Parallel Identification of New Genes in *Saccharomyces cerevisiae*

Guy Oshiro,[1] Lisa M. Wodicka,[2] Michael P. Washburn,[3] John R. Yates III,[3,4] David J. Lockhart,[2,5] and Elizabeth A. Winzeler[1,4,6]

[1]*Genomics Institute of the Novartis Research Foundation, San Diego, California 92121, USA;* [2]*Aventa Biosciences Corporation, San Diego, California 92121, USA;* [3]*Torrey Mesa Research Institute, San Diego, California 92121, USA;* [4]*Department of Cell Biology, The Scripps Research Institute, San Diego, California 92121, USA;* [5] *Salk Institute for Biological Studies, Laboratory of Genetics, La Jolla, California 92037, USA.*

Short open reading frames (ORFs) occur frequently in primary genome sequence. Distinguishing bona fide small genes from the tens of thousands of short ORFs is one of the most challenging aspects of genome annotation. Direct experimental evidence is often required. Here we use a combination of expression profiling and mass spectrometry to verify the independent transcription of 138 and the translation of 50 previously nonannotated genes in the *Saccharomyces cerevisiae* genome. Through combined evidence, we propose the addition of 62 new genes to the genome and provide experimental support for the inclusion of 10 previously identified genes.

[The following individuals kindly provided reagents, samples, or unpublished information as indicated in the paper: V. Velculescu. Supplementary material is available online at http://www.genome.org.]

The complete genomic sequence of the budding yeast, *Saccharomyces cerevisiae,* was determined in 1996 (Goffeau et al. 1996). This was the first eukaryotic genome completely sequenced and served as an important test case for sequencing, annotation, and analyses of other larger genomes. Altogether, 6275 putative genes were identified in the original annotation effort (Goffeau et al. 1996). Because yeast is very AT rich and stop codons are frequently encountered, any open reading frame (ORF) predicted to encode >100 amino acids was automatically annotated as a gene. The cutoff of 100 amino acids was chosen because the likelihood of a misidentified ORF in the genome increases dramatically if shorter regions are allowed. Approximately 260,000 ORFs from 2 to 99 codons are found in the yeast genome. There are 9524 ORFs of 25 to 99 codons present in the intergenic regions (Basrai et al. 1997), or 64,085 if one considers ORFs within and overlapping the 6275 genes. Because only a minor fraction of these small ORFs are real genes, ORFs encoding proteins with <100 amino acids were omitted from the original annotation unless evidence for the gene had been found by direct experimentation. There are currently only 224 known genes (3.5% of the genome) in the yeast genome that code for proteins <100 amino acids in length (Cherry et al. 1998; Mewes et al. 1999). Many of these smaller genes encode proteins that play important roles in the yeast cell, such as mating pheromones, transporters, transcriptional regulators, and ribosomal proteins. In contrast, genes encoding small proteins in other sequenced organisms constitute up to 10% of their genomes (Basrai et al. 1997). By extrapolation, we suspect that there may be an additional 400 genes encoding small proteins lurking within the yeast genome.

Because computational methods do not reliably predict small genes and their small size makes them an elusive target for mutagenic screens, other experimental techniques are required to facilitate their identification. One method that has been used for such a purpose is the serial analysis of gene expression (SAGE) (Velculescu et al. 1997). In this technique, small 9-bp sequence tags are isolated from defined regions near the 3′ ends of different cDNAs. The 9-bp sequences are then concatenated, polymerase chain reaction (PCR) amplified, cloned, and sequenced. Estimations of the abundance of a transcript are made by sequencing and counting each SAGE tag. This technique does not rely on a priori gene predictions, and in one study of yeast ~160 cDNA tags were detected that were convincingly mapped to nonannotated open reading frames (NORFs) of 60–98 codons (Velculescu et al. 1997). This result highlights the fact that genes that encode small proteins may have been missed in the original annotation effort. As a result of the SAGE study, 27 new annotated genes were added to the *Saccharomyces* Genome Database (SGD) on the basis of the combination of their strong SAGE expression profile and homology with proteins in other organisms (Cherry et al. 1998). Data for additional NORFs were also collected, but the results were inconclusive: Either the SAGE signal was weak or the SAGE tag was deemed too close to another ORF. In this study, we searched for novel genes in the yeast genome by first using genome-wide transcriptional profiling with oligonucleotide arrays containing probes to many of the larger SAGE-identified NORFs and then by whole genome proteomic analysis (Lockhart and Winzeler 2000; Washburn et al. 2001).

RESULTS

Identification of Expressed NORFS

We designed the Affymetrix Yeast S98 Array to query 6996 ORFs, as well as 93 tRNAs, 63 small nuclear RNAs, 5 ribosomal RNAs, 418 Ty elements, and 150 intergenic regions >5 kb (gap regions) within the yeast genome selected after probes for the NORFS were picked. Probes to 6075 yeast genes recognized by

[6]Corresponding author.
E-MAIL winzeler@scripps.edu; FAX (858) 784-9860.
Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.226802.

12:1210–1220 ©2002 by Cold Spring Harbor Laboratory Press ISSN 1088-9051/01 $5.00; www.genome.org

either the *Saccharomyces* Genome Database or MIPS (Munich Information Center for Protech Sequences) as of December 1998 were included on the S98 array (Mewes et al. 1997; Cherry et al. 1998). In addition to the recognized genes, probes that specifically interrogate 921 small NORFs were also included (see Materials and Methods section for NORF and probe selection). Evidence from the aforementioned SAGE study indicated that a significant fraction of these NORFs might be transcribed and thus should be included on the array (Velculescu et al. 1997). To increase the chance of observing expression of these NORFs, we grew yeast in a variety of different growth conditions. These included treatments with hydroxyurea, nocodazole, methyl methane sulfonate (MMS), and ultraviolet (UV) light, along with a heat and cold shock. After treatment, RNA was extracted from the yeast cells, labeled, and hybridized to high-density oligonucleotide arrays using standard methods (Wodicka et al. 1997). Replicate hybridizations were conducted for each of the nine different conditions and measurements of the expression levels for each of the 6996 genes and NORFs were taken. The transcriptional response of genes that were differentially expressed is shown in Figure 1. Several major patterns are readily discernible from the global view including a massive transcriptional response triggered by DNA damage caused by exposure to UV light or MMS (cluster V), an induction of a different class of genes in response to growth in glycerol media (cluster XVI), and repression of another class of genes in the presence of the DNA-damaging agents MMS and UV light (XVIII).

Affymetrix uses an algorithm to call a gene present (expressed) or absent (not expressed) on the basis of the behavior of the probe set that interrogates each gene. Eighty-seven percent (5525) of the known genes were called "present" (expressed) by Affymetrix GeneChip software in at least two of the 18 experiments, in good agreement with previous data (Wodicka et al. 1997). Of the 5525 genes, 3802 (62%) genes were determined to be present at a level of at least one copy per cell by normalizing the average difference of each gene to genes with a known copy number in the cell (Wodicka et al. 1997). This group of "expressed genes" included 19 of the 20 SAGE-identified small ORFs that had previously been given "gene" designations in SGD or MIPS and that were included on the array (Table 1), thus indicating that hybridization data could be used to confirm SAGE data. In contrast to the annotated genes, we found very little signal for gap regions: Only 18% of the gap regions were called "present," and at more than one copy per cell in one condition; these regions may also contain transcribed NORFs.

We next asked if there was clear evidence for the expression of any NORFs included on the array. Altogether, 323 of the 921 NORFs queried on the array were called "present" by the Affymetrix GeneChip software at a level of at least one copy per cell (Avg Diff > 100) in one condition (see http://pub.gnf.org/~ewinzeler/identification_of_new_gene.htm). This fraction (35%) is lower than that found for annotated genes (62%), indicating that some proportion of the NORFs are most likely not transcribed. However, 59% of the expressed NORFS (192/323) have a codon adaptation index >0.1, indicating that these genes are likely to be transcribed at moderate to high levels within the cell (Sharp and Li 1987).

