

7-1-2010

# High-throughput genomic/proteomic studies : finding structure and meaning by similarity

George Sidney Davidson

Follow this and additional works at: [https://digitalrepository.unm.edu/biol\\_etds](https://digitalrepository.unm.edu/biol_etds)

---

## Recommended Citation

Davidson, George Sidney. "High-throughput genomic/proteomic studies : finding structure and meaning by similarity." (2010).  
[https://digitalrepository.unm.edu/biol\\_etds/22](https://digitalrepository.unm.edu/biol_etds/22)

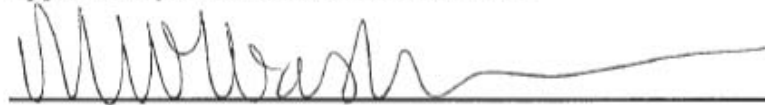
This Dissertation is brought to you for free and open access by the Electronic Theses and Dissertations at UNM Digital Repository. It has been accepted for inclusion in Biology ETDs by an authorized administrator of UNM Digital Repository. For more information, please contact [disc@unm.edu](mailto:disc@unm.edu).

George S. Davidson  
*Candidate*

Biology  
*Department*

This dissertation is approved, and it is acceptable in quality and form for publication:

*Approved by the Dissertation Committee:*

  
\_\_\_\_\_, Chairperson

  
\_\_\_\_\_

  
\_\_\_\_\_

  
\_\_\_\_\_

\_\_\_\_\_

\_\_\_\_\_

\_\_\_\_\_

\_\_\_\_\_

\_\_\_\_\_

**HIGH-THROUGHPUT GENOMIC/PROTEOMIC STUDIES,  
FINDING STRUCTURE AND MEANING BY SIMILARITY**

**BY**

**GEORGE S. DAVIDSON**

B.A. Mathematical Sciences, Rice University, 1974  
Master of Statistics, Texas A&M University, 1977

DISSERTATION

Submitted in Partial Fulfillment of the  
Requirements for the Degree of

**Doctor of Philosophy  
Biology**

The University of New Mexico  
Albuquerque, New Mexico

**July, 2010**

## Dedication

To my family, extending back to the first stirrings of carbon chemistry, and reaching forward to my children Ashley and Meredith and recently to my new granddaughter, Madison; but most especially to my wife Maureen, who has been my life companion, teacher, and confidant.

## Acknowledgements

I would like to thank Dr. Maggie Werner-Washburne, my advisor, for her support, and enthusiasm for research as a way of life, and for keeping me on track (no easy job!). I also want to acknowledge the support and kindness of my committee members, Drs. Mary Anne Nelson, Richard Cripps, and Shawn Martin; thank you for your help and insights. I'd also like to acknowledge Dr. Vicky Peck and thank her for showing me the world of microbial genomics and for first suggesting I write about microarrays for her class; I particularly thank Dr. Stuart Kim who 'got it' even before I really learned how to explain what 'it' was with respect to analyzing high-throughput data with VxInsight. I thank Dr. Cheryl Wilman who taught so many of us about leukemia and cancer research and who funded my research with her laboratory and who encouraged my addiction to opera. Dr. William (Bill) Camp deserves special thanks for encouraging my interest in science and for being such a supportive manager (and for reminding me, numerous times, that biologists like to publish in light-weight journals, like *Science*). Of course, I must acknowledge and thank Brian Wylie who developed VxInsight and Chuck Meyers who managed the Sandia National Laboratories LDRD Office, which funded the original research. I thank all of my teachers, but especially Mrs. Young, who taught me to read, and Dr. Robert Glew who showed me Biochemistry (both of which I use daily). Finally, I thank the muse because beauty and our relationship with it are always transcendent and sublime; her touch always causes my hair to stand on end; hence it is fit that Sappho should have the final say, Ἔρος δ' ἐτίναξέ μοι φρένας, ὡς ἄνεμος κατ' ὄρος δρύσιν ἐμπέτων.

**HIGH-THROUGHPUT GENOMIC/PROTEOMIC STUDIES,  
FINDING STRUCTURE AND MEANING BY SIMILARITY**

**BY**

**GEORGE S. DAVIDSON**

**ABSTRACT OF DISSERTATION**

Submitted in Partial Fulfillment of the  
Requirements for the Degree of

**Doctor of Philosophy  
Biology**

The University of New Mexico  
Albuquerque, New Mexico

**July, 2010**

# **HIGH-THROUGHPUT GENOMIC/PROTEOMIC STUDIES, FINDING STRUCTURE AND MEANING BY SIMILARITY**

by

**George S. Davidson**

B.A. Mathematical Sciences, Rice University, 1974

Master of Statistics, Texas A&M University, 1977

Ph.D. Biology, University of New Mexico, 2010

## **Abstract**

The post-genomic challenge was to develop high-throughput technologies for measuring genome scale mRNA expression levels. Analyses of these data rely on computers in an unprecedented way to make the results accessible to researchers. My research in this area enabled the first compendium of microarray experiments for a multi-cellular eukaryote, *Caenorhabditis elegans*. Prior to this research approximately 6% of the *C. elegans* genome had been studied, and little was known about global expression patterns in this organism. Here I cluster data from 553 different microarray experiments and show that the results are stable, statistically significant and highly enriched for specific biological functions. These enrichments allow identification of gene function for the majority of *C. elegans* genes. Tissue specific expression patterns are discovered suggesting the role of particular proteins in digestion, tumor suppression, protection from bacteria and from heavy metals. I report evidence that genome instability in males involves transposons, and find co-expression patterns between sperm proteins, protein kinases and phosphatases suggesting that sperm, that are transcriptionally inactive cells,

commonly use phosphorylation to regulate protein activities. My subsequent research addresses protein concentrations and interactions, beginning with a simultaneous comparison of multiple data sets to analyze *Saccharomyces cerevisiae* gene-expression (cell cycle and exit from stationary phase/ $G_0$ ) and protein-interaction studies. Here, I find that  $G_1$ -regulated genes are not co-regulated during exit from stationary phase, indicating that the cells are not synchronized. The tight clustering of other genes during exit from stationary-phase does indicate that the physiological responses during  $G_0$  exit are separable from cell-cycle events. Subsequently, I report *in vivo* proteomic research investigating population phenotypes in stationary phase cultures using the yeast Green Fluorescent Protein-fusion library (4156 strains) together with flow cytometry. Stationary phase cultures consist of dense quiescent (Q) and less dense non-quiescent (NQ) fractions. The Q-cell fraction is generally composed of daughter cells with high concentrations of proteins involved in the citric acid cycle and the electron transport chain, for example Cit1p. The NQ fraction has subpopulations of cells that can be separated by the low and high concentrations of these mitochondrial proteins, i.e., NQ cells often have double intensity peaks: a bright fraction and a much dimmer fraction, which is the case for Cit1p. The Q fraction uses oxygen 6 times as rapidly as the NQ fraction, and 1.6 times as rapidly as exponentially growing cells. NQ cells are less reproductively capable than Q cells, and show evidence of reactive oxygen species stress. These phenotypes develop as early as 20-24 hours after the diauxic shift, which is as early as we can make a differentiating measurement using fluorescence intensities. Finally, I propose a new way to analyze multidimensional flow cytometry data, which may lead to better understanding of Q/NQ cell differentiation.



# Table of Contents

Dedication.....	iii
Acknowledgements.....	iv
Abstract.....	vi
Table of Contents.....	viii
Table of Figures.....	xi
Table of Tables.....	xiii
Chapter 1: Introduction.....	1
The Third Way.....	1
Microarray and Flow Cytometry Measurements.....	2
Fluorescent Markers in Both Measurement Techniques.....	3
Examples of the Third Way.....	3
Revisiting Steps along the Third Way.....	5
Overview of the Research in Chapters 2-4.....	5
Chapter 2: A Gene Expression Map for <i>Caenorhabditis elegans</i> .....	8
Abstract.....	9
Introduction.....	9
The Experiments and Findings.....	10
References and Notes.....	26
Chapter 3: Comparative Analysis of Multiple Genome-Scale Data Sets.....	30

Abstract.....	31
Introduction.....	32
Results.....	34
Discussion.....	46
Methods.....	49
Acknowledgements.....	54
References.....	54
Web Site References.....	58
 Chapter 4: The Proteomics of Quiescent and Non-Quiescent Cell Differentiation in Yeast Stationary-Phase Cultures.....	 59
Abstract.....	60
Introduction.....	61
Materials and Methods.....	64
Results.....	70
Discussion.....	79
Acknowledgements.....	83
References.....	84
Tables.....	90
Figure Legends.....	94
Figures.....	96
Supplemental Figures.....	106
 Chapter 5: Discussion and Conclusion.....	 119
The Challenge: to Develop Analysis for High-Throughput Methods.....	119

Toward a Thorough Proteomic Analysis of GFP-Fusion Strain Flow Data.....	121
The Bigger Challenge .....	125
Conclusion .....	126
References.....	128
Appendix I – Earth Mover Distances from Cit1p to 38 Genes in Chapter 4.....	135
Appendix II – Top, Middle, and Bottom 20 Genes from Cit1p.....	143

## Table of Figures

Figure 2-1. Types of experiments and VxInsight terrain map .....	11
Figure 2-2. VxInsight map with biological groups and statistical significance.....	12
Figure 2-3. Biological categories in VxInsight mounts .....	16
Figure 2-4. Transposon mounts .....	23
Figure 3-1. $\alpha$ -Factor-arrest data set ordinated and visualized in VxInsight.....	35
Figure 3-2. VxInsight-generated ordination of exit from stationary-phase data set.....	36
Figure 3-3. Location of G <sub>1</sub> -regulated genes in two different gene-expression data sets ..	37
Figure 3-4. Location of ribosomal protein genes in two gene-expression data sets .....	39
Figure 3-5. Protein-protein interaction maps .....	41
Figure 3-6. Interactions among proteins encoded by G <sub>1</sub> -regulated genes .....	43
Figure 3-7. Protein-protein interactions between Nup116p and other proteins.....	45
Figure 4-1. EXP and SP distributions of median peak intensities .....	96
Figure 4-2. Histogram of fluorescence for Cys3p:GFP and Cit1p:GFP fusion strains ....	97
Figure 4-3. Distribution of Cit1p:GFP and DHE (ROS) fluorescence intensity .....	98
Figure 4-4. Fluorescence intensities .....	99
Figure 4-5. Reproductive capability as measured by colony forming units .....	100
Figure 4-6. Oxygen consumption measurements of s288c (prototrophic) cells .....	101
Figure 4-7. GFP protein abundance in mother:daughter pairs.....	102
Figure 4-8. Flow cytometry analysis of Cit1p:GFP fluorescence intensity .....	103
Figure 4-9. Our current model for cell differentiation in yeast cultures.....	104
Figure 4S-1. Correlation plot between our EXP data and that of Newman <i>et al.</i> .....	106
Figure 4S-2. Flow cytometry histograms for 38 separated into Q and NQ fractions .....	113

Figure 4S-3. Q/NQ ratios of median fluorescence for 38 strains with 2 peaks in SP.....	114
Figure 4S-4. MoFlo plates: upper and lower fraction.....	115
Figure 4S-5. Colony formation for NQ fractions separated by GFP and ROS.....	116
Figure 4S-6. Analysis of petite colony formation of NQ fraction .....	117
Figure 4S-7. Mother:daughter analysis.....	118
Figure 5- 1. Forward scatter and log side scatter for stationary phase Cit1p strain.....	122
Figure 5-2. Stationary phase GDPHp strain GFP and log side scatter .....	123
Figure 5-3. Stationary phase HTB1p strain GFP and log side scatter .....	123
Figure 5- 4. Gray-scale rendering of the Earth Mover Distances .....	124
Figure 5-5. VxInsight finds three subclusters within the 38 genes from Chapter 4 .....	125

## Table of Tables

Table 2- 1 Characteristics of the gene groups.....	14
Table 2-2. Heat shock induction levels for 10 genes in mount 36.....	24
Table 4-1. Most abundant proteins in EXP and SP .....	90
Table 4-2. GO process of proteins expressed 2-fold or higher in SP than in EXP .....	92
Table 4-3. Comparison of Q, cycling G1, and NQ daughters and G1 mother cells .....	93

## **Chapter 1: Introduction**

### **The Third Way**

A complex, dynamic system may be approached in two very different ways. The whole of the system may be studied to understand large scale structure and system level transformation. The microscopic study of dividing cells offers such an example: a mother cell slowly changes size and the arrangements of visible organelles, separates its visible structures into two parts and finally into two cells. This approach gives great insight into cellular life and the large-scaled cyclic, systemic processes involved. Puzzles about how the cell accomplishes these changes motivate a second, completely different approach. This second approach uses a bottom up, biochemical and mechanistic strategy. It begins with the details and assembles a theory from parts to wholes.

The top down approach is particularly weak in explaining the mechanisms involved, while the bottom up approach explains detailed interactions but faces its own difficulties in synthesizing a theory of how all of the parts ultimately make and maintain the whole cell. There is a third way; one which begins with a system level approach, but captures a large collection of fine-scaled detail, such that the bottom up approach can also be employed to better understand the mechanisms and their system-wide interactions. This third approach has enabled very rapid progress in genomics, which is the combination of large-scale sequencing with systematic computational analysis of the genomes and their interaction within and between cells (Akil, et al., 2010) .

## Microarray and Flow Cytometry Measurements

Two separate-high-throughput measurement technologies have enabled rapid progress in the field of genomics. First, microarrays, a simplified and greatly scaled-up version of Southern/northern blotting (Alwine, Kemp, & Stark, 1977; Schena & Davis, 2000; E. Southern, 2006; E. M. Southern, 1975) allowed the simultaneous, but indirect measurement of mRNA concentrations from each transcribed gene<sup>1</sup>; Chapters 2 and 3 involve microarray experiments. Second, the use of flow cytometry (Coulter, 1956; Ferry, Farr, & Hartman, 1949; Fulwyler, 1965; Gucker & Okonski, 1949; Hulett, Bonner, Barrett, & Herzenberg, 1969; Melamed, Kamensk, & Boyse, 1969; Shapiro, 2003) enables the near simultaneous measure of concentrations of fluorescently tagged proteins in tens of thousands of cells (Huh, et al., 2003), one at a time as they pass through a micro cuvette with a laser and detectors for the induced fluorescence.

Together, these approaches allow genomic and proteomic changes to be measured and ordered into groups of coordinately changing molecular concentrations. Equally importantly, these molecular groups often suggest experiments that extend our bottom up knowledge of cellular mechanisms. Because protein concentrations are particularly important, and because the concentration can be directly measured for each cell, this more challenging technique is particularly useful; Chapter 4 exploits this technology.

---

<sup>1</sup> The direct measurement of mRNA concentrations after separation by electrophoresis is referred to as a northern blot, see Alwine *et al.* (1977). While microarrays are used to measure the concentration of mRNA species, the measurement is not made directly using the original RNA; rather, the mRNA is reverse transcribed back to complementary DNA (cDNA). The concentration of cDNA molecules is measured by the microarray. Consequently, a microarray exists somewhere between Southern and northern blotting. Because they measure DNA, not RNA, they are perhaps more like a Southern blot than a northern blot, but the distinction is the subject of controversy.



## **Fluorescent Markers in Both Measurement Techniques**

In the case of the microarrays discussed in Chapters 2 and 3, the reverse transcription from sampled mRNA to cDNA included either Cy3 or Cy5 fluorescent bases. Measurement of the fluorescence intensity at each DNA probe location indicates the number of hybridized cDNA molecules and hence the concentration of the original mRNA. For flow cytometry Huh et al. (2003), created a library having strain-specific gene fusions of the wild type genes and an exogenous gene encoding green fluorescent protein (GFP), originally cloned from *Aequorea victoria* (Tsien, 1998). Consequently, the transcribed mRNA is translated into a fusion protein, the expression of which remains under the control of native transcription factors and regulation. By observing fluorescent intensity, protein concentration can be inferred by microscopy or in flow cytometry by photon counters.

## **Examples of the Third Way**

The yeast *Saccharomyces cerevisiae* was the first eukaryote to have its genome sequenced (Goffeau, et al., 1996). Subsequently, *Caenorhabditis elegans* became the first multi-cellular organism to be fully sequenced (The C. elegans Sequencing Consortium, 1998). Combining the sequence information with microarrays allowed the simultaneous measurement of mRNA concentrations for thousands of genes. These data together with new computational analyses opened the cell cycle to allow testing of both system-level hypotheses and of fine scaled interactions (DeRisi, et al., 1996; Spellman, et al., 1998).

That opportunity joined together what had previously been two communities of separate expertise. The system level research revealed global similarities in the gene expression profiles through repeated cell cycles. The detailed molecular knowledge of

cell biologists then offered a way to propose probable functions for unstudied genes. Combining the patterns of similar expression allowed groups of genes to be clustered together, then deep knowledge of a few specific genes could be used to impute related functions for the genes having similar expression profiles (which then became laboratory testable hypotheses).

Progress in genomics was greatly accelerated because microarray technology, a simplified and greatly scaled-up version of Southern blotting (DeRisi, et al., 1996; Schena, Shalon, Davis, & Brown, 1995; E. M. Southern, 1975) could be automated with inexpensive, array-printing robots (DeRisi, et al., 1996). These robots were widely replicated and large collections of array experiments became available, enabling a new kind of system level genomic analysis: the compendium approach (Hughes, et al., 2000; Kim, et al., 2001), where array data from very different experiments were combined and jointly clustered.

With the increasing number of arrays, statistical power to detect subtle patterns increased. With the wider set of experimental conditions, more of the possible cellular states were sampled. Combined together these compendium data sets allowed greater precision in gene clustering. As a result, higher quality estimates for gene function became possible. Without the combination of top down analysis using cluster by similarity of co-expression and the use of detailed, bottom up knowledge we would not have been able to impute gene functions for the majority of the genes in *C. elegans* in Chapter 2.

## **Revisiting Steps along the Third Way**

Importantly, the genomics community developed as an open, data sharing community where results and whole data sets are available for sharing from online repositories; see for example (SGD project, May 1, 2010; Tweedie, et al., 2009; WormBase web site) among many others. The open availability encouraged the development of new tools and approaches for reanalyzing previously published data. Equally importantly, different data sets could be combined for meta analyses (Werner-Washburne, et al., 2002) and further development of new algorithms and software packages (George S. Davidson, et al., 2007; Gentleman, et al., 2004; Martin, Davidson, May, Faulon, & Werner-Washburne, 2004; Reich, et al., 2006; SGD, 2010; The MathWorks, 2010; Tibshirani, Hastie, Narasimhan, & Chu, 2002; Wu, Chen, Hastie, Sobel, & Lange, 2009). Sharing and reuse of earlier data sets has been an important element of the progress in our understanding of genomics.

## **Overview of the Research in Chapters 2-4**

My research has involved the search for biologically relevant order in huge collections of high-throughput data by means of similarity measurements. This research began with the analysis of microarray data sets from *Saccharomyces cerevisiae* and *Caenorhabditis elegans* experiments. My contributions have combined statistical analyses and computer programming to work with the data and with annotation databases. While these methods were applied to microarray data (as in Chapter 2), I extended them to compare gene expression studies with protein interaction data (Chapter 3). Importantly, the research in Chapter 4 goes beyond expression data and beyond *in vitro* protein-protein interactions to study actual *in vivo* protein concentration differences

between exponentially (EXP) growing cells and cells from stationary phase (SP) cultures. This research reveals specific phenotype differences between quiescent (Q cells) and non-quiescent (NQ cells) in the stationary cultures.

Chapters 2 through 4 document the evolution of this research. My paper (Kim, et al., 2001) in Chapter 2 is an analysis of a compendium of 553 arrays taken from *C. elegans* experiments. This paper describes the first compendium expression study of a multicellular organism. Consequently, it is the first compendium study to address expression changes through developmental stages and processes.

My paper in Chapter 3 (Werner-Washburne, et al., 2002) combined expression studies with protein-protein interaction data to jointly analyze both types of experiments using the visual data analysis environment, VxInsight (G. S. Davidson, Wylie, & Boyack, 2001). This paper combined four *S. cerevisiae* high-throughput data sets: two protein interaction studies (Ito, et al., 2001; Schwikowski, Uetz, & Fields, 2000); our own stationary phase-expression data, and cell cycle expression changes following release from alpha arrest (Spellman, et al., 1998). This paper was the origin of my suite of robust methods for microarray analyses (George S. Davidson, et al., 2007), and for further research into the mechanisms of G<sub>0</sub>, and the rapid sampling equipment that enabled the study of mRNA changes in the earliest few seconds after refeeding stationary phase yeast (Allen, et al., 2006; Aragon, et al., 2005; Aragon, Quinones, Thomas, Roy, & Werner-Washburne, 2006; Aragon, et al., 2008), a time when the cells make extensive use of previously sequestered, protein-bound mRNAs, which are not detected with traditional protocols.

The sequestered mRNA and in general the imperfect correlation between protein concentrations and mRNA concentrations motivated the need to study stationary phase cells by direct, high-throughput proteomic measurements reported in Chapter 4. These experiments exploit flow cytometry (Coulter, 1956; Ferry, et al., 1949; Fulwyler, 1965; Gucker & Okonski, 1949; Hulett, et al., 1969; Melamed, et al., 1969; Shapiro, 2003). The experiments measure protein concentrations in 10-30,000 cells (observed cell by cell) across 4156 strains; where each strain has a single gene modified to express mRNA from the native gene immediately followed by continued transcription of the gene for green fluorescent protein (GFP), originally cloned from *Aequorea victoria* (Tsien, 1998). The proteomic results are verified and extended by microscopy, reproductive capacity measurements, density gradient separations followed by further flow measurements, and by metabolic measurements to reveal new information about quiescent and non-quiescent cells. As expected, compared to exponentially growing cells, both non-quiescent and quiescent cells have greater accumulations of proteins involved in the citric acid cycle and the electron transport chain. However, the quiescent cells have a much higher concentration of these proteins raising the question, are the non-quiescent cells able to respire. Direct measurements of oxygen consumption indicate that quiescent cells consume oxygen about 6 times faster than non-quiescent cells, and exponentially growing cells are using oxygen 4 times faster than non-quiescent cells.

## Chapter 2: A Gene Expression Map for *Caenorhabditis elegans*

This chapter has previously appeared in substantially the same form as: Stuart K. Kim,<sup>1</sup> Jim Lund,<sup>1</sup> Moni Kiraly,<sup>1</sup> Kyle Duke,<sup>1</sup> Min Jiang,<sup>1</sup> Joshua M. Stuart,<sup>2</sup> Andreas Eizinger,<sup>1</sup> Brian N. Wylie,<sup>3</sup> George S. Davidson<sup>3</sup>, “A Gene Expression Map for *Caenorhabditis elegans*”, *Science*, New Series, Vol. 293, No. 5537 (Sep. 14, 2001), pp. 2087-2092.

<sup>1</sup>Department of Developmental Biology and Genetics, Stanford University Medical School, Stanford, CA 94305, USA. <sup>2</sup>Stanford Medical Informatics, 251 Campus Drive, MSOB X-215, Stanford, CA 94305, USA. <sup>3</sup>Computation, Computers and Mathematics Center, Sandia National Laboratories, Albuquerque, NM 87185-0318, USA.

GSD contributions: VxInsight data processing and statistical computations; paper sections on VxInsight and analysis; extensive responses to reviewers; Supplemental Online Material common look and feel.

## **Abstract**

We have assembled data from *Caenorhabditis elegans* DNA microarray experiments involving many growth conditions, developmental stages, and varieties of mutants. Co-regulated genes were grouped together and visualized in a three-dimensional expression map that displays correlations of gene expression profiles as distances in two dimensions and gene density in the third dimension. The gene expression map can be used as a gene discovery tool to identify genes that are co-regulated with known sets of genes (such as heat shock, growth control genes, germ line genes, and so forth) or to uncover previously unknown genetic functions (such as genomic instability in males and sperm caused by specific transposons).

## **Introduction**

The completion of the *C. elegans* genome sequence has identified nearly all of the genes in the genome (19,282 genes) (1), but the function for most of these genes remains mysterious. A scant 6% of them have been studied with the use of classical genetic or biochemical approaches (1135 genes), and only about 53% show homology to genes in other organisms (10,303 genes) (2). The current challenge is to develop high-throughput functional genomics procedures to study many genes in parallel in order to elucidate gene function on a global scale (3–8). In one approach, a compendium of gene expression profiles was assembled from a large number of yeast DNA microarray experiments (9), which made it possible to ascribe potential functions to previously unknown genes by comparing their expression results to those of genes with known functions. Here, we have established a compendium of gene expression profiles for an animal, *C. elegans*. We

combined data from many DNA microarray experiments in order to identify sets of co-regulated genes. In each experiment, RNA from one sample was used to generate Cy3-labeled cDNA, and RNA from another sample was used to prepare Cy5-labeled cDNA. The two cDNA probes were simultaneously hybridized to a single DNA microarray and the ratio of the Cy3 to Cy5 hybridization intensities was measured.

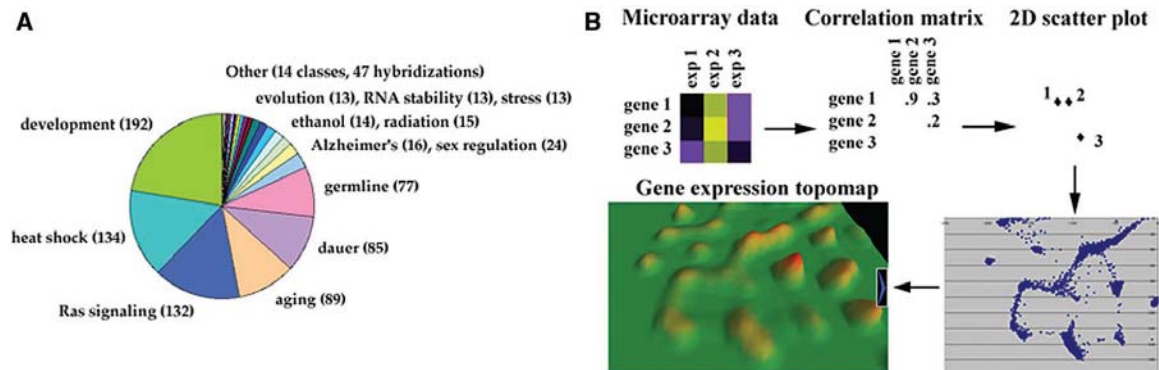
We have combined data from 553 experiments performed in collaboration with 30 different laboratories (10), including 179 experiments with microarrays containing 11,917 genes (63% of the genome) and 374 experiments using microarrays that have 17,817 genes (94% of the genome). The experiments compare RNA between mutant and wild-type strains or between worms grown under different conditions. Figure 2-1(A) shows the types of experiments that have been done to date, including experiments on wild-type development, heat shock, Ras signaling, aging, the dauer stage, sex regulation, and germ line gene expression (6, 7, 10).

### **The Experiments and Findings**

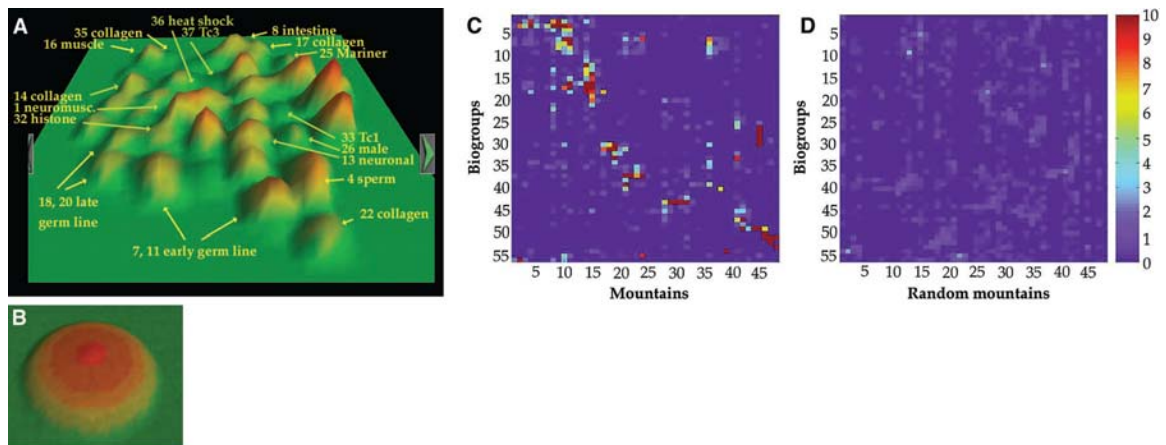
To find out which genes are co-expressed, we first assembled a gene expression matrix in which each row represents a different gene (17,817 genes) and each column corresponds to a different microarray experiment (553 experiments) (Fig. 2-1(B)). The matrix contains the relative expression level for each gene in each experiment (expressed as  $\log_2$  of the normalized Cy3/Cy5 ratios). We calculated the Pearson correlation coefficient between every pair of genes. For each gene, the similarity between it and the 20 genes with the strongest (positive) correlations were used to assign that gene to an  $x$ - $y$  coordinate in a two-dimensional scatter plot with the use of force-directed placement. In this  $x$ - $y$  ordination step, genes are positioned relative to each other under the influence of



attractive and repulsive forces. Each gene is attracted to other genes with a force proportional to their similarity in gene expression, but a constant force also repels each gene from groups of other genes. We then used a computer program called VxInsight to visualize the spatial distribution of the genes, resulting in a display in which genes with a high correlation are placed near to each other on a two-dimensional scatter plot. [Forced-directed placement and data mining with VxInsight are described in (11, 12), available Online at [www.cs.sandia.gov/projects/VxInsight.html](http://www.cs.sandia.gov/projects/VxInsight.html), and Link 1 at *Science Online* (13)]. As a further visual cue, the two-dimensional scatter plot is converted into a three-dimensional terrain map in which the  $z$  axis denotes the density of genes within an area (Fig. 2-2(A)).



**Figure 2-1. Types of experiments and VxInsight terrain map.** (A) Pie chart shows types of experiments used to generate the gene expression terrain map (10). Numbers in parentheses refer to the number of microarray hybridizations done for that experiment class, out of a total of 553 different microarray hybridizations. Some microarray hybridizations fall into multiple classes. (B) Construction of the gene expression terrain map by VxInsight. Expression data involving 17,661 genes and 553 experiments are shown. In the expression matrix, yellow denotes increased relative gene expression and blue denotes decreased gene expression. Only three genes and three experiments are shown for simplicity. The expression data are used to calculate Pearson correlations between every pair-wise combination of genes. The most correlated genes in the correlation matrix are used to construct a two-dimensional scatter plot. The scatter plot is converted to a gene expression terrain map showing the gene correlations in three dimensions, where the altitude of a mountain corresponds to density of the genes, denoted by red, yellow, and green.



**Figure 2-2. VxInsight map with biological groups and statistical significance.** (A) *Caenorhabditis elegans* gene expression terrain map created by VxInsight at lowest resolution, showing three-dimensional representation of 44 gene mountains derived from 553 microarray hybridizations and consisting of 17,661 genes (representing 98.6% of the genes present on the DNA microarrays) (31). Selected gene classes that are enriched in specific mountains are shown. (B) Terrain map derived from randomized data. (C and D) We created 56 lists of genes with similar biological function (biogroup), such as genes involved in meiosis, mitosis, translation, DNA synthesis, etc. We then counted the number of genes that overlap in the biogroup with that of the gene expression mountain. We calculated the probability of seeing the observed number of overlaps or more by chance (P value) for each biogroup-mountain pair assuming a hypergeometric distribution. Overlap P values for each biogroup with each mountain (C) and with randomly constructed mountains of the same size as the original mountain (D) are shown. Scale shows the  $\log_{10}$  (P value). The list of biogroups and the mountains are shown in Web table 2 and Web table 3 (13), respectively. The biogroups and mountains are ordered so that neighbors have similar mountain profiles.

The gene expression map shows gene expression clusters for nearly all of the genes (17,661 genes, 93% of the genome) formed by numerous, diverse microarray experiments (Fig. 2-2 (A)) (14). The raw *C. elegans* expression data can be downloaded from (13), and copies of VxInsight can be downloaded from <http://cmgm.stanford.edu/~kimlab/topomap/vxinsight.htm>. Genes were assigned to individual gene expression clusters (terrain map mountains), and each cluster was numbered according to size, from mount 0 (2703 genes) to mount 43 (5 genes) (Table 2-1). Each mountain contains sets of highly correlated genes, and the mountain width denotes the overall level of correlation of the genes in that mountain. Mountain altitude

signifies the number of genes present in that mountain. It is not yet clear how well gene expression correlations between genes in different mountains can guide the relative placement of one mountain to other mountains on the map.

To assess the significance of the topographical patterns shown in Fig. 2-2 (A), we first randomized the expression table by shuffling the values within each row and then reclustered the genes. We observed no appreciable structure in the randomized terrain map (Fig. 2-2 (B)), suggesting that the geography observed in the actual expression map (Fig. 2-2 (A)) has biological significance. Then, to assess the stability of the gene expression terrain map, we either rederived the map from random starting positions or added a small amount of noise to the data and noted that there was a high degree of overlap between the various derived maps [Web Links 2 and 3 (13)]. To determine which correlations are dependent on specific sets of experiments, we split the experiments into two non-overlapping sets, formed two new expression maps, and compared gene correlations on one map with those on the other. We observed that many genes have similar neighbors in both maps [Web Link 4 (13)].

**Table 2- 1 Characteristics of the gene groups. The R value is a measure of the correlation of the expression patterns of the genes in a mountain. For each mountain, the Pearson correlation between each gene and every other gene in that mountain was calculated. R is the median of all of these Pearson correlations. Large mountains tend to have lower R because genes on opposite sides of the mountain have lower correlations. Unless otherwise noted, representation factors are significant at  $P < 0.001$  (17). The probability was determined using either the exact hypergeometric probability or using the normal distribution approximation, when appropriate.**

Mount	No. of genes	R	Functional groups (representation factor)
0	2703	0.11	
1	1818	0.15	Muscle (4.0X); neuronal (2.7X); PDZ genes (2.9X)
2	1465	0.15	Germ line-enriched (3.8X); oocyte (4.6X)
3	1363	0.13	Reverse transcriptase (3.0X)
4	1195	0.41	Sperm-enriched genes (21X); protein kinases (6.8X); protein phosphatases (15X); major sperm proteins (13X)
5	978	0.22	
6	909	0.21	Neuronal genes (6.5X)
7	810	0.43	Germ line-enriched (12X); oocyte (9.0X); meiosis (11X); mitosis (4.4X)
8	803	0.21	Intestine (13X); <i>Entemebahistolytica</i> N-acetylmuraminidase (12X); protease (6.4X); carboxylesterase (7.3X); lipases (10X); antibacterial proteins (17X); UGT (2.8X)
9	786	0.16	
10	635	0.19	
11	587	0.38	Germ line-enriched (13X); oocyte (13X); meiosis (8X); mitosis (10X); histone H1 (18X); retinoblastoma complex (26X)
12	462	0.29	
13	396	0.10	Neuronal genes (3.1X; $P < 0.006$ ); reverse transcriptase (4.0X)
14	353	0.38	Collagen (2.6X; $P < 0.005$ )
15	247	0.37	
16	230	0.40	Muscle (24X); collagen (29X)
17	210	0.37	Collagen (9.6X)
18	190	0.38	Germ line (2.4X); oocyte (4.1X); biosynthesis (2.6X); protein synthesis (9.7X)
19	189	0.29	Amino acid metabolism (5.5X); lipid metabolism (5.0X); cytochrome P450 (12X)
20	160	0.46	Germ line-enriched (7.5X); biosynthesis (10X); protein expression (16X); heat shock (10X)
21	154	0.30	Lipid metabolism (10X)
22	151	0.58	Collagen (8X)
23	143	0.53	Protein expression (19X); energy generation (8.6X)
24	133	0.37	Amino acid metabolism (3.9X); lipid metabolism (8.5X); fatty acid oxidation (22X)
25	102	0.44	Mariner transposases (173X)
26	95	0.43	Male-enriched genes (9.5X)
27	87	0.48	Amino acid metabolism (8X); energy generation (8.8X)
28	61	0.28	
29	40	0.53	
30	36	0.41	Protein expression (7.7X)
31	25	0.36	
32	24	0.47	Nucleosomal histones (226X)
33	27	0.43	Tc1 transposon (538X)
34	17	0.44	
35	15	0.59	Collagen (60X)
36	10	0.71	Heat shock (337X)
37	11	0.77	Tc3 transposon (1600X)
38	8	0.44	
39	8	0.42	
40	8	0.43	Protein expression (23X)
41	7	0.45	Protein expression (26X)
42	6	0.33	
43	5	0.69	

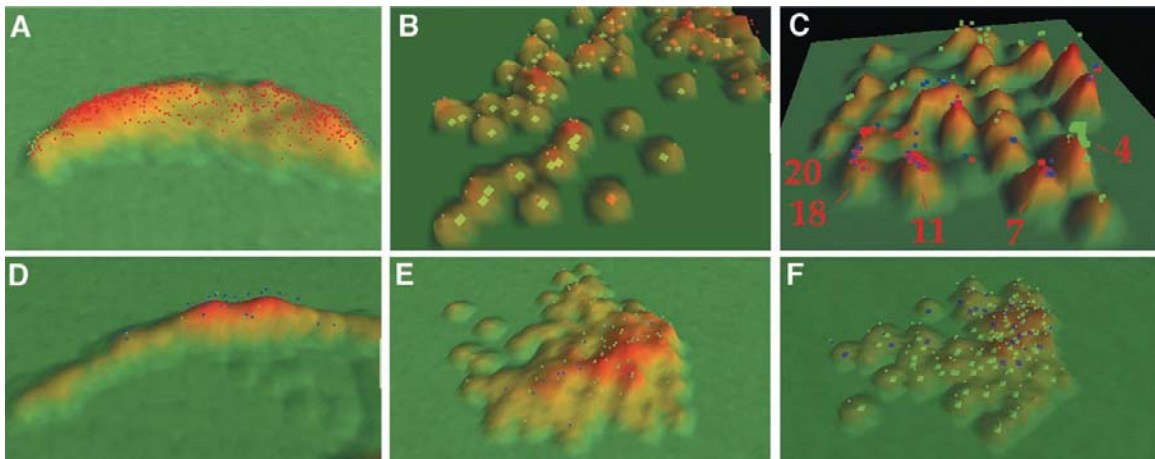
Lastly, we showed that the observed overlaps between clusters on the gene expression terrain map and groups of genes with similar biological functions are much higher than would be expected by random chance (Fig. 2-2 (C and D)) (13, 15). This demonstrates that there are strong biological patterns embedded in the expression data and that the clustering produced by VxInsight has biological relevance. A wide variety of other algorithms [such as hierarchical clustering (16)] could have been used in addition to VxInsight to cluster genes on the basis of their expression profiles. We chose to use VxInsight because depicting gene correlation data in three dimensions is extremely useful to visualize patterns of gene expression in large data sets.

We studied the genes in each mountain to find patterns suggesting the underlying biological property for that group of genes. We also looked through 56 sets of genes that were previously known to function together (Web table 1) and found that 46 showed enrichment in one or more of the gene expression mountains (Fig. 2-2 (C)). Some of the gene expression mountains grouped genes together that were expressed in similar tissues (such as muscle, neuron, germ line), whereas other mountains grouped genes that had similar cellular functions (for example, histones, ribosomal genes, collagens). Overall, we were able to infer a potential physiological importance for 30 of the 44 mountains by showing that specific mountains were enriched for particular sets of genes. The functional interactions suggested by the gene expression terrain map are based entirely on expression data. Thus, in addition to biochemistry and genetics, one could now infer gene functions with the use of gene expression data.

Several mountains were highly enriched for genes from particular tissues or organs. For example, previous microarray experiments identified a total of 650 sperm-

enriched genes (6). Of these, 583 genes (89%) are present in mount 4 (1195 genes in total), which is 21 times (21X) more than the number of genes expected due to random chance [defined as the representation factor (17)] (Fig. 2-3 (A) and Web table 1).

