have similar GC percent and length to avoid sequence composition bias. Two approaches are used to generate background sequences:

1. GC percent and length matched background sequences: Exonic sequences for *Escherichia coli* (Ec) (release version GCF_000005845.2_ASM584v2) and the coding sequence for *Candida albicans* (Ca) (SC5314, assembly 22) and

Drosophila melanogaster (Dm) (release version 6.30) were used as templates (Skrzypek et al., 2017; Thurmond et al., 2019). Three second-order Markov models were trained using each of the three genome sequence datasets. For each foreground sequence, one of the three Markov models was randomly utilized for generating length and GC percent matched background sequence. The background sequence generated by the Markov model was excluded if it was found in the foreground sequence list and if the background sequence did not match the GC percent of the foreground sequence.

2. Dinucleotide shuffled background sequences: Similar to Step 1, each foreground sequence is permuted to create the background sequence. This approach maintains the dinucleotide frequency of the foreground sequence. P. Clote's implementation of Erikson and Altschul's dinucleotide shuffling algorithm was utilized in this approach (Altschul and Erickson, 1985).

Each of these approaches generated a separate background sequence set with an equal number of sequences as those in the foreground sequences. As a result, there are two times as many background sequences as foreground sequences.

Building the feature table

Each sequence is characterized by two Poisson based metrics for sequence similarity and DNA shape for each nucleotide position. Similarity metrics use the overlap of k-mer patterns (subsequence) between two sequences with the count of k-mer overlap modeled as a Poisson probability (Van Helden, 2004). Poisson additive score (PAS) calculates the additive effect of kmer overlap while the Poisson product score (PPS) characterized the k-mer overlap as independent events. DNA shapes were calculated using the DNAShapeR tool (Chiu et al., 2016). DNAShapeR outputs five shape values for each position of the input sequence. The shape values measure electrostatic potential (EP), helix twist (HelT), minor groove width (MGW), propeller twist (ProT), and Roll for a sequence. For an n length DNA sequence, 5n shape values were computed. The feature table was saved in a binary feather file format.

Training and cross validation of the SVM model

The background data was combined with the foreground sequence data to train the support vector machine classifier model. This dataset, consisting of n foreground sequences and 3n background sequences, was split into two categories: (a) training and testing data and (b) validation data. 80% of the samples were randomly chosen for training the model and tuning the model's parameters. The remaining 20% of the samples were reserved for evaluating the model's accuracy. This cross-validation approach minimizes the issue of model overfitting, where the model is hyper tuned for training data and does not perform well on yet unseen real-world data sets. Model parameter C was tuned using a randomized grid search over a range of values. Permutation testing was used to test whether model classifications are significant or not (Ojala and Garriga, 2009). Lastly, model accuracy was measured using the validation dataset and plotted as a confusion matrix, which is a two-by-two matrix of true vs. predicted input labels. The diagonal elements of the confusion matrix represent the number of inputs which were correctly

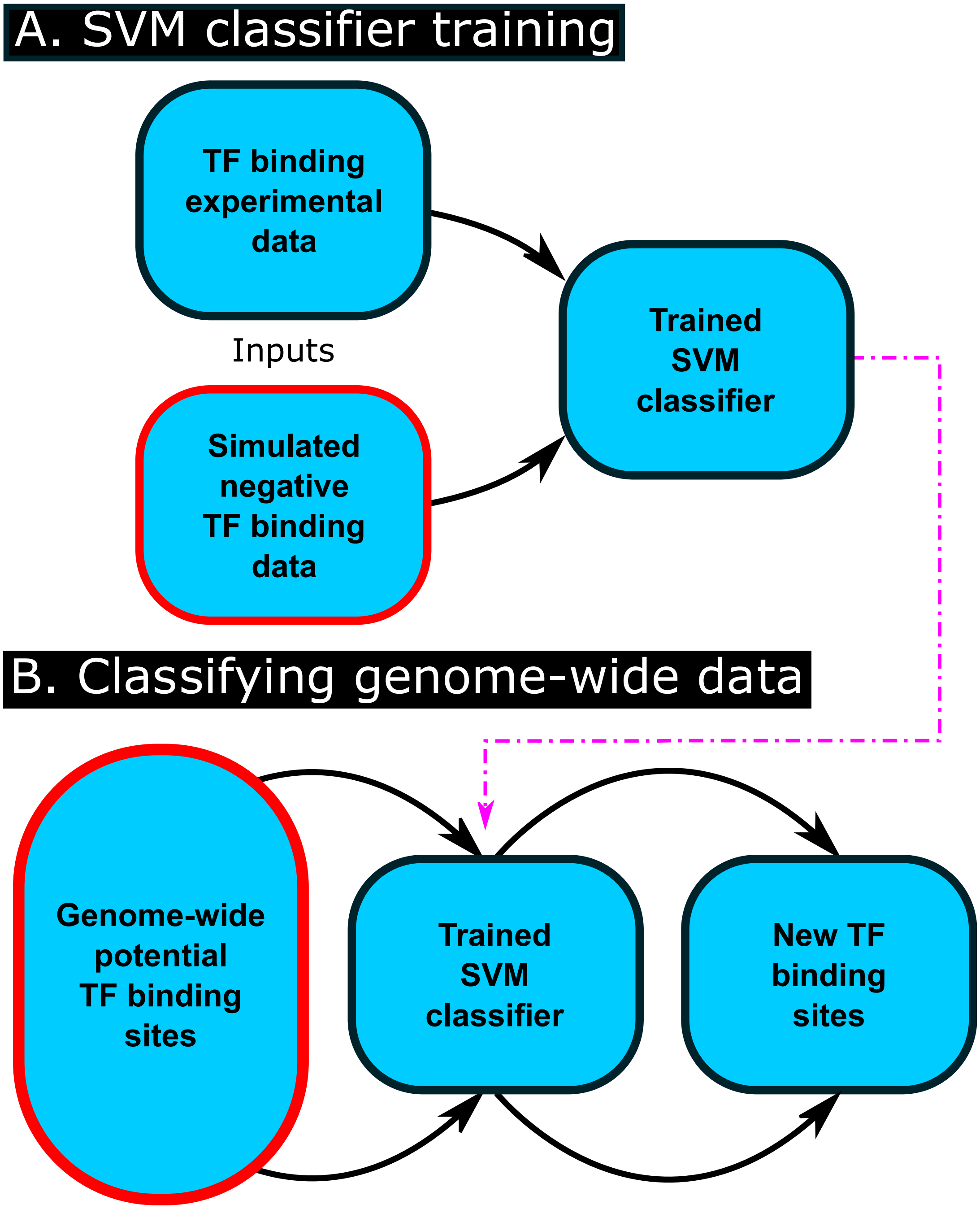
A. SVM classifier training
TF binding experimental data
Inputs
Simulated negative TF binding data
Trained SVM classifier
B. Classifying genome-wide data
Genome-wide potential TF binding sites
Trained SVM classifier
New TF binding sites

Figure 1.11: Methods overview. The top panel (A) denotes the training process for the support machine learning classifier. The bottom panel (B) provides the process to classify genome-wide TF binding sites identified through BLASTn. The red outlined boxes denote novel implementations to generate negative data to create an informative classifier. Whereas in the bottom B panel, BLASTn was utilized to identify all regions matching the consensus TF motif for all master regulators.

classified, while the off-diagonal elements represent misclassification errors. For each TF, the trained model is saved in a binary feather file format, allowing for reuse of the trained model on new datasets. Once the model is saved in a file, the feather file can be used to directly classify new inputs, without having to train the model. **Figure 1.11** provides an overview of SVM model training and testing phases.

Classifying genome-wide matches of TF binding regions as True binding sites

The genome-wide search of TF binding sites using the consensus sequence for each TF was performed using BLASTn (Altschul et al., 1990). In total, 1,914,803, 526,808, 1,168,070, 5,184,038, 2,762,292, and 42,335 potential TF binding sites were returned by BLASTn for Bcr1, Brg1, Efg1, Ndt80, Rob1, and Tec1, respectively. These binding regions (search space) represent all possible regions where each TF can bind to the *C. albicans* genome. A feature table is created for all binding sites in the search space, as specified in the building feature table section above. Using the feature table and the trained SVM model, every region in the search space is classified as either a true binding site or a false binding site. The binding sites classified as True by the model are saved in a BED file format.

Associating TFBS to target genes

After obtaining high-confidence positive TFBS predictions, intergenic predictions were identified using an intersect function with f parameter set to 1.0, for full intersection over the full length of TFBSs using Pybedtools (Dale et al., 2011; Quinlan and Hall, 2010). Each of the intergenic TFBSs were associated to their closest ORFs using the Pybedtool closest function with D="b" parameter; the genome file was supplied to -b. With these results, model predictions were converted from TFBS genome locations to gene domain.

Identifying significant targets as putative drug targets

For each biofilm master TF, we accumulated its target's gene expression (log fold change (LFC) values) from Fox et al. (2015), along with systematic gene name and gene description from CGD (Fox et al., 2015; Skrzypek et al., 2017). The LFC values were averaged across all four biofilm growth stages, and the mean LFC values were converted to Z scores by subtracting the global mean and dividing by standard deviation of mean expression values. Since these Z scores are now assumed to follow a standard Normal distribution, they were converted to p-values using Scipy's norm.cdf() function (Virtanen et al., 2020). To reduce false positive significant target genes, the Benjamini-Hochberg correction with a false discovery rate of 0.05 was implemented on Z score p-values using statsmodels fdrcorrection() function (Benjamini and Hochberg, 1995). Significant target genes, whose FDR corrected p-values are below 0.05.

1.8 References

Al Balushi, A., Khamis, F., Klaassen, C.H.W., Gangneux, J.-P., Van Hellemond, J.J., Petersen, E., 2018. Double Infection With Leishmania tropica and L. major in an HIV Patient Controlled With High Doses of Amphotericin B. Open Forum Infect. Dis. 5, ofy323. https://doi.org/10.1093/ofid/ofy323

Albert, R., Barabási, A.-L., 2002. Statistical mechanics of complex networks. Rev. Mod. Phys. 74, 47–97. https://doi.org/10.1103/RevModPhys.74.47

Altschul, S.F., Erickson, B.W., 1985. Significance of nucleotide sequence alignments: a method for random sequence permutation that preserves dinucleotide and codon usage. Mol. Biol. Evol. 2, 526–538.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. J. Mol. Biol. 215, 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Benjamini, Y., Hochberg, Y., 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J. R. Stat. Soc. Ser. B Methodol. 57, 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Bensen, E.S., Martin, S.J., Li, M., Berman, J., Davis, D.A., 2004. Transcriptional profiling in *Candida albicans* reveals new adaptive responses to extracellular pH and functions for Rim101p. Mol. Microbiol. 54, 1335–1351. https://doi.org/10.1111/j.1365-2958.2004.04350.x

Berman, B.P., Nibu, Y., Pfeiffer, B.D., Tomancak, P., Celniker, S.E., Levine, M., Rubin, G.M., Eisen, M.B., 2002. Exploiting Transcription Factor Binding Site Clustering to Identify Cis-Regulatory Modules Involved in Pattern Formation in the Drosophila Genome. Proc. Natl. Acad. Sci. - PNAS 99, 757–762. https://doi.org/10.1073/pnas.231608898

Bonhomme, J., Chauvel, M., Goyard, S., Roux, P., Rossignol, T., d'Enfert, C., 2011. Contribution of the glycolytic flux and hypoxia adaptation to efficient biofilm formation by *Candida albicans*. Mol. Microbiol. 80, 995–1013. https://doi.org/10.1111/j.1365-2958.2011.07626.x

Brandes, U., Fleischer, D., 2005. Centrality Measures Based on Current Flow, in: Diekert, V., Durand, B. (Eds.), STACS 2005, Lecture Notes in Computer Science. Springer, Berlin, Heidelberg, pp. 533–544. https://doi.org/10.1007/978-3-540-31856-9_44

Buenrostro, J.D., Wu, B., Chang, H.Y., Greenleaf, W.J., 2015. ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. Curr. Protoc. Mol. Biol. 109, 21.29.1-21.29.9. https://doi.org/10.1002/0471142727.mb2129s109

Camacho, D.M., Collins, K.M., Powers, R.K., Costello, J.C., Collins, J.J., 2018. Next-Generation Machine Learning for Biological Networks. Cell 173, 1581–1592. https://doi.org/10.1016/j.cell.2018.05.015

Chiu, T.-P., Comoglio, F., Zhou, T., Yang, L., Paro, R., Rohs, R., 2016. DNAshapeR: an R/Bioconductor package for DNA shape prediction and feature encoding. Bioinforma. Oxf. Engl. 32, 1211–1213. https://doi.org/10.1093/bioinformatics/btv735

Copping, V.M.S., Barelle, C.J., Hube, B., Gow, N.A.R., Brown, A.J.P., Odds, F.C., 2005. Exposure of *Candida albicans* to antifungal agents affects expression of SAP2 and SAP9 secreted proteinase genes. J. Antimicrob. Chemother. 55, 645–654. https://doi.org/10.1093/jac/dki088

Cuellar-Partida, G., Buske, F.A., McLeay, R.C., Whitington, T., Noble, W.S., Bailey, T.L., 2012. Epigenetic priors for identifying active transcription factor binding sites. Bioinformatics 28, 56–62. https://doi.org/10.1093/bioinformatics/btr614

Dale, R.K., Pedersen, B.S., Quinlan, A.R., 2011. Pybedtools. Bioinformatics 27, 3423–3424. https://doi.org/10.1093/bioinformatics/btr539

Enjalbert, B., Smith, D.A., Cornell, M.J., Alam, I., Nicholls, S., Brown, A.J.P., Quinn, J., 2006. Role of the Hog1 stress-activated protein kinase in the global transcriptional response to stress in the fungal pathogen *Candida albicans*. Mol. Biol. Cell 17, 1018–1032. https://doi.org/10.1091/mbc.e05-06-0501

Fan, J., Chaturvedi, V., Shen, S.-H., 2002. Identification and phylogenetic analysis of a glucose transporter gene family from the human pathogenic yeast *Candida albicans*. J. Mol. Evol. 55, 336–346. https://doi.org/10.1007/s00239-002-2330-4

Fanning, S., Xu, W., Solis, N., Woolford, C.A., Filler, S.G., Mitchell, A.P., 2012. Divergent targets of *Candida albicans* biofilm regulator Bcr1 in vitro and in vivo. Eukaryot. Cell 11, 896–904. https://doi.org/10.1128/EC.00103-12

Fitzpatrick, D.A., O'Gaora, P., Byrne, K.P., Butler, G., 2010. Analysis of gene evolution and metabolic pathways using the *Candida* Gene Order Browser. BMC Genomics 11, 1–14. https://doi.org/10.1186/1471-2164-11-290

Fox, E.P., Bui, C.K., Nett, J.E., Hartooni, N., Mui, M.C., Andes, D.R., Nobile, C.J., Johnson, A.D., 2015. An expanded regulatory network temporally controls *Candida albicans* biofilm formation. Mol. Microbiol. 96, 1226–1239. https://doi.org/10.1111/mmi.13002

Fox, E.P., Nobile, C.J., 2012. A sticky situation: untangling the transcriptional network controlling biofilm development in *Candida albicans*. Transcription 3, 315–322. https://doi.org/10.4161/trns.22281

García-Sánchez, S., Aubert, S., Iraqui, I., Janbon, G., Ghigo, J.-M., d'Enfert, C., 2004. *Candida albicans* Biofilms: a Developmental State Associated With Specific and Stable Gene Expression Patterns. Eukaryot. Cell 3, 536–545. https://doi.org/10.1128/EC.3.2.536-545.2004

García-Sánchez, S., Mavor, A.L., Russell, C.L., Argimon, S., Dennison, P., Enjalbert, B., Brown, A.J.P., 2005. Global roles of Ssn6 in Tup1- and Nrg1-dependent gene regulation in the fungal pathogen, *Candida albicans*. Mol. Biol. Cell 16, 2913–2925. https://doi.org/10.1091/mbc.e05-01-0071

Gretes, M.C., Poole, L.B., Karplus, P.A., 2012. Peroxiredoxins in Parasites. Antioxid. Redox Signal. 17, 608–633. https://doi.org/10.1089/ars.2011.4404