Identification of Independently Transcribed NORFs

Although genome-wide expression profiling provides direct experimental verification that genomic regions are transcribed into RNA, expression profiling does have some limitations. A potential source of false positives in our analyses is the indeterminate length of the 3′ or 5′ untranslated regions of yeast genes. Because there is no highly conserved polyadenylation signal in yeast to demarcate the 3′ end of a transcript and promoter regions are difficult to predict, it is possible that the transcripts that hybridized to NORF probes actually originated at the promoters of adjacent larger genes. To address this probability, we identified NORFs that were separated by at least 500 nucleotides (nt) from the nearest upstream or downstream gene or were located at least 150 nt from neighboring genes and showed transcriptional patterns uncorre-

Table 1. SAGE-Identified ORFs

SAGE name	Size (bp)	SAGE tag	Number of occurrences	Affymetrix or SGD designation	Present at more than one copy per cell
NORF1	198	TTCGTTCACT	94	YDR524C-A	True
NORF2	243	GCTCTCCCCC	73	RPL36B	True
NORF3	189	TGTACGCATT	16	YOL013W-A	False
NORF4	177	TTTTATTATC	15	RPL29	True
NORF5	204	CTTCTCTTTT	12	YML058w-a	True
NORF6	252	TTTCCTATAA	11	YMR122w-a	True
NORF7	192	TCTAGTCGCC	10	YLR262C-A	True
NORF8	174	ATCGTTTTAT	8	YOR298C-A	True
NORF9	267	GGCCAATGGT	8	YDR363W-A	True
NORF10	255	ACCCTGTCAT	7	YBR085C-A	True
NORF11	87	AAAAGATCAT	7	Not probed	N.D.
NORF12	279	CAGAAAATGG	6	MRS11	True
NORF13	183	TGACATTCTT	6	NPR087W	True
NORF14	141	TAGACATCTA	6	YBR126W-A	True
NORF15	216	TGCCCTGGCC	5	YER007C-A	True
NORF16	291	GGTTTTGGCG	4	YCL057C-A	True
NORF17	114	CCATACAGGT	4	Not probed	N.D.
NORF18	258	CCAAATCAAA	3	YDL130W-A	True
NORF19	399	AAGCGGTACT	3	Not probed	N.D.
NORF20	198	AACGCTTTTC	3	YBR056W-A	True
NORF21	240	GAGGATAGAG	3	YBR058C-A	True
NORF22	243	CAATGAACCG	3	RPL38B	True
NORF23	90	TCTTTATATA	3	Not probed	N.D.
NORF24	108	CGCCTCCAGT	3	Not probed	N.D.
NORF25	81	TACGTAAGTT	3	Not probed	N.D.
NORF26	93	GATTTAAACT	3	Not probed	N.D.
NORF27	222	GCGCCTCCAA	2	SOM1	True
NORF28	78	CAATGGCCCA	2	Not probed	N.D.
NORF29	264	TTGAGGAACG	2	MAK31	True
NORF30	204	GCTAAGAACC	2	YDL085C-A	True

SAGE, serial analysis of gene expression; ORF, open reading frame; SGD, *Saccharomyces* Genome Database; NORF, nonannotated open reading frame.

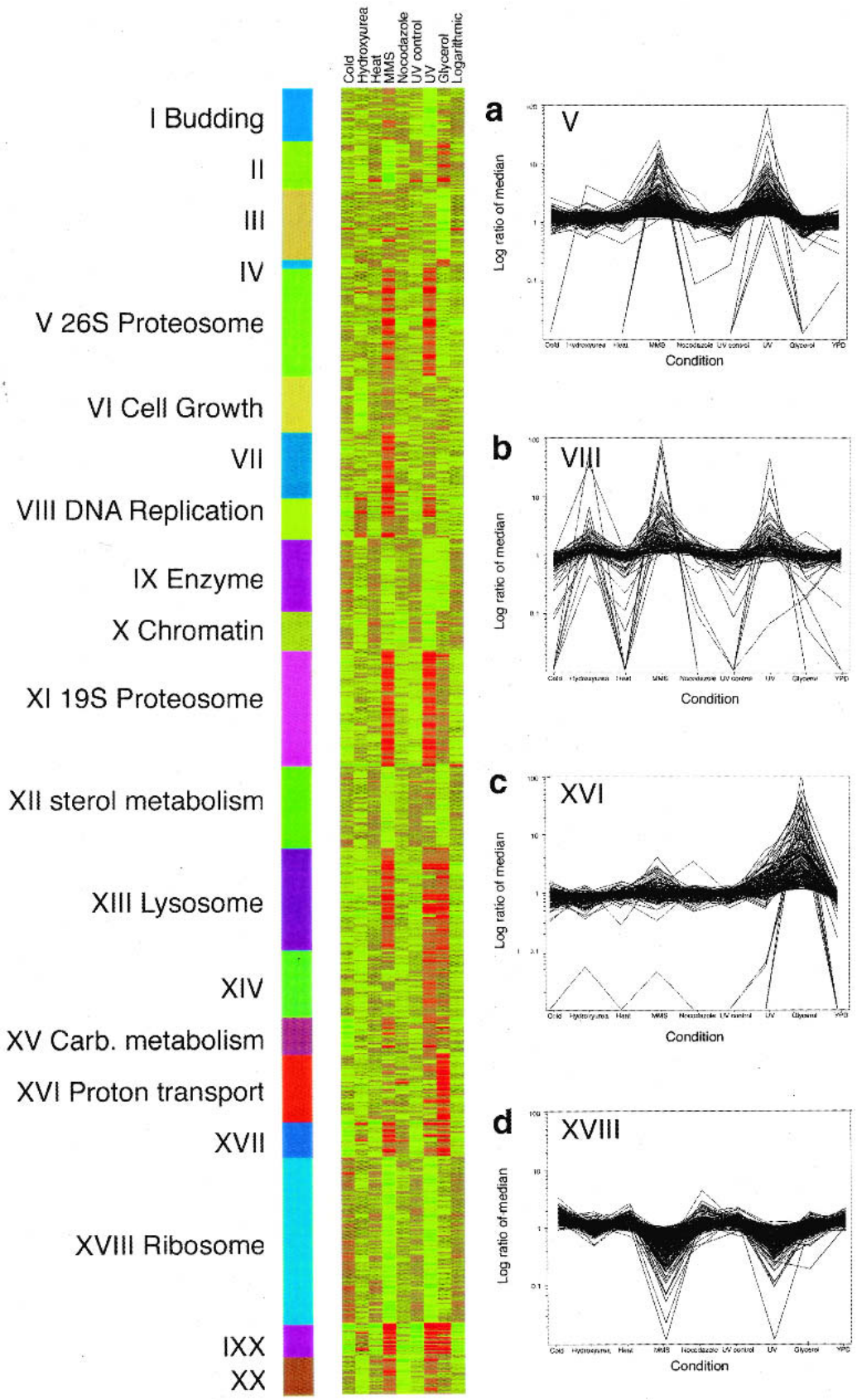

Figure 1 Transcriptional clusters identified by expression profiling over nine conditions. The data from the 18 different arrays were normalized such that the mean average difference for all genes was 200 (approximately two copies per cell). For clustering, the signals for each gene were normalized so that the median for all conditions was one. Representative clusters are shown in *a–d*, including clusters in which genes are induced after treatment with methyl methane sulfonate (MMS) and ultraviolet light (UV), induced after treatment with hydroxyurea (VIII), expressed on growth in glycerol-containing media (XVI), and repressed after treatment with MMS or UV (XVIII). For highly expressed genes, the fold change is likely to be underestimated because of the nonlinear response of the fluorescence signal at high concentrations. All data can be downloaded from http://pub.gnf.org/~ewinzeler/identification_of_new_gene.htm.