The sperm-enriched genes were defined using microarrays containing only 63% of the genome, and 848 of the genes in mount 4 were present on these microarrays (and, thus, were available to be identified as sperm enriched). Thus, highly sperm-enriched genes (99.9% confidence level) composed about 69% of mount 4. Much of the remainder of mount 4 consisted of genes that are sperm enriched but at a lower level; 775 genes in mount 4 were sperm-enriched at the 95% confidence level (88% of mount 4 out of 848 genes).



**Figure 2-3. Biological categories in VxInsight mounts.** (A) Mount 4 (sperm). Sperm-enriched and MSP genes are shown in red and green, respectively. (B) Enlarged view of MSP genes (green) and sperm-enriched genes (red) in mount 4. (C) Germ line genes in mounts 7, 11, 18, and 20. Sperm-enriched (green), oocyte-enriched (blue) and germ line enriched genes (red) from (6) are shown. Numbers refer to mountains. (D) Mount 8 (intestine). Intestinal (green) and protease (blue) genes are shown. (E) Mount 16 (muscle). Muscle (blue) and collagen (green) genes are shown. (F) Mount 26 (male). Male-enriched (green) and lectins (blue) are shown.

The major sperm protein (MSP) genes, which are genes encoding proteins that bind each other in forming the sperm cytoskeleton and are required for sperm motility

(Fig. 2-3 (A and B)) [see movie (13)] (18), clustered together at one end of mount 4. As noted previously, protein kinases and phosphatases are enriched in sperm (6). These gene classes were also highly enriched in mount 4; specifically, 103 of 361 protein kinase genes (6.8X higher than random chance) and 67 of 106 protein phosphatases (15X) are present in mount 4 (Web table 1). Because sperm are unusual cells in that they are transcriptionally and translationally inactive, the high abundance of protein kinases and phosphatases in mount 4 suggests that sperm commonly use protein phosphorylation to regulate protein activity.

Previous microarray experiments identified 258 oocyte-enriched genes and 508 genes enriched in both sperm and oocytes (germ line-intrinsic genes) (6). The germ line-enriched and oocyte-enriched genes were concentrated in three mountains: mount 7 (12X and 9X, respectively), mount 11 (13X and 13X), and mount 18 (2.4X and 4.1X). Additional germ line-enriched genes were also concentrated in mount 20 (7.5X) [Fig. 2-3(C) and movies at (13)]. These four mountains contain 483 of the 766 germ line- and oocyte-enriched genes (63%). Of the remaining 283 germ line-enriched genes, 161 (21%) were found in mount 2, which is a large mountain containing many genes involved in diverse biosynthetic pathways.

These four mountains segregate the germ line genes according to their different biological roles. For example, the first two (mount 7 and mount 11) were highly enriched for meiosis and mitosis genes and, therefore, may reflect genes expressed in the early germ line. We identified a set of 23 genes known to be involved in meiosis; 12 are in mount 7 (11X representation factor) and six are in mount 11 (8X) (Web table 1). The list of meiosis genes contains six involved in forming the synaptonemal complex, and all are

contained in mount 7 (19). We identified a set of 80 genes known to be involved in mitosis (Web table 1). Of these, 16 are in mount 7 (4.4X) and 26 are in mount 11 (10X). The list of mitosis genes contains five that are orthologs of components of the mammalian retinoblastoma (Rb) tumor suppressor complex. The Rb tumor suppressor complex regulates cell growth and division by controlling gene expression throughout the cell cycle (20). In *C. elegans*, this complex consists of LIN-35 (Rb), HDA-1 (histone deacetylase), and RBA-1/RBA-2 (both RbAP48) (21). All four genes encoding proteins in the Rb tumor suppressor complex were present in mount 11. In addition to these four genes, *lin-9* is implicated in Rb complex formation as *lin-9* mutants have a similar phenotype to *lin-35*, *hda-1* and *rba-2* mutants (synthetic multivulva) (22). We observed that *lin-9* was clustered with the Rb complex genes in mount 11. Thus, both mutant phenotype and microarray expression data indicate that *lin-9* may play a functional role in the Rb complex.

Mount 18 and mount 20 were both enriched for protein expression and biosynthesis genes, respectively. We identified 478 genes involved in various biosynthetic pathways, such as energy generation, nucleotide synthesis, carbohydrate metabolism, fatty acid oxidation, and amino acid synthesis (Web table 1). The biosynthesis genes were mildly enriched in mount 18 (2.6X) and strongly concentrated in mount 20 (10X). Then, we identified 390 genes involved in protein synthesis, such as genes encoding tRNA synthetases, ribosomal proteins, chaperones, heat shock proteins, protein translocation components, and RNA processing proteins (Web table 1). These protein synthesis genes are enriched in mount 18 (9.7X) and mount 20 (16X). Biosynthesis and protein expression are highly active during oogenesis, as small germ



line cells enlarge into enormous oocytes ready to begin growth of the new embryo. Thus, genes clustered in mount 18 and 20 may correspond to late germ line genes.

Eight genes are known to be expressed primarily in the intestine (Web table 1). Five of the intestinal genes were expressed in mount 8, which is 13X the number expected given the size of this mount (803 genes) (Fig. 2-3 (D)). Additional genes in mount 8 are likely to be expressed in the intestine because they encode proteins involved in digestion or protection from bacterial infection. Mount 8 contained five genes that are similar to *Entameba histolytica* N-acetylmuraminidase (a bacterial lysozyme, 12X enriched), suggesting that these genes may be expressed in the *C. elegans* intestine to digest bacterial cell walls. There were 32 protease genes in mount 8 (out of 116 proteases in the genome, 6.4X enriched) that could be expressed in the intestine to break down bacterial proteins. Carboxylesterases are enzymes used by the intestine to metabolize carbohydrates and sugars; 12 (out of a total of 36 carboxylesterases in the genome, 7.3X enriched) are expressed in mount 8 including *ges-1*, which is known to be expressed in the intestine (23). Lipases are enzymes used by the intestine to digest lipids; 15 of the 32 lipases in the *C. elegans* genome are contained in mount 8 (10X enriched). Mount 8 contained the gene *nuc-1*, which encodes a deoxyribonuclease (DNase) expressed by the intestine for digestion of bacterial DNA (24). Two genes encoding proteins similar to the mammalian low-density lipoprotein (LDL) receptor were present in mount 8 and could function in the intestine to bind sterols in the lumen and internalize them into intestinal cells. Mount 8 contained two genes that encode insulin-related peptides that might be expressed in the intestine to regulate uptake of nutrients.

Another function of the intestine is that it protects against bacterial infection and from ingestion of harmful chemicals. Mount 8 contained seven out of nine genes that encode antibacterial proteins similar to granulysin of cytotoxic T cells (17X enrichment). These genes may be expressed in the intestine to protect the worm from bacterial infections. Mount 8 contained a metallothionein gene (*mtl-2*), which is known to be expressed in the intestine and function to bind and inactivate heavy metals (25). Mount 8 contained eight genes encoding UDP-*N*-acetylglucosamine: alpha-3-D-mannoside beta-1, 2-Nacetylglucosaminyltransferase I (where UDP is uridine 59-diphosphate) out of a total of 64 such genes in the genome (2.8-fold enrichment), including *gly-14*, which is known to be expressed in the intestine (26). These genes encode enzymes that are of major importance in the modification and subsequent inactivation of toxic compounds. They could be expressed in the intestine to protect the worm from harmful chemicals.

Thirty-nine genes are known to be expressed primarily in muscle (Web table 1). These genes were enriched in mount 1 (4.1X) and mount 16 (24X). Mount 1 is a large mountain with diverse types of genes, and it was also enriched for many neuronal proteins. In mount 1, the known muscle genes included primarily receptors, extra-cellular proteins, or receptor-associated proteins such as *egl-19* (which encodes a voltage-dependent calcium channel), *unc-52* (which encodes a component of the basement membrane), or *egl-30* (which encodes a G<sub>alpha</sub> protein) (Fig. 2-3 (F)) (27–29). Mount 16 included genes that make the muscle filaments themselves, such as those encoding myosin light chain, myosin heavy chain, paramyosin, and two types of troponin (Fig. 2-3 (E)).

We examined 88 genes that are known to be enriched in neuronal cells. These neuronal genes were clustered in mount 1 (2.7X), mount 6 (6.5X), and mount 13 (3.1X). Both muscle and neuronal genes are clustered in mount 1, and the known muscle or neuronal genes in mount 1 tended to encode receptors or receptor-associated proteins. One possibility is that these genes function in synaptic transmission at neuromuscular junctions. For example, PDZ-containing proteins are expressed in synapses and appear to have a role in clustering or localizing neurotransmitter receptors in both the pre- and postsynaptic densities (30). There are 58 genes with PDZ domains in *C. elegans*, and 17 of these were concentrated in mount 1 along with other neuromuscular genes (2.9X enriched). In addition to neuronal genes, mount 13 was enriched for retrotransposons (4.0X), suggesting that retrotransposons might be active in worm neurons.

Previous microarray experiments comparing adult males with adult hermaphrodites identified 1651 male-enriched genes, consisting not only of the sperm genes (enriched in mount 4) but also genes expressed in the soma such as in the male copulatory organ or in male-specific neurons (7). Many of the male-enriched genes were clustered in mount 4, corresponding to sperm-enriched genes. The male-enriched genes were also enriched in mount 26 (9.5X) (Fig. 2-3 (F)). Of the 95 genes in mount 26, 83 are male-enriched (87%) and are likely expressed in the male soma. Mount 26 contained 15 genes that encode cell surface markers (C-type lectins), suggesting that these genes may function to distinguish the extra-cellular surfaces of male and hermaphrodite cells.

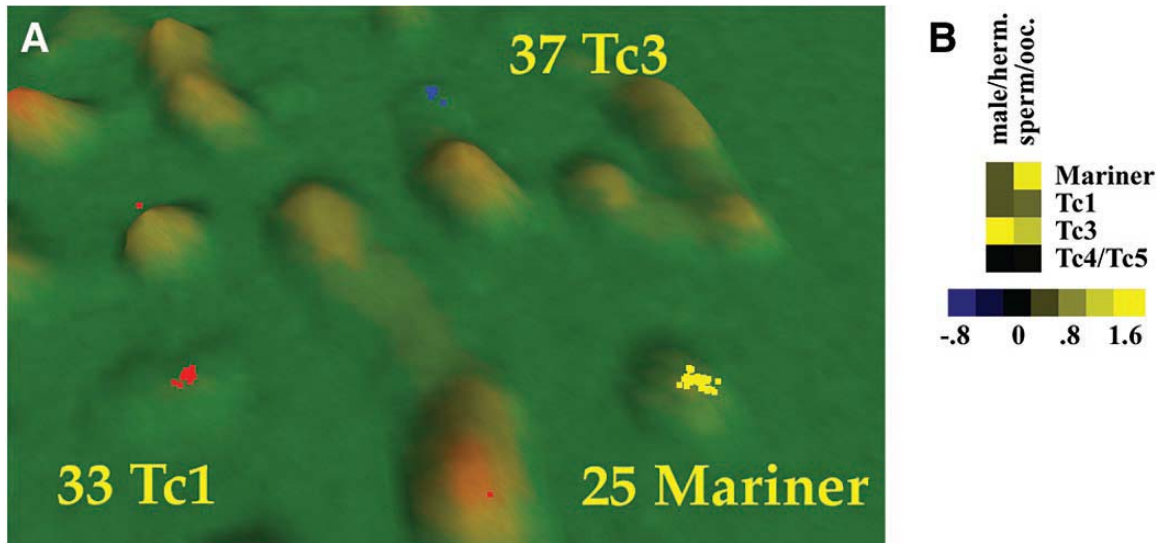
The second general pattern of gene clusters observed in the gene expression terrain map corresponds to sets of genes that form functional modules, such as genes that act in one biochemical pathway or encode similar types of proteins. For example, mount

20 and mount 36 were both enriched for heat shock genes. In particular, 7 of the 10 genes in mount 36 encode heat shock proteins (337X enriched). The remaining three genes (F26H11.3, F58E10.4, and Y43F8B.2A) were not previously known to be involved in the heat shock response. We performed another set of heat shock microarray experiments and found that all three are heat shock-regulated at the 99% confidence level (Table 2-2). Thus, direct experimental evidence confirmed the genetic relation suggested by the juxtaposition of three unknown genes with known heat shock protein genes.

Mount 32 is highly enriched for histone genes (226X); of the 24 genes in this mountain, 22 are histone genes that comprise the nucleosomal core (H2A, H2B, H3, and H4). The other type of histone (H1) is not part of the nucleosome itself but serves as a linker between nucleosomal subunits on chromatin. There are five histone H1 genes, and three of these are in mount 11 (18X) along with early germ line genes.

The 99 transposons in the *C. elegans* genome consist mainly of Mariner elements, Tc1, Tc3, Tc4, and Tc5 (Web table 1). In most cases, transposons of the same type fell into the same cluster, as was expected because different members of each transposon type have nearly identical sequences and would be expected to cross-hybridize. The Mariner transposons fell into mount 25, most Tc1 copies were in mount 33, and Tc3 copies were in mount 37 (Fig. 2-4 (A)). Tc4 and Tc5 show more sequence heterogeneity and were spread out in mounts 0, 1, 3, and 9. The expression map showed that the Tc1, Tc3, and Mariner transposon families were expressed differently from each other, suggesting different types of developmental regulation. To begin to elucidate this developmental control, we examined the expression profiles for the transposons in the published microarray data (6, 7). We found that average expression of Mariner transposons was

high in sperm relative to oocytes, suggesting that this transposon may have a higher mobilization rate in the male compared with the hermaphrodite germ line (Fig. 2-4 (B)). We also found that the average expression of Tc3 was high in the male soma, as it is enriched in males versus hermaphrodites but not in sperm versus oocytes.



**Figure 2-4. Transposon mounts.** (A) Transposon clusters in the gene expression terrain map. Tc1 (red), Tc3 (blue), and Mariner (yellow) transposons are indicated. Numbers refer to mountains. (B) Transposon expression in males and sperm. Because different copies of each type of transposon have nearly identical sequences, expression for all genes of each type of transposon are averaged together. Web fig. 4 has expression for individual transposon copies. Male/herm., experiments comparing adult male to adult hermaphrodite RNAs (7); sperm/oocyte, experiments comparing *fem-3(gf)* to *fem-1(lf)* worms (6). Yellow and blue denote high- and low-expression levels, respectively.

Additional sets of genes that cluster in the same mountain on the gene expression terrain map are shown in Table 2-1 and listed in Web table 1. Further investigation is likely to reveal many more clusters of genes on the terrain map.

The gene expression database provides higher resolution than individual microarray experiments because the expression patterns of particular groups of genes are refined by a multitude of experiments. For example, the germ line microarray

experiments (6) identified 758 genes that are enriched in the hermaphrodite germ line, but the gene expression terrain map was able to subdivide these genes into four mountains (mounts 7, 11, 18, and 20) enriched for genes with distinct biological roles. Furthermore, the position of genes within a mountain in the terrain map often provides information about its function, as we frequently observed that genes with similar function were placed close to each other in a section of one mountain. This level of detail was not observed in microarray experiments comparing only two worm samples (31).

**Table 2-2. Heat shock induction levels for 10 genes in mount 36**

. Heat shocks were for 15 min at 33°C, and RNA expression levels were measured 30 min after heat shock. Results show average expression levels (6 SE) from four independent experiments. HSP, heat shock protein

Gene	Induction 6SE	Protein
C12C8.1	65.3 619.3	HSP70
F44E5.4	82.5 624.8	HSP70
F44E5.5	109.7 641.8	HSP70
<i>hsp-16.11</i>	39.7 614.7	HSP-16
<i>hsp-16.1</i>	52.3 615.0	HSP-16
<i>hsp16-2</i>	68.7 622.6	HSP-16
<i>hsp16-41</i>	39.0 65.0	HSP-16
F26H11.3	11.1 62.3	Bromodomain protein
F58E10.4	5.1 61.4	Similar to <i>S.cerevisiae</i> YNL155W
Y43F8B.2A	10.5 61.6	Similar to Y43F8B.M

The ability to identify candidate genes whose function can subsequently be confirmed by experimental testing depends greatly on the resolution of the terrain map. Some sets of genes (such as the heat shock genes, sperm enriched genes, nucleosomal histone genes, and ribosomal genes) show tight clustering in which genes that are known to be functionally related are adjacent to each other on the gene expression map. Other

groups of genes (such as the retinoblastoma complex genes) may be loosely clustered together in the same expression mountain.

Although the sperm versus oocyte experiments were specifically designed to identify sperm and oocyte genes (hypothesis testing), the terrain map also grouped genes even when they were not specific targets of any of the experiments in the database (undirected knowledge discovery). For example, none of the experiments were specifically designed to reveal expression in muscle, intestine, or neurons, or to show expression by the histone, collagen, or transposon genes (Fig. 2-1 (A)). Nevertheless, these genes form discrete clusters or mountains on the terrain map, most likely because they showed serendipitous co-regulation in one or more of the experiments in the large database. In many cases, mountains on the gene expression terrain map reveal unexpected interactions between genes. These types of unexpected gene clusters are best revealed using undirected data mining of a global gene expression database rather than testing specific hypotheses.

*Caenorhabditis elegans* is a powerful model system to analyze biological processes with the use of functional genomics approaches. In addition to global expression studies, efforts are under way to determine the mutant phenotype of most *C. elegans* genes using RNA interference and to identify protein binding interactions on a whole genome level using a high-throughput, yeast two-hybrid approach (32–36). Thus, there is a rapid accumulation of expression data, mutant phenotypes, and protein binding interactions, making it possible to begin to elucidate cellular, developmental, and organismic processes on a global scale.

## References and Notes

1. The *C. elegans* Sequencing Consortium, *Science* **282**, 2012 (1998).
2. M. C. Costanzo *et al.*, *Nucleic Acids Res.* **28**, 73 (2000).
3. M. Schena, D. Shalon, R. W. Davis, P. O. Brown, *Science* **270**, 467 (1995).
4. D. Shalon, S. J. Smith, P. O. Brown, *Genome Res.* **6**, 639 (1996).
5. J. DeRisi *et al.*, *Nature Genet.* **14**, 457 (1996).
6. V. Reinke *et al.*, *Mol. Cell* **6**, 605 (2000).
7. M. Jiang *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **98**, 218 (2001).
8. A. A. Hill, C. P. Hunter, B. T. Tsung, G. Tucker-Kellogg, E. L. Brown, *Science* **290**, 809 (2000).
9. T. R. Hughes *et al.*, *Cell* **102**, 109 (2000).
10. S. K. Kim, unpublished data. Personal communications from colleagues are as follows: V. Ambros (Dartmouth College), P. Anderson (Univ. of Wisconsin), I. Callard (Boston Univ.), C. Conley (NASA Ames), D. Eisenmann (Univ. of Maryland), S. Emmons (Albert Einstein Univ.), A. Fire (Carnegie Institute), M. Hengartner (Univ. of Zurich, Switzerland), T. Johnson (Univ. of Colorado), J. Kimble (Univ. of Wisconsin), J. Lee (Yonsei Univ., Korea), P. Larsen (Univ. of Los Angeles), C. Link (Univ. of Colorado), G. Lithgow (Univ. of Manchester, England), S. Mango (Univ. of Utah), S. McIntire (Univ. of California, San Francisco), W. Shafer (Univ. of California, San Diego), R. Menzel (Free Univ., Berlin), R. Padgett (Rutgers Univ.), J. Thomas (Univ. of Washington), K. Thomas (Univ. of Missouri), L. Vassilieva (Univ. of Utah), and D. Zarkower (Univ. of Minnesota).



11. G. S. Davidson, B. Hendrickson, D. K. Johnson, C. E. Meyers, B. N. Wylie, *J. Intelligent Inform. Syst.* **11**, 259 (1998).
12. G. S. Davidson, B. N. Wylie, K. W. Boyack, *Cluster Stability and the Use of Noise in Interpretation of Clustering*, IEEE Symposium on Information Visualization, 2001,23-30.
13. Web Figures, tables, movies, and text are available on *Science Online* at [www.sciencemag.org/cgi/content/full/293/5537/2087/DC1](http://www.sciencemag.org/cgi/content/full/293/5537/2087/DC1).
14. We compared the gene clustering results with the use of VxInsight to those using hierarchical clustering, which is a standard method to cluster genes based on Pearson correlation coefficients (27). We obtained similar results using the two methods and found that there was strong overlap between mountains formed using VxInsight and gene clusters using hierarchical clustering.
15. Using a conservative Bonferroni correction, the probability of observing one of the red dots in Fig. 2C is approximately  $10^{-6}$ . The actual significance of the entire result is much more than this because there are 64 different overlaps with this level of significance, whereas the random solution contains no overlaps at this significance level.
16. M. B. Eisen, P. T. Spellman, P. O. Brown, D. Botstein, *Proc. Natl. Acad. Sci. U.S.A.* **95**, 14863 (1998).
17. The representation factor shows whether genes from one list (list A) are enriched in another list (list B), assuming that genes behave independently. The representation factor is defined as:  $(\text{number of genes in common between both lists})/(\text{number of genes in the genome})/(\text{number of genes in list A})(\text{number of genes in list B})$ .
18. S. W. L'Hernault, T. M. Roberts, *Methods Cell Biol.* **48**, 273 (1995).
19. A. M. Villeneuve, personal communication.

20. N. Dyson, *Genes Dev.* **12**, 2245 (1998).
21. X. Lu, H. R. Horvitz, *Cell* **95**, 981 (1998).
22. E. L. Ferguson, H. R. Horvitz, *Genetics* **123**, 109 (1989).
23. L. G. Edgar, J. D. McGhee, *Dev. Biol.* **114**, 109 (1986).
24. C. J. Lyon, C. J. Evans, B. R. Bill, A. J. Otsuka, R. J. Aguilera, *Gene* **252**, 147 (2000).
25. J. H. Freedman, L. W. Slice, D. Dixon, A. Fire, C. S. Rubin, *J. Biol. Chem.* **268**, 2554 (1993).
26. S. Chen, S. Zhou, M. Sarkar, A. M. Spence, H. Schachter, *J. Biol. Chem.* **274**, 288 (1999).
27. R. Y. Lee, L. Lobel, M. Hengartner, H. R. Horvitz, L. Avery, *EMBO J.* **16**, 6066 (1997).
28. L. Brundage *et al.*, *Neuron* **16**, 999 (1996).
29. T. M. Rogalski, B. D. Williams, G. P. Mullen, D. G. Moerman, *Genes Dev.* **7**, 1471 (1993).
30. S. K. Kim, *Curr. Opin. Cell Biol.* **9**, 853 (1997).
31. There are 156 genes that are present on the DNA microarrays but not represented on the gene expression terrain map, either because there is a large amount of missing data or they show almost no variation across experiments.
32. A. G. Fraser *et al.*, *Nature* **408**, 325 (2000).
33. P. Gonczy *et al.*, *Nature* **408**, 331 (2000).
34. F. Piano, A. J. Schetterdagger, M. Mangone, L. Stein, K. J. Kemphues, *Curr. Biol.* **10**, 1619 (2000).

35. I. Maeda, Y. Kohara, M. Yamamoto, A. Sugimoto, *Curr. Biol.* **11**, 171 (2001).
36. A. J. M. Walhout *et al.*, *Science* **287**, 116 (2000).
37. We would like to especially thank S. Scherer (Acacia Biosciences) for guidance and advice on this project, M. Werner-Washburne for help in applying VxInsight to microarray analysis, and A. Owen and L. Lazzeroni for helpful advice on statistics. We thank J. Ryu, P. Roy, and J. Shaw for critical comments on the manuscript. We thank the programmers at the Stanford Microarray Database for their help in the microarray analyses, and Proteome for annotation of *C. elegans* genes. Supported by grants from the National Institute for General Medical Sciences, National Center for Research Resources, Merck Genome Research Institute, Aventis, and by Laboratory Directed Research and Development, Sandia National Laboratories, U.S. Department of Energy (DE-AC04-94AL85000).

### **Chapter 3: Comparative Analysis of Multiple Genome-Scale Data Sets**

This chapter has previously appeared in substantially the same form as: Margaret Werner-Washburne,<sup>1,3</sup> Brian Wylie,<sup>2</sup> Kevin Boyack,<sup>2</sup> Edwina Fuge,<sup>1</sup> Judith Galbraith,<sup>1</sup> Jose Weber,<sup>1</sup> and George Davidson<sup>2</sup>, “Comparative Analysis of Multiple Genome-Scale Data Sets”, *Genome Research*, 2002 12: 1564-1573

<sup>1</sup>*Biology Department, University of New Mexico, Albuquerque, New Mexico 87131, USA;* <sup>2</sup>*Sandia National Laboratories, Albuquerque, New Mexico 87185, USA*

GSD contributions: VxInsight data processing, statistical analysis and modification for multi-data set comparisons; paper sections on VxInsight analysis, and response to reviewers.

## **Abstract**

The ongoing analyses of published genome-scale data sets is evidence that different approaches are required to completely mine this data. We report the use of novel tools for both visualization and data set comparison to analyze yeast gene-expression (cell cycle and exit from stationary phase/  $G_0$ ) and protein-interaction studies. This analysis led to new insights about each data set. For example,  $G_1$ -regulated genes are not co-regulated during exit from stationary phase, indicating that the cells are not synchronized. The tight clustering of other genes during exit from stationary-phase data set further indicates the physiological responses during  $G_0$  exit are separable from cell-cycle events. Comparison of the two data sets showed that ribosomal-protein genes cluster tightly during exit from stationary phase, but are found in three significantly different clusters in the cell-cycle data set. Two protein-interaction data sets were also compared with the gene-expression data. Visual analysis of the complete data sets showed no clear correlation between co-expression of genes and protein interactions, in contrast to published reports examining subsets of the protein-interaction data. Neither two-hybrid study identified a large number of interactions between ribosomal proteins, consistent with recent structural data, indicating that for both data sets, the identification of false-positive interactions may be lower than previously thought.

[Supplemental material is available online at <http://www.genome.org> and at [http://biology.unm.edu/biology/maggiww/Public\\_Html/Visualcomparison.htm](http://biology.unm.edu/biology/maggiww/Public_Html/Visualcomparison.htm), including data sets and download information for VxInsight.]

## Introduction

Enormous amounts of data are generated by high-throughput, genome-scale studies. Currently, data sets are available in which the quality of the data is so good that numerous re-analyses have yet to mine all the information present in them. Because of the size of genome-scale data sets, it is currently difficult, if not impossible, for the average researcher to ask global questions about a single data set, much less compare several data sets simultaneously. For this data to be completely mined, improved methods for integration and analysis of this information will be necessary to extract information from within and between the data sets and to develop hypotheses on the basis of these analyses (Aach et al. 2000). Toward that end, we performed a comparative analysis of four data sets from the yeast *Saccharomyces cerevisiae*, using the ordination and visualization tool VxInsight (Viswave).

As a model system for which the entire genome has been known since 1996 (Goffeau et al. 1996), *S. cerevisiae* has been the subject of several genome-scale studies, including gene expression (Lasharki et al. 1997; Chu et al. 1998; Eisen et al. 1998; Ferea et al. 1999; Gasch et al. 2000), protein-protein interactions (Schwikowski et al. 2000; Ito et al. 2001), and gene deletions (Winzeler et al. 1999). Research using yeast and other model systems is now poised to reveal even greater insight into cellular dynamics. As information about localization, modification, and abundance of all the proteins in the cell is obtained, it will become possible to reconstruct the dynamic interactions between all the major levels of organization in living organisms.

The data sets that we used for this comparative analysis include the following: transcriptional analysis of exit from stationary phase and the cell cycle after release from  $\alpha$ -factor arrest (Spellman et al. 1998) and two protein-protein interaction data sets (Schwikowski et al. 2000; Ito et al. 2001). We chose these gene-expression data sets because stationary phase, or  $G_0$ , is an offshoot of the mitotic cell cycle, and cells exiting  $G_0$  reenter mitosis at  $G_1$  (Werner-Washburne et al. 1993). In addition, starvation-induced  $G_0$  arrest is commonly used to synchronize eukaryotic cells to study reentry into the cell cycle (Callard and Mazzolini 1997; Zeise et al. 1998; Hildebrand and Dahlin 2000).

It is important to understand the relationship between the quiescent state and the cell cycle because most solid tumors are derived from  $G_0$  cells, and the proof-of-principal for chemotherapeutics is the ability to restore  $G_0$  arrest (Clark and Gillespie 1997; Zeitler et al. 1997; Joshi et al. 1998; Pajic et al. 2000). Additionally, a variety of important pathogens, such as *Mycobacterium tuberculosis* and *Cryptococcus neoformans*, are relatively difficult to treat because they reside in the body for extended periods of time as quiescent antibiotic-resistant cells (Tomee et al. 1997; Murray 1999). Finally, pathogens used as bio-weapons are usually stored and disseminated as quiescent cells. Thus, the importance of the  $G_0$  state and the relative lack of information about this phase of the life cycle underscore the importance of identifying the differences and similarities between the mitotic cell cycle and exit from  $G_0$ .

In the visual comparison reported here, we were able to detect significant differences in gene clusters between the two gene-expression data sets, indicating that yeast cells exiting starvation-induced quiescence are not synchronous and that expression of ribosomal protein genes during the cell cycle shows three distinct patterns. Overlaying

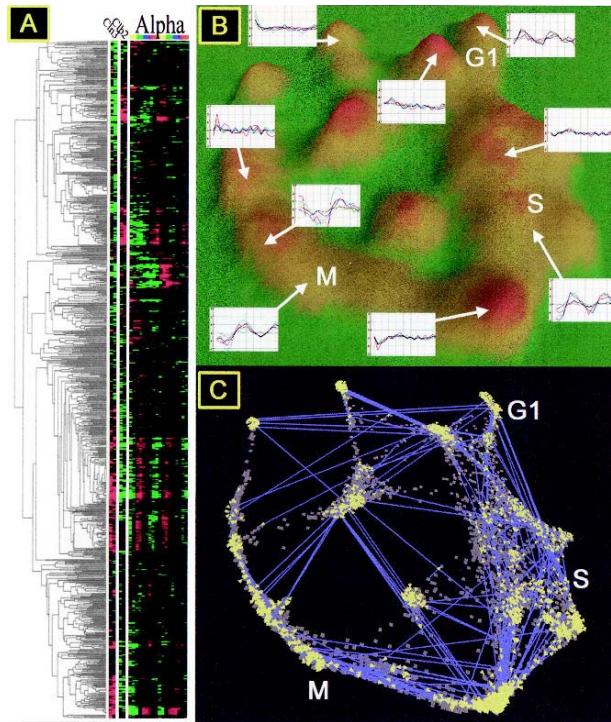
protein-interaction data led to the rapid detection of differences in the data sets and the finding that neither protein-interaction data set detected interactions between ribosomal proteins in the same subunit, which is consistent with recently published structural data, and indicates that the two-hybrid assay may be less prone to false-positives than previously thought.

## **Results**

### ***Data Set Topographies***

Ordination of genes of the  $\alpha$ -factor arrest/cell-cycle data into clusters (18 experiments per 6000 genes; Spellman et al. 1998), as described in Methods, resulted in a circular pattern (Fig. 3-1 (B, C)). Hills or ridges of G<sub>1</sub>-, S-, M-, and M- G<sub>1</sub>-regulated genes are found on the circumference of the circle, although not all of the groups of genes on the circumference of the ordination are cell-cycle regulated (see Web Supplement). In addition, M and G<sub>1</sub> clusters, with genes with expressions that are approximately opposite, are located on opposite sides of the topography. The two inner groups contain genes with regulation that is fairly constant throughout the cell cycle, including many genes involved in secretion, sterol biosynthesis, golgi function, and other constitutive pathways.

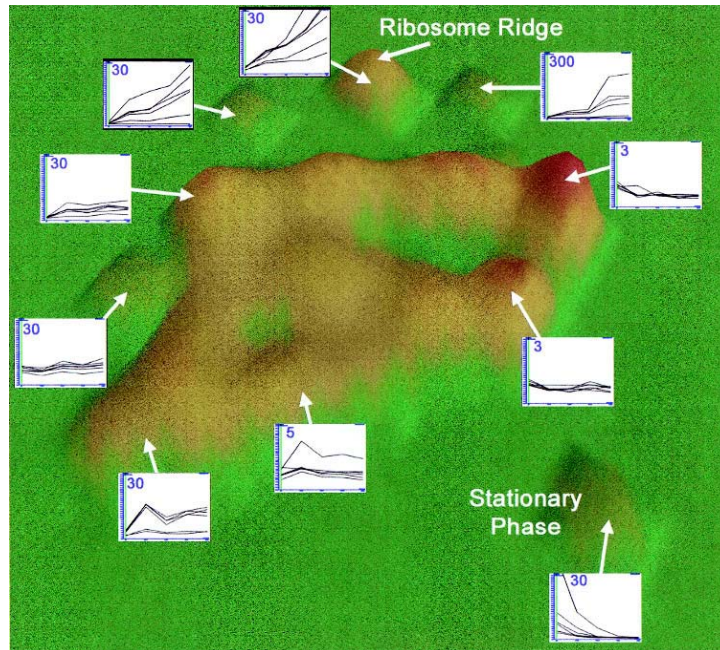




**Figure 3-1.  $\alpha$ -Factor-arrest data set (18 time points) ordinated and visualized in VxInsight.** (A) Cell cycle gene expression after  $\alpha$ -factor arrest and the dendrogram indicating similarities of gene expression as presented by Spellman et al. (Reprinted, with permission, from Spellman et al. 1998.) (B) Three dimensional topography in which mountains are formed over clusters of genes. The height of the mountain corresponds to the number of genes beneath it. Typical expression profiles for genes in each mountain are provided. G<sub>1</sub>, S, and M: Genes in these clusters are induced during the G<sub>1</sub>, S, or M phase of the cell cycle, respectively. (C) Ordination of genes (dots) that underlie the topography with links (blue lines with yellow arrows at each end) showing strong similarities (Pearson's  $R > 0.887$ ) that exist between genes in different clusters.

In the topography of the exit from stationary-phase data set, the 45 genes with mRNAs that accumulate in stationary phase are clustered in a hill at the bottom right of the topography (Fig. 3-2). Genes with mRNAs that accumulate rapidly as cultures exit stationary phase are found at the top and left sides of the topography. Background normalized data from membrane hybridizations were used for this analysis. Although there is variation in each of the expression profiles as a function of membrane and hybridization order, these differences were not significant, and normalization of this data

by several methods did not affect the clusters, although it did have an effect on the overall topography (data not shown).

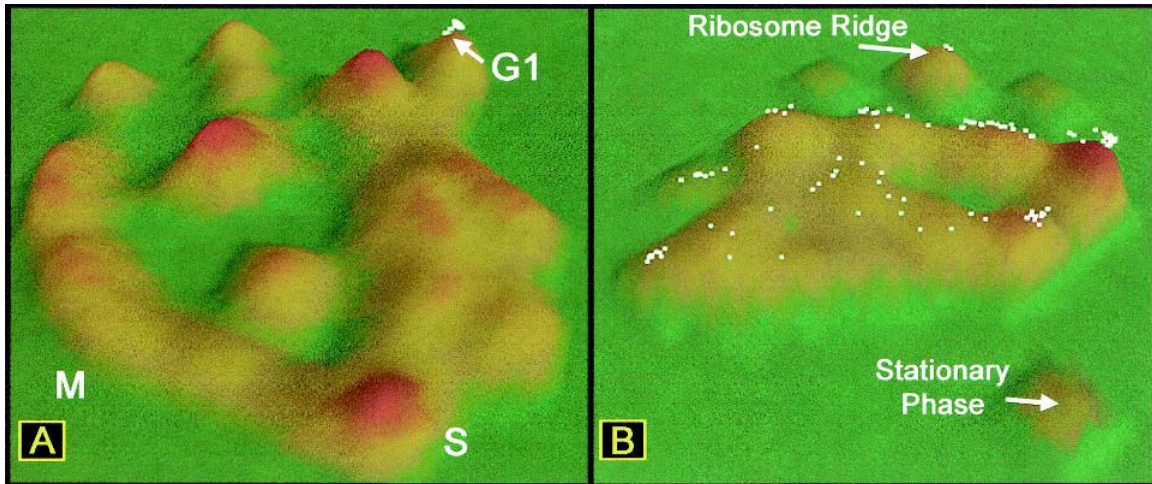


**Figure 3-2. VxInsight-generated ordination of exit from stationary-phase data set.** Examples of gene expression within each hill or cluster are shown. Along the x-axis of insert graphs are time points (0, 15, 30, 45, and 60 min) after re-feeding. The y-axis of insert graphs indicates the fold-increase or decrease from time equals 0, which is an average of four to five replicates for each time point. Numbers in the insert graphs indicate the maximum value of the y-axis, which indicates relative expression values obtained using GeneSpring (Silicon Genetics; see Methods). Data were generated as described (Methods).

### *Visual Queries of Two Gene-Expression Data Sets*

Using microarray data to develop hypotheses about related biological processes requires the ability to make comparative queries of multiple data sets. For this analysis, we chose to investigate the relationships between the processes of the mitotic cell cycle and exit from stationary phase in yeast. Cells in stationary-phase cultures are small and unbudded and are considered to be in the  $G_0$  state of the cell cycle. We asked whether cell cycle-regulated genes that clustered in the cell-cycle data set (Fig. 3-1 (B, C)) also

clustered in the exit from stationary-phase data set (Fig. 3-2). A set of  $G_1$ -regulated genes in the cell cycle topography (each represented as a white dot, see Fig. 3-3 (A)) was selected, and the positions of these genes were identified in the stationary-phase exit topography (Fig. 3-3 (B)). The selected  $G_1$ -induced genes, which are tightly clustered during the cell cycle, were randomly positioned in the stationary-phase exit topography.



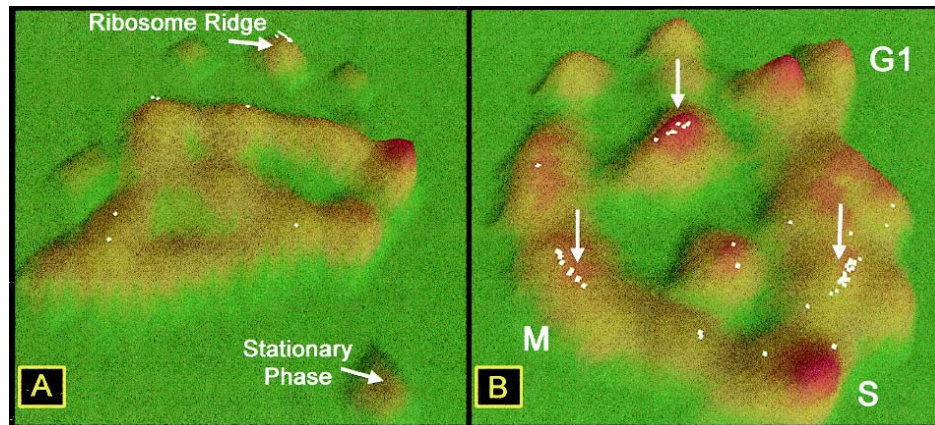
**Figure 3-3. Location of  $G_1$ -regulated genes in two different gene-expression data sets.** (A) Dots represent selected  $G_1$ -regulated genes in  $\alpha$ -factor–arrest cell-cycle data (Spellman et al. 1998). (B) Location of the same genes in the ordination of stationary-phase exit data.

To determine whether genes were  $G_1$  regulated, each gene was assigned a value that reflected how purely its expression coincided with  $G_1$ , which allowed us to rank order the subset of classical cell-cycle genes. We then examined groups of these genes. Of the 10 strongest  $G_1$ -regulated genes— including CLB6, SWI4, MCD1, RNR1, MNN1, YOX1, POL30, CLN2, SVS1, and TOS4—one half of these genes were randomly distributed, and one half were clustered ( $P < 0.001$ ) in the exit from stationary-phase data set (see supplemental data). When the positions of these genes were evaluated in the exit from stationary-phase topography, POL30 and MCD1 clustered with the genes with

induction that occurs almost immediately on refeeding, including CLN3 and most of the ribosomal protein genes. SWI4 clustered with genes with mRNAs that accumulate in the first 15 min and then remain fairly constant. In contrast, five of the most G<sub>1</sub>-like genes cluster in a region in which mRNA abundance fluctuates as a function of the particular membrane, but overall, the gene expression remains constant from hybridization to hybridization for the same membrane. These genes are CLB6, RNR1, CLN2, TOS4, and SVS1. The probability of finding these genes clustered in a region of 516 genes is highly significant ( $P < 0.001$ ).