Gulati, M., Nobile, C.J., 2016. *Candida albicans* biofilms: development, regulation, and molecular mechanisms. Microbes Infect. 18, 310–321. https://doi.org/10.1016/j.micinf.2016.01.002

Inukai, S., Kock, K.H., Bulyk, M.L., 2017. Transcription factor–DNA binding: beyond binding site motifs. Curr. Opin. Genet. Dev. 43, 110–119. https://doi.org/10.1016/j.gde.2017.02.007

Kantorovitz, M.R., Kazemian, M., Kinston, S., Miranda-Saavedra, D., Zhu, Q., Robinson, G.E., Göttgens, B., Halfon, M.S., Sinha, S., 2009. Motif-Blind, Genome-Wide Discovery of cis-Regulatory Modules in Drosophila and Mouse. Dev. Cell 17, 568–579. https://doi.org/10.1016/j.devcel.2009.09.002

Karababa, M., Coste, A.T., Rognon, B., Bille, J., Sanglard, D., 2004. Comparison of gene expression profiles of *Candida albicans* azole-resistant clinical isolates and laboratory strains exposed to drugs inducing multidrug transporters. Antimicrob. Agents Chemother. 48, 3064–3079. https://doi.org/10.1128/AAC.48.8.3064-3079.2004

Lin, L., Yee, S.W., Kim, R.B., Giacomini, K.M., 2015. SLC transporters as therapeutic targets: emerging opportunities. Nat. Rev. Drug Discov. 14, 543–560. https://doi.org/10.1038/nrd4626

Liu, P., Honig, B., Mann, R.S., Rohs, R., West, S.M., Sosinsky, A., 2009. The role of DNA shape in protein-DNA recognition. Nat. Lond. 461, 1248–1253. https://doi.org/10.1038/nature08473

Liu, T.T., Lee, R.E.B., Barker, K.S., Lee, R.E., Wei, L., Homayouni, R., Rogers, P.D., 2005. Genome-wide expression profiling of the response to azole, polyene, echinocandin, and pyrimidine antifungal agents in *Candida albicans*. Antimicrob. Agents Chemother. 49, 2226–2236. https://doi.org/10.1128/AAC.49.6.2226-2236.2005

Lohse, M.B., Gulati, M., Johnson, A.D., Nobile, C.J., 2018. Development and regulation of single- and multi-species *Candida albicans* biofilms. Nat. Rev. Microbiol. 16, 19–31. https://doi.org/10.1038/nrmicro.2017.107

Maglott, D., Ostell, J., Pruitt, K.D., Tatusova, T., 2007. Entrez Gene: gene-centered information at NCBI. Nucleic Acids Res. 35, D26-31. https://doi.org/10.1093/nar/gkl993

Maguire, S.L., ÓhÉigeartaigh, S.S., Byrne, K.P., Schröder, M.S., O'Gaora, P., Wolfe, K.H., Butler, G., 2013. Comparative genome analysis and gene finding in *Candida* species using CGOB. Mol. Biol. Evol. 30, 1281–1291. https://doi.org/10.1093/molbev/mst042

Mancera, E., Nocedal, I., Hammel, S., Gulati, M., Mitchell, K.F., Andes, D.R., Nobile, C.J., Butler, G., Johnson, A.D., 2021. Evolution of the complex transcription network controlling biofilm formation in *Candida* species. eLife 10. https://doi.org/10.7554/eLife.64682

Mathelier, A., Wasserman, W.W., 2013. The Next Generation of Transcription Factor Binding Site Prediction. PLoS Comput. Biol. 9, e1003214–e1003214. https://doi.org/10.1371/journal.pcbi.1003214

Mathelier, A., Xin, B., Chiu, T.-P., Yang, L., Rohs, R., Wasserman, W.W., 2016. DNA Shape Features Improve Transcription Factor Binding Site Predictions In Vivo. Cell Syst. 3, 278-286.e4. https://doi.org/10.1016/j.cels.2016.07.001

Mayer, F.L., Wilson, D., Hube, B., 2013. *Candida albicans* pathogenicity mechanisms. Virulence 4, 119–128. https://doi.org/10.4161/viru.22913

McCormack, P.L., Perry, C.M., 2005. Caspofungin. Drugs 65, 2049–2068. https://doi.org/10.2165/00003495-200565140-00009

Nett, J.E., Lepak, A.J., Marchillo, K., Andes, D.R., 2009. Time Course Global Gene Expression Analysis of an In Vivo *Candida* Biofilm. J. Infect. Dis. 200, 307–313. https://doi.org/10.1086/599838

Nobile, C.J., Fox, E.P., Nett, J.E., Sorrells, T.R., Mitrovich, Q.M., Hernday, A.D., Tuch, B.B., Andes, D.R., Johnson, A.D., 2012. A Recently Evolved Transcriptional Network Controls Biofilm Development in *Candida albicans*. Cell 148, 126–138. https://doi.org/10.1016/j.cell.2011.10.048

Nobile, C.J., Johnson, A.D., 2015. *Candida albicans* biofilms and human disease. Annu. Rev. Microbiol. 69, 71–92. https://doi.org/10.1146/annurev-micro-091014-104330

Nobile, C.J., Mitchell, A.P., 2005. Regulation of Cell-Surface Genes and Biofilm Formation by the *C. albicans* Transcription Factor Bcr1p. Curr. Biol. 15, 1150–1155. https://doi.org/10.1016/j.cub.2005.05.047

Ojala, M., Garriga, G.C., 2009. Permutation Tests for Studying Classifier Performance. IEEE, pp. 908–913. https://doi.org/10.1109/ICDM.2009.108

Ondov, B.D., Treangen, T.J., Melsted, P., Mallonee, A.B., Bergman, N.H., Koren, S., Phillippy, A.M., 2016. Mash: fast genome and metagenome distance estimation using MinHash. Genome Biol. Online Ed. 17. https://doi.org/10.1186/s13059-016-0997-x

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, É., 2011. Scikit-learn: Machine Learning in Python. J. Mach. Learn. Res. https://doi.org/10.5555/1953048.2078195

Quinlan, A.R., Hall, I.M., 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26, 841–842. https://doi.org/10.1093/bioinformatics/btq033

Rogers, J.M., Bulyk, M.L., 2018. Diversification of transcription factor–DNA interactions and the evolution of gene regulatory networks. Wiley Interdiscip. Rev. Syst. Biol. Med. 10, e1423-n/a. https://doi.org/10.1002/wsbm.1423

Rybak, J.M., Fortwendel, J.R., Rogers, P.D., 2019. Emerging threat of triazole-resistant Aspergillus fumigatus. J. Antimicrob. Chemother. 74, 835–842. https://doi.org/10.1093/jac/dky517

Schweizer, A., Rupp, S., Taylor, B.N., Röllinghoff, M., Schröppel, K., 2000. The TEA/ATTS transcription factor CaTec1p regulates hyphal development and virulence in *Candida albicans*. Mol. Microbiol. 38, 435–445. https://doi.org/10.1046/j.1365-2958.2000.02132.x

Seneviratne, C.J., Wang, Y., Jin, L., Abiko, Y., Samaranayake, L.P., 2008. *Candida albicans* biofilm formation is associated with increased anti-oxidative capacities. Proteomics 8, 2936–2947. https://doi.org/10.1002/pmic.200701097

Singh, R.P., Prasad, H.K., Sinha, I., Agarwal, N., Natarajan, K., 2011. Cap2-HAP complex is a critical transcriptional regulator that has dual but contrasting roles in regulation of iron homeostasis in *Candida albicans*. J. Biol. Chem. 286, 25154–25170. https://doi.org/10.1074/jbc.M111.233569

Sinha, S., van Nimwegen, E., Siggia, E.D., 2003. A probabilistic method to detect regulatory modules. Bioinformatics 19, i292–i301. https://doi.org/10.1093/bioinformatics/btg1040

Skene, P.J., Henikoff, S., 2017. An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites. eLife 6. https://doi.org/10.7554/eLife.21856

Skrzypek, M.S., Binkley, J., Binkley, G., Miyasato, S.R., Simison, M., Sherlock, G., 2017. The *Candida* Genome Database (CGD): incorporation of Assembly 22, systematic identifiers and visualization of high throughput sequencing data. Nucleic Acids Res. 45, D592–D596. https://doi.org/10.1093/nar/gkw924

Slavic, K., Krishna, S., Derbyshire, E.T., Staines, H.M., 2011. Plasmodial sugar transporters as anti-malarial drug targets and comparisons with other protozoa. Malar. J. 10, 165. https://doi.org/10.1186/1475-2875-10-165

Soll, D.R., Daniels, K.J., 2016. Plasticity of *Candida albicans* Biofilms. Microbiol. Mol. Biol. Rev. 80, 565–595. https://doi.org/10.1128/MMBR.00068-15

Stormo, G.D., 2013. Modeling the specificity of protein-DNA interactions. Quant. Biol. 1, 115–130. https://doi.org/10.1007/s40484-013-0012-4

Thurmond, J., Goodman, J.L., Strelets, V.B., Attrill, H., Gramates, L.S., Marygold, S.J., Matthews, B.B., Millburn, G., Antonazzo, G., Trovisco, V., Kaufman, T.C., Calvi, B.R., 2019. FlyBase 2.0: the next generation. Nucleic Acids Res. 47, D759–D765. https://doi.org/10.1093/nar/gky1003

Uppuluri, P., Chaturvedi, A.K., Srinivasan, A., Banerjee, M., Ramasubramaniam, A.K., Köhler, J.R., Kadosh, D., Lopez-Ribot, J.L., 2010. Dispersion as an Important Step in the *Candida albicans* Biofilm Developmental Cycle. PLoS Pathog. 6, e1000828–e1000828. https://doi.org/10.1371/journal.ppat.1000828

Van Helden, J., 2004. Metrics for comparing regulatory sequences on the basis of pattern counts. Bioinformatics 20, 399–406. https://doi.org/10.1093/bioinformatics/btg425

Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S.J., Brett, M., Wilson, J., Millman, K.J., Mayorov, N., Nelson, A.R.J., Jones, E., Kern, R., Larson, E., Carey, C.J., Polat, İ., Feng, Y., Moore, E.W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E.A., Harris, C.R., Archibald, A.M., Ribeiro, A.H., Pedregosa, F., van Mulbregt, P., 2020. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat. Methods 17, 261–272. https://doi.org/10.1038/s41592-019-0686-2

Wardle, F.C., Tan, H., 2015. A ChIP on the shoulder? Chromatin immunoprecipitation and validation strategies for ChIP antibodies [version 1; peer review: 2 approved]. F1000 Res. 4, 235–235. https://doi.org/10.12688/f1000research.6719.1

Yeater, K.M., Chandra, J., Cheng, G., Mukherjee, P.K., Zhao, X., Rodriguez-Zas, S.L., Kwast, K.E., Ghannoum, M.A., Hoyer, L.L., 2007. Temporal analysis of *Candida albicans* gene expression during biofilm development. Microbiol. Soc. Gen. Microbiol. 153, 2373–2385. https://doi.org/10.1099/mic.0.2007/006163-0

Zhou, T., Shen, N., Yang, L., Abe, N., Horton, J., Mann, R.S., Bussemaker, H.J., Gordân, R., Rohs, R., 2015. Quantitative modeling of transcription factor binding specificities using DNA shape. Proc. Natl. Acad. Sci. - PNAS, From the Cover 112, 4654–4659. https://doi.org/10.1073/pnas.1422023112

Chapter 2

A CUT&RUN method and data analysis workflow for genome-wide analysis of transcription factor-DNA interactions in *Candida albicans*

2.1 Abstract

Regulatory transcription factors control many important biological processes including cellular differentiation, responses to environmental perturbations and stresses, and host-pathogen interactions. Determining the genome-wide binding of regulatory transcription factors to DNA is essential to understanding the function of transcription factors in these often complex biological processes. Cleavage Under Targets and Release Using Nuclease (CUT&RUN) is a modern method for genome-wide mapping of in vivo protein-DNA binding interactions that is an attractive alternative to the traditional and widely used chromatin immunoprecipitation followed by sequencing (ChIP-seq) method. CUT&RUN is amenable to a higher throughput experimental setup and has a substantially higher dynamic range with lower per-sample sequencing costs compared to ChIP-seq. Here, we describe a comprehensive CUT&RUN protocol and accompanying data analysis workflow that is tailored for genome-wide analysis of transcription factor-DNA binding interactions in the human fungal pathogen *Candida albicans*. This detailed protocol describes all necessary experimental procedures, from epitope tagging of transcription factor coding genes, to library preparation for sequencing; additionally, it includes our customized computational workflow for CUT&RUN data analysis.

2.2 Introduction

Candida albicans is a clinically relevant polymorphic human fungal pathogen that exists in a variety of different modes of growth, such as the planktonic (free-floating) form and as communities of tightly adhered cells protected by an extracellular matrix, known as the biofilm form (Nobile et al., 2012). Like other developmental and cellular processes, biofilm development is an important *C. albicans* virulence trait that is known to be controlled at the transcriptional level by regulatory transcription factors (TFs) that bind to DNA in a sequence-specific manner (Nobile et al., 2012). To understand the complex biology of this important fungal pathogen, effective methods to determine the genome-wide localization of specific TFs during distinct developmental and cellular processes is increasingly valuable.

Chromatin immunoprecipitation followed by sequencing (ChIP-seq) has been widely used to investigate protein-DNA interactions in *C. albicans* and has largely replaced the more traditional chromatin immunoprecipitation followed by microarray (ChIP-chip) approach. Both ChIP-seq and ChIP-chip, however, require a large number of input cells, which can be a complicating factor when investigating TFs in the context of specific growth forms, such as biofilms. In addition, the chromatin immunoprecipitation (ChIP) assay often yields a significant amount of background signal throughout the genome, requiring a high level of enrichment for the target of interest to sufficiently separate signal from noise. While the ChIP-chip assay is largely out of date today, the sequencing depths necessary for ChIP-seq has made this assay prohibitively expensive for many researchers, particularly those studying multiple TFs and chromatin-associated proteins.

Cleavage Under Targets and Release Using Nuclease (CUT&RUN) is an attractive alternative to ChIP-seq. It was developed by the Henikoff lab in 2017 to circumvent the limitations of ChIP-seq, while providing high resolution genome-wide mapping of TFs and chromatin-associated proteins (Henikoff). CUT&RUN relies on the targeted digestion of chromatin within permeabilized nuclei using tethered micrococcal nucleases, which is followed by sequencing of the digested DNA fragments. Since DNA fragments are specifically generated at the loci that are bound by a protein of interest, rather than being generated throughout the genome via random fragmentation as in ChIP assays, the CUT&RUN approach results in greatly reduced background signal, and thus requires one tenth of the sequencing depth compared to ChIP-seq. These improvements ultimately lead to significantly lower sequencing costs as well as lower numbers of input cells.

Here we describe a robust CUT&RUN protocol that has been adapted and optimized for determining the genome-wide localization of TFs in *C. albicans* cells isolated from biofilms and planktonic cultures. We also present a thorough data analysis pipeline, enabling the processing and analysis of the resulting sequence data, that requires users to have minimal expertise in coding or bioinformatics. Briefly, we begin with our epitope tagging procedure for TF coding genes, describe the harvesting of biofilm and planktonic cells, and describe the isolation of intact permeabilized nuclei from the cells. We describe how the nuclei are incubated with primary antibodies against the specific protein or epitope tagged protein of interest, the tethering of the chimeric A/G-micrococcal nuclease (pAG-MNase) fusion proteins to the primary antibodies, genomic DNA recovery after chromatin digestion, and preparation of the genomic DNA library for sequencing. We also describe our data analysis pipeline, which takes raw DNA sequencing reads in FASTQ format and implements all required processing steps to provide a complete list of significantly enriched loci that are bound by the TF of interest (targeted by the primary antibody). We note that multiple steps of our library preparation protocol have been specifically adapted and optimized for CUT&RUN analysis of TFs (as opposed to nucleosomes). While the data presented in this manuscript were generated using our TF-specific adaptations of a commercially available CUT&RUN kit (the EpiCypher CUTANA ChIC/CUT&RUN kit, Catalog # 14-1048), we have also validated our protocols using individually sourced components (i.e., pAG-MNase enzyme and magnetic DNA purification beads) as well as in-house prepared buffers, which significantly reduce experimental cost. Our comprehensive experimental and data analysis protocols are described in detail below in a step-by-step format.