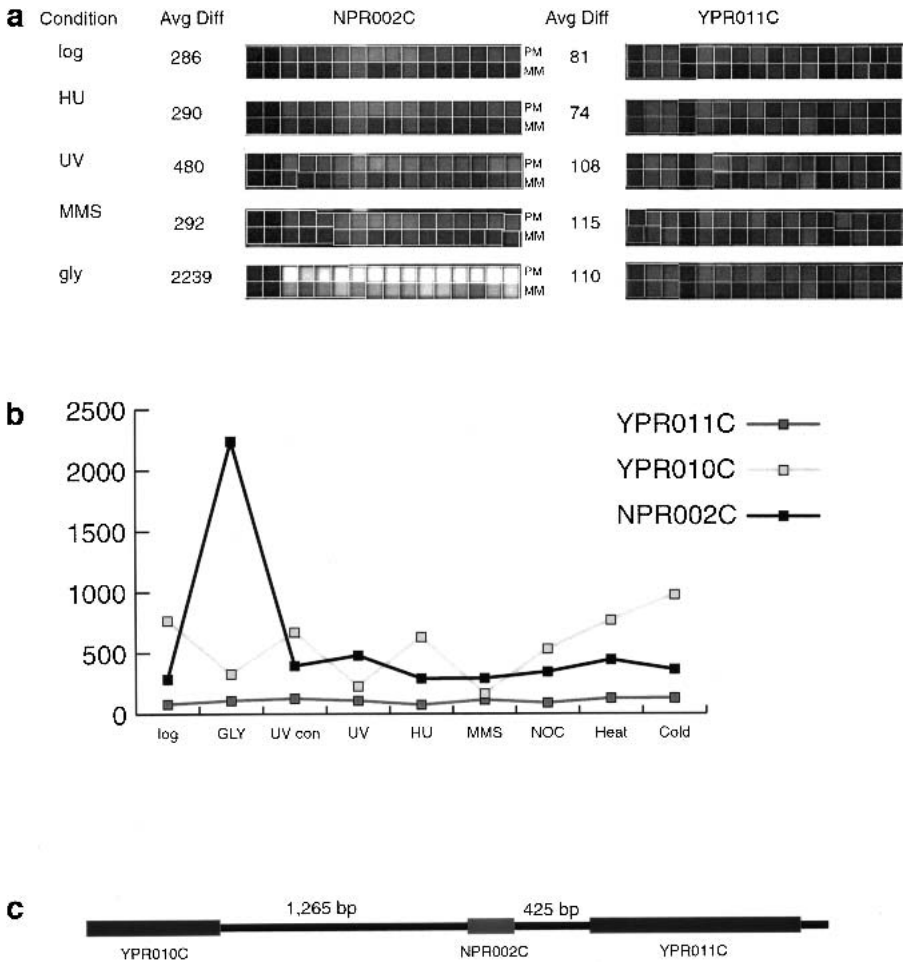

Figure 2 Transcriptional profile of the nonannotated open reading frame (NORF) *NPR002C* and the flanking neighboring genes *YPR010C* and *YPR011C*. (*a*) Array hybridization images. Each open reading frame (ORF) and NORF is represented on the S98 array by 16 oligonucleotide pairs. One member of each pair corresponds to a perfectly matched sequence from the ORF (PM); the other pair member contains a single-base mismatch in a central position (MM). The difference in intensity between the perfectly matched and the mismatched sequences (PM-MM) is used to calculate an "average difference intensity" for each ORF in each experiment. Array probe hybridization images for NORF *NPR002C* and ORF *YPR011C* from control cells in logarithmic phase growth, cells treated with HU, UV, MMS, and cells grown in glycerol containing media-treated cells are shown along with the average difference (Avg Diff) intensity values. (*b*) The average difference intensity of each gene graphed across all the conditions tested in this study. (*c*) Chromosomal view of *NPR002C*, *YPR011C*, and *YPR010C* with the distance in nucleotides between the NORF and ORF printed above the gap regions. The correlation of expression profiles between *NPR002C* and the upstream gene *YPR011C* and the downstream gene *YPR010C* is 0.13 and −0.32, respectively.

lated with those of neighboring genes ($r < .6$). We found 138 NORFs that satisfied these criteria. The entire list is available in Supplemental Table 1 available online at http://www.genome.org. The correlation and distance criteria are conservative and could result in a number of false negatives because coregulated genes are often juxtaposed in the genome (Cohen et al. 2000) and untranslated regions >150 nt are rare in yeast (Olivas et al. 1997). An example of one of the NORFs that meet the strict criteria is shown in Figure 2. *NPR002C* is expressed under all conditions and is significantly induced on growth in glycerol-containing media (Fig. 2). The physically adjacent genes *YPR011C* and *YPR010C* are not expressed in the same way as *NPR002C*, showing no up-regulation on growth in glycerol. Northern blot analysis of *NPR002C* and *YPR011C* confirms the differential expression patterns observed in the GeneChip analysis (Fig. 3). Furthermore, the size of the transcripts on the Northern blots shows that the *NPR002C* mRNA is not simply an extension of the mRNA of neighboring genes.

Functional Assignment of Independently Transcribed NORFs

The expression pattern for a gene can provide clues to its function. In fact, in cases such as yeast in which gene redundancy is common, transcriptional profiling may be more informative than mutagenesis. We used this "guilt by association" method to assign a function to some of the 138 NORFs that were transcribed in a manner independent of adjacent genes. Of the 138 NORFS, 120 were identified as being differentially expressed using a nonparametric Kruskal-Wallis test over the nine different growth conditions. The expression results for the 120 NORFs were combined with the data for the 3392 expressed genes that were determined to be differentially expressed using the same nonparametric Kruskal-Wallis test ($P < .05$). The entire group was subjected to K-means clustering based on the Pearson correlation coefficient. For the 20 clusters, we found significant overlap with 11 MIPS functional categories (Mewes et al. 1997), including proteosome function (V, induction after treatment with MMS or UV light), ribonucleotide reductase function (VIII, induced in hydroxyurea), and ribosome function (XVIII repressed in MMS and UV).

Some of the functional classifications were not surprising. For example, yeast prefer fermentation to cellular respiration to generate ATP. Growth in media with a nonfermentable carbon source, such as glycerol, forces a switch to oxidative respiration. In the cluster containing genes induced after treatment with glycerol, we found 13 of the 16 genes known to have roles in proton transport (cluster XVI $P = 7.9 \times 10^{-13}$) and 11 of the 21 genes with known roles in TCA intermediate metabolism ($P = 1.5 \times 10^{-10}$).

We also confirmed that a major transcriptional response to DNA damaging agents is the up-regulation of genes involved in protein degradation (Jelinsky and Samson 1999, Jelinsky et al. 2000) and a down-regulation of genes involved in protein synthesis (Fig. 1b). In fact, 29 of the 35 genes known to play a role in the function of the 26S proteosome were found in clusters V or XI, two similar clusters showing

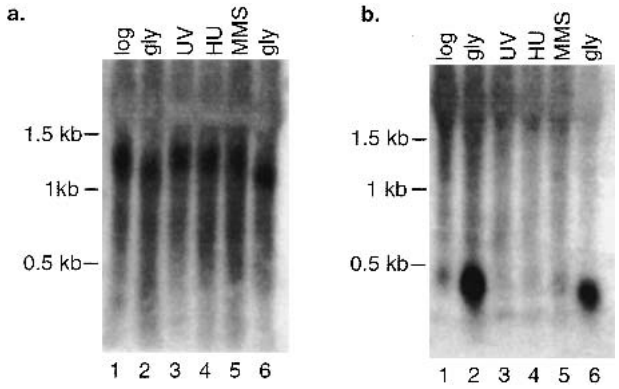

Figure 3 Northern blot analysis of *NPR002C* and *YPR011C*. (*a*) Expression of *YPR011C* across various conditions. RNA was extracted and total yeast RNA was separated by electrophoresis in an agarose gel, blotted, and hybridized with a polymerase chain reaction (PCR) amplicon of *YPR011C*. (*b*) The same blot was then stripped and hybridized with a PCR amplicon of *NPR002C*.

the most overlap with genes having a role in the function of the 26S or 19S proteosome, respectively. On the other hand, 102 of the 123 genes encoding proteins comprising the cytosolic ribosome were found in cluster XVIII ($P = .0$).

Hydroxyurea is known to interfere with the activity of ribonucleotide reductase (RNR) (Rittberg and Wright 1989). We expected, and indeed found, that all four members of the RNR gene family were induced by hydroxyurea and located in the same cluster of 108 genes (VIII) ($P = 7.6 \times 10^{-4}$). The RNR genes were also induced in response to MMS and UV light, although not as strongly as in hydroxyurea. This is probably because the cell needs extra dexoynucleoside triphosphates (dNTPs) for DNA replication and repair processes (Elledge et al. 1993; Huang and Elledge 1997). Another gene that shows a profile similar to the RNR genes is YML058W-A/HUG1 (hydroxyurea and UV and gamma radiation induced), which was originally identified in the aforementioned SAGE study and originally named NORF5 (Velculescu et al. 1997). HUG1 is known to interact with genes in the MEC1 DNA damage checkpoint (Basrai et al. 1999). In addition, in support of the functional assignments we found that the clusters could be used to identify transcription factor binding sites relevant to a particular cluster by searching for sequences that are overrepresented in regions upstream of genes in a transcriptional cluster (Table 2) (Cho et al. 1998; Hughes et al. 2000).