During the cell cycle, CLN3 is induced first, followed by POL30 and MCD1, which are co-expressed with CLN1 (Stanford Genome Database). Although we had hypothesized that at least some of the patterns of gene expression might be conserved between the cell cycle and exit from stationary phase, the small subset of highly G<sub>1</sub>-regulated genes does not follow this temporal relationship. Early, morphological data had indicated that the cells in stationary-phase cultures did not exit stationary phase synchronously (Johnston et al. 1977). The induction of CLN3, POL30, and MCD1 almost immediately on refeeding and the relatively random distribution of the majority of other strongly G<sub>1</sub>-regulated genes in the exit from stationary-phase data set are consistent with the hypothesis that cells exiting stationary phase are not synchronous. Further analysis will be required to determine the conditions under which cells exiting stationary phase can be completely synchronized. Despite the lack of co-regulation of cell-cycle genes, there are clusters of genes with expression that increased or decreased dramatically during exit from stationary phase.

To determine whether genes co-expressed during exit from stationary phase might also be co-expressed in the cell-cycle data set, we investigated the small subunit ribosomal-protein (RPS) genes. Fifty-three of the 59 RPS genes are found in a ridge in the exit data set (Fig. 3-4 (A)). When the positions of all the RPS genes are identified in the cell-cycle topography, they are not clustered in one group but are located mostly in three different groups of genes (Fig. 3-4 (B)), with gene-expression profiles that are significantly different ( $P < 0.0001$ ). We conclude from this that RPS gene expression, which is tightly coregulated during exit from stationary phase and during other stress conditions (Gasch et al. 2000), shows at least three distinct patterns of expression during the mitotic cell cycle.



**Figure 3-4. Location of ribosomal protein genes (RPS genes) in two gene-expression data sets.** (A) Location of RPS genes in exit from stationary phase data. Fifty-three of 59 RPS genes are localized in the upper middle cluster. (B) Localization of the same RPS genes in cell-cycle data set. Arrows indicate three major groups of RPS genes.

The clustering of these genes into three groups is interesting because many ribosomal protein genes are duplicated and found as highly conserved gene pairs. Thus, any separation of these pairs of genes may have evolutionary implications. Of the 46

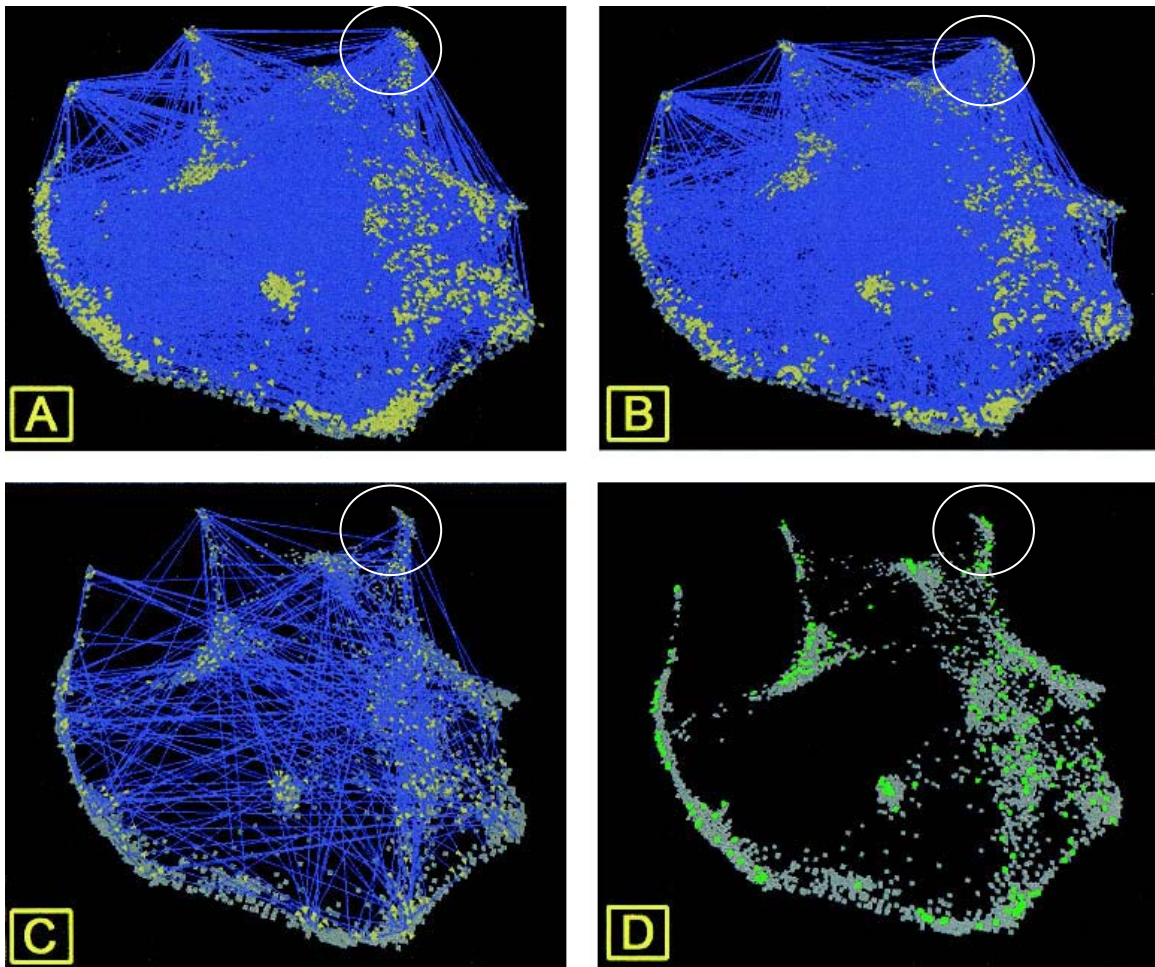
genes comprising 23 pairs of ribosomal protein genes that were present in the three clusters, there was an almost a threefold higher chance of members of a pair being in different clusters (34 of 48) compared with finding them in the same cluster (12 of 48; data not shown). Additional experiments will be required to determine the correlation of expression with protein abundance and, thus, whether the differences in ribosomal gene expression during the cell cycle have an effect on ribosome function or biogenesis.

### ***Visual Analysis of Protein-Protein Interactions***

Hypothesizing that the cell would use ‘just in time’ production of interacting proteins throughout the cell cycle as part of its regulation and control repertoire, we evaluate the extent to which co-expressed genes were found to encode interacting proteins. We incorporated information from two protein-protein interaction data sets (Schwikowski et al. 2000; Ito et al. 2001) in the cell-cycle topography (Fig. 3-5). Ito’s data sets including 4549 interactions (1532 nonduplicated interactions) in the full data set (Ito et al. 2001) are based on yeast two-hybrid assays, whereas Schwikowski’s data set, reporting 2709 interactions (1157 nonduplicated interactions), was gathered from yeast two-hybrid, biochemical, and genetic data (Schwikowski et al. 2000). Interacting pairs of proteins are visualized as lines drawn between two genes on the topography. Because the protein-protein interaction data is binary—that is, proteins either interact or they do not—the relative strength of the interactions is not a parameter that can be used for visualization.

The impression from both data sets is that the complete set of interacting proteins creates a network over the entire expression topography (Fig. 3-5 (A,B); see supplemental data). At this level of analysis, differences in the structure of the data can

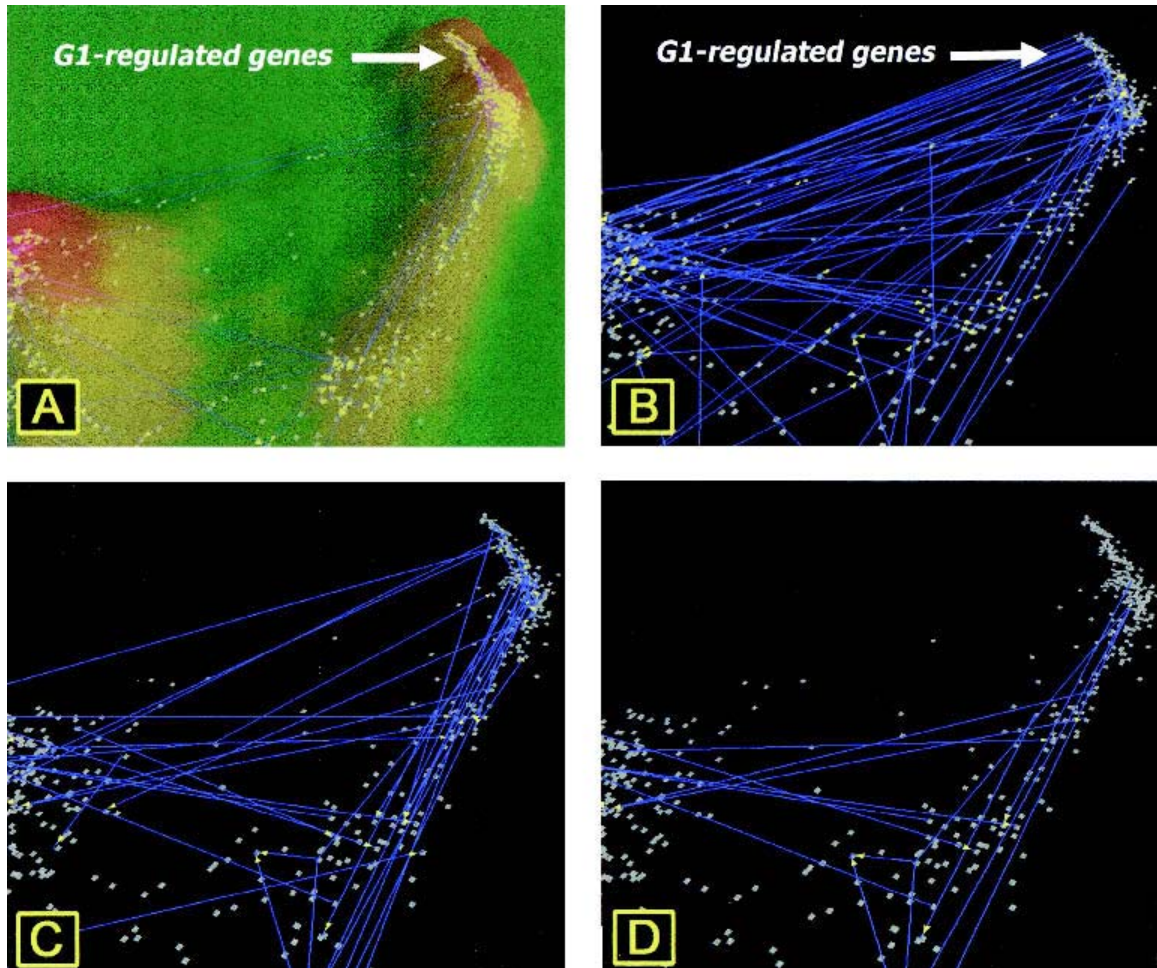
be detected only at the margins. When the protein interactions that are common to both data sets are visualized in VxInsight, the previously reported lack of overlap in the two data sets (Ito et al. 2001) can be clearly seen (only 19% of Schwikowski and 8.3% of Ito's full data sets are in common; Fig. 3-5 (C, D)). Visualization of only the genes encoding interacting proteins common to both data sets (Fig. 3-5(D)) shows that relatively large segments of the topography contain no interacting proteins.



**Figure 3-5. Protein-protein interaction maps as a function of the cell-cycle gene-expression topography.** Lines are drawn between genes encoding interacting proteins and the G1-regulated gene cluster is circled for clarity. (A) Schwikowski's complete data set. (B) Ito's full data set. (C) Protein-protein interactions reported from both data sets. (D) Genes encoding interacting proteins common to both data sets. In A and B, genes encoding proteins involved in interactions are indicated by yellow pyramids.

In both data sets, many interactions are observed between proteins encoded by tightly clustered G<sub>1</sub> phase–regulated genes (Fig. 3-6). Although both data sets contain G<sub>1</sub>-regulated genes that interact with each other, there is little overlap between the data sets (Fig. 3-6 (D)). Ito's data set (Fig. 3-6 (B)) includes many interactions between proteins encoded by genes in the G<sub>1</sub> cluster and an adjacent cluster, containing genes that are not cell-cycle regulated. In contrast, the interactions reported in Schwikowski's data set (Fig. 3-6 (C)) more closely parallel the connections based on strong similarities of gene expression (Fig. 3-6 (D)). In the region of M phase–regulated genes, both data sets report interacting proteins that parallel the strong similarities in gene expression, but with little overlap between the data sets (data not shown). In examining the G<sub>1</sub>-regulated genes reported to be involved in interactions in both data sets, Ito's data set is much more likely to contain genes of unknown function (33 of 78; 42%) than is Schwikowski's data set (5 of 50; 10%; data not shown). Furthermore, there are no genes in the main G<sub>1</sub>-regulated cluster that encode interactive proteins common to both data sets (Fig. 3-6 (D)).





**Figure 3-6. Interactions among proteins encoded by  $G_1$ -regulated genes.** (A) Topographical presentation of  $G_1$ - regulated gene cluster with connections between genes showing strong similarities ( $R > 0.887$ ) of expression between genes. (B) Genes encoding interacting proteins from Ito's full data set. (C) Genes encoding interacting proteins reported from Schwikowski's data set. (D) Protein interactions in common to the two data sets. Connections between genes in B–D indicate interactions occurring between proteins encoded by the specific genes.

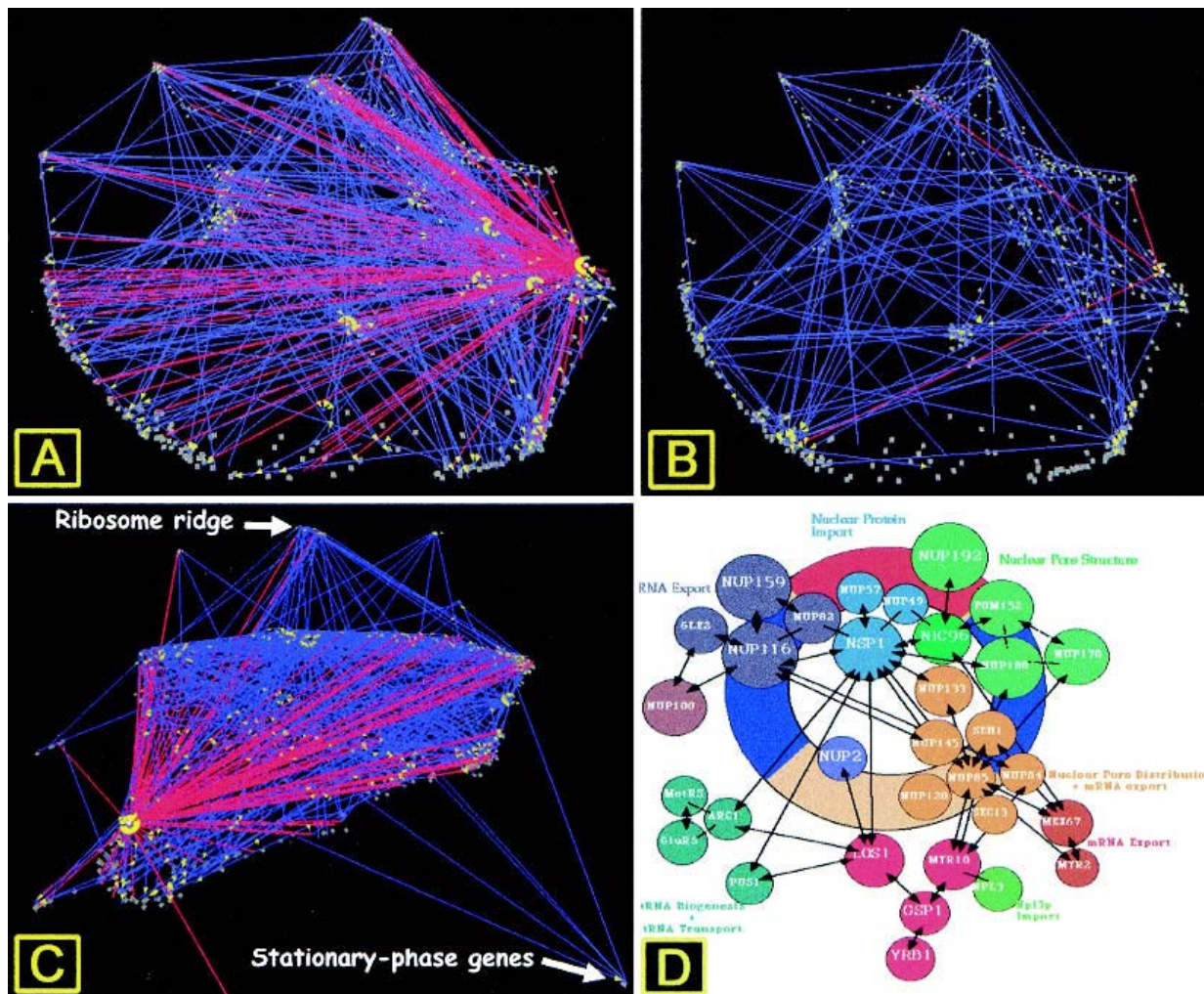
Looking at genes within the  $G_1$ -regulated gene cluster that are reported to interact in each data set, Schwikowski reports an interaction between MSH6 and PMS1, both involved in mismatch repair, whereas Ito reports an interaction between RFL2 and CAC1, both subunits of chromatin assembly factor (CAF-1). The lack of overlap in the two data sets and the presence of reasonable interacting pairs in both data sets indicate

that for the present time, the data sets are most useful when examined concurrently, as was performed in a recent paper (Ge et al. 2001). We conclude from this analysis that the differences in results of both studies could be indicative of the range of detection in the two-hybrid assay and the difficulty in obtaining sample sizes large enough to include the entire set of interactions.

The structures of the two data sets are also distinct. Several genes have significantly more interactions in the Ito data set (Fig. 3-7(A)) than in the Schwikowski data set (Fig. 3-7 (B)). One of these, Nup116p, a nuclear pore protein, is reported to have 125 interactions in the Ito full data set, 15 in the core data set (interactions observed three separate times), and three in the Schwikowski data set (which includes data from the Munich Information Center for Protein Sequences). Nup116p has been shown genetically or biochemically to interact with 15 proteins ([www.Proteome.com](http://www.Proteome.com)), including many involved in nuclear pore function (Fig. 3-7 (D)). Based on information from the Munich Information Center for Protein Sequences, Schwikowski reported three Nup116p-interacting proteins: Kap95p, Kap104p, and Gle2p. Ito, based solely on two-hybrid data, also identified three of these interacting proteins, Gle2p, Nup 82p, and Nup100p, in the full data set (Fig. 3-7 (B)).

Interestingly, when interactions reported in Ito's full data set for Nup116p are visualized as a function of gene expression during exit from stationary phase (Fig. 3-7 (C)), it is striking that there are no interactions between Nup116p and proteins encoded by stationary-phase genes and only three interactions with proteins encoded by genes with expression that increases rapidly after refeeding, including those in ribosome ridge. If Nup116p interactions were randomly distributed, more than nine interactions would

have been expected with proteins encoded by these genes. In ribosome ridge alone, ~125 proteins (of 290) are known to be ribosomal, and nine other proteins are predicted to be nuclear, yet there are only two interactions with proteins encoded by genes in this cluster. Further experiments will be necessary to determine whether this interaction pattern is accurate or reflective of a higher than expected rate of false negatives (Ito et al. 2001) with this assay.



**Figure 3-7. Protein-protein interactions between Nup116p and other proteins.** (A) Ito's full data set: cell-cycle expression topography. (B) Schwikowski's full data set: cell-cycle topography. (C) Ito's full data set: exit from stationary phase topography. (D) Diagram of Nup116p interactions in the nuclear pore from the Munich Information Center for Protein Sequences ([http://vms.gsf.de/htbin/search\\_code/YMR047C](http://vms.gsf.de/htbin/search_code/YMR047C)). (Reprinted, with permission, from E. Hurt, BZH; Universitaet Heidelberg.)

### ***Relative Absence of Ribosomal-Protein Interactions in the Protein-Interaction Data Sets***

Because of the strong similarity in gene expression among the ribosomal protein genes (RPS and RPL genes) during exit from stationary phase, we were interested in examining the interactions among proteins encoded by genes found in ribosome ridge in the exit from stationary-phase data set. Surprisingly, although there was a high degree of similarity of gene expression and some interactions reported between nonribosomal proteins in ribosome ridge, there was only one interaction reported between ribosomal proteins (see Web Supplement). The absence of interactions among these proteins was surprising but consistent with recent structural data, indicating that ribosomal proteins interact primarily with ribosomal RNA and not with each other (Spahn et al. 2001). This observation, which is in contrast to results from immunoprecipitation– mass spectroscopy analysis of protein complexes in which ribosomal proteins are common contaminants (Gavin et al. 2002), actually strengthens the confidence in both two-hybrid data sets, indicating that the level of identification of false-positive interactions (Schwikowski et al. 2000), at least among some groups of proteins, is relatively low.

### **Discussion**

An integrative approach to cell function requires the tools to compile and integrate information from different levels of cellular organization (Ideker et al. 2001). We have shown the utility of visual comparison of distinct types of genome-scale data sets. In this process, we were able to conclude that  $G_1$ - regulated genes were not coordinately regulated during exit from stationary phase, indicating that cells exiting

stationary phase are not synchronous or that a subset of  $G_1$ -regulated genes is required for this process.

The hypothesis that the cells in stationary-phase cultures are not synchronous is supported by the observation of different sizes of cells in stationary-phase cultures (Werner-Washburne et al. 1993) and previous studies of reentry into the cell cycle indicating that cells do not bud until they reach a critical size (Johnston et al. 1977). In addition, one report indicated that mammalian cells are not synchronized when induced to grow by refeeding (Cooper 1998), although  $G_0$  arrest by serum starvation is a method commonly used to synchronize mammalian cells (Callard and Mazzolini 1997; Zeise et al. 1998; Hildebrand and Dahlin 2000). If yeast cells can be synchronized during exit from stationary phase; for example, by isolating small unbudded cells, it should be possible to distinguish those changes in gene expression that are physiological in nature (e.g., induction of ribosomal protein genes) from those that are specific for the cell-cycle transition (e.g., expression of cell cycle-regulated genes). The discovery of different genes required for the physiological response and the cell-cycle response could easily lead to the development of novel drug-targeting strategies that are specific for quiescent cells.

The lack of overlap in the two protein-interaction data sets from yeast (Schwikowski et al. 2000; Ito et al. 2001) has been a puzzle to researchers interested in proteomics; to date no clear reason for these differences has been determined. One suggestion was that the size of the cloned genes might have been a factor (Hazbun and Fields 2001). In our analysis, there was no clear reason to exclude data from either data set. A study of the relationship between cell-cycle expression and protein-interaction data

was recently published (Ge et al. 2001) in which the protein-interaction data were combined. This is consistent with our conclusions for the two data sets analyzed here. We hypothesize that the differences between the two data sets could be caused by the ability of two-hybrid analysis to detect a very wide range of interactions, and that the sample size, even in genome-scale analyses, may be too small to detect all of the interactions in one or even in several experiments.

The process of analysis presented here, although extremely useful to researchers interested in the quiescent state, is also meant to serve as an example that can be used by biologists interested in other questions. For example, is it possible to evaluate differences between distinct, but related, developmental pathways by identifying genes that cluster in one expression data set but not in another? Is it possible to identify protein interactions that occur only under specific growth conditions by identifying those conditions in which interacting proteins are clustered as a function of gene expression?

As multi-data set analyses become more common, they will also lead to changes in experimental design, for example, the increased use of time-course experiments and coordination or parallelization of assays for gene expression and protein interactions, abundance, and/or modifications. Additional pressure for these types of experiments will come from the need for complete characterization of complex processes, such as regulatory pathways, involving every level of cellular and multi-cellular organization. Because it is also unlikely that any one level of cellular organization will provide all the critical elements for diagnostics, both basic and applied research will fuel the continued development of more functional and intuitive software tools for this analysis.

## Methods

### *Exit From Stationary Phase: Growth Conditions, RNA Isolation, and Microarray Analysis*

Overnight cultures of yeast cells (S288C) were inoculated into rich glucose-based medium (YPD) and incubated at 30°C with shaking. At day 7, cells were harvested, washed, re-suspended to an OD<sub>600</sub> of 2 in fresh YPD and returned to 30°C. Samples (~40 OD<sub>600</sub> units) were taken at t = 0, 15, 30, 45, and 60 min after cells were re-suspended in fresh rich medium. Cells were harvested by centrifugation at 4°C and washed once with ice-cold water. Cell pellets were stored at -70°C until use.

Total RNA from ~40 OD units of cells was extracted using a modified Gentra protocol. Briefly, cell pellets were re-suspended in 300 µL of cell lysis buffer (Gentra) to which ~0.2 gm of acid-washed beads had been added. The cells were lysed by vortexing for 30 sec followed by 30 sec on ice (six repetitions). DNA and protein were precipitated from the supernatant, and the RNA was further purified with a phenol/ chloroform extraction and DNase treatment.

Radiolabeled ([<sup>33</sup>P]-dCTP) cDNA “probe” was obtained by reverse transcription of total RNA (2 µg) following the protocol from Research Genetics ([www.resgen.com](http://www.resgen.com)). cDNA was purified to remove unincorporated nucleotides, and total incorporated counts were measured by scintillation counting. The entire probe was then hybridized to nylon membranes containing 6144 yeast open reading frames (Research Genetics). Five sets of nylon membranes were hybridized per experiment (one time point per membrane set per hybridization).

Hybridization was detected by phosphor imaging, and the scanned images were uploaded into Research Pathways Image software (Research Genetics) and as

background-subtracted counts into GeneSpring (Silicon Genetics) and VxInsight (Viswave). Data were normalized using the 50th percentile of all measurements as a positive control. Each measurement was divided by this synthetic positive control to obtain relative expression values.

Replicate experiments were performed by stripping the nylon membranes and reprobing (following the protocol from Research Genetics) with a new reverse transcription reaction obtained from the original RNA extracts. Four to five replicates were performed for each time point.

### ***Data Preparation and Analysis with VxInsight***

Gene expression values in tab-delimited data files were used to compute all pairwise correlations between genes. For each gene, the 20 strongest positive correlations were retained and used for clustering. Because the significance of correlations is nonlinear (a change of 0.05 is much more significant for larger correlations than for smaller ones), the correlations were transformed to a T-statistic, which reflects the statistical rareness of the correlation numbers. In each case, the two gene names and the T-statistic for their correlation were passed to the VxOrd clustering program. The algorithm used by VxOrd places genes on a two-dimensional plane with respect to their similarities (i.e., the T-statistics). It minimizes the potential energy of particles (genes) attracted to each other by forces proportional to their similarities and repulsed from each other by a local force proportional to the density of genes in the immediate region of each gene. The details of the ordination are described more fully elsewhere (Davidson et al. 2001). The hills represent gene clusters, which are determined by similarities in gene expression. The topographical distance between genes and clusters is a function of the



similarity of expression between the genes, and the height of the hills in VxInsight corresponds to the number of genes beneath them.

We decided to identify as strongly correlated, all gene pairs that could have true correlations,  $\rho$  exceeding 0.95. To find the appropriate critical value for R, the sample correlation rather than the assumed underlying true correlation,  $\rho$ , we used the approach described in Davidson et al. (2001). Briefly, if two genes have some true long-term correlations (e.g.,  $\rho = 0.95$ ) and we measure these two genes with only 18 microarray experiments, our particular sample correlation will often fall below  $R = 0.95$ . For any critical value we might choose, there would be a risk of some rare set of 18 experiments yielding a sample correlation less than our selected value. However, we can control that risk by choosing a critical value such that the chance of seeing one of those misleading sample correlations is acceptably small. So, for example, in our analysis we were willing to accept the chance of missing a pair of strongly correlated genes (with a true long-term correlation,  $\rho \geq 0.95$ ) only one time in 20. The analysis described in Davidson et al. (2001) indicates that the critical value for the observed sample correlations should be  $R > 0.887$ . Gene pairs passing this test are identified as being strongly correlated in our analysis.

### ***Identification of Highly Correlated, G<sub>1</sub>-Regulated Genes***

Genes that are strongly up-regulated in G<sub>1</sub>-phase in the  $\alpha$ -factor arrest/cell cycle data set show sharp increases in the third through fifth experiment and then again in the 11th through 13th experiment and are much lower at all other times (Spellman et al. 1998). To generate a list of these genes, we computed the dot product of the expression of

every gene with a vector having +1 values where  $G_1$ -regulated genes would be expected to be up-regulated, and -1 values elsewhere. These dot products were sorted and the largest of them were used to identify the strongest  $G_1$ -regulated genes.

### ***Testing the Significance of the Clustering for Ribosomal-Protein Genes***

To answer the question “Are two mountains in the VxInsight map significantly different from each other?” we compared the empirical distribution of pair-wise correlations in each mountain, and also the distributions of correlations between the two mountains. There are three ways clusters could systematically differ from each other:

1. Expression correlations within each of the two mountains could be very different from each other and also different from the intermountain correlations.
2. The correlations might be vaguely similar in each of the mountains, but their intermountain correlations could be noticeably different from the correlations in either mountain.
3. The correlations in each mountain could be noticeably different from each other, but the intermountain correlations could have some intermediate value, such that the intermountain correlations could not be detected as being different from either of the mountains, even if the mountains were, themselves, statistically different.

The first case corresponds to strongly separated clusters, the second to weakly separated clusters, and the third case corresponds to a gradual gradation from one cluster into another. However, there is only one way that the genes can be incorrectly separated into different groups: that is if all three groupings are found to be indistinguishable. If the gene expressions for genes in, and between, the two mountains were really

indistinguishable (the null hypothesis), then analysis of variance (ANOVA) should fail to detect a significant difference between the means of the three sets of correlations. We tested a number of clusters using ANOVA to assure ourselves that the clustering was significant.

Briefly, we started with two nonintersecting gene lists, GroupA and GroupB. We computed all possible correlations between the genes in GroupA, all possible correlations between genes in GroupB, and finally the correlations between every gene in GroupA with every gene in GroupB. These individual correlations were transformed to their corresponding T-statistics, which are directly related to the P values associated with observing the correlations when the expressions are not actually correlated. ANOVA was performed to test if the mean correlations for these three different groups were significantly different. Under the null hypothesis, one would rarely (the ANOVA P value) see large F-statistics from this analysis. On the other hand, ANOVA should uncover a difference if the genes in the two VxInsight clusters were correctly separated into different groups. That is, we expect ANOVA to yield a very small P value when the expressions for genes in either mountain are more like the expressions for genes in the same mountain than they are for genes in the other mountain. Further, when the correlations between the two clusters are different from the correlations in at least one of the mountains, ANOVA should also allow us to reject the null hypothesis. In either case, we would conclude that the VxInsight clusters are not artifacts.

## **Acknowledgements**

This paper is dedicated to the memory of Judith Galbraith. We would like to thank Andreas Wagner for careful reading of the manuscript and the members of our laboratories for extremely helpful discussions. This work was funded by grants from National Science Foundation (MCB-0092374) to M.W.W., National Institutes of Health Initiatives for Minority Student Development (NIH-IMSD 1R25 GM60201-01) to J.W., and by Laboratory Directed Research and Development, Sandia National Laboratories, U.S. Department of Energy (DEAC04- 94AL85000).

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked “advertisement” in accordance with 18 USC section 1734 solely to indicate this fact.

## **References**

- Aach, J., Rindone, W., and Church, G. 2000. Systematic management and analysis of yeast gene expression data. *Genome Res.* 10: 431–445.
- Callard, D. and Mazzolini, L. 1997. Identification of proliferation-induced genes in *Arabidopsis thaliana*: Characterization of a new member of the highly evolutionarily conserved histone H2A.F/Z variant subfamily. *Plant Physiol.* 115: 1385–1395.
- Chu, S., DeRisi, J., Eisen, M., Mulholland, J., Botstein, D., Brown, P.O., and Herskowitz, I. 1998. The transcriptional program of sporulation in budding yeast. *Science* 282: 699–705.

Clark, W. and Gillespie, D.A.F. 1997. Transformation by v-Jun prevents cell cycle exit and promotes apoptosis in the absence of serum growth factors. *Cell Growth Differ.* 8: 371–380.

Cooper, S. 1998. Mammalian cells are not synchronized in G<sub>1</sub>-phase by starvation or inhibition: Considerations of the fundamental concept of G<sub>1</sub>-phase synchronization. *Cell Prolif.* 31: 9–16.

Davidson, G.S., Wylie, B.N. and Boyack, K. 2001. Cluster stability and the use of noise in interpretation of clustering. *Proc. IEEE Information Visualization 2001*, 23–30.

Eisen, M.B., Spellman, P.T., Brown, P.O., and Botstein, D. 1998. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci.* 95: 14863–14868.

Ferea, T L., Botstein, D., Brown, P.O., and Rosenzweig, R.F. 1999. Systemic changes in gene expression patterns following adaptive evolution in yeast. *Proc. Natl. Acad. Sci.* 96: 9721–9726.

Gasch, A., Spellman, P., Kao, C., Carmel-Harel, O., Eisen, M., Storz, G., Botstein, D., and Brown, P. 2000. Genomic expression programs in the response of yeast cells to environmental changes. *Mol. Biol. Cell* 11:, 4241–4257.

Gavin, A.C., Bosche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J.M., Michon, A.M., Cruciat, C.M., et al. Functional organization of the yeast proteome by systematic analysis of protein complexes. 2002. *Nature* 415: 141–147.

Ge, H., Liu, Z., Church, G.M., and Vidal, M. 2001. Correlation between transcriptome and interactome mapping data from *Saccharomyces cerevisiae*. *Nat. Genet.* 29: 482–486

Goffeau, A., Barrell, B.G., Bussey, H., Davis, R.W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J.D., Jacq, C., Johnston, M., et al. 1996. Life with 6000 genes. *Science* 274: 546–567.

Hazbun, T.R. and Fields, S. 2001. Networking proteins in yeast. *Proc. Natl. Acad. Sci.* 98: 4277–4278.

Hildebrand, M. and Dahlin, K. 2000. Nitrate transporter genes from the diatom *Cylindrotheca fusiformis* (Bacillariophyceae): mRNA levels controlled by nitrogen source and by the cell cycle. *J. Phycol.* 36: 702–713.

Ideker, T., Galitski, T., and Hood, L. 2001. A new approach to decoding life: Systems biology. *Annu. Rev. Genomics Hum. Genet.* 2: 343–372.

Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M., and Sakaki, Y. 2001. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl. Acad. Sci.* 98: 4569–4574.

Johnston, G.C., Pringle, J.R., and Hartwell, L.H. 1977. Coordination of growth with cell division in the yeast *Saccharomyces cerevisiae*. *Exp. Cell Res.* 105: 79–98.

Joshi, U.S., Chen, Y.Q., Kalemkerian, G.P., Adil, M.R., Kraut, M., and Sarkar, F.H. 1998. Inhibition of tumor cell growth by p21(WAF1) adenoviral gene transfer in lung cancer. *Cancer Gene Ther.* 5: 183–191.

Lasharki, D.A., DeRisi, J.L., McCusker, J.H., Namath, A.F., Gentile, C., Hwang, S.Y., Brown, P.O., and Davis, R.W. 1997. Yeast microarrays for genome wide parallel genetic and gene expression analysis. *Proc. Natl. Acad. Sci.* 94: 13057–13062.

Murray, P.J. 1999. Defining the requirements for immunological control of mycobacterial infections. *Trends Microbiol.* 7: 366– 372.

Pajic, A., Spitkovsky, D., Christoph, B., Kempkes, B., Schuhmacher, M., Staeger, M.S., Brielmeier, M., Ellwart, J., Kohlhuber, F., Bornkamm, G.W., et al. 2000. Cell cycle activation by c-myc in a Burkitt lymphoma model cell line. *Int. J. Cancer* 87: 787–793.

Schwikowski, B., Uetz, P., and Fields, S. 2000. A network of protein-protein interactions in yeast. *Nat. Biotechnol.* 18: 1257–1261.

Spahn, C.M.T., Beckmann, R., Eswar, N., Penczek, P.A., Sali, A., Blobel, G., and Frank, J. 2001. Structure of the 80S ribosome from *Saccharomyces cerevisiae*: tRNA-ribosome and subunit-subunit interactions. *Cell* 107: 373–386.

Spellman, P.T., Sherlock, G., Zhang, M.Q., Iyer, V.R., Anders, K., Eisen, M.B., Brown, P.O., Botstein, D., and Futcher, B. 1998. Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell* 9: 3273–3297.

Tomee, J.F.C., Hiemstra, P.S., Heinzl Wieland, R., and Kauffman, H.F. 1997. Antileukoprotease: An endogenous protein in the innate mucosal defense against fungi. *J. Infect. Dis.* 176: 740–747.

Werner-Washburne, M., Braun, E., Johnston, G.C., and Singer, R.A. 1993. Stationary phase in the yeast *Saccharomyces cerevisiae*. *Microbiol. Rev.* 57: 383–401.

Winzler, E.A., Shoemaker, D.D., Astromoff, A., Liang, H., Anderson, K., Andre, B., Bangham, R., Benito, R., Boeke, J.D., Bussey, H., et al. 1999. Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis. *Science* 285: 901–906.

Zeise, E., Kuhl, N., Kunz, J., and Rensing, L. 1998. Nuclear translocation of stress protein Hsc70 during S phase in rat C6 glioma cells. *Cell Stress Chaperones* 3: 94–99.

Zeitler, H., Ko, Y., Glodny, B., Totzke, G., Appenheimer, M., Sachinidis, A., and Vetter, H. 1997. Cell-cycle arrest in G<sub>0</sub>/G<sub>1</sub> phase of growth factor-induced endothelial cell proliferation by various calcium channel blockers. *Cancer Detect. Prev.* 21, 332–339.

### **Web Site References**

<http://www.Proteome.com>; Nup116p has been shown genetically or biochemically to interact with 15 proteins. [http://vms.gsf.de/htbin/search\\_code/YMR047C](http://vms.gsf.de/htbin/search_code/YMR047C); Munich Information Center for Protein Sequences. <http://genome-www.stanford.edu/Saccharomyces/>; Stanford Genome Database

Received November 26, 2001; accepted in revised form July 31, 2002.



## Chapter 4: The Proteomics of Quiescent and Non-Quiescent Cell Differentiation in Yeast Stationary-Phase Cultures

*Authors.* George S. Davidson,<sup>\*,†</sup> Ray M. Joe,<sup>\*</sup> Sushmita Roy,<sup>‡</sup> Osorio Meirelles,<sup>\*</sup> Chris P. Allen,<sup>§</sup> Melissa R. Wilson,<sup>\*</sup> Swagata Chakraborty,<sup>\*</sup> Anne E. Dodson,<sup>\*</sup> Elaine E. Manzanilla,<sup>\*</sup> Mark Carter,<sup>§</sup> Susan Young,<sup>§</sup> Bruce Edwards,<sup>§,||</sup> Larry Sklar,<sup>§,||</sup> Phillip H. Tapia,<sup>\*</sup> Margaret Werner-Washburne<sup>\*,||</sup>

*Affiliations:* <sup>\*</sup>Biology Department; <sup>‡</sup>Computer Science Department, University of New Mexico, Albuquerque, NM, 87131; <sup>§</sup>Department of Cytometry; <sup>||</sup>Department of Pathology, University of New Mexico School of Medicine, Albuquerque, NM, 87131; <sup>†</sup>Sandia National Laboratories, Albuquerque, NM, 81185.

