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

Recommended Citation

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.
George S. Davidson

Biology

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

Approved by the Dissertation Committee:

[Signatures]

Chairperson

[Signatures]

[Signatures]
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 ἀνεμος 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 multicellular eukaryote, *Caenorhabditis elegans*. Prior to this research approximately 6% of the *C. elegans* genome had been studied, and little was know 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/G0) and protein-interaction studies. Here, I find that G1-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 G0 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 Caenorhabditis elegans ................................................... 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
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 G1-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 G1-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 et al ........ 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 gene1; 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, Kamentsk, & 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.

1 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 \textit{Saccharomyces cerevisiae} and \textit{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 \textit{in vitro} protein-protein interactions to study actual \textit{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 though 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 G0, 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*


1Department of Developmental Biology and Genetics, Stanford University Medical School, Stanford, CA 94305, USA. 2Stanford Medical Informatics, 251 Campus Drive, MSOB X-215, Stanford, CA 94305, USA. 3Computation, 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 log2 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. [Force-
directed placement and data mining with VxInsight are described in (11, 12), available
Online at 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.](image)

(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 log10 (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 redrew 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.

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

**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 lineD 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 mountain (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 *Entemeba 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-\(\text{N}\)-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\(_{\alpha}\) 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

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

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


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).


13. Web Figures, tables, movies, and text are available on *Science* Online at 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.


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: (number of genes in common between both lists)(number of genes in the genome)/(number of genes in list A)(number of genes in list B).


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.


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,1,3 Brian Wylie2 Kevin Boyack,2 Edwina Fuge,1 Judith Galbraith,1 Jose Weber,1 and George Davidson2, “Comparative Analysis of Multiple Genome-Scale Data Sets”, Genome Research, 2002 12: 1564-1573

1Biology Department, University of New Mexico, Albuquerque, New Mexico 87131, USA; 2Sandia 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/ G₀) and protein-interaction studies. This analysis led to new insights about each data set. For example, G₁-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 G₀ 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/maggieww/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 α-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 G₀, is an offshoot of the mitotic cell cycle, and cells exiting G₀ reenter mitosis at G₁ (Werner-Washburne et al. 1993). In addition, starvation-induced G₀ 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 G₀ cells, and the proof-of-principal for chemotherapeutics is the ability to restore G₀ 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 Cryptoccus 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 G₀ 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 G₀.

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 α-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 G₁-, S-, M-, and M- G₁–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 G₁ 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. α-Factor-arrest data set (18 time points) ordinated and visualized in VxInsight. (A) Cell cycle gene expression after α-factor arrest and the dendogram 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. G1, S, and M: Genes in these clusters are induced during the G1, 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 G0 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 α-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 G1-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 G1-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 G1-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.

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 G1 phase–regulated genes (Fig. 3-6). Although both data sets contain G1-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 G1 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 G1-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 G1-regulated cluster that encode interactive proteins common to both data sets (Fig. 3-6 (D)).
Figure 3-6. Interactions among proteins encoded by G1-regulated genes. (A) Topographical presentation of G1-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 G1-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), 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). (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 G1- 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 G1-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 G0 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$_{600}$ of 2 in fresh YPD and returned to 30°C. Samples (~40 OD$_{600}$ 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/ choloroform extraction and DNase treatment.

Radiolabeled ($[^{33}P]$-dCTP) cDNA “probe” was obtained by reverse transription of total RNA (2 μg) following the protocol from Research Genetics (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, G1-Regulated Genes**

Genes that are strongly up-regulated in G1-phase in the α-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 G₁-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 G₁-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


**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; 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,*,† Ray M. Joe,*, Sushmita Roy,† Osorio Meirelles,*, Chris P. Allen,§ Melissa R. Wilson,* Swagata Chakraborty,* Anne E. Dodson,* Elaine E. Manzanilla,*, Mark Carter,§ Susan Young,§ Bruce Edwards,§Larry Sklar,§ Phillip H. Tapia,*, Margaret Werner-Washburne*‡

Affiliations: *Biology Department; †Computer Science Department, University of New Mexico, Albuquerque, NM, 87131; §Department of Cytometry; ‖Department of Pathology, University of New Mexico School of Medicine, Albuquerque, NM, 87131; ‡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

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 genically unstable (Aragon *et al*., 2008).
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-electrophoreograms 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 MgCl2, 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 4th, 8th, 12th, 16th, 20th, and 24th 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 5x10^6 cells were suspended in 500 μl of filter sterilized (0.22 μm) 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 ± 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 1x10^8 S288c upper and lower fraction cells per sample were pelleted and resuspended in 100 μl of the YPD+a, supernatant that had been filter sterilized. 1 μl 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 1x10^6 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 ± one standard deviation for each sample.

**Rate of oxygen consumption assay.** Rates of oxygen consumption were determined using
the BD™ 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 1x10^8 cells/ml; 200 µl 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 pO₂ using the following equation: pO₂ = (DR/NRF – 1)/Ksv, 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 Ksv is the Stern-Volmer constant, which was
calculated using the following equation and then converted to units of atm⁻¹: Ksv = (DR –
1)/\rho_{O_2}A$, where $\rho_{O_2}A$ is the partial pressure of oxygen at ambient conditions. $\rho_{O_2}A$ 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 (2x10^{-4}L). Rates were determined from the slope of the regression line for Time(s) vs. mol O₂/cell. Final rates were calculated as mol O₂/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 log₂ 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](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, Δlog-FI, the difference
between SP and EXP log-FI was computed for each of the three technical replicates. A
regression of Δlog-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 ≥ 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 ≥ 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), Enol1p: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≤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,
especially 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
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 thermostolerant 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


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


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.


**Table 4-1. Most abundant proteins in EXP and SP**

<table>
<thead>
<tr>
<th>Systematic name</th>
<th>Gene Name</th>
<th>Function</th>
<th>Localization</th>
<th>Log2EXP</th>
<th>Log2SP</th>
</tr>
</thead>
<tbody>
<tr>
<td>YLR044C</td>
<td>PDC1</td>
<td>Pyruvate decarboxylase</td>
<td>c, n</td>
<td>10.39</td>
<td>8.41</td>
</tr>
<tr>
<td>YHR174W</td>
<td>ENO2</td>
<td>Enolase II, a phosphopyruvate hydratase</td>
<td>c, m, pm</td>
<td>10.12</td>
<td>9.47</td>
</tr>
<tr>
<td>YGR192C</td>
<td>TDH3</td>
<td>Glyceraldehyde-3-phosphate dehydrogenase</td>
<td>c, n, pm</td>
<td>10.01</td>
<td>8.75</td>
</tr>
<tr>
<td>YKL060C</td>
<td>FBA1</td>
<td>Fructose 1,6-bisphosphate aldolase</td>
<td>c, m</td>
<td>9.97</td>
<td>8.42</td>
</tr>
<tr>
<td>YPR080W</td>
<td>TEF1</td>
<td>Translational elongation factor EF-1 alpha</td>
<td>c</td>
<td>9.16</td>
<td>8.34</td>
</tr>
<tr>
<td>YLL024C</td>
<td>SSA2</td>
<td>ATP binding protein</td>
<td>c, n, pm</td>
<td>9.16</td>
<td>8.24</td>
</tr>
<tr>
<td>YBR118W</td>
<td>TEF2</td>
<td>Translational elongation factor EF-1 alpha</td>
<td>c</td>
<td>9.09</td>
<td>7.81</td>
</tr>
<tr>
<td>YJR009C</td>
<td>TDH2</td>
<td>Glyceraldehyde-3-phosphate dehydrogenase</td>
<td>c, m, pm</td>
<td>8.91</td>
<td>7.59</td>
</tr>
<tr>
<td>YAL005C</td>
<td>SSA1</td>
<td>ATPase</td>
<td>c, n</td>
<td>8.71</td>
<td>9.05</td>
</tr>
<tr>
<td>YLR249W</td>
<td>YEF3</td>
<td>Translational elongation factor 3</td>
<td>c</td>
<td>8.54</td>
<td>6.62</td>
</tr>
<tr>
<td>YDL055C</td>
<td>PSA1</td>
<td>GDP-mannose pyrophosphorylase</td>
<td>c</td>
<td>8.5</td>
<td>6.47</td>
</tr>
<tr>
<td>YMR186W</td>
<td>HSC82</td>
<td>Cytoplasmic chaperone of the Hsp90 family</td>
<td>c, m, pm</td>
<td>8.49</td>
<td>6.85</td>
</tr>
<tr>
<td>YLR109W</td>
<td>AHP1</td>
<td>Thiol-specific peroxiredoxin</td>
<td>c, pm</td>
<td>8.23</td>
<td>5.78</td>
</tr>
<tr>
<td>YBR196C</td>
<td>PGI1</td>
<td>Glycolytic enzyme phosphoglucone isomerase</td>
<td>c, m, pm</td>
<td>7.7</td>
<td>5.64</td>
</tr>
<tr>
<td>YMR205C</td>
<td>PFK2</td>
<td>Beta subunit of hetero-octameric phosphofructokinase</td>
<td>c, m</td>
<td>7.66</td>
<td>6.16</td>
</tr>
<tr>
<td>YDR382W</td>
<td>RPP2B</td>
<td>Ribosomal protein P2 beta</td>
<td>c</td>
<td>7.52</td>
<td>6.1</td>
</tr>
<tr>
<td>YDL130W</td>
<td>RPP1B</td>
<td>Ribosomal protein P1 beta</td>
<td>c</td>
<td>7.5</td>
<td>5.66</td>
</tr>
<tr>
<td>YDL081C</td>
<td>RPP1A</td>
<td>Ribosomal stalk protein P1 alpha</td>
<td>c</td>
<td>7.42</td>
<td>6.08</td>
</tr>
<tr>
<td>YBR189W</td>
<td>RPS9B</td>
<td>Protein component of the small (40S) ribosomal subunit</td>
<td>c</td>
<td>7.32</td>
<td>5.98</td>
</tr>
<tr>
<td>YGL135W</td>
<td>RPL1B</td>
<td>N-terminally acetylated protein component of the large (60S) ribosomal subunit</td>
<td>c</td>
<td>7.33</td>
<td>5.76</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-------</td>
<td>-------</td>
<td>------------------------------------------------------------------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
</tr>
<tr>
<td>YGL253W</td>
<td>91</td>
<td>YGL253W HXK2 Hexokinase isoenzyme 2</td>
<td>c, m, n</td>
<td></td>
<td>00</td>
</tr>
<tr>
<td>YDR070C</td>
<td>YGL253W HXK2 Putative protein of unknown function</td>
<td>m</td>
<td></td>
<td>6.97</td>
<td></td>
</tr>
<tr>
<td>YDL185W</td>
<td>YGL253W HXK2 The A subunit of the V-ATPase V1 domain</td>
<td>c</td>
<td></td>
<td>6.74</td>
<td></td>
</tr>
<tr>
<td>YJL138C</td>
<td>YGL253W HXK2 Translation initiation factor eIF4A</td>
<td>c</td>
<td></td>
<td>6.83</td>
<td></td>
</tr>
<tr>
<td>YDL184C</td>
<td>YGL253W HXK2 Ribosomal protein L47 of the large (60S) ribosomal subunit</td>
<td>c</td>
<td></td>
<td>6.74</td>
<td></td>
</tr>
<tr>
<td>YFL014W</td>
<td>YGL253W HXK2 Heat-shock protein that protects membranes from desiccation</td>
<td>c, n, pm</td>
<td></td>
<td>3.89</td>
<td></td>
</tr>
<tr>
<td>YLR061W</td>
<td>YGL253W HXK2 Protein component of the large (60S) ribosomal subunit</td>
<td>c</td>
<td></td>
<td>6.62</td>
<td></td>
</tr>
<tr>
<td>YHR183W</td>
<td>YGL253W HXK2 6-phosphogluconate dehydrogenase (decarboxylating)</td>
<td>c, m</td>
<td></td>
<td>6.27</td>
<td></td>
</tr>
</tbody>
</table>
Table 4-2. GO process of proteins expressed 2-fold or higher in SP (87 proteins) than in EXP (121 proteins)

