Characterizing Gene Functions with an Overexpression ORF Collection in *Saccharomyces cerevisiae*

Simon Ascher

Dean's Scholars Biology Honors Thesis

P.I. Vishy Iyer

May 31st, 2008

TABLE OF CONTENTS

ABSTRACT

My research project aims to discover new eukaryotic gene functions using yeast as a model organism. Yeast (*Saccharomyces cerevisiae*), are a prototypical eukaryotic model because of its high degree of genetic similarity to humans, fast generation time, and relatively low-cost maintenance. Identifying gene function in eukaryotes, such as humans, is an important, broad step in mapping gene and protein networks, predicting phenotypes, and understanding disease causation. This project uses a gain-of-function approach to characterizing new gene functions. We use a pool of transformed yeast with each yeast cell "over-expressing" a single gene carried by a plasmid, such that the quantity of the protein product encoded by the gene increases. The pool contains yeast transformed with plasmids representative of approximately 93% of all yeast genes. A plasmid is an exogenous piece of DNA that can be inserted into cells and engineered to carry specific genes of interest. Each plasmid in this pool has a galactose promoter that regulates the expression of a target gene, and only in the presence of galactose is that gene constitutively expressed. The yeast is treated with galactose to induce over-expression of a specific gene.

This project aims to induce overexpression in a pool of yeast and monitor the change in abundance of each plasmid on a genome-wide scale using DNA microarrays. A DNA microarray measures the activity of thousands of yeast genes using the knowledge of complementary binding between nucleotides. By analyzing the patterns in which sets of plasmids are enriched and which drop out during overexpression, their functions may be inferred and characterized. Another aim of this project is to troubleshoot the

overexpression screen: from growing the yeast, to isolating the DNA from the yeast cells, and to amplifying the plasmid DNA.

After several troubleshooting experiments, we found the PCR yield for yeast grown in galactose to be lower than that of yeast grown in glucose and yeast grown in raffinose. Additionally, PCR yield decreased the longer yeast grew in galactose, such that 12 hours of galactose induction yielded the lowest PCR. This could be due to the overexpression plasmids being lost from the yeast cells over time, the plasmids undergoing recombination with the genome, or plasmid DNA being lost or degraded during the DNA prep or PCR procedures. Our results also show changes in relative abundances of plasmid DNA as yeast change media from raffinose to galactose. Additionally, we found functional enrichment of distinct gene sets in yeast pools grown in raffinose and yeast pools grown in galactose. This illustrates the value this overexpression screen has in sorting genes into their functional networks, which in turn provides information in characterizing genes of unknown function.

I. INTRODUCTION

MOTIVATION

Since the arrival of sequenced genomes, identifying genes and their functions has become an important challenge to researchers. While the definition of gene function can be ambiguous, for the purposes of this paper knowing a gene's function means knowing what product it codes for (e.g. proteins), the biochemical activities of its products, and what pathways and cellular processes it is involved in. Gene function has an inherently important predictive value; knowledge of the function of a gene can predict observable phenotypes. Finding conserved gene functions between species facilitates an understanding of their evolutionary relationship. In addition, many diseases have strong genetic components, and the information provided from characterized genes could help as a predictive tool in medicine as well as in disease targeting and prevention.

The budding yeast *S. cerevisiae* is one of the main model organisms researchers utilize in hopes of creating an encyclopedia of the genome. As of March 2007, there are still 1253 uncharacterized yeast genes listed on the *S. cerevisiae* genome database (SGD), which is nearly 21% of all known yeast genes[1]. Furthermore, the rate of gene characterization over the past several years has been noticeably slow moving (see Figure 1). Therefore, there continues to be a demand for analyzing gene function in yeast. Genome-wide (genomic) techniques offer a systematic, high-throughput approach to determining both the functions of genes and the regulatory networks that modulate their activity.

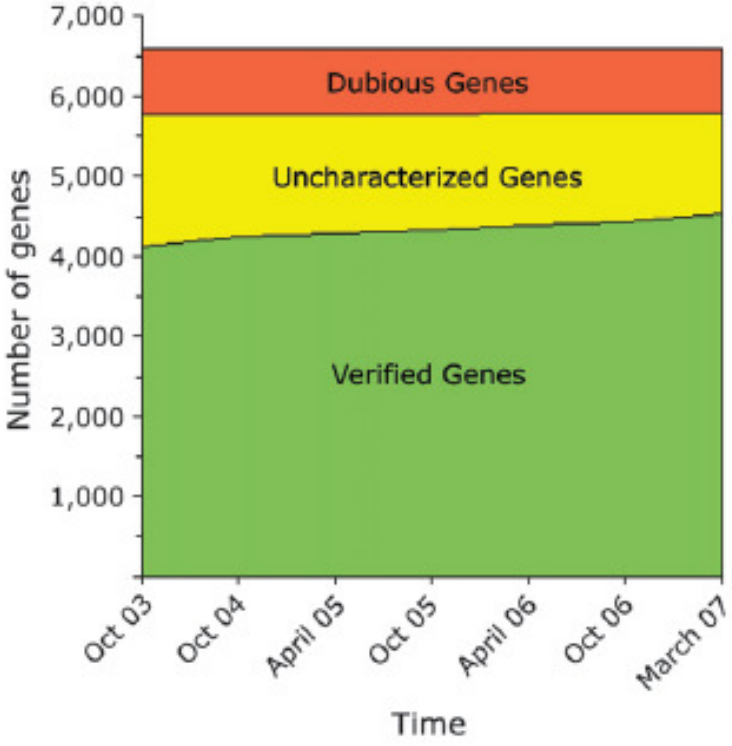


Figure 1: Distribution of genes with functions classified as "Dubious", "Uncharacterized", or "Verified" by SGD since October 2003. Notice the interface between verified and uncharacterized genes has a flattened slope.[1]

Along these lines, our lab has begun working with a yeast overexpression collection in order to functionally characterize genes. These gain-of-function mutants express a gene at a level much higher than normal, creating large amounts of mRNA. Typically, overexpression is achieved by coupling a gene to a strong promoter sequence that upregulates expression and making multiple copies of the gene in a cell by using multicopy plasmids (explained later) or by integrated copies into the genome. We use an overexpression plasmid collection developed by Gelperin et al[2]. Yeast cells are transformed with plasmids carrying a specific gene's open-reading frame (coding sequence of a gene) that is regulated by the powerful GAL1 promoter. With the plasmid

collection, a pool of yeast cells can be transformed such that each cell is regulated to overexpress only one gene.

This collection is a valuable, high-throughput tool because it has an overexpression plasmid for over 5,800 yeast genes and offers a flexible system for gain-of-function analyzes[2]. Specifically, our lab plans to determine the function of the overexpressed genes by treating the pool with a variety of lethal stimuli and screening for cells whose singular overexpressed gene allowed survivability. Those specific cells that survive can then be analyzed genetically using DNA microarray technology in order to determine which gene is overexpressed and what function it has that allowed the cell to survive. This paper provides a background for overexpression screening in yeast, as well as results from experiments aimed at optimizing our current overexpression collection and preliminary experiments comparing gene activation and drop out before and after galactose induction.

YEAST AS A MODEL SYSTEM

The yeast *Saccharomyces cerevisiae* is a unicellular eukaryote with ideal properties for biological research (see Figure 2). Yeast offers scientists advantages such as rapid growth, discrete cells, ability for replica plating, ease in mutant isolation, well-defined genetics, and an accommodating DNA transformation system (explained later)[3]. In addition, yeast is a cheap, nonpathogenic, and commercially available biological model.

S. cerevisiae has the unique ability to grow in both haploid and diploid states, which allows for a wide range of genetic analyses aimed at identifying new genes and determining the function of known genes. Recessive mutations, whose phenotypes are typically masked by dominant alleles in heterozygous diploids, can be detected more easily in haploids. Meanwhile, yeast diploid strains can undergo simple genetic tests, for example complementation tests[4]. Complementation occurs when two recessive mutant strains are crossed together and have a wild-type phenotype, implying that the two mutations do not affect the same gene.

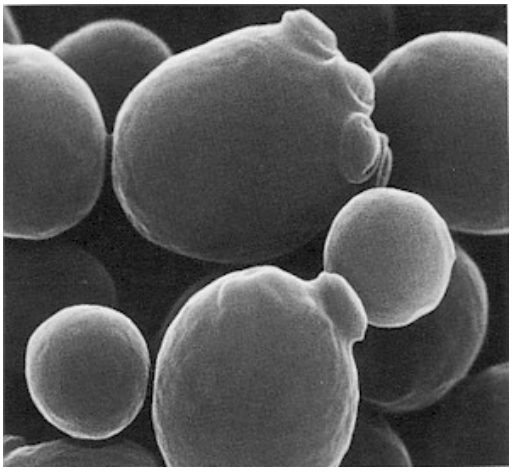


Figure 2: Picture of the budding yeast *S. cerevisiae* (http://www.genomenewsnetwork.org)

The close homology between yeast and humans adds to its practicality as a model organism. Many human genes related to disease have orthologues in yeast[4], meaning the same genes have been conserved through time in both yeast and humans while maintaining the same or similar function. More specifically, at least 31% of proteins encoded by yeast genes have human homologs, while nearly 50% of human genes implicated in heritable diseases have yeast homologs[5]. Additionally, there is high conservation of cell cycle, signal transduction, metabolic, and regulatory mechanisms

between yeast and most eukaryotic organisms[4]. The simplicity of yeast genetics, combined with its close homology to higher-order eukaryotes, allows for mammalian genes to be introduced into yeast and analyzed for their function[6].

Because of these suitable characteristics, *S. cerevisiae* has been at the forefront of many scientific breakthroughs. It was the first eukaryote to have its genome completely sequenced, to undergo transformation by plasmids, and to have gene knockouts constructed[4]. With its entire genome sequenced, yeast has shifted molecular genetics research to a systems-level approach of functionally characterizing parts of the genome. Now, much of yeast research aims to analyze of genome-scale data using microarray technology, define gene and protein interaction networks, and characterize gene function on an individual and global scale[7].

TRANSFORMATION

S. cerevisiae is readily accessible to genetic modification through DNA transformation. Transformation is the uptake and expression of foreign DNA in a living cell. This is accomplished using DNA molecules called vectors, which carry regulatory sequences and a fragment of foreign DNA called an insert.

Yeast transformation studies utilize plasmids as vectors. Plasmids are small, circular DNA molecules that are able to carry copies, or clones, of specific genes along with specific regulatory sequences that allow for the conditional expression of the clone within the yeast cell. Often they may be multicopy plasmids, meaning there are multiple

copies of a plasmid within a cell. Plasmids also carry: an origin of replication, allowing for replication of the plasmid independent of the cell's DNA replication; a selectable marker, allowing for the selection of transformed cells against wild type cells; and a reporter gene, allowing for detection of plasmid expression. With respect to the plasmid library used in this paper, the plasmid BG1805 carries the GAL1 promoter for regulating expression of the insert – a yeast ORF (open reading frame), the URA3 gene for selection, and a fusion protein tag as a reporter (see Figure 3). These aspects of the plasmid are described below.