*Corresponding Author.* Margaret Werner-Washburne  
Mailing Address: MSC03-2020; Biology Department; University of New Mexico;  
Albuquerque, NM 87131  
Phone Number: 505-277-9338  
Fax Number: 505-277-0304  
Email Address: [maggieww@unm.edu](mailto:maggieww@unm.edu)

Running Head: Proteomics of SP yeast cultures

*Abbreviations List* exponential growth phase (EXP), stationary phase (SP), quiescent (Q), non-quiescent (NQ), reactive oxygen species (ROS), green fluorescent protein (GFP), dihydroethidium (DHE), differential interference microscopy (DIC), Stanford Genome Database (SGD).

GSD contributions: Data analysis, literature research covering our genes of interest, overall organization, first drafts and revisions of the paper. Statistical methods and computer analysis of raw data files through Bayesian models of the multidimensional data, first application of the Earth Mover Distance metric to the analysis of flow cytometry data.

## **Abstract**

Yeast cultures enter stationary phase in rich, glucose-based medium in response to carbon starvation. During this process, differentiation of two major subpopulations of cells, termed quiescent and non-quiescent, has been observed. Differences in mRNA abundance between exponentially growing and stationary-phase cultures and quiescent and non-quiescent cells have been identified. To measure changes in protein abundance between exponential and stationary-phase cultures, the yeast GFP-fusion library (4156 strains) was examined during exponential and stationary-phases, using high-throughput flow cytometry (HyperCyt®). About 5% of proteins in the library showed 2-fold or greater changes in median fluorescence intensity (abundance) between the two conditions. We identified and characterized 38 strains exhibiting two distinct peaks of fluorescence-intensity in SP and determined that the two fluorescence peaks identified quiescent and non-quiescent cells, indicating these are the two major subpopulations. Most proteins that distinguished quiescent and non-quiescent cells were more abundant in quiescent cells and were involved in mitochondrial function, consistent with the 6-fold increase in respiration observed in quiescent cells. Examination of the induction of quiescent-cell specific proteins found symmetry in protein accumulation in dividing cells after glucose exhaustion and led to a new model for the differentiation of quiescent and non-quiescent cells.

## Introduction

The yeast *Saccharomyces cerevisiae* is a major model system that is seldom considered for studies of cellular differentiation, especially the differentiation of cell types within the same culture. However, when yeast cultures, grown in rich, glucose-based medium, exhaust glucose, two cell fractions: quiescent (Q) and non-quiescent (NQ), do differentiate and, by two days after glucose exhaustion (3 days after inoculation), are separable by density-gradient centrifugation (Allen *et al.*, 2006).

Q cells, in contrast to cells in the NQ fraction, are uniform, unbudded, bright (refractile) by phase-contrast microscopy, relatively dense, stress-resistant, and most (>90%) are virgin daughters. They are synchronous when re-fed and nearly 100% reproductively competent. They contain thousands of mRNAs in insoluble protein-RNA complexes from which specific mRNAs are released in a stress-specific manner (Aragon *et al.*, 2006).

The NQ fraction, in contrast, contains budded and unbudded cells comprised of approximately equal numbers of mothers and daughters, and few sequestered mRNAs. This fraction is not synchronous when re-fed, but retains viability while rapidly losing reproductive capacity, independent of replicative age, making it a model for, among other things, the viable but unculturable state (Lewis, 2007). Of NQ cells that can reproduce, 40% form petite colonies, consistent with previous reports of genomic rearrangements and transpositions in stationary phase (SP) or glucose-limited cultures (Dunham *et al.*, 2002; Coyle and Kroll, 2008). The most abundant, soluble mRNAs in NQ cells encode proteins involved in DNA recombination and repair and Ty-element transposition, consistent with their being genomically unstable (Aragon *et al.*, 2008). By

14-days post-inoculation, about 50% of NQ cells are apoptotic. The differences between Q and NQ cells and the preponderance of virgin daughters in Q fractions raise questions about the origins and differentiation of these populations, especially, the virgin daughters in Q and NQ cell fractions.

Large, robust, transcriptome data sets are available for Q and NQ cells (Aragon *et al.*, 2008), but there are no extensive proteomic data for these fractions. Until this paper, the only proteomic data available were from two-dimensional, polyacrylamide gel-electrophoretograms from studies of protein synthesis in cultures grown to stationary phase in rich medium (Fuge *et al.*, 1994). That analysis demonstrated that, although protein synthesis decreases as cultures approached stationary phase, major changes in protein synthesis are observed immediately after the cultures exhaust glucose at the diauxic shift. Because only a small percentage of total cellular proteins can be visualized in this assay, proteomic-level insight into the origins and differentiation of Q and NQ cells requires a more comprehensive proteomic assay.

To obtain quantitative data for abundance of more than 2/3 of yeast proteins, the yeast GFP-fusion library (4156 strains, each tagged at the 3' end (coding strand) of the ORF with a GFP-encoding gene) (Huh *et al.*, 2003; Howson *et al.*, 2005) was screened, in triplicate, during exponential (EXP) and stationary phase (SP), using high-throughput flow cytometry (HyperCyt®) (Edwards *et al.*, 2004). The GFP-fusion library was developed as a tool for *in vivo* analysis of protein abundance and localization at the level of the proteome. The strain library, which represents about 75% of all yeast genes, has been validated and used to localize proteins in cells in exponential phase cultures (Huh *et al.*, 2003). It has also been used to examine the relationship between mRNA and protein

abundance (Newman *et al.*, 2006) and this and similar libraries have been used to model the factors that contribute to differences in protein abundance at the cellular level (Raser and O'Shea, 2004, 2005; Newman *et al.*, 2006). However, to our knowledge, the entire library has not previously been used to examine differences in protein abundance between two environmental conditions, such as EXP in rich, glucose-based medium and SP.

We report here that flow-cytometry analysis of approximately 25,000 GFP-fusion strain samples in EXP and SP revealed that only 3% of GFP-fusion proteins showed a two-fold or greater change in abundance between EXP and SP. Abundant EXP proteins are involved in biosynthetic processes while abundant SP proteins are involved in mitochondrial function. To find GFP-fusion proteins that might distinguish Q from NQ cells, we identified 38 strains with distinct double peaks of fluorescence in the flow cytometry data from unfractionated SP cultures. All 38 exhibited higher fluorescence intensity in the Q fraction. Most of these strains carried GFP-fusions in mitochondrial proteins, many of which are involved in respiration. This observation is consistent with our finding that respiration was significantly higher in Q than NQ cells. Examination of *Cit1p*:GFP and *Acs1p*:GFP strains, which express GFP-fusion proteins almost exclusively in Q cells, revealed that daughter cells produced after the diauxic shift express the same level of GFP protein as the mother, i.e., dim NQ mothers produce dim NQ daughters while bright, GFP-producing mothers produce bright daughters. This observation leads to a new model for the production of Q and NQ cells in stationary-phase cultures.

## Materials and Methods

**Growth conditions.** For the GFP HyperCyt® screen, individual strains from the Yeast GFP Collection (Huh *et al.*, 2003), constructed from the parental strain ATCC 201388: *MATa his3Δ1 leu2Δ0 met15Δ0 ura3Δ0* (S288C) (Brachmann *et al.*, 1998), were replicated into 96 well plates containing YPD + A (2% yeast extract, 1% peptone, 2% glucose, 0.04 mg/mL adenine) and 50 µg/ml ampicillin; (Rose *et al.*, 1990) using pin tools. The plates were covered with Breathe Easy sealing membranes (Sigma Aldrich cat #380059) and the strains were cultured at 30°C with aeration either overnight (for exponential growth) or for 7 days (for stationary-phase growth). For the 38 subpopulation strain analysis, wild-type (S288c) and the yeast GFP-fusion set (Huh *et al.*, 2003) were used for analysis. Strains were cultured in YPD + A (2% yeast extract, 1% peptone, 2% glucose, 0.04 mg/mL adenine, and 50 µg/ml ampicillin) at 30°C for 7 days for stationary phase growth.

**Cell Separation and Harvest.** Percoll™ (GE Healthcare) density gradients were made using a solution of one part 1.5M NaCl per 8 parts of Percoll™ (vol/vol) (Allen *et al.*, 2006). The gradients were formed using 10-ml aliquots of this solution in 15 ml Corex tubes which were centrifuged at 24,700 g for 15 min at 4°C. In order to separate the fractions, 5 ml samples of 7 day stationary-phase yeast cultures were pelleted, resuspended in 500 µl of 50 mM Tris HCl buffer pH 7.5 and overlaid onto these gradients. The gradients were then centrifuged at 400 g for 60 min at 25 °C in a tabletop centrifuge with a swinging bucket rotor (Allegra X12-R, Beckman). The resulting fractions were collected and washed in 13 ml of Tris buffer. The pellets were

resuspended in 1 ml of Tris buffer and cell density was ascertained using the Z2 Coulter Counter (Beckman). The cells were again pelleted and then suspended in 100 µl of their own respective stationary phase conditioned media for analysis.

***High-throughput flow-cytometric screening.*** Three steps were used to prepare the samples for high-throughput screening. First, dilution plates were prepared by transferring 90 µL of peptide dilution flow buffer (30mM HEPES buffer, pH7.4, 110mM NaCl, 10mM KCl, 1mM MgCl<sub>2</sub>, 10mM Glucose and 0.1% BSA) into each well of the 384-well plates (Greiner Bio-one Cat #781280) using the Biomek NXMC (Beckman Coulter, Fullerton, CA.) liquid handling robot. Second, 10 µL of each yeast strain were transferred from the 96-well growth plates into three adjacent wells of the 384-well dilution plates using the Biomek NXS8 (Beckman Coulter) liquid handling robot. This step created a 1:10 dilution and generated three technical replicates for each sample. The 4<sup>th</sup>, 8<sup>th</sup>, 12<sup>th</sup>, 16<sup>th</sup>, 20<sup>th</sup>, and 24<sup>th</sup> columns of the dilution plates did not contain samples, just buffer alone. These columns served as a wash well used between different samples to minimize sample carryover. Third, the cells were sampled with a HyperCyt® (Edwards *et al.*, 2004; Young *et al.*, 2005) autosampler controlled by HyperSip software and interrogated for GFP fluorescence with a CYAN ADP (Dako Cytomation, Ft. Collins, CO) flow cytometer using excitation at 488 nm and collection of fluorescent emissions with a 530/40 nm filter set. The data were processed using IDLeQuery software and the median channel fluorescence for each sample was calculated and used for subsequent analyses.

***Low-throughput flow cytometry and MoFlo-based cell sorting.*** For re-analysis of the 38

strains and Q/NQ fractions, approximately  $5 \times 10^6$  cells were suspended in 500  $\mu$ l of filter sterilized (0.22  $\mu$ m) Tris buffer in 2-ml flow tubes. These were analyzed for GFP fluorescence intensity using the Accuri C6 Flow Cytometer with the FL-1 channel. 30,000 events were acquired for each of 3 technical replicates. Data were analyzed with IDLeQuery software. For single-cell growth studies, cells were sorted based on fluorescence and 144 cells positioned per YPD agar plate using a MoFlo cell sorter (Coulter). Three plates were sorted per sample, e.g., high GRE low ROS, and results are means  $\pm$  standard deviation

***DHE assay for quantification of ROS.*** Dihydroethidium (DHE) stock solution (Invitrogen) was diluted 1:10 in PBS (Fluka) for a working solution. Approximately  $1 \times 10^8$  S288c upper and lower fraction cells per sample were pelleted and resuspended in 100  $\mu$ l of the YPD+a, supernatant that had been filter sterilized. 1  $\mu$ l DHE working solution was added to each sample and incubated for 6 min at room temperature in the dark. The samples were washed three times in PBS. The samples were diluted to  $1 \times 10^6$  cells/ml in Isoton II, and 30,000 cells per sample were analyzed with a FACScan flow cytometer (CLONTECH Laboratories, Inc.) using 488 nm excitation and collecting fluorescent emission with filters at 585/42 nm for FL-1 parameter.

***Microscopy.*** The fluorescent images were obtained using an Axioskop 2 mot *plus* microscope (Carl Zeiss). All of the images were taken with a 50 ms exposure time for the DIC image, 2000 ms exposure time for the Rhodamine filter to detect DHE staining, and 2000 ms and automatic exposure times for the FITC filter to detect GFP. The automatic exposure image was acquired for the purpose of identifying localization of



protein in case the protein expression was too bright or dim for clarity in the 2000 ms image. Axiovision 4.7 software was used to compile and analyze the images.

***Assay for Reproductive Capacity (Colony-forming Units, CFUs).*** Yeast strains were grown to 7 days post-inoculation in YPD and separated into Q and NQ fractions by density gradient centrifugation. For FACS-enabled positioning of NQ and Q cells, samples were sorted using the MoFl cell sorter (Coulter). For each sample, 144 cells (high GFP/low ROS, high ROS/low GFP, and low GFP/low ROS) were positioned on solid, YPD medium. At least 3 plates of 144 cells were obtained per sorted sample, e.g., high GFP/low ROS, and incubated for 2-3 days at 30 °C. The reported values represent the mean  $\pm$ one standard deviation for each sample.

***Rate of oxygen consumption assay.*** Rates of oxygen consumption were determined using the BD<sup>TM</sup> Oxygen Biosensor System (BD Biosciences), which is a 96-well plate containing a fluorophore that fluoresces in the absence of oxygen. Quiescent and nonquiescent cells were separated as described earlier, all samples were diluted to a concentration of  $1 \times 10^8$  cells/ml; 200  $\mu$ l was placed in each well and coated with mineral oil. Fluorescence was measured every minute for one hour, using a microplate reader and SoftMax Pro software. Relative fluorescence was normalized to the average signal of three wells containing conditioned media at each time point. Normal fluorescence units were converted to  $\rho O_2$  using the following equation:  $\rho O_2 = (DR/NRF - 1)/K_{sv}$ , where DR (dynamic range) is the ratio of the signal at zero oxygen to the signal at ambient condition, which was calculated using 100 mM sodium sulfite in PBS buffer; NRF is the normalized relative fluorescence; and  $K_{sv}$  is the Stern-Volmer constant, which was calculated using the following equation and then converted to units of  $\text{atm}^{-1}$ :  $K_{sv} = (DR -$

1)/ $\rho_{O_2A}$ , where  $\rho_{O_2A}$  is the partial pressure of oxygen at ambient conditions.  $\rho_{O_2A}$  was calculated by multiplying the mole fraction of oxygen at ambient conditions (0.209) by the total pressure in Albuquerque (85 kPa). Once the  $\rho_{O_2}$  of each time point for each well was calculated, it was converted to moles of oxygen by dividing by Henry's constant (756.5133 atm·L/mol at 25 °C for air) and multiplying by the volume ( $2 \times 10^{-4}$  L). Rates were determined from the slope of the regression line for Time(s) vs. mol  $O_2$ /cell. Final rates were calculated as mol  $O_2$ /cell/sec and represented as an average of three biological replicates.

***Correlation-based reproducibility analysis comparing GFP measurements across laboratories.*** We compared fluorescence intensities of exponentially growing cells from our laboratory to those from Newman *et al.* (2007). After excluding proteins with no measurements in either data set we had a total of 2,735 proteins. Abundances of these proteins were correlated using Spearman's correlation (0.6554), Pearson's correlation (0.91) and Pearson's correlation on Savage scores of abundances (0.8290).

***Gene Ontology Relations.*** All fluorescence intensity data for all strains in all three replicates in stationary and exponential phases were  $\log_2$  transformed and averaged before computing stationary phase data/exponential phase ratios for each strain. For ratio values greater than two, the Gene Ontology (GO) terms were tabulated using the Gene Ontology Term Finder Database, <http://www.yeastgenome.org/cgi-bin/GO/goTermFinder.pl>.

***IDLeQuery identification of subpopulations in strains.*** HyperCyt® measurements were analyzed with the flow cytometry software, IDLeQuery, provided by the

University of New Mexico Flow Cytometry Facility (Young *et al.*, 2005). Raw count data were gated and binned for plotting. IDLeQuery was used to plot relative distributions of forward scatter and side scatter intensity (the latter were  $\log_{10}$  transformed).

***Slope Differentiation Identification (SDI) Algorithm.*** To identify GFP fusion strains having 2 fluorescence peaks, the stationary phase side scatter (SS) data set was divided into 100 bins; each bin was averaged to compute log-FI. The EXP side scatter (SS) data set was similarly processed to yield log-SS. Then for each bin,  $\Delta\log\text{-FI}$ , the difference between SP and EXP log-FI was computed for each of the three technical replicates. A regression of  $\Delta\log\text{-FI}$  (from the difference between SP and EXP) vs. log-SS (from SP) was computed and the median of the regression slope across the three replicates was used to compute the SDI measure. Near-zero SDI values indicates low correlation, which is suggestive of a single peak of fluorescence intensity in both samples. Higher SDI values occur when there is not a good overlap of peaks, either there are single, non-overlapping peaks in both samples or there are 2 peaks in one sample. Evaluation of the highest 78 strains identified by SDI revealed that 71 (91%) were strains that exhibited one peak in EXP and two peaks in SP (not shown).

***k-means clustering-based two peak identification.*** To identify proteins with two fluorescence peaks, k(=20)-means clustering was performed on each data set using the ratio of side scatter to forward-scatter. The average profile for each cluster was computed, followed by visual identification of clusters with broad or jagged profiles. This analysis identified one cluster of 80 SP samples and one cluster with 99 EXP

samples. Samples from these clusters were compared with candidate heterogeneous strains identified with other methods to identify strains found by all three methods.

## Results

### *The yeast GFP-fusion library was sampled in triplicate for both EXP and SP cultures.*

Although many cDNA microarray experiments are available for EXP cultures, and some studies on cells in SP cultures have been published (Allen *et al.*, 2006; Aragon *et al.*, 2006; Aragon *et al.*, 2008), the only information about changes in protein abundance between EXP and SP is protein abundance and synthesis from 2-D gel analysis of radioactively labeled and unlabeled proteins (Fuge *et al.*, 1994). To better quantify the change in protein abundance in cultures between these phases on a proteome scale, we analyzed the yeast GFP-fusion library (4156 strains, each carrying the GFP gene inserted into a known 3' region of a different gene) (Huh *et al.*, 2003) and the HyperCyt® high-throughput flow cytometer (Edwards *et al.*, 2004). In this assay, strains producing a GFP-fusion protein under the control of a native promoter were assayed in triplicate under the two conditions (~ 25,000 samples).

We found that fluorescence measurements in EXP and SP samples were extremely robust ( $R^2 = 0.995$ ) for 96-well plates containing the same strains sampled more than a month apart (see supplemental data). Previous studies, using an identical GFP fusion set, reported similar reproducibility ( $R^2 = 0.997$ ), i.e., measurement reproducibility between replicate experiments for the same strain (Newman *et al.*, 2006). Comparison of the abundance of 2735 proteins between our results and those of Newman *et al.* gave  $R = 0.91$ , indicating that reproducibility between laboratories is also excellent. Newman found that GFP measurements and tandem affinity purification (TAP)-tag

measurements for those proteins were closely correlated ( $R^2=0.80$ ), comparable to the precision achieved with duplicate western blots ( $R^2=0.77$ ). We conclude from these results that GFP fluorescence measurements are highly reproducible, even between laboratories and that there is strong evidence that GFP intensity is a true measure of protein abundance for the fusion protein.

***Of the top 20 most abundant proteins, 12 (60%) were among the most abundant in both EXP and SP.*** In comparing the top 20 most abundant proteins in EXP and SP, regardless of the change in expression, the 12 proteins that were found in common (Table 4-1) are involved in glycolysis (5 proteins), cell wall biosynthesis (1), translation (including the two translation elongation factors Tef1p and Tef2p that encode EF- alpha elongation factor and Yef3p), nuclear transport (Ssa1p and Ssa2p), and Hsc82p, involved in proteasome assembly (Imai *et al.*, 2003; Le Tallec *et al.*, 2007). Proteins that were among the 20 most abundant in EXP but not SP were Ahp1p, a thiol-specific peroxiredoxin that protects against oxidative damage (Lee *et al.*, 1999) and 3 proteins that are part of the ribosomal stalk. Also included were Pgi1p, which catalyzes the inter-conversion of glucose-6-phosphate and fructose-6-phosphate and is required for cell cycle progression, and Pfk2p, a subunit of phosphofructokinase that is required for glucose induction of cell cycle-related genes (Aguilera, 1986). Gene ontology analysis showed that the proteins with high abundance in EXP were involved in biosynthetic processes, especially translation (40%) (Table 4-1; supplemental data).

Proteins that were most abundant in SP, in addition to those that were in common between EXP and SP, included two ribosomal large-subunit proteins, associated with increased fitness (Rpl41a) and, surprisingly, decreased longevity

(Rpl22a) (SGD, <http://www.yeastgenome.org/>). Abundant proteins were also involved in an NADPH-generating step of the pentose phosphate pathway (Gnd1p), required for resistance to oxidative stress, and glucose phosphorylation (Hxk2p), required for competitive fitness and growth on fermentable carbon sources. Finally, abundant proteins included the vacuolar ATPase (Tfp1p), required for resistance to oxidative stress, and the translation initiation factor, eIF4a (Tif2p), a DEA(D/H)-box RNA helicase (SGD) that is a current target for cancer therapeutics (Lindqvist and Pelletier, 2009; Li *et al.*, 2010). Thus, the proteins that were specifically abundant in SP cultures were generally involved aging and stress responses.

***Characteristics of changes in protein abundance in EXP and SP.*** For cells undergoing such a major metabolic shift, moving from 2% glucose to essentially no fermentable carbon, only 5% of the 4156 GFP-fusion proteins showed changes in abundance  $\geq 2$ -fold under the two conditions: 121 proteins were more abundant in EXP and 87 were more abundant in SP (Figure 4-1). Interestingly, proteins that showed large increases in abundance in cells in SP cultures compared with EXP cultures were typically low abundance proteins in EXP, while many of the proteins with significant increases in abundance in EXP compared with SP were relatively high abundance in cells in SP cultures. In addition, only four of 121 proteins with two-fold or higher abundance in EXP had unknown functions (3.3%). Twenty-one of the 87 proteins (24.1%) with two-fold or higher abundance in SP were of unknown function, suggesting that SP proteins have received relatively less attention than the processes involved in exponential growth. We conclude from this that the EXP to SP transition requires relatively few major changes in protein abundance, suggesting that biochemical regulation may play a major

role in responding to these dramatically different conditions. Secondly, proteins required at higher levels in SP are generally not abundant in EXP, suggesting that new functions might be required for survival in SP. Finally, the significant difference in percentage of abundant SP proteins with unknown function may be indicative of the relatively understudied nature of this part of the yeast life cycle.

***In EXP and SP, proteins involved in different processes increase in abundance.*** Gene Ontology analyses (SGD) revealed that proteins increasing at least 2-fold in SP cultures were involved in respiration, including ATP synthesis and electron transport (Table 4-2; supplemental data), but did not include all the proteins in particular multi-protein complexes. Nine proteins were involved in stress response, primarily oxidative stress, including Hsp12p, and the two superoxide dismutases Sod1p and Sod2p. A similar number of proteins were involved in chromatin silencing, modification, and histone acetylation (see supplemental data). These results are consistent with previous findings that mitochondrial function is important for Q cell survival and that Q cells are stress resistant and genomically stable (Allen *et al.*, 2006; Aragon *et al.*, 2008).

***Some GFP-producing strains exhibited two distinct fluorescent populations in SP.*** We have shown previously that there are two major cellular fractions in SP cultures: Q and NQ (Allen *et al.*, 2006). In searching for an efficient Q/NQ screen, we examined the set of strains with the highest fluorescence intensity in SP and found they were primarily mitochondrial fusion proteins. We then did a microscopic screen of mitochondrial proteins and identified Cit1p:GFP, a citrate synthase, which clearly had two populations of cells in SP and determined that Cit1p:GFP exhibited two fluorescent peaks in SP but

not in EXP (Figure 4-2) (note that traditional median-based analyses would miss these peaks, and would report a biologically misleading intensity). Density gradient separation of cells from SP into NQ and Q cells clearly showed greater abundance of Cit1p:GFP in the Q fraction (Figure 4-3). We tested whether Cit1p:GFP abundance could be used to separate Q and NQ cells by fluorescence-activated cell sorting and found that, based on reproductive capacity and petite formation, Cit1p:GFP-producing cells were essentially identical to Q cells and dim Cit1p:GFP cells were similar to the NQ fraction (see supplement).

Because there are two major subpopulations of cells in SP cultures, we wanted to identify other proteins that had 2 peaks of fluorescence in SP. We wanted to determine how many proteins showed this distribution and, based on the function of these proteins, what they revealed about the physiological differences between the cell types. Three different methods were used to identify strains with two peaks of fluorescence intensity: visual evaluation of the flow-cytometry output for 4156 of ~12,500 samples; k-means clustering; and a statistical method we called Slope-Differentiation Identification (SDI) (see Materials and Methods). Thirty-eight strains were predicted by all three methods to have multiple intensity peaks and were examined further.

***Q and NQ cells were differentiated by the fluorescence peaks of all 38 strains.*** For all 38 strains exhibiting two peaks of fluorescence, GFP-fusion proteins were more abundant in Q cells (Figure 4-4). In addition, 58% (22 of 38) carried mitochondrially localized GFP-fusion proteins (Figure 4-4). Because respiration and oxidative phosphorylation are also the most significant processes for the proteins that increase two-fold or more from EXP to SP, we conclude that the changes in GFP-fusion protein expression in SP were



driven by increases in protein abundance in Q cells. Additionally, we conclude from microscopic analysis that Cit1p:GFP abundance differences between Q and NQ cells are observed in both mothers and daughters in these fractions, i.e., it is not a function of replicative age.

Q:NQ median fluorescence ratios ranged from 37 for cytoplasmic Hsp12p:GFP, which is involved in membrane stabilization during desiccation, to 1.4 for Cox6p:GFP, a cytochrome C oxidase protein. In general, most of the strains with Q:NQ fluorescence ratios  $\geq 5$  produced GFP-fusion proteins that were mitochondrially localized, with the exception of three following strains: the heat shock protein Hsp12p ; the nuclear-localized acetyl Co-A synthetase involved in histone acetylation (Acs1P); and a putative membrane protein of unknown function that associates with lipid rafts and is involved in secretion of proteins with non-classical signal sequences (Nce102p) (SGD). Another protein, Inh1p, is an ATPase inhibitor with typical mitochondrial localization, suggesting that, while mitochondrial profiles in Q cells are robust, ATPase function may be down-regulated. We conclude from these results that abundant proteins in Q cells are consistent with mitochondrial maintenance and long-term survival.

***Most NQ populations in the 38 strains exhibited 2 distinct peaks of fluorescence.*** In the evaluation of separated Q and NQ fractions from the 38 strains described above, we were somewhat surprised to find that separation by density did not result in single peaks in both Q and NQ fractions. In fact, 29 of the 38 strains, carrying mostly mitochondrially localized GFP-fusions (Figure 4-4) showed two peaks of fluorescence intensity in the NQ fraction (see supplement). These strains typically had a larger, low-fluorescence peak and a smaller higher-fluorescence peak, with a slightly lower fluorescence intensity than

that of the Q cell fraction (Figure 4-3). One strain, expressing Htb1p:GFP, a histone 2B GFP fusion, showed two peaks in the Q fraction in 2 of 3 analyses.

To study the subpopulations in NQ-fractions, we examined several strains, including Kgd1p:GFP (a component of alpha-ketoglutarate dehydrogenase), Fmp16p:GFP (a mitochondrial protein of unknown function), Eno1p:GFP (cytoplasmic enolase), Sbp1p:GFP (RNA-binding protein), and Ndi1p:GFP (NADH:ubiquinone oxidoreductase) (SGD). To identify cells with reactive oxygen species (ROS), NQ fractions were also stained with DHE (dihydroethidium). Three subpopulations were observed prior to sorting: cells with high ROS and low GFP, cells with high GFP and low ROS, and cells with both low ROS and low GFP (Figure 4-3). A fourth subpopulation, observed during flow cytometry, had intermediate GFP and low ROS and exhibited colony formation that was intermediate between the high GFP and high ROS cells (see supplement). Cells with both high ROS and high GFP were not observed microscopically, which was confirmed by flow cytometry.

For each cell-sorting experiment, at least 3 x 144 individual cells were plated from each of the three populations. Sorted cells were evaluated for colony formation/reproductive capacity (Figure 4-5) and petite formation (see supplement). A representative experiment, using the mitochondrially localized Kgd1p:GFP showed that cells in the NQ fraction with high levels of GFP were similar in viability and colony-forming units to Q cells (Figure 4-5). In contrast, cells with high ROS and no GFP showed significant reduction in colony-forming units, typical of NQ cells. Finally, cells containing little or no GFP-fusion protein and low ROS exhibited an intermediate loss of reproductive or colony-forming capacity. Hence, while high ROS does correlate with

loss of reproductive capacity in NQ cells, cells that are low ROS, low GFP show loss of reproductive capacity, suggesting other factors are likely to be involved in this phenotype. The production of petite colonies, indicative of mutation in mitochondrial proteins, showed a similar pattern, with high GFP low ROS cells producing few petite colonies while cells with either high ROS and low GFP, or low ROS alone frequently produced similar numbers of petites (see supplement). Thus, the high GFP low ROS cells found in the less-dense NQ fraction have several characteristics of Q cells, including genome stability, reproductive capacity, and mitochondrial integrity, leading to the conclusion that increased density may not be necessary for quiescence.

***Q cells have greater mitochondrial function than cells in the NQ fraction.*** To determine whether there were significant differences in respiration between Q and NQ cells, we evaluated oxygen utilization using a BD™ Oxygen Biosensor System (BD Biosciences). SP cultures utilized oxygen at a rate 63% of that for EXP cultures (Figure 4-6). Because cells in SP cultures were assayed in their own medium, which is depleted of carbon, the low rate of respiration was not surprising. Separated Q cells consume six times more oxygen than NQ cells ( $p \leq 5.5E-6$ ) and 1.6 times more oxygen than is used by EXP cultures. This result is consistent with the differences in mitochondrial protein abundances observed above and suggests that most cells in the NQ fraction do not respire or have extremely low levels of respiration in SP.

***Changes in fluorescence intensity shows populations diverge in the first 24 hours after glucose exhaustion.*** We do not yet know the process leading to the differentiation of Q and NQ cells. To begin studying this process, we examined cultures producing the

mitochondrial protein Cit1p:GFP by flow cytometry from 1–7 days after inoculation, the time during which cultures are in the post-diauxic phase, non-fermentable carbon sources are still available, and Q and NQ cells are first observed (Figure 4-7). Initially, there was a general increase in Cit1p:GFP abundance in the whole population, shown by a shift in the single peak to higher fluorescence intensity from 3 hours prior to 9 hours after glucose exhaustion at the diauxic. By 10 hours post diauxic, a second, dimmer population appears. The second peak becomes larger and shifts to lower fluorescence intensity (decreased protein concentration) through the time course, and corresponds to cells from an NQ fraction. The high fluorescence intensity peak continues to increase in fluorescence up to 24 hours post diauxic and then broadens by 144 hours. This peak typically represents the Q fraction. We conclude from this time course data that cells in the post diauxic phase are dynamic and that Cit1p:GFP has the potential to give valuable information in studying this process, especially from 2 or 3 days post diauxic (3-4 days post inoculation) to SP or 7 days post inoculation.

***In post-diauxic populations containing Cit1p:GFP or Acs1p:GFP, almost 100% of mother:daughter pairs are either both bright or both dim.*** Cells from cultures producing Cit1p:GFP or the nuclear protein Acs1p:GFP, both of which typically have bright Q cells and dim NQ cells, were examined by fluorescence microscopy at days 3, 5, and 7 after inoculation. In these cultures, ~40% of the NQ cells (less dense fraction) were budded while none of cells in the more dense or Q fraction were budded. We discovered that, at day 3, when mother:daughter relationships could be clearly determined, essentially all of the mother cells showed symmetry with respect to protein abundance (Figure 4-8). That is, bright, GFP-producing mothers gave rise to bright daughters and

dim mothers gave rise to dim daughters. For Acs1p:GFP, symmetric protein expression during cell division was found 98.7% of the time (n= 228) and for Cit1p:GFP, symmetric protein expression during cell division was found 100% of the time (n = 209). Similar results were found for days 5 and 7 (see supplemental data). Previous examination of virgin daughters in NQ fractions showed that they have the same characteristics, with respect to reproductive capacity and petites as the NQ mother cells (Allen *et al.*, 2006). We conclude from this result that at least during the post-diauxic phase, cells are committed to becoming Q or NQ and produce daughters that are committed to that fate. Cell division takes place predominantly in the less-dense fraction, and the Q fraction is mostly virgin daughters, so mother cells seem to be unlikely to become dense again after division and probably transition to NQ cells. Finally, because these cultures can be started from a single yeast cell (i.e., one cell type begets two types), cell fate must be fixed at some point prior to our observation of symmetry of protein abundance. Because Q and NQ cells can be re-grown to produce both Q and NQ (mother and daughter) cells, we hypothesize that this switch is epigenetic. We do not yet know what controls cell fate in yeast post-diauxic cultures, but this observation clearly deserves more study.

## **Discussion**

We have demonstrated the utility of analyzing the yeast GFP-fusion library with high-throughput flow cytometry to uncover underlying phenotypes and population structure and to interrogate previously intractable biological processes. We quantified protein abundance in EXP and SP and examined the origins of Q and NQ cell phenotypes. We identified tools for in-depth studies of these cells and demonstrated that Q/NQ differentiation is more complex than previously thought.

These studies revealed the heterogeneity of NQ fractions with respect to protein accumulation and reproductive capacity and the relative homogeneity of Q cells, consistent with previous studies (Allen *et al.*, 2006). However, Q cells in the *HTB1:GFP* strain sometimes exhibited 2 fluorescent peaks (see supplement). Because, in prototrophic cells, DNA content analysis of Q cells revealed a single peak and Q cells are extremely synchronous (Allen *et al.*, 2006), we suspect this is an artifact. We are currently investigating the basis for this heterogeneity.

These results helped refine our model of this process (Figure 4-9). The significance of mitochondrial function for Q cells is consistent with previous studies (Martinez *et al.*, 2004; Aragon *et al.*, 2008), but the ability to study these cells through fluorescence differences, especially in mitochondrial proteins, led to discoveries. The surprising finding of symmetry in protein expression in post-diauxic cells is novel and suggests that cell fate is determined prior to Cit1p or Acs1p:GFP accumulation. Because Q and NQ cells can be re-grown to produce Q and NQ cells in SP, the cell fate determinant is likely to be an epigenetic change. However, once a cell has become NQ, its contribution to future generations becomes much less likely because of hypermutability, loss of reproductive capacity and mitochondrial function, and, ultimately, apoptosis. Nevertheless, NQ cells can contribute significantly to species survival. The high viability and loss of reproductive capacity in NQ cells suggests they have two major roles: providing nutrients to Q cells and genetic novelty to the species. Nature ensures a physical connection between NQ and Q cells through flocculation of wild-type yeast and this has recently been suggested to provide self-self recognition and the ability to form biofilms (Smukalla *et al.*, 2008). We have shown that Q cells sequester

many of the mRNAs that are abundant in NQ cells that would translate into proteins required during DNA damage, thus, in the absence of NQ cells, Q cells can become NQ. What we do not yet know is whether Q cells, like *C. elegans* egg cells (Andux and Ellis, 2008), are programmed to enter apoptosis over time to extend the lifespan of the remaining population of quiescent cells.

Published characteristics of Q cells, NQ daughters and mothers and cycling G1 cells reveal important differences between these cells (Table 4-3). For example, both Q and NQ fractions contain virgin, daughter cells that differ significantly in sequestration of mRNA in protein-mRNA complexes, mitochondrial function, reproductive capacity (Allen *et al.*, 2006; Aragon *et al.*, 2008). Loss of reproductive capacity has been suggested to be due to replication stress (Burhans and Weinberger, 2007), implying that NQ cells may have poor checkpoint control. It is our hope that more comparative analyses will identify the regulatory-level differences between Q and NQ daughters as well as Q and G1-cycling cells.

Two processes: metabolic cycling and slow growth, have been suggested to relate to Q cell differentiation in SP cultures. The process of metabolic cycling is observed in some yeast strains under specific conditions of starvation followed by chemostat growth under low glucose conditions (Tu *et al.*, 2005). These cells show respiration during G1 and fermentation during the rest of the cell cycle. Recently, it was shown that glycogen and trehalose accumulation correlate with transient density increases in the CEN.PK strain used to study cycling during cycling and the post-diauxic phase (Shi *et al.*, 2010). However, the CEN.PK strain background, which is best for demonstrating metabolic cycling, does not maintain a dense cell fraction for much more than 24 hours – as

compared with one month or longer for our prototrophic strains (Allen *et al.*, 2006; Li *et al.*, 2009). We also observed much smaller difference in density in a *glc3* mutant than did Shi *et al* (Allen *et al.*, 2006). In metabolic cycling studies, only 50% of the cells divide and it is not yet known whether the non-dividing fraction of cells are NQ-like or whether cycling cells exhibit differences in Cit1p:GFP expression. Certainly, the oscillation between an oxidative, G1 phase and reductive (fermentative) S-M phases in metabolic cycling is reminiscent of the apparent oxidative capacity of the Q cells and the lack of respiration in the NQ cells, although NQ cells are typically on a path towards apoptosis. If Q/NQ differentiation and metabolic cycling are highly related processes, this will be a wonderful example of why it is important to examine a process from several directions and with a keen eye to the evolutionary and environmental ramifications. However, there are enough differences to suggest that, while related, these two processes lead to very different outcomes.

Other important, recent studies under different growth conditions provide additional and valuable insight into this differentiation process. The first study examined differentiation of yeast cells in synthetic complete medium and, among other important findings, concluded that Q cells were genomically unstable (Madia *et al.*, 2009) (in contrast to our finding in YPD that NQ cells were hypermutable (Aragon, 2008)). Because, we and others (Burtner *et al.*, 2009) have found that cells begin to die within days in SC medium, the instability of Q cells in SC is consistent with our hypothesis that, under stress conditions, Q cells can become NQ cells (Aragon *et al.*, 2006). A second study of yeast grown in high-glucose concentrations (700g/L) showed these cells enter an uncoupling phase allowing fermentation without growth (Benbadis



*et al.*, 2009). Uncoupled cultures develop two cell populations with similarities to Q and NQ cells but, after prolonged uncoupling, only the dense fraction remains. Interestingly, this phenotype is observed in *sch9* mutants in SC, suggesting regulatory pathways involved in this phenotype. The appearance of quiescent-like cells under high glucose conditions suggests that glucose exhaustion alone does not induce this differentiation.

Finally, if glucose exhaustion does not regulate this differentiation, what does? There has long been a hypothesis that quiescent-like cells were present in low abundance in EXP, since a small but thermotolerant population is present in most populations (Elliott and Futcher, 1993). While this idea is appealing, elutriated cells from EXP cultures, which would be likely Q analogs, are not as synchronous during the first cell cycle and certainly not for two cycles (Spellman *et al.*, 1998) as Q cells from SP cultures, suggesting that elutriated cells are not identical to Q cells. An important and answerable question is whether slow growth, quorum sensing, or a combination of these or other signals induces this differentiation. Quorum sensing has been demonstrated in yeast, which produce aromatic alcohols in response to nitrogen starvation that induces pseudohyphal growth (Chen and Fink, 2006; Sprague and Winans, 2006). Finally, because of the clear evolutionary pressures for survival, we should not underestimate the potential complexity of this process, including the presence of other, as yet undiscovered regulators and components that will entertain researchers for years to come.