Seventy-six of the NORFs were found in transcriptional clusters that had a significant overlap with MIPS functional categories (see Supplemental Table 1 available online at http://www.genome.org). For example, *NOL015W* and *NPR002C* may be involved in respiration because they are found in a cluster that includes many other genes that are required for energy metabolism and that are significantly induced in cells grown with glycerol as their sole carbon source. Although *NOL015W* was unclassified when the array was designed, it has since been shown by biochemical purification of the F_1F_0-ATP synthase complex (Arnold et al. 1998) to encode a subunit of the ATP synthase (ATP19), consistent with the functional assignment on the basis of its expression behavior. The list of 138 NORFs that shows evidence of independent transcription as well as codon adaptation indices, expression levels, and potential cellular roles is in Supplemental Table 1 (available online at http://www.genome.org).

Computational Evidence of Gene Conservation

Evidence of independent transcription does not necessarily indicate that a NORF is a real gene: The transcript may not be translated into a protein, and there may be multiple small ORFs in regions that are transcriptionally active. Therefore a computational approach was used to provide further evidence that the NORFs detected by transcriptional profiling encoded real genes. Homology searches were conducted against the nonredundant protein databases to determine whether any of the transcribed NORFs encoded proteins that appear to have been conserved across multiple species. All 323 NORFs were searched against the National Center for Biotechnology Information (NCBI) nonredundant protein database with a

Table 2. Regulatory Elements Identified in Expression Clusters

Sequence	Frequency	K-means cluster	Potential function of genes with sequence elements	*P* value	Potential binding factor	Consensus	Refs
GTGGCAAA	27/281	V	Ubiquitin-dependent protein degradation	1.2×10^{-11}	RPN4	GGTGGCAA	(Jelinsky et al. 2000; Mannhaupt et al. 1999)
AAAATTTT	173/435	XVIII	Nucleolus/transcription from pol I promoter	5.9×10^{-38}	Unknown	Unknown	
GCGATGAG	47/435	XVIII	Nucleolus/transcription from pol I promoter	3.6×10^{-23}	Unknown	Unknown	
TCCGTACA	28/435	XVIII	Cytosolic ribosome	9.9×10^{-18}	RAP1	Unknown	(Kurtz and Shore 1991; Moehle and Hinnebusch 1991)
CCAATCA	23/170	XVI	Hydrogen/energy transport	1.2×10^{-9}	HAP2	CCAAT	(Ozsarac et al. 1997)
VAAAGGG	30/303	XI	Unknown	1.1×10^{-11}	Unknown	Unknown	

The sequence GTGGCAAA was overrepresented upstream of genes in the cluster containing genes with a potential functional role in the 26S proteosome. This sequence is the consensus-binding site for Rpn4p, a key regulator of proteosome function, and it is found in a number of genes involved in protein degradation (Mannhaupt et al. 1999). The sequence CCAATCA was overrepresented upstream of genes in the cluster of genes, along with hydrogen-transporting ATP synthase genes that have a putative mitochondrial function. This sequence contains the consensus HAP2 binding-site CCAAT (Ozsarac et al. 1997).
ATP, adenosine triphosphate.

Table 3. NORFs with Homologous Sequences

NORF ID	Proposed ORF designation	Chromosomal location	Size (aa)	CAI	Transcripts per cell	*P* value	Accession no.
NCL001W	YCL002W-A	Chr III: 110356-110526	56	0.09	0.7	1.1×10^{-17}	gi|*2131206*
NCR009C	YCR024C-B	Chr III: 162857-162591	88	0.15	29.4	1.7×10^{-58}	gi|*283181*
NHR002C	YHR007C-A	Chr VIII: 122758-122543	71	0.12	N.D.	4.1×10^{-43}	gi|*4517*
NHR014W	YHR050W-A	Chr VIII: 209467-209637	56	0.12	0.9	2.9×10^{-34}	gi|*3575*
NHR017W	YHR132W-B	Chr VIII: 370055-370343	95	0.10	0.8	1.0×10^{-56}	gi|*6862571*
NIL001W	YIL002W-A	Chr IX: 350298-350507	69	0.17	1.5	2.3×10^{-2}	gi|*3766232*
NIL005W	YIL046W-A	Chr IX: 268308-268472	54	0.10	0.9	1.5×10^{-8}	gi|*2860888*
NKR013C	YKR099C-A	Chr XI: 638722-638531	63	0.14	0.8	7.0×10^{-6}	gi|*6322033*
NLR036W	YLR154W-B	Chr XII: 455884-456024	46	0.13	1.5	2.1×10^{-5}	gi|*4669161*
NMR046W	YMR175W-A	Chr XIII: 611313-611507	64	0.13	N.D.	1.3×10^{-37}	gi|*7493896*
NNL005C	YNL024C-A	Chr XIV: 586816-586598	72	0.12	3.8	8.3×10^{-22}	gi|*12833197*
NOR009W	YOR072W-A	Chr XV: 464470-464631	53	0.10	N.D.	5.8×0^{-11}	gi|*7493881*
NOR043C	YOR192C-C	Chr XV: 704223-703987	78	0.16	0.6	7.2×10^{-27}	gi|*6322129*
NPR087W	YPR170W-A	Chr XVI: 883373-883558	61	0.11	5.6	5.3×10^{-36}	gi|*7493886*

A Smith-Waterman search was performed for each NORF against the National Center for Biotechnology Information (NCBI) nonredundant protein database (NR) (October 2001) with a BLOSUM62 weight matrix, a gap opening penalty of 12, and gap extension penalty of 2. CAI, codon adaptation index.

Smith-Waterman algorithm. Alignments of 14 NORFs with a *P* value <.05 are listed in Table 3. An example of a NORF that shows strong sequence conservation throughout evolution is *NNL005C*. We found that *NNL005C* shares significant homology with a gene found in both mouse and *Drosophila* (Fig. 4). The sequences share 63% identity and 78% similarity in amino-acid sequence across the entire length of the coding sequences. The conservation in sequence indicates functional constraints on the sequences.

Others have applied this comparative genomics approach more systematically, first by collecting random shotgun sequences from related yeast species and second by comparing the predicted proteomes of these species with that of *S. cerevisiae*. Blandin et al. identified 50 NORFs conserved between *S. cerevisiae* and hemiascomyces (Blandin et al. 2000). Of these 50, 13 (11 as NORFs and two as genes) were probed on the yeast expression array and eight (seven NORFs, one gene) were found to be actively transcribed. Cliften et al. identified 11 NORFs conserved within the *Saccharomyces* genus (Cliften et al. 2001). Of these 11, two were probed on the array and transcripts were detected for one. This fraction is similar to that observed for annotated genes (62%). Although the sample size is small for generalizations, these results confirm the value of the comparative genomics approach. One other gene, *YLR363W-A*, was identified by mass spectrometry as described below. These ORFs that are transcribed and that encode proteins that are homologous with proteins from other species are listed in Tables 3 and 4 and have been given a gene designation according to *S. cerevisiae* systematic nomenclature.

Evidence of Translation

Because a priori gene predictions are not needed, mass spectrometry represents an attractive alternative technology for the unbiased detection of new translated ORFs. We used tandem mass spectrometry coupled with in-line two-dimensional liquid chromatography, multidimensional protein identification technology (MudPIT) (Washburn et al. 2001), to characterize the proteins present in logarithmically growing yeast cultures. First, total yeast protein was extracted and proteolyzed into peptides with the proteases endoproteinase Lys-C and trypsin. The resulting complex peptide mixture was then applied to a biphasic microcapillary column packed with strong cation exchange and reverse-phase material from which peptides were sequentially eluted directly into a tandem mass spectrometer (Washburn et al. 2001). This MudPIT dataset was searched with the SEQUEST algorithm (Eng et al. 1994) against a concatenated database that contained the annotated ORF sequences (yeast_orfs.fasta database from the NCBI), as well as all ORFs from 25–99 amino acids (64,085), including 1187 NORFS identified by the SAGE

```
mouse        1  --MSAIFNFQSLLTVILLLICTCAYIRSLAPSILDRNK-TGLLGIFWKCA
Drosophila   1  MLPSALFNFHSLLSVILLLICTCAYLRSLFPSLIDRNK-TGFMGTFWKLA
NNL005C      1  --MSALFNFRSLLQVILLLICSCSYVHGQWPSLLDRYKNHEVLGAFWKMA
consensus    1    .**.***.***.*******.*.*.....**..**.*  ....*  ***.*

mouse       48  RIGERKSPYVAICCIVMAFSILFIQ
Drosophila  50  RIGERKSPWVGAACLIMAFTVLFWS
NNL005C     49  RVGERASPYVSLACILMAISQFNS
consensus   51  *.***.**.*...*..**.....
```

Figure 4 Homologs of NORF *NNL005C* are found in other species. CLUSTAWL alignment of homologous protein sequences from the mouse RIKEN cDNA 0610041E09 gene, *Drosophila CG14199* gene, and the yeast NORF *NNL005C*. The mouse sequence scores ($P < 8.3 \times 10^{-22}$) and the *Drosophila* sequence scores ($P < 2.0 \times 10^{-20}$).