2.3 Protocol

1. Epitope Tagging of *C. albicans* Strains

> 1.1. Upload the gene of interest, along with the 1 kb upstream and downstream flanking sequences, from the *Candida* Genome Database (http://www.candidagenome.org/) to Benchling.com.

1.2. Design a gRNA by highlighting 50 bp upstream and downstream from the stop codon and click the gRNA selection tool on the right. Select “Design and Analyze Guides”. Use the Ca22 (*Candida albicans* SC5314 (diploid)) genome and an NGG (SpCas9, 3' side) PAM for the guide parameters. Click finish.

1.2.1. On the subsequent page, confirm the target region for gRNA design and press the green button.

1.2.2. Sort gRNAs by the "On-Target Score".

NOTE: Benchling.com computes on-target and off-target scores to quantify the specificity of the gRNAs. An ideal guide has an on-target score of >60, an off-target score of approximately 33, and overlaps the stop codon. This enables high gRNA specificity, while ablating gRNA targeting after GFP integration. gRNAs with an off-target score of approximately 50 indicates allelic variation, and thus only one allele will be recognized by the gRNA.

1.2.3. Add these sequences to the 5' (CGTAAACTATTTTTAATTTG) and 3' (GTTTTAGAGCTAGAAATAGC) ends of the 20 bp gRNA, creating a 60 bp primer/oligonucleotide. Alternatively, copy the 20 bp sequence to the gRNA calculator supplied by Nguyen et al.

1.2.4. Order the 60 bp custom gRNA oligonucleotide.

1.2.5. Amplify the "universal A fragment" with 100 µM AHO1096 (GACGGCACGGCCACGCGTTTAAACCGCC) and 100 µM AHO1098 (CAAATTAAAAATAGTTTACGCAAG) and the "unique B fragment" with the custom 60 bp gRNA oligonucleotide (100 µM) from step 1.2.4 and 100 µM AHO1097 (CCCGCCAGGCGCTGGGGTTTAAACACCG) using pADH110 (Addgene ID# 90982) and pADH139 (Addgene ID# 90987), respectively, as template DNA. Use the provided cycling conditions and PCR reaction mixes.

	A Fragment	B Fragment
dH_2O	75.5 µL	75.5 µL
5X HF Buffer	20 µL	20 µL
10 mM dNTP	2 µL	2 µL
FWD Primer (100 µM)	0.5 µL (AHO1096)	0.5 µL unique gRNA
REV Primer (100 µM)	0.5 µL (AHO1098)	0.5 µL (AHO1097)
DNA (1 ng/µL)	1 µL pADH110	1 µL pADH139
Phusion Polymerase	0.5 µL	0.5 µL

Temp (°C)	Time	Cycles
98	30 s	1
98	20 s	30
58	20 s	
72	30 s	
72	15 s	1
8	hold	1

Total Reaction Volume	100 µL	100 µL

1.2.6. Confirm successful amplification by checking 5 µL of PCR on a 1% agarose gel. The A and B fragment amplicons are approximately 1 kb each.

1.2.7. Mix 1 µL each of A and B fragments together and stitch together with the PCR reaction mix and cycling conditions to create the full-length C fragment.

C Fragment PCR Reaction Mix	
dH$_2$O	74.5 µL
5x HF Buffer	20 µL
10 mM dNTPs	2 µL
Universal A	1 µL
Unique B	1 µL
Phusion Polymerase	0.5 µL

Cycling Condition 1		
Temp (°C)	Time	Cycles
98	30 s	1
98	10 s	5
58	20 s	
72	60 s	

1.2.8. Add 0.5 µL of 100 mM AHO1237 (AGGTGATGCTGAAGCTATTGAAG) and 0.5 µL of 100 mM AHO1453 (ATTTTAGTAACAGCTTCGACAATCG) to each PCR reaction, mix well by pipetting, and complete the following cycling conditions.

Cycling Condition 2		
Temp (°C)	Time	Cycles
98	30 s	1
98	10 s	30
66	20 s	
72	60 s	
72	30 s	1
8	hold	1

1.2.9. Confirm proper stitching and amplification of the C fragment by checking 5 µL of the amplicon on a 1% agarose gel. The expected fragment size is 2 kb.

NOTE: If stitching and amplification results in multiple non-specific bands or smearing, perform a PCR cleanup of the A and B fragments and repeat from step 1.2.7.

1.2.10. Store C fragment at -20 °C until use.

1.3. Add the entire CTG optimized monomeric eGFP with linker sequence (RIPLING) published in Ennis et al. (pCE1, Addgene ID# 174434) immediately upstream of the stop codon in your gene of interest on Benchling.com, creating a C-terminal translational fusion. This construct will be used to design oligonucleotides for amplifying donor DNA (dDNA) from pCE1.

1.3.1. Design a forward oligonucleotide with 18-22 bp homology to the linker sequence and > 50 bp homology to the 3' end of the open reading frame.

NOTE: The 18-22 bp homology creates an annealing Tm for amplification between 55-58 °C. If the full-length oligonucleotide forms primer dimers, adjust homology to the linker sequence/GFP or the genome accordingly.

1.3.2. Create a reverse oligonucleotide with 18-22 bp homology to the 3' end of GFP and > 50 bp homology to the downstream non-coding sequence.

1.3.3. Order these oligonucleotides and amplify the dDNA with touch down PCR cycling protocol.

GFP Amplification PCR Reaction	
dH_2O	37.25 µL
5x HF Buffer	10 µL
10 mM dNTPs	1 µL
pCE1 (1 ng/µL)	1 µL
Forward Primer (100 µM)	0.25 µL
Reverse Primer (100 µM)	0.25 µL
Phusion Polymerase	0.25 µL

Temp (°C)	Time	Cycles	
98	30 s	1	
98	10 s	10	Δ-1 C / cycle
65	20 s		
72	30 s		
98	10 s	25	
57	20 s		
72	30 s		
72	15 s	1	
8	hold	1	

1.4. Design two sets of colony PCR (cPCR) oligonucleotides for confirming integration of GFP by amplifying across the flanking integration sites. First select the forward dDNA oligo on Benchling.com and click the Primer button on the right.

1.4.1. Click “Create Primers” and then “Wizard”. Click “Tm Param” and confirm algorithm is set to “SantaLucia 1998”. Click “Use Selection” to input the coordinates of the forward oligonucleotide designed in 1.3.1 as the target sequence.

1.4.2. Set the optimal primer Tm to 55 °C and the max amplicon size to 900 bp and then click the blue “Generate Primers” button at the top right.

1.4.3. Select the oligonucleotide pair with the lowest penalty score and confirm that the primers amplify across the 5’ integration site. The forward cPCR primer should lie upstream of the forward dDNA primer sequence, and the reverse cPCR primer should lie fully within the eGFP tag or the linker sequence.

1.4.4. Repeat steps 1.4-1.4.3 with the reverse dDNA oligonucleotide to create the second set of cPCR oligonucleotides that amplify across the 3’ integration site. The forward cPCR primer should lie fully within the eGFP tag or the linker sequence, and the reverse cPCR primer should lie downstream of the reverse dDNA primer sequence.

1.4.5. Order these oligonucleotides.

1.5. Digest 2,500 ng of pADH140, which contains Cas9 (Addgene ID# 90988), with Fast Digest MssI for each gene to be GFP-tagged. Each digestion reaction should be 15 µL total, adjust volume of water added accordingly based on the pADH140 plasmid concentration. Store digested plasmid at -20 °C until use.

dH_2O	Variable
pADH140	Variable
FastDigest Buffer	1.5 µL
MssI	0.8 µL
Total Volume	15 µL

Temp (°C)	Time
37	1 h
65	15 min
8	hold

1.6. Denature 12 µL 10 mg/mL salmon sperm DNA for each gene that will be GFP-tagged at 99 °C for 10 min and rapidly cool to ≤ 4 °C. Store at -20 °C until use

1.7. Streak a *C. albicans* LEU2 hemizygous nourseothricin-sensitive strain onto YPD plates and incubate at 30 °C for two days.

1.8. Select a single colony and transfer to 4 mL liquid YPD. Incubate 12-16 h at 30 °C with shaking at 250 rpm.

1.9. Measure the optical density at 600nm wavelength (OD600) of the overnight (12-16 h) cell culture using a disposable cuvette (1 mL, 1 cm path length).

1.10. Dilute overnight culture into an Erlenmeyer flask to an OD600 of 0.1 in YPD. The volume depends on the number of transformation reactions. Account for 5 mL per reaction and include an additional 5 mL for checking the OD600 later.

1.11. Incubate the diluted overnight culture in a shaking incubator at 30 °C with shaking at 250 rpm until it reaches an OD600 of 0.5-0.8.

1.12. Centrifuge at 4000 x g for 5 min then remove and discard the supernatant.

1.13. Resuspend cell pellet in 1 mL sterile water via gentle pipette mixing with filter tips and transfer to a sterile 1.5 mL microfuge tube.

1.14. Pellet the cells by centrifuging at 4000 x g for 1 min then remove and discard the supernatant. Resuspend in 1 mL sterile water and repeat for a total of two washes.

1.15. Resuspend the pellet in 1/100th of the volume used in step 1.10. For example, if step 1.10 used 15 mL, resuspend pellet in 150 µL sterile water.

1.16. In a separate tube for each transformation reaction, mix 50 µL C fragment, 50 µL dDNA, 2500 ng MssI digested pADH140, and 10 µL denatured salmon sperm DNA.

1.17. Add 50 µL of the cell slurry from step 1.15 and mix by pipetting.

1.18. Make a stock of PLATE mix for n + 1 transformations.

1.19. Add 1 mL PLATE mix to the cell/DNA mixture and mix by inverting 3-5 times.

NOTE: Tap the bottom of the tubes while inverted to dislodge any liquid that remains trapped at the bottom of the tube.

1.20. Place in an incubator at 30 °C overnight (12-16 h) without shaking.

1.21. Heat shock the cells for 15 min at 44 °C in a water bath.

1.22. Centrifuge the 1.5 mL tubes at 5000 x g for 2 min.

1.23. Remove the PLATE mix by vacuum aspiration with sterile pipette tips, being careful to avoid disturbing the cell pellet.

1.24. Resuspend the cells in 1 mL YPD, pellet by centrifugation at 4000 x g for 1 min, then remove and discard the supernatant. Repeat for a second wash, then resuspend the cells in 1 mL YPD and transfer to a 10 mL round bottom disposable culture tube containing an additional 1 mL of YPD (2 mL final volume). Recover cells at 30 °C with shaking at 250 rpm for 5 h.

1.25. Centrifuge tubes at 4000 x g for 5 min, remove and discard supernatant.

1.26. Resuspend the cell pellet in 100 µL sterile water and plate on YPD supplemented with 200 µg/mL nourseothricin (NAT200). Incubate at 30 °C for 2-3 days.

1.27. Aliquot 100 µL of 20 mM NaOH into the wells of a 96 well PCR plate, with each well corresponding to an individual colony that grew on the NAT200 plates. Using a sterile toothpick or pipette tip, pick individual transformed colonies and patch them onto a new NAT200 plate and swirl remaining cells into a well with 20 mM NaOH. Repeat for the remaining colonies. This creates the cell lysate used as the DNA template for PCR amplification.

1.28. Seal the PCR plate and incubate for 10 min at 99 °C in a thermal cycler with heated lid.

1.29. Make two cPCR reaction mixes with the oligonucleotides designed in steps 1.4-1.4.5. Scale up the number of reactions as needed. Perform the PCR reaction with the cell lysate made in steps 1.27 and 1.28 and run 20 µL from each well on a 1% agarose gel. Colonies with amplification of the two cPCR primer sets properly incorporated the GFP dDNA.

cPCR Reaction	
dH_2O	11.66 µL
DreamTaq Green Buffer	2.2 µL
5M Betaine	4.4 µL
$MgCl_2$	0.44 µL
10 mM dNTP	0.44 µL
DreamTaq	0.22 µL
Forward Primer (100 µM)	0.22 µL
Reverse Primer (100 µM)	0.22 µL
Lysate	2.2 µL

Temp (°C)	Time	Cycles
94	30 s	1
94	10 s	35
55	30 s	
72	1 min	
72	15 s	1
8	hold	1

1.30. Re-streak colonies that incorporated GFP on SC media lacking leucine. Incubate in a 30 °C incubator for 2-3 days. Pick individual colonies and patch onto YPD and YPD supplemented with 400 µg/mL nourseothricin (NAT400) plates. Colonies that fail to grow on NAT400 plates after 24 h successfully lost the CRISPR components.

1.31. Confirm that the GFP-tag is retained by repeating step 1.29 from the YPD patch plate. If the correct bands are present, inoculate 4 mL YPD and grow overnight (12-16 h) as described in step 1.8.

1.32. Mix the overnight culture of the new GFP-tagged strain from 1.31 with filter-sterilized 50% glycerol in a 1:1 ratio in a sterile cryotube. Store at -80 °C and re-streak on YPD plates as needed.

NOTE: We highly recommend validating the GFP-tagged strains by confirming nuclear localization of the tagged TF via fluorescent microscopy and confirming a wild-type phenotype in an appropriate phenotypic assay. For example, if studying biofilm regulators confirm that the strain forms wild-type biofilms using the phenotypic assay described in steps 2.1-2.7.

2. Sample Preparation using Biofilm or Planktonic Cultures

NOTE: For biofilms, follow steps 2.1-2.7. For planktonic cells follow steps 2.1-2.3 and 2.8-2.9.

2.1 Streak *C. albicans* eGFP-tagged strain(s) onto YPD agar plates and incubate at 30 °C for 2-3 days.

2.2 Using a single isolated colony from the agar plate, inoculate into 4 mL of YPD liquid medium. Incubate at 30 °C with shaking overnight (12-16 h).

2.3 Determine the OD600 of the overnight culture(s).

2.4 Inoculate a sterile 12-well untreated cell culture plate with the overnight cell culture to a final OD600 of 0.5 (equivalent to 2 x 107 cells/mL) in RPMI-1640 medium to a final volume of 2 mL. Incubate for 90 min at 37 °C in an ELMI microplate incubator with shaking at 250 rpm.

NOTE: We recommend using one 12-well cell culture plate per strain with one well uninoculated as a control for medium contamination.

2.5 Remove unadhered cells by aspiration using sterile pipette tips attached via flexible plastic tubing to a vacuum trap apparatus. Wash the adhered cells once with 2 mL sterile 1x PBS solution. Add 2 mL fresh RPMI-1640 medium to wells and incubate for 24 h at 37 °C with shaking at 250 rpm.

NOTE: Change pipette tips between wells of different strains and/or conditions. Take care not to scrape the bottom of the well with the tip while aspirating.

2.6 At the end of the 24 h incubation, collect and pool the liquid and biofilm material from each of the 11 inoculated wells from step 2.4 into a single sterile 50 mL conical tube. Repeat as necessary with independent pools if processing more than one strain or growth condition concurrently.

NOTE: Scrape the bottoms and edges of each well with a pipette filter tip to dislodge cells that remain adhered to the surface. Use the pipette to homogenize the biofilms.

2.7 Pellet samples by centrifuging at 4000 x g for 5 mins. Decant as much of the supernatant as possible, taking care to minimize disruption of the pellet. Snap-freeze pellet in liquid nitrogen and store at -80 °C immediately after collection or continue directly to step 3 (Isolation of Nuclei).

2.8 For planktonic cultures, back dilute to OD600 of 0.1 in 50 mL volume of RPMI-1640 liquid media and incubate at 30 °C shaking at 225 rpm for 2-5 h until OD600 is between 0.5 to 0.8. Cells should go through at least two doublings before being harvested.

NOTE: Conditions used for planktonic cultures can be adjusted as needed.

2.9 Pellet samples by centrifuging at 4000 x g for 5 mins. Decant as much of the supernatant as possible, taking care to minimize disruption of the pellet. Snap-freeze pellet in liquid nitrogen and store at -80 °C immediately after collection or continue directly to step 3 (Isolation of Nuclei).