## **Acknowledgements**

We thank Benjamin Tu and Linda Breeden for helpful discussions. This work was supported by National Science Foundation (NSF) grant MCB-0092364 (to M.W.W.). R.M.J., P.H.T, M.R.W., A.E.D., and E.E.M were supported by a National Institutes of Health (NIH) for Maximizing Student Diversity (IMSD) GM-060201. R.M.J. was also supported by NIH GM-0975149, and E.E.M had further supported under a Louis Stokes Alliance for Minority Participation (LSAMP) Bridge to the Doctorate (BD) fellowship grant through NSF HRD-0832947. Sandia National Laboratories is a multi-program laboratory operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

## References

- Aguilera, A. (1986). Deletion of the phosphoglucose isomerase structural gene makes growth and sporulation glucose dependent in *Saccharomyces cerevisiae*. *Molecular & General Genetics* 204, 310-316.
- Allen, C., Buttner, S., Aragon, A.D., Thomas, J., Meirelles, O., Jaetao, J., Benn, D., Ruby, S., Veenhuis, M., Madeo, F., and Werner-Washburne, M. (2006). Isolation of quiescent and non-quiescent cells from stationary-phase yeast cultures. *Journal of Cell Biology* 174, 89-100.
- Andux, S., and Ellis, R.E. (2008). Apoptosis maintains oocyte quality in aging *Caenorhabditis elegans* females. *Plos Genetics* 4.
- Aragon, A.D., Quiñones, G.A., Thomas, E.V., S.Roy, and Werner-Washburne, M. (2006).

Release of extraction-resistant mRNA in stationary-phase *S. cerevisiae* produces a massive increase in transcript abundance in response to stress. *Genome Biology* 7 R9 doi:10.1186/gb-2006-1187-1182-r1189

Aragon, A.D., Rodriguez, A.L., Meirelles, O., Roy, S., Davidson, G.S., Tapia, P.H., Allen, C., Joe, R., Benn, D., and Werner-Washburne, M. (2008). Characterization of differentiated quiescent and nonquiescent cells in yeast stationary-phase cultures. *Mol Biol of the Cell* 19, 1271-1280.

Benbadis, L., Cot, M., Rigoulet, M., and Francois, J. (2009). Isolation of two cell populations from yeast during high-level alcoholic fermentation that resemble quiescent and nonquiescent cells from the stationary phase on glucose. *Fems Yeast Research* 9, 1172-1186.

Brachmann, C.B., Davies, A., Cost, G.J., Caputo, E., Li, J.C., Hieter, P., and Boeke, J.D. (1998). Designer deletion strains derived from *Saccharomyces cerevisiae* S288C: a useful set of strains and plasmids for PCR-mediated gene disruption and other applications. *Yeast* 14, 115-132.

Burhans, W.C., and Weinberger, M. (2007). DNA replication stress, genome instability and aging. *Nucleic Acids Research* 35, 7545-7556.

Burtner, C.R., Murakami, C.J., Kennedy, B.K., and Kaerberlein, M. (2009). A molecular mechanism of chronological aging in yeast. *Cell Cycle* 8, 1256-1270.

Chen, H., and Fink, G.R. (2006). Feedback control of morphogenesis in fungi by aromatic alcohols. *Genes & Development* 20, 1150-1161.

Coyle, S., and Kroll, E. (2008). Starvation induces genomic rearrangements and starvation-

resilient phenotypes in yeast. *Molecular Biology and Evolution* 25, 310-318.

Dunham, M.J., Badrane, H., Ferea, T., Adams, J., Brown, P.O., Rosenzweig, F., and Botstein, D. (2002). Characteristic genome rearrangements in experimental evolution of *Saccharomyces cerevisiae*. *Proc. Natl. Acad. Sci. USA* 99, 16144-16149.

Edwards, B.S., Oprea, T., Prossnitz, E.R., and Sklar, L.A. (2004). Flow cytometry for high-throughput, high-content screening. *Current Opinion in Chemical Biology* 8, 392-398.

Elliott, B., and Futcher, B. (1993). Stress resistance of yeast cells is largely independent of cell cycle phase. *Yeast* 9, 33-42.

Fuge, E.K., Braun, E.L., and Werner-Washburne, M. (1994). Protein synthesis in long-term stationary-phase cultures of *Saccharomyces cerevisiae*. *J Bact* 176, 5802-5813.

Howson, R., Huh, W.K., Ghaemmaghami, S., Falvo, J.V., Bower, K., Belle, A., Dephoure, N., Wykoff, D.D., Weissman, J.S., and O'Shea, E.K. (2005). Construction, verification and experimental use of two epitope-tagged collections budding yeast strains. *Comparative and Functional Genomics* 6, 2-16.

Huh, W.-K., Falvo, J.V., Gerke, L.C., Carroll, A.S., Howson, R.W., Weissman, J.S., and O' Shea, E.K. (2003). Global analysis of protein localization in budding yeast. *Nature* 425, 686-691.

Imai, J., Maruya, M., Yashiroda, H., Yahara, I., and Tanaka, K. (2003). The molecular chaperone Hsp90 plays a role in the assembly and maintenance of the 26S proteasome. *Embo Journal* 22, 3557-3567.

Le Tallec, B., Barrault, M.B., Courbeyrette, R., Guerois, R., Marsolier-Kergoat, M.C.,

and Peyroche, A. (2007). 20S proteasome assembly is orchestrated by two of chaperones in yeast distinct pairs and in mammals. *Molecular Cell* 27, 660-674.

Lee, J., Spector, D., Godon, C., Labarre, J., and Toledano, M.B. (1999). A new antioxidant with alkyl hydroperoxide defense properties in yeast. *Journal of Biological Chemistry* 274, 4537-4544.

Lewis, K. (2007). Persister cells, dormancy and infectious disease. *Nature Reviews Microbiology* 5, 48-56.

Li, L., Lu, Y., Qin, L.-X., Bar-Joseph, Z., Werner-Washburne, M., and Breeden, L.L. (2009). Budding yeast SSD1-V regulates transcript levels of many longevity genes and extends chronological life span in purified quiescent cells. *Mol Biol of the Cell* 20, 3851-3864.

Li, W., Dang, Y.J., Liu, J.O., and Yu, B.A. (2010). Structural and stereochemical requirements of the spiroketal group of hippuristanol for antiproliferative activity. *Bioorganic & Medicinal Chemistry Letters* 20, 3112-3115.

Lindqvist, L., and Pelletier, J. (2009). Inhibitors of translation initiation as cancer therapeutics. *Future Medicinal Chemistry* 1, 1709-1722.

Madia, F., Wei, M., Yuan, V., Hu, J., Gattazzo, C., Pham, P., Goodman, M.F., and Longo, V.D. (2009). Oncogene homologue Sch9 promotes age-dependent mutations by a superoxide and Rev1/Pol $\zeta$ -dependent mechanism. *Journal of Cell Biology* 186, 509-523.

Martinez, M., Roy, S., Archuleta, A., Wentzell, P., Santa Anna-Arriola, S., Rodriguez, A., Aragon, A., Quiñones, G., Allen, C., and Werner-Washburne, M. (2004). Genomic analysis

of stationary-phase and exit in *Saccharomyces cerevisiae*: Gene expression and identification of novel essential genes. *Molecular Biology of the Cell* 15, 5295-5305.

Newman, J.R.S., Ghaemmaghami, S., Ihmels, J., Breslow, D.K., Noble, M., DeRisi, J.L., and Weissman, J.S. (2006). Single-cell proteomic analysis of *S. cerevisiae* reveals the architecture of biological noise. *Nature* 441, 840-846.

Raser, J.M., and O'Shea, E.K. (2004). Control of stochasticity in eukaryotic gene expression. *Science* 304, 1811-1814.

Raser, J.M., and O'Shea, E.K. (2005). Noise in gene expression: origins, consequences, and control. *Science* 309, 2010-2013.

Rose, M.D., Winstron, F., and Hieter, P. (1990). *Methods in Yeast Genetics a Laboratory Course Manual*. Cold Spring Harbor Laboratory Press: Cold Spring Harbor, NY.

Shi, L., Sutter, B., Xinyue, Y., and Tu, B. (2010). Trehalose is a key determinant of the quiescent metabolic state that fuels cell cycle progression upon return to growth. *Mol Biol of the Cell* *in press*.

Smukalla, S., Caldara, M., Pochet, N., Beauvais, A., Guadagnini, S., Yan, C., Vinces, M.D., Jansen, A., Prevost, M.C., Latge, J.P., Fink, G.R., Foster, K.R., and Verstrepen, K.J. (2008). FLO1 Is a Variable Green Beard Gene that Drives Biofilm-like Cooperation in Budding Yeast. *Cell* 135, 726-737.

Spellman, P.T., Sherlock, G., Zhang, M.Q., Iyer, V.R., Anders, K., Eisen, M.B., Brown, P.O., Botstein, D., and Futcher, B. (1998). Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol Biol of the*

Cell 9, 3273-3297.

Sprague, G.F., and Winans, S.C. (2006). Eukaryotes learn how to count: quorum sensing by yeast. *Genes & Development* 20, 1045-1049.

Tu, B.P., Kudlicki, A., Rowicka, M., and McKnight, S.L. (2005). Logic of the yeast metabolic cycle: temporal compartmentalization of cellular processes. *Science* 310, 1152-1158.

Young, S.M., Bologna, C., Prossnitz, E.R., Oprea, T.I., Sklar, L.A., and Edwards, B.S. (2005). High-throughput screening with HyperCyt (R) flow cytometry to detect small molecule formylpeptide receptor ligands. *Journal of Biomolecular Screening* 10, 374-382.

Tables

<b>Table 4-1. Most abundant proteins in EXP and SP</b>						
	<b>Systematic name</b>	<b>Gene Name</b>	<b>Function<sup>1</sup></b>	<b>Localization</b>	<b>Log<sub>2</sub>EXP</b>	<b>Log<sub>2</sub>SP</b>
<b>Abundant in both EXP and SP</b>	YLR044C	<i>PDC1</i>	Pyruvate decarboxylase	c, n	10.39	8.41
	YHR174W	<i>ENO2</i>	Enolase II, a phosphopyruvate hydratase	c, m, pm	10.12	9.47
	YGR192C	<i>TDH3</i>	Glyceraldehyde-3-phosphate dehydrogenase	c, n, pm	10.01	8.75
	YKL060C	<i>FBA1</i>	Fructose 1,6-bisphosphate aldolase	c, m	9.97	8.42
	YPR080W	<i>TEF1</i>	Translational elongation factor EF-1 alpha	c	9.16	8.34
	YLL024C	<i>SSA2</i>	ATP binding protein	c, n, pm	9.16	8.24
	YBR118W	<i>TEF2</i>	Translational elongation factor EF-1 alpha	c	9.09	7.81
	YJR009C	<i>TDH2</i>	Glyceraldehyde-3-phosphate dehydrogenase	c, m, pm	8.91	7.59
	YAL005C	<i>SSA1</i>	ATPase	c, n	8.71	9.05
	YLR249W	<i>YEF3</i>	Translational elongation factor 3	c	8.54	6.62
	YDL055C	<i>PSA1</i>	GDP-mannose pyrophosphorylase	c	8.5	6.47
	YMR186W	<i>HSC82</i>	Cytoplasmic chaperone of the Hsp90 family	c, m, pm	8.49	6.85
<b>More abundant in EXP</b>	YLR109W	<i>AHP1</i>	Thiol-specific peroxiredoxin	c, pm	8.23	5.78
	YBR196C	<i>PGI1</i>	Glycolytic enzyme phosphoglucose isomerase	c, m, pm	7.7	5.64
	YMR205C	<i>PFK2</i>	Beta subunit of heterooctameric phosphofructokinase	c, m	7.66	6.16
	YDR382W	<i>RPP2B</i>	Ribosomal protein P2 beta	c	7.52	6.1
	YDL130W	<i>RPP1B</i>	Ribosomal protein P1 beta	c	7.5	5.66
	YDL081C	<i>RPP1A</i>	Ribosomal stalk protein P1 alpha	c	7.42	6.08
	YBR189W	<i>RPS9B</i>	Protein component of the small (40S) ribosomal subunit	c	7.32	5.98
	YGL135W	<i>RPL1B</i>	N-terminally acetylated protein component of the large (60S) ribosomal subunit	c	7.33	5.76



<b>More abundant in SP</b>	YGL253W	<i>HXK2</i>	Hexokinase isoenzyme 2	c, m, n	6.97	7.04
	YDR070C	<i>FMP16</i>	Putative protein of unknown function	m	4.43	6.9
	YDL185W	<i>TFP1</i>	The A subunit of the V-ATPase V1 domain	c	6.73	6.74
	YJL138C	<i>TIF2</i>	Translation initiation factor eIF4A	c	6.83	6.61
	YDL184C	<i>RPL41A</i>	Ribosomal protein L47 of the large (60S) ribosomal subunit	c	6.74	6.55
	YFL014W	<i>HSP12</i>	Heat-shock protein that protects membranes from desiccation	c, n, pm	3.89	6.54
	YLR061W	<i>RPL22A</i>	Protein component of the large (60S) ribosomal subunit	c	6.62	6.54
	YHR183W	<i>GND1</i>	6-phosphogluconate dehydrogenase (decarboxylating)	c, m	6.27	6.5

**Table 4-2. GO process of proteins expressed 2-fold or higher in SP (87 proteins) than in EXP (121 proteins)**

<b>Processes for proteins higher in SP</b>	<b>No. in each GO category</b>	<b>P-value</b>
oxidative phosphorylation	11	2.8E-09
generation of precursor metabolites and energy	17	1.2E-06
electron transport chain	8	1.4E-06
respiratory electron transport chain	8	1.4E-06
ATP synthesis coupled electron transport	8	1.4E-06
mitochondrial ATP synthesis coupled electron transport	8	1.4E-06
oxidation reduction	8	1.4E-06
cofactor metabolic process	16	5.2E-06
<b>Processes for proteins higher in EXP</b>	<b>No. in each GO category</b>	<b>P-value</b>
translation	49	5.5E-22
biosynthetic process	79	4.1E-11
cellular protein metabolic process	62	1.7E-10
protein metabolic process	63	3.0E-10
cellular biosynthetic process	74	7.6E-09
primary metabolic process	98	4.2E-07

**Table 4-3. Comparison of Q, cycling G1, and NQ daughters and G1 mother cells**

Q cells	Cycling G1 cells	NQ unbudded daughters	NQ mother G1 cells
100 % form colonies <sup>1</sup>	ND	~50% form colonies	~50% form colonies
No petite colonies produced (genomically stable) <sup>2</sup>	ND	~40% petite colonies (genomically unstable)	~40% petite colonies (genomically unstable)
Q cells give rise to Q daughters in the post-diauxic phase <sup>3</sup>	ND	NQ cells give rise to NQ daughters in the post-diauxic phase	
Produced concurrently with NQ unbudded daughters <sup>1</sup>		Produced concurrently with Q daughters	
Respiration <sup>3</sup>	ND	Little or no respiration	
Low ROS, no apoptosis <sup>1</sup>	ND	50% with high ROS by day 7, 50% apoptotic by day 14	
Typically high density (gm/cm <sup>3</sup> ) <sup>1</sup>	Low density	Low density	Low density
High glycogen, trehalose <sup>1,4</sup>	Low	No glycogen, low trehalose?	
Synchronous for almost 2 cell divisions, lag phase 1.5 hours at 7d <sup>1,5</sup>	Not as synchronous as Q cells, shorter lag phase	In NQ cells, <i>atp17</i> and <i>atp18</i> showed variability in petites and, cells from petite colonies produced both petite and non-petite colonies, suggesting epigenetic regulation of petites in these strains.	
First daughter, no delayed G1 (mother and daughter bud concurrently) <sup>5</sup>	Daughters have delayed G1	ND (not synchronous populations)	
~2000 mRNAs in insoluble protein-mRNA complexes selectively released in response to different stresses <sup>6</sup>	Few insoluble mRNAs	Few insoluble mRNAs	
No observed effect of <i>atp17</i> and <i>atp18</i> mutants <sup>2</sup>	ND	In NQ cells, <i>atp17</i> and <i>atp18</i> showed variability in petites and, cells from petite colonies produced both petite and non-petite colonies, suggesting epigenetic regulation of the petite phenotype in these strains.	

<sup>1</sup> (Allen et al., 2006)

<sup>2</sup> (Aragon et al., 2008)

<sup>3</sup> (This work)

<sup>4</sup> (Shi, et al, 2010)

<sup>5</sup> (Allen, unpublished data)

<sup>6</sup> (Aragon et al., 2006)

## Figure Legends

**Figure 4-1.** EXP and SP distributions of median peak intensities measured by high-throughput flow cytometry for strains from the Yeast GFP-fusion library in EXP and SP. Diagonal, parallel lines identify strains whose difference between SP and EXP is greater than 2-fold. A list of these genes can be found in supplementary data. (◇) 87 GFP-fusion proteins with  $\geq 2$ -fold increases in SP. (X) 121 GFP-fusion proteins with  $\geq 2$ -fold increases in EXP.

**Figure 4-2.** Histogram of fluorescence intensity distributions for Cys3p:GFP and Cit1p:GFP fusion strains from EXP and unfractionated SP cultures. Flow cytometry measurements were collected as described in Materials and Methods.

**Figure 4-3.** Distribution of Cit1p:GFP and DHE (ROS) fluorescence intensity in fractionated Q and NQ fractions. Fluorescence was detected by flow cytometry (A and B) and microscopy (C-F). (A) Cit1p:GFP fluorescence-intensity histogram for the NQ fraction. (B) GFP fluorescence intensity histogram for the Q fraction. (C and E) NQ fraction of Cit1p:GFP stained with DHE (red) indicating reactive oxygen species. (C) Fluorescence of Cit1p:GFP NQ cells stained with DHE overlaid on the DIC image. (E) Cit1p:GFP alone for the same NQ fraction in C. (D and F) Q fraction. (D) Q fraction of Cit1p:GFP cells stained with DHE (red) overlaid on the DIC image. (F) Cit1p:GFP alone for the same Q cells as in D. White scale bars in C-F indicate 5 microns.

**Figure 4-4.** Fluorescence intensities from flow cytometry measurements of fractionated Q and NQ populations from 38 GFP fusion strains grouped by cellular localization (SGD). Results are the average for 3 technical replicates.

**Figure 4-5.** Reproductive capability as measured by colony forming units for biological replicates of wild type (S288c) NQ and Q fractions and Kgd1p:GFP fusion strains sorted into GFP bright (GFP+), DHE bright (ROS+), and GFP and DHE dim (ROS-GFP-). Cells that were both GFP and DHE bright were not observed.

**Figure 4-6.** Oxygen consumption measurements of s288c (prototrophic) cells from unfractionated EXP and SP cultures and fractionated NQ and Q fractions. The actual rate for EXP was 13.7  $\mu\text{mol}/\text{cell}/\text{sec}$ ; SP was 3.7  $\mu\text{mol}/\text{cell}/\text{sec}$ ; NQ was 3.6  $\mu\text{mol}/\text{cell}/\text{sec}$ ; and Q was 21.8  $\mu\text{mol}/\text{cell}/\text{sec}$ . The difference between NQ and Q respiration was significant ( $p \leq 5.5\text{E-}6$ ).

**Figure 4-7.** GFP protein abundance in mother:daughter pairs observed by fluorescence microscopy for two GFP fusion proteins Cit1p:GFP and Acs1p:GFP 3 days post-inoculation (2 days after glucose exhaustion). Insert: Examples of Cit1p:GFP bright ► bright, dim ► dim, bright mother ► dim daughter, and dim mother ► bright daughter. Bright mother ► dim daughter was seen extremely rarely, and dim mother ► bright daughter, not at all. Symmetric and asymmetric abundance refers to whether mothers and daughter exhibit similar levels of GFP-fusion proteins.

**Figure 4-8.** Flow cytometry analysis of Cit1p:GFP fluorescence intensity as a function of time after glucose exhaustion in post-diauxic phase cultures. X-axis is not to scale. Peaks represent number of events at specific fluorescence intensities.

**Figure 4-9.** Our current model for cell differentiation in yeast cultures grown in rich, glucose-based medium (YPD) to SP. In the post-diauxic phase after glucose exhaustion, mother:daughter pairs are symmetric with respect to GFP protein abundance. Dividing cells are typically, but not always, in the less-dense fraction. Dividing cells, both GFP-expressing (Q) and dim (NQ), are predominantly in the less dense, fraction, consistent with the recent finding that density is a function of trehalose concentration (Shi, 2010). Because ~ 90% the cells in the Q fraction are daughters, most of the mother cells originally found in the Q fraction are hypothesized to become NQ cells. We do not yet know if mother cells found in the dense Q fraction stay dense during cell division or are a select group of mother cells that can become dense again after cell division. Our model predicts that NQ cells in SP cultures do not generate Q cells or become quiescent unless they are re-grown and produce Q and NQ progeny.

Figures

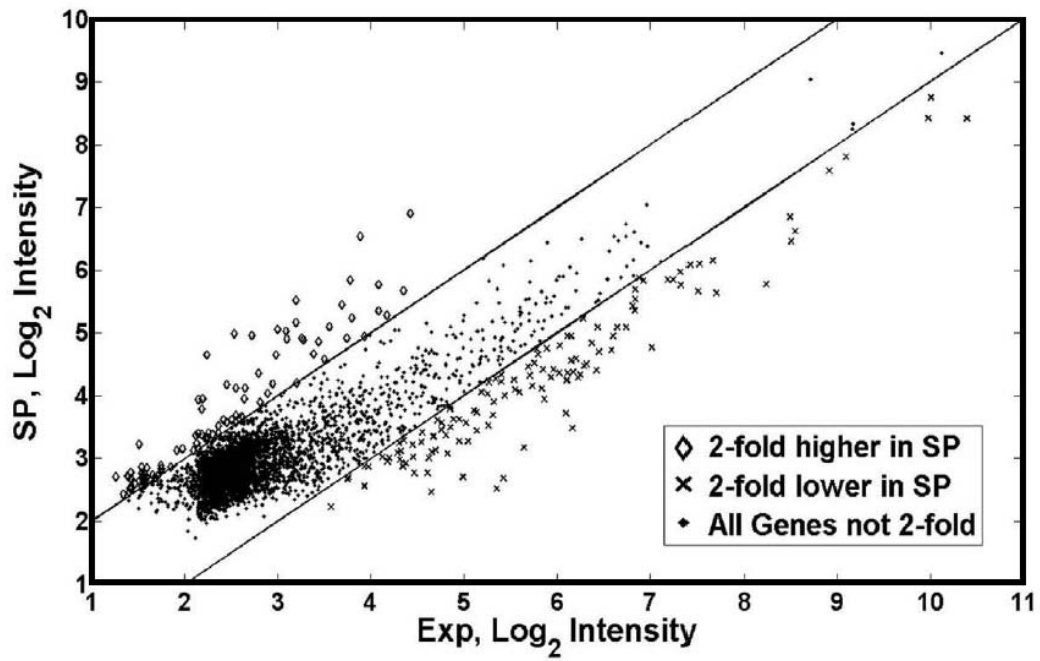


Figure 4-1. EXP and SP distributions of median peak intensities

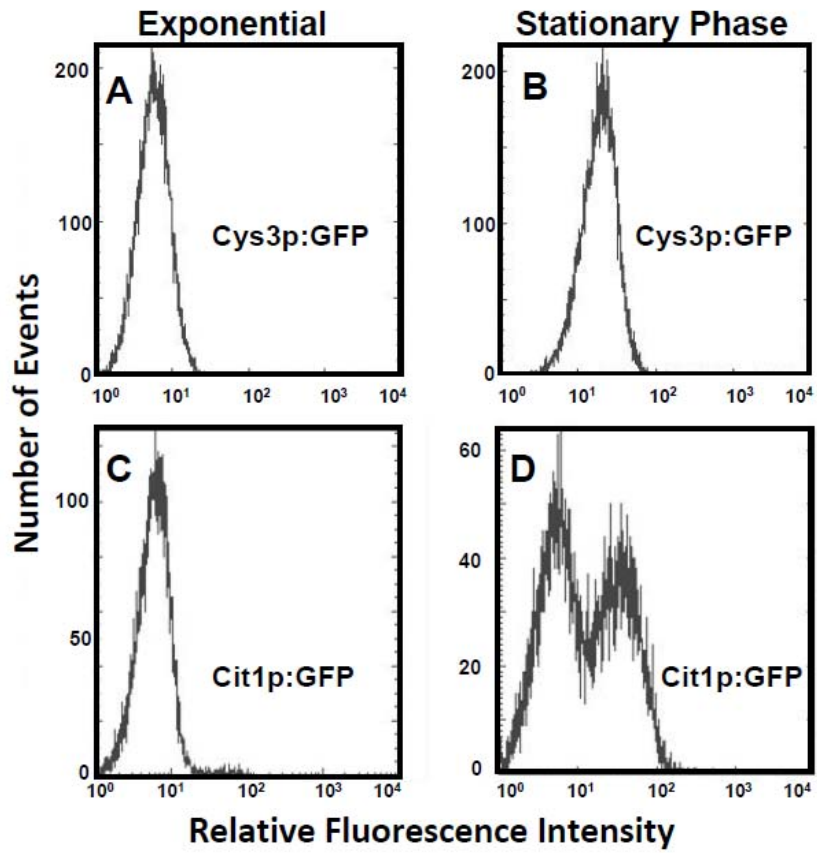


Figure 4-2. Histogram of fluorescence intensity distributions for Cys3p:GFP and Cit1p:GFP fusion strains