Table 4. NORFs Found in Other Studies

NORF ID	Proposed ORF	Chromosomal location	Size (AA)	CAI	Upstream gene distance	Upstream gene correlation	Function
NBL011C*	YBL029C-A	Chr II: 164734-164450	94	0.125	2746	−0.46	N/A
NDR019C*	YDR079C-A	Chr IV: 603805-603587	72	0.119	3448	−0.43	N/A
NDR156C*[c]	YDR379C-A	Chr IV: 1233506-1233267	79	0.161	6398	0.52	N/A
NGR072W*	YGR161W-B	Chr VII: 810222-810500	92	0.087	2543	−0.48	N/A
NJL008W*	YJL062W-A	Chr X: 316419-316676	85	0.106	1153	0.09	N/A
NJL020C[a]	YJL133C-A	Chr X: 159545-159321	74	0.235	4431	−0.14	c
NLR022W*	YLR099W-A	Chr XII: 341326-341589	87	0.072	5292	0.59	h
NOL015W[b]	YOL077W-A	Chr XV: 185437-185643	68	0.2	226	0.45	e
NOL017W*	YOL086W-A	Chr XV: 159172-159444	90	0.154	8897	0.28	f

The nucleotide distance between each NORF and its nearest upstream gene was calculated. The pairwise correlation coefficient of expression of each NORF with its nearest upstream gene was computed. The possible functional classification of each NORF on the basis of the expression profile is also listed. NORFs with an asterisk (*) are conserved in other hemiascomycetes yeast species (Blandin et al. 2000).
[a]NJL020C is conserved in *Saccharomyces kluyveri* (Cliften et al. 2001).
[b]YOL077W-A was discovered by the biochemical purification of the F_1F_0-ATP synthase complex (Arnold et al. 1998).
[c]NORFs are detected by mass spectrometric analysis.
Potential functional classifications: a: 26S proteosome, b: chromatin, c: enzyme, d: glutamate metabolism, e: hydrogen transporting, f: mitochondrion, g: nucleolus/transcription, h: organelle organization, i: ribonucleoside diphosphate, j: ribosome, and k: sterol metabolism.

study (Velculescu et al. 1997). The overall results of our MudPIT analyses were comparable to those previously published (Washburn et al. 2001) in which approximately one fourth of the predicted, annotated proteins in the yeast genome were detected and identified in a highly automated fashion (data not shown). The protein products of 22 SAGE NORFs were also detected, and 11 of these were in the set of 323 detectable transcripts (Table 5). An example of a mass spectra matching

Table 5. NORFs Identified by MudPIT Proteomic Analyses

NORF ID	Proposed ORF designation	Chromosomal location	Size (AA)	CAI[a]	Peptide identified	Transcripts per cell[b]
NAL010C	YAL063C-A	Chr I: 22400-22688	96	0.17	R.YRNKEKGKFISLCK.N	1.8
NBR028W[c]	YBR126C-A	Chr II: 490808-491014	68	0.16	R.LHQLDGIPHA.-[e,f]	N.D.
NCR024W	YCR095W-A	Chr III: 289632-289790	52	0.08	H.TKVNKKSSMHAFLCKIYK.R	0.7
NDR129W	YDR320W-B	Chr IV: 1108476-1108613	45	0.13	L.NSLLFLPICFCLLQLKATCAVR.V	2.0
NDR156C	YDR379C-A	Chr IV: 1233506-1233267	79	0.16	K. DFTTIEHLLRVGNK.K K.ENQVNFVNYIHEEFGK.Y	1.4
NGR097C	YGR169C-C	Chr VII: 836660-836382	92	0.13	K.ERDALLTAEEEQLGK.G K.ERDALLTAEEEQLGKGGK.G K.QRAQMEQLEAEEASK.W X.QRAQMEQLEAEEASKWEQGSRK.E	1.0
NHL007C	YHL048C-A	Chr VIII: 5796-5662	44	0.08	G.RARMGGLIVKHRFN.H	1.2
NHR007W	YHR032W-A	Chr VIII: 175186-175365	59	0.06	G.NFKGFAMWHATGKH.H	0.7
NIL001W	YIL002W-A	Chr IX: 350298-350507	69	0.17	K.DILDVLNLLK.G	1.5
NIR003C	YIR018C-A	Chr IX: 385698-385561	45	0.10	K.RYLEIMSTASAQ.F	1.0
NIR008W	YIR021W-A	Chr IX: 398511-398723	70	0.11	K.SDFKKHSKE.I	0.7
NLR127C	YLR361C-A	Chr XII: 849678-849382	98	0.12	R.TGGHRPQISDEEVSK.R	0.8
NMR066W	YMR247W-A	Chr XIII: 769282-769425	47	0.10	S.AKLLSGIMALLFNGKSLLRP.I	0.5
NNL014W	YNL042W-B	Chr XIV: 547109-547366	85	0.11	V.RVATYICQKNESR.F	0.5
NNL029W	YN067W-B	Chr XIV: 499414-499554	46	0.11	L.MWCTGVVSKTALLTGNFFFS.S	0.4
NNL042C	YNL146C-A	Chr XIV: 351577-351383	64	0.13	S.AYYVSQVLRICKEMPYR.D	0.2
NNL058W	YNL277W-A	Chr XIV: 116677-116865	62	0.05	M.CHILPPLR.S	−0.2
NOL015W[d]	YOL077W-A	Chr XV: 185437-185643	68	0.20	L.GLLGLLVVPNPFK.S	2.2
NOL020W	YOL097W-A	Chr XV: 136219-136404	61	0.12	Q.SMICSSEHENLTCK.Y	0.2
NOL049W	YOL155W-A	Chr XV: 27083-27217	44	0.08	G.SFNKCVTGYSCRMAIHYY.V	0.0
NOR002C	YOR034C-A	Chr XV: 397667-397425	80	0.13	R.IWVREKGRKCSFFFF.S	0.8
NPL013C	YPL119C-A	Chr XVI: 324286-324023	87	0.11	R.NIFEIGGLLLQ.S	0.5

[a]CAI values were calculated according to Sharp and Li (1987).
[b]The transcriptional expression level of each NORF in approximate copies of transcripts per cell in a log phase cell.
[c]Probes to NBR028W were not selected for inclusion on the S98 array.
[d]YOL077W-A was discovered by the biochemical purification of the F_1F_0-ATP synthase complex (Arnold et al. 1998).
[e]The '-' indicates the C-terminus of the protein.
[f]The '.' after an amino acid indicates the cleavage sites of the peptide. The sequence between the periods in each cell indicates the actual peptide identified by tandem mass spectrometry. (ND) there was no detectable expression in a log phase culture.
MudPIT, multidimensional protein identification technology.