<table>
<thead>
<tr>
<th>Processes for proteins higher in SP</th>
<th>No. in each GO category</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>oxidative phosphorylation</td>
<td>11</td>
<td>2.8E-09</td>
</tr>
<tr>
<td>generation of precursor metabolites and energy</td>
<td>17</td>
<td>1.2E-06</td>
</tr>
<tr>
<td>electron transport chain</td>
<td>8</td>
<td>1.4E-06</td>
</tr>
<tr>
<td>respiratory electron transport chain</td>
<td>8</td>
<td>1.4E-06</td>
</tr>
<tr>
<td>ATP synthesis coupled electron transport</td>
<td>8</td>
<td>1.4E-06</td>
</tr>
<tr>
<td>mitochondrial ATP synthesis coupled electron transport</td>
<td>8</td>
<td>1.4E-06</td>
</tr>
<tr>
<td>oxidation reduction</td>
<td>8</td>
<td>1.4E-06</td>
</tr>
<tr>
<td>cofactor metabolic process</td>
<td>16</td>
<td>5.2E-06</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Processes for proteins higher in EXP</th>
<th>No. in each GO category</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>translation</td>
<td>49</td>
<td>5.5E-22</td>
</tr>
<tr>
<td>biosynthetic process</td>
<td>79</td>
<td>4.1E-11</td>
</tr>
<tr>
<td>cellular protein metabolic process</td>
<td>62</td>
<td>1.7E-10</td>
</tr>
<tr>
<td>protein metabolic process</td>
<td>63</td>
<td>3.0E-10</td>
</tr>
<tr>
<td>cellular biosynthetic process</td>
<td>74</td>
<td>7.6E-09</td>
</tr>
<tr>
<td>primary metabolic process</td>
<td>98</td>
<td>4.2E-07</td>
</tr>
<tr>
<td>Q cells</td>
<td>Cycling G1 cells</td>
<td>NQ unbudded daughters</td>
</tr>
<tr>
<td>---------</td>
<td>-----------------</td>
<td>-----------------------</td>
</tr>
<tr>
<td>100% form colonies&lt;sup&gt;1&lt;/sup&gt;</td>
<td>ND</td>
<td>~50% form colonies</td>
</tr>
<tr>
<td>No petite colonies produced (genomically stable)&lt;sup&gt;2&lt;/sup&gt;</td>
<td>ND</td>
<td>~40% petite colonies (genomically unstable)</td>
</tr>
<tr>
<td>Q cells give rise to Q daughters in the post-diauxic phase&lt;sup&gt;3&lt;/sup&gt;</td>
<td>ND</td>
<td>NQ cells give rise to NQ daughters in the post-diauxic phase</td>
</tr>
<tr>
<td>Produced concurrently with NQ unbudded daughters&lt;sup&gt;1&lt;/sup&gt;</td>
<td>Produced concurrently with Q daughters</td>
<td></td>
</tr>
<tr>
<td>Respiration&lt;sup&gt;3&lt;/sup&gt;</td>
<td>ND</td>
<td>Little or no respiration</td>
</tr>
<tr>
<td>Low ROS, no apoptosis&lt;sup&gt;1&lt;/sup&gt;</td>
<td>ND</td>
<td>50% with high ROS by day 7, 50% apoptotic by day 14</td>
</tr>
<tr>
<td>Typically high density (gm/cm&lt;sup&gt;3&lt;/sup&gt;)&lt;sup&gt;1&lt;/sup&gt;</td>
<td>Low density</td>
<td>Low density</td>
</tr>
<tr>
<td>High glycogen, trehalose&lt;sup&gt;1,4&lt;/sup&gt;</td>
<td>Low</td>
<td>No glycogen, low trehalose?</td>
</tr>
<tr>
<td>Synchronous for almost 2 cell divisions, lag phase 1.5 hours at 7d&lt;sup&gt;1,5&lt;/sup&gt;</td>
<td>Not as synchronous as Q cells, shorter lag phase</td>
<td>In NQ cells, atp17 and atp18 showed variability in petites and, cells from petite colonies produced both petite and non-petite colonies, suggesting epigenetic regulation of petites in these strains.</td>
</tr>
<tr>
<td>First daughter, no delayed G1 (mother and daughter bud concurrently)&lt;sup&gt;5&lt;/sup&gt;</td>
<td>Daughters have delayed G1</td>
<td>ND (not synchronous populations)</td>
</tr>
<tr>
<td>~2000 mRNAs in insoluble protein-mRNA complexes selectively released in response to different stresses&lt;sup&gt;6&lt;/sup&gt;</td>
<td>Few insoluble mRNAs</td>
<td>Few insoluble mRNAs</td>
</tr>
<tr>
<td>No observed effect of atp17 and atp18 mutants&lt;sup&gt;2&lt;/sup&gt;</td>
<td>ND</td>
<td>In NQ cells, atp17 and atp18 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.</td>
</tr>
</tbody>
</table>

<sup>1</sup> (Allen et al., 2006)<br><sup>2</sup> (Aragon et al., 2008)<br><sup>3</sup> (This work)<br><sup>4</sup> (Shi, et al, 2010)<br><sup>5</sup> (Allen, unpublished data)<br><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 ≥ 2-fold increases in SP. (X) 121 GFP-fusion proteins with ≥ 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 μmol/cell/sec; SP was 3.7 μmol/cell/sec; NQ was 3.6 μmol/cell/sec; and Q was 21.8 μmol/cell/sec. The difference between NQ and Q respiration was significant (p≤ .5.5E-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.
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

ATP14 NQ

ATP14 Q

ATP4 NQ

ATP4 Q

COX6 NQ

COX6 Q

ILV2 NQ

ILV2 Q

AIP1 NQ

AIP1 Q

Gamma subunit of the F1 sector of mitochondrial F1F0-ATPase, which is a large, evolutionally conserved enzyme complex required for ATP synthesis.

Subunit b of the stator stalk of mitochondrial F1F0-ATP synthase, which is a large, evolutionally conserved enzyme complex required for ATP synthesis. Phosphorylated.

Subunit b of the F0 sector of mitochondrial F1F0-ATP synthase, which is a large, evolutionally conserved enzyme complex required for ATP synthesis.

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.