3. Isolation of Nuclei

NOTE: Prepare Resuspension Buffer, Ficoll Buffer, and SPC Buffer-PI fresh on the day of the experiment. To resuspend pellets, gently pipette mix cell or nuclei pellets (using either 200 µL or 1 mL pipette tips) to avoid damaging cells or nuclei. Turn on heat block and pre-heat block to 30 °C before beginning the nuclei isolation. All pipette tips and tubes for the remainder of this protocol should be certified DNA/RNA and DNase/RNase-free, and we recommend the use of filter-tips for all pipetting steps.

3.1. Resuspend pellet(s) in 1 mL of room temperature Resuspension Buffer-PI and transfer to a sterile 1.5 mL microfuge tube.

3.2. Pellet at 2000 x g for 2 min in a table-top centrifuge and remove supernatant.

NOTE: Remove supernatant using either 200 µL or 1 mL pipette tips, taking care to minimize disruption of the pellets.

3.3. Resuspend pellet(s) in 200 µL room temperature Resuspension Buffer-PI.

3.4. From the resuspended pellet, transfer a 5 µL aliquot into a new PCR tube and store at 4 °C for later use.

NOTE: This will used as a control during a quality control step to evaluate the quality of the isolated nuclei.

3.5. Pellet at 2000 x g for 2 min and remove and discard the supernatant using a pipette. Repeat wash step for a total of two washes with 200 µL Resuspension Buffer-PI.

3.6. Centrifuge at 2000 x g for 2 min and remove supernatant. Add 300 µL of Resuspension Buffer-PI & 10 µL of Zymolyase solution (50mg/ml). Incubate for 30 min in 30 °C heat block.

NOTE: Alternatively, a water bath heated to 30 °C can also be used instead of a heat block. During the 30 min incubation step, complete step 4 (Concanavalin A Bead Activation) to save time.

CRITICAL STEP: After the 30 min incubation, transfer 5 µL aliquot into a new PCR tube. Stain the 5 µL of isolated nuclei and the 5 µL aliquot of intact cells stored at 4 °C at step 3.4 using calcofluor white (a fluorescent cell wall staining dye) and a nucleic acid stain per the manufacturer's instruction. Visually inspect the integrity and purity of the isolated nuclei using a fluorescence microscope. Intact control cells should show prominent cell wall staining by the calcofluor dye, while the isolated nuclei will show prominently stained intact nuclei without cell walls.

3.7. Centrifuge at 2000 x g for 5 min at 4 °C and then remove supernatant.

NOTE: Keep samples and buffers on ice from this point forward.

3.8. Resuspend pellet in 500 µL ice-cold Resuspension Buffer-PI using 1 mL filter tip by pipetting gently up and down 5 times. Centrifuge at 2000 x g for 5 min at 4 °C then remove supernatant using a 1 mL pipette. Resuspend the pellet with 1 mL freshly made ice-cold Ficoll Buffer.

3.9. Centrifuge samples at 5000 x g for 10 min at 4 °C and then remove supernatant. Resuspend the pellet in 500 µL ice-cold SPC-PI Buffer.

NOTE: From this point onward, handle the nuclei extra gently to avoid damaging them.

3.10. Centrifuge samples at 5000 x g for 10 min at 4 °C and remove as much of the supernatant as possible without disrupting the pellet. Place the tubes containing pelleted nuclei on ice and proceed to section 4 or snap-freeze pellet in liquid nitrogen and store at -80 °C immediately after collection.

NOTE: We recommend proceeding to section 4 immediately, if possible. Avoid multiple freeze-thawing of isolated nuclei as it is known to increase DNA damage leading to poor quality results.

4. Concanavalin A Bead Activation

CRITICAL STEP: From this point forward, users have the choice to continue with the protocol using a commercially available CUT&RUN Kit or to source key components individually and prepare buffers in-house. If using the kit, all buffers and reagents used below are included in the kit unless otherwise noted. If sourcing components independently, we also provide individual catalog numbers for all reagents used in Table 1.

NOTE: Chill all buffers on ice before use.

4.1. Gently resuspend the ConA beads using a pipette. Transfer 22 µL of ConA bead suspension per sample to be processed in a single 1.5 mL microfuge tube.

NOTE: When performing CUT&RUN for a total of 10 samples, for example, transfer 220 µL of ConA bead suspension to a 1.5 mL microfuge tube.

4.2. Place the tube on a magnetic rack until the bead slurry is clear, remove and discard the supernatant using a pipette.

4.3. Remove the tube containing the ConA beads from the magnetic rack and immediately add 200 µL ice-cold Bead Activation Buffer and gently mix using a pipette. Place the tube on the magnetic rack until the bead slurry is clear, remove and discard the supernatant using a pipette. Repeat this step for a total of two washes.

4.4. Resuspend beads in 22 µL ice-cold Bead Activation Buffer per sample of nuclei to be processed. Keep beads on ice until needed.

5. Binding Nuclei to Activated Beads

NOTE: Chill all Buffers on ice before use. All Buffers supplemented with protease inhibitors should be prepared on the day of the experiment. We recommend using strip tubes to facilitate handling of the 0.2 mL tubes in the subsequent steps.

5.1. Resuspend the pelleted nuclei from step 3 in 100 µL of ice-cold SPC-PI Buffer and transfer to a new 8-tube 0.2 mL strip. Add 20 µL of the activated beads to each sample and gently pipette mix. Incubate at room temperature (RT) for 10 min without any agitation.

5.2. Place tubes on magnetic rack until slurry is clear, remove and discard supernatant using a pipette. Remove tubes from the magnetic rack and add 200 µL ice-cold Wash Buffer to each sample. Resuspend beads by gently pipetting up and down several times. Transfer 100 µL aliquots from each sample into a new 8-tube 0.2 mL strip.

CRITICAL STEP: Each CUT&RUN sample is divided into two separate aliquots. One of the aliquots is used for negative control antibody (e.g., IgG negative control antibody) and the other is used for target antibody against protein of interest (e.g., anti-GFP antibody). Both datasets are required for the computational pipeline to accurately identify enrichment signals that are specific to the TF of interest.

NOTE: We have performed a control CUT&RUN experiment using anti-GFP antibodies with an untagged strain and found the results to be comparable to the use of IgG antibodies in a GFP-tagged strain; therefore, we recommend using the standard IgG control for all experiments.

6. Primary Antibody Binding

NOTE: pAG-MNase fusion protein binds well to rabbit, goat, donkey, guinea pig and mouse IgG antibodies. Generally, most commercial ChIP-seq certified commercial antibodies are compatible with CUT&RUN. The amount of primary antibody used depends on the efficiency of the antibody and a titration of the antibody (e.g., 1:50, 1:100, 1:200, and 1:400 final dilution) may be necessary if antibody of interest has not been previously tested in ChIP or CUT&RUN experiments. Chill all buffers on ice before use. All buffers used for antibody binding steps should be prepared on the day of the experiment.

6.1. Place tubes on magnetic rack and wait until slurry is completely clear, remove and discard supernatant using a pipette. Add 50 µL Antibody Buffer and gently pipette mix.

6.2. Add 3 µL anti-GFP polyclonal antibody (or 0.5 ug if using untested antibodies).

NOTE: While some CUT&RUN protocols report increased yield by adding a secondary antibody prior to pAG-MNase addition, we did not observe a significant improvement in our hands and thus do not include a secondary antibody.

6.3. Incubate tubes on nutator at 4 °C for 2 h.

6.4. Quickly centrifuge the tubes at 100 x g, place the tubes on a magnetic rack, and once the slurry is clear remove and discard the supernatant using a pipette.

6.5. While the tubes containing the beads are still on the magnetic rack, add 200 μL ice-cold Cell Permeabilization Buffer directly onto the beads. Remove and discard supernatant using a pipette. Repeat for a total of two washes with the ice-cold Cell Permeabilization Buffer.

6.6. Add 50 μL ice-cold Cell Permeabilization Buffer to each tube and gently pipette mix.

NOTE: Beads are often clumpy at this point but can easily be dispersed by gently mixing using a 200 μL pipette.

7. Binding of pAG-MNase to Antibody

7.1. Add 2.5 μL CUTANA pAG-MNase (20X stock) to each sample, and gently pipette mix. Incubate samples (slightly elevated at ~ 45-degree angle) on a nutator at 4 °C for 1 h.

7.2. Quickly centrifuge the strip tubes at 100 x g, place the tubes on a magnetic rack, and once the slurry is clear, remove and discard the supernatant using a pipette.

NOTE: This step is critical. Carry-over antibody remaining in cap or sides of the tubes after this step will significantly increase the amount of background signal.

7.3. While the tubes containing the beads are still on the magnetic rack, add 200 μL ice-cold Cell Permeabilization Buffer, allow the slurry to clear, and remove and discard the supernatant using a pipette. Repeat step for a total of two washes with the Cell Permeabilization Buffer.

7.4. Add 100 μL ice-cold Cell Permeabilization Buffer to samples and gently pipet up and down.

8. Targeted Chromatin Digestion and Release

8.1. Incubate the tubes containing the sample(s) in a wet ice bath for 5 min. Using a multichannel pipet, add 3 μL of 100 mM CaCl2 into each sample. Gently pipet up and down and immediately return the tubes to the wet ice bath, incubate for 30 min.

8.2. Add 66 μL Stop Buffer to each sample, and gently vortex to mix. Incubate samples for 10 min at 37 °C in a dry bath.

NOTE: We recommend adding 1.5 pg of heterologous *E. coli* spike-in DNA per sample in the 2X Stop Buffer. Addition of 1.5 pg of *E. coli* spike-in DNA results in 1,000-10,000 mapped spike-in reads for 1-10 million mapped experimental reads. The spike-in DNA is used to calibrate for sequencing depth and is especially important for comparing samples in a series. Addition of spike-in *E. coli* is highly recommended but not essential. The commercial CUT&RUN kit includes *E. coli* spike-in DNA, but it can also be purchased separately.

8.3. Place the tubes on the magnetic rack and transfer 160 μL of the supernatant into a 1.5 mL microfuge tube. Transfer 80 μL of sample into a new 2 mL microfuge tube and store at -20 °C in the event the experiment fails, and you need a backup sample. Proceed to step 9 with the remaining 80 µL collected sample.

9. Cleanup of Collected DNA Sample

NOTE: Incubate DNA Purification Beads at room temperature for at least 30 min before use. Pre-chill 100% isopropanol on ice before moving forward. When pipette mixing the samples, pipette up and down at least 10 times.

9.1. Vortex DNA Purification Beads to homogenize the bead solution. Add 50 µL (~0.6X sample volume) resuspended beads to each sample. Pipette mix and incubate samples on a nutator for at least 5 min at room temperature.

NOTE: The ratio of DNA purification beads to sample used is critical. Using 0.6X volume of DNA Purification Bead solution relative to the sample allows the magnetic beads to bind to large DNA fragments released from damaged nuclei. CUT&RUN enriched DNA fragments are much smaller compared to these large DNA fragments and are thus retained in the supernatant in step 9.1.

9.2. Place the tubes on a magnetic rack and transfer 130 µL of the supernatant containing your DNA to a Simport 0.2 mL 8-tube strip. Add an additional 30 µL (Total volume of 160 µL) of DNA Purification Beads to the sample(s).

9.3. Add 170 µL (~1X sample volume) of ice-cold 100% isopropanol, mix well by pipetting up and down at least 10 times and incubate on ice for 10 min.

NOTE: It is critical that 100% ice-cold isopropanol is used for this step for the DNA purification beads to efficiently capture the CUT&RUN enriched small fragments.

9.4. Place the tubes on the magnetic rack and once the slurry has cleared, carefully remove, and discard the supernatant using a pipette.

9.5. While the tubes are on the magnetic rack, add 200 µL of freshly prepared room temperature 80% ethanol to the tubes and incubate at room temperature for 30 s. Carefully remove and discard the supernatant using a pipette. Repeat step for a total of two washes with 80% ethanol.

9.6. Spin the tubes briefly at 100 x g then place the tubes back on the magnetic rack and remove any leftover ethanol using pipette after the slurry has cleared. Air dry the beads for up to 5 min while the tubes remain on the magnetic rack with the lid open.

NOTE: Do not exceed 5 min drying time as it can significantly reduce the final DNA yield.

9.7. Remove the tubes from the magnetic rack and elute the DNA from the beads by adding 17 µL of 0.1X TE pH8. Mix well then incubate for at least 5 min at room temperature.

9.8. Place the tubes on the magnetic rack until slurry becomes clear. Once the slurry has cleared, carefully transfer 15 µL of the supernatant to a sterile 0.2 mL PCR tube.

9.9. Measure the concentration of the collected DNA using a Qubit fluorometer following the manufacturer's protocol.

NOTE: Typically, the concentration of the collected DNA is ~1ng per µL. Sometimes, the concentration of the collected DNA is too low to quantify using Qubit. This is not an indicator of a failed experiment. Proceed with library preparation regardless of the concentration of the collected DNA.

9.10. Proceed to Library Preparation for Sequencing section or store samples at -20 °C until ready to process.

Library Preparation for Sequencing

NOTE: The following steps use the NEB Ultra II DNA Library Prep Kit. When handling steps requiring the Ultra II Ligation Master Mix, avoid touching sample tubes and always keep them on ice.

10. End Repair and Adaptor Ligation

10.1. Using 0.1X TE pH8 bring up total volume of CUT&RUN DNA to 50 µL. Make a master mix of 3 µL End Prep Enzyme Mix and 7 µL End Prep Reaction Buffer per sample. Add 10 µL of master mix to CUT&RUN DNA and mix thoroughly by pipetting up and down.

10.2. Perform a quick spin at 100 x g to collect all liquid from the sides of the tube. Place the tubes in a thermocycler, with the heated lid set to ≥ 75 °C, and run the following program:

Temp	Time	Total Number of Cycles
20 ºC	30 min	1
50 ºC	60 min	1
4 ºC	Hold	1

NOTE: Depending on the starting input DNA concentrations collected from section 9, follow the required adaptor dilution from the table below.

Input DNA	Adaptor Dilution	Working Adaptor Conc
101 ng-1 ug	No dilution	15 µM
5-100 ng	10-fold	1.5 µM
< 5 ng	25-fold	0.6 µM

10.3. Add 2.5 µL Adaptor per sample and mix thoroughly by pipetting up and down at least 10 times.

NOTE: It is critical that the adaptor is added to the sample and mixed thoroughly before the ligation master mix is added.

10.4. Make a master mix of 30 µL Ultra II Ligation Master Mix, 1 µL Ligation Enhancer. Add 31 µL of the master mix to the sample(s). Mix thoroughly by pipetting up and down at least 10 times.

10.5. Incubate at 20 oC for 15 min in a thermocycler with the heated lid off.

NOTE: It is critical that samples be always kept on ice and transferred to the thermocycler only after the thermocycler has already reached 20 ^{o}C.

10.6. Perform a quick spin at 100 x g to collect all liquid from the sides of the tube, then add 3 µL of USER Enzyme and incubate tubes in thermocycler at 37 °C for 15 min with the heated lid set to ≥ 47 °C.

NOTE: This is a safe stopping point, samples can be stored at -20°C or continued directly to step 11.

11. Cleanup of Adaptor-Ligated DNA without Size Selection

NOTE: Place magnetic beads at room temperature for at least 30 min before using.

11.1. Add 154.4 µL (~1.6X sample volume) of the DNA Purification Beads to the Adaptor Ligation reaction from step 10. Pipet mix and incubate samples on benchtop for at least 5 min at room temperature.

11.2. Place the tubes on the magnetic rack and once the slurry has cleared, carefully remove, and discard the supernatant using a pipette.

11.3. Add 200 µL of freshly prepared room temperature 80% ethanol to the tubes and incubate at room temperature for 30 s. Carefully remove and discard the supernatant using a pipette, repeat step for a total of two washes with 80% ethanol.

11.4. Spin the tubes briefly at 100 x g. Place the tubes back on the magnetic rack and remove any leftover ethanol using a pipette. Air dry the beads for up to 5 min while the tubes remain on the magnetic rack with the lid open.