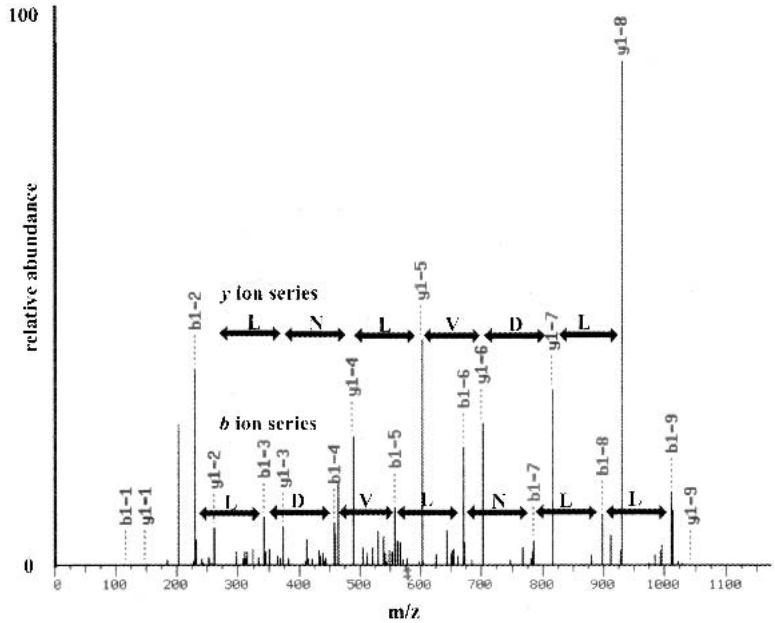

Figure 5 Mass spectra for a peptide from the NORF *NIL001W*. A multidimensional protein identification technology (MudPIT) analysis of the soluble proteome of *BJ5460* was performed and the results analyzed via SEQUEST (Eng et al. 1994) using a concatenated database containing ORFs and NORFs. In the MudPIT analyses, a collision-induced dissociation tandem mass spectrum for (M + 2H) ^{2}+ ion of the peptide DILDVLNLLK at *m/z* 578.5 from the NORF *NIL001W* was detected and identified. An eight-ion *b* and seven-ion *y* series are shown in red and blue, respectively, and the corresponding amino acid difference between each ion is shown. The SEQUEST result for the tandem mass spectrum shown had an Xcorr of 3.1276 and a ΔCn of 0.2292, indicating complete confidence in the SEQUEST result.

a NORF is shown in Figure 5. An intense string of seven ions from the *y* ion series and a less intense string of eight ions from the *b* ion series resulted in an excellent SEQUEST match and complete confidence in the identification (Fig. 5). The additional peptides described in Table 3 yielded comparable SEQUEST results indicating the probable translation of each of the NORFS listed. In the search of the 62,898 remaining 25–99 amino-acid proteome, spectra mapping to 28 small (<100 amino acids) NORFs were identified. Twenty-one of these small ORFs were within annotated genes, and eight were intergenic (Table 6), one of which had been identified in a previous study (Blandin et al. 2000). The fact that 11 proteins were identified in the 323 expressed ORFs (3%), 11 additional proteins in the set of (864) SAGE NORFs (which were either not expressed or not included on the array [1.2%]), and only 28 additional hits in a search of the entire 25–99 amino acid potential proteome (.04%) indicates a tremendous enrichment of proteins detected by mass spectrometry within the set of 323 "expressed" NORFs. NORFs detected by mass spectometry are listed in Tables 5 and 6 and have been given gene designations according to the systematic nomenclature.

DISCUSSION

Array-based expression profiling allows a greater proportion of the genome to be queried than does mass spectrometry at present. It also allows more conditions to be rapidly examined than does SAGE. Also, in contrast to the use of computational searches, genome-wide expression profiling provides direct experimental verification that genomic regions are transcribed into RNA. Although expression patterns are not conclusive evidence of gene function because multiple functional categories may be represented in a particular cluster, they indicate directions for future research. Although expression profiling is a powerful strategy, there are certain inherent limitations. First, the oligonucleotide probes to putative genes on the S98 yeast array were chosen on the basis of experimental and computational data indicating that they might be expressed. A more unbiased search could be performed using arrays that cover the entire genome or all possible intergenic regions. Such "tiling" arrays have been successfully used to identify new genes and further define exon boundaries in humans (Shoemaker et al. 2001; Kapranov et al. 2002). At present, >500,000 probes can be placed on a typical Affymetrix array, allowing a more unbiased exploration of the genome. Another limitation of expression profiling is that it provides only a snapshot of the transcriptome at a specific time in response to specific stimuli. Many genes may be expressed only under very specific conditions and not under the range of conditions sampled here.

It is also possible that some of the RNAs detected are not translated into protein products. For example, we found that the signal for the 35S ribosomal RNA was often significant (average difference values up to 1500), indicating that even with oligo-dT priming for the construction of cDNA, some untranslated RNAs were amplified and labeled. It is unlikely that signals detected for many of the small ORFs are caused by contamination with genomic DNA, consistent with the observation that the average difference signal values for regions of the genome (so-called "gap regions"), which are not predicted to encode a gene, are generally quite low. In theory, whether an RNA was polyadenylated and thus likely to be translated could be confirmed by looking for further increases in signal at particular probe sets after direct labeling of RNA in the absence of poly(A)$^{+}$ purification (Wodicka et al. 1997). Alternative technologies, such as random transposon mutagenesis, may also be valuable ways to distinguish translated from nontranslated RNAs (Ross-Macdonald et al. 1999; Kumar et al. 2002). Finally, there may be multiple small ORFs within a single transcribed intergenic region and determining which one encodes the protein is difficult.

Proteomic analysis provides the most direct way to distinguish translated from nontranslated RNAs, although current proteomic methodologies have limitations. Although MudPIT has been shown to be the proteomic technology with the most comprehensive ability to detect and identify proteins with a broad range of isoelectric point (pI), hydrophobicity, size, and abundance (Washburn et al. 2001), MudPIT is not a completely comprehensive method. A recent characterization of the dynamic range of MudPIT indicated that it is capable of detecting and identifying a protein at 100 copies per cell in the background of proteins at 1,000,000 copies per cell (Wolters et al. 2001a). Although it is likely that MudPIT detects and identifies some proteins at <100 copies per cell, the number of proteins detected and identified at <100 copies per cell will likely be few. As with most techniques, low abundance proteins are harder to detect simply because of their scarcity.

Table 6. Translated ORFs Identified in an Unbiased Search of Yeast Proteome

Proposed ORF designation	Location	Peptide identified	CAI	Chromosome	Size (A.A.)
YBR221W-A	intergenic	K.RISLGMINTVVSLDR.-	0.103	Chr II: 666497-666598	23
YBR196C-A	intergenic	V.VLSKEKILLKKAYYAK.T	0.087	Chr II: 614589-614488	34
YBR121C-A	within YBR121C different frame	F.KKLVLLNQLSRQLVKQ.L	0.116	Chr II: 482443-482288	52
YBL039C-A	within YBL039C different frame	N.RWLTFTMLILLIT.S	0.103	Chr II: 144994-144914	27
YDR003W-A	intergenic (3′ of YDR003W)	M.TCGIENSYKSAEK.K	0.131	Chr IV: 454778-454897	40
YDR118W-A	within YDR118W/APC4 different frame	K.RIPSVSKR.K	0.106	Chr IV: 687761-687874	38
YDR371C-A	opposite	-.MGSMILDITGNSM.S	0.073	Chr IV: 1219602-1219501	34
Multiple locations	intergenic	V.DFYSNIINKNLR.L	0.104	Chr V: 443764-443633	44
YER090C-A	opposite	F.LFLARNNEHSHKK.Y	0.17	Chr V: 338407-338321	29
YFR009W-A	within YFR009W different frame	T.KWFTESTCKSLLTNTD.T	0.095	Chr VI: 163868-164122	85
YFR010W-A	opposite of YFR011c	L.FVTIQWLALIGQKTLSQ.F	0.116	Chr VI: 166720-166905	62
YGL041W-A	opposite of YGL042C	K.KLVNLDGTSANEENTMKPWQMK.I K.SGIQLGPEQLAPLMTVLGLEK.K	0.109	Chr VII: 419038-419283	82
YGR035W-A	intergenic	P.EAPLIIRGK.G	0.095	Chr VII: 557559-557777	73
YGL210W-A	intergenic	K.STAHTQSSGSPIKR.S	0.121	Chr VIII: 93078-93305	76
YGL014C-A	opposite of YGL014W	R.RRAISELRILR.N	0.109	Chr VII: 466394-466236	53
YHR073C-A	opposite of YHR073W	K.YLGSTSCPLLR.J	0.109	Chr VIII: 245503-245426	26
YHL015W-A	intergenic	L.REPLYLANLKIKVHIYRMKR.-	0.253	Chr VIII: 74695-74775	27
YHR073W-A	within YHR073W different frame	G.KRDHILHCPLAAY.S	0.065	Chr VIII: 242869-243042	58
YJL197C-A	opposite of YJL197W	K.KDLSLSVTLIDVYC.S	0.08	Chr X: 66085-65807	93
YKL145W-A	within YKL145W different frame	-.MGHLVLVR.H	0.036	Chr XI: 174960-175049	30
YKL100W-A	opposite of YKL100C	L.PDVIFVAHR.N	0.109	Chr XI: 253802-253888	29
YLR163W-A	opposite of YLR163C	Y.SLSLSIALLSKTDLVK.I	0.065	Chr XII: 492814-492924	37
YLR363W-A[a]	YLR262W-A	K.SSSLTETTERLVASK.V	0.281	Chr XII: 853459-853713	85
YLR364C-A	opposite of YLR366W	I.RVFIGSLPMLDLKNR.V	0.086	Chr XII: 855643-855524	40
YMR013C-A	YMR013C/SEC59 different frame	R.GPLLPLIYLINK.S	0.085	Chr XIII: 296619-296473	49
YOR293C-A	intergenic	L.LFLNHVVR.R	0.069	Chr XV: 868145-867996	50
YOL083C-A	opposite of YOL083W	R.VILITHLNV.M	0.144	Chr XV: 16660-166463	46
YPR160W-A	within GPHI/YPR160W different frame	S.MVSLLKRLLTVTRWK.L	0.134	Chr XVI: 861929-862006	26

[a]Identified by homology (Blandin et al. 2000).