Actin cytoskeletal component, interacts with the actin depolymerizing factor coflin; required for actin coflin localization to cortical patches; contains WD repeats.
IDH2 NQ

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

IDH2 Q

Type 2C protein phosphatase, dephosphorylates histones (see also P53p) to limit maximal kinase activity induced by genomic stress, dephosphorylates T109 phosphorylated G5620P (see also P53p), role in DNA checkpoint inactivation.

ACS1 NQ

Acetyl-CoA synthetase isoform which, along with Asp2, is the nuclear target of acetyl-CoA for threonine acetylation, expressed during growth on nonfermentable carbon sources and under anaerobic conditions.

ACS1 Q

CIT1 NQ

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

CIT1 Q

PCS60 NQ

Parasexual AMP-binding protein, localized to both the parasexual peripheral membrane and matrix, expression is highly inducible by osmotic shock, similar to E. coli long chain acyl-CoA synthetase.

PCS60 Q

PDS60 NQ

H. sapiens mitochondrial glycerol dehydrogenase complex, required for the conversion of glycerol to 3,3,5-trimethylaminotriol (TMT), also required for all protein C37a1 modification, expression is regulated by levels of 3,3,5-trimethylaminotriol.
LEU4 NQ

TPS1 NQ

COX4 NQ

LEU4 Q

TPS1 Q

COX4 Q

Alpha-ketoisovalerate synthase (3-ketoisovalerate synthase), the main isocitrate responsible for the first step in the isocitrate isocitrate pathway.

Synthase subunit of trehalose-6-phosphate synthase/hydrolases complex, which synthesizes the storage carbohydrate trehalose, also found in a monomeric form, expression is induced by the stress response and regulated by the Rnr1-MAP pathway.

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

ATP5 NQ

ATP15 NQ

GDH2 NQ

ATP5 Q

ATP15 Q

GDH2 Q

Subunit 6 of the stator stalk of mitochondrial F1/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 F1/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.
KGD1 NQ
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.

KGD1 Q

COX15 NQ
Protein required for the hydroxylation of heme C to form heme A, which is an essential prosthetic group for cytochrome c oxidase.

COX15 Q

SAC6 NQ
Fibrin, actin-binding protein, cooperates with Sso to cap protofilaments in the organization and maintenance of the actin cytoskeleton.

SAC6 Q

PIL1 NQ
Primary component of eosomes, which are large immobile cell sorting structures associated with endocytosis; null mutants show activation of Pho85/Sykp stress resistance pathways; detected in phosphorylated state in mitochondria.

PIL1 Q

HSP12 NQ
Plasma membrane localized protein that protects membranes from desiccation; induced by heat shock, oxidative stress, chemoresistance, stationary phase entry, glucose depletion, osmotic and alcohol; regulated by the HOG and Rsa/Rfa pathways.

HSP12 Q

YAT2 NQ
Carboxyl esterases; has similarity to Yatp, which is a carboxyl ester esterase associated with the mitochondrial outer membrane.

YAT2 Q
ABP1 NQ

ABP1 Q

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

HTB1 NQ

HTB1 Q

Histone H3B, core histone protein required for chromatin assembly and chromosomal function; nearly identical to HTB3. Rad54-like subfamily b mediated ubiquitination regulates transcriptional activation, meiotic DSB formation and H3 methylation.

INH1 NQ

INH1 Q

Protein that inhibits ATP hydrolysis by the F1-ATP synthase; inhibitory function is enhanced by stabilizing proteins SF1p and SF2; has similarity to SF1p; 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 saturated 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 domain; involved in secretion of proteins that lack classical secretory signal sequences; component of the detergent-insoluble glycerol-3-phosphate-stained complexes (DIX).

PBI2 NQ

PBI2 Q

Cytotoxic inhibitor of vacuolar protease B1, required for efficient vacuolar inheritance; with Rhodobacter capsulatus protein complex (LMA), which assists in priming SNARE molecules and promotes vacuolar 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 media plates and grown at 30°C for 2-3 days.

Wild type (S288c) 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 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 plated 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 (Fruhwirth-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.
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


SGD project. (May 1, 2010). Saccharomyces Genome Database http://www.yeastgenome.org/ (May 1, 2010).


## 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)

<table>
<thead>
<tr>
<th></th>
<th>YNR001C (CIT1)</th>
<th>YDL215C (GDH2)</th>
<th>YBR222C (PCS60)</th>
<th>YER024W (YAT2)</th>
<th>YIL125W (KGD1)</th>
<th>YAL054C (ACS1)</th>
<th>YFL014W (HSP12)</th>
<th>YPL078C (ATP4)</th>
<th>YNL104C (LEU4)</th>
<th>YDR148C (KGD2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>364</td>
<td>386</td>
<td>388</td>
<td>401</td>
<td>499</td>
<td>527</td>
<td>558</td>
<td>565</td>
<td>578</td>
</tr>
<tr>
<td>2</td>
<td>364</td>
<td>0</td>
<td>339</td>
<td>374</td>
<td>452</td>
<td>497</td>
<td>577</td>
<td>573</td>
<td>578</td>
<td>663</td>
</tr>
<tr>
<td>3</td>
<td>386</td>
<td>339</td>
<td>0</td>
<td>408</td>
<td>313</td>
<td>580</td>
<td>446</td>
<td>473</td>
<td>644</td>
<td>817</td>
</tr>
<tr>
<td>4</td>
<td>388</td>
<td>374</td>
<td>408</td>
<td>0</td>
<td>461</td>
<td>475</td>
<td>491</td>
<td>703</td>
<td>510</td>
<td>651</td>
</tr>
<tr>
<td>5</td>
<td>401</td>
<td>452</td>
<td>313</td>
<td>461</td>
<td>0</td>
<td>606</td>
<td>335</td>
<td>397</td>
<td>596</td>
<td>843</td>
</tr>
<tr>
<td>6</td>
<td>499</td>
<td>497</td>
<td>580</td>
<td>475</td>
<td>606</td>
<td>0</td>
<td>667</td>
<td>830</td>
<td>711</td>
<td>893</td>
</tr>
<tr>
<td>7</td>
<td>527</td>
<td>577</td>
<td>446</td>
<td>491</td>
<td>335</td>
<td>667</td>
<td>0</td>
<td>606</td>
<td>517</td>
<td>831</td>
</tr>
<tr>
<td>8</td>
<td>558</td>
<td>573</td>
<td>473</td>
<td>703</td>
<td>397</td>
<td>830</td>
<td>606</td>
<td>0</td>
<td>738</td>
<td>906</td>
</tr>
<tr>
<td>9</td>
<td>565</td>
<td>578</td>
<td>644</td>
<td>510</td>
<td>596</td>
<td>711</td>
<td>517</td>
<td>738</td>
<td>0</td>
<td>570</td>
</tr>
<tr>
<td>10</td>
<td>578</td>
<td>663</td>
<td>817</td>
<td>651</td>
<td>843</td>
<td>893</td>
<td>831</td>
<td>906</td>
<td>570</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>610</td>
<td>619</td>
<td>616</td>
<td>761</td>
<td>620</td>
<td>874</td>
<td>711</td>
<td>459</td>
<td>558</td>
<td>671</td>
</tr>
<tr>
<td></td>
<td>619</td>
<td>679</td>
<td>681</td>
<td>729</td>
<td>608</td>
<td>873</td>
<td>651</td>
<td>655</td>
<td>420</td>
<td>604</td>
</tr>
<tr>
<td></td>
<td>653</td>
<td>663</td>
<td>717</td>
<td>565</td>
<td>730</td>
<td>805</td>
<td>608</td>
<td>874</td>
<td>314</td>
<td>604</td>
</tr>
<tr>
<td></td>
<td>668</td>
<td>650</td>
<td>717</td>
<td>565</td>
<td>730</td>
<td>805</td>
<td>608</td>
<td>874</td>
<td>314</td>
<td>604</td>
</tr>
<tr>
<td></td>
<td>710</td>
<td>710</td>
<td>881</td>
<td>965</td>
<td>827</td>
<td>938</td>
<td>1063</td>
<td>921</td>
<td>984</td>
<td>790</td>
</tr>
<tr>
<td></td>
<td>713</td>
<td>713</td>
<td>565</td>
<td>746</td>
<td>822</td>
<td>823</td>
<td>905</td>
<td>842</td>
<td>822</td>
<td>656</td>
</tr>
<tr>
<td></td>
<td>739</td>
<td>739</td>
<td>611</td>
<td>739</td>
<td>858</td>
<td>729</td>
<td>898</td>
<td>760</td>
<td>635</td>
<td>830</td>
</tr>
<tr>
<td></td>
<td>779</td>
<td>779</td>
<td>682</td>
<td>765</td>
<td>772</td>
<td>809</td>
<td>786</td>
<td>729</td>
<td>883</td>
<td>537</td>
</tr>
<tr>
<td></td>
<td>783</td>
<td>783</td>
<td>638</td>
<td>774</td>
<td>819</td>
<td>772</td>
<td>867</td>
<td>736</td>
<td>770</td>
<td>541</td>
</tr>
<tr>
<td></td>
<td>799</td>
<td>799</td>
<td>619</td>
<td>764</td>
<td>717</td>
<td>896</td>
<td>815</td>
<td>797</td>
<td>988</td>
<td>643</td>
</tr>
<tr>
<td></td>
<td>805</td>
<td>805</td>
<td>1015</td>
<td>1095</td>
<td>986</td>
<td>1082</td>
<td>1187</td>
<td>1124</td>
<td>1074</td>
<td>982</td>
</tr>
<tr>
<td></td>
<td>835</td>
<td>835</td>
<td>949</td>
<td>1105</td>
<td>957</td>
<td>1127</td>
<td>1151</td>
<td>1112</td>
<td>1161</td>
<td>921</td>
</tr>
<tr>
<td></td>
<td>858</td>
<td>858</td>
<td>952</td>
<td>821</td>
<td>726</td>
<td>729</td>
<td>1014</td>
<td>608</td>
<td>991</td>
<td>738</td>
</tr>
<tr>
<td></td>
<td>868</td>
<td>868</td>
<td>908</td>
<td>1054</td>
<td>822</td>
<td>1077</td>
<td>1123</td>
<td>1022</td>
<td>1137</td>
<td>605</td>
</tr>
</tbody>
</table>