NOTE: Do not exceed 5 min drying time as it can significantly reduce the final DNA yield.

11.5. Remove the tubes from the magnetic rack and elute the DNA from the beads by adding 17 µL of or 0.1X TE pH 8. Mix well then incubate for at least 5 min at room temperature.

11.6. Place the tubes on the magnetic rack until slurry becomes clear. Once the slurry has cleared, carefully transfer 15 µL of the supernatant to a sterile 0.2 mL PCR tube.

12. PCR Enrichment of Adaptor-Ligated DNA

12.1. Make a master mix of 25 µL Ultra II Q5 Master Mix and 5 µL Universal PCR Primer/i5 Primer (10 µM) per sample.

NOTE: Prepare one extra sample-worth of master mix to account for pipetting losses.

12.2. Add 30 µL of master mix to the 15 µL Adaptor Ligated DNA sample. Add 5 µL of unique Index Primer/i7 Primer (10 µM) to each sample to bring the final volume to a total of 50 µL. Mix thoroughly by pipetting up and down at least 10 times.

12.3. Perform the following PCR cycling conditions:

	Temp	Time	Total Number of Cycles
Initial Denaturation	98°C	45 s	1
Denaturation	98°C	15 s	14
Annealing/Extension	65°C	10 s	
Final Extension	65°C	5 min	1
Hold	4°C	Hold	1

13. Cleanup of PCR Amplified DNA with Size Selection

NOTE: Incubate DNA Purification Beads at room temperature for at least 30 min before use. Mixing steps involve pipetting up and down at least 10 times.

13.1. Vortex DNA Purification Beads to resuspend. Add 35 µL (~0.7X sample volume) of resuspended beads to the PCR amplified DNA samples. Mix and incubate samples on nutator for at least 5 min at room temperature.

13.2. Place the tubes on magnetic rack and once the slurry is clear, transfer the supernatant containing your DNA to a new Simport 0.2 mL 8-well PCR strip tube.

13.3. Add 119 µL (~1.4X sample volume) beads to the sample, mix by pipetting up and down. Incubate samples on nutator for at least 5 min at room temperature.

13.4. Place the tubes on the magnetic rack and once the slurry has cleared, carefully remove, and discard the supernatant using a pipette.

13.5. Add 200 µL of freshly prepared room temperature 80% ethanol to the tubes and incubate at room temperature for 30 s. Carefully remove and discard the supernatant using a pipette, repeat step for a total of two washes with 80% ethanol.

13.6. Spin the tubes briefly at 100 x g. Place the tubes back on the magnetic rack and remove any leftover ethanol using a pipette. Air dry the beads for up to 5 min while the tubes remain on the magnetic rack with the lid open.

NOTE: Do not exceed 5 min drying time as it can significantly reduce the final DNA yield.

13.7. Remove the tubes from the magnetic rack and elute the DNA from the beads by adding 14 µL of or 0.1X TE pH 8. Mix well then incubate for at least 5 min at room temperature.

13.8. Place the tubes on the magnetic rack until slurry becomes clear. Once the slurry has cleared, carefully transfer 13 µL of the supernatant to a sterile 0.2 mL PCR tube.

14. Size Select Libraries by Polyacrylamide Gel Electrophoresis

14.1. Prepare fresh 1x TBE and insert pre-made commercial 10% acrylamide TBE gel in the gel electrophoresis apparatus filled with 1x TBE.

14.2. In the first well add 2 µL of Low Range DNA ladder.

14.3. Mix 3 µL of 6X Loading Dye with 13 µL of sample previously collected on step 13.8.

14.4. Carefully add 15 µL into each well of the gel.

NOTE: We recommend leaving one well in the gel empty between each sample, if possible. This reduces the likelihood of sample cross contamination.

14.5. Run gel for 90 min at 70 V.

14.6. Remove the gel cast from the gel box. Open the gel cast per manufacturer's instructions.

14.7. Gently remove the gel from the gel cast per manufacturer's instructions and place it inside a gel holding tray or a PCR-tube box cover filled with 100 mL of 1x TBE.

NOTE: Make sure to gently remove gel from cast to avoid ripping the gel as it is thin and fragile. It is critical to soak the users gloves and the gel itself with 1x TBE whenever handling the gel.

14.8. Add 10 µL of Sybr Gold to the tray and gently swirl for a brief period. Cover with foil to protect from light and incubate statically at room temperature for 10 min.

14.9. Rinse gel twice with 100 mL of in-house deionized tap water.

14.10. Image the gel under blue light illumination using an amber filter cover (Figure 1).

NOTE: Successful libraries show a smear between 100-500 bp. There will also be a prominent ~125 adaptor-dimer band. The presence of adaptor-dimer is not an indicator of poor library quality or failure. This amount of adaptor-dimer is unavoidable for CUT&RUN experiments against low-abundance TFs and is a consequence of low amount of input material used to prepare these libraries. It is critical that ultraviolet light is not used to avoid damaging the DNA.

14.11. As shown in Figure 1, for each individual library, cut the gel slightly above the ~125 bp prominent adaptor-dimer band (making sure to avoid touching the prominent adaptor-dimer band) and below the 400 bp ladder mark.

NOTE: It is critical to avoid the ~125 adaptor-dimer band. Even tiny amounts of adaptor-dimers will significantly reduce library quality.

14.12. Puncture the bottom of 0.65 mL tube using a 22-gauge needle and place the punctured tube inside a sterile 2 mL microfuge tube. Transfer slice of gel to the punctured tube inside of the 2 mL microfuge tube.

14.13. Centrifuge the 2 mL microfuge tube containing the 0.65 mL punctured tube and sample at 10,000 x g for 2 min to collect the gel slurry inside the 2 mL microfuge tube. The punctured tube should now be empty and can be discarded.

NOTE: If the punctured tube still has any gel remaining inside, place the punctured tube back inside the 2 mL microfuge tube and centrifuge again at 10,000 x g for 2 min.

14.14. To the gel slurry inside the 2 mL microfuge tube, add 300 µL of ice-cold gel elution Buffer and mix on a nutator platform at room temperature for a minimum of 3 h or overnight (12-16 h).

14.15. Transfer all liquid and gel slurry to 0.22 µM filter column. Centrifuge at 10,000 x g for 1 min. Collected volume should be ~ 300 µL.

14.16. Add 450 µL (~1.5X sample volume) of DNA Purification Beads and incubate at room temperature for 5 min on shaking nutator. After the 5 min, place sample on magnetic rack until slurry is clear.

NOTE: Incubate DNA Purification Beads at room temperature for at least 30 min before use. Mixing steps involve pipetting up and down at least 10 times.

14.17. Remove and discard 500 µL of supernatant, making sure not to disrupt the beads.

14.18. Remove sample from magnetic rack and mix beads by pipetting up and down several times. Transfer 200 µL of sample into new PCR strip tube.

14.19. Place the strip-tubes on the magnetic rack and once the slurry has cleared, carefully remove, and discard the supernatant using a pipette.

14.20. Add 200 µL of freshly prepared room temperature 80% ethanol to the tubes and incubate at room temperature for 30 s. Carefully remove and discard the supernatant using a pipette, repeat step for a total of two washes with 80% ethanol.

14.21. Spin the tubes briefly at 100 x g. Place the tubes back on the magnetic rack and remove any leftover ethanol using a pipette. Air dry the beads for up to 5 min while the tubes remain on the magnetic rack with the lid open.

NOTE: Do not exceed 5 min drying time as it can significantly reduce the final DNA yield.

14.22. Remove the tubes from the magnetic rack and elute the DNA from the beads by adding 17 µL of 0.1X TE pH 8. Mix well then incubate for at least 5 min at room temperature.

14.23. Place the tubes on the magnetic rack until slurry becomes clear. Once the slurry has cleared, carefully transfer 15 µL of the supernatant to a sterile 0.2 mL PCR tube.

14.24. Measure the final library quantity using Qubit.

14.25. This final library is now ready for sequencing.

NOTE: We typically pool together up to 48 libraries and sequence the pooled libraries using an Illumina NextSeq platform with paired-end 40 bp read lengths.

CUT&RUN sequence analysis

From here onwards, we present our computational protocol to analyze CUT&RUN sequence data. The protocol begins with setting up the computational virtual environment and walks users through executing the commands on their local machine. This protocol will work on all computational resources such as local machines, AWS instances, high-performance computing clusters, etc.

Analysis software prerequisites

15. Users can download the source code for CUT&RUN analysis from GitHub - https://github.com/akshayparopkari/cut_run_analysis

15.1. The workflow will work best on a system running MacOS or various Linux OS. Windows users can run the workflow using GitBash (https://gitforwindows.org/).

15.2. Users can directly download the code from GitHub page by clicking on the green “Code” button, followed by clicking on “Download ZIP” option. Then, users can unzip the folder to a relevant location on their local machine.

15.3. (Run only once) Install Conda environment – This workflow uses the Conda command line tool environment to install all required software and tools. Conda software can be accessed at https://docs.conda.io/en/latest/miniconda.html.

15.4. (Run only once) Once Conda is installed, users must create a virtual environment using the supplementary file 2 provided with the following command –

```
conda create --name <env> --file Supplementary_File_2.txt
```

15.5. Users will need to activate the virtual environment, every time they want to execute/run this workflow using –

```
conda activate <env>
```

15.6. Organize your input raw FASTQ files in a single folder, ideally one folder per CUT&RUN experiment

Generate genome file for alignment (run only once for each genome file)

16. Create a folder to save all *C. albicans* genome files

16.1. `mkdir ca_genome_files`

17. Download *C. albicans* genome assembly 21 from *Candida* Genome Database, using either `wget` or `curl` tool (Skrzypek et al., 2016).

17.1.
```
wget
http://www.candidagenome.org/download/sequence/C_albicans_SC5314/Assembly21/current/C_albicans_SC5314_A21_current_chromosomes.fasta.gz
```

17.2.
```
curl -O
http://www.candidagenome.org/download/sequence/C_albicans_SC5314/
```

```
Assembly21/current/C_albicans_SC5314_A21_current_chromosomes.fasta.gz
```

NOTE: We use *C. albicans* assembly 21 here to compare CUT&RUN results with previously published ChIP-chip results, which were aligned to Assembly 21. Users can download other assembly versions and run similar commands to generate relevant genome files for their alignment needs.

18. Generate Bowtie2 index database (database name: ca21)

18.1.
```
bowtie2-build
C_albicans_SC5314_A21_current_chromosomes.fasta.gz ca21
```

18.2.
```
bowtie2-inspect -s ca21
```

<u>Run CUT&RUN analysis pipeline</u>

19. Familiarize yourself with the parameters of the pipeline by reading through the help section.

19.1.
```
bash cut_n_run_pipeline.sh -h
```

20. Execute cut_n_run_pipeline.sh file with relevant parameters. An example to execute the script is shown below.

20.1.
```
bash cut_n_run_pipeline.sh /path/to/input/folder 4 y y y
y y > /path/to/output.log 2>&1
```

<u>Organize output files</u>

21. Merge significant peaks from all replicates called by Macs2 located in `/path/to/input/folder/peakcalling/macs2` using BedTools merge function (Quinlan & Hall, 2010).

21.1.
```
cat
/path/to/input/folder/peakcalling/macs2/all_replicate_files
sort -k1,1 -k2,2n | mergeBed -c 4,5,6,7,8,9 -o
last,mean,first,mean,mean,mean > /path/to/merged_output.bed
```

22. Remove matches to blacklisted genomic regions BedTools subtract function.

22.1.
```
subtractBed -a /path/to/merged_output.bed -b
/path/to/Supplementary_File_1.bed -A >
/path/to/merged_output_no_blacklist_hits.bed
```

NOTE: `Supplementary_File_1.bed` contains the blacklisted regions in *C. albicans* genome. Users can skip this step to keep signal contained within these blacklist regions.

23. Merge BigWig files from replicates using UCSC bigWigMerge function (Kent et al. 2010).

23.1. Convert the BedGraph output from bigWigMerge to BigWig using UCSC bedGraphToBigWig function.

2.3.1 Results

We describe a robust CUT&RUN protocol adapted and optimized for investigating the genome-wide localization of specific TFs in *C. albicans* biofilms and planktonic cultures (see **Figure 2.1** for an overview of the experimental approach). We also describe a thorough data analysis pipeline to analyze the resulting CUT&RUN sequence data that requires users to have minimal expertise in coding or bioinformatics (see **Figure 2.2** for an overview of the analysis pipeline). Contrary to the ChIP-chip and ChIP-seq methods, CUT&RUN is carried out using intact permeabilized nuclei prepared from a significantly reduced number of input cells, without formaldehyde crosslinking. Isolating intact nuclei from *C. albicans* spheroplasts is a critical step in the protocol. Efficient spheroplasting via digestion of the *C. albicans* cell wall using zymolyase can be challenging since the enzymatic digestion reaction conditions must be optimized for each cell type. Thus, to ensure a successful CUT&RUN experiment with high-quality sequencing results, we use an early quality control step and verify the presence of intact nuclei using a standard fluorescence microscope. We regularly assess cell wall digestion and nuclear integrity by visualizing both control intact cells and isolated nuclei stained with the cell wall dye calcofluor white and the nucleic acid strain SYTO 13. In contrast to the isolated intact nuclei, where cell wall staining by calcofluor white is not observed, both nuclei and the cell walls are fluorescently labeled in the intact control cells (**Figure 2.3A**). Lastly, prior to sequencing, we evaluate fragment size distribution of CUT&RUN libraries using a bioanalyzer. This quality control step is a reliable measure in assessing the quality of CUT&RUN libraries. As seen in **Figure 2.3B**, successful libraries generated for experiments investigating TFs show high enrichment for fragments smaller than 280 bp. We also recommend assessing the final pooled libraries using a bioanalyzer to ensure the complete removal of contaminating adapter dimers (**Figure 2.3C**). In our experience, 5-10 million paired-end reads per-library provide sufficient sequencing depth for most TF CUT&RUN experiments in *C. albicans*.

We validated this CUT&RUN protocol and accompanying data analysis pipeline by investigating two TFs, Ndt80 and Efg1, controlling *C. albicans* biofilm formation. As shown in **Figure 2.3D**, both Ndt80 and Efg1 are bound at intergenic regions (highlighted in red) flanking the TEC1 ORF. These intergenic regions surrounding TEC1 were previously shown to be highly enriched for Ndt80 and Efg1 binding during biofilm formation by ChIP-chip4. A systematic comparative analysis indicated that our CUT&RUN protocol successfully identified the majority of the previously known binding events for Ndt80 and Efg1 during biofilm formation4 (**Figure 2.4**). Furthermore, we were able to identify many new TF binding events that were not captured in the previously published ChIP-chip experiments (**Figure 2.4**). Overall, both Ndt80 and Efg1 bound to loci overlapping with previously published ChIP-chip data, as well as to loci identified only using the CUT&RUN method (overlaps between our CUT&RUN data and previously published ChIP-chip data for Ndt80 and Efg1 are summarized in the Venn diagrams in **Figure 2.4**). Nonetheless, our CUT&RUN method identified most of the significant peaks identified by ChIP-chip along with additional peaks missed using the ChIP-chip method. In summary, these results show that the CUT&RUN protocol described here is a robust method optimized for investigating *C. albicans* TF-DNA binding interactions from low-abundance samples.