It is important that searches for small genes with small NORFs be attempted for any genome for which there is sequence available, and other methods have been proposed, including random transposon mutagenesis (Kumar et al. 2002). This is because as the volume of sequence data grows, primary data are seldom considered and researchers become dependent on databases and catalogues that process, sort, and serve the sequence data. Because the index for many of these databases is the annotated gene, a NORF is effectively lost from consideration in many queries. There may be important signaling molecules, drug targets, or tumor suppressors in this collection of nonannotated genes. The comprehensive identification of all the transcribed RNAs and proteins in a genome will be a difficult task and is likely to be accomplished incrementally, especially as no method is perfectly suited to the task. In this work, we have shown the feasibility of using both expression profiling as well as mass spectrometry for the identification of new genes.

MATERIALS AND METHODS

Selection of Yeast NORFs to Include on the S98 Yeast Chip

The genome sequence and annotations were downloaded in November of 1998 (Mewes et al. 1997; Cherry et al. 1998). Approximately 1458 potential NORFs (>43 amino acids) were identified in the initial SAGE study (Velculescu et al. 1997). In 1187 cases, the SAGE tag mapped to a single region of the genome. Oligonucleotide probes for 1187 NORFs were selected and then subjected to a computational screen that favored a subset of sequences with similar GC content and thermodynamic properties and eliminated probes with possible secondary structure or sequence similarity to other probes. Probes specific to this subset of 921 potential NORFs were then synthesized on the S98 array by a process of photolithography and combinatorial chemistry following standard Affymetrix protocols (Pease et al. 1994).

Strains, Media, and Growth Conditions

S. cerevisiae strain *BY4741* (MAT *a* his3Δ1 leu2Δ0 met15Δ0 ura3Δ0) was used in this study. To limit the variables in expression profiling, a single large logarithmically growing culture (*BY4741*) was split into nine subcultures. Logarithmically growing cells were obtained by growing yeast cells to early log phase (3×10^6 cells/mL) in yeast extract-peptone-dextrose-(YPD) rich medium at 30°C. For arrest in the S phase of the cell cycle, hydroxyurea (0.1 M) was added to early log phase cells, and the culture was incubated at 30°C for an additional 3.5 h. For arrest in the G2/M phase of the cell cycle, nocodazole (15 µg/mL) was added to early log phase cells, and the culture was incubated at 30°C for an additional 100 min. For cold shock and heat shock, yeast cells were shifted to either 37°C or 15°C for 20 min. For MMS exposure, MMS (0.1%) was added to early log phase cells, and the culture was incubated at 30°C for an additional hour. For exposure to UV irradiation, cells were spread on the surface of YPD plates, irradiated (Stratagene; UV Stratalinker 2400) at 60 J/m^2, and then incubated for an additional hour before harvesting the cells from the plates (Kiser and Weinert 1996; Basrai et al. 1999). To control for the additional handling steps, an additional con-

trol was performed: Control cells were subjected to the same collection procedure without the UV exposure. For growth in a nonfermentable carbon source, an early log phase culture was resuspended in YP + 3% glycerol and incubated at 30°C for seven generations. Harvested cells were washed once with water before freezing at −70°C. The growth state and cell-cycle stage of the harvested cells were confirmed by microscopic analyses.

Yeast Expression Profiling

Total yeast RNA was isolated by using a hot phenol extraction method (Wodicka et al. 1997). All array hybridizations were performed in duplicate as previously described (Wodicka et al. 1997). Hybridizations were performed at 45°C for 16 hr. Microarray analysis was performed essentially as previously described. Briefly, 5 µg total RNA was converted to cDNA and used as a template to generate biotinylated cRNA. cRNA was fragmented and hybridized to Affymetrix S98 Yeast arrays as described in the standard protocol outlined in the `GeneChip` Expression Analysis Technical Manual (Affymetrix). After sample hybridization, arrays were washed and scanned at a resolution of 3 µM using a commercially available confocal laser scanner (Affymetrix).

Data Processing

Scanned image files were visually inspected for artifacts and analyzed with `GeneChip` 3.1 (Affymetrix). The data were normalized by setting the mean hybridization signal for each sample equal to 200. Initial data processing was accomplished with Affymetrix `GeneChip` software. Expression correlations were calculated with the correlation function within `MatLab` (Mathworks) and ad hoc `Perl` scripts. Clustering and data filtering was performed using `GeneSpring` 4.0 (Silicon Genetics).

Northern Blot Analysis

Northern analysis was performed with the Northern Max Kit from Ambion. Thirty µg of glyoxylated total RNA was separated in a 1% TBE agarose gel, blotted to Brightstar Plus membrane filter (Ambion) and hybridized to labeled PCR products. PCR products were labeled with (^{32}P)dCTP by random priming (Roche). Hybridizations were performed at 42°C for 16 hr. The resulting blots were washed at 42°C and imaged using a Molecular Dynamics Storm imager and autoradiographic film.

MudPIT Analysis

Whole protein extracts of *S. cerevisiae* strains *BJ5460*, *BY4741*, and *S288C* grown in rich media to mid-log phase at 30°C were prepared as described previously (Wolters et al. 2001). The samples were subjected to MudPIT analysis on a quaternary Hewlett Packard 1100 series HPLC that was directly coupled to a Finnigan LCQ ion trap mass spectrometer equipped with a nano-liquid chromatography ionization source as described previously (Washburn et al. 2001; Wolters et al. 2001). The SEQUEST algorithm (Eng et al. 1994) was run on each of the datasets using a database that contained the yeast_orfs.fasta database from the NCBI concatenated with 1458 potential NORFS identified in the initial SAGE study (Velculescu et al. 1997). The SEQUEST results were interpreted as described previously (Washburn et al. 2001; Wolters et al. 2001). Briefly, for specific identification of peptides from NORFS, the matches of tandem mass spectra for which the top scoring peptide was from a NORF were analyzed if the ΔCn was at least 0.1. When this was the case, the Xcorr was then analyzed in a charge-state dependent fashion. Xcorr and ΔCn are scoring values by which a user can judge the quality of a SEQUEST result (Eng et al. 1994). The same criteria for Xcorr were used for matches to NORFS as those described previously for other matches in which a +1 peptide had to be at least partially tryptic and with an Xcorr of at least 1.9, a +2 peptide had to be at least partially tryptic with an Xcorr between 2.2 and 3.0, a +2 peptide with an Xcorr >3.0 was accepted regardless of its tryptic nature, and a +3 peptide had to be at least partially tryptic with an Xcorr of at least 3.75. When a tandem mass spectra to a NORF was detected and passed the above criteria, the match was visually assessed for complete confidence as described previously (Washburn et al. 2001; Wolters et al. 2001).

ACKNOWLEDGMENTS

We thank Pete Schultz and Steve Kay for supporting this research, Mike Mittmann at Affymetrix for help with the design of the S98 Array, Victor Velculescu for providing a list of the NORFS, and Katy Donaldson for critical reading of the manuscript. John R. Yates acknowledges funding from the National Institutes of Health (R33CA81665–01 and RR11823–03); Elizabeth Winzeler from the Ellison Medical Foundation (EMF ID-NS-0050–01); and Michael P. Washburn acknowledges support from the genome training grant T32HG000035–05.

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

Arnold, I., Pfeiffer, K., Neupert, W., Stuart, R.A., and Schagger, H. 1998. Yeast mitochondrial F1F0-ATP synthase exists as a dimer: Identification of three dimer-specific subunits. *Embo J* **17:** 7170–7178.

Basrai, M.A., Hieter, P., and Boeke, J.D. 1997. Small open reading frames: Beautiful needles in the haystack. *Genome Res.* **7:** 768–771.