135
<table>
<thead>
<tr>
<th>Gene</th>
<th>YGL187C (COX4)</th>
<th>YFR033C (QCR6)</th>
<th>YER141W (COX15)</th>
<th>YLR295C (ATP14)</th>
<th>YOR027W (STI1)</th>
<th>YOR136W (IDH2)</th>
<th>YMR092C (AIP1)</th>
<th>YDR381W (YRA1)</th>
<th>YDR129C (SAC6)</th>
<th>YDR070C (FMP16)</th>
<th>YGL062W (PYC1)</th>
<th>YBR126C (TPS1)</th>
<th>YDR224C (HTB1)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>870</td>
<td>1089</td>
<td>1120</td>
<td>1003</td>
<td>1070</td>
<td>1257</td>
<td>1118</td>
<td>1093</td>
<td>1031</td>
<td>726</td>
<td>933</td>
<td>1144</td>
<td>1178</td>
</tr>
<tr>
<td></td>
<td>933</td>
<td>1144</td>
<td>1178</td>
<td>1104</td>
<td>1142</td>
<td>1202</td>
<td>1239</td>
<td>1171</td>
<td>1245</td>
<td>939</td>
<td>979</td>
<td>783</td>
<td>970</td>
</tr>
<tr>
<td></td>
<td>979</td>
<td>783</td>
<td>970</td>
<td>1017</td>
<td>1026</td>
<td>1039</td>
<td>996</td>
<td>1012</td>
<td>826</td>
<td>994</td>
<td>988</td>
<td>1005</td>
<td>924</td>
</tr>
<tr>
<td></td>
<td>988</td>
<td>1005</td>
<td>924</td>
<td>1037</td>
<td>781</td>
<td>1214</td>
<td>781</td>
<td>684</td>
<td>843</td>
<td>1031</td>
<td>1071</td>
<td>957</td>
<td>971</td>
</tr>
<tr>
<td></td>
<td>1071</td>
<td>957</td>
<td>971</td>
<td>1086</td>
<td>957</td>
<td>1079</td>
<td>869</td>
<td>1046</td>
<td>800</td>
<td>1128</td>
<td>1077</td>
<td>1100</td>
<td>898</td>
</tr>
<tr>
<td></td>
<td>1077</td>
<td>1100</td>
<td>898</td>
<td>1002</td>
<td>828</td>
<td>1110</td>
<td>693</td>
<td>1091</td>
<td>1062</td>
<td>1359</td>
<td>1102</td>
<td>967</td>
<td>931</td>
</tr>
<tr>
<td></td>
<td>1102</td>
<td>967</td>
<td>931</td>
<td>1171</td>
<td>957</td>
<td>1144</td>
<td>932</td>
<td>919</td>
<td>931</td>
<td>1209</td>
<td>1110</td>
<td>968</td>
<td>987</td>
</tr>
<tr>
<td></td>
<td>1110</td>
<td>968</td>
<td>987</td>
<td>1180</td>
<td>1068</td>
<td>1161</td>
<td>1049</td>
<td>1009</td>
<td>1030</td>
<td>1242</td>
<td>1134</td>
<td>1016</td>
<td>1020</td>
</tr>
<tr>
<td></td>
<td>1134</td>
<td>1016</td>
<td>1020</td>
<td>1228</td>
<td>997</td>
<td>1246</td>
<td>1016</td>
<td>843</td>
<td>946</td>
<td>1188</td>
<td>1152</td>
<td>1082</td>
<td>951</td>
</tr>
<tr>
<td></td>
<td>1152</td>
<td>1082</td>
<td>951</td>
<td>1260</td>
<td>1039</td>
<td>1161</td>
<td>1082</td>
<td>1008</td>
<td>1350</td>
<td>1465</td>
<td>1154</td>
<td>921</td>
<td>1138</td>
</tr>
<tr>
<td></td>
<td>1154</td>
<td>921</td>
<td>1138</td>
<td>1187</td>
<td>1264</td>
<td>1208</td>
<td>1255</td>
<td>1238</td>
<td>1131</td>
<td>1122</td>
<td>1360</td>
<td>1132</td>
<td>1224</td>
</tr>
<tr>
<td></td>
<td>1360</td>
<td>1132</td>
<td>1224</td>
<td>1399</td>
<td>1332</td>
<td>1366</td>
<td>1320</td>
<td>1260</td>
<td>1295</td>
<td>1453</td>
<td>1417</td>
<td>1208</td>
<td>1196</td>
</tr>
</tbody>
</table>
EMD from Cit1p to each of the 38 genes of interest in Chapter 4 (Part 2 of 4)