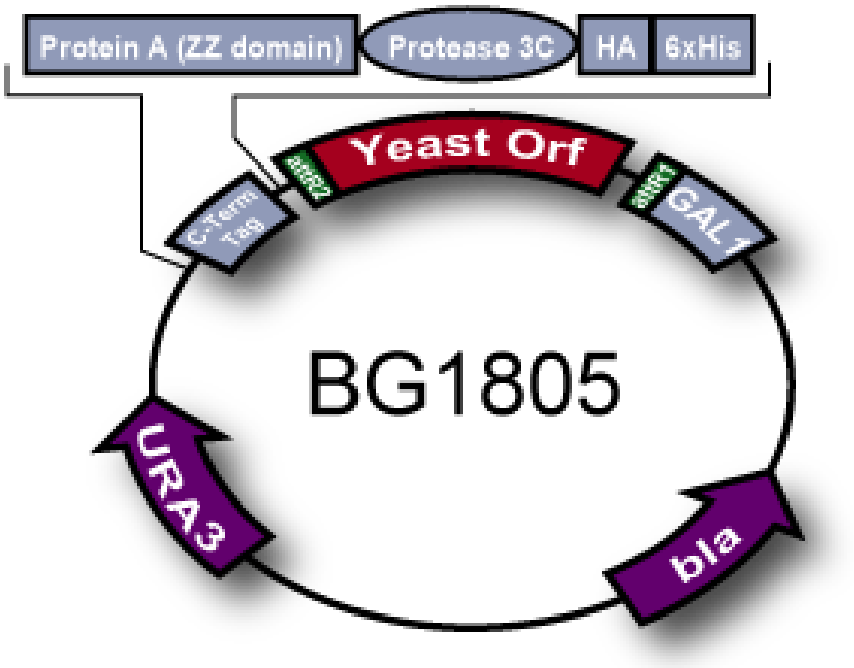


Figure 3: Schematic map for vector BG1805, the 2μ (micron) multicopy plasmid used to create the ORF collection used in this paper. Vector includes URA3 for selection, GAL1 promoter for constitutive expression of the yeast ORF in galactose media, yeast ORF insertion site surrounded by attB1 and attB2 sequences, and a fusion protein tag on the C-terminal end of the ORF protein[2].

The regulatory sequence upstream of the insertion site allows for specific expression of the DNA fragment inserted on a plasmid. The GAL1 promoter is a powerful promoter sequence in that, when activated, it strongly activates expression of

the downstream genes. It is repressed in glucose media, meaning the genomic insert downstream of it is not expressed when the yeast cell is growing in glucose. If the yeast is switched to galactose media, the galactose sugar activates the GAL1 promoter, leading to constitutive (constant) transcription of the insert. Thus, the GAL1 promoter strictly regulates expression of the DNA fragment for each plasmid based on the presence of either glucose or galactose in the media.

Cells containing the plasmid are often selected for with nutritional markers, such as the URA3 gene that codes for uracil, but they can also be selected for with drug and color markers. Selection is an important aspect of using a plasmid library when trying to isolate the transformed cells of interest among a population of yeast either lacking the plasmid or carrying different plasmids. In the case of using URA3 for selection, the yeast cells are ura3 mutant strains, meaning the cells' genome carries a mutant ura3 gene that cannot make uracil, while the plasmid carries the wild type URA3 gene. Yeast cells are grown in the absence of uracil, an essential amino acid that ura3 mutants cannot produce. Thus, only yeast cells carrying the plasmid with the URA3 gene are selected for and will survive.

Reporter genes are necessary since cells that carry the plasmid of interest may or may not be expressing the genes it carries. Common reporter genes express a visually detectable protein product, such as a fluorescent protein. Another way to report gene expression is through fusion proteins. Fusion proteins are two or more proteins connected together by having the genes that code for them expressed directly in-frame to one

another. For plasmids, the stop codon from the DNA fragment insert is removed and a second gene is added in-frame and downstream to the insertion site on the vector (see Figure 3 above). Thus, the fusion gene expresses the protein coded by the insert and the protein coded by the sequence downstream of the insert together, making a fusion protein. In the experiments for this paper, a polyhistidine tag is fused to the C-terminal end of the inserted ORF, creating a single polypeptide that can be detected.

Since yeast have the ability to replicate artificial, extrachromosomal DNA and have a naturally high rate of homologous recombination (explained below), the plasmid carrying the transgene may be self-replicating inside the yeast cell or integrate into the yeast genome by way of homologous recombination (see Figure 4). Homologous recombination is the exchange of DNA between to strands of similar sequence; it involves the alignment of the similar sequences and a crossover of DNA between the two strands. Thus, while homologous recombination is a natural event, it is used as a molecular technique in integrating plasmids. The plasmid BG1805 that is used to transform yeast cells in our experiments is self-replicating and does not integrate.

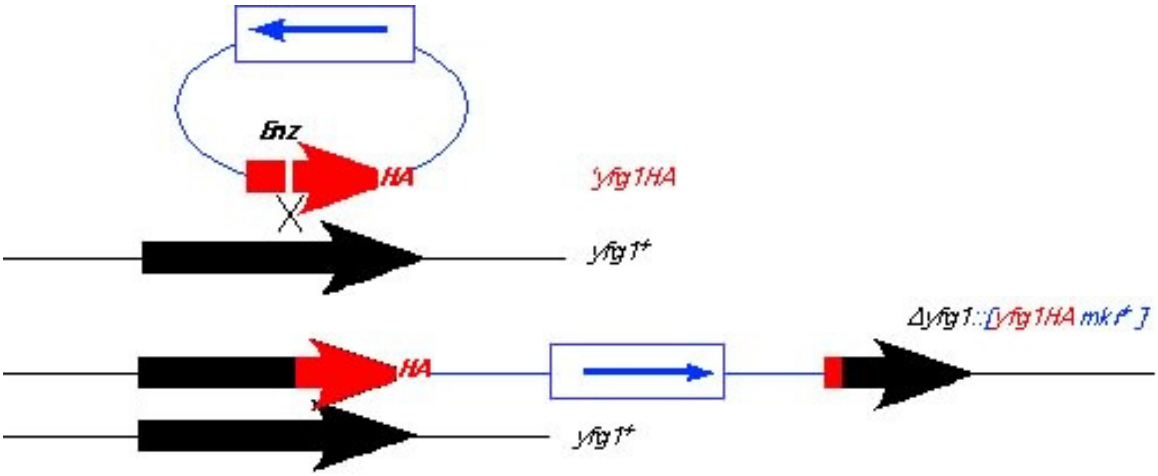


Figure 4: Example of plasmid integration into the genome by way of homologous recombination (http://www-rcf.usc.edu/~forsburg/).

CONSTRUCTING THE PLASMID

To investigate the function of genes in a controlled setting, current studies depend on precisely cloned copies of genes that allow for easy protein expression. Open reading frames (ORFs) of only the coding sequence are inserted into a plasmid downstream to a promoter sequence and upstream to a reporter gene. Currently, large ORF clone collections are made using site-specific recombination-based cloning technology[8,9]. While there are several known systems of site-specific recombination[10], the expression vectors used in this paper were derived from Invitrogen's Gateway cloning system.

Gateway cloning uses homologous recombination between *att* sequences to transfer ORFs from one plasmid to another. Two distinct versions of the *att* recombination sequence surround the insertion site, such that the upstream *att* site cannot recombine with the downstream *att* site[9]. Thus, DNA can be cloned into a vector with one proper orientation. Two recombination reactions occur to make a desired vector expressing a gene of interest (see Figure 5). ORFs with flanking *att* sites are first

recombined into non-expressing vectors, or entry clones, after which the entry clone can then recombine with a variety of expression vectors (plasmids with different promoters upstream to the insertion site). While Gateway is standardized for high throughput cloning, insensitive to DNA concentrations, and allows for the expression of ORF-encoded proteins with either N-terminal or C-terminal tags, there is reduced efficiency with larger ORF fragments[8].

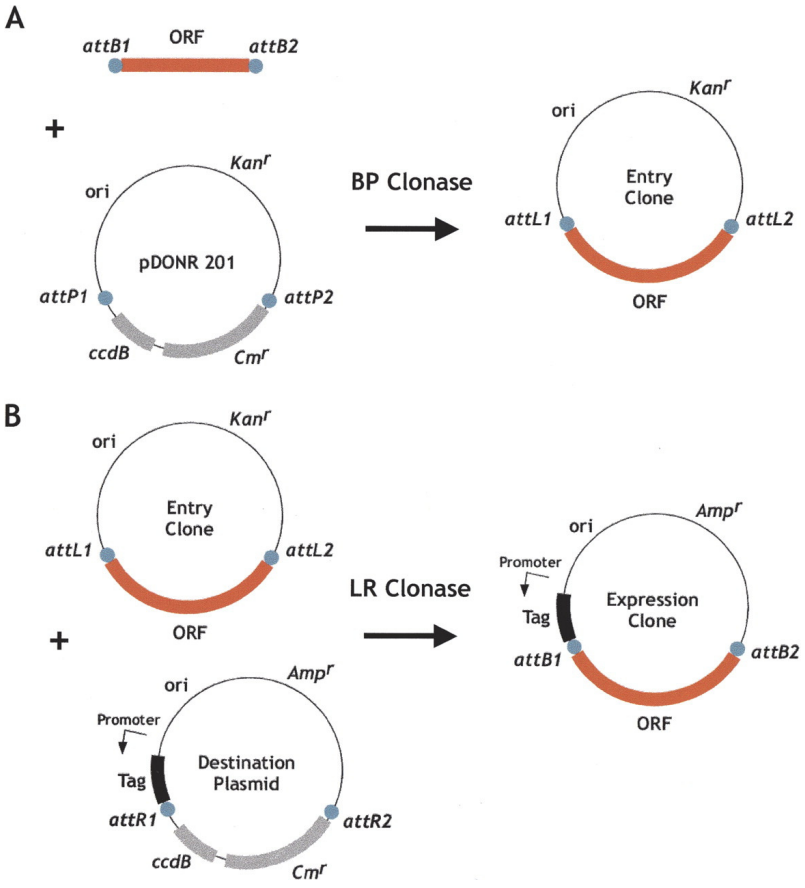


Figure 5: Overview of the Gateway site-specific recombination cloning system. (*A*) Cloning of ORF *attB*-PCR products by Gateway BP Clonase-mediated recombination. The blue circles represent *att* recombination sites. (*B*) Transfer of ORF coding sequences from the Entry vector to create an expression clone by Gateway LR Clonase-mediated recombination[8].

DNA MICROARRAYS

The development of DNA microarrays opened the door to the high-throughput, systematic genetic analyses that define functional genomics[16]. A DNA microarray is a high-density array of synthesized oligonucleotides (small single strands of DNA) printed onto a glass slide. Each oligonucleotide is one spot on the array and represents any sequence of the genome, from transcription binding domains, to intergenic regions (non-coding sequences), to ORFs (coding sequences). The key is that the oligonucleotides are complementary to the target sequence, allowing for complementary binding between sample DNA and the array. They can be used to measure relative levels of DNA in two populations (e.g. cells in glucose versus cells in galactose) of nucleic acid. Each group of DNA or cDNA is labeled with a fluorescent dye (often Cy3 or Cy5) and simultaneously hybridized to the array of oligonucleotides. The fluorescent intensity of each spot on the array is measured to give a ratio of Cy3 to Cy5 signals, giving a quantitative indication of the relative abundance of each target DNA sequence in the two populations. Green spots indicate the population tagged with Cy3 has relatively more abundance of the DNA associated with that oligonucleotide sequence, red spots the same for the Cy5 tagged population, and yellow spots indicate a relatively equal abundance of the target DNA in both populations.