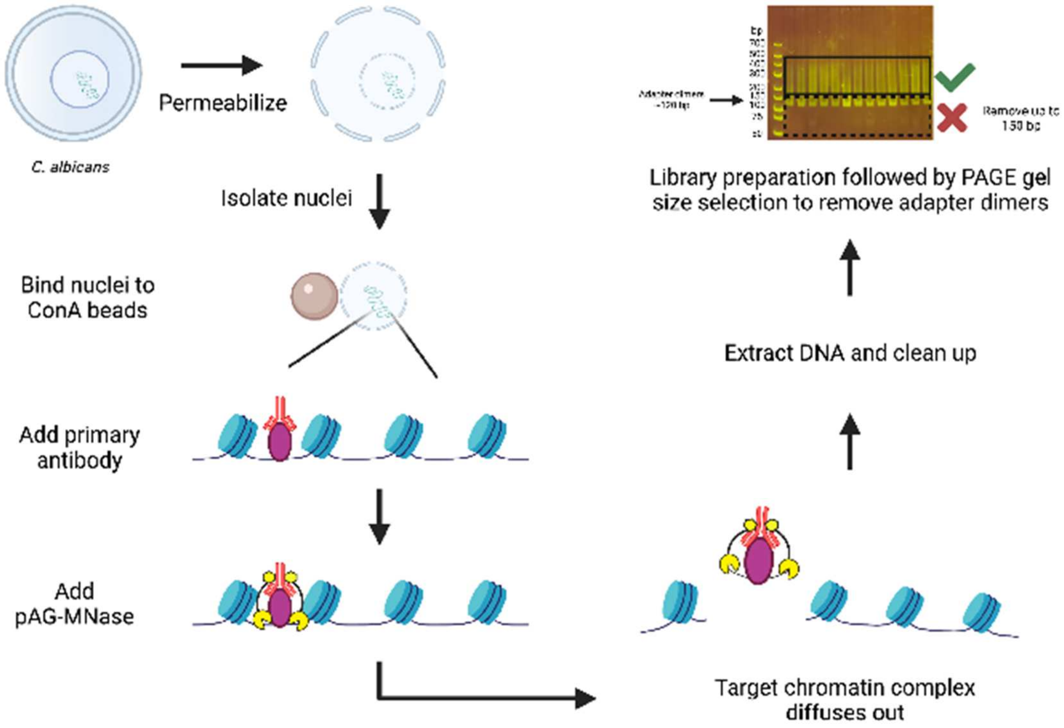

Figure 2.1: Schematic of the CUT&RUN protocol. First, *C. albicans* cells are permeabilized to isolate intact nuclei. ConA beads are activated, and the intact nuclei are then bound to the activated ConA beads. Antibody of interest is added to the bead-bound nuclei and incubated at 4 ^{o}C. Next, pAG-MNase is added and allowed to bind to the target antibody. Only after the addition of $CaCl_2$, pAG-MNase is activated and targeted chromatin digestion proceeds until the addition of the chelating regent to inactivate pAG-MNase. The pAG-MNase bound antibody complex is allowed to diffuse out of the permeabilized nuclei and the resulting DNA extracted and cleaned-up. Sequencing ready libraries are prepared from the CUT&RUN enriched DNA fragments. The resulting libraries are then run on a 10% PAGE gel to separate and remove contaminating adaptor-dimers from the CUT&RUN enriched fragments which are now ligated with sequencing adaptor.

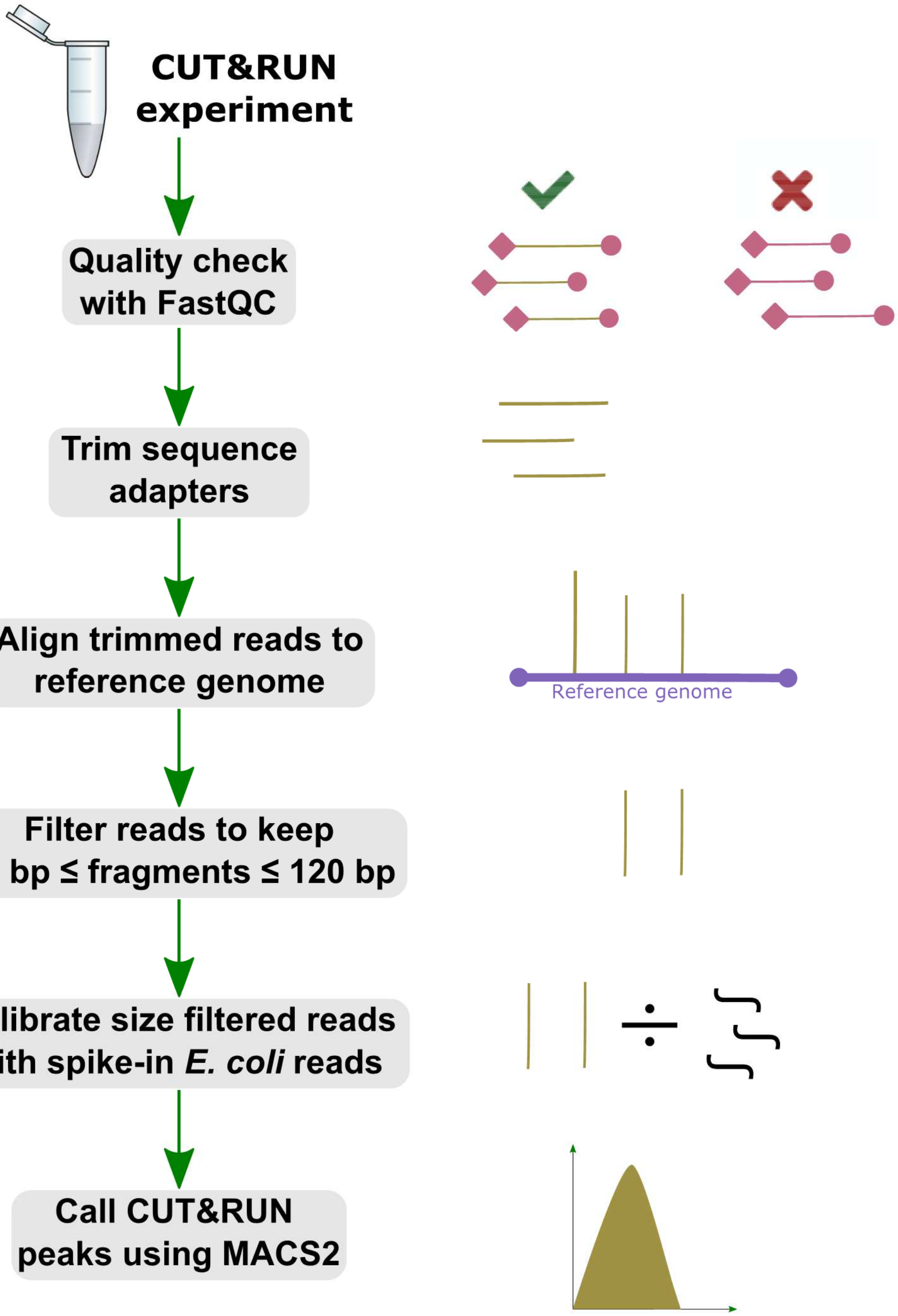
CUT&RUN experiment
Quality check with FastQC
Trim sequence adapters
Align trimmed reads to reference genome
Reference genome
Filter reads to keep 20 bp ≤ fragments ≤ 120 bp
Calibrate size filtered reads with spike-in *E. coli* reads
÷
Call CUT&RUN peaks using MACS2

Figure 2.2: Schematic of CUT&RUN sequence data analysis. The workflow starts by performing quality check on raw FASTQ files using FASTQC, followed by trimming to remove sequencing adapters. The trimmed reads are then aligned to reference genome and the aligned reads are filtered based on their size to enrich for transcription factor-sized binding signals (20 bp ≤aligned read ≤120 bp). Size selected reads are then calibrated against spike-in *Escherichia coli* reads, and lastly, calibrated reads are used to call peaks using MACS2.

2.4 Discussion and Summary

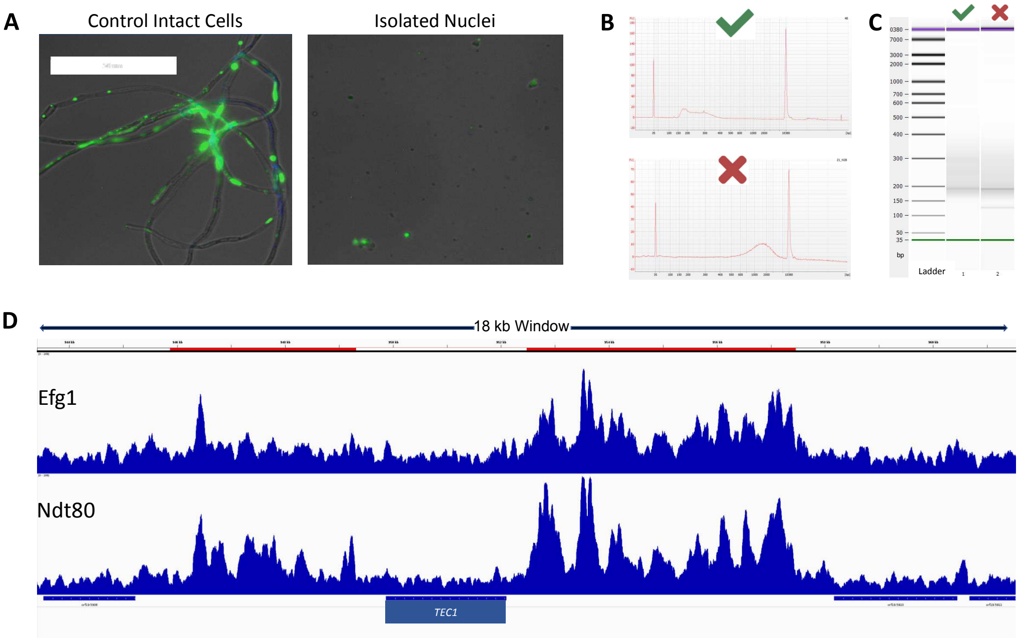

Figure 2.3: Quality control steps critical for successful CUT&RUN experiments. (A) Cells were stained with calcofluor white and syto 13 green fluorescent nucleic acid stain before and after nuclei isolation and visualized using EVOS fluorescent microscope. (B) CUT&RUN libraries are analyzed using Bioanalyzer. Successful CUT&RUN TF libraries (indicated by the green checkmark) are enriched for short-fragments smaller than 200 bp and is the best indicator of success. (C) 48 CUT&RUN libraries pooled together and analyzed using Bioanalyzer. High quality pooled CUT&RUN libraries (indicated by the green checkmark) are free of adaptor-dimers (lane 1) while low-quality pooled libraries (indicated to the red 'X') retain small amounts of adaptor-dimers (lane 2). (D) Representative IGV tracks from CUT&RUN datasets showing significant enrichment for Efg1 and Ndt80 binding at the intergenic regions flanking Tec1.

This protocol presents a comprehensive experimental and computational pipeline for genome-wide localization of regulatory transcription factors in *C. albicans* and is designed to be highly accessible to anyone with standard microbiology and molecular biology training. By leveraging the high dynamic range and low sample input requirements of the CUT&RUN assay and optimizing the protocol for localization of TF-DNA binding interactions in *C. albicans*, we have developed a powerful and affordable alternative to traditional ChIP-seq approaches. When compared to ChIP-seq, this protocol is more amenable to high throughput, requires substantially lower input cell numbers, does not

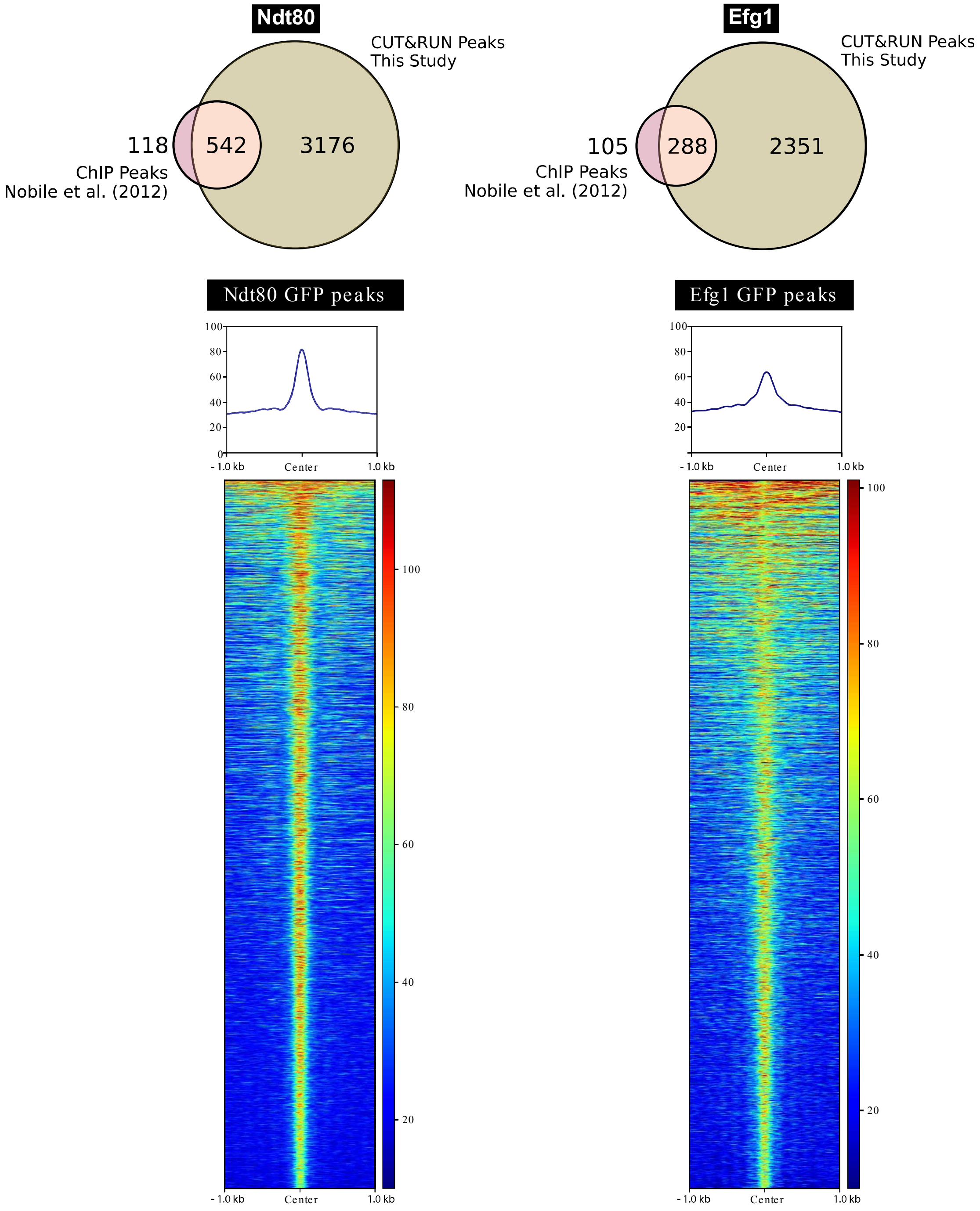

Figure 2.4: Evaluation of Ndt80 and Efg1 enriched peaks identified using our CUT&RUN protocol and data analysis pipeline on *C. albicans* biofilm cells. The Venn diagrams in the top row illustrate the degree of overlap between Ndt80 and Efg1 binding sites identified via CUT&RUN with previously published ChIP-chip data. The bottom row highlights CUT&RUN signals for all binding events for Ndt80 and Efg1 as colored heatmaps (red = high peak signal, blue = low/no peak signal) and the signal intensity as a profile plot above the heatmaps. 1000 bp region upstream (-1 kb) and downstream (+1 kb) are also shown in the heatmap.

require the use of toxic crosslinking agents, and requires ten-fold fewer sequencing reads per sample to produce high-quality results (Hainer et al., 2019; Meers et al., 2019; Skene & Henikoff, 2017; Skene et al, 2018). To further reduce the per-sample cost of this protocol, we have included buffer recipes and a detailed reagent list to enable in-house preparation of all necessary buffers and economical bulk sourcing of other essential reagents. Since *C. albicans* biofilm formation, phenotypic switching, and commensalism are all regulated by complex interwoven transcriptional networks, this robust, facile, and affordable methodology provides a powerful new tool for understanding these and many other cellular processes in this important fungal pathogen (Rodriguez et al., 2020).