<table>
<thead>
<tr>
<th></th>
<th>YBR039W (ATP3)</th>
<th>YCR088W (ABP1)</th>
<th>YDL181W (INH1)</th>
<th>YGR019W (UGA1)</th>
<th>YLL041C (SDH2)</th>
<th>YHR051W (COX6)</th>
<th>YGR086C (PIL1)</th>
<th>YOR065W (CIT1)</th>
<th>YOR317W (FAA1)</th>
<th>YMR108W (ILV2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>610</td>
<td>619</td>
<td>653</td>
<td>663</td>
<td>668</td>
<td>710</td>
<td>713</td>
<td>739</td>
<td>779</td>
<td>783</td>
</tr>
<tr>
<td>12</td>
<td>531</td>
<td>679</td>
<td>557</td>
<td>650</td>
<td>454</td>
<td>881</td>
<td>565</td>
<td>611</td>
<td>682</td>
<td>638</td>
</tr>
<tr>
<td>13</td>
<td>616</td>
<td>681</td>
<td>457</td>
<td>717</td>
<td>602</td>
<td>965</td>
<td>746</td>
<td>739</td>
<td>765</td>
<td>774</td>
</tr>
<tr>
<td>14</td>
<td>761</td>
<td>729</td>
<td>708</td>
<td>565</td>
<td>758</td>
<td>827</td>
<td>822</td>
<td>858</td>
<td>772</td>
<td>819</td>
</tr>
<tr>
<td>15</td>
<td>620</td>
<td>608</td>
<td>547</td>
<td>730</td>
<td>709</td>
<td>938</td>
<td>823</td>
<td>729</td>
<td>809</td>
<td>772</td>
</tr>
<tr>
<td>16</td>
<td>874</td>
<td>873</td>
<td>798</td>
<td>805</td>
<td>807</td>
<td>1063</td>
<td>905</td>
<td>898</td>
<td>786</td>
<td>867</td>
</tr>
<tr>
<td>17</td>
<td>711</td>
<td>651</td>
<td>545</td>
<td>608</td>
<td>747</td>
<td>921</td>
<td>842</td>
<td>760</td>
<td>729</td>
<td>736</td>
</tr>
<tr>
<td>18</td>
<td>459</td>
<td>655</td>
<td>583</td>
<td>874</td>
<td>693</td>
<td>984</td>
<td>822</td>
<td>635</td>
<td>883</td>
<td>770</td>
</tr>
<tr>
<td>19</td>
<td>558</td>
<td>420</td>
<td>553</td>
<td>314</td>
<td>745</td>
<td>790</td>
<td>656</td>
<td>604</td>
<td>537</td>
<td>541</td>
</tr>
<tr>
<td>20</td>
<td>671</td>
<td>604</td>
<td>775</td>
<td>604</td>
<td>811</td>
<td>534</td>
<td>695</td>
<td>830</td>
<td>802</td>
<td>868</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>0</td>
<td>455</td>
<td>439</td>
<td>590</td>
<td>529</td>
<td>868</td>
<td>453</td>
<td>313</td>
<td>568</td>
<td>472</td>
</tr>
<tr>
<td>22</td>
<td>455</td>
<td>0</td>
<td>467</td>
<td>464</td>
<td>808</td>
<td>815</td>
<td>815</td>
<td>353</td>
<td>491</td>
<td>489</td>
</tr>
<tr>
<td>23</td>
<td>439</td>
<td>467</td>
<td>0</td>
<td>522</td>
<td>583</td>
<td>955</td>
<td>494</td>
<td>460</td>
<td>444</td>
<td>528</td>
</tr>
<tr>
<td>24</td>
<td>590</td>
<td>464</td>
<td>522</td>
<td>0</td>
<td>701</td>
<td>843</td>
<td>514</td>
<td>577</td>
<td>425</td>
<td>517</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>529</td>
<td>808</td>
<td>583</td>
<td>701</td>
<td>0</td>
<td>951</td>
<td>541</td>
<td>575</td>
<td>710</td>
<td>697</td>
</tr>
<tr>
<td>26</td>
<td>868</td>
<td>815</td>
<td>955</td>
<td>843</td>
<td>951</td>
<td>0</td>
<td>894</td>
<td>967</td>
<td>979</td>
<td>1062</td>
</tr>
<tr>
<td>27</td>
<td>453</td>
<td>535</td>
<td>494</td>
<td>514</td>
<td>541</td>
<td>894</td>
<td>0</td>
<td>422</td>
<td>334</td>
<td>518</td>
</tr>
<tr>
<td>28</td>
<td>313</td>
<td>491</td>
<td>460</td>
<td>577</td>
<td>575</td>
<td>967</td>
<td>422</td>
<td>0</td>
<td>417</td>
<td>281</td>
</tr>
<tr>
<td>29</td>
<td>568</td>
<td>489</td>
<td>444</td>
<td>425</td>
<td>710</td>
<td>979</td>
<td>334</td>
<td>417</td>
<td>0</td>
<td>390</td>
</tr>
<tr>
<td>30</td>
<td>472</td>
<td>545</td>
<td>528</td>
<td>517</td>
<td>697</td>
<td>1062</td>
<td>518</td>
<td>281</td>
<td>390</td>
<td>0</td>
</tr>
<tr>
<td>31</td>
<td>676</td>
<td>717</td>
<td>516</td>
<td>456</td>
<td>649</td>
<td>1030</td>
<td>426</td>
<td>615</td>
<td>370</td>
<td>591</td>
</tr>
<tr>
<td>32</td>
<td>997</td>
<td>878</td>
<td>1109</td>
<td>1071</td>
<td>1154</td>
<td>445</td>
<td>1004</td>
<td>1104</td>
<td>1119</td>
<td>1233</td>
</tr>
<tr>
<td>33</td>
<td>955</td>
<td>954</td>
<td>1067</td>
<td>923</td>
<td>963</td>
<td>333</td>
<td>881</td>
<td>1061</td>
<td>1026</td>
<td>1149</td>
</tr>
<tr>
<td>34</td>
<td>1084</td>
<td>880</td>
<td>959</td>
<td>809</td>
<td>1188</td>
<td>1096</td>
<td>1198</td>
<td>1168</td>
<td>1066</td>
<td>1093</td>
</tr>
<tr>
<td>35</td>
<td>819</td>
<td>679</td>
<td>911</td>
<td>603</td>
<td>1027</td>
<td>692</td>
<td>774</td>
<td>924</td>
<td>839</td>
<td>931</td>
</tr>
<tr>
<td>36</td>
<td>1112</td>
<td>994</td>
<td>1196</td>
<td>1158</td>
<td>1290</td>
<td>484</td>
<td>1213</td>
<td>1247</td>
<td>1310</td>
<td>1344</td>
</tr>
<tr>
<td>37</td>
<td>1275</td>
<td>1209</td>
<td>1356</td>
<td>1397</td>
<td>1387</td>
<td>752</td>
<td>1402</td>
<td>1425</td>
<td>1502</td>
<td>1530</td>
</tr>
<tr>
<td>38</td>
<td>689</td>
<td>726</td>
<td>653</td>
<td>691</td>
<td>717</td>
<td>1122</td>
<td>388</td>
<td>481</td>
<td>394</td>
<td>491</td>
</tr>
<tr>
<td>Gene</td>
<td>YLR295C (ATP14)</td>
<td>YOR027W (STI1)</td>
<td>YOR136W (IDH2)</td>
<td>YMR092C (AIP1)</td>
<td>YDR381W (YRA1)</td>
<td>YDR129C (SAC6)</td>
<td>YDR070C (FMP16)</td>
<td>YGL062W (PYC1)</td>
<td>YBR126C (TPS1)</td>
<td>YDR224C (HTB1)</td>
</tr>
<tr>
<td>--------------</td>
<td>-----------------</td>
<td>----------------</td>
<td>----------------</td>
<td>----------------</td>
<td>----------------</td>
<td>----------------</td>
<td>----------------</td>
<td>----------------</td>
<td>----------------</td>
<td>----------------</td>
</tr>
<tr>
<td>YLR295C</td>
<td>779</td>
<td>730</td>
<td>798</td>
<td>900</td>
<td>1038</td>
<td>1129</td>
<td>1041</td>
<td>865</td>
<td>1034</td>
<td>931</td>
</tr>
<tr>
<td>YOR027W</td>
<td>792</td>
<td>723</td>
<td>635</td>
<td>662</td>
<td>844</td>
<td>1228</td>
<td>632</td>
<td>570</td>
<td>444</td>
<td>466</td>
</tr>
<tr>
<td>YOR136W</td>
<td>1256</td>
<td>1115</td>
<td>978</td>
<td>1084</td>
<td>1200</td>
<td>1407</td>
<td>1337</td>
<td>1276</td>
<td>1189</td>
<td>1220</td>
</tr>
<tr>
<td>YMR092C</td>
<td>752</td>
<td>785</td>
<td>585</td>
<td>802</td>
<td>771</td>
<td>1269</td>
<td>664</td>
<td>538</td>
<td>554</td>
<td>527</td>
</tr>
<tr>
<td>YDR381W</td>
<td>816</td>
<td>844</td>
<td>674</td>
<td>902</td>
<td>828</td>
<td>1286</td>
<td>616</td>
<td>629</td>
<td>573</td>
<td>691</td>
</tr>
<tr>
<td>YDR129C</td>
<td>691</td>
<td>761</td>
<td>703</td>
<td>835</td>
<td>804</td>
<td>1246</td>
<td>700</td>
<td>532</td>
<td>643</td>
<td>558</td>
</tr>
<tr>
<td>YDR070C</td>
<td>1167</td>
<td>1343</td>
<td>1022</td>
<td>1334</td>
<td>912</td>
<td>1544</td>
<td>1272</td>
<td>1242</td>
<td>1308</td>
<td>1324</td>
</tr>
<tr>
<td>YGL062W</td>
<td>929</td>
<td>1158</td>
<td>962</td>
<td>974</td>
<td>698</td>
<td>1305</td>
<td>691</td>
<td>882</td>
<td>858</td>
<td>927</td>
</tr>
<tr>
<td>YBR126C</td>
<td>1020</td>
<td>1240</td>
<td>955</td>
<td>1124</td>
<td>820</td>
<td>1509</td>
<td>859</td>
<td>883</td>
<td>903</td>
<td>939</td>
</tr>
<tr>
<td>YDR224C</td>
<td>1169</td>
<td>1374</td>
<td>1037</td>
<td>1161</td>
<td>933</td>
<td>1735</td>
<td>1104</td>
<td>1045</td>
<td>1051</td>
<td>1037</td>
</tr>
</tbody>
</table>
## EMD from Cit1p to each of the 38 genes of interest in Chapter 4 (Part 3 of 4)

<table>
<thead>
<tr>
<th></th>
<th>YNL015W (PBI2)</th>
<th>YDR928C (ATP5)</th>
<th>YPR149W (NCE102)</th>
<th>YAL054C (GCV3)</th>
<th>YBL056W (PTC3)</th>
<th>YGL187C (COX4)</th>
<th>YFR033C (QCR6)</th>
<th>YER141W (COX15)</th>
<th>YLR295C (ATP14)</th>
<th>YOR027W (STI1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>21</td>
<td>799</td>
<td>805</td>
<td>835</td>
<td>858</td>
<td>868</td>
<td>870</td>
<td>933</td>
<td>979</td>
<td>988</td>
<td>1071</td>
</tr>
<tr>
<td>22</td>
<td>619</td>
<td>1015</td>
<td>949</td>
<td>952</td>
<td>908</td>
<td>1089</td>
<td>1144</td>
<td>783</td>
<td>1005</td>
<td>957</td>
</tr>
<tr>
<td>23</td>
<td>764</td>
<td>1095</td>
<td>1105</td>
<td>821</td>
<td>1054</td>
<td>1120</td>
<td>1178</td>
<td>970</td>
<td>924</td>
<td>971</td>
</tr>
<tr>
<td>24</td>
<td>717</td>
<td>986</td>
<td>957</td>
<td>726</td>
<td>822</td>
<td>1003</td>
<td>1104</td>
<td>1017</td>
<td>1037</td>
<td>1086</td>
</tr>
<tr>
<td>25</td>
<td>896</td>
<td>1082</td>
<td>1127</td>
<td>729</td>
<td>1077</td>
<td>1070</td>
<td>1142</td>
<td>1026</td>
<td>781</td>
<td>957</td>
</tr>
<tr>
<td>26</td>
<td>815</td>
<td>1187</td>
<td>1151</td>
<td>1014</td>
<td>1123</td>
<td>1257</td>
<td>1202</td>
<td>1039</td>
<td>1214</td>
<td>1079</td>
</tr>
<tr>
<td>27</td>
<td>797</td>
<td>1124</td>
<td>1112</td>
<td>608</td>
<td>1022</td>
<td>1118</td>
<td>1239</td>
<td>996</td>
<td>781</td>
<td>869</td>
</tr>
<tr>
<td>28</td>
<td>988</td>
<td>1074</td>
<td>1161</td>
<td>991</td>
<td>1137</td>
<td>1093</td>
<td>1171</td>
<td>1012</td>
<td>684</td>
<td>1046</td>
</tr>
<tr>
<td>29</td>
<td>643</td>
<td>982</td>
<td>921</td>
<td>738</td>
<td>605</td>
<td>1031</td>
<td>1245</td>
<td>826</td>
<td>843</td>
<td>800</td>
</tr>
<tr>
<td>30</td>
<td>793</td>
<td>601</td>
<td>529</td>
<td>987</td>
<td>376</td>
<td>726</td>
<td>939</td>
<td>994</td>
<td>1031</td>
<td>1128</td>
</tr>
</tbody>
</table>