DNA microarrays can be used to measure target DNA samples for a variety of purposes (see Figure 6). In this paper, DNA microarrays with oligonucleotides representing every coding and non-coding sequence in the yeast genome are hybridized with genomic DNA to determine differences in gene content. This will accurately

identify the presence or absence of overexpressed genes in the yeast cells grown in galactose (overexpression activated) compared to the same cells grown in glucose (overexpression repressed). In other words, the microarrays determine which overexpressed strains are present and in how much abundance between two pools of cells grown in glucose and galactose. This demonstrates how DNA microarrays have the ability to analyze thousands of DNA samples simultaneously by hybridization-based assay.

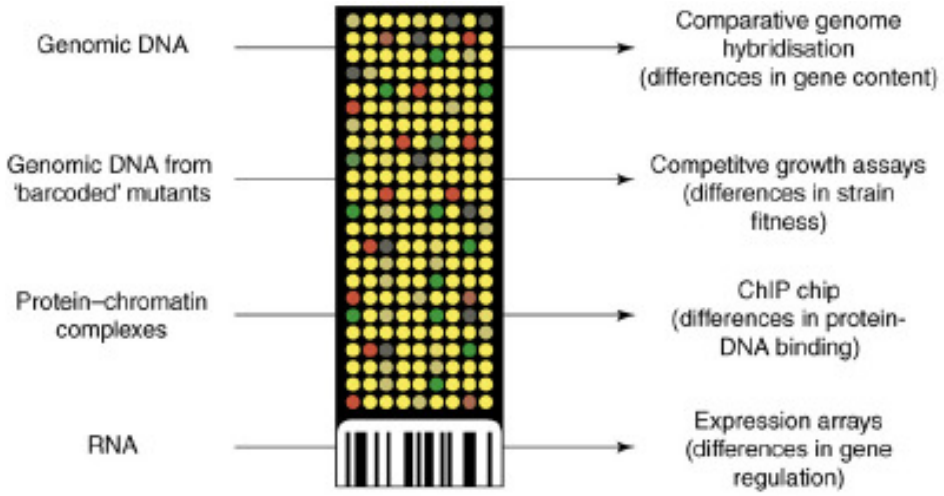


Figure 6: Representation of the whole-genome investigations enabled by DNA microarrays. To the left are the specific inputs, to the right are the genetic aspects being measured[17].

OVEREXPRESSION SCREENING

Overexpression screening looks at how increased expression of genes affects viability of cells in certain environments, allowing for easy selection among a pool of gain-of-function mutants. Overexpression screens are important because the increasing the dosage of a gene confers a mutant phenotype often different than loss-of-function mutants of the same gene[13]. The gene of interest can be readily identified in

overexpression screens by isolating the plasmid from the yeast cells, partially sequencing the insertion site, and comparing the product to the genome[4]. Additionally, deletion screens in haploids cannot examine essential genes because the mutant cells will be unviable, while overexpression screens can target both essential and non-essential genes.

Along these lines, overexpression of a single gene can either rescue viability in a lethal environment or compromise cell growth. The latter case is called dosage lethality screening, and previous research suggests that approximately 15% of all yeast genes, or 769 genes, conferred toxicity (i.e. limited cell growth) when overexpressed in normal conditions[11,12]. The former case is called multicopy suppression screening, or gene dosage resistance screening, and defines the strategy our lab hopes to use with the ORF plasmid collection created by Gelperin et al. In other words, the plasmid collection will be used to screen for genes that confer resistance to growth inhibitors (sometimes lethal) when overexpressed. This is based on the idea that the gain in a certain biochemical or cellular function due to the singular overexpressed gene increases the tolerance of the cell to higher levels of the inhibitory stimuli than wild type cells[18].

MORF COLLECTION

Gelperin et al. developed what they termed a "moveable ORF (MORF) library" of 5854 yeast expression plasmids, with each plasmid expressing a sequence-verified ORF under regulated control by the GAL1 promoter sequence. It is considered "moveable" since the ORFs can be moved in and out of various expression vectors. Previous plasmid collections of this type have not had as much gene coverage due to mutations during

cloning, and not as much protein coverage because the fusion proteins were attached to the N terminus of the insert, which is known to interfere with proteins in the secretory pathway. This is a significant issue when attempting to systematically target the genome because as many as 20%–30% of eukaryotic proteins have been estimated to be membrane or secreted proteins[2]. However, one potential concern raised with the galactose-induced overexpression is the known galactose-induced toxicity of yeast cells[14]. This may be a limiting factor during yeast cell growth. The new collection is based on the 2005 annotation of the yeast genome and is made with high-efficiency and high fidelity Gateway cloning procedures, providing the most complete collection of ORFs available for any organism. The fusion protein is attached to the C-terminal end of the insert, allowing for efficient purification of all ORF products including transmembrane and secreted proteins. Therefore, Gelperin's MORF collection consists of a library of transformed yeast strains expressing the cloned ORFs as a C-terminal ORF fusion protein under the GAL1 promoter system.

AIM OF THIS STUDY

My role in this study was two-fold – troubleshooting the overexpression screen and preliminary analysis of gene functional enrichment after galactose induction.

AIM 1: First, I tried troubleshooting the initial steps of establishing an overexpression screening system so that the MORF collection can be utilized as a tool to

characterize genes of unknown function. Specifically, there are four main procedural steps that were optimized:

The first step is galactose induction of the yeast cells. This is accomplished by inoculating the pool of transformed yeast cells into glucose media, in which the GAL1 promoter is repressed, transferring the cells to raffinose media, an intermediate stage, and finally to galactose media, in which the GAL1 promoter is activated and the ORFs are overexpressed. The yeast should grow in each media, with slowed growth in galactose due to it being a less desirable carbon source and activating overexpression of over 700 genes that inhibit cell growth for as many mutant yeast strains in the pool[11].

After galactose induction, the second step involves preparation of genomic and plasmid DNA from the cells collected from each media. With genomic and plasmid DNA isolated from the yeast cells, the third step requires amplification of the ORF insert by using *att*B specific primers in PCR (explained in Materials and Methods section). Amplification of the ORF inserts allows for hybridization to the DNA microarray, the fourth step. Again, the microarray slides consist of oligonucleotide spots representing every gene and intergenic region of the yeast genome. I describe my attempts to troubleshoot and optimize each of these steps. Once all steps can be readily performed with accuracy and little error, the pool of transformed yeast can then be subjected to growth inhibitory stimuli and undergo overexpression screening.

AIM 2: A second aim was to compare transformed yeast cells grown in raffinose and galactose by DNA microarray analysis in order to determine which overexpression strains had enhanced survivability and which overexpression strains limited growth when induced. DNA from the pool of overexpression yeast strains was taken after growing in glucose, in raffinose, and 8 hours of galactose induction. The genomic and plasmid DNA was isolated and purified, then the ORF inserts were amplified using PCR. Then, the copies of the ORF inserts were hybridized to DNA microarrays. The first slide compared glucose versus raffinose yeast plasmid DNA; ideally the oligonucleotide spots should all be yellow, indicating that the relative abundance of each plasmid stays the same in the pool in glucose and raffinose. The second slide compared glucose versus 8 hours galactose; in this case after 8 hours of induction certain ORFs may confer enhanced survivability or slowed growth, leading to a change in the relative abundance of that plasmid within the pool of yeast strains. By comparing relative abundance of raffinose to glucose and galactose to glucose, our lab has the information to directly compare raffinose to galactose. Each spot on the array is known to represent a particular yeast ORF, and through analysis we can compare the top 5% and bottom 5% spot intensities for both slides, which in turn indicate which plasmids were selected against or selected for. The list of genes in the top and bottom percentiles can then be run in functional enrichment programs that characterize relationships among the inputted genes.

II. METHODS

GALACTOSE INDUCTION OF MOVEABLE ORF COLLECTION

The protocol used in these experiments was taken from OpenBiosystems yeast ORF collection manual[2]. First, inoculate 20-mL 2% SD-URA media with 200-uL of yeast overexpression pool for overnight growth (12 hours) in a 30°C shaking incubator. Then, inoculate 20-mL –URA, 2% raffinose with 0.8-mL of the SD-URA yeast growth, creating a 1:25 dilution, and grow for 12 hours in a 30°C shaking incubator. After, dilute with –URA, 2% raffinose media for an OD = 0.3. Grow at 30°C, shaking, until OD = 0.8 (~ 6 hours), and add 3x –URA, 6% galactose for a final concentration of –URA, 2% galactose (1:2 dilution). Grow for 12 hours in a 30°C shaking incubator and harvest galactose induced yeast cells for DNA preparation. Cells can be spun down and stored as a pellet at –20°C.

To make 20-mL 2% synthetic dextrose (SD)-URA media:

2-mL 20% glucose
2-mL 10x yeast nitrogen base, without amino acids
2-mL 10x –URA dropout supplement
14-mL ddH20

For –URA, 2% raffinose and –URA, 2% galactose media, include the corresponding sugar instead of glucose.

YEAST DNA PREP: ETHANOL PRECIPITATION

The initial DNA prep used for isolated genomic and plasmid DNA from the yeast cells came from the supplemental section of the Current Protocols manual (1989). The yeast cell pellet is thawed at room temperature and resuspended with 0.5-mL of sorbitol solution. Adding 50-uL of 2-3mg zymolyase enzyme plus 50-uL of 0.28M β-

mercaptoethanol and incubating the cells at 37oC for 1 hour in a shaking incubator turn the yeast cells into spheroplasts. This step degrades the yeast cell wall and leaves a spheroplast ready for lysis. The spheroplasts are spun down into a pellet and resuspended with 0.5-mL Tris/EDTA solution. Then, the spheroplasts undergo lysis by adding 50-uL of 10% SDS and incubating at 65^{o}C for 20 minutes. The proteins and cell debris are precipitated by adding 200-uL of 5 M potassium acetate, setting the lysed cells on ice for at least 30 minutes, and then microcentrifuging for 3 minutes at room temperature. Next, precipitation the genomic and plasmid DNA takes place by taking the clear supernatant from the white protein precipitate and adding 95% ethanol. The precipitated DNA should look like threads of fiber. The DNA is microcentrifuged for 10 seconds at room temperature, the supernatant is aspirated, and the resulting DNA pellet is partially dried with a speedvac for 5 minutes. Subsequently, 300-uL of TE buffer is added and the DNA pellet is dissolved either overnight or by incubating at 65^{o}C for 10 minutes and frequently finger-flicking the tube. Once the DNA is dissolved into the TE buffer solution, the remaining RNA is degraded with the addition of 5-uL of 1mg/mL RNase A and ensuing incubation for 1 hour at 37^{o}C. Later, 0.5-mL of 100% isopropanol is added and the contents are gently mixed until the DNA precipitates into a single clump. Should the DNA not precipitate, the tube is spun down in a microcentrifuge for 10 minutes at full speed at room temperature, the DNA pellet washed with 70% ethanol twice, and then the pellet dried in a speedvac for 5 minutes. Finally, the dried DNA pellet is completely dissolved in 125-mL of TE buffer.