TFs are not as abundant as histones or other chromatin-associated proteins, thus creating a unique challenge for investigating TF-DNA binding interactions via CUT&RUN. To address this challenge, we made critical adjustments and optimizations to the standard CUT&RUN experimental protocol (Skene & Henikoff, 2017). Since most successful CUT&RUN experiments targeting TFs yield a small amount of DNA that is too dilute to quantify and is often enriched for fragments smaller than 150 bp in our protocol, we optimized the End Repair and dA-Tailing reaction conditions to favor these smaller fragments (Hainer et al., 2019; Skene & Henikoff, 2017; Zhu et al., 2019). Even with this optimization step, we found that the PCR-amplified libraries contained a significant proportion of adapter dimers, which could not be completely removed using magnetic bead-based DNA size selection methods. To address this issue, we additionally included a PAGE gel size selection step to generate final sequencing ready libraries that are largely devoid of adapter dimers. This is a critical step of our CUT&RUN protocol, as removing adapter dimers while retaining the smaller TF-derived CUT&RUN fragments is essential for obtaining high quality results. Furthermore, our computational pipeline filters the sequencing data to focus on the smaller reads that are derived from TF-DNA binding interactions in the CUT&RUN assay. Due to these TF-specific adjustments, our protocol is not recommended for the profiling of large chromatin associated complexes such as nucleosomes. While it is theoretically possible to adapt our protocol for this purpose by following the standard library preparation protocol included with the NEBNext Ultra II Library Prep kit, one would still need to adjust the post-sequencing size selection included in our computational pipeline. Specifically, in the size filtering section in the code file `cut_n_run_pipeline.sh`, users would need to replace the current value of “14400” (120 bp * 120 bp) with the square of the desired fragment length to enable our analysis pipeline to analyze the sequencing results generated for other types of chromatin-DNA binding interactions.

Another key step in a successful CUT&RUN experiment includes choosing optimal post-sequencing data analysis parameters. While most of our computational pipeline is designed to be standardized and applicable to the study of any regulatory TF of interest in *C. albicans*, there are two important considerations that the user should evaluate while running the pipeline. The first consideration is whether to include or remove duplicate reads from the sequencing data prior to identification of bound target sites. Since low abundance TFs will typically yield sequencing data containing a significant percentage of reads that are derived from PCR duplication during the library amplification step, removing PCR duplicates can have a significant negative impact on the results. However, with highly abundant TFs or chromatin associated proteins, PCR duplicates typically represent a smaller portion of the total number of reads and are often removed to suppress background noise in the data. Ultimately, this decision to keep or remove PCR duplicates is dependent on the TF of interest and the depth of sequencing data obtained, and thus we automatically generate independent output files for data derived with or without PCR duplicate reads so the user can decide which output files yield the best results for each experiment. The second consideration is whether to identify and remove problematic loci that yield significant, yet highly variable, enrichment in both experimental (antibody against protein of interest) and negative control (IgG) samples. Our peak-calling algorithm uses MACS2 (https://pypi.org/project/MACS2/) to identify significantly enriched loci in both the experimental and control samples and excludes those that appear in both. While this typically eliminates these problematic loci, we have noticed that some of these loci occasionally appear as significant peaks in certain experiments, but based on our experience, we do not believe that these are true positive sites of TF enrichment. Thus, we provide an optional filtering step to remove these problematic loci, which we refer to as "blacklisted" loci. Our list of blacklisted loci primarily contains highly repetitive sequence elements and regions such as telomeric repeats and centromeres that have historically yielded false positive results in our previous genome-wide binding assays. We note that this is a very conservative list of loci that we have high confidence in assigning as problematic; however, each user should evaluate whether this filter is appropriate for their experiment(s) on a case-by-case basis.

CUT&RUN has become a popular choice for investigating protein-DNA interactions in higher eukaryotes as well as in the model yeast *Saccharomyces cerevisiae*, and we have successfully adapted this methodology to investigate genome-wide TF-DNA binding interactions in the clinically relevant fungal pathogen *C. albicans*. This protocol provides detailed methods for all necessary experimental and computational procedures, from engineering strains that express epitope-tagged TFs, through to the computational analysis of the resulting CUT&RUN sequencing data. Overall, this protocol and the accompanying data analysis pipeline produce robust TF-DNA binding profiles, even when using complex multimorphic populations of cells isolated from low abundance biofilm samples and provides superior data quality at a lower overall cost when compared to ChIP-seq methodologies.

2.5 References

Ennis, C. L., Hernday, A. D., & Nobile, C. J. (2021). A Markerless CRISPR-Mediated System for Genome Editing in *Candida* auris Reveals a Conserved Role for Cas5 in the Caspofungin Response. Microbiology spectrum, 9(3), e01820-21.

Hainer, S.J., Fazzio, T.G. High-resolution chromatin profiling using CUT&RUN. Current Protocols in Molecular Biology. 126 (1), e85, doi: 10.1002/cpmb.85 (2019).

Meers, M.P., Bryson, T.D., Henikoff, J.G., Henikoff, S. Improved CUT&RUN chromatin profiling tools. eLife. 8, doi: 10.7554/elife.46314 (2019).

Nguyen, N., Quail, M. M., & Hernday, A. D. (2017). An efficient, rapid, and recyclable system for CRISPR-mediated genome editing in *Candida albicans*. MSphere, 2(2), e00149-17.

Nobile, C. J., Fox, E. P., Nett, J. E., Sorrells, T. R., Mitrovich, Q. M., Hernday, A. D., ... & Johnson, A. D. (2012). A recently evolved transcriptional network controls biofilm development in *Candida albicans*. Cell, 148(1-2), 126-138.

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics, 26(6), 841-842

Skene, P. J., & Henikoff, S. (2017). An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites. Elife, 6, e21856.

Skene, P.J., Henikoff, J.G., Henikoff, S. Targeted in situ genome-wide profiling with high efficiency for low cell numbers. Nature Protocols. 13 (5), 1006–1019, doi: 10.1038/nprot.2018.015 (2018).

Skrzypek, M. S., Binkley, J., Binkley, G., Miyasato, S. R., Simison, M., & Sherlock, G. (2016). The *Candida* Genome Database (CGD): incorporation of Assembly 22, systematic identifiers and visualization of high throughput sequencing data. Nucleic acids research, gkw924.

Rodriguez, D. L., Quail, M. M., Hernday, A. D., & Nobile, C. J. (2020). Transcriptional circuits regulating developmental processes in *Candida albicans*. Frontiers in Cellular and Infection Microbiology, 10.

Zhu, Q., Liu, N., Orkin, S. H., & Yuan, G. C. (2019). CUT&RUNTools: a flexible pipeline for CUT&RUN processing and footprint analysis. Genome biology, 20(1), 1-12.

Chapter 3

A computational workflow for the analysis of 3’ Tag-Seq data

3.1 Abstract

RNA-sequencing (RNA-seq) is a ubiquitous tool to profile genome-wide changes in gene expression. RNA-seq uses high-throughput sequencing technology to quantify the amount of RNA in a biological sample. With the increasing popularity of RNA-seq, many variations on the protocol have been proposed to extract unique and relevant information from biological samples. 3’ Tag-Seq (also called TagSeq, 3′ Tag-RNA-Seq, and Quant-Seq 3′ mRNA-Seq) is one RNA-seq variation, where the 3’ end of the transcript is selected and amplified to yield one copy of cDNA from each transcript in the biological sample.

We present a simple, easy, and publicly available computational workflow to analyze 3’ Tag-Seq data. The workflow begins by trimming sequence adapters from raw FASTQ files. The trimmed sequence reads are checked for quality using FastQC, aligned to the reference genome, and read counts are obtained using STAR. Differential gene expression analysis is performed using DESeq2, based on differential analysis of gene count data. The outputs of this workflow are MA plots and tables of significant and differentially expressed genes and UpSet plots.

This protocol is intended for users interested in analyzing 3’ Tag-Seq data. As such, transcript length-based normalizations are not performed within the workflow. Future updates to this workflow could include custom analyses based on the gene counts table and data visualization enhancements.

3.2 Introduction

RNA sequencing (RNA-seq) is a widely used tool to detect genome-wide changes in gene expression (Wang et al., 2009). It was first used in *Saccharomyces cerevisiae* to identify the gene expression patterns of all genes, exons, and their boundaries across the genome (Nagalakshmi et al., 2008). For humans, as well as many model organisms like yeast, fruit flies, and mice, RNA-seq has been instrumental in providing high-resolution, functionally relevant, genome annotations (Cherry et al., 2012; Gnerre et al., 2011; International Human Genome Sequencing Consortium, 2004; Matthews et al., 2015). Briefly, the RNA-seq protocol begins with the isolation of total RNA, which is typically enriched/selected for polyadenylated (polyA) RNA or alternatively depleted for ribosomal RNA (rRNA) (Zhao et al., 2018). After this step, double-stranded complementary DNA (cDNA) is synthesized via reverse transcription from the RNA, resulting in cDNA libraries. The cDNA molecules are fragmented, and sequencing adapters are added to the cDNA fragments. Then, the cDNA fragments are subjected to high-throughput sequencing to generate short sequence reads, which are used for downstream analyses.

Whole transcript RNA-seq workflows produce data providing information on quantifying gene expression, novel transcripts, alternatively spliced genes, and allele specific expression (Wang et al., 2009). Although the extent of this information is typically valuable to the researcher, oftentimes, the goal of many biological research projects is simply to identify changes in gene expression patterns between conditions. Given this simplified goal, classical RNA-seq protocols may be more complex and more expensive than is necessary towards the goal of obtaining genome-wide gene expression changes (Ma et al., 2019; Wang et al., 2009). To simplify classical RNA-seq protocols, recent advances in the field have provided alternative RNA-seq methods that are used today to address specific biological questions (Moll et al., 2014; Morrissy et al., 2011). One of these alternative methods is 3' Tag-Seq (also called TagSeq, 3′ Tag-RNA-Seq, and Quant-Seq 3′ mRNA-Seq). In this method, cDNA libraries are reverse transcribed only from the 3'-end of the mRNAs, resulting in a single copy of cDNA arising from each transcript (Ma et al., 2019; Moll et al., 2014; Torres et al., 2008). Compared to classical RNA-seq methods, 3' Tag-Seq is simpler, quicker, and lower cost, and provides sufficient sequencing depth for differential gene expression analysis (Ma et al., 2019). These benefits make 3' Tag-Seq the ideal choice for researchers whose end goals are to identify changes in patterns of gene expression between two or more conditions.

Analysis of classical RNA-seq data was identified as one of the early challenges in dealing with these complex sequencing datasets (Wang et al., 2009). Ultimately, analysis of RNA-seq data is highly dependent on the experimental design used to create the sequencing libraries (Conesa et al., 2016). Consequently, there is no "one size fits all" workflow for the analysis of RNA-seq output reads. Here, we present a simple, easy, and publicly available computational workflow to analyze 3' Tag-Seq data. The workflow begins by trimming sequencing adapters from raw FASTQ files (Cock et al., 2010; Conesa et al., 2016). The trimmed sequence reads are checked for quality using FastQC, aligned to the reference genome, and read counts are obtained using STAR ("Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data," n.d.; Dobin et al., 2013). Differential gene expression analysis is performed using DESeq2, based on differential analysis of gene count data (Love et al., 2014). The outputs of this workflow are MA plots and tables of significant and differentially expressed genes.

3.3 Basic Protocol 1

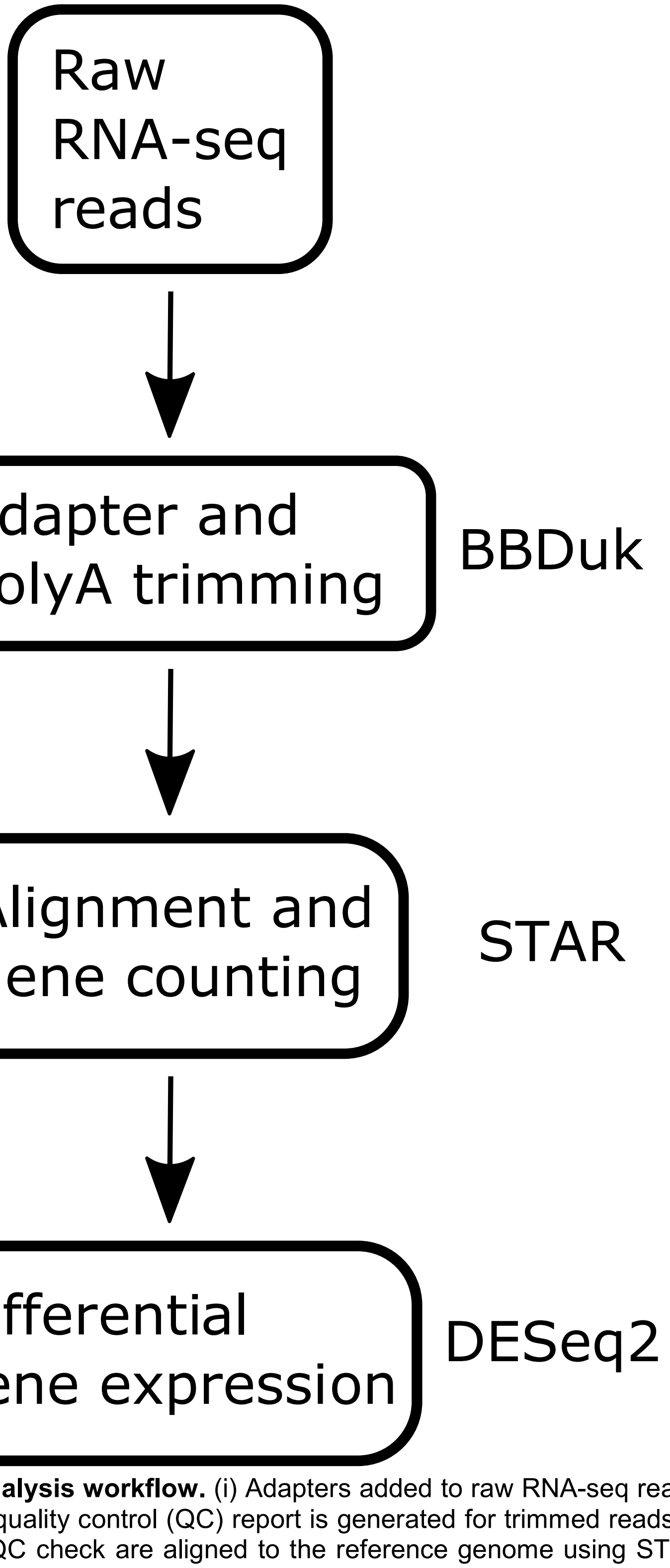
Raw RNA-seq reads
Adapter and polyA trimming
BBDuk
Alignment and gene counting
STAR
Differential gene expression
DESeq2

Figure 3.1: RNA-seq analysis workflow. (i) Adapters added to raw RNA-seq reads are trimmed using BBDuk tool. (ii) A quality control (QC) report is generated for trimmed reads using FastQC. (iii) Reads passing the QC check are aligned to the reference genome using STAR and a gene count table is created. (iv) The gene count table is used to run differential gene expression analysis using DESeq2, and the DESeq2 output is saved as a table to a file for downstream usage.

In the following section, we describe in detail our 3' Tag-Seq analysis workflow. **Figure 3.1** provides a summary of the computational workflow. Briefly the pipeline begins with the processing of raw RNA-seq FASTQ files and ends with a table output of differential gene expression (Love et al., 2014).

Necessary resources

Hardware

An internet connected computer.

Software

The workflow uses the Conda command line tool environment to install all required software and tools. Conda software can be accessed at https://docs.conda.io/en/latest/miniconda.html. The workflow is saved in a bash script file called pipeline.sh. The source code and documentation can be found on GitHub at https://github.com/akshayparopkari/RNAseq/wiki.

Other Requirements

1. Access to computational cluster and login information
2. Basic knowledge of Linux
3. Raw FASTQ sequencing data
4. Sample metadata

3.3.1 Downloading RNAseq workflow on local machine

On Linux and macOS, users can use the in-built Terminal application, and on Windows, users can download and use Git Bash (https://gitforwindows.org/).

1. Navigate to desired directory to download this folder on your machine

```
git clone https://github.com/akshayparopkari/RNAseq.git
```

NOTE: Alternatively, users can click on the green "Code" button on GitHub page - https://github.com/akshayparopkari/RNAseq - followed by clicking on "Download ZIP" option. Then, users can unzip the downloaded folder and save it to a relevant location on their local machines.

2. Make script files executable

```
cd RNAseq/

chmod u+x pipeline.sh

chmod u+x format_counts_table.py
```

3.3.2 Loading the Conda virtual environment

Conda enables virtual environments that contain the required software packages/libraries to be installed and set up. In this instance, the RNAseq Conda environment contains the BBMap suite, STAR alignment software, and FASTQC tool (“Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data,” n.d.; Bushnell, 2014; Dobin et al., 2013). Additionally, required Python and R libraries and their dependencies are also installed.