For example, YNR001C (CIT1) has an EMD of 799 to YNL015W (PBI2), and 988 to YNL104C (LEU4).
<table>
<thead>
<tr>
<th>Gene Name</th>
<th>Value 1</th>
<th>Value 2</th>
<th>Value 3</th>
<th>Value 4</th>
<th>Value 5</th>
<th>Value 6</th>
<th>Value 7</th>
<th>Value 8</th>
<th>Value 9</th>
<th>Value 10</th>
<th>Value 11</th>
</tr>
</thead>
<tbody>
<tr>
<td>YDR381W (YRA1)</td>
<td>760</td>
<td>1399</td>
<td>1351</td>
<td>1430</td>
<td>1319</td>
<td>1624</td>
<td>1777</td>
<td>434</td>
<td>1250</td>
<td>562</td>
<td></td>
</tr>
<tr>
<td>YDR129C (SAC6)</td>
<td>858</td>
<td>1431</td>
<td>1351</td>
<td>1373</td>
<td>1254</td>
<td>1581</td>
<td>1777</td>
<td>592</td>
<td>951</td>
<td>388</td>
<td></td>
</tr>
<tr>
<td>YDR070C (FMP16)</td>
<td>1261</td>
<td>1679</td>
<td>1603</td>
<td>1464</td>
<td>1713</td>
<td>1727</td>
<td>1755</td>
<td>1391</td>
<td>1111</td>
<td>1344</td>
<td></td>
</tr>
<tr>
<td>YGL062W (PYC1)</td>
<td>728</td>
<td>1483</td>
<td>1212</td>
<td>1611</td>
<td>1203</td>
<td>1659</td>
<td>1786</td>
<td>711</td>
<td>1489</td>
<td>932</td>
<td></td>
</tr>
<tr>
<td>YBR126C (TPS1)</td>
<td>878</td>
<td>1718</td>
<td>1512</td>
<td>1715</td>
<td>1531</td>
<td>1878</td>
<td>2008</td>
<td>664</td>
<td>1503</td>
<td>783</td>
<td></td>
</tr>
<tr>
<td>YDR224C (HTB1)</td>
<td>975</td>
<td>2024</td>
<td>1786</td>
<td>1588</td>
<td>1670</td>
<td>2112</td>
<td>2224</td>
<td>999</td>
<td>1474</td>
<td>884</td>
<td></td>
</tr>
</tbody>
</table>
EMD from Cit1p to each of the 38 genes of interest in Chapter 4 (Part 4 of 4)

<table>
<thead>
<tr>
<th></th>
<th>31 YOR136W (IDH2)</th>
<th>32 YMR092C (AIP1)</th>
<th>33 YDR381W (YRA1)</th>
<th>34 YDR129C (SAC6)</th>
<th>35 YDR070C (FMP16)</th>
<th>36 YGL062W (PYC1)</th>
<th>37 YBR126C (TPS1)</th>
<th>38 YDR224C (HTB1)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>YNR001C (CIT1)</td>
<td>1077</td>
<td>1102</td>
<td>1110</td>
<td>1134</td>
<td>1152</td>
<td>1154</td>
<td>1160</td>
<td>1417</td>
</tr>
<tr>
<td>YDL215C (GDH2)</td>
<td>1100</td>
<td>967</td>
<td>968</td>
<td>1016</td>
<td>1082</td>
<td>921</td>
<td>1132</td>
<td>1208</td>
</tr>
<tr>
<td>YBR222C (PCS60)</td>
<td>898</td>
<td>931</td>
<td>987</td>
<td>1020</td>
<td>951</td>
<td>1138</td>
<td>1224</td>
<td>1196</td>
</tr>
<tr>
<td>YER024W (YAT2)</td>
<td>1002</td>
<td>1171</td>
<td>1180</td>
<td>1228</td>
<td>1260</td>
<td>1187</td>
<td>1399</td>
<td>1380</td>
</tr>
<tr>
<td>YIL125W (KGD1)</td>
<td>828</td>
<td>957</td>
<td>1068</td>
<td>997</td>
<td>1039</td>
<td>1264</td>
<td>1332</td>
<td>1293</td>
</tr>
<tr>
<td>YAL054C (ACS1)</td>
<td>1110</td>
<td>1144</td>
<td>1161</td>
<td>1246</td>
<td>1161</td>
<td>1208</td>
<td>1366</td>
<td>1347</td>
</tr>
<tr>
<td>YFL014W (HSP12)</td>
<td>693</td>
<td>932</td>
<td>1049</td>
<td>1016</td>
<td>1082</td>
<td>1255</td>
<td>1320</td>
<td>1236</td>
</tr>
<tr>
<td>YPL078C (ATP4)</td>
<td>1091</td>
<td>919</td>
<td>1009</td>
<td>843</td>
<td>1008</td>
<td>1238</td>
<td>1260</td>
<td>1306</td>
</tr>
<tr>
<td>YNL104C (LEU4)</td>
<td>1062</td>
<td>931</td>
<td>1030</td>
<td>946</td>
<td>1350</td>
<td>1131</td>
<td>1295</td>
<td>1315</td>
</tr>
<tr>
<td>YDR148C (KGD2)</td>
<td>1359</td>
<td>1209</td>
<td>1242</td>
<td>1188</td>
<td>1465</td>
<td>1122</td>
<td>1453</td>
<td>1616</td>
</tr>
<tr>
<td>YBR039W (ATP3)</td>
<td>1256</td>
<td>752</td>
<td>816</td>
<td>691</td>
<td>1167</td>
<td>929</td>
<td>1020</td>
<td>1169</td>
</tr>
<tr>
<td>YCR088W (ABP1)</td>
<td>1115</td>
<td>785</td>
<td>844</td>
<td>761</td>
<td>1343</td>
<td>1158</td>
<td>1240</td>
<td>1374</td>
</tr>
<tr>
<td>YDL181W (INH1)</td>
<td>978</td>
<td>585</td>
<td>674</td>
<td>703</td>
<td>1022</td>
<td>962</td>
<td>955</td>
<td>1037</td>
</tr>
<tr>
<td>YGR019W (UGA1)</td>
<td>1084</td>
<td>802</td>
<td>902</td>
<td>835</td>
<td>1334</td>
<td>974</td>
<td>1124</td>
<td>1161</td>
</tr>
<tr>
<td>YLL041C (SDH2)</td>
<td>1200</td>
<td>771</td>
<td>828</td>
<td>804</td>
<td>912</td>
<td>698</td>
<td>820</td>
<td>933</td>
</tr>
<tr>
<td>YHR051W (COX6)</td>
<td>1407</td>
<td>1269</td>
<td>1286</td>
<td>1246</td>
<td>1544</td>
<td>1305</td>
<td>1509</td>
<td>1735</td>
</tr>
<tr>
<td>YGR086C (PIL1)</td>
<td>1337</td>
<td>664</td>
<td>616</td>
<td>700</td>
<td>1272</td>
<td>691</td>
<td>859</td>
<td>1104</td>
</tr>
<tr>
<td>YOR065W (CYT1)</td>
<td>1276</td>
<td>554</td>
<td>629</td>
<td>532</td>
<td>1242</td>
<td>882</td>
<td>883</td>
<td>1045</td>
</tr>
<tr>
<td>YOR317W (FAA1)</td>
<td>1189</td>
<td>554</td>
<td>573</td>
<td>643</td>
<td>1308</td>
<td>858</td>
<td>903</td>
<td>1051</td>
</tr>
<tr>
<td>YMR108W (ILV2)</td>
<td>1220</td>
<td>527</td>
<td>691</td>
<td>558</td>
<td>1324</td>
<td>927</td>
<td>939</td>
<td>1037</td>
</tr>
<tr>
<td>YNL015W (PB12)</td>
<td>1186</td>
<td>736</td>
<td>760</td>
<td>858</td>
<td>1261</td>
<td>728</td>
<td>878</td>
<td>975</td>
</tr>
<tr>
<td>YDR298C (ATP5)</td>
<td>1609</td>
<td>1461</td>
<td>1399</td>
<td>1431</td>
<td>1679</td>
<td>1483</td>
<td>1718</td>
<td>2024</td>
</tr>
<tr>
<td>YPR149W (NCE102)</td>
<td>1587</td>
<td>1353</td>
<td>1351</td>
<td>1351</td>
<td>1603</td>
<td>1212</td>
<td>1512</td>
<td>1786</td>
</tr>
<tr>
<td>YAL044C (GCV3)</td>
<td>717</td>
<td>1345</td>
<td>1430</td>
<td>1373</td>
<td>1464</td>
<td>1611</td>
<td>1715</td>
<td>1588</td>
</tr>
<tr>
<td>YBL056W (PTC3)</td>
<td>1526</td>
<td>1266</td>
<td>1319</td>
<td>1254</td>
<td>1713</td>
<td>1203</td>
<td>1531</td>
<td>1670</td>
</tr>
<tr>
<td>YGL187C (COX4)</td>
<td>1560</td>
<td>1619</td>
<td>1624</td>
<td>1581</td>
<td>1727</td>
<td>1659</td>
<td>1878</td>
<td>2112</td>
</tr>
<tr>
<td>YFR033C (QC6R)</td>
<td>1659</td>
<td>1798</td>
<td>1777</td>
<td>1777</td>
<td>1755</td>
<td>1786</td>
<td>2008</td>
<td>2224</td>
</tr>
<tr>
<td>YER141W (COX15)</td>
<td>1437</td>
<td>496</td>
<td>434</td>
<td>592</td>
<td>1391</td>
<td>711</td>
<td>664</td>
<td>999</td>
</tr>
<tr>
<td>YLR295C (ATP14)</td>
<td>1110</td>
<td>1026</td>
<td>1250</td>
<td>951</td>
<td>1111</td>
<td>1489</td>
<td>1503</td>
<td>1474</td>
</tr>
<tr>
<td>YOR027W (ST11)</td>
<td>1224</td>
<td>290</td>
<td>562</td>
<td>388</td>
<td>1344</td>
<td>932</td>
<td>783</td>
<td>884</td>
</tr>
<tr>
<td>YOR136W (IDH2)</td>
<td>0</td>
<td>1255</td>
<td>1379</td>
<td>1356</td>
<td>1133</td>
<td>1626</td>
<td>1584</td>
<td>1332</td>
</tr>
<tr>
<td>YMR092C (AIP1)</td>
<td>1255</td>
<td>0</td>
<td>495</td>
<td>295</td>
<td>1144</td>
<td>967</td>
<td>709</td>
<td>815</td>
</tr>
<tr>
<td>YDR381W (YRA1)</td>
<td>1379</td>
<td>495</td>
<td>0</td>
<td>586</td>
<td>1303</td>
<td>941</td>
<td>720</td>
<td>1018</td>
</tr>
<tr>
<td>Gene Symbol</td>
<td>YDR129C (SAC6)</td>
<td>YDR070C (FMP16)</td>
<td>YGL062W (PYC1)</td>
<td>YBR126C (TPS1)</td>
<td>YDR224C (HTB1)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>--------------</td>
<td>----------------</td>
<td>-----------------</td>
<td>----------------</td>
<td>----------------</td>
<td>----------------</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Value</td>
<td>1356</td>
<td>1133</td>
<td>1626</td>
<td>1584</td>
<td>1332</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>295</td>
<td>1144</td>
<td>967</td>
<td>709</td>
<td>815</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>586</td>
<td>1303</td>
<td>941</td>
<td>720</td>
<td>1018</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>1233</td>
<td>1035</td>
<td>825</td>
<td>914</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1233</td>
<td>0</td>
<td>1358</td>
<td>1235</td>
<td>631</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1035</td>
<td>1358</td>
<td>0</td>
<td>631</td>
<td>992</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>825</td>
<td>1235</td>
<td>631</td>
<td>631</td>
<td>728</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>914</td>
<td>1099</td>
<td>992</td>
<td>0</td>
<td>728</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## 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).