YEAST DNA PREP: LITHIUM-CHLORIDE BEADING

In later experiments, the Li-Cl beading DNA preparation was used to test for comparison to the aforementioned DNA prep procedure when isolating DNA from single yeast colonies. This is Promega's protocol initially designed for RNA preps, called "Rapid PCR Sequencing of Plasmid DNA Directly from Colonies of *S. cerevisiae.*" At first, a 1cm patch of cells, or equivalently 5-mL of inoculated media with an OD around 2, is first washed with 20mM Tris-HCl pH 7.5, 0.1 M NaCl. Then, 0.2-mL of a lithium-chloride buffer is added to resuspend the yeast cell pellet. Next, 0.2-mL of phenol:chloroform:isoamyl alcohol and 0.2-mL of 0.45-0.50mm glass beads are added and the mixture is vortexed in order to lyse the yeast cells and precipitate the proteins and cell debris. After microcentrifuging for 5 minutes at full speed at room temperature, the supernatant consisting of DNA and RNA is aspirated to a new tube. The DNA is precipitated using ethanol and leaving at room temperature for 5 minutes. Then, the DNA is pelleted by microcentrifuging for 15 minutes at full speed at 4^{o}C and subsequently purified by washing with 70% ethanol. After drying the DNA pellet with a speedvac for 5 minutes, it is resuspended with 0.1-mL TE buffer. Finally, adding 1-uL of 10mg/mL RNase-A and incubating at 37^{o}C for 30 minutes degrades the RNA and leaves genomic and plasmid DNA.

PCR AMPLIFICATION

PCR is performed in a Tetrad™ 96-well thermal cycler with a heated lid. The polymerase chain reaction uses plasmid-specific primers to amplify the ORF inserts. The forward and reverse primers were designed from the *att*B sequences upstream and

downstream of the insert, respectively named attB1-F and attB2-R primers, such that only plasmid DNA will undergo amplification. For each DNA sample, the initial PCR reaction included 100ng of DNA, 8-uL of 10x $MgCl_2$, 10-uL of 10x PCR buffer, 6-uL of 100µM AttB1-F and 6-uL of 100µM AttB2-R, 2-uL of amino-allyl dUTP mix, 1-uL of 2x Taq polymerase, and 62-uL of ddH_20 for a 100-uL PCR reaction. The PCR conditions consisted of denaturation at 94^oC for 30 seconds, annealing at 55^oC for 30 seconds, and extension at 72^oC for 3 minutes for 33 cycles.

The PCR is performed using an amino-allyl dUTP mix, such that the copies of the ORF inserts are made with a dUTP mix that can be tagged with reporter dyes (Cy3 and Cy5) during DNA microarray hybridization. Regular dNTP mix does not bind to the reporter dye and thus cannot be used for fluorescent spotting of the array of oligonucleotides.

DNA MICROARRAY HYBRIDIZATION AND SCANNING

Once the PCR is complete, the DNA is eluted using QIAquick plasmid mini-prep. First add 600-uL of PB buffer to 95-uL of PCR product and centrifuge in a spin column for thirty seconds. This step binds the DNA to the membrane. Then, discard the flow-through and add 650-uL of PE buffer and centrifuge in the spin column for another thirty seconds. Discard the flow through and repeat this step only this time centrifuge for one minute. This purifies the DNA. Now, the membrane should be completely dry. Add 10-uL of 0.1M sodium bicarbonate pH 9.0, incubate at room temperature for five minutes,

and centrifuge at full speed for one minute. Check the concentration of the eluted DNA using a nanodrop.

Next, prepare the DNA for labeling in the dark. First mix Cy3 and Cy5 solutions each with 1-uL of DMSO. Then, mix one 10-uL sample of DNA with the 1-uL of Cy3 dye solution, and then mix the other 10-uL sample of DNA with the 1-uL of Cy5 dye solution. Leave the mixtures in a dark place at room temperature for one hour to allow the dye to bind to the amino-allyl dUTPs of the PCR products.

After the hour of dye coupling to eluted DNA, continue to work in the dark and mix well with 600-uL of PB buffer, centrifuge in a spin column at full speed for 10 seconds, and discard the flow through. The membranes should look colored, indicating the amino-allyl dUTPs effectively coupled with the Cy3 and Cy5 dyes. Then, add 650-uL of PE buffer, centrifuge at full speed for 10 seconds, and discard flow through. Repeat this step only this time centrifuge at full speed for one minute to completely dry the membrane. Add 18-uL of elution buffer (EB) directly onto the membrane without the pipette tip touching it, incubate at room temperature for five minutes, and centrifuge at full speed for one minute. Now, the DNA and dye have coupled and are ready to hybridize to the DNA microarray.

Before proceeding, the microarray slide needs to be post-processed. This cleans the slide and rinses the poly-L-lysine coating off of it, allowing for efficient hybridization to the oligonucleotide spots. Post-processing can be performed while the DNA samples

are coupling to the Cy3 and Cy5 dyes. First, fill a dry, clean 4-L basin with ddH_20 to ¾ volume, heat it to boiling for 2-3 minutes, and stir with a stir bar to de-gas the water. It can be covered with aluminum foil for faster heating. Next, mark the corners of the slide around the visible spots with a glass etcher to know where to put the lifter slip during hybridization since the array spots will no longer be visible after post-processing. Stack the slides that will be used for hybridization into a slide rack. In the dry hood, measure 335-mL of 1-methyl-2-pyrrolidinone into a clean, dry slide dish. It is toxic, so keep under the hood. Additionally, it should be clear and if yellow do not use. Add 5.5g of succinic anhydride to the solution and stir with a stir bar until it completely dissolves. Upon dissolve, immediately mix the solution with 15-mL of 1M sodium borate pH 8.0. Then, plunge the slide rack rapidly into the blocking solution and vigorously shake up and down for about 30 seconds. Make sure the tops of the slides are under solution at all times. Put a metal lid on the glass dish with slide rack and shake on a rotator for 15 minutes. Meanwhile, reduce the heat of the boiling water in the basin to about 95^{o}C and get rid of its stir bar. After 15 minutes of the slides rotating in the blocking solution, take the slide rack out, quickly drain excess blocking solution, and completely submerge the slide rack into the boiling water in order to remove any organic solvent. Gently swish back and forth and incubate for 90 seconds. Obtain a second slide dish and fill it with 95% ethanol to ¾ volume. Quickly transfer the slide rack from boiling water to the 95% ethanol dish and plunge the slide rack 15-20 times to mix. Quickly transfer the slide rack to the centrifuge and spin for 4 minutes at 750rpm at 25^{o}C. Finally, remove the slides from the slide rack and store in a plastic slide box.

With a post-processed slide and two eluted, dye-coupled DNA samples, hybridization can take place. Still in the dark, mix the 18-uL Cy3 labeled sample with the 18-uL Cy5 labeled sample, and then add to the solution the following: 1-uL tRNA (5ug/ul), 1-uL polyA (10ug/ul), 1-uL of 1M HEPES buffer, 8.7-uL 20x SSC, and 1.3-uL 10% SDS. Mix the solution well and spin down for a total volume of 50-uL. Next, incubate it at 100^{o}C for 2 minutes and cool down at room temperature. Spin and check to see if a RNA precipitant is there, if so only use the supernatant. Then, obtain a hybridization chamber and a lifter slip and thoroughly clean them with ethanol and kimwipes. A cleaning duster may also be used to take off any dust or apparent particles on it. Take the post-processed slide from the plastic slide box and put into the chamber with the label-side up. After, put the lifter slip on top of the glass-etched corners, with the lifter slip's two rough white strips facing down on the slide (this can be checked by using forceps to feel the rough side versus the smooth side). Add 15-uL of 3x SSC into each of the 2 holes of the hybridization chamber as well as several drops of 3x SSC on all four corners of the slide. Then, pipette the 50-uL of the DNA coupled to Cy3 and Cy5 dyes under the lifter slip onto the slide. Pipette slowly but continuously so no air bubbles form. Put the hybridization chamber together, clamp both ends, and place the chamber in a 65^{o}C water bath for 16-18 hours, keeping the slide horizontal at all times.

After 16-18 hours of hybridization, the microarray slide is washed before scanning. First, prepare two wash solutions in two separate clean, dry slide dishes. The first wash solution is 340-mL ddH20, 10-mL 20x SSC, and 1-mL 10% SDS. The second wash solution is 350-mL ddH20 and 1-mL 20x SSC. Add slide racks to each slide dish.

Then, carefully remove each array in hybridization chambers from the 65^{o}C, keeping the slides horizontal at all times. Open the chamber, remove the array, and submerge it into wash solution 1. Tilt it to gently dump off the lifter slip. Be sure to not expose the array surface to air due to drying and Cy5 dye oxidation (reducing its fluorescence). Then, put the array in the slide rack and repeat for all additional arrays. When all array slides are racked in wash solution 1, plunge the rack up and down about 20 times. Next, partially lift the slide rack out of wash solution 1 and pull out a slide by its edge, drain it briefly, and place the slide in the rack of wash solution 2. One by one, quickly transfer each slide from wash solution 1 to wash solution 2 to minimized carryover of SDS and minimize exposure to air. Once all the slides are in the slide rack in wash solution 2, plunge 20 times and centrifuge for 2 minutes at 900rpm at 22^{o}C. Finally, the slides are ready to be scanned using the GenePix 4000a microarray scanner.

In the dark, insert the slide into the scanner label-side up and closest to you. During scanning, change the "lines to average" to 2 and PMT levels for 635nm (Cy5 – red) and 532nm (Cy3 – green) channels to 600. Use the preview scan command and change the PMT levels so that the intensity ratio is approximately 1.00. After full scans, the array images are then saved and gridded using GenePix Pro software. The gridded array can then be analyzed using either Acuity software or the LAD (Longhorn Array Database). In these experiments, the analysis dealt with median of ratios calculations (explained in the Results section). The median of ratios for each spot on the array representing a gene was then ranked in percentile order on Microsoft Excel. The Results are then stored on LAD.

III. RESULTS

AIM 1: TROUBLESHOOTING

YEAST CELL GROWTH IN GALACTOSE

Transformed yeast strains underwent galactose induction up to 12 hours according to the method listed above. A steady increase in the optical density of yeast-inoculated media was evident from 6 hours of galactose induction to each hour thereafter, up to 12 hours, indicating cells were growing. The 12-hour mark was chosen due to the time needed for galactose to activate overexpression and induce novel phenotypes. Were the cells picked at 3 hours or 6 hours, there may have only been but one generation of new yeast cells, and there may be little time for the overexpressed ORFs to take affect.

DNA PREP YIELD

To check whether or not the DNA prep methods listed above properly isolated and purified DNA, the DNA prep solutions were run on a 1% agarose gel stained with ethidium bromide, as well as tested on a nanodrop. The nanodrop can determine the concentration and purity of double stranded nucleic acid in a sample solution. Meanwhile, performing gel electrophoresis on a DNA sample can show the presence of DNA fragments of different lengths and sizes – ethidium bromide intercalates into DNA and can be detected under ultraviolet (UV) light. After calculating the concentration of nucleic acid in solution using the nanodrop, 100ng of the DNA prep solution and 1.25-ul of 6x loading dye was added to ddH_20 for a final volume of 6.25-ul, which went into each lane of the gel. In addition, a 100-bp DNA ladder was used to assess the approximate lengths of the DNA fragments. If genomic DNA is present, the gel will show an intense

band high on the DNA ladder. If there is no band, or if there are multiple smaller bands (further down the DNA ladder), then the DNA prep did not isolate and purify DNA.

At first, following the ethanol precipitation procedure yielded low nucleic acid concentrations on the nanodrop and faded bands on the gel. We then changed the zymolyase concentration from 2 mg/ml to 20mg/ml in the DNA prep method. Zymolyase is an enzyme that helps destroy the yeast cell wall, creating spheroplasts that can then be lysed. Thereafter, the DNA preps yielded higher nucleic acid concentrations on the nanodrop and more intense bands on the gel, especially for the yeast pool grown in galactose (see Figure 7). An important note to consider is that the gel only shows genomic DNA bands. While the yeast contains a high-copy number of plasmids with ORF inserts (multiple copies of the plasmid per cell), there are too few to distinguish on a gel.