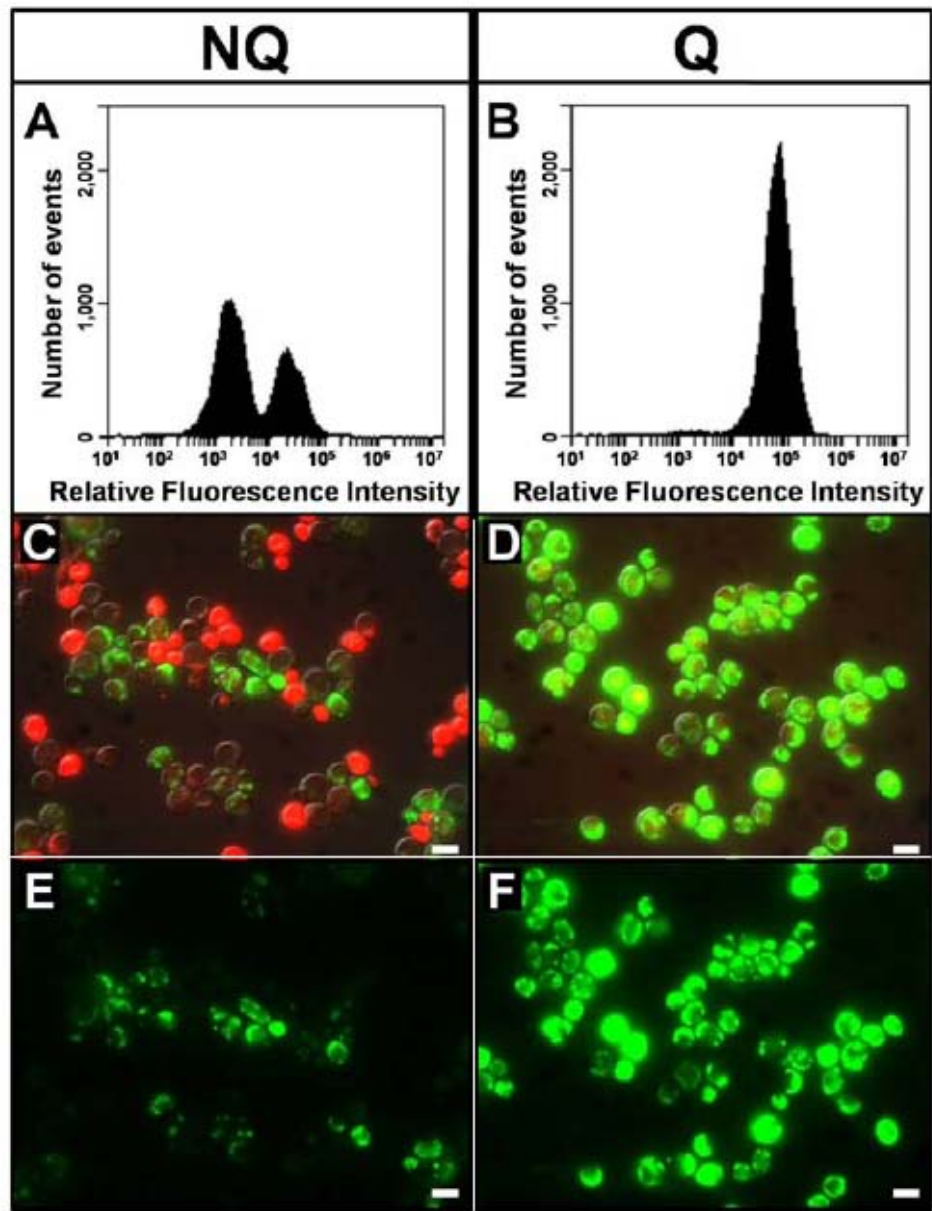


Figure 4-3. Distribution of Cit1p:GFP and DHE (ROS) fluorescence intensity



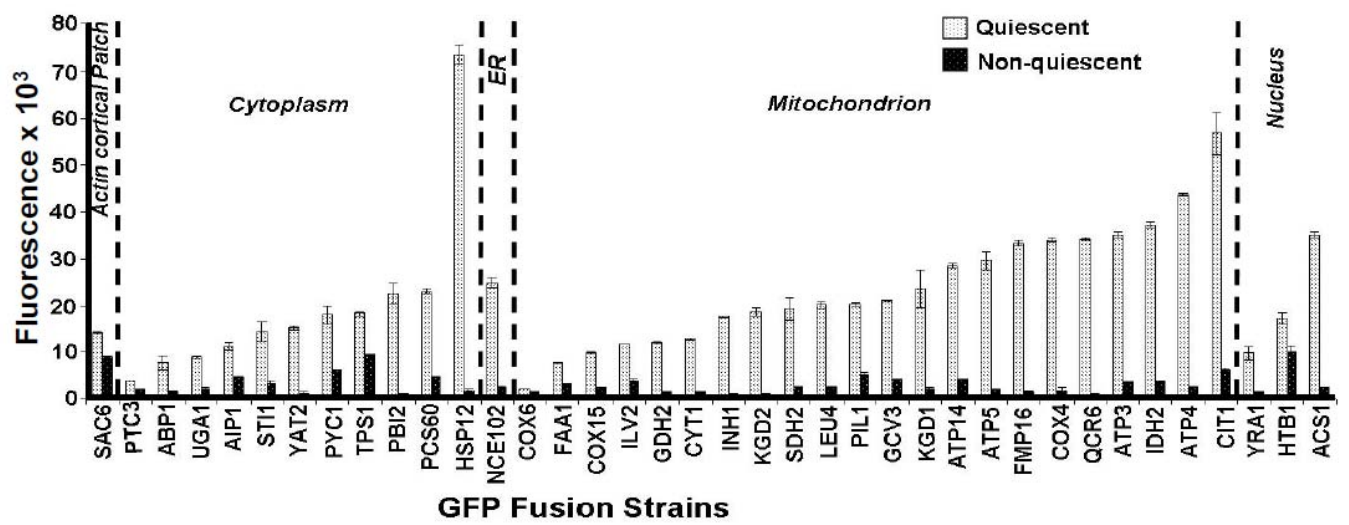


Figure 4-4. Fluorescence intensities

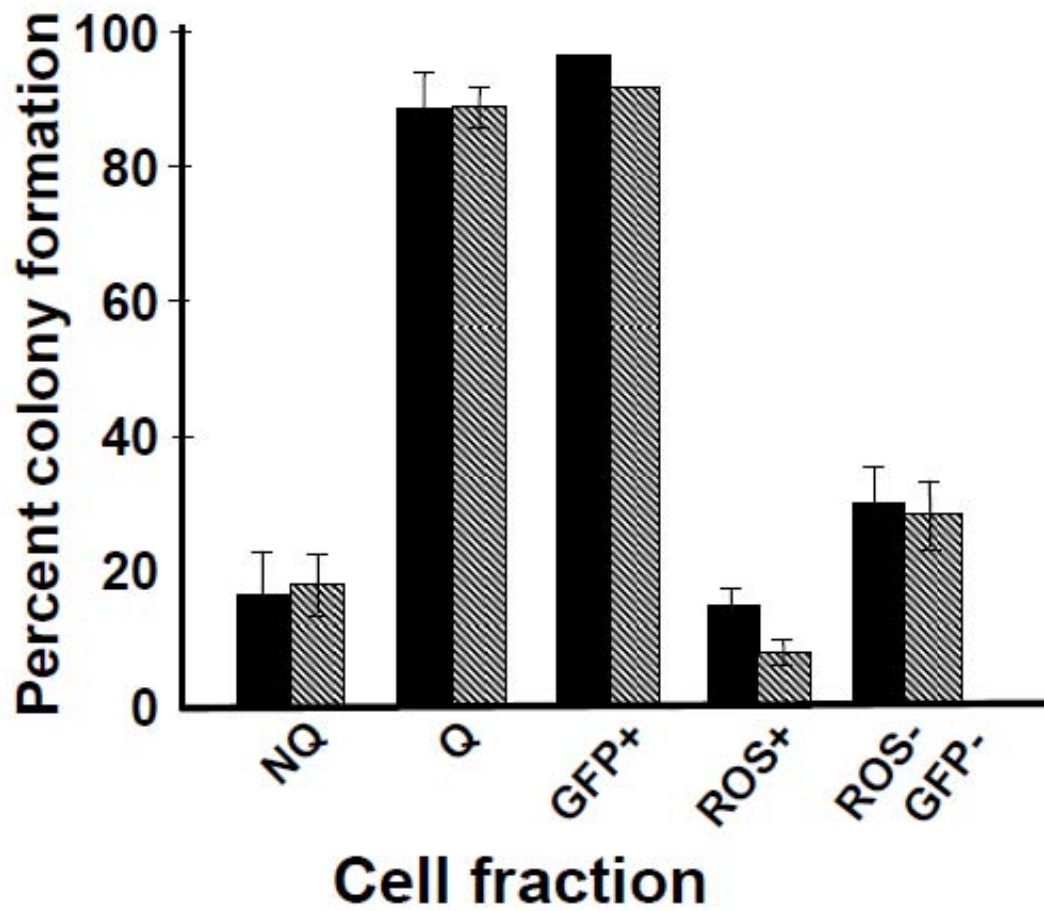


Figure 4-5. Reproductive capability as measured by colony forming units

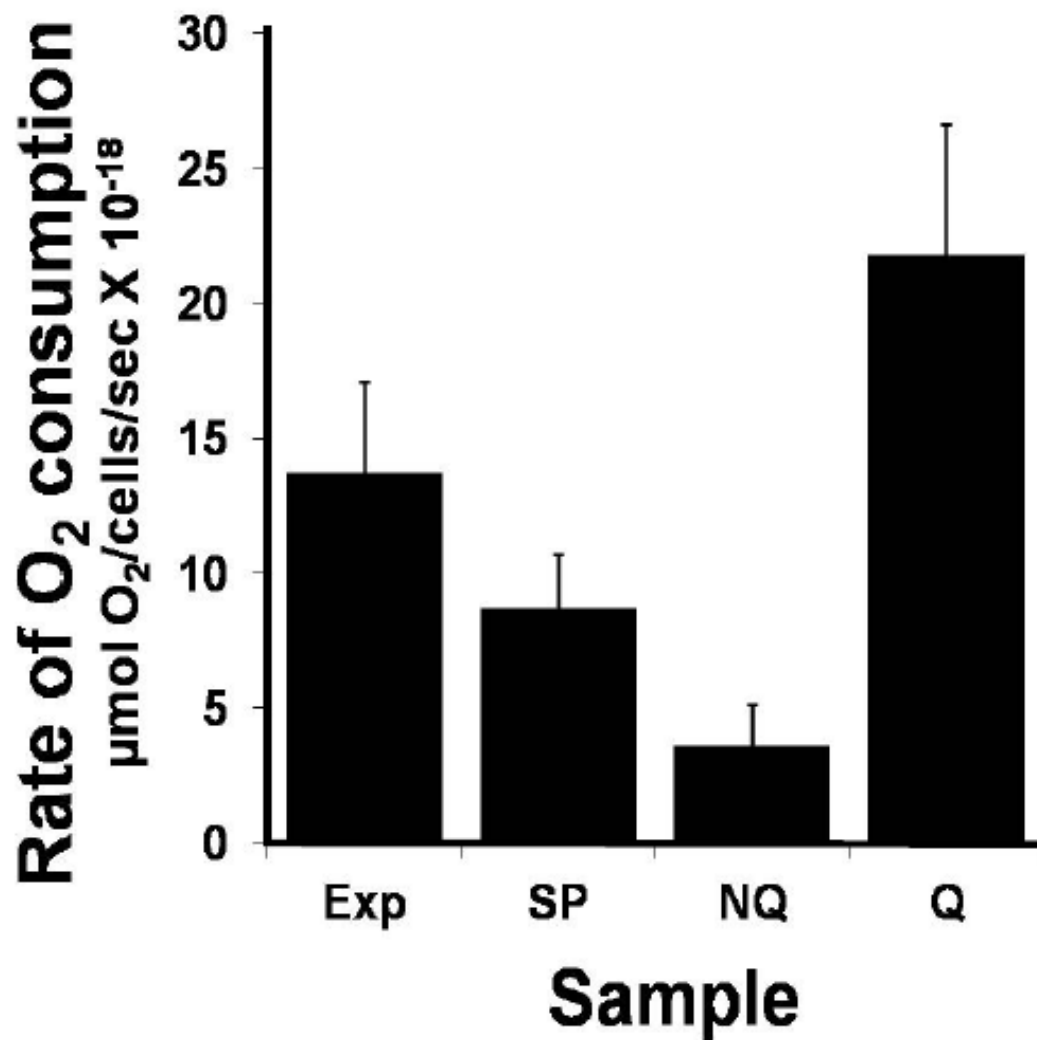


Figure 4-6. Oxygen consumption measurements of s288c (prototrophic) cells

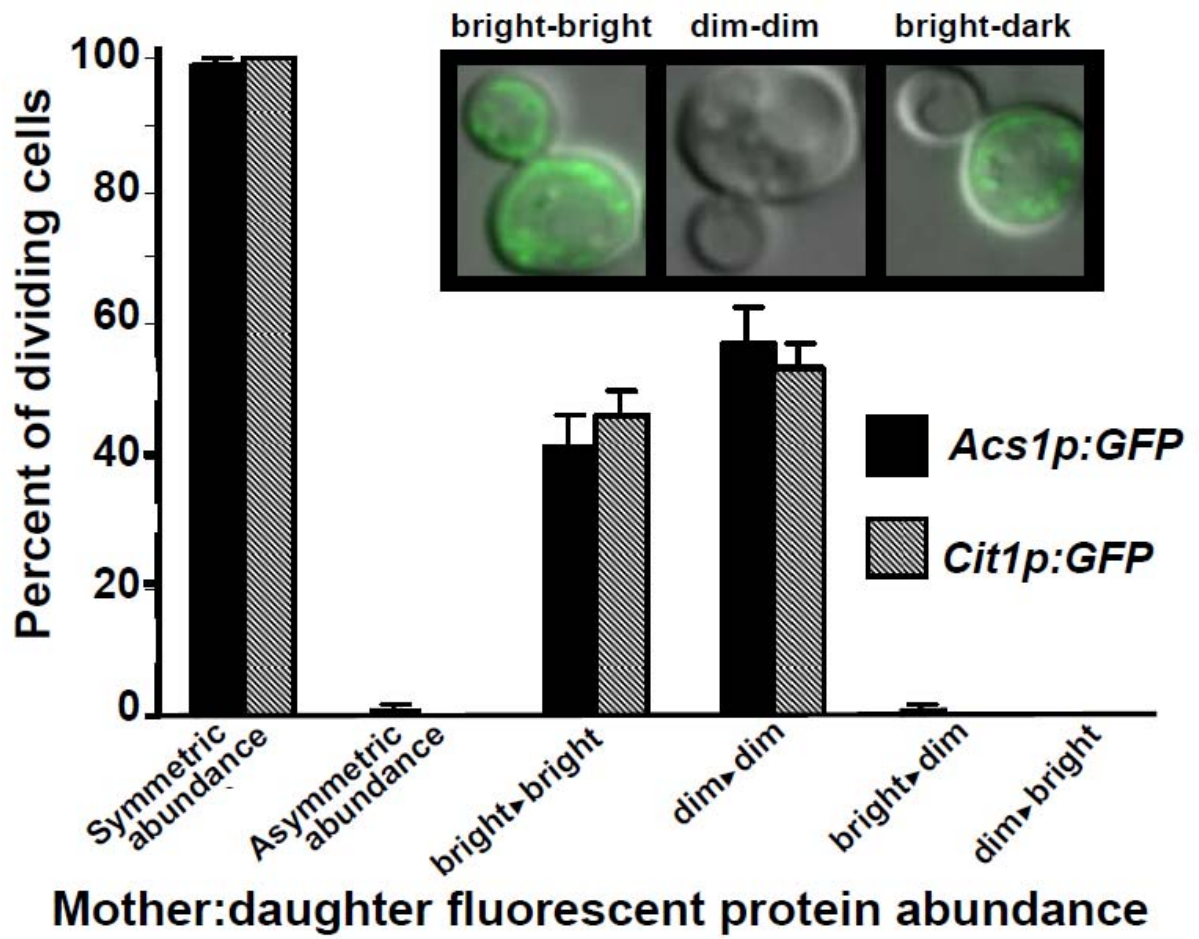


Figure 4-7. GFP protein abundance in mother:daughter pairs

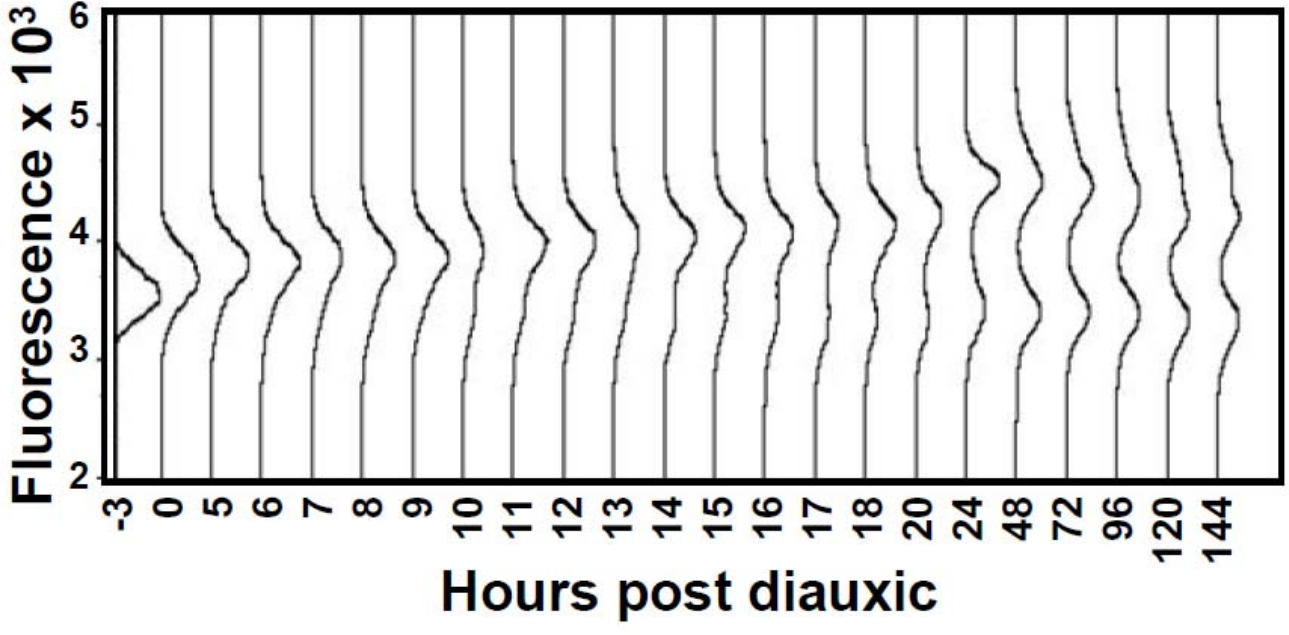


Figure 4-8. Flow cytometry analysis of Cit1p:GFP fluorescence intensity

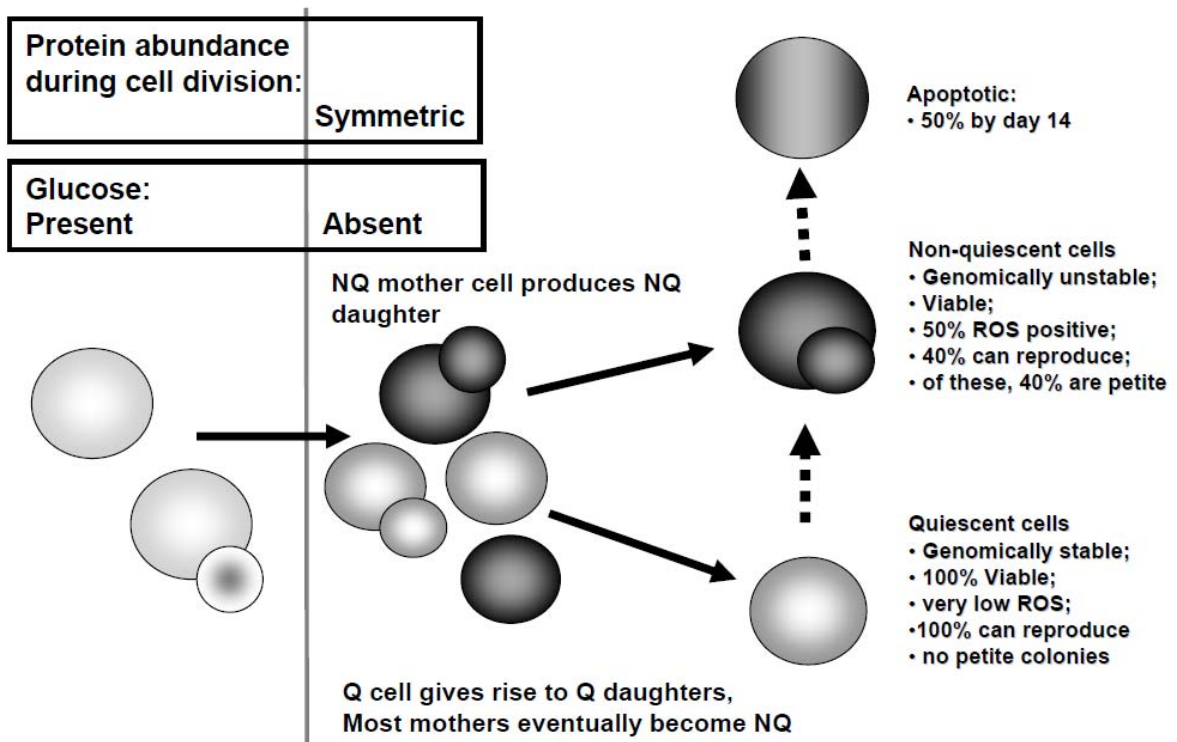


Figure 4-9. Our current model for cell differentiation in yeast cultures

## Supplemental Figure Legends

Figure 4-S1. Correlation plot between our EXP data and that of Newman et al.

Figure 4-S2. Flow cytometry histograms for 38 strains separated into Q and NQ fractions 7 days post-inoculation (SP).

Figure 4-S3. Ratios of median fluorescence measurements for separated Q and NQ fractions from 38 strains

Figure 4-S4. Example of 144 positioned cells sorted by the MoFlo cell sorter. The number of colonies and petite to wild type colonies is typical for Q/NQ separations.

Figure 4-S5. Reproductive capacity (cfu) of NQ fraction of strains sorted by relative GFP and DHE (ROS) fluorescence and plated by the MoFlo cell sorter.

Figure 4-S6. Petite colonies from NQ fraction of strains sorted as above.

Figure 4-S7. Mother:daughter protein abundance for day 3, 5, and 7 post-inoculation in NQ populations of Cit1p:GFP and Acs1p:GFP.

Supplemental Figures

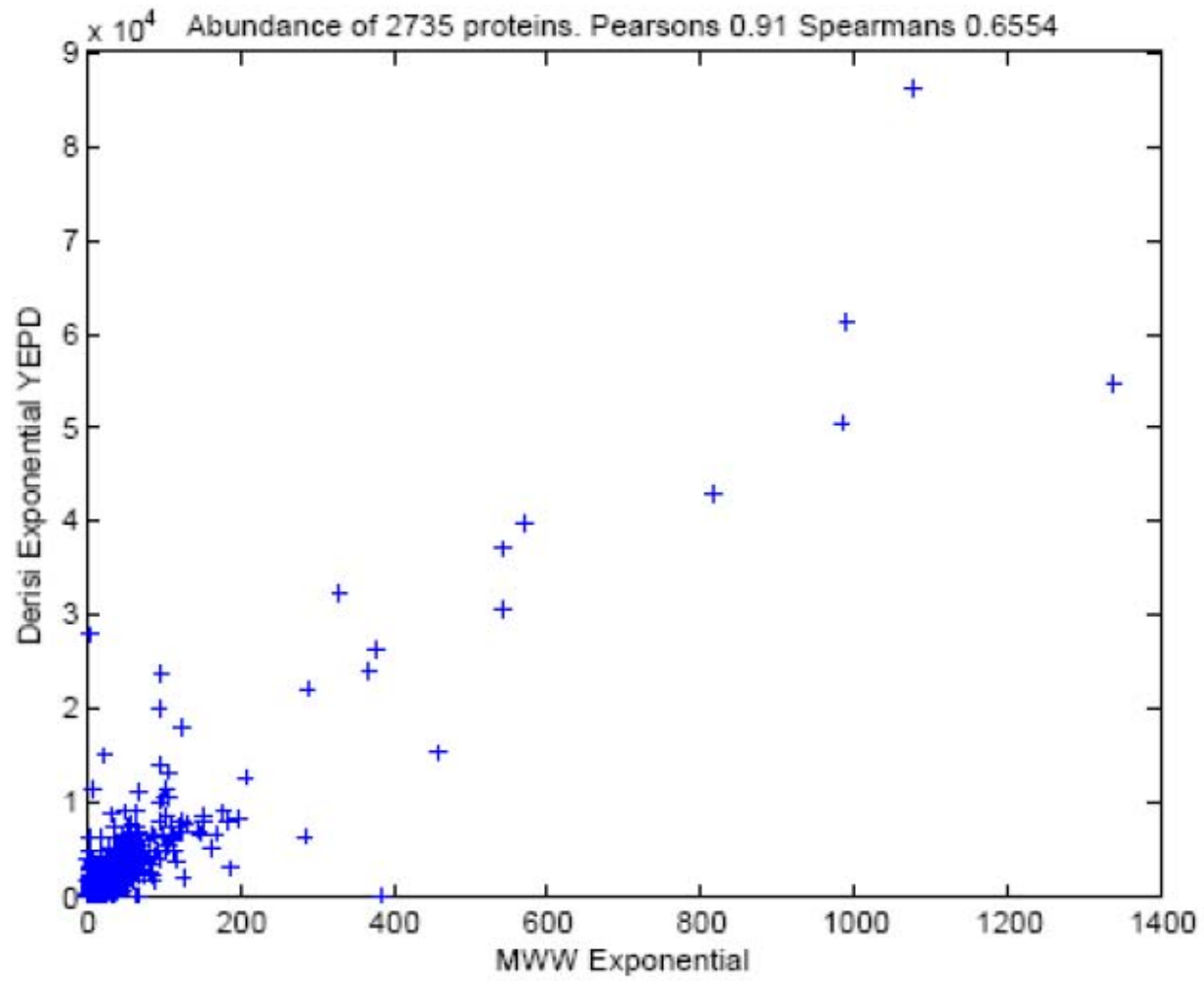
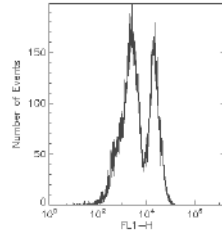


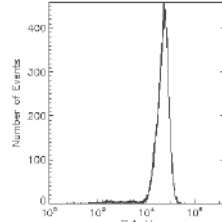
Figure 4S-1. Correlation plot between our EXP data and that of Newman *et al.*



ATP3 NQ

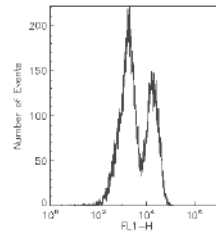


ATP3 Q

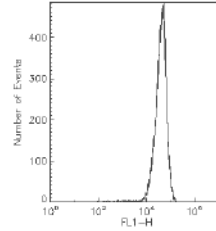


Gamma subunit of the F1 sector of mitochondrial F1F0 ATP synthase, which is a large, evolutionarily conserved enzyme complex required for ATP synthesis

ATP4 NQ

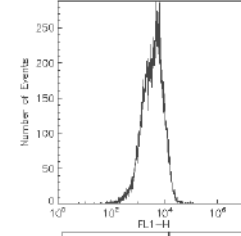


ATP4 Q

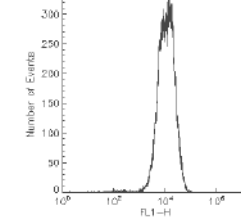


Subunit b of the stator stalk of mitochondrial F1F0 ATP synthase, which is a large, evolutionarily conserved enzyme complex required for ATP synthesis; phosphorylated

ILV2 NQ

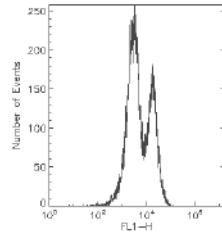


ILV2 Q

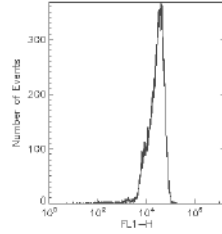


Acetolactate synthase, catalyzes the first common step in isoleucine and valine biosynthesis and is the target of several classes of inhibitors, localizes to the mitochondria, expression of the gene is under general amino acid control

ATP14 NQ

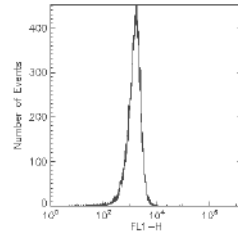


ATP14 Q

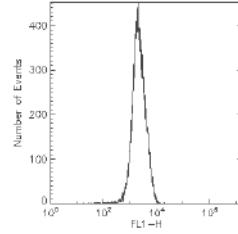


Subunit h of the F0 sector of mitochondrial F1F0 ATP synthase, which is a large, evolutionarily conserved enzyme complex required for ATP synthesis

COX6 NQ

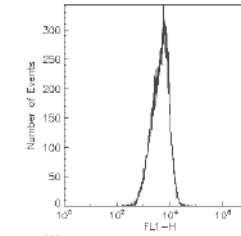


COX6 Q

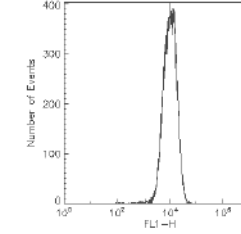


Subunit VI of cytochrome c oxidase, which is the terminal member of the mitochondrial inner membrane electron transport chain; expression is regulated by oxygen levels

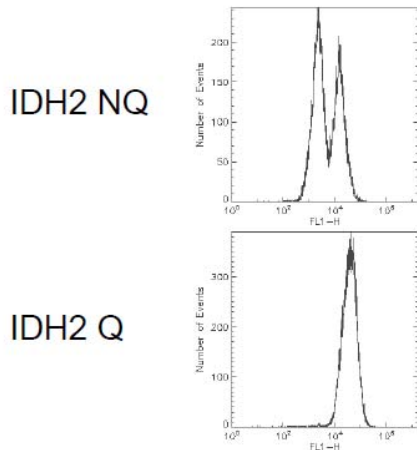
AIP1 NQ



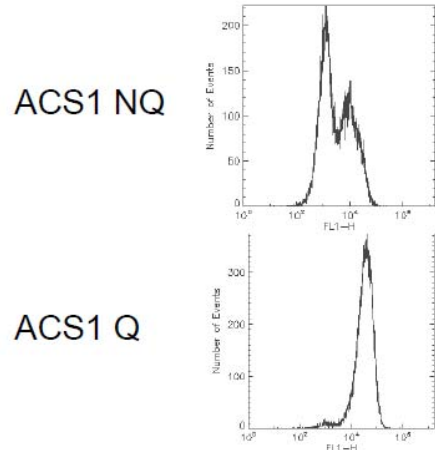
AIP1 Q



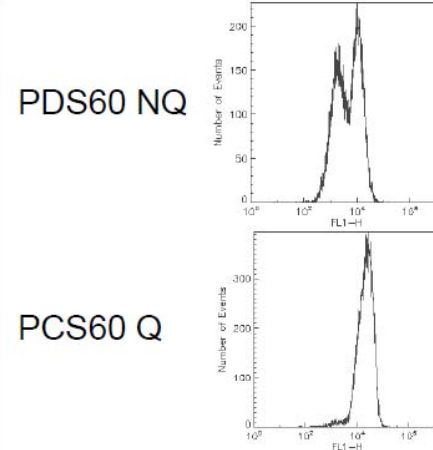
Actin cortical patch component, interacts with the actin depolymerizing factor cofilin; required to restrict cofilin localization to cortical patches; contains WD repeats



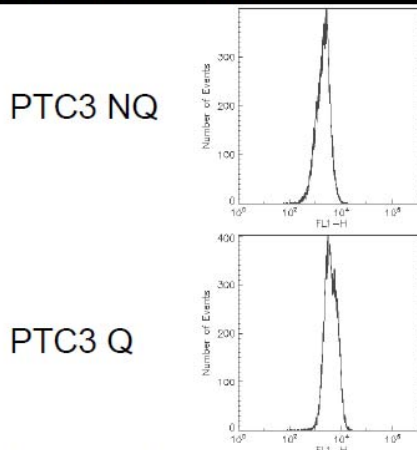
Subunit of mitochondrial NAD(+)-dependent isocitrate dehydrogenase, which catalyzes the oxidation of isocitrate to alpha-ketoglutarate in the TCA cycle; phosphorylated



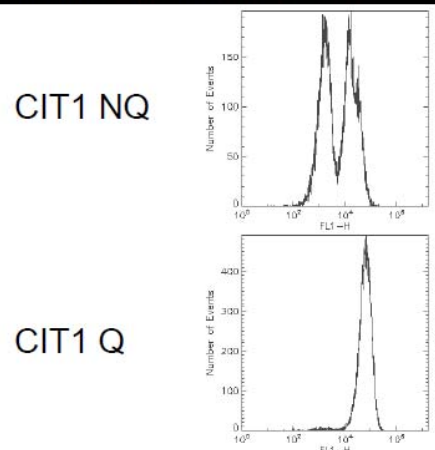
Acetyl-coA synthetase isoform which, along with Acs2p, is the nuclear source of acetyl-coA for histone acetylation; expressed during growth on nonfermentable carbon sources and under aerobic conditions



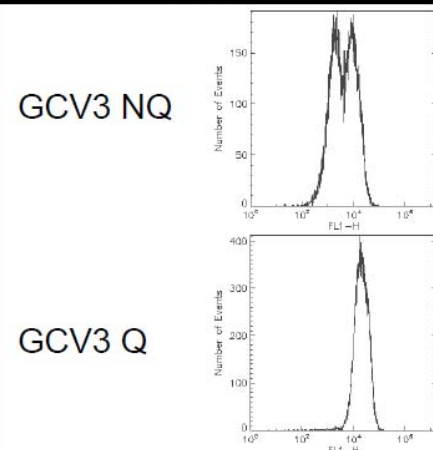
Peroxisomal AMP-binding protein, localizes to both the peroxisomal peripheral membrane and matrix; expression is highly inducible by oleic acid, similar to E. coli long chain acyl-CoA synthetase



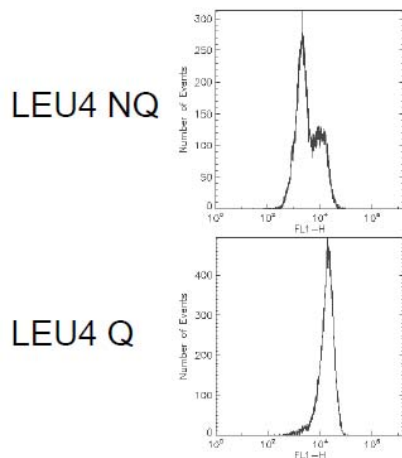
Type 2C protein phosphatase; dephosphorylates Hog1p (see also Ptc2p) to limit maximal kinase activity induced by osmotic stress; dephosphorylates T169 phosphorylated Cdc28p (see also Ptc2p); role in DNA checkpoint inactivation



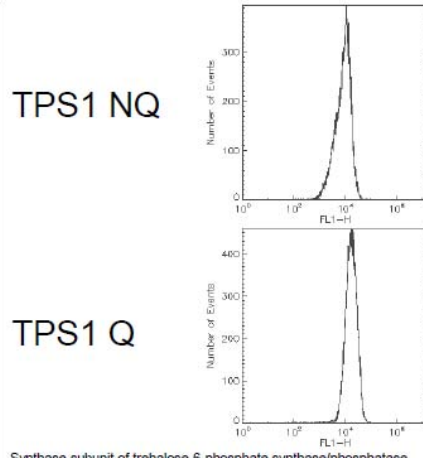
Citrate synthase, catalyzes the condensation of acetyl coenzyme A and oxaloacetate to form citrate, the rate-limiting enzyme of the TCA cycle; nuclear encoded mitochondrial protein



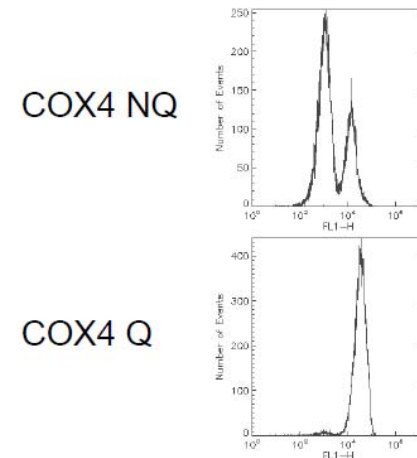
H subunit of the mitochondrial glycine decarboxylase complex, required for the catabolism of glycine to 5,10-methylene-THF; also required for all protein lipoylation; expression is regulated by levels of 5,10-methylene-THF



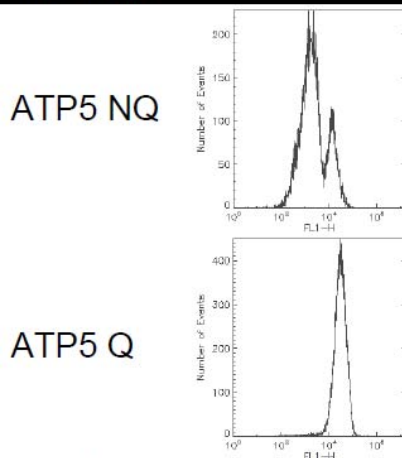
Alpha-isopropylmalate synthase (2-isopropylmalate synthase); the main isozyme responsible for the first step in the leucine biosynthesis pathway



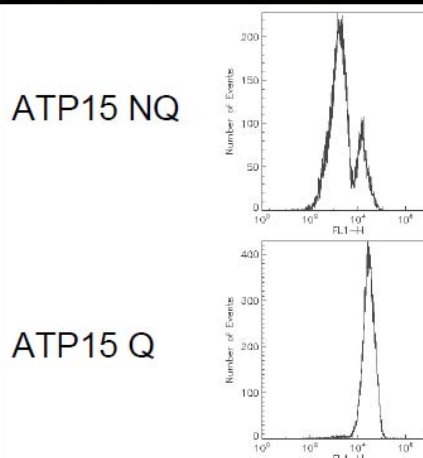
Synthase subunit of trehalose-6-phosphate synthase/phosphatase complex, which synthesizes the storage carbohydrate trehalose; also found in a monomeric form; expression is induced by the stress response and repressed by the Ras-cAMP pathway



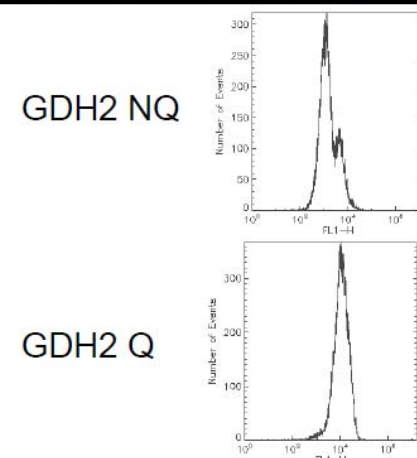
Subunit IV of cytochrome c oxidase, which is the terminal member of the mitochondrial inner membrane electron transport chain; N-terminal 25 residues of precursor are cleaved during mitochondrial import; phosphorylated



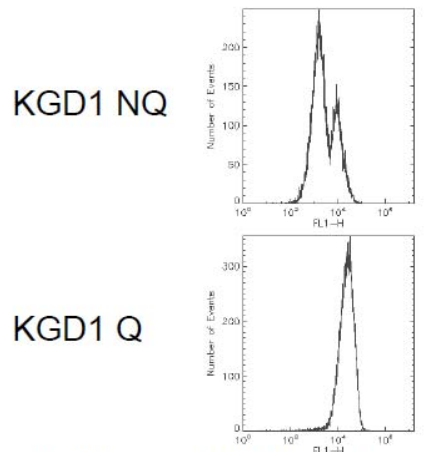
Subunit 5 of the stator stalk of mitochondrial F1F0 ATP synthase, which is an evolutionarily conserved enzyme complex required for ATP synthesis; homologous to bovine subunit OSCP (oligomycin sensitivity-conferring protein); phosphorylated



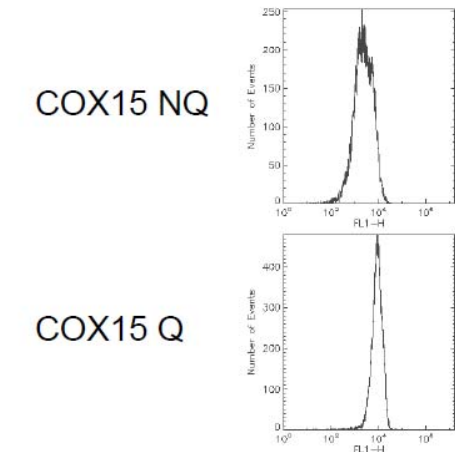
Epsilon subunit of the F1 sector of mitochondrial F1F0 ATP synthase, which is a large, evolutionarily conserved enzyme complex required for ATP synthesis; phosphorylated



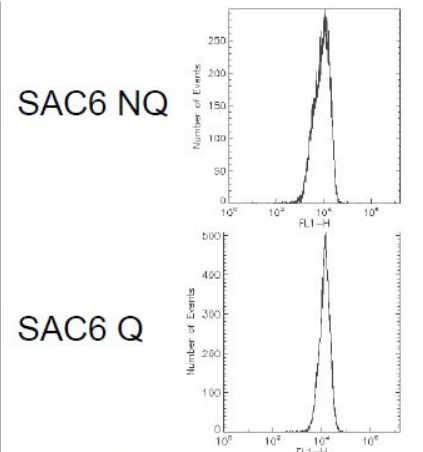
NAD(+)-dependent glutamate dehydrogenase, degrades glutamate to ammonia and alpha-ketoglutarate; expression sensitive to nitrogen catabolite repression and intracellular ammonia levels



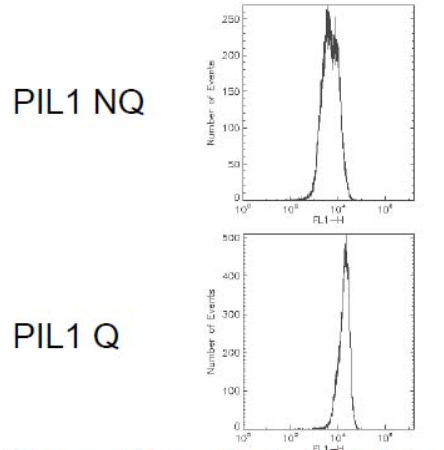
Component of the mitochondrial alpha-ketoglutarate dehydrogenase complex, which catalyzes a key step in the tricarboxylic acid (TCA) cycle, the oxidative decarboxylation of alpha-ketoglutarate to form succinyl-CoA



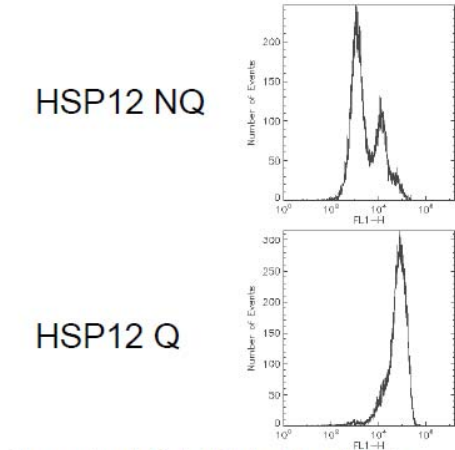
Protein required for the hydroxylation of heme O to form heme A, which is an essential prosthetic group for cytochrome c oxidase



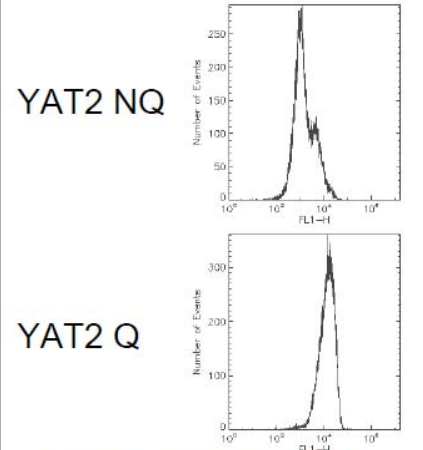
Fimbrin, actin-binding protein; cooperates with Scp1p (calponin/transgelin) in the organization and maintenance of the actin cytoskeleton



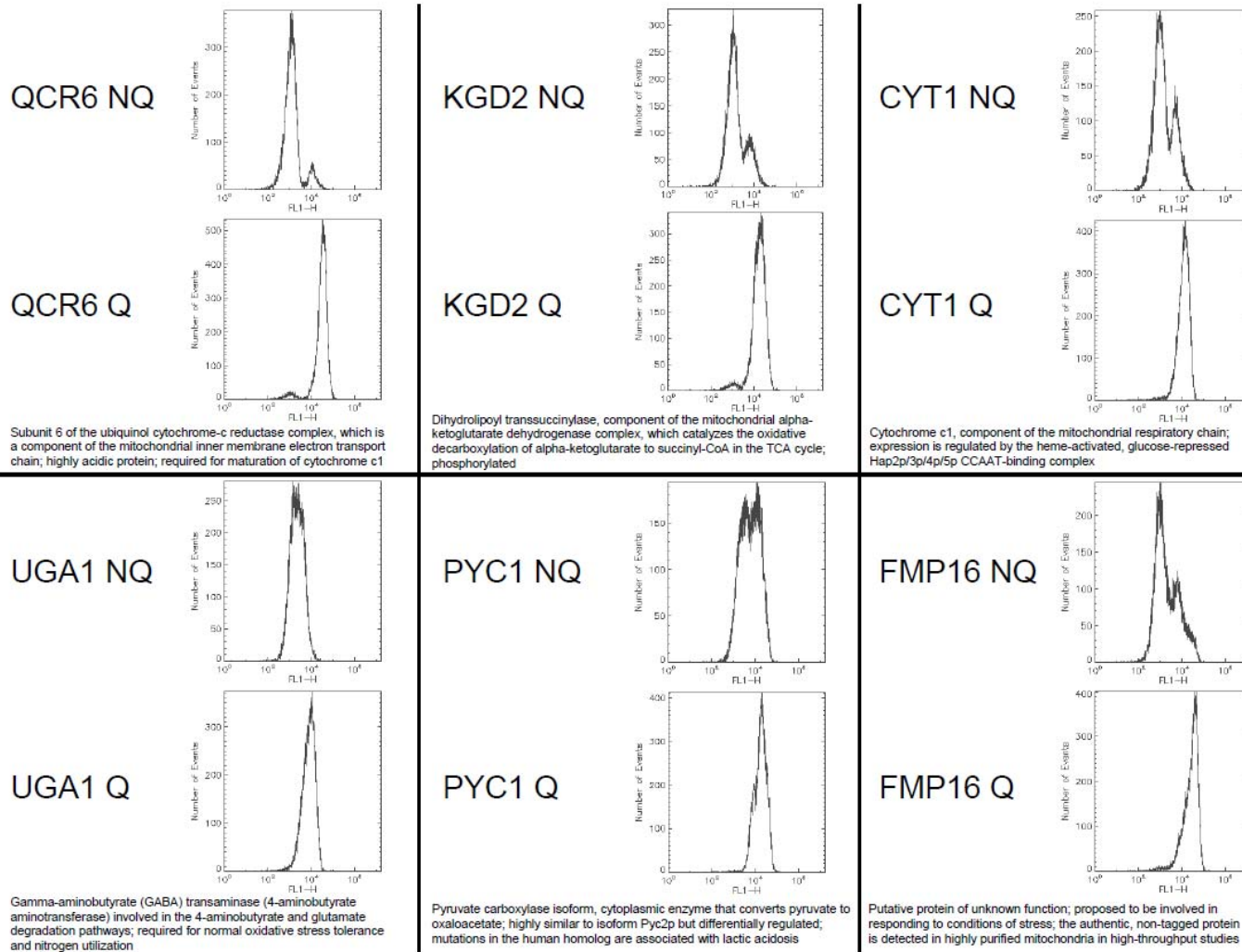
Primary component of eisosomes, which are large immobile cell cortex structures associated with endocytosis; null mutants show activation of Pkc1p/pkc1p stress resistance pathways; detected in phosphorylated state in mitochondria



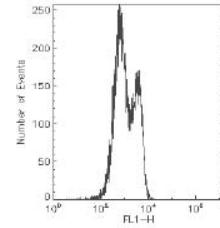
Plasma membrane localized protein that protects membranes from desiccation; induced by heat shock, oxidative stress, osmotic stress, stationary phase entry, glucose depletion, oleate and alcohol; regulated by the HOG and Ras-Pka pathways



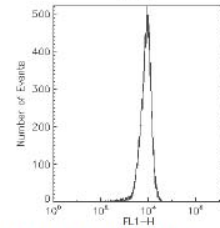
Carnitine acetyltransferase; has similarity to Yat1p, which is a carnitine acetyltransferase associated with the mitochondrial outer membrane



ABP1 NQ

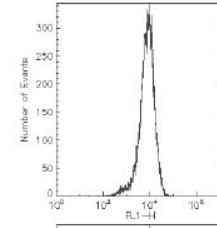


ABP1 Q

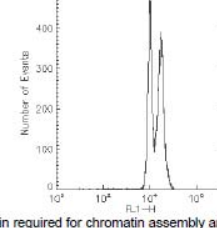


Actin-binding protein of the cortical actin cytoskeleton, important for activation of the Arp2/3 complex that plays a key role actin in cytoskeleton organization

HTB1 NQ

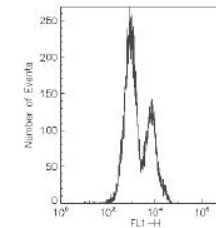


HTB1 Q

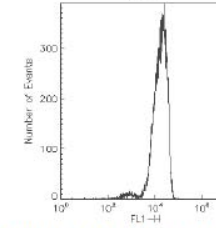


Histone H2B, core histone protein required for chromatin assembly and chromosome function, nearly identical to HTB2; Rad5p-5re1p-Lge1p mediated ubiquitination regulates transcriptional activation, meiotic DSB formation and H3 methylation

INH1 NQ

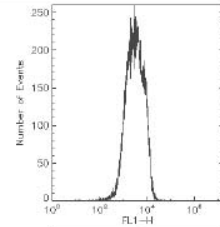


INH1 Q

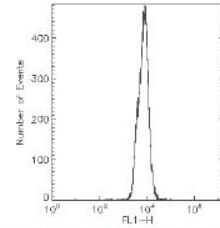


Protein that inhibits ATP hydrolysis by the F1F0-ATP synthase; inhibitory function is enhanced by stabilizing proteins Sst1p and Sst2p; has similarity to Sst1p; has a calmodulin-binding motif and binds calmodulin in vitro

FAA1 NQ

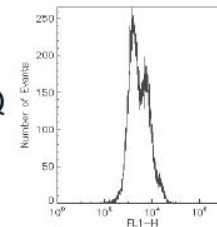


FAA1 Q

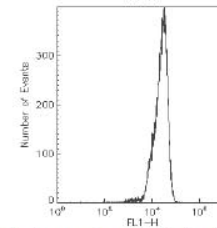


Long chain fatty acyl-CoA synthetase with a preference for C12:0-C16:0 fatty acids; involved in the activation of imported fatty acids; localized to both lipid particles and mitochondrial outer membrane; essential for stationary phase

NCE102 NQ

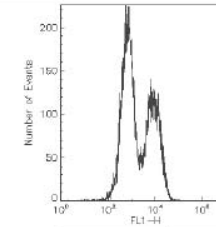


NCE102 Q

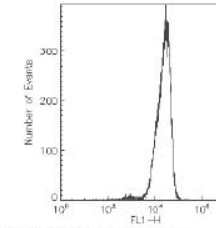


Protein of unknown function; contains transmembrane domains; involved in secretion of proteins that lack classical secretory signal sequences; component of the detergent-insoluble glycolipid-enriched complexes (DIGs)

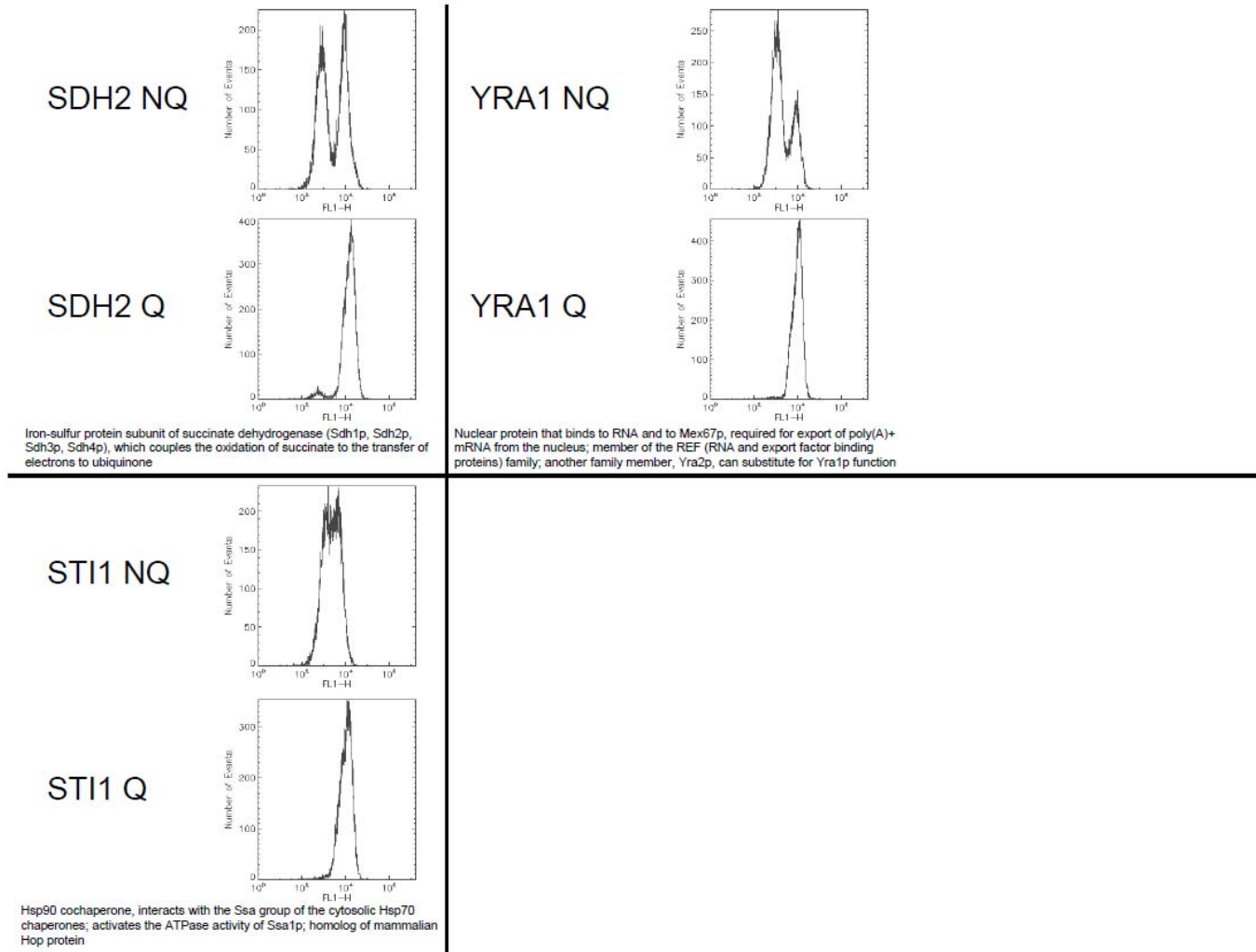
PBI2 NQ



PBI2 Q



Cytosolic inhibitor of vacuolar proteinase B, required for efficient vacuole inheritance; with thioredoxin forms protein complex LMA1, which assists in priming SNARE molecules and promotes vacuole fusion