<table>
<thead>
<tr>
<th>Rank Order</th>
<th>Experiment</th>
<th>FromORF</th>
<th>ToORF</th>
<th>EMD</th>
<th>In 38?</th>
<th>gene</th>
<th>Alt. Names</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Top Twenty</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>SP</td>
<td>YNR001C</td>
<td>YNR001C</td>
<td>0</td>
<td>Yes</td>
<td>CIT1</td>
<td>CS1 LYS6</td>
<td>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</td>
</tr>
<tr>
<td>2</td>
<td>SP</td>
<td>YNR001C</td>
<td>YKL085W</td>
<td>265.82308</td>
<td>MDH1</td>
<td></td>
<td></td>
<td>Mitochondrial malate dehydrogenase, catalyzes interconversion of malate and oxaloacetate; involved in the tricarboxylic acid (TCA) cycle; phosphorylated</td>
</tr>
<tr>
<td>3</td>
<td>SP</td>
<td>YNR001C</td>
<td>YOR374W</td>
<td>269.504048</td>
<td>ALD4</td>
<td>ALD7 ALDH2</td>
<td>Mitochondrial aldehyde dehydrogenase, required for growth on ethanol and conversion of acetaldehyde to acetate; phosphorylated; activity is K⁺ dependent; utilizes NADP+ or NAD+ equally as coenzymes; expression is glucose repressed</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>SP</td>
<td>YNR001C</td>
<td>YDL215C</td>
<td>363.855737</td>
<td>Yes</td>
<td>GDH2</td>
<td>GDH-B GDHB</td>
<td>Mitochondrial aldehyde dehydrogenase, required for growth on ethanol and conversion of acetaldehyde to acetate; phosphorylated; activity is K⁺ dependent; utilizes NADP+ or NAD+ equally as coenzymes; expression is glucose repressed</td>
</tr>
<tr>
<td>5</td>
<td>SP</td>
<td>YNR001C</td>
<td>YDR529C</td>
<td>370.02237</td>
<td>QCR7</td>
<td>COR4 CRO1 UCR7</td>
<td>NAD(+)−dependent glutamate dehydrogenase, degrades glutamate to ammonia and alpha-ketoglutarate; expression sensitive to nitrogen catabolite repression and intracellular ammonia levels</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>SP</td>
<td>YNR001C</td>
<td>YBR222C</td>
<td>386.286402</td>
<td>Yes</td>
<td>PCS60</td>
<td>FAT2</td>
<td>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</td>
</tr>
<tr>
<td>7</td>
<td>SP</td>
<td>YNR001C</td>
<td>YER024W</td>
<td>388.155838</td>
<td>Yes</td>
<td>YAT2</td>
<td></td>
<td>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</td>
</tr>
<tr>
<td>8</td>
<td>SP</td>
<td>YNR001C</td>
<td>YIL125W</td>
<td>400.927413</td>
<td>Yes</td>
<td>KGD1</td>
<td>OGD1</td>
<td>Carnitine acetyltransferase; has similarity to Yat1p, which is a carnitine acetyltransferase associated with the mitochondrial outer membrane</td>
</tr>
<tr>
<td>9</td>
<td>SP</td>
<td>YNR001C</td>
<td>YHR008C</td>
<td>415.9496</td>
<td>Yes</td>
<td>SOD2</td>
<td></td>
<td>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</td>
</tr>
<tr>
<td>10</td>
<td>SP</td>
<td>YNR001C</td>
<td>YBR269C</td>
<td>427.501672</td>
<td></td>
<td>FMP21</td>
<td></td>
<td>Mitochondrial superoxide dismutase, protects cells against oxygen toxicity; phosphorylated</td>
</tr>
<tr>
<td>11</td>
<td>SP</td>
<td>YNR001C</td>
<td>YHR137W</td>
<td>438.103701</td>
<td></td>
<td>AR09</td>
<td></td>
<td>Putative protein of unknown function; the authentic, non-tagged protein is detected in highly purified mitochondria in high-throughput studies</td>
</tr>
<tr>
<td>12</td>
<td>SP</td>
<td>YNR001C</td>
<td>YBL015W</td>
<td>497.391923</td>
<td></td>
<td>ACH1</td>
<td></td>
<td>Aromatic aminotransferase II, catalyzes the first step of tryptophan, phenylalanine, and tyrosine catabolism</td>
</tr>
<tr>
<td>13</td>
<td>SP</td>
<td>YNR001C</td>
<td>YDR389C</td>
<td>542.715889</td>
<td>Yes</td>
<td>LIG1</td>
<td></td>
<td>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</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>SP</td>
<td>YNR001C</td>
<td>YAL054C</td>
<td>498.578034</td>
<td>Yes</td>
<td>ACS1</td>
<td>FUN44</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>SP</td>
<td>YNR001C</td>
<td>YHR001W-A</td>
<td>510.716937</td>
<td></td>
<td>QCR10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>SP</td>
<td>YNR001C</td>
<td>YPL262W</td>
<td>519.574543</td>
<td></td>
<td>FUM1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>SP</td>
<td>YNR001C</td>
<td>YMR189W</td>
<td>523.939536</td>
<td></td>
<td>GCV2</td>
<td>GSD2</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>SP</td>
<td>YNR001C</td>
<td>YFL014W</td>
<td>526.61456</td>
<td>Yes</td>
<td>HSP12</td>
<td>GLP1 HOP5</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>SP</td>
<td>YNR001C</td>
<td>YML120C</td>
<td>532.560443</td>
<td></td>
<td>NDI1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>SP</td>
<td>YNR001C</td>
<td>YLR393W</td>
<td>535.712799</td>
<td></td>
<td>ATP10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>SP</td>
<td>YNR001C</td>
<td>YPL111W</td>
<td>543.917882</td>
<td></td>
<td>CAR1</td>
<td>LPH15 cargA</td>
<td></td>
</tr>
</tbody>
</table>