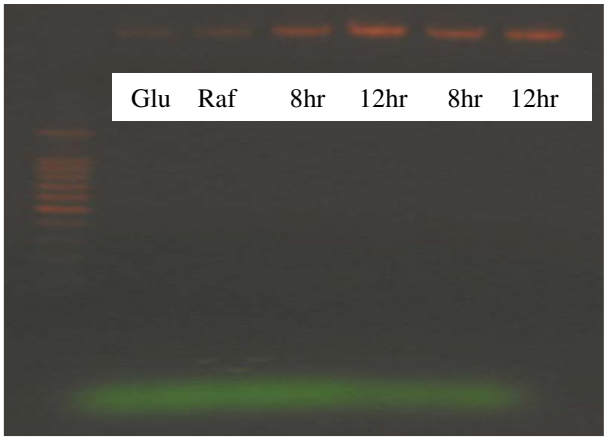


Figure 7: Ethanol precipitated DNA prep run on a 1% agarose gel. Notice the DNA preps from yeast grown in glucose and raffinose had fainter bands than the pair of DNA preps from yeast grown after 8 hours of galactose induction and pair of DNA preps from yeast grown after 12 hours of galactose induction. This illustrates the isolation of genomic DNA from the cells.

Later on, when PCR yield of yeast grown at later time points of galactose induction became a persistent issue (explained in PCR yield section), we proposed that the method of DNA prep only yielded genomic DNA and not plasmid DNA. The plasmid DNA is the essential fragment of DNA that carries the ORF inserts for amplification. Therefore, we utilized a different DNA prep method, the lithium chloride DNA prep using beads. In these experiments, overexpression strains SKO1 and RFA1 (each strain overexpresses the SKO1 and RFA1 gene respectively) were used as positive controls since previous Western blot analysis of these strains showed normal results. Additionally, rather than using a pool of overexpression strains, single colonies from the pool (individual strains) were used. As explained in the Discussion section, it was proposed that perhaps the overexpression strains were losing their plasmids in the pool but may not if grown separately. This DNA prep also yielded high nucleic acid concentrations on the nanodrop and intense bands on the gel (see Figure 8).

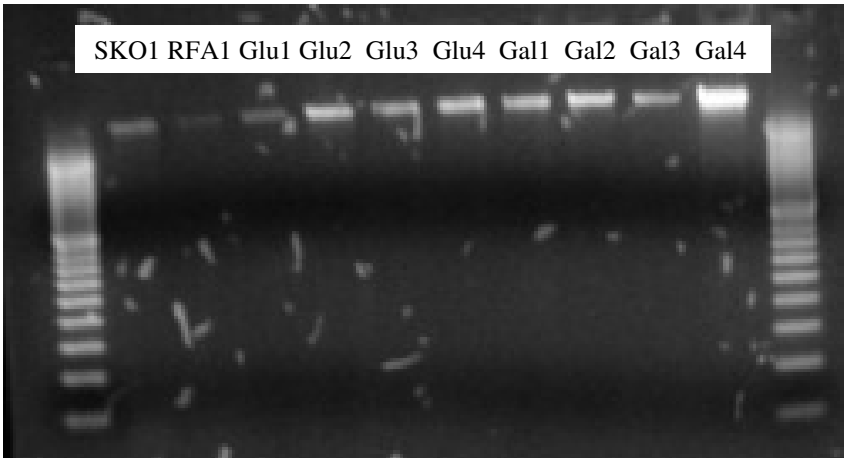


Figure 8: Lithium Chloride DNA prep run on a 1% agarose gel. Lanes include positive controls (SKO1 and RFA1) and DNA preps from four single colonies of yeast grown in glucose (Glu1 – Glu4) as well as four single colonies of yeast grown under 12 hours of galactose induction (Gal1 – Gal4).

PCR YIELD

The PCR products were run on a 1% agarose gel stained with ethidium bromide. After calculating the concentration of nucleic acid in the PCR reaction using the nanodrop, 100ng of the PCR product and 1.25-ul of 6x loading dye were added to ddH_20 for a final volume of 6.25-ul, which went into each lane of the gel. In addition, a 100-bp DNA ladder was used to assess the approximate lengths of the PCR fragments. The PCR of DNA preps from pools of overexpression strains should show a smear of bands on the gel under UV light. Each ORF insert is amplified in the polymerase chain reaction, and because each ORF fragment varies in length, the resulting PCR product solution carries a variety of amplified ORFs that in turn create a smear of bands on the gel. In addition to the smear, there will also be a band at the bottom (lowest part of the DNA ladder) that represents the primers and primer dimers from the PCR.

Each PCR run had a positive and a negative control. The negative control used ddH_20 instead of DNA, therefore there should only be a bright band at the bottom representing the unused primers and primer dimers. The positive control used a DNA sample known to work; once yeast cells grown in raffinose had been prepped for DNA and had consistent PCR smear on the gel, a raffinose DNA prep was used as a positive control. The controls tested for systematic errors in preparing the PCR reactions.

At first, the PCR reactions did not yield any product on the gel. While PCR can be optimized in many ways, primer concentration, annealing temperature, magnesium concentration, and starting DNA template were the main conditions our lab looked to

optimize. The first optimization experiment dealt with primer concentration. The PCR method called for 6-ul of 100μM of each primer, but 6-ul of 20 μM of each primer was found to yield more product (see Figure 9).

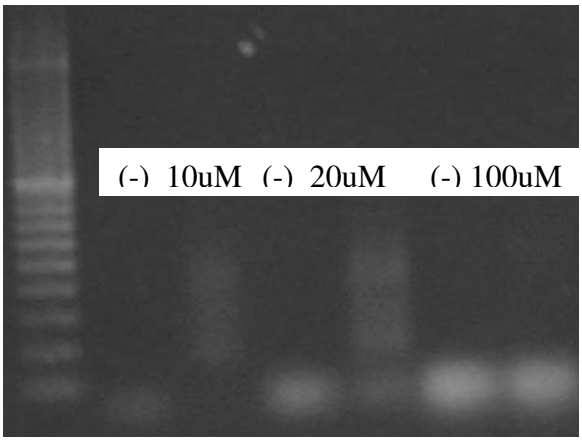


Figure 9: PCR optimization of primer concentration. Each "(-)" represents a negative control. With the strongest intensity smear, it is evident that 20μM proved to be the optimal primer concentration for the AttB1-F and AttB2-R primers.

Instead of 100ng as the starting template amount of DNA, 500ng proved to have much higher yields of PCR product (see Figure 10).

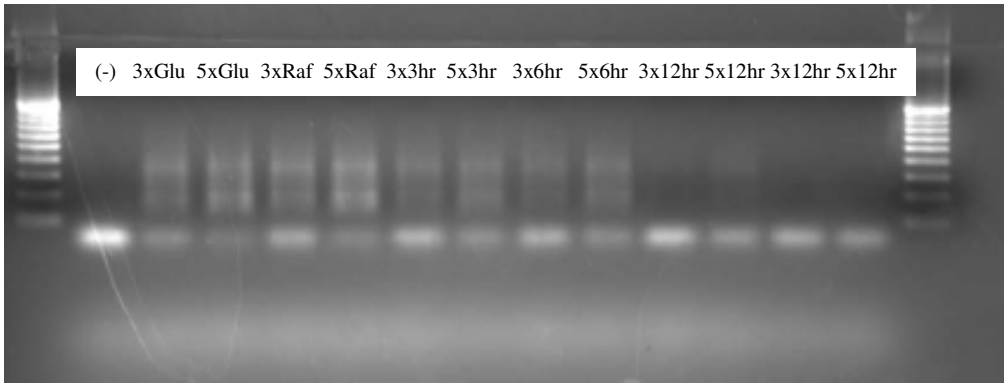


Figure 10: Comparison of PCR yields with starting DNA template of 300ng and 500ng. Each lane labeled "3x" represents a PCR with a starting template of 300ng, and each lane labeled "5x" represents a PCR with a starting template of 500ng. "3hr" "6hr" and "12hr" represent the amount of time the yeast pools underwent galactose induction. It is evident from this figure that the longer time the yells undergo induction, the lower the PCR yield of amplified ORFs.

Figure 10 also illustrates the critical issue with PCR of yeast grown in galactose: while yeast grew in galactose and yielded strong genomic DNA preps, there was a negative correlation between time under galactose induction and PCR product concentration. This proves problematic when the most informative overexpression changes should be found in yeast grown for 12 hours under galactose induction. Therefore, additional attempts at PCR optimization were carried out. An experiment varying magnesium concentration and annealing temperature showed that 1.5mM of $MgCl_2$ and an annealing temperature of 51^{o}C proved to yield higher concentrations of PCR product than the initial 2mM of $MgCl_2$ and 55^{o}C annealing temperature. Nevertheless, even the most optimal conditions of 20μM primer concentration, 500ng of starting DNA template, 1.5mM of $MgCl_2$, and 51^{o}C annealing temperature could not yield the amount of PCR product for yeast pools grown in 12 hours of galactose as high as the yields in glucose, raffinose, or earlier time points of galactose induction (see Figure 11).

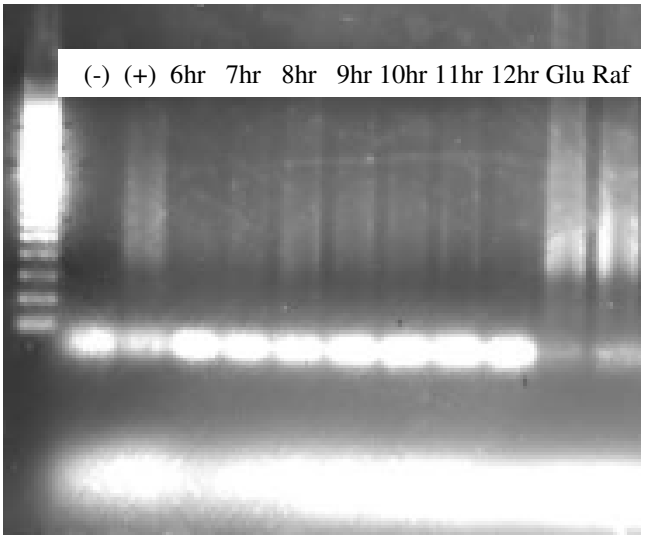


Figure 11: PCR results from an experiment comparing hourly time points of yeast grown in galactose from 6hours to 12hours (labeled *x*hr on the gel). The results clearly show the decrease in PCR yield over time, and the overall lower PCR yield of galactose-induced yeast compared to yeast grown in glucose and raffinose.

The troubleshooting then looked to the yields of DNA prep and PCR from single colonies to determine if the problem of low PCR yields in galactose was due to the some error in creating the pool of strains. The lithium chloride DNA prep was used for the single colonies (refer back to Figure 8).

Using the same optimized PCR conditions for the pooled samples, the single colonies grown on glucose plates, raffinose plates, and galactose plates showed an absence of plasmids (see Figure 12). For individual strains, the PCR should yield a single, or possibly several, distinct bands rather than a smear. (The colonies represent one strain, and each strain was only transformed with one plasmid carrying one ORF insert). However, only about 1 out of every 10 colonies grown in glucose and raffinose had a PCR yield, while none of the colonies grown in galactose had any bands. Assuming the DNA prep isolated and purified plasmid DNA along with genomic DNA, the PCRs showing no band for a colony implies that the cells have somehow lost the plasmid.