Basrai, M.A., Velculescu, V.E., Kinzler, K.W., and Hieter, P. 1999. NORF5/HUG1 is a component of the MEC1-mediated checkpoint response to DNA damage and replication arrest in *Saccharomyces cerevisiae*. *Mol. Cell Biol.* **19:** 7041–7049.

Blandin, G., Durrens, P., Tekaia, F., Aigle, M., Bolotin-Fukuhara, M., Bon, E., Casaregola, S., de Montigny, J., Gaillardin, C., Lepingle, A. et al. 2000. Genomic exploration of the hemiascomycetous yeasts: 4. The genome of *Saccharomyces cerevisiae* revisited. *FEBS Lett* **487:** 31–36.

Cherry, J.M., Adler, C., Ball, C., Chervitz, S.A., Dwight, S.S., Hester, E.T., Jia, Y., Juvik, G., Roe, T., Schroeder, M. et al. 1998. SGD: *Saccharomyces* Genome Database. *Nucleic Acids Res.* **26:** 73–79.

Cho, R.J., Campbell, M.J., Winzeler, E.A., Steinmetz, L., Conway, A., Wodicka, L., Wolfsberg, T.G., Gabrielian, A.E., Landsman, D., Lockhart, D.J. et al. 1998. A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol. Cell* **2:** 65–73.

Cliften, P.F., Hillier, L.W., Fulton, L., Graves, T., Miner, T., Gish, W.R., Waterston, R.H., and Johnston, M. 2001. Surveying *Saccharomyces* genomes to identify functional elements by comparative DNA sequence analysis. *Genome Res.* **11:** 1175–1186.

Cohen, B.A., Mitra, R.D., Hughes, J.D., and Church, G.M. 2000. A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression. *Nat. Genet.* **26:** 183–186.

Elledge, S.J., Zhou, Z., Allen, J.B., and Navas, T.A. 1993. DNA damage and cell cycle regulation of ribonucleotide reductase. *Bioessays* **15:** 333–339.

Eng, J.K., McCormack, A.L., and Yates, J.R.I. 1994. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* **5:** 976–989.

Goffeau, A., Barrell, B.G., Bussey, H., Davis, R.W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J.D., Jacq, C., Johnston, M. et al. 1996. Life with 6000 genes. *Science* **274:** 546, 563–567.

Huang, M. and Elledge, S.J. 1997. Identification of RNR4, encoding a second essential small subunit of ribonucleotide reductase in *Saccharomyces cerevisiae*. *Mol. Cell Biol.* **17:** 6105–6113.

Hughes, T.R., Marton, M.J., Jones, A.R., Roberts, C.J., Stoughton, R., Armour, C.D., Bennett, H.A., Coffey, E., Dai, H., He, Y.D. et al. 2000. Functional discovery via a compendium of expression profiles. *Cell* **102:** 109–126.

Jelinsky, S.A. and Samson, L.D. 1999. Global response of *Saccharomyces cerevisiae* to an alkylating agent. *Proc. Natl. Acad. Sci.* **96:** 1486–1491.
Jelinsky, S.A., Estep, P., Church, G.M., and Samson, L.D. 2000. Regulatory networks revealed by transcriptional profiling of damaged *Saccharomyces cerevisiae* cells: rpn4 links base excision repair with proteasomes. *Mol. Cell Biol.* **20:** 8157–8167.
Kapranov, P., Cawley, S.E., Drenkow, J., Bekiranov, S., Strausberg, R.L., Fodor, S.P., and Gingeras, T.R. 2002. Large-scale transcriptional activity in chromosomes 21 and 22. *Science* **296:** 916–919.
Kiser, G.L. and Weinert, T.A. 1996. Distinct roles of yeast MEC and RAD checkpoint genes in transcriptional induction after DNA damage and implications for function. *Mol. Biol. Cell* **7:** 703–718.
Kumar, A., Harrison, P.M., Cheung, K.H., Lan, N., Echols, N., Bertone, P., Miller, P., Gerstein, M.B., and Snyder, M. 2002. An integrated approach for finding overlooked genes in yeast. *Nat. Biotechnol.* **20:** 58–63.
Kurtz, S. and Shore, D. 1991. RAP1 protein activates and silences transcription of mating-type genes in yeast. *Genes Dev.* **5:** 616–628.
Lockhart, D.J. and Winzeler, E.A. 2000. Genomics, gene expression and DNA arrays. *Nature* **405:** 827–836.
Mannhaupt, G., Schnall, R., Karpov, V., Vetter, I., and Feldmann, H. 1999. Rpn4p acts as a transcription factor by binding to PACE, a nonamer box found upstream of 26S proteasomal and other genes in yeast. *FEBS Lett* **450:** 27–34.
Mewes, H.W., Albermann, K., Heumann, K., Liebl, S., and Pfeiffer, F. 1997. MIPS: A database for protein sequences, homology data and yeast genome information. *Nucleic Acids Res.* **25:** 28–30.
Mewes, H.W., Heumann, K., Kaps, A., Mayer, K., Pfeiffer, F., Stocker, S., and Frishman, D. 1999. MIPS: A database for genomes and protein sequences. *Nucleic Acids Res.* **27:** 44–48.
Moehle, C.M. and Hinnebusch, A.G. 1991. Association of RAP1 binding sites with stringent control of ribosomal protein gene transcription in *Saccharomyces cerevisiae*. *Mol. Cell Biol.* **11:** 2723–2735.
Olivas, W.M., Muhlrad, D., and Parker, R. 1997. Analysis of the yeast genome: Identification of new non-coding and small ORF-containing RNAs. *Nucleic Acids Res.* **25:** 4619–4625.
Ozsarac, N., Straffon, M.J., Dalton, H.E., and Dawes, I.W. 1997. Regulation of gene expression during meiosis in *Saccharomyces cerevisiae*: SPR3 is controlled by both ABFI and a new sporulation control element. *Mol. Cell Biol.* **17:** 1152–1159.
Pease, A.C., Solas, D., Sullivan, E.J., Cronin, M.T., Holmes, C.P., and Fodor, S.P. 1994. Light-generated oligonucleotide arrays for rapid DNA sequence analysis. *Proc. Natl. Acad. Sci.* **91:** 5022–5026.
Rittberg, D.A. and Wright, J.A. 1989. Relationships between sensitivity to hydroxyurea and 4-methyl-5-amino-1–formylisoquinoline thiosemicarbazone (MAIO) and ribonucleotide reductase RNR2 mRNA levels in strains of *Saccharomyces cerevisiae*. *Biochem. Cell Biol.* **67:** 352–357.
Ross-Macdonald, P., Coelho, P.S., Roemer, T., Agarwal, S., Kumar, A., Jansen, R., Cheung, K.H., Sheehan, A., Symoniatis, D., Umansky, L. et al. 1999. Large-scale analysis of the yeast genome by transposon tagging and gene disruption. *Nature* **402:** 413–418.
Sharp, P.M. and Li, W.H. 1987. The codon adaptation index—a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.* **15:** 1281–1295.
Shoemaker, D.D., Schadt, E.E., Armour, C.D., He, Y.D., Garrett-Engele, P., McDonagh, P.D., Loerch, P.M., Leonardson, A., Lum, P.Y., Cavet, G. et al. 2001. Experimental annotation of the human genome using microarray technology. *Nature* **409:** 922–927.
Velculescu, V.E., Zhang, L., Zhou, W., Vogelstein, J., Basrai, M.A., Bassett, Jr., D.E., Hieter, P., Vogelstein, B., and Kinzler, K.W. 1997. Characterization of the yeast transcriptome. *Cell* **88:** 243–251.
Washburn, M.P., Wolters, D., and Yates, 3rd, J.R., 2001. Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat. Biotechnol.* **19:** 242–247.
Wodicka, L., Dong, H., Mittmann, M., Ho, M.H., and Lockhart, D.J. 1997. Genome-wide expression monitoring in *Saccharomyces cerevisiae*. *Nat. Biotechnol.* **15:** 1359–1367.
Wolters, D.A., Washburn, M.P., and Yates, 3rd, J.R. 2001. An automated multidimensional protein identification technology for shotgun proteomics. *Anal. Chem.* **73:** 5683–5690.

WEB SITE REFERENCES

http://pub.gnf.org/~ewinzeler/identification_of_new_gene.htm; Genomics Institute of the Novartis Research Foundation site.

Received December 7, 2001; accepted in revised form May 17, 2002.