**Figure 4S-2. Flow cytometry histograms for 38 strains (fluorescence intensity vs. number of events) separated into Q and NQ fractions. These strains all showed 2 fluorescence peaks in unseparated SP cultures.**

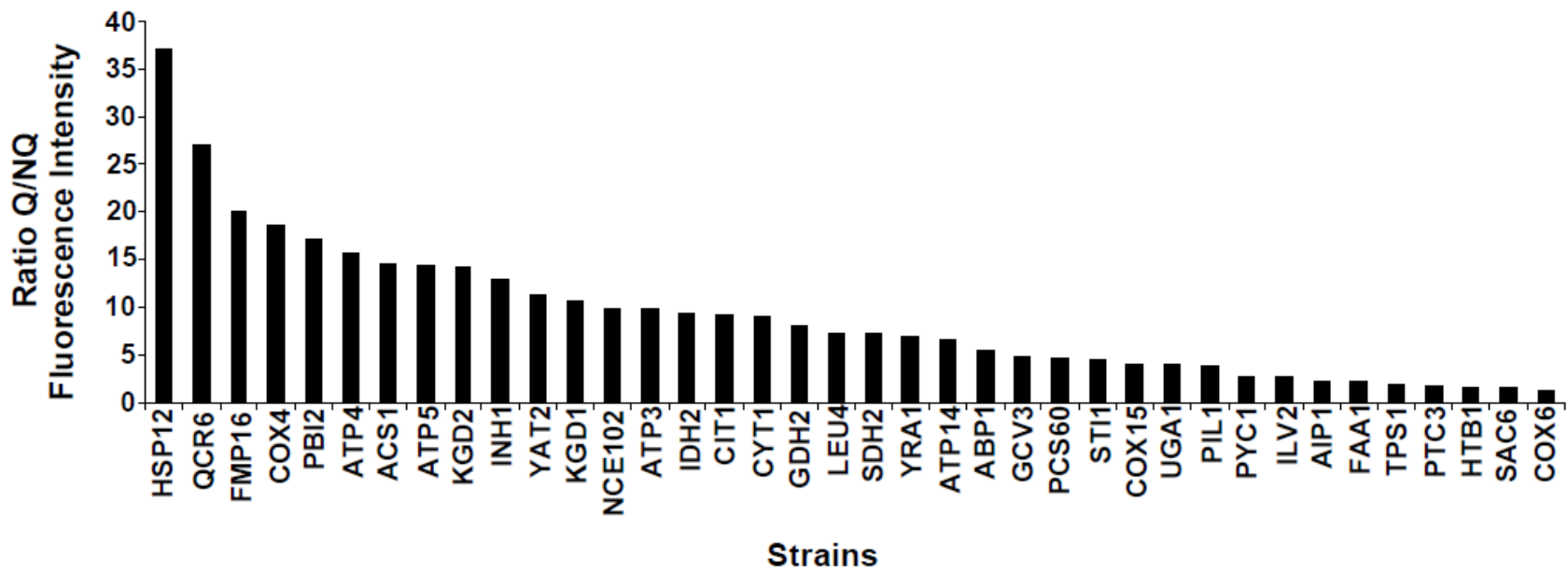
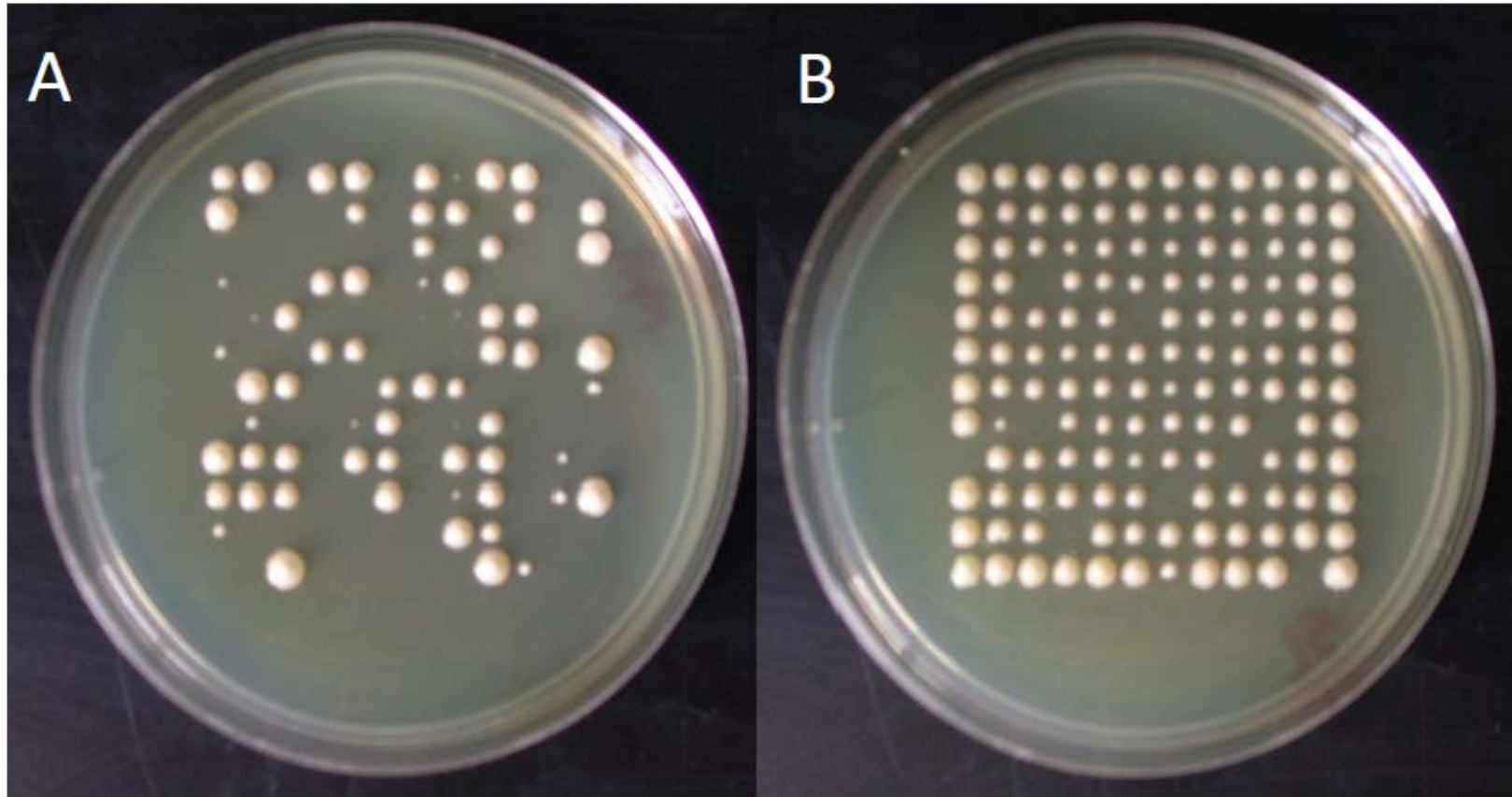


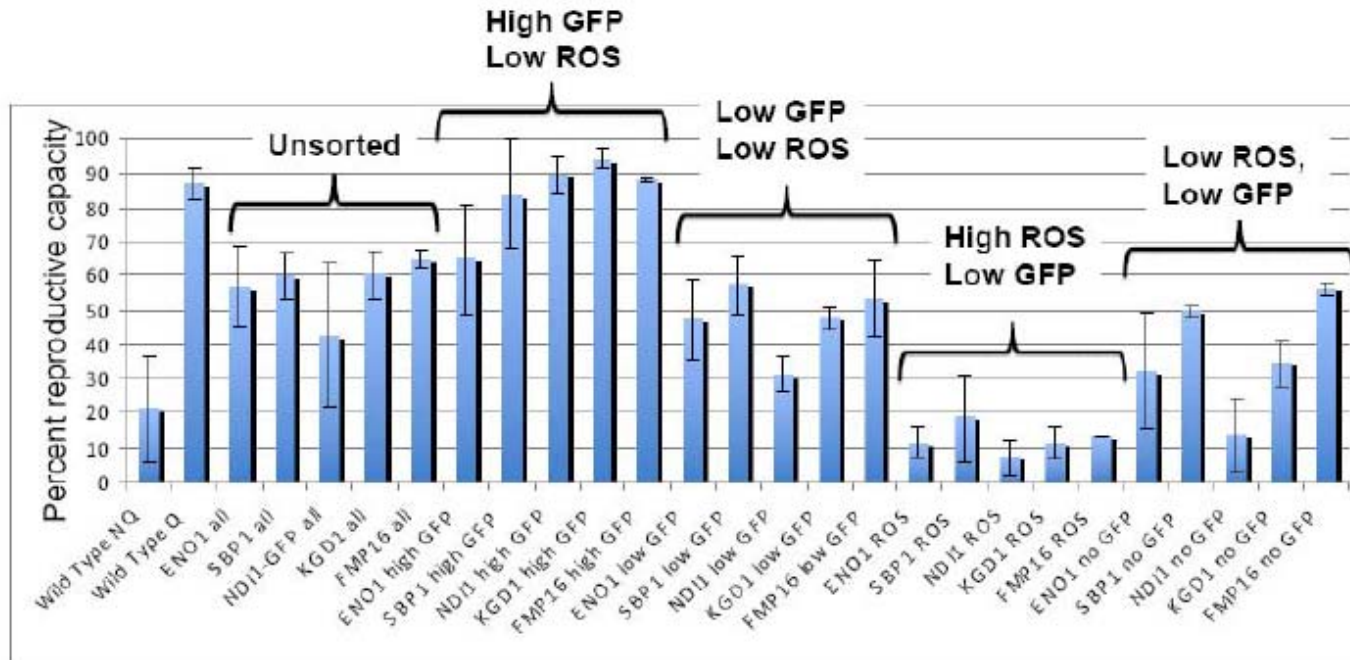
Figure 4S-3. Q/NQ ratios of median fluorescence for 38 strains with 2 fluorescence peaks in SP.





**Figure 4S-4. MoFlo plates: A) upper fraction and B) lower fraction. Similar results were obtained for Q and NQ fractions of S288c prototrophs and Cit1p:GFP strains sorted by fluorescence intensity.**

**High ROS -> low reproductive capacity (cfu) but low ROS/low GFP cells also show loss of reproductive capacity**



Cells were grown to stationary phase (7d), separated using density centrifugation, stained with DHE for ROS, sorted using a MoFlo FACS flow cytometer, and plated on YPD+A plates and grown at 30°C for 2-3 days.

Wild type (S288c) nq/q are positive controls for reproductive fitness. Unsorted fractions encompass entire Q/NQ population. High GFP samples are a subpopulation expressing high GFP. Low GFP samples are a subpopulation expressing a low amount of GFP. High ROS samples are a subpopulation (from the low GFP subpopulation) that exhibit a high amount of oxidative stress. The low ROS/GFP samples are a subpopulation that contained no GFP or ROS detectable to the FACS.

**Figure 4S-5. Colony formation for NQ fractions separated by GFP and ROS. Each sample (144 cells) was plate on 3 plates.**

## For most strains, petite formation is a function of GFP and independent of ROS

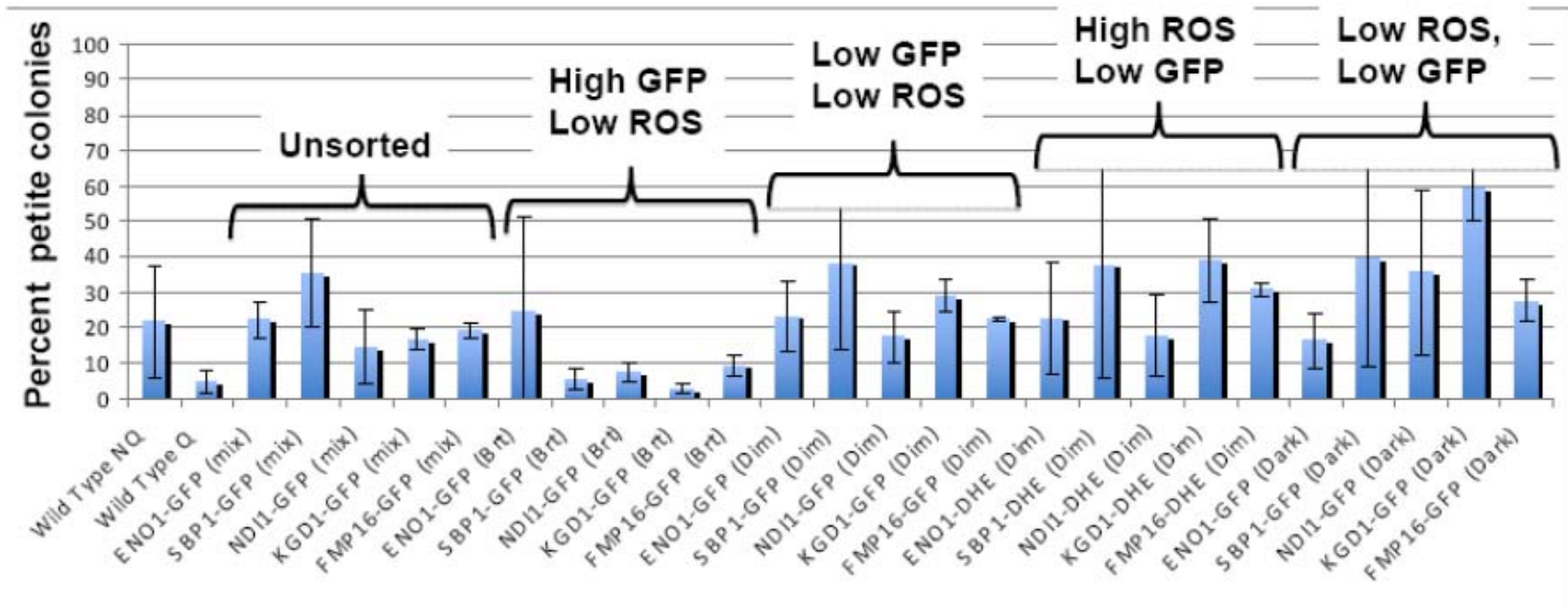


Figure 4S-6. Analysis of petite colony formation of NQ fractions separated by GFP and ROS.

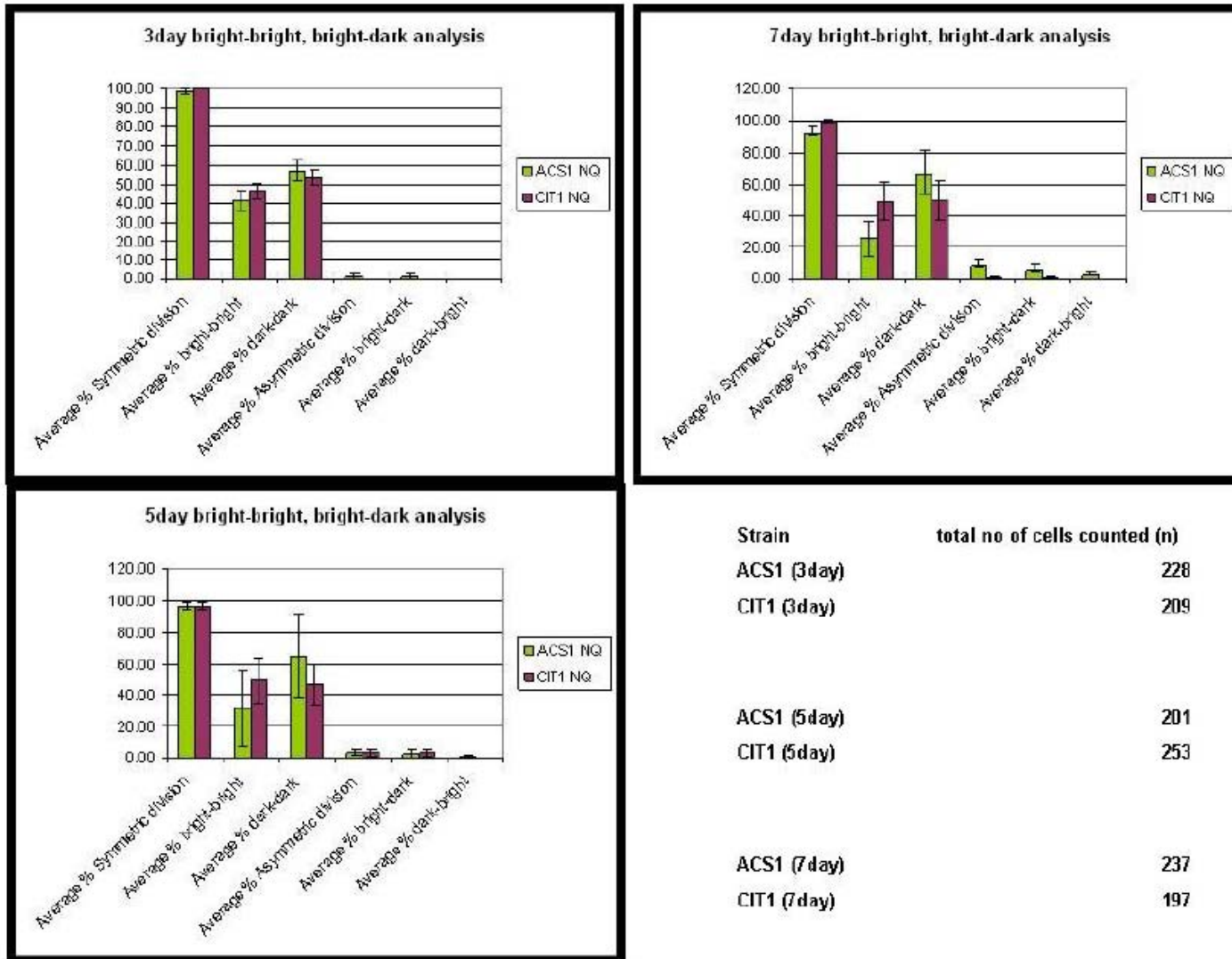


Figure 4S-7. Mother:daughter analysis: symmetric vs. asymmetric GFP protein abundance in the post-diauxic phase.

## Chapter 5: Discussion and Conclusion

Genome and transcriptome analysis led to a more complete description of cellular processes, for example, the gene expression levels throughout the yeast cell-cycle (Spellman, et al., 1998) and the tissue-specific gene expression patterns identified in *C. elegans* (Kim, et al., 2001). Still, the picture is incomplete without knowledge of protein/gene and protein/protein interactions (Costanzo, et al., 2010) and of protein concentrations and localizations (Schubert, et al., 2006), which are not directly revealed by gene expression measurements (Li, et al., 2004; Rual, et al., 2005; Yu, et al., 2008). The combination of genome and transcriptome analyses do explain some levels of cellular function, but they expose other complexities, which can only be answered with new experiments and direct measurement of *in vivo* protein concentrations and localizations. In our experiments, the combination of genomics and new high-throughput proteomic methods were key elements in achieving a more complete understanding of these complexities.

### **The Challenge: to Develop Analysis for High-Throughput Methods**

As explained in Chapter 2, the post-genomic challenge was the development of methods to exploit the new technologies. New experimental methods were developed that differed from earlier methods by their scope and scale (thousands of genes studied simultaneously). These methods required new analysis tools and relied on computers in an unprecedented way to make the results accessible to researchers. Progress in understanding how to cluster genes based on similar gene expression combined with deeper knowledge gained by previous and ongoing analysis of specific genes allowed us

to leverage this knowledge to understand the functions of the many other genes with similar expression patterns (a process I've called the third way, or the middle outward approach in contrast to top down descriptive biology and bottom up biochemical studies of cellular processes).

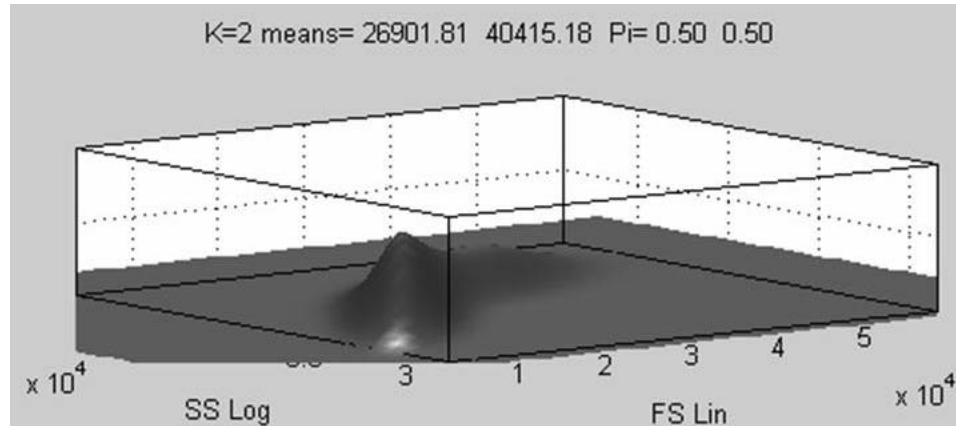
The original challenge to develop methods and analysis tools for microarrays is largely met, and the basic approach in Chapter 2 has been extended far beyond my initial vision. Those analysis methods identified groups of functionally related genes by clustering them together based on similar gene expression profiles across a broad compendium of experimental results. Stuart et al. (2003), recognized that the same approach could be used across not just a compendium of single species results, but across collections of microarray experiments from many species. In their work, VxInsight clusters of homologous genes clearly reveal evolved units of functionality that have been preserved across species.

Srinivasan et al. (2005) went even further with the same concept. Directly using the sequenced genomes of over two hundred microbial species, they computed a similarity between genes based on the number of times pairs of homologous genes appear together in the species. VxInsight clusters reveal genes that have moved together or been lost together at speciation points through evolutionary time. Consequently, genes that have remained together through descent are likely to be involved in the same functional network, a hypothesis that they were able to verify. Most interestingly, they were able to use this method to predict phenotypes. In short, while developed to meet the analysis challenge presented in Chapter 2, my tools have been used in genomics more broadly

than anticipated; however, the original challenge remains only partially met with respect to high-throughput proteomic data sets such as those in Chapter 4.

### **Toward a Thorough Proteomic Analysis of GFP-Fusion Strain Flow Data**

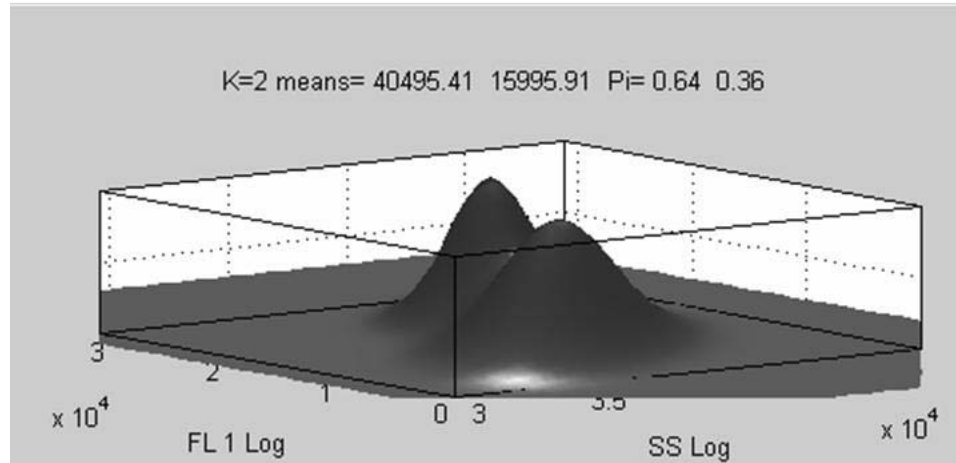
The hypotheses in Chapter 4 came from questions driven by prior microarray analyses, but the results we found rely on a very diverse set of experimental approaches (cell viability studies, characterization of differential morphologies by optical and density gradients, oxygen consumption measurements, and direct measurements of protein concentrations by flow cytometry). Certainly, there is a need for more thorough analyses of the multi-dimensional flow cytometry data. For example, these data are known to contain much more information encoded in the distribution of cells across at least five independent measurement dimensions (forward scattering, side scattering, and at least three fluorescent channels), see Figure 5-1. The statistical methods in Chapter 4 can be developed much further, but new visualization and computer analysis tools will be required to understand the full range of information in these massive data sets (Fruhirth-Schnatter & Pyne, 2010; Pyne, et al., 2009). The multi-dimensional analysis, sketched below, of the probability density functions describing our GFP flow data, is an example of such a combination of mathematics, computing, and visualization.



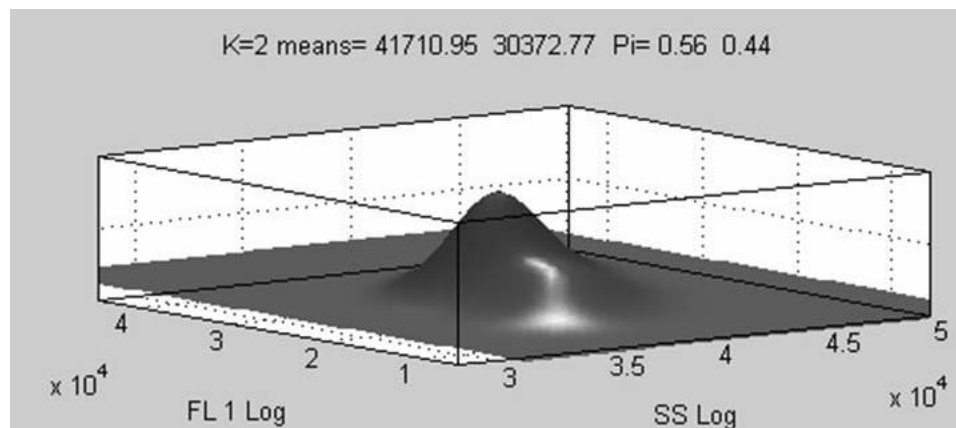
**Figure 5- 1. Forward scatter (FS) and log side scatter (SS) for stationary phase Cit1p strain.** This representation uses a mixture of two Gaussian distributions. Note that most of the cells are relatively small (low FS) and relatively uniform (low SS); however some of the larger cells (large FS) are much more granular (higher SS). This information can be gleaned without considering the GFP channel.

The next steps toward more complete analysis of our flow cytometry data are becoming clearer. For example, it is possible to use dissimilarities between multidimensional results like those shown in Figure 5-2 and 5-3 to identify information that was not originally obvious in the work discussed in Chapter 4. For instance, the multidimensional distributions of the 38 genes identified in Chapter 4 were compared using the dissimilarity metric known as the Earth Mover's Distance (EMD) (Rubner, Tomasi, & Guibas, 2000), which can be thought of as the work involved to move the probability mass as found in one distribution until it exactly matches the second distribution. EMD can be seen as a classical transportation problem, consequently the resulting dissimilarity measure can be found using efficient algorithms (Ling & Okada, 2007; Pele & Werman, 2009).





**Figure 5-2. Stationary phase GDPHp strain GFP (here, FL 1 Log) and log side scatter (here, SS Log) distributions for sampled cells.**

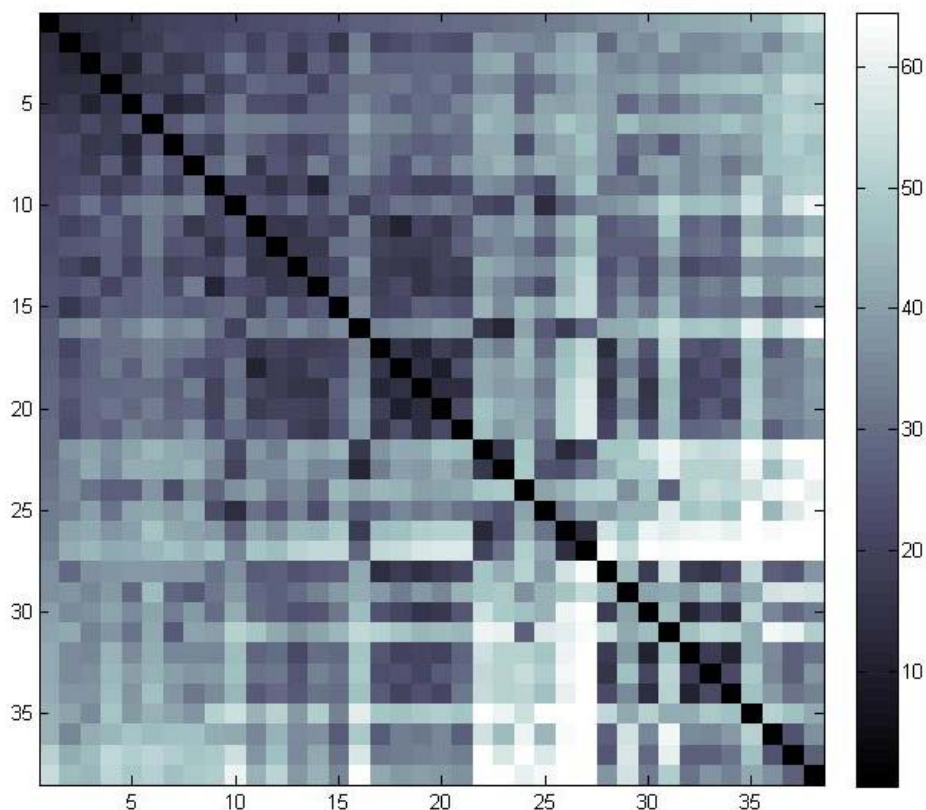


**Figure 5-3. Stationary phase HTB1p strain GFP (here FL 1 Log) and log side scatter (here SS Log) distributions for sampled cells.** HTB1p, a core histone protein regulating transcriptional activity. Note that the brighter GFP cells are also more granular for this histone protein, presumably the brightness reflects greater concentrations of the histone protein, perhaps reflecting less tightly organized chromatin in the dimmer, less granular cells, which may be indicative of an unsuccessful transition from glucose metabolism to oxidative phosphorylation.

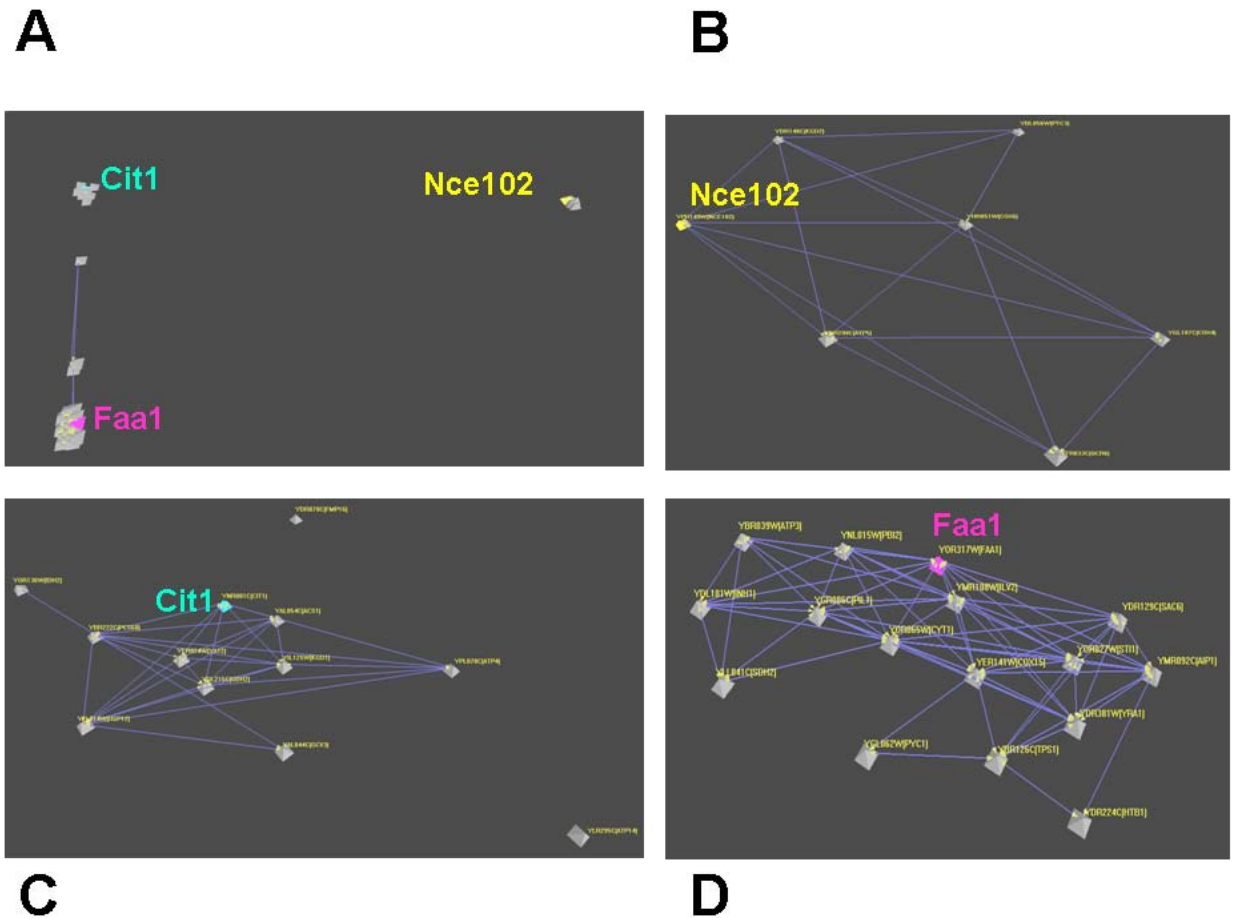
Figure 5-4 shows a gray-scale rendering of the EMD between each pair of the 38 strains singled out in Chapter 4 as being potential flags for quiescent cells (see Appendix I for a list of these distances). These EMD values were computed using the three dimensional densities: GFP x Side Scatter x Forward Scatter, each having 10 bins for a

total of 10 x 10 x 10 histogram bins. The rows and columns have been sorted by increasing distance from Cit1p.

To look for more structure, the table of pairwise EMD values can be compared by computing row similarities using Pearson correlation. Each EMD value has already, jointly, taken into account all of the histogram cells. Applying Pearson row by row (that is, gene by gene) computes the overall similarity of the Earth Mover Distances between each of the compared genes to all of the others. Computing all such comparisons allows one to display the similarities using VxInsight, as shown in Figure 5-5. Interestingly, as suggested by Figure 5-4, VxInsight finds at least three groups within the set of 38 strains growing in stationary phase, which we had not previously recognized.



**Figure 5- 4. Gray-scale rendering of the Earth Mover Distances (EMD) between each of 38 genes of interest in Chapter 4. Rows (and similarly columns) have been sorted to order the genes by increasing distance from Cit1p. Brighter pixels indicate larger distances (greater dissimilarity), so the diagonal is always black representing an EMD of zero between a gene and itself. The plot suggests the 38 genes may fall into at least three groups by distance from Cit1p.**



**Figure 5-5.** VxInsight finds (A) three subclusters within the 38 genes from Chapter 4; (B) close-up view of the Nce102p subcluster; (C) and of the Cit1p subcluster; (D) and the Faa1p subcluster. Lines indicate the two genes are closely similar.

### The Bigger Challenge

Beyond learning to make better use of the multidimensional proteomic data from the flow cytometers, the bigger challenge is to continue to develop analysis methods that simultaneously use new (and old) time course studies in combination with the increasingly complete model organism databases to more thoroughly analyze the data and extract biological implications from the results. Specifically, this middle outward way must be extended to exploit automatically the detailed knowledge in these databases (for example, the Gene Ontology projects have been useful in this manner, but a much deeper approach is required).

If particularly rapid progress occurred in genomics as a result of collaboration between computer scientists, statisticians, and biologists, then the next burst of progress is likely to depend on widening the collaboration to machine learning researchers, a broader group of applied mathematicians, and to knowledge engineers. This much wider collaborative effort will be required to build the smart tools that will relate experimental data collected in specific laboratories with the broader knowledge reported in the literature, summarized in the model organism databases and in the more general gene and protein data resources.

Of course, discipline-specific research remains to be accomplished as individual parts of the collaboration. However, the art and skill involved in engineering successful collaborations must also be explicitly addressed because these collaborations are not easy to develop. They take time to put together, they require commitment, trust, and sometimes excruciating honesty between collaborators. Also, they require a surprisingly long period of working together before the participants begin to share a common language, realize what is possible, and discover what each discipline offers the others. Consequently, the real challenge will be how to initiate and continue the training of the researchers required to develop the tools to continue exploiting the middle outward approach.

## **Conclusion**

As discussed above, there are good theoretical reasons to believe that biologists will find ways to deal with even greater accumulations of details, will continue to extract a more comprehensive body of knowledge about cellular machinery, and will develop more and more powerful technologies. For example, consider how high-throughput, multi-level analysis is used to understand mechanisms in tissue complex eukaryotes or to analyze cellular microenvironments in cancer research. Biology itself, and the practice of biological research, remain as exciting as they ever were. They will, however, become much more integrated with the research programs and goals of other disciplines, which will present difficult but not insurmountable challenges. The integration of statistical data analysis, computing and information visualization with biology first motivated my

interactions with biologists, and has been the story of my research and of this dissertation; it has been a fruitful journey.

## References

- Akil, H., Brenner, S., Kandel, E., Kendler, K. S., King, M. C., Scolnick, E., et al. (2010). The future of psychiatric research: genomes and neural circuits. *Science*, 327(5973), 1580-1581.
- Allen, C., Buttner, S., Aragon, A. D., Thomas, J. A., Meirelles, O., Jaetao, J. E., et al. (2006). Isolation of quiescent and nonquiescent cells from yeast stationary-phase cultures. *Journal of Cell Biology*, 174, 89-100.
- Alwine, J. C., Kemp, D. J., & Stark, G. R. (1977). Method for detection of specific RNAs in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with DNA probes. *Proceedings of the National Academy of Sciences of the United States of America*, 74(12), 5350-5354.
- Aragon, A. D., Quinones, G., Allen, C., Thomas, J., Roy, S., Davidson, G., et al. (2005). An automated, pressure-driven sampling device for harvesting from liquid cultures for genomic and biochemical analyses. *Journal of Microbiological Methods*, doi:10.1016/j.mimet.2005.08.015.
- Aragon, A. D., Quinones, G. A., Thomas, E. V., Roy, S., & Werner-Washburne, M. (2006). Release of extraction-resistant mRNA in stationary phase *Saccharomyces cerevisiae* produces a massive increase in transcript abundance in response to stress. *Genome Biology*, 7(2).
- Aragon, A. D., Rodriguez, A. L., Meirelles, O., Roy, S., Davidson, G. S., Tapia, P. H., et al. (2008). Characterization of differentiated quiescent and nonquiescent cells in yeast stationary-phase cultures. *Molecular Biology of the Cell*, 19(3), 1271-1280.