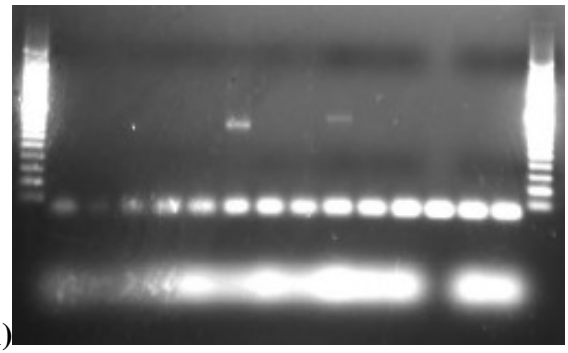
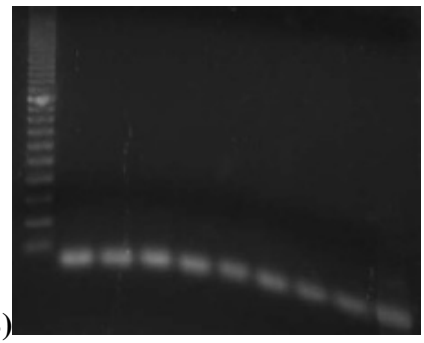
A)  **B)** 

Figure 12: PCR yields from single colonies plated from the yeast overexpression pool. A) The first 6 lanes represent yeast overexpression colonies grown on glucose plates. Notice only the last lane of glucose colonies showed a PCR band on the gel. The last 8 lanes represent yeast overexpression colonies grown on raffinose plates. Notice only the 3rd lane of the raffinose colonies showed a PCR band on the gel. B) Each of the 9 lanes represented the PCR product from a yeast overexpression colony grown on galactose plates. Notice none of the colonies yielded a PCR product on the gel.

As explained later in the Discussion section, this notion seems counterintuitive since the yeast cells carrying the plasmids of interest are selected for by growing them in –URA media. Again, the plasmids carry the URA3 gene and the yeast cells are ura3 mutants, implying that the gene coding for uracil on the genome is not functional and the gene coding for uracil on the plasmid is functional. Therefore, only cells carrying the plasmids should grow in media lacking uracil.

Because single colonies did not yield a PCR product, colony PCR of SKO1 and RFA1 *E. coli* overexpression strains was performed. This ultimately tested to see if the PCR reaction was working at all, since DNA preps are not required for colony PCR of *E. coli.* Through previous experiments, the *E. coli* were known to carry the designated plasmids. If the PCR did not yield a product, then the troubleshooting should focus primarily on the PCR steps. If the PCR does work, then the troubleshooting should focus primarily on the DNA prep steps. The colony PCR successfully worked on the *E. coli* strains (see Figure 13).

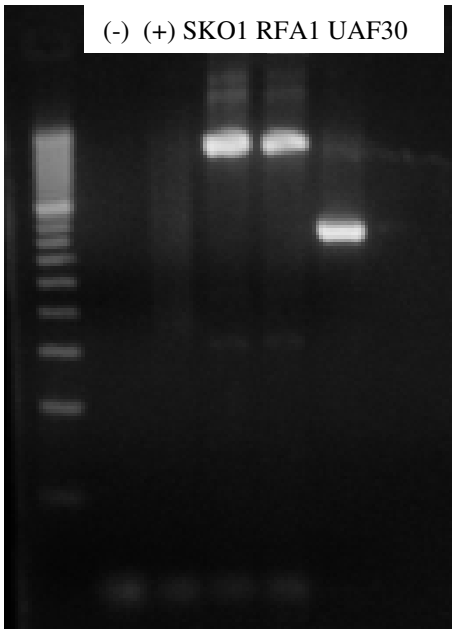


Figure 13: Colony PCR results. The (+) control lane was PCR of DNA from yeast grown in raffinose. The SKO1 and RFA1 colony PCR products correspond to the predicted lengths of each ORF fragment (~2-kb). The UAF30 was a single yeast colony strain that proved to have good PCR product. This illustrates the PCR works and the problem likely lies in the DNA prep method.

In summary, the yeast ORF collection was found to grow readily in glucose, raffinose, and galactose media. Minor changes in the DNA prep method yielded strong genomic DNA bands on agarose gels and high concentrations of nucleic acid on the nanodrop. The genomic DNA bands were more intense for yeast cells grown in galactose. Notably, the tests for DNA prep yield do not definitively show the presence of plasmid DNA, only nucleic acid concentration and purity (nanodrop) and genomic DNA presence (gel). The PCR was optimized in primer concentration, starting DNA template amount, magnesium concentration, and annealing temperature. Even so, the PCR yield for yeast grown in galactose was lower than that of yeast grown in glucose and yeast grown in raffinose. Additionally, PCR yield decreased the longer yeast grew in galactose, such that 12 hours of galactose induction yielded the lowest PCR. Single colony PCR of plated yeast showed few carried the plasmid. Colony PCR of *E. coli* overexpression strains

indicates the PCR reaction works, and that the cause of low PCR yield is likely attributed to problems in isolating and purifying plasmid DNA from the yeast cells.

AIM 2: DNA MICROARRAY ANALYSIS

Even though PCR yield for galactose-induced yeast cells was low, we continued and performed DNA microarray hybridization and analysis. Three microarrays were hybridized, scanned, and analyzed:

1. Comparison of concentrations of amplified ORF inserts (PCR product) from yeast strains grown in glucose (labeled Cy3) to yeast strains grown in raffinose (labeled Cy5).
2. Comparison of concentrations of amplified ORF inserts from yeast strains grown in glucose (labeled Cy3) to yeast strains grown under 6 hours of galactose induction (labeled Cy5).
3. Comparison of concentrations of amplified ORF inserts from yeast strains grown in glucose (labeled Cy3) to yeast strains grown under 8 hours of galactose induction (labeled Cy5).

The 6 hour and 8 hour time points were chosen due to those points having the highest PCR yield.

After using the GenePix scanner to scan the DNA microarray slides, then using the software program GenePix Pro to grid the slides such that each spot is matched up with the gene or intergenic sequence it represents on the computer, the spots were analyzed using Acuity to calculate the median of ratios for each spot on the array. The calculated

median of ratios values were then ranked in percentiles using Microsoft Excel. The spots representing intergenic sequences were ignored and discarded. The ranked percentiles were then compared across microarrays.

The median of ratios is a value found by calculating the median of pixel-by-pixel ratios of pixel intensities, with the median background subtracted for each spot on the array. The ratio is Red/Green (Cy5/Cy3), such that the higher the ratio the more red intensity on the spot, and the lower the ratio the more green intensity on the spot. Equivalent intensities (or a median of ratios close to 1) indicate a yellow spot. Because glucose was the standard Cy3 label for each microarray, a lower median of ratios value means the spot had more DNA hybridization from the DNA sample of yeast grown in glucose compared to the other sample on the array (raffinose, or galactose for 6 hours, or galactose for 8 hours).

The top 5% of the ranked percentiles represents the 5% of spots on the array with highest median of ratios values. Spots with higher median of ratios represent genes with more ORF fragments hybridized to it from the Cy5 sample compared to the Cy3 sample. Because the ORFs are inserts on the designed plasmids and the PCR uses plasmid-specific primers, a higher amount of ORF hybridizing to a spot indicates a higher amount of that particular plasmid in the pool. In other words, the relative abundance of the plasmids carrying those ORFs in the top 5% is significantly higher in the pools stained with Cy5. This means that the plasmids were competitively selected for, or enriched, in

the pool. Fewer of these plasmids were lost over time, and the cells carrying the plasmids grew better and faster than the other cells.

The bottom 5% of the ranked percentiles represents the 5% of spots on the array with lowest median of ratios values. Spots with lower median of ratios represent genes with fewer ORF fragments hybridized to it from the Cy5 sample compared to the Cy3 sample. Because the ORFs are inserts on the designed plasmids and the PCR uses plasmid-specific primers, a lower amount of ORF hybridizing to a spot indicates a lower amount of that particular plasmid in the pool. In other words, the relative abundance of the plasmids carrying those ORFs in the bottom 5% is significantly lower in the pools stained with Cy5. This means that the plasmids were competitively selected against, or lost, in the pool. More of these plasmids were lost over time, and the cells carrying the plasmids were inhibited in growth.

While knowledge of the top 5% and bottom 5% for each array shows comparisons to glucose, comparing the top and bottom 5% across microarray experiments proves even more informative. Because glucose is the Cy3 sample in each experiment, the comparisons ultimately compare Cy5 sample intensities. In other words, comparing the top 5% median of ratios values from the glucose vs. raffinose array to the top 5% median of ratios values from glucose vs. 6-hour galactose array directly compares the top 5% most abundant ORFs in the raffinose and 6-hour galactose samples. Those spots that overlap indicate the same ORF was in the top 5% abundance of all the ORF plasmids present for both raffinose and 6-hour galactose samples. Those spots in the top 5% of

raffinose ORF abundance and not in the top 5% of 6-hour galactose ORF abundance (and vice versa) indicate the plasmids either dropped out or increased in amount when the yeast pool switched from raffinose growth to galactose growth. The percentile comparisons included:

- Top 5% raffinose vs. top 5% 6-hour galactose
- Top 5% raffinose vs. top 5% 8-hour galactose
- Bottom 5% raffinose vs. bottom 5% 6-hour galactose
- Bottom 5% raffinose vs. bottom 5% 8-hour galactose

Clearly, there are other comparisons that can be made, illustrating the richness of information derived from comparing percentile rankings of median of ratios across arrays. For example, one can compare the top 5% and bottom 5% of the 6-hour galactose array to that of the 8-hour galactose array. Our data shows there to be 121 (out of 273) overlapping ORFs in the top 5% and only 16 (out of 273) overlapping ORFs in the bottom 5%. This illustrates the importance of the different time points in galactose – approximately 44% of the highest intensity spots overlapped and approximately 6% of the lowest intensity ORFs overlapped. Obtaining microarray analysis of yeast induced by galactose for 12 hours could prove to be even more informative. The results of the aforementioned percentile comparisons are illustrated in Figure 14.