1. Create Conda environment using Supplemental File 1

```
conda create –n RNAseq –file Supplemental_File_1.txt
```

NOTE: Users only need to create the environment once. For subsequent analysis, users can activate the environment to run the analysis using the following command –

```
conda activate RNAseq
```

3.3.3 Creating an input data folder

The main script of 3’ Tag-Seq is the `pipeline.sh` file. This single bash script contains all the preprocessing steps - QC filtering with BBDuk, QC check with FastQC and, finally, alignment and gene counting with STAR. `pipeline.sh` takes in a single input which is a folder/directory with:

1. all raw FASTQ sequence files AND
2. the sample metadata Excel file

The raw FASTQ sequence files may either be compressed (using gzip) or uncompressed. The file names must start with the sample ID, followed by the underscore and the rest of the file name. For example, projectname_date_L001.fastq.gz should be named sampleid_projectname_date_L001.fastq.gz. The first part of the file name before the first underscore is how the script knows which sample it is processing. The sample metadata file contains all metadata associated with the input samples including sample ID, genotype, condition, treatment, time, etc. For this repository, the sample metadata file must contain at least two columns - SampleID and Condition. The table below is an example of a sample metadata file, where the first two columns SampleID and Condition are required, and the third column FASTQ_file and beyond is optional, but highly recommended. A comprehensive metadata file also enables convenient sample submission to a sequence read archive (SRA), once your manuscript is published. Table 1 represents a sample metadata file.

Table 3.1: Sample metadata file. Users can use this file as a template to generate their metadata file.

SampleID	Condition	FASTQ_file	Other_Sample_Info
Sample1A	WT	Sample1A_S8_L001_R1_001.fastq.gz	...
Sample1B	Mutant	Sample1B_S8_L001_R1_001.fastq.gz	...
Sample2A	WT	Sample2A_S8_L001_R1_001.fastq.gz	...

Sample2B	Mutant	Sample2B_S8_L001_R1_001.fastq.gz	...
Sample3A	WT	Sample3A_S8_L001_R1_001.fastq.gz	...
Sample3B	Mutant	Sample3B_S8_L001_R1_001.fastq.gz	...
...	...	...	...

NOTE: The input directory must contain raw FASTQ files and a sample metadata Excel file. Users may implement a user-defined project structure to organize their RNA-seq data. Please see Cookiecutter Data Science project (https://cookiecutter.readthedocs.io/en/1.7.2/) for ideas on how to best organize computational data.

Transferring data to/from cloud computing resource to a local machine via command line

Below is common usage of secure copy `scp` function which one of the commands used for transferring files to/from cloud computing resource. The other command is secure file transfer protocol `sftp`. Please refer to cloud computing resource wiki for detailed instructions on sftp function.

```
scp FROM TO
```

where FROM is the source location and TO is the destination location.

Third party GUI apps

Users can also use third party clients to transfer files to/from MERCED. FileZilla (https://filezilla-project.org/) for Linux and Windows or Cyberduck (https://filezilla-project.org/) for MacOS and Windows are alternative to using `scp` or sftp to transfer files with drag and drop.

3.3.4 Running the RNAseq pipeline

Users must activate the RNAseq Conda environment, before attempting to executing the pipeline. For more information, please see section 3.3.2.

1. Run the RNAseq pipeline

```
INPUTFOLDER=’’path/to/your/input/folder’’  # enter your data
folder with FASTQ files here

bash pipeline.sh "$INPUTFOLDER" > "$INPUTFOLDER"/preprocess.log
```

3.3.5 Output files

pipeline.sh outputs many files, which can be useful to dig deeper into specific samples to address any discrepancy in the data. The three important files to check are:

1. `gene_raw_counts.txt`, which is a tab-separated file of raw gene counts for all samples with gene names as the rows and samples as columns

2. `deseq2_lfc.txt`, which is a tab-separated file from the DESeq2 analysis
3. `MA_plot.pdf`, which is a PDF depicting volcano plots of log fold changes against mean gene expression

More information about other output files:

1. All "_trimmed.fastq" ending files are trimmed sequences from BBmap and are saved in the trim_log directory
2. All ".bam" ending files are alignment files generated by STAR and are saved in the STAR_log directory
3. All "ReadsPerGene.out.tab" ending files are gene count files for each sample generated by STAR and are saved in the STAR_log directory
4. All "Log.out", "Log.final.out" and "Log.progress.out" ending files are intermediary alignment files generated by STAR and are saved in the STAR_log directory.

3.3.6 Visualizing overlaps in multiple experimental conditions

Users can use the `overlap_upsetR.R` script to visualize overlaps in genes for multiple experimental conditions. The overlap is represented as an UpSet plot (Lex et al., 2014). UpSet plots are an extension of Venn diagrams and are useful when there are more than three categories/sets of conditions/samples. The `overlap_upsetR.R` takes one input – either "up" or "down" – to calculate overlap between various samples/conditions. Users need to supply an input directory in the code on line 38, and run the following command to get the output UpSet plot -

1. To visualize genes upregulated in multiple conditions/samples
   a. `overlap_upsetR.R up`

2. To visualize genes downregulated in multiple conditions/samples
   a. `overlap_upsetR.R down`

Figure 3.2 is an example of UpSet plot showing upregulated genes overlapping various combinations of four experimental conditions.

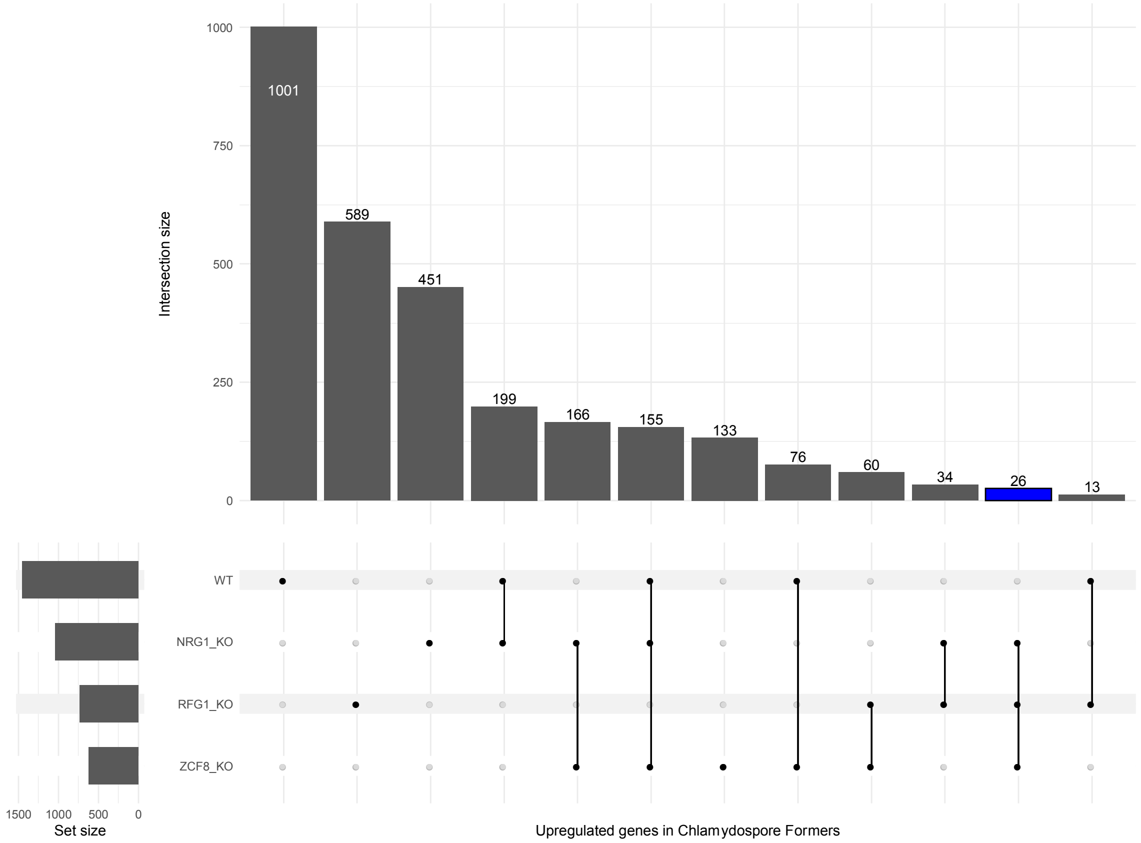

Figure 3.2: Example UpSet plot output. Three transcription factors (TFs) – Nrg1, Rfg1 and Zcf8 – knockout (KO) RNAseq experimental data is shown. The bar plot on the top represents the overlap of upregulated genes in each of the three experimental samples as well as control WT sample. The horizontal bar on the left highlights the number of significantly upregulated genes in each of the four experimental conditions. The black circles and lines in the bottom show the categories for which overlap is calculated and shown by the bar chart on the top. E.g., the fourth bar on the top represents 199 significantly upregulated genes observed in both WT and Nrg1-KO conditions. The blue bar is highlighting overlaps for the three experimental conditions.

3.4 Support Protocol 1

During the alignment step, STAR utilizes genome index files for mapping sequenced reads to a reference genome. This protocol describes how to generate genome indices for the *Candida albicans* Assembly 21 genome as an example. These steps can be used to generate genome indices for your reference genome of choice.

1. In your home folder, download *C. albicans* chromosomal sequences from the *Candida* Genome Database - http://www.candidagenome.org/ (Skrzypek et al., 2017).

```
wget
http://www.candidagenome.org/download/sequence/C_albicans_SC5314/
Assembly21/current/C_albicans_SC5314_A21_current_chromosomes.fast
a.gz
```

```
gunzip C_albicans_SC5314_A21_current_chromosomes.fasta.gz
```

2. Download *C. albicans* genome annotation GTF file from *Candida* Genome Database.

```
wget
http://www.candidagenome.org/download/gff/C_albicans_SC5314/Assem
bly21/C_albicans_SC5314_A21_current_features.gtf
```

```
gunzip C_albicans_SC5314_A21_current_features.gtf
```

3. Activate 3'Tag-Seq Conda environment

```
module load anaconda3
```

```
source activate RNA-seq
```

4. Generate STAR genomes

```
mkdir ca_genome/
```

```
cd ca_genome/
```

```
STAR --runMode genomeGenerate --genomeDir ./ --genomeFastaFiles
~/ C_albicans_SC5314_A21_current_chromosomes.fasta
```

5. STAR will generate output index files in the `ca_genome` folder.

3.5 Summary

We presented an RNAseq workflow which takes in raw FASTQ files and provides differential gene expression table as the output. The workflow is specifically designed to handle data generated using 3' Tag-Seq experimental protocol. We also provide an R script which can visualize overlapping genes among multiple experimental conditions. The code is published on GitHub which allows users to dig into the source code and potentially modify the code to their specific use case.

3.6 References

Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data [WWW Document], n.d. URL https://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (accessed 11.7.21).

Bushnell, B., 2014. BBMap: A Fast, Accurate, Splice-Aware Aligner (No. LBNL-7065E). Lawrence Berkeley National Lab. (LBNL), Berkeley, CA (United States).

Cherry, J.M., Hong, E.L., Amundsen, C., Balakrishnan, R., Binkley, G., Chan, E.T., Christie, K.R., Costanzo, M.C., Dwight, S.S., Engel, S.R., Fisk, D.G., Hirschman, J.E., Hitz, B.C., Karra, K., Krieger, C.J., Miyasato, S.R., Nash, R.S., Park, J., Skrzypek, M.S., Simison, M., Weng, S., Wong, E.D., 2012. Saccharomyces Genome Database: the genomics resource of budding yeast. Nucleic Acids Research 40, D700–D705. https://doi.org/10.1093/nar/gkr1029

Cock, P.J.A., Fields, C.J., Goto, N., Heuer, M.L., Rice, P.M., 2010. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Research 38, 1767–1771. https://doi.org/10.1093/nar/gkp1137

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., Szcześniak, M.W., Gaffney, D.J., Elo, L.L., Zhang, X., Mortazavi, A., 2016. A survey of best practices for RNA-seq data analysis. Genome Biol 17, 1–19. https://doi.org/10.1186/s13059-016-0881-8

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., Gingeras, T.R., 2013. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29, 15–21. https://doi.org/10.1093/bioinformatics/bts635

Gnerre, S., MacCallum, I., Przybylski, D., Ribeiro, F.J., Burton, J.N., Walker, B.J., Sharpe, T., Hall, G., Shea, T.P., Sykes, S., Berlin, A.M., Aird, D., Costello, M., Daza, R., Williams, L., Nicol, R., Gnirke, A., Nusbaum, C., Lander, E.S., Jaffe, D.B., 2011. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proc Natl Acad Sci U S A 108, 1513–1518. https://doi.org/10.1073/pnas.1017351108

International Human Genome Sequencing Consortium, 2004. Finishing the euchromatic sequence of the human genome. Nature 431, 931–945. https://doi.org/10.1038/nature03001

Lex, A., Gehlenborg, N., Strobelt, H., Vuillemot, R., Pfister, H., 2014. UpSet: Visualization of Intersecting Sets. IEEE Transactions on Visualization and Computer Graphics 20, 1983–1992. https://doi.org/10.1109/TVCG.2014.2346248

Love, M.I., Huber, W., Anders, S., 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 15, 1–21. https://doi.org/10.1186/s13059-014-0550-8

Ma, F., Fuqua, B.K., Hasin, Y., Yukhtman, C., Vulpe, C.D., Lusis, A.J., Pellegrini, M., 2019. A comparison between whole transcript and 3’ RNA sequencing methods using Kapa and Lexogen library preparation methods. BMC Genomics 20, 1–12. https://doi.org/10.1186/s12864-018-5393-3

Matthews, B.B., Dos Santos, G., Crosby, M.A., Emmert, D.B., St Pierre, S.E., Gramates, L.S., Zhou, P., Schroeder, A.J., Falls, K., Strelets, V., Russo, S.M., Gelbart, W.M., FlyBase Consortium, 2015. Gene Model Annotations for Drosophila melanogaster: Impact of High-Throughput Data. G3 (Bethesda) 5, 1721–1736. https://doi.org/10.1534/g3.115.018929

Moll, P., Ante, M., Seitz, A., Reda, T., 2014. QuantSeq 3′ mRNA sequencing for RNA quantification. Nat Methods 11, i–iii. https://doi.org/10.1038/nmeth.f.376

Morrissy, S., Zhao, Y., Delaney, A., Asano, J., Dhalla, N., Li, I., McDonald, H., Pandoh, P., Prabhu, A.-L., Tam, A., Hirst, M., Marra, M., 2011. Tag-Seq: Next-Generation Tag Sequencing for Gene Expression Profiling, in: Tag-Based Next Generation Sequencing. John Wiley & Sons, Ltd, pp. 211–241. https://doi.org/10.1002/9783527644582.ch13

Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M., Snyder, M., 2008. The Transcriptional Landscape of the Yeast Genome Defined by RNA Sequencing. Science 320, 1344–1349. https://doi.org/10.1126/science.1158441

Skrzypek, M.S., Binkley, J., Binkley, G., Miyasato, S.R., Simison, M., Sherlock, G., 2017. The *Candida* Genome Database (CGD): incorporation of Assembly 22, systematic identifiers and visualization of high throughput sequencing data. Nucleic acids research 45, D592–D596. https://doi.org/10.1093/nar/gkw924

Torres, T.T., Metta, M., Ottenwälder, B., Schlötterer, C., 2008. Gene expression profiling by massively parallel sequencing. Genome Res 18, 172–177. https://doi.org/10.1101/gr.6984908

Wang, Z., Gerstein, M., Snyder, M., 2009. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 10, 57–63. https://doi.org/10.1038/nrg2484

Zhao, S., Zhang, Y., Gamini, R., Zhang, B., von Schack, D., 2018. Evaluation of two main RNA-seq approaches for gene quantification in clinical RNA sequencing: polyA+ selection versus rRNA depletion. Sci Rep 8, 4781. https://doi.org/10.1038/s41598-018-23226-4