- Costanzo, M., Baryshnikova, A., Bellay, J., Kim, Y., Spear, E. D., Sevier, C. S., et al. (2010). The genetic landscape of a cell. *Science*, 327(5964), 425-431.
- Coulter, W. (1956). *High speed automatic blood cell counter and cell size analyzer*. Paper presented at the National Electronics Conference.
- Davidson, G. S., Martin, S., Boyack, K. W., Wylie, B. N., Martinez, J., Aragon, A., et al. (2007). Robust methods for microarray analysis. In M. Akay (Ed.), *Genomics and Proteomics Engineering in Medicine and Biology* (pp. 99-128). Hoboken, New Jersey: John Wiley & Sons, Inc.
- Davidson, G. S., Wylie, B. N., & Boyack, K. (2001). Cluster stability and the use of noise in interpretation of clustering. *Proc. IEEE Information Visualization, 2001*, 23-30.
- DeRisi, J., Penland, L., Brown, P. O., Bittner, M. L., Meltzer, P. S., Ray, M., et al. (1996). Use of a cDNA microarray to analyze gene expression patterns in human cancer. *Nature Genetics*, 14, 457-460.
- Ferry, R., Farr, L. J., & Hartman, M. (1949). The preparation and measurement of the concentration of dilute bacterial aerosols. *Chemical Reviews*, 44(2), 389-417.
- Fruhwrth-Schnatter, S., & Pyne, S. (2010). Bayesian inference for finite mixtures of univariate and multivariate skew-normal and skew-t distributions. *Biostatistics*, 11(2), 317-336.
- Fulwyler, M. J. (1965). Electronic separation of biological cells by volume. *Science*, 150(3698), 910-911.
- Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., et al. (2004). Bioconductor: open software for computational biology and bioinformatics. *Genome Biology*, 5, R80.

- Goffeau, A., Barrell, B. G., Bussey, H., Davis, R. W., Dujon, B., Feldmann, H., et al. (1996). Life with 6000 genes. *Science*, 274(5287), 546-567.
- Gucker, F. T., & Okonski, C. T. (1949). An improved photoelectronic counter for colloidal particles, suitable for size-distribution studies. *Journal of Colloid Science*, 4(6), 541-560.
- Hughes, T., Marton, M., Jones, A., Roberts, C., Stoughton, R., Armour, C., et al. (2000). Functional discovery via a compendium of expression profiles. *Cell*, 102, 109-126.
- Huh, W.-K., Falvo, J. V., Gerke, L. C., Carroll, A. S., Howson, R. W., Weissman, J. S., et al. (2003). Global analysis of protein localization in budding yeast. [10.1038/nature02026]. *Nature*, 425(6959), 686-691.
- Hulett, H. R., Bonner, W. A., Barrett, J., & Herzenberg, L. A. (1969). Cell sorting - automated separation of mammalian cells as a function of intracellular fluorescence. *Science*, 166(3906), 747-749.
- Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M., & Sakaki, Y. (2001). A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proceedings of the National Academy of Sciences of the United States of America*, 98(8), 4569-4574.
- Kim, S. K., Lund, J., Kiraly, M., Duke, K., Jiang, M., Stuart, J. M., et al. (2001). A gene expression map for *Caenorhabditis elegans*. *Science*, 293(5537), 2087-2092.
- Li, S. M., Armstrong, C. M., Bertin, N., Ge, H., Milstein, S., Boxem, M., et al. (2004). A map of the interactome network of the metazoan *C-elegans*. *Science*, 303(5657), 540-543.



- Ling, H., & Okada, K. (2007). An efficient earth mover's distance algorithm for robust histogram comparison. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29(5), 840-853.
- Martin, S., Davidson, G., May, E., Faulon, J., & Werner-Washburne, M. (2004). Inferring genetic networks from microarray data. *3rd International IEEE (CSB)*, 566-569.
- Melamed, M. R., Kamenski, L. A., & Boyse, E. A. (1969). Cytotoxic test automation - a live-dead cell differential counter. *Science*, 163(3864), 285-286.
- Pele, O., & Werman, M. (2009). *Fast and robust earth mover's distances*. Paper presented at the The Ninth Asian Conference on Computer Vision.
- Pyne, S., Hu, X. L., Wang, K., Rossin, E., Lin, T. I., Maier, L. M., et al. (2009). Automated high-dimensional flow cytometric data analysis. *Proceedings of the National Academy of Sciences of the United States of America*, 106(21), 8519-8524.
- Reich, M., Liefeld, T., Gould, J., Lerner, J., Tamayo, P., & Mesirov, J. P. (2006). GenePattern 2.0. *Nat Genet*, 38(5), 500-501.
- Rual, J. F., Venkatesan, K., Hao, T., Hirozane-Kishikawa, T., Dricot, A., Li, N., et al. (2005). Towards a proteome-scale map of the human protein-protein interaction network. *Nature*, 437(7062), 1173-1178.
- Rubner, Y., Tomasi, C., & Guibas, L. J. (2000). The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, 40(2), 99-121.
- Schena, M., & Davis, R. (2000). Technology standards for microarray research. In M. Schena (Ed.), *Microarray Biochip Technology* (pp. 1-18). Natick, MA: Eaton Publishing, Biotechniques Book Division.

- Schena, M., Shalon, D., Davis, R. W., & Brown, P. O. (1995). Quantitative monitoring of gene-expression patterns with a complementary-DNA microarray. *Science*, 270, 467-470.
- Schubert, W., Bonnekoh, B., Pommer, A. J., Philipsen, L., Bockelmann, R., Malykh, Y., et al. (2006). Analyzing proteome topology and function by automated multidimensional fluorescence microscopy. *Nature Biotechnology*, 24(10), 1270-1278.
- Schwikowski, B., Uetz, P., & Fields, S. (2000). A network of protein-protein interactions in yeast. *Nature Biotechnology* 18(12), 1257-1261.
- SGD. (2010). SMD: Microarray Resources : Software and Tools, <http://smd.stanford.edu/resources/restech.shtml> (5/1/2010).
- SGD project. (May 1, 2010). Saccharomyces Genome Database <http://www.yeastgenome.org/> (May 1, 2010).
- Shapiro, H. M. (2003). *Practical Flow Cytometry* (4 ed.). Hoboken, New Jersey: John Wiley & Sons, Inc.
- Southern, E. (2006). Southern blotting. *Nature Protocols*, 1(2), 518-525.
- Southern, E. M. (1975). Detection of specific sequences among DNA fragments separated by gel-electrophoresis. *Journal of Molecular Biology*, 98(3), 503-517.
- Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., et al. (1998). Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Molecular Biology of the Cell*, 9(12), 3273-3297.

- Srinivasan, B. S., Caberoy, N. B., Suen, G., Taylor, R. G., Shah, R., Tengra, F., et al. (2005). Functional genome annotation through phylogenomic mapping. [10.1038/nbt1098]. *Nat Biotech*, 23(6), 691-698.
- Stuart, J. M., Segal, E., Koller, D., & Kim, S. K. (2003). A gene-coexpression network for global discovery of conserved genetic modules. *Science*, 302(5643), 249-255.
- The *C. elegans* Sequencing Consortium. (1998). Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science*, 282(5396), 2012-2018.
- The MathWorks. (2010). Bioinformatics Toolbox 3.5, <http://www.mathworks.com/products/bioinfo/> (5/1/2010).
- Tibshirani, R., Hastie, T., Narasimhan, B., & Chu, G. (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences of the United States of America*, 99(10), 6567-6572.
- Tsien, R. Y. (1998). The green fluorescent protein. *Annual Review of Biochemistry*, 67, 509-544.
- Tweedie, S., Ashburner, M., Falls, K., Leyland, P., McQuilton, P., Marygold, S., et al. (2009). FlyBase: enhancing *Drosophila* gene ontology annotations. *Nucleic Acids Research*, 37, D555-D559.
- Werner-Washburne, M., Wylie, B., Boyack, K., Fuge, E., Galbraith, J., Weber, J., et al. (2002). Comparative analysis of multiple genome-scale data sets. *Genome Research*, 12(10), 1564-1573.
- WormBase web site. <http://www.wormbase.org>, release WS204, date 5/1/2010.

Wu, T. T., Chen, Y. F., Hastie, T., Sobel, E., & Lange, K. (2009). Genome-wide association analysis by LASSO penalized logistic regression. *Bioinformatics*, 25(6), 714-721.

Yu, H. Y., Braun, P., Yildirim, M. A., Lemmens, I., Venkatesan, K., Sahalie, J., et al. (2008). High-quality binary protein interaction map of the yeast interactome network. *Science*, 322(5898), 104-110.

**Appendix I – Earth Mover Distances from Cit1p to 38 Genes in Chapter 4**

**EMD from Cit1p to each of the 38 genes of interest in Chapter 4 (Part 1 of 4)**

	1	2	3	4	5	6	7	8	9	10
	YNR001C (CIT1)	YDL215C (GDH2)	YBR222C (PCS60)	YER024W (YAT2)	YIL125W (KGD1)	YAL054C (ACS1)	YFL014W (HSP12)	YPL078C (ATP4)	YNL104C (LEU4)	YDR148C (KGD2)
YNR001C (CIT1)	0	364	386	388	401	499	527	558	565	578
YDL215C (GDH2)	364	0	339	374	452	497	577	573	578	663
YBR222C (PCS60)	386	339	0	408	313	580	446	473	644	817
YER024W (YAT2)	388	374	408	0	461	475	491	703	510	651
YIL125W (KGD1)	401	452	313	461	0	606	335	397	596	843
YAL054C (ACS1)	499	497	580	475	606	0	667	830	711	893
YFL014W (HSP12)	527	577	446	491	335	667	0	606	517	831
YPL078C (ATP4)	558	573	473	703	397	830	606	0	738	906
YNL104C (LEU4)	565	578	644	510	596	711	517	738	0	570
YDR148C (KGD2)	578	663	817	651	843	893	831	906	570	0
YBR039W (ATP3)	610	531	616	761	620	874	711	459	558	671
YCR088W (ABP1)	619	679	681	729	608	873	651	655	420	604
YDL181W (INH1)	653	557	457	708	547	798	545	583	553	775
YGR019W (UGA1)	663	650	717	565	730	805	608	874	314	604
YLL041C (SDH2)	668	454	602	758	709	807	747	693	745	811
YHR051W (COX6)	710	881	965	827	938	1063	921	984	790	534
YGR086C (PIL1)	713	565	746	822	823	905	842	822	656	695
YOR065W (CYT1)	739	611	739	858	729	898	760	635	604	830
YOR317W (FAA1)	779	682	765	772	809	786	729	883	537	802
YMR108W (ILV2)	783	638	774	819	772	867	736	770	541	868
YNL015W (PBI2)	799	619	764	717	896	815	797	988	643	793
YDR298C (ATP5)	805	1015	1095	986	1082	1187	1124	1074	982	601
YPR149W (NCE102)	835	949	1105	957	1127	1151	1112	1161	921	529
YAL044C (GCV3)	858	952	821	726	729	1014	608	991	738	987
YBL056W (PTC3)	868	908	1054	822	1077	1123	1022	1137	605	376

YGL187C (COX4)	870	1089	1120	1003	1070	1257	1118	1093	1031	726
YFR033C (QCR6)	933	1144	1178	1104	1142	1202	1239	1171	1245	939
YER141W (COX15)	979	783	970	1017	1026	1039	996	1012	826	994
YLR295C (ATP14)	988	1005	924	1037	781	1214	781	684	843	1031
YOR027W (STI1)	1071	957	971	1086	957	1079	869	1046	800	1128
YOR136W (IDH2)	1077	1100	898	1002	828	1110	693	1091	1062	1359
YMR092C (AIP1)	1102	967	931	1171	957	1144	932	919	931	1209
YDR381W (YRA1)	1110	968	987	1180	1068	1161	1049	1009	1030	1242
YDR129C (SAC6)	1134	1016	1020	1228	997	1246	1016	843	946	1188
YDR070C (FMP16)	1152	1082	951	1260	1039	1161	1082	1008	1350	1465
YGL062W (PYC1)	1154	921	1138	1187	1264	1208	1255	1238	1131	1122
YBR126C (TPS1)	1360	1132	1224	1399	1332	1366	1320	1260	1295	1453
YDR224C (HTB1)	1417	1208	1196	1380	1293	1347	1236	1306	1315	1616

## EMD from Cit1p to each of the 38 genes of interest in Chapter 4 (Part 2 of 4)

	11 YBR039W (ATP3)	12 YCR088W (ABP1)	13 YDL181W (INH1)	14 YGR019W (UGA1)	15 YLL041C (SDH2)	16 YHR051W (COX6)	17 YGR086C (PIL1)	18 YOR065W (CYT1)	19 YOR317W (FAA1)	20 YMR108W (ILV2)
YNR001C (CIT1)	610	619	653	663	668	710	713	739	779	783
YDL215C (GDH2)	531	679	557	650	454	881	565	611	682	638
YBR222C (PCS60)	616	681	457	717	602	965	746	739	765	774
YER024W (YAT2)	761	729	708	565	758	827	822	858	772	819
YIL125W (KGD1)	620	608	547	730	709	938	823	729	809	772
YAL054C (ACS1)	874	873	798	805	807	1063	905	898	786	867
YFL014W (HSP12)	711	651	545	608	747	921	842	760	729	736
YPL078C (ATP4)	459	655	583	874	693	984	822	635	883	770
YNL104C (LEU4)	558	420	553	314	745	790	656	604	537	541
YDR148C (KGD2)	671	604	775	604	811	534	695	830	802	868
YBR039W (ATP3)	0	455	439	590	529	868	453	313	568	472
YCR088W (ABP1)	455	0	467	464	808	815	535	491	489	545
YDL181W (INH1)	439	467	0	522	583	955	494	460	444	528
YGR019W (UGA1)	590	464	522	0	701	843	514	577	425	517
YLL041C (SDH2)	529	808	583	701	0	951	541	575	710	697
YHR051W (COX6)	868	815	955	843	951	0	894	967	979	1062
YGR086C (PIL1)	453	535	494	514	541	894	0	422	334	518
YOR065W (CYT1)	313	491	460	577	575	967	422	0	417	281
YOR317W (FAA1)	568	489	444	425	710	979	334	417	0	390
YMR108W (ILV2)	472	545	528	517	697	1062	518	281	390	0
YNL015W (PBI2)	676	717	516	456	649	1030	426	615	370	591
YDR298C (ATP5)	997	878	1109	1071	1154	445	1004	1104	1119	1233
YPR149W (NCE102)	955	954	1067	923	963	333	881	1061	1026	1149
YAL044C (GCV3)	1084	880	959	809	1188	1096	1198	1168	1066	1093
YBL056W (PTC3)	819	679	911	603	1027	692	774	924	839	931
YGL187C (COX4)	1112	994	1196	1158	1290	484	1213	1247	1310	1344
YFR033C (QCR6)	1275	1209	1356	1397	1387	752	1402	1425	1502	1530
YER141W (COX15)	689	726	653	691	717	1122	388	481	394	491

YLR295C (ATP14)	779	730	798	900	1038	1129	1041	865	1034	931
YOR027W (STI1)	792	723	635	662	844	1228	632	570	444	466
YOR136W (IDH2)	1256	1115	978	1084	1200	1407	1337	1276	1189	1220
YMR092C (AIP1)	752	785	585	802	771	1269	664	538	554	527
YDR381W (YRA1)	816	844	674	902	828	1286	616	629	573	691
YDR129C (SAC6)	691	761	703	835	804	1246	700	532	643	558
YDR070C (FMP16)	1167	1343	1022	1334	912	1544	1272	1242	1308	1324
YGL062W (PYC1)	929	1158	962	974	698	1305	691	882	858	927
YBR126C (TPS1)	1020	1240	955	1124	820	1509	859	883	903	939
YDR224C (HTB1)	1169	1374	1037	1161	933	1735	1104	1045	1051	1037



## EMD from Cit1p to each of the 38 genes of interest in Chapter 4 (Part 3 of 4)

	21 YNL015W (PBI2)	22 YDR298C (ATP5)	23 YPR149W (NCE102)	24 YAL044C (GCV3)	25 YBL056W (PTC3)	26 YGL187C (COX4)	27 YFR033C (QCR6)	28 YER141W (COX15)	29 YLR295C (ATP14)	30 YOR027W (STI1)
YNR001C (CIT1)	799	805	835	858	868	870	933	979	988	1071
YDL215C (GDH2)	619	1015	949	952	908	1089	1144	783	1005	957
YBR222C (PCS60)	764	1095	1105	821	1054	1120	1178	970	924	971
YER024W (YAT2)	717	986	957	726	822	1003	1104	1017	1037	1086
YIL125W (KGD1)	896	1082	1127	729	1077	1070	1142	1026	781	957
YAL054C (ACS1)	815	1187	1151	1014	1123	1257	1202	1039	1214	1079
YFL014W (HSP12)	797	1124	1112	608	1022	1118	1239	996	781	869
YPL078C (ATP4)	988	1074	1161	991	1137	1093	1171	1012	684	1046
YNL104C (LEU4)	643	982	921	738	605	1031	1245	826	843	800
YDR148C (KGD2)	793	601	529	987	376	726	939	994	1031	1128
YBR039W (ATP3)	676	997	955	1084	819	1112	1275	689	779	792
YCR088W (ABP1)	717	878	954	880	679	994	1209	726	730	723
YDL181W (INH1)	516	1109	1067	959	911	1196	1356	653	798	635
YGR019W (UGA1)	456	1071	923	809	603	1158	1397	691	900	662
YLL041C (SDH2)	649	1154	963	1188	1027	1290	1387	717	1038	844
YHR051W (COX6)	1030	445	333	1096	692	484	752	1122	1129	1228
YGR086C (PIL1)	426	1004	881	1198	774	1213	1402	388	1041	632
YOR065W (CYT1)	615	1104	1061	1168	924	1247	1425	481	865	570
YOR317W (FAA1)	370	1119	1026	1066	839	1310	1502	394	1034	444
YMR108W (ILV2)	591	1233	1149	1093	931	1344	1530	491	931	466
YNL015W (PBI2)	0	1211	1033	1105	814	1352	1525	551	1153	666
YDR298C (ATP5)	1211	0	446	1266	755	344	541	1271	1268	1445
YPR149W (NCE102)	1033	446	0	1285	649	606	828	1145	1294	1295
YAL044C (GCV3)	1105	1266	1285	0	1073	1174	1315	1356	969	1234
YBL056W (PTC3)	814	755	649	1073	0	835	1095	1046	1158	1158
YGL187C (COX4)	1352	344	606	1174	835	0	371	1471	1262	1609
YFR033C (QCR6)	1525	541	828	1315	1095	371	0	1663	1435	1815
YER141W (COX15)	551	1271	1145	1356	1046	1471	1663	0	1193	476
YLR295C (ATP14)	1153	1268	1294	969	1158	1262	1435	1193	0	1093
YOR027W (STI1)	666	1445	1295	1234	1158	1609	1815	476	1093	0
YOR136W (IDH2)	1186	1609	1587	717	1526	1560	1659	1437	1110	1224
YMR092C (AIP1)	736	1461	1353	1345	1266	1619	1798	496	1026	290

YDR381W (YRA1)	760	1399	1351	1430	1319	1624	1777	434	1250	562
YDR129C (SAC6)	858	1431	1351	1373	1254	1581	1777	592	951	388
YDR070C (FMP16)	1261	1679	1603	1464	1713	1727	1755	1391	1111	1344
YGL062W (PYC1)	728	1483	1212	1611	1203	1659	1786	711	1489	932
YBR126C (TPS1)	878	1718	1512	1715	1531	1878	2008	664	1503	783
YDR224C (HTB1)	975	2024	1786	1588	1670	2112	2224	999	1474	884

## EMD from Cit1p to each of the 38 genes of interest in Chapter 4 (Part 4 of 4)

	31 YOR136W (IDH2)	32 YMR092C (AIP1)	33 YDR381W (YRA1)	34 YDR129C (SAC6)	35 YDR070C (FMP16)	36 YGL062W (PYC1)	37 YBR126C (TPS1)	38 YDR224C (HTB1)
YNR001C (CIT1)	1077	1102	1110	1134	1152	1154	1360	1417
YDL215C (GDH2)	1100	967	968	1016	1082	921	1132	1208
YBR222C (PCS60)	898	931	987	1020	951	1138	1224	1196
YER024W (YAT2)	1002	1171	1180	1228	1260	1187	1399	1380
YIL125W (KGD1)	828	957	1068	997	1039	1264	1332	1293
YAL054C (ACS1)	1110	1144	1161	1246	1161	1208	1366	1347
YFL014W (HSP12)	693	932	1049	1016	1082	1255	1320	1236
YPL078C (ATP4)	1091	919	1009	843	1008	1238	1260	1306
YNL104C (LEU4)	1062	931	1030	946	1350	1131	1295	1315
YDR148C (KGD2)	1359	1209	1242	1188	1465	1122	1453	1616
YBR039W (ATP3)	1256	752	816	691	1167	929	1020	1169
YCR088W (ABP1)	1115	785	844	761	1343	1158	1240	1374
YDL181W (INH1)	978	585	674	703	1022	962	955	1037
YGR019W (UGA1)	1084	802	902	835	1334	974	1124	1161
YLL041C (SDH2)	1200	771	828	804	912	698	820	933
YHR051W (COX6)	1407	1269	1286	1246	1544	1305	1509	1735
YGR086C (PIL1)	1337	664	616	700	1272	691	859	1104
YOR065W (CYT1)	1276	538	629	532	1242	882	883	1045
YOR317W (FAA1)	1189	554	573	643	1308	858	903	1051
YMR108W (ILV2)	1220	527	691	558	1324	927	939	1037
YNL015W (PBI2)	1186	736	760	858	1261	728	878	975
YDR298C (ATP5)	1609	1461	1399	1431	1679	1483	1718	2024
YPR149W (NCE102)	1587	1353	1351	1351	1603	1212	1512	1786
YAL044C (GCV3)	717	1345	1430	1373	1464	1611	1715	1588
YBL056W (PTC3)	1526	1266	1319	1254	1713	1203	1531	1670
YGL187C (COX4)	1560	1619	1624	1581	1727	1659	1878	2112
YFR033C (QCR6)	1659	1798	1777	1777	1755	1786	2008	2224
YER141W (COX15)	1437	496	434	592	1391	711	664	999
YLR295C (ATP14)	1110	1026	1250	951	1111	1489	1503	1474
YOR027W (STI1)	1224	290	562	388	1344	932	783	884
YOR136W (IDH2)	0	1255	1379	1356	1133	1626	1584	1332
YMR092C (AIP1)	1255	0	495	295	1144	967	709	815
YDR381W (YRA1)	1379	495	0	586	1303	941	720	1018

YDR129C (SAC6)	1356	295	586	0	1233	1035	825	914
YDR070C (FMP16)	1133	1144	1303	1233	0	1358	1235	1099
YGL062W (PYC1)	1626	967	941	1035	1358	0	631	992
YBR126C (TPS1)	1584	709	720	825	1235	631	0	728
YDR224C (HTB1)	1332	815	1018	914	1099	992	728	0

## Appendix II – Top, Middle, and Bottom 20 Genes from Cit1p

Genes sorted by increasing Earth Mover's Distance from Cit1p (top, middle, and bottom of the list).

Rank Order	Experiment	FromORF	ToORF	EMD	In 38?	gene	Alt. Names	Description
<b>Top Twenty</b>								
1	SP	YNR001C	YNR001C	0	Yes	CIT1	CS1 LYS6	Citrate synthase, catalyzes the condensation of acetyl coenzyme A and oxaloacetate to form citrate; the rate-limiting enzyme of the TCA cycle; nuclear encoded mitochondrial protein
2	SP	YNR001C	YKL085W	265.82308		MDH1		Mitochondrial malate dehydrogenase, catalyzes interconversion of malate and oxaloacetate; involved in the tricarboxylic acid (TCA) cycle; phosphorylated
3	SP	YNR001C	YOR374W	269.504048		ALD4	ALD7 ALDH2	Mitochondrial aldehyde dehydrogenase, required for growth on ethanol and conversion of acetaldehyde to acetate; phosphorylated; activity is K <sup>+</sup> dependent; utilizes NADP <sup>+</sup> or NAD <sup>+</sup> equally as coenzymes; expression is glucose repressed
4	SP	YNR001C	YDL215C	363.855737	Yes	GDH2	GDH-B GDHB	NAD(+)-dependent glutamate dehydrogenase, degrades glutamate to ammonia and alpha-ketoglutarate; expression sensitive to nitrogen catabolite repression and intracellular ammonia levels
5	SP	YNR001C	YDR529C	370.02237		QCR7	COR4 CRO1 UCR7	Subunit 7 of the ubiquinol cytochrome-c reductase complex, which is a component of the mitochondrial inner membrane electron transport chain; oriented facing the mitochondrial matrix; N-terminus appears to play a role in complex assembly
6	SP	YNR001C	YBR222C	386.286402	Yes	PCS60	FAT2	Peroxisomal AMP-binding protein, localizes to both the peroxisomal peripheral membrane and matrix, expression is highly inducible by oleic acid, similar to E. coli long chain acyl-CoA synthetase
7	SP	YNR001C	YER024W	388.155838	Yes	YAT2		Carnitine acetyltransferase; has similarity to Yat1p, which is a carnitine acetyltransferase associated with the mitochondrial outer membrane
8	SP	YNR001C	YIL125W	400.927413	Yes	KGD1	OGD1	Component of the mitochondrial alpha-ketoglutarate dehydrogenase complex, which catalyzes a key step in the tricarboxylic acid (TCA) cycle, the oxidative decarboxylation of alpha-ketoglutarate to form succinyl-CoA
9	SP	YNR001C	YHR008C	415.9496		SOD2		Mitochondrial superoxide dismutase, protects cells against oxygen toxicity; phosphorylated
10	SP	YNR001C	YBR269C	427.501672		FMP21		Putative protein of unknown function; the authentic, non-tagged protein is detected in highly purified mitochondria in high-throughput studies
11	SP	YNR001C	YHR137W	438.103701		ARO9		Aromatic aminotransferase II, catalyzes the first step of tryptophan, phenylalanine, and tyrosine catabolism
12	SP	YNR001C	YBL015W	497.391923		ACH1		Protein with CoA transferase activity, particularly for CoASH transfer from succinyl-CoA to acetate; has minor acetyl-CoA-hydrolase activity; phosphorylated; required for acetate utilization

and for diploid pseudohyphal growth

13	SP	YNR001C	<b>YAL054C</b>	498.578034	Yes	<b>ACS1</b>	FUN44	Acetyl-coA synthetase isoform which, along with Acs2p, is the nuclear source of acetyl-coA for histone acetylation; expressed during growth on nonfermentable carbon sources and under aerobic conditions
14	SP	YNR001C	<b>YHR001W-A</b>	510.716937		<b>QCR10</b>		Subunit of the ubiquinol-cytochrome c oxidoreductase complex which includes Cobp, Rip1p, Cyt1p, Cor1p, Qcr2p, Qcr6p, Qcr7p, Qcr8p, Qcr9p, and Qcr10p and comprises part of the mitochondrial respiratory chain
15	SP	YNR001C	<b>YPL262W</b>	519.574543		<b>FUM1</b>		Fumarase, converts fumaric acid to L-malic acid in the TCA cycle; cytosolic and mitochondrial distribution determined by the N-terminal targeting sequence, protein conformation, and status of glyoxylate shunt; phosphorylated in mitochondria
16	SP	YNR001C	<b>YMR189W</b>	523.939536		<b>GCV2</b>	GSD2	P subunit of the mitochondrial glycine decarboxylase complex, required for the catabolism of glycine to 5,10-methylene-THF; expression is regulated by levels of 5,10-methylene-THF in the cytoplasm
17	SP	YNR001C	<b>YFL014W</b>	526.61456	Yes	<b>HSP12</b>	GLP1 HOR5	Plasma membrane localized protein that protects membranes from desiccation; induced by heat shock, oxidative stress, osmotic stress, stationary phase entry, glucose depletion, oleate and alcohol; regulated by the HOG and Ras-Pka pathways
18	SP	YNR001C	<b>YML120C</b>	532.560443		<b>NDI1</b>		NADH:ubiquinone oxidoreductase, transfers electrons from NADH to ubiquinone in the respiratory chain but does not pump protons, in contrast to the higher eukaryotic multisubunit respiratory complex I; phosphorylated; homolog of human AMID
19	SP	YNR001C	<b>YLR393W</b>	535.712799		<b>ATP10</b>		Mitochondrial inner membrane protein required for assembly of the F <sub>0</sub> sector of mitochondrial F <sub>1</sub> F <sub>0</sub> ATP synthase, interacts genetically with ATP6
20	SP	YNR001C	<b>YPL111W</b>	543.917882		<b>CAR1</b>	LPH15 cargA	Arginase, responsible for arginine degradation, expression responds to both induction by arginine and nitrogen catabolite repression; disruption enhances freeze tolerance

[...] Middle Twenty

1914	SP	YNR001C	<b>YPL212C</b>	922.678441		<b>PUS1</b>		tRNA:pseudouridine synthase, introduces pseudouridines at positions 26-28, 34-36, 65, and 67 of tRNA; nuclear protein that appears to be involved in tRNA export; also acts on U2 snRNA
1915	SP	YNR001C	<b>YPL183C</b>	922.707192		<b>RTT10</b>		Cytoplasmic protein with a role in regulation of Ty1 transposition
1916	SP	YNR001C	<b>YJR053W</b>	922.852167		<b>BFA1</b>	IBD1	Component of the GTPase-activating Bfa1p-Bub2p complex involved in multiple cell cycle checkpoint pathways that control exit from mitosis
1917	SP	YNR001C	<b>YMR178W</b>	922.974697		<b>MMT1</b>		
1918	SP	YNR001C	<b>YBR166C</b>	923.083917		<b>TYR1</b>		Prephenate dehydrogenase involved in tyrosine biosynthesis, expression is dependent on phenylalanine levels
1919	SP	YNR001C	<b>YJL010C</b>	923.182093		<b>NOP9</b>		Essential subunit of U3-containing 90S preribosome involved in production of 18S rRNA and assembly of small ribosomal subunit; also part of pre-40S ribosome and required for its export into cytoplasm; binds RNA and contains pumilio domain
1920	SP	YNR001C	<b>YBR279W</b>	923.263052		<b>PAF1</b>		Component of the Paf1p complex that binds to and modulates the activity of RNA polymerases I and II; required for expression of a

							subset of genes, including cell cycle-regulated genes; homolog of human PD2/hPAF1
1921	SP	YNR001C	<b>YLR035C-A</b>	923.391488	<b>MLH2</b>		
1922	SP	YNR001C	<b>YOR205C</b>	923.710765	<b>GEP3</b>	AIM40 FMP38 LRC5	Protein of unknown function; null mutant is defective in respiration and interacts synthetically with prohibitin (phb1); the authentic, non-tagged protein is detected in purified mitochondria in high-throughput studies
1923	SP	YNR001C	<b>YBR159W</b>	923.725569	<b>IFA38</b>		Microsomal beta-keto-reductase; contains oleate response element (ORE) sequence in the promoter region; mutants exhibit reduced VLCFA synthesis, accumulate high levels of dihydroshingosine, phytosphingosine and medium-chain ceramides
1924	SP	YNR001C	<b>YPL093W</b>	923.736015	<b>NOG1</b>		Putative GTPase that associates with free 60S ribosomal subunits in the nucleolus and is required for 60S ribosomal subunit biogenesis; constituent of 66S pre-ribosomal particles; member of the ODN family of nucleolar G-proteins
1925	SP	YNR001C	<b>YIL110W</b>	923.841715	<b>MNI1</b>		Putative S-adenosylmethionine-dependent methyltransferase of the seven beta-strand family; deletion mutant exhibits a weak vacuolar protein sorting defect, enhanced resistance to caspofungin, and is synthetically lethal with MEN mutants
1926	SP	YNR001C	<b>YKR086W</b>	923.879343	<b>PRP16</b>	PRP23 RNA16	RNA helicase in the DEAH-box family involved in the second catalytic step of splicing, exhibits ATP-dependent RNA unwinding activity
1927	SP	YNR001C	<b>YHR066W</b>	924.06706	<b>SSF1</b>		Constituent of 66S pre-ribosomal particles, required for ribosomal large subunit maturation; functionally redundant with Ssf2p; member of the Brix family
1928	SP	YNR001C	<b>YER080W</b>	924.165588	<b>AIM9</b>	FMP29	Putative protein of unknown function; the authentic, non-tagged protein is detected in highly purified mitochondria in high-throughput studies; null mutant displays elevated frequency of mitochondrial genome loss
1929	SP	YNR001C	<b>YMR295C</b>	924.447111	<b>IBI2</b>		Protein of unknown function that associates with ribosomes; green fluorescent protein (GFP)-fusion protein localizes to the cell periphery and bud; YMR295C is not an essential gene
1930	SP	YNR001C	<b>YDL161W</b>	924.502035	<b>ENT1</b>		Epsin-like protein involved in endocytosis and actin patch assembly and functionally redundant with Ent2p; binds clathrin via a clathrin-binding domain motif at C-terminus
1931	SP	YNR001C	<b>YLR423C</b>	924.643327	<b>ATG17</b>	APG17	Scaffold protein responsible for phagophore assembly site organization; regulatory subunit of an autophagy-specific complex that includes Atg1p and Atg13p; stimulates Atg1p kinase activity
1932	SP	YNR001C	<b>YLR005W</b>	924.726589	<b>SSL1</b>		Component of the core form of RNA polymerase transcription factor TFIIF, which has both protein kinase and DNA-dependent ATPase/helicase activities and is essential for transcription and nucleotide excision repair; interacts with Tfb4p
1933	SP	YNR001C	<b>YLR367W</b>	924.996678	<b>RPS22B</b>		Protein component of the small (40S) ribosomal subunit; nearly identical to Rps22Ap and has similarity to E. coli S8 and rat S15a ribosomal proteins

<b>[...] Bottom Twenty</b>							
3830	SP	YNR001C	<b>YLR048W</b>	1970.51702	<b>RPS0B</b>	NAB1B YST2	Protein component of the small (40S) ribosomal subunit, nearly identical to Rps0Ap; required for maturation of 18S rRNA along with Rps0Ap; deletion of either RPS0 gene reduces growth rate,

							deletion of both genes is lethal
3831	SP	YNR001C	<b>YDL184C</b>	1973.0256	<b>RPL41A</b>	RPL47A	Ribosomal protein L47 of the large (60S) ribosomal subunit, identical to Rpl41Bp and has similarity to rat L41 ribosomal protein; comprised of only 25 amino acids; rpl41a rpl41b double null mutant is viable
3832	SP	YNR001C	<b>YLR287C-A</b>	1973.60828	<b>RPS30A</b>		Protein component of the small (40S) ribosomal subunit; nearly identical to Rps30Bp and has similarity to rat S30 ribosomal protein
3833	SP	YNR001C	<b>YGR192C</b>	1977.29566	<b>TDH3</b>	GLD1 HSP35 HSP36 SSS2	Glyceraldehyde-3-phosphate dehydrogenase, isozyme 3, involved in glycolysis and gluconeogenesis; tetramer that catalyzes the reaction of glyceraldehyde-3-phosphate to 1,3 bis-phosphoglycerate; detected in the cytoplasm and cell wall
3834	SP	YNR001C	<b>YGR285C</b>	1982.17295	<b>ZUO1</b>		Cytosolic ribosome-associated chaperone that acts, together with Ssz1p and the Ssb proteins, as a chaperone for nascent polypeptide chains; contains a DnaJ domain and functions as a J-protein partner for Ssb1p and Ssb2p
3835	SP	YNR001C	<b>YBR118W</b>	1993.45624	<b>TEF2</b>	EF-1 alpha	Translational elongation factor EF-1 alpha; also encoded by TEF1; functions in the binding reaction of aminoacyl-tRNA (AA-tRNA) to ribosomes; may also have a role in tRNA re-export from the nucleus
3836	SP	YNR001C	<b>YDL081C</b>	1998.62389	<b>RPP1A</b>	RPLA1	Ribosomal stalk protein P1 alpha, involved in the interaction between translational elongation factors and the ribosome; accumulation of P1 in the cytoplasm is regulated by phosphorylation and interaction with the P2 stalk component
3837	SP	YNR001C	<b>YGL031C</b>	1998.89097	<b>RPL24A</b>	RPL30A	Ribosomal protein L30 of the large (60S) ribosomal subunit, nearly identical to Rpl24Bp and has similarity to rat L24 ribosomal protein; not essential for translation but may be required for normal translation rate
3838	SP	YNR001C	<b>YGL189C</b>	2033.57593	<b>RPS26A</b>	RPS26	Protein component of the small (40S) ribosomal subunit; nearly identical to Rps26Bp and has similarity to rat S26 ribosomal protein
3839	SP	YNR001C	<b>YKL117W</b>	2058.09957	<b>SBA1</b>		Co-chaperone that binds to and regulates Hsp90 family chaperones; important for pp60v-src activity in yeast; homologous to the mammalian p23 proteins and like p23 can regulate telomerase activity
3840	SP	YNR001C	<b>YDR418W</b>	2071.19489	<b>RPL12B</b>		Protein component of the large (60S) ribosomal subunit, nearly identical to Rpl12Ap; rpl12a rpl12b double mutant exhibits slow growth and slow translation; has similarity to E. coli L11 and rat L12 ribosomal proteins
3841	SP	YNR001C	<b>YCR012W</b>	2094.01123	<b>PGK1</b>		3-phosphoglycerate kinase, catalyzes transfer of high-energy phosphoryl groups from the acyl phosphate of 1,3-bisphosphoglycerate to ADP to produce ATP; key enzyme in glycolysis and gluconeogenesis
3842	SP	YNR001C	<b>YPR080W</b>	2130.28168	<b>TEF1</b>	EF-1 alpha	Translational elongation factor EF-1 alpha; also encoded by TEF2; functions in the binding reaction of aminoacyl-tRNA (AA-tRNA) to ribosomes; may also have a role in tRNA re-export from the nucleus
3843	SP	YNR001C	<b>YHR174W</b>	2146.25996	<b>ENO2</b>		Enolase II, a phosphopyruvate hydratase that catalyzes the conversion of 2-phosphoglycerate to phosphoenolpyruvate during glycolysis and the reverse reaction during gluconeogenesis; expression is induced in response to glucose



3844	SP	YNR001C	<b>YPR163C</b>	2174.90351	<b>TIF3</b>	RBL3 STM1	Translation initiation factor eIF-4B, has RNA annealing activity; contains an RNA recognition motif and binds to single-stranded RNA
3845	SP	YNR001C	<b>YDR382W</b>	2189.96099	<b>RPP2B</b>	RPL45 YPA1	Ribosomal protein P2 beta, a component of the ribosomal stalk, which is involved in the interaction between translational elongation factors and the ribosome; regulates the accumulation of P1 (Rpp1Ap and Rpp1Bp) in the cytoplasm
3846	SP	YNR001C	<b>YDL130W</b>	2197.61243	<b>RPP1B</b>	RPL44' RPLA3	Ribosomal protein P1 beta, component of the ribosomal stalk, which is involved in interaction of translational elongation factors with ribosome; accumulation is regulated by phosphorylation and interaction with the P2 stalk component
3847	SP	YNR001C	<b>YKL060C</b>	2200.10785	<b>FBA1</b>	LOT1	Fructose 1,6-bisphosphate aldolase, required for glycolysis and gluconeogenesis; catalyzes conversion of fructose 1,6 bisphosphate to glyceraldehyde-3-P and dihydroxyacetone-P; locates to mitochondrial outer surface upon oxidative stress
3848	SP	YNR001C	<b>YER131W</b>	2202.6556	<b>RPS26B</b>		Protein component of the small (40S) ribosomal subunit; nearly identical to Rps26Ap and has similarity to rat S26 ribosomal protein
3849	SP	YNR001C	<b>YGL135W</b>	2312.07257	<b>RPL1B</b>	SSM2	N-terminally acetylated protein component of the large (60S) ribosomal subunit, nearly identical to Rpl1Ap and has similarity to E. coli L1 and rat L10a ribosomal proteins; rpl1a rpl1b double null mutation is lethal