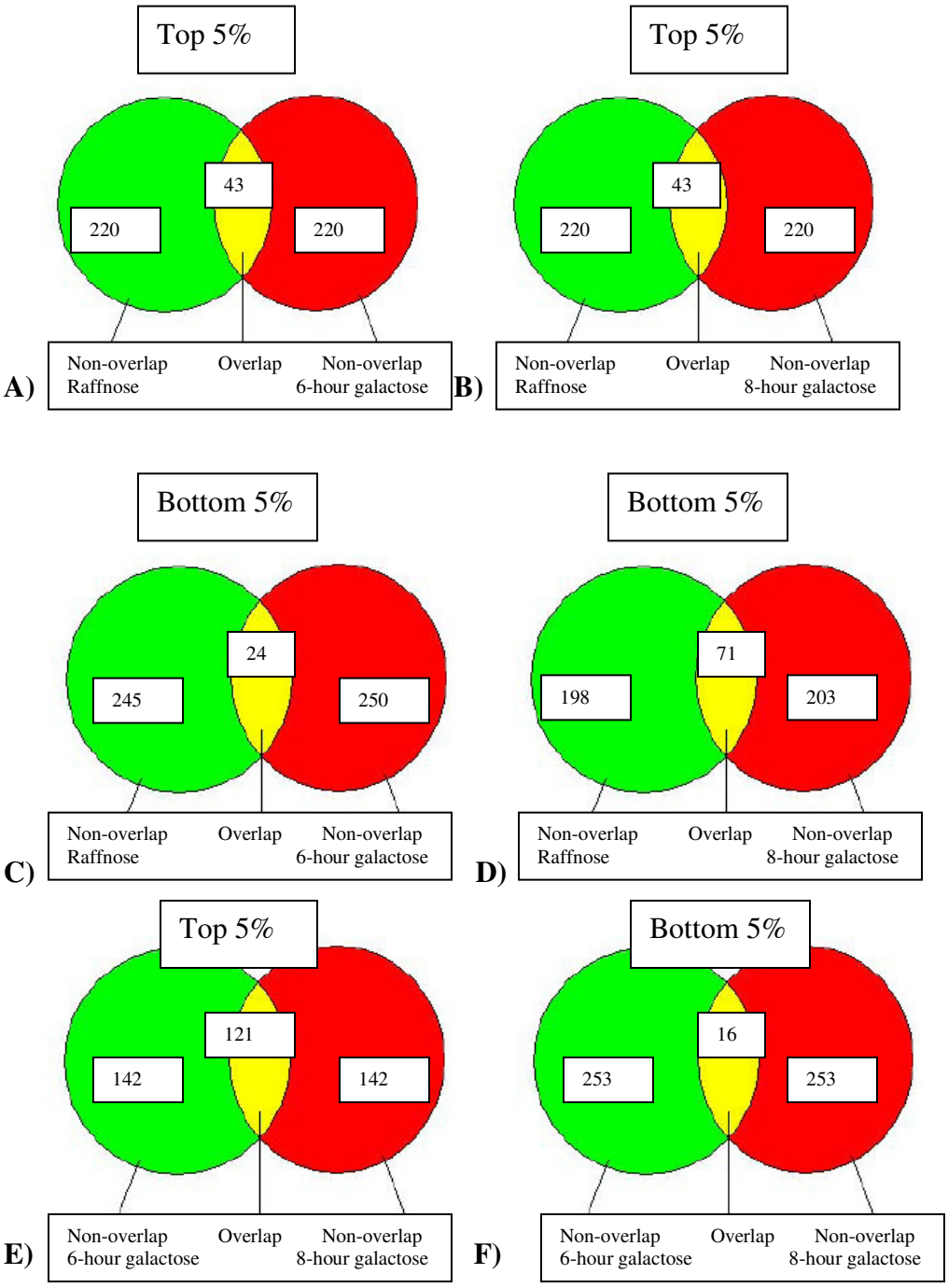


Figure 14: Percentile comparison results. A) Comparing top 5% of median of ratios scores between raffinose and 6-hour galactose, 43 out of 263 ORFs overlapped. B) Comparing top 5% of median of ratios scores between raffinose and 8-hour galactose, 43 out of 263 ORFs overlapped. Note that some of the 43 overlapping ORFs are different between the 6-hour and 8-hour galactose yeast. C) Comparing bottom 5% of median of ratios scores between raffinose and 6-hour galactose; 24 out of 269-274 ORFs overlapped. D) Comparing bottom 5% of median of ratios scores between raffinose and 8-hour galactose; 71 out of 269-274 ORFs overlapped. E) Comparing top 5% of median

of ratios scores between 6-hour galactose and 8-hour galactose; 121 out of 273 ORFs overlapped. F) Comparing bottom 5% of median of ratios scores between 6-hour galactose and 8-hour galactose; 16 out of 273 ORFs overlapped.

In addition to percentile comparisons, the sets of overlapping and non-overlapping ORFs can be characterized for functional enrichments using web-based cluster interpreters for yeast genes, such as FunSpec or the GO (gene ontology) term finder on the SGD database. This shows whether genes corresponding to certain functions were enriched in both media (overlapping ORFs) or in one media only (non-overlapping ORFs). The latter case signifies two possible scenarios: functional enrichment *found* in repressed overexpression strains (in raffinose) is *lost* when overexpression was activated (in galactose), or functional enrichment *not found* in repressed overexpression strains (in raffinose) is *gained* when overexpression was activated (in galactose). The GO term finder results are shown in Tables 1-3:

Table 1: Functional enrichments of overlapping and non-overlapping ORF sets from top 5% of raffinose vs. 6-hour galactose and raffinose vs. 8-hour galactose DNA microarray comparisons

ORF Set	Biological Process	Molecular Function	Cellular Component
Top 5% overlapping raf vs. 6-hour gal [43 ORFs]	No significant term could be found (10 of unknown ORFs biological process)	No significant term could be found (13 ORFs of unknown molecular function)	No significant term could be found (12 ORFs of unknown cellular component)
Top 5% non-overlapping raf (in raf vs. 6-hour gal) [220 ORFs]	Sulfate assimilation: 2.3% Sulfate utilization: 2.3% Response to stimulus: 21.5%	Catalytic activity: 43.4%	Cell: 88.6% Cell part: 88.6%
Top 5% non-overlapping 6-hour gal (in raf vs. 6-hour gal) [220 ORFs]	No significant term could be found (61 ORFs of unknown biological process)	No significant term could be found (94 ORFs of unknown molecular function)	No significant term could be found (52 ORFs of unknown cellular component)
Top 5% overlapping raf vs. 8-hour gal [43 ORFs]	No significant term could be found (7 ORFs of unknown biological process)	No significant term could be found (12 ORFs of unknown molecular function)	No significant term could be found (8 ORFs of unknown cellular component)
Top 5% non-overlapping raf (in raf vs. 8-hour gal) [220 ORFs]	Sulfate assimilation: 2.3% Sulfate utilization: 2.3% Response to stimulus: 21.0%	Catalytic activity: 42.9%	No significant term could be found (29 ORFs of unknown cellular component)
Top 5% non-overlapping 8-hour gal (in raf vs. 8-hour gal) [220 ORFs]	No significant term could be found (49 ORFs of unknown biological process)	No significant term could be found (85 ORFs of unknown molecular function)	No significant term could be found (34 ORFs of unknown cellular component)

Table 2: Functional enrichments of overlapping and non-overlapping ORF sets from bottom 5% of raffinose vs. 6-hour galactose and raffinose vs. 8-hour galactose DNA microarray comparisons

ORF Set	Biological process	Molecular function	Cellular Component
Bottom 5% overlapping raf vs. 6-hour gal [24 ORFs]	No significant term could be found (5 ORFs of unknown biological process)	No significant term could be found (7 ORFs of unknown molecular function)	Proton-transporting domain: 12.5%
Bottom 5% non-overlapping raf (in raf vs. 6-hour gal) [245 ORFs]	No significant term could be found (112 ORFs of unknown biological process)	Structural constituent of Ribosome: 10.7% Structural molecule activity: 12.3%	Ribosomal subunit: 10.7% Mitochondrial part: 14.3% Large ribosomal subunit: 7.0% Ribosome: 11.9% Organellar ribosome: 4.9% Mitochondrial ribosome: 4.9%
Bottom 5% non-overlapping 6-hour gal (in raf vs. 6-hour gal) [250 ORFs]	Cellular process: 78.4%	Cation antiport activity: 1.2%	Cell part: 91.6% Cell: 91.6% Intracellular: 85.2% Intracellular part: 84.0% Intracellular organelle: 70.0% Organelle: 70.0% Cytoplasm: 65.6%
Bottom 5% overlapping raf vs. 8-hour gal [71 ORFs]	No significant term could be found (29 of unknown ORFs biological process)	No significant term could be found (35 ORFs of unknown molecular function)	Mitochondrial part: 19.7%
Bottom 5% non-overlapping raf (in raf vs. 8-hour gal) [198 ORFs]	No significant term could be found (89 of unknown ORFs biological process)	Structural constituent of Ribosome: 9.6%	Ribosomal subunit: 9.6%
Bottom 5% non-overlapping 8-hour gal (in raf vs. 8-hour gal) [203 ORFs]	No significant term could be found (70 ORFs of unknown biological process)	No significant term could be found (92 ORFs of unknown molecular function)	Fungal-type cell wall: 6.5% Cell wall: 6.5% External encapsulating structure: 6.5%

Table 3: Functional enrichments of overlapping and non-overlapping ORF sets from top 5% and bottom 5% of 6-hour galactose vs. 8-hour galactose DNA microarray comparisons

ORF Set	Biological Process	Molecular Function	Cellular Component
Top 5% overlapping 6-hour gal vs. 8-hour gal [121 ORFs]	No significant term could be found (28 ORFs of unknown biological process)	No significant term could be found (41 ORFs of unknown molecular function)	No significant term could be found (24 ORFs of unknown cellular component)
Top 5% non-overlapping 6-hour gal (in raf vs. 6-hour gal) [142 ORFs]	No significant term could be found (43 ORFs of unknown biological process)	No significant term could be found (66 ORFs of unknown molecular function)	Protein-DNA complex: 2.8% Nucleosome: 2.8%
Top 5% non-overlapping 8-hour gal (in raf vs. 6-hour gal) [142 ORFs]	No significant term could be found (28 ORFs of unknown biological process)	No significant term could be found (56 ORFs of unknown molecular function)	No significant term could be found (18 ORFs of unknown cellular component)
Bottom 5% overlapping 6-hour gal vs. 8-hour gal [16 ORFs]	No significant term could be found (5 ORFs of unknown biological process)	No significant term could be found (7 ORFs of unknown molecular function)	No significant term could be found (2 ORFs of unknown cellular component)
Bottom 5% non-overlapping 6-hour gal [257 ORFs]	Cellular process: 79.5%	Ion-transmembrane transporter activity: 6.6% Substrate-specific transporter activity: 10.5% Hydrogen ion transmembrane transporter activity: 3.9% Cation transmembrane transporter activity: 5.8%	Cell part: 91.1% Cell: 91.1% Intracellular: 84.9% Intracellular part: 83.7% Intracellular organelle: 69.8% Organelle: 69.8% Cytoplasm: 65.9% Membrane: 25.2%
Bottom 5% non-overlapping 8-hour gal [257 ORFs]	No significant term could be found (61 ORFs of unknown biological process)	No significant term could be found (94 ORFs of unknown molecular function)	No significant term could be found (52 ORFs of unknown cellular component)

IV. DISCUSSION

With the sequences of human and model organisms completed, systematic genetic screens can now be performed to identify and characterize genes. Due to the well-defined genetics and ease of transformation in yeast, ORF plasmid collections have become a highly regarded genome-level development in characterizing gene functions. Using a collection representative of the genome has led not only to the characterization of individual genes, but also to a clearer framework of cellular, protein, and genomic functional networks. This project utilized a movable ORF collection containing 93% of all verified yeast ORFs at the time of completion (2005)[1]. With use of DNA microarray technology, this collection can facilitate direct linkage of the activities of many genes to a quantifiable stimulus.

Our results illustrate the ability to use this ORF collection to categorize groups of ORFs by their known and unknown functions, and in turn both infer new functions and identify new functional relationships. We induced overexpression in a pool of transformed yeast cells representative of the entire ORF collection and analyzed the patterns of the relative enrichment and loss of each particular plasmid. The changes in abundance could be do to the rate of change in the number of the high-copy plasmids in each cell, or a change in the number of each transformed cell in the pool. This was accomplished by amplifying the ORF inserts and probing the ORFs on a DNA microarray. The ORFs ranked in the top 5% and bottom 5% in relative abundance compared to the entire ORF collection were: compared to ORFs in the top 5% and

bottom 5% from other microarray tests, grouped into overlapping and non-overlapping sets of ORFs, and analyzed for related functions based on the members of each cluster (refer back to Figure 14). In other words, by monitoring relative changes in the amount of each transformed yeast strain and each overexpression plasmid, we generated sets of ORFs and classified each sets functional relationships.