For diploid pseudohyphal growth, 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. 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. 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. 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. 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. 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. Mitochondrial inner membrane protein required for assembly of the F0 sector of mitochondrial F1F0 ATP synthase, interacts genetically with ATP6. Arginase, responsible for arginine degradation, expression responds to both induction by arginine and nitrogen catabolite repression; disruption enhances freeze tolerance. 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. Cytoplasmic protein with a role in regulation of Ty1 transposition. Component of the GTPase-activating Bfa1p-Bub2p complex involved in multiple cell cycle checkpoint pathways that control exit from mitosis. Prephenate dehydrogenase involved in tyrosine biosynthesis, expression is dependent on phenylalanine levels. 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. Component of the Pax1p complex that binds to and modulates the activity of RNA polymerases I and II; required for expression of a
| S. No. | Organism | Gene ID | RPKM | H.s.
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1921</td>
<td>SP</td>
<td>YNR001C</td>
<td>YLR035C-A</td>
<td>923.391488</td>
</tr>
<tr>
<td>1922</td>
<td>SP</td>
<td>YNR001C</td>
<td>YOR205C</td>
<td>923.710765</td>
</tr>
<tr>
<td>1923</td>
<td>SP</td>
<td>YNR001C</td>
<td>YBR159W</td>
<td>923.725569</td>
</tr>
<tr>
<td>1924</td>
<td>SP</td>
<td>YNR001C</td>
<td>YPL093W</td>
<td>923.736015</td>
</tr>
<tr>
<td>1925</td>
<td>SP</td>
<td>YNR001C</td>
<td>YIL110W</td>
<td>923.841715</td>
</tr>
<tr>
<td>1926</td>
<td>SP</td>
<td>YNR001C</td>
<td>YKR086W</td>
<td>923.879343</td>
</tr>
<tr>
<td>1927</td>
<td>SP</td>
<td>YNR001C</td>
<td>YHR066W</td>
<td>924.06706</td>
</tr>
<tr>
<td>1928</td>
<td>SP</td>
<td>YNR001C</td>
<td>YER080W</td>
<td>924.165588</td>
</tr>
<tr>
<td>1929</td>
<td>SP</td>
<td>YNR001C</td>
<td>YMR295C</td>
<td>924.447111</td>
</tr>
<tr>
<td>1930</td>
<td>SP</td>
<td>YNR001C</td>
<td>YDL161W</td>
<td>924.502035</td>
</tr>
<tr>
<td>1931</td>
<td>SP</td>
<td>YNR001C</td>
<td>YLR423C</td>
<td>924.643327</td>
</tr>
<tr>
<td>1932</td>
<td>SP</td>
<td>YNR001C</td>
<td>YLR005W</td>
<td>924.726589</td>
</tr>
<tr>
<td>1933</td>
<td>SP</td>
<td>YNR001C</td>
<td>YLR367W</td>
<td>924.996678</td>
</tr>
</tbody>
</table>

* Subset of genes, including cell cycle-regulated genes; homolog of human PD2/hPAF1

Protein of unknown function; null mutant is defective in respiration and interacts with prohibitin (phb1); the authentic, non-tagged protein is detected in purified mitochondria in high-throughput studies

Microsomal beta-keto-reductase; contains oleate response element (ORE) sequence in the promoter region; mutants exhibit reduced VLCFA synthesis, accumulate high levels of dihydrosphingosine, phytosphingosine and medium-chain ceramides

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

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

RNA helicase in the DEAH-box family involved in the second catalytic step of splicing, exhibits ATP-dependent RNA unwinding activity

Constituent of 66S pre-ribosomal particles, required for ribosomal large subunit maturation; functionally redundant with Ssf2p; member of the Brix family

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

Protein of unknown function that associates with ribosomes; green fluorescent protein (GFP)-fusion protein localizes to the cell periphery and bud; YMR229C is not an essential gene

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

Scaffold protein responsible for phagophore assembly site organization; regulatory subunit of an autophagy-specific complex that includes Atg1p and Atg13p; stimulates Atg1p kinase activity

Component of the core form of RNA polymerase transcription factor TFIH, which has both protein kinase and DNA-dependent ATPase/helicase activities and is essential for transcription and nucleotide excision repair; interacts with Tfb4p

Protein component of the small (40S) ribosomal subunit; nearly identical to Rps22Ap and has similarity to E. coli S8 and rat S15a ribosomal proteins

Protein component of the small (40S) ribosomal subunit, nearly identical to Rps0Ap; required for maturation of 18S rRNA along with Rps8Ap; deletion of either RPS8 gene reduces growth rate,
<table>
<thead>
<tr>
<th>Gene ID</th>
<th>Strain</th>
<th>Chromosome</th>
<th>Start</th>
<th>End</th>
</tr>
</thead>
<tbody>
<tr>
<td>YDL184C</td>
<td>SP YNR001C</td>
<td>3</td>
<td>1973.0256</td>
<td>RPL41A RPL47A</td>
</tr>
<tr>
<td>YLR287C-A</td>
<td>SP YNR001C</td>
<td>3</td>
<td>1973.60828</td>
<td>RPS30A</td>
</tr>
<tr>
<td>YGR192C</td>
<td>SP YNR001C</td>
<td>3</td>
<td>1977.29566</td>
<td>TDH3 GLD1 HSP35 HSP36 SSS2</td>
</tr>
<tr>
<td>YGR285C</td>
<td>SP YNR001C</td>
<td>3</td>
<td>1982.17295</td>
<td>ZUO1</td>
</tr>
<tr>
<td>YBR118W</td>
<td>SP YNR001C</td>
<td>3</td>
<td>1993.45624</td>
<td>TEF2 EF-1 alpha</td>
</tr>
<tr>
<td>YDL081C</td>
<td>SP YNR001C</td>
<td>3</td>
<td>1998.62389</td>
<td>RPP1A RPLA1</td>
</tr>
<tr>
<td>YGL031C</td>
<td>SP YNR001C</td>
<td>3</td>
<td>1998.89097</td>
<td>RPL24A RPL30A</td>
</tr>
<tr>
<td>YGL189C</td>
<td>SP YNR001C</td>
<td>3</td>
<td>2033.57593</td>
<td>RPS26A RPS26</td>
</tr>
<tr>
<td>YKL117W</td>
<td>SP YNR001C</td>
<td>3</td>
<td>2058.09957</td>
<td>SBA1</td>
</tr>
<tr>
<td>YDR418W</td>
<td>SP YNR001C</td>
<td>3</td>
<td>2071.19489</td>
<td>RPL12B</td>
</tr>
<tr>
<td>YCR012W</td>
<td>SP YNR001C</td>
<td>3</td>
<td>2094.01123</td>
<td>PGK1</td>
</tr>
<tr>
<td>YPR080W</td>
<td>SP YNR001C</td>
<td>3</td>
<td>2130.28168</td>
<td>TEF1 EF-1 alpha</td>
</tr>
<tr>
<td>YHR174W</td>
<td>SP YNR001C</td>
<td>3</td>
<td>2146.25996</td>
<td>ENO2</td>
</tr>
</tbody>
</table>

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.

Protein component of the small (40S) ribosomal subunit; nearly identical to Rps30Bp and has similarity to rat S30 ribosomal protein.

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.

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.

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.

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.

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.
<table>
<thead>
<tr>
<th>Entry</th>
<th>Cell Type</th>
<th>Accession</th>
<th>Protein 1</th>
<th>Accession</th>
<th>Protein 2</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>3844</td>
<td>SP</td>
<td>YNR001C</td>
<td>YPR163C</td>
<td>2174.90351</td>
<td>TIF3</td>
<td>RBL3  STM1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>![](Translation initiation factor eIF-4B, has RNA annealing activity; contains an RNA recognition motif and binds to single-stranded RNA)</td>
</tr>
<tr>
<td>3845</td>
<td>SP</td>
<td>YNR001C</td>
<td>YDR382W</td>
<td>2189.96099</td>
<td>RPP2B</td>
<td>RPL45 YPA1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>![](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)</td>
</tr>
<tr>
<td>3846</td>
<td>SP</td>
<td>YNR001C</td>
<td>YDL130W</td>
<td>2197.61243</td>
<td>RPP1B</td>
<td>RPL44' RPLA3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>![](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)</td>
</tr>
<tr>
<td>3847</td>
<td>SP</td>
<td>YNR001C</td>
<td>YKL060C</td>
<td>2200.10785</td>
<td>FBA1</td>
<td>LOT1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>![](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)</td>
</tr>
<tr>
<td>3848</td>
<td>SP</td>
<td>YNR001C</td>
<td>YER131W</td>
<td>2202.6556</td>
<td>RPS26B</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>![](Protein component of the small (40S) ribosomal subunit; nearly identical to Rps26Ap and has similarity to rat S26 ribosomal protein)</td>
</tr>
<tr>
<td>3849</td>
<td>SP</td>
<td>YNR001C</td>
<td>YGL135W</td>
<td>2312.07257</td>
<td>RPL1B</td>
<td>SSM2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>![](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)</td>
</tr>
</tbody>
</table>