Through web-based gene ontology programs we were able to distinguish the number of unknown ORFs in clusters without significant functional enrichment, and distinguish clusters of ORFs with significant functional enrichment in biological process, molecular function, and cellular components (refer back to Tables 1-3). The groups of ORFs with significant functional enrichments can be searched for ORFs of unknown function, allowing for gene functions to be inferred or identified. Meanwhile, the groups of ORFs with insignificant functional enrichments can become starting points of clusters to be further analyzed, with the new knowledge that these ORFs may have related functions.

This ORF collection in particular is flexible in its method of screening, providing limitless possibilities in future experimentation. Use of stimuli such as radiation, toxins, heat shock, and starvation as lethal screens on the pool of overexpression strains may prove advantageous in further characterizing unknown gene functions. The lethal overexpression screens can determine which overexpressed ORFs confer survivability, and with the knowledge of the physiological effects of each stimulus on the yeast cells, gene functions may be inferred. In addition, this ORF offers advantages in finding

relationships in other realms, from biochemical networks of proteins to cellular structures.

These future directions rely on reproducible experiments using this particular collection of transformed yeast pools. The other aim of this project attempted to troubleshoot the methods of screening with this ORF collection in order to limit any systematic errors and create an easy-to-use ORF collection with highly reproducible experiments. The methods include: inducing overexpression, isolating DNA, and amplifying the ORF inserts before probing DNA microarrays. While the process of switching the yeast from glucose to raffinose to galactose media yielded steady cell growth, and both DNA prep methods yielded consistent genomic DNA, there were continually low PCR yields from yeast grown in galactose combined with a decline in PCR yield from yeast grown in galactose over longer periods of time. Somehow the induced cells (cells grown in galactose) do not retain the activated overexpression plasmids from cell growth to DNA prep to PCR. This problem is critical in creating reproducible results because low PCR product leads to low hybridization of the DNA samples to the microarray, which leads to higher background noise resulting in higher variance in spot intensities. This hinders analysis as the relative abundance of each ORF in the pool is measured by the intensity of each sample hybridizing to spots on the array.

This problem was considered at each of the three stages. In the cell growth stage, we know that research has shown galactose to be toxic to yeast[14], and that nearly 15% of all ORFs inhibit growth when overexpressed[11,12]. Even so, the galactose toxicity and

inhibiting ORFs do not account for the stark drop in PCR yield in yeast grown in galactose considering the *same* amount of yeast was prepared from glucose and raffinose inoculations, nor does it account for a gradual decline in PCR yield from 6 hours to 12 hours of galactose induction. Rather, these findings only suggest yeast cells take longer to grow and that there will be fewer strains in the pool when overexpression is induced.

In the DNA prep stage, a major issue is the ambiguity of the testing done to determine the presence of plasmid DNA. The results from the nanodrop only indicate the concentration of nucleic acid – DNA strands may be degraded, or the DNA sample may only be genomic DNA and not plasmid DNA. Meanwhile, gel electrophoresis of the DNA samples assumed that the DNA prep isolated and purified both genomic and plasmid DNA because only genomic DNA bands can be seen on the gel. The plasmids are too few in number before PCR amplification to be visualized on an agarose gel.

In the PCR stage, multiple experiments resulted in evidence suggesting the problem does not come from amplification. Optimization of multiple conditions in the PCR stage did not improve PCR yield from yeast grown in galactose, and single colony PCR showed a majority of the colonies lacking any plasmid, suggesting there was either some consist problem in amplifying the ORF insert from the plasmid DNA or that the plasmid DNA was not in the DNA prep samples. Then, colony PCR of *E. coli* strains carrying plasmids from the same ORF collection yielded excellent PCR yield (refer back to Figure 13). The colony PCR method of *E. coli* was the same as all the other PCR methods on the yeast pools and yeast colonies, yet *E. coli* did not require the DNA prep

stage while yeast did. Therefore, our PCR troubleshooting indicated the PCR stage worked well, and that the plasmid DNA was not found or degraded in the post-DNA prep samples.

Our results could not conclude why only yeast grown in galactose differs in plasmid amplification compared to yeast grown in glucose and raffinose. This is the next critical step in troubleshooting this ORF collection. In theory, the low PCR yields suggest the galactose-induced yeast is either losing their plasmids at a higher rate during growth, or losing the plasmid DNA during DNA prep. The first case seems counterintuitive since the ORF collection was designed to select for transformed yeast cells by having all transformed yeast be *ura3* mutants, incorporating the URA3 gene onto the plasmid, and growing the yeast strictly in –URA media. If a cell "lost its plasmids", the cell would die due to the inability to produce uracil. One possibility is that there may be recombination between the URA3 plasmid gene and its homologous region on the genome, relaxing the selective pressure to keep the overexpression plasmids. The second case also seems counterintuitive since all harvested yeast cells – from glucose, raffinose, and galactose media – underwent DNA preparation together. Any systematic error in the DNA prep leading to loss or degradation of plasmid DNA in one sample should do the same for other samples, yet yeast grown in glucose and raffinose had consistently high PCR yields. Even so, there may be adjustments necessary in the galactose induction yeast cell growth protocol, or perhaps steps in the DNA preparation to assure higher retention of plasmid DNA.

It is evident from the wealth of information, from functional enrichment in groups of ORFs to clustering of unknown ORFs, that overexpression screens are promising tools in systematic gene characterization. Gain-of-function screens can show novel phenotypes resulting from increased gene dosages and allow analysis of essential gene dosage modifications (as compared to the numerous well-established gene deletion screens). The ORF collection used here was optimized in the galactose-induction, DNA preparation, and ORF insert amplification stages, but there continues to be issues in plasmid DNA retention and/or plasmid DNA isolation in overexpression-induced cells. Nevertheless, we have shown through DNA microarray analysis novel functional clusters of unknown ORFs based on substantial changes in the relative abundance of each ORF in the yeast pool when overexpression is induced. In addition, we found new ORF sets made up of ORFs of both known and unknown function that are characterized by significant functional enrichments. With further analysis of these functional enrichments and unknown ORF clustering, our findings may lead to the characterization of unknown gene functions and the establishment of more defined gene networks in *S. cerevisiae.*

LITERATURE CITED

1. Pena-Castillo, Lourdes, and Timothy R. Hughes. "Why are There Still 1000 Uncharacterized Yeast Genes?" Genetics 176 (2007): 7-14.
2. Gelperin, Daniel M., Michael A. White, Martha L. Wilkinson, Yoshiko Kon, Li A. Kung, Kevin J. Wise, Nelson Lopez-Hoyo, Lixia Jiang, Stacy Piccirillo, Haiyuan Yu, Mark Gerstein, Mark E. Dumont, Eric M. Phizicky, Michael Snyder, and Elizabeth J. Grayhack. "Biochemical and Genetic Analysis of the Yeast Proteome with a Movable ORF Collection." Genes & Development 19 (2005): 2816-2826. Pubmed. Yeast ORF collection used in our experiments was developed by Eric Phizicky (University of Rochester) and Mike Snyder (Yale University).

   Available through Open Biosystems:
   http://www.openbiosystems.com/GeneExpression/Yeast/ORF/
3. Broach, J. R., Jones, E. W. and Pringle, J. R. (Eds.): *The Molecular and Cellular Biology of the Yeast* Saccharomyces*, Vol. 1. Genome Dynamics, Protein Synthesis, and Energetics*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY (1991).
4. Forsburg, Susan L. "The Art and Design of Genetic Screens: Yeast." Nature Reviews Genetics 2 (2001): 659-668.
5. Hartwell, L H. "Yeast and Cancer." Bioscience Reports 24 (2004): 523-544.
6. Sherman, Fred. "Getting Started with Yeast." Methods in Enzymology 350 (2002): 3-41.
7. Dolinski, Kara, and David Botstein. "Changing Perspectives in Yeast Research Nearly a Decade After the Genome Sequence." Genome Research 15 (2005): 1611-1619.
8. Walhout, A. J., G. F. Temple, M. A. Brasch, J. L. Hartley, M. A. Lorson, S. Van Den Heuvel, and M. Vidal. "GATEWAY Recombinational Cloning: Application to the Cloning of Large Numbers of Open Reading Frames or ORFeomes." Methods in Enzymology 328 (2000): 575-592.
9. Hartley, J. L., G. F. Temple, and M. A. Brasch. "DNA Cloning Using in Vitro Site-Specific Recombination." Genome Research 10 (2000): 1788-1795.
10. Marsischky, Gerald, and Joshua LaBaer. "Many Paths to Many Clones: a Comparative Look At High-Throughput Cloning Methods." Genome Research 14 (2004): 2020-2028.
11. Sopko, Richelle, Dongqing Huang, Nicolle Preston, Gordon Chua, Balazs Papp, Kimberly Kafadar, Mike Snyder, Stephen G. Oliver, Martha Cyert, Timothy Hughes, Charles Boone, and Brenda Andrews. "Mapping Pathways and Phenotypes by Systematic Gene Overexpression." Molecular Cell 21 (2006): 319-330. Pubmed. Elsevier Ltd.
12. Boyer, Jeanne, Gwenael Badis, Cecile Fairhead, Emmanuel Talla, Florence Hantraye, Emmanuelle Fabre, Gilles Fischer, Christophe Hennequin, Romain Koszul, Ingrid Lafontaine, Odile Ozier-Kalogeropoulos, Miria Ricchetti, Guy-Franck Richard, Agnes Thierry, and Bernard Dujon. "Large-Scale Exploration of Growth Inhibition Caused by Overexpression of Genomic Fragments in Saccharomyces Cerevisiae." Genome Biology 5 (2004): r72.

13. Chua, Gordon, Quald D. Morris, Richelle Sopko, Mark D. Robinson, Owen Ryan, Esther T. Chan, Brendan J. Fray, Brenda J. Andrews, Charles Boone, and Timothy Hughes. "Identifying Transcription Factor Functions and Targets by Phenotypic Activation." Proc Natl Acad Sci USA 103 (2006): 12045-12050. Pubmed.
14. De Jongh, William A., Christoffer Bro, Simon Ostergaard, Birgitte Regenberg, Lisbeth Olsson, and Jens Nielsen. "The Roles of Galactitol, Galactose-1-Phosphate, and Phosphoglucomutase in Galactose-Induced Toxicity in Saccharomyces Cerevisiae." Biotechnology and Bioengineering (2008). Pubmed. Epub. Published ahead of print
15. Dolinski, Kara, and David Botstein. "Changing Perspectives in Yeast Research Nearly a Decade After the Genome Sequence." Genome Research 15 (2005): 1611-1619.
16. Kumar, Anuj, and Michael Snyder. "Emerging Technologies in Yeast." Nature Reviews Genetics os 2 (2001): 302-312.
17. Borneman, Anthony R., Paul J. Chambers, and Isak S. Pretorius. "Yeast Systems Biology: Modelling the Winemaker's Art." Trends in Biotechnology 25 (2007): 349-355.
18. Jones, Grace M., Jim Stalker, Sean Humphray, Anthony West, Tony Cox, Jane Rogers, Ian Dunham, and Gregory Prelich. "A Systematic Library for Comprehensive Overexpression Screens in Saccharomyces Cerevisiae." Nature Methods 5 (2008): 239-241.