Cell, Vol. 88, 243–251, January 24, 1997, Copyright ©1997 by Cell Press

Characterization of the Yeast Transcriptome

Victor E. Velculescu,*† Lin Zhang,‡ Wei Zhou,‡ Jacob Vogelstein,† Munira A. Basrai,§ Douglas E. Bassett Jr.,*§‖ Phil Hieter,*§ Bert Vogelstein,*†‡ and Kenneth W. Kinzler†
*Program in Human Genetics and Molecular Biology
†Oncology Center
‡Howard Hughes Medical Institute
§Department of Molecular Biology and Genetics
The Johns Hopkins University School of Medicine
Baltimore, Maryland 21231
‖National Center for Biotechnology Information
National Library of Medicine
Bethesda, Maryland 20894

Summary

We have analyzed the set of genes expressed from the yeast genome, herein called the transcriptome, using serial analysis of gene expression. Analysis of 60,633 transcripts revealed 4,665 genes, with expression levels ranging from 0.3 to over 200 transcripts per cell. Of these genes, 1981 had known functions, while 2684 were previously uncharacterized. The integration of positional information with gene expression data allowed for the generation of chromosomal expression maps identifying physical regions of transcriptional activity and identified genes that had not been predicted by sequence information alone. These studies provide insight into global patterns of gene expression in yeast and demonstrate the feasibility of genome-wide expression studies in eukaryotes.

Introduction

It is by now axiomatic that the phenotype of an organism is largely determined by the genes expressed within it. These expressed genes can be represented by a "transcriptome" conveying the identity of each expressed gene and its level of expression for a defined population of cells. Unlike the genome, which is essentially a static entity, the transcriptome can be modulated by both external and internal factors. The transcriptome thereby serves as a dynamic link between an organism's genome and its physical characteristics.

The transcriptome, as defined above, has not been characterized in any eukaryotic or prokaryotic organism, largely because of technological limitations. Some general features of gene expression patterns, however, were elucidated two decades ago through RNA-DNA hybridization measurements (Bishop et al., 1974; Hereford and Rosbash, 1977). In many organisms, it was thus found that at least three classes of transcripts could be identified with either high, medium, or low levels of expression, and the number of transcripts per cell was estimated (Lewin, 1980). These data, of course, provided little information about the specific genes that were members of each class. Data on the expression levels of individual genes have accumulated as new genes have been discovered. In only a few instances, however, have the absolute levels of expression of particular genes been measured and compared to other genes in the same cell type.

Description of any cell's transcriptome would therefore provide new information useful for understanding numerous aspects of cell biology and biochemistry. In this paper, we provide the first description of a transcriptome, determined in S. cerevisiae cells. This organism was chosen because it is widely used to clarify the biochemical and physiologic parameters underlying eukaryotic cellular functions and because it is the only eukaryote for which the entire genome has been defined at the nucleotide level (Goffeau et al., 1996).

Results

Characteristics and Rationale of SAGE Approach

Several methods have recently been described for the high throughput evaluation of gene expression (Nguyen et al., 1995; Schena et al., 1995; Velculescu et al., 1995). We used SAGE (serial analysis of gene expression) because it can provide quantitative gene expression data without the prerequisite of a hybridization probe for each transcript. The SAGE technology is based on two main principles (Figure 1). First, a short sequence tag (9–11 bp) is generated that contains sufficient information to identify uniquely a transcript, provided that it is derived from a defined location within that transcript. Second, many transcript tags can be concatenated into a single molecule and then sequenced, revealing the identity of multiple tags simultaneously. The expression pattern of any population of transcripts can be quantitatively evaluated by determining the abundance of individual tags and identifying the gene corresponding to each tag.

Genome-wide Expression

In order to maximize representation of genes involved in normal growth and cell-cycle progression, SAGE libraries were generated from yeast cells in three states: log phase, S phase–arrested, and G2/M phase–arrested. In total, SAGE tags corresponding to 60,633 total transcripts were identified (including 20,184 from log phase, 20,034 from S phase–arrested, and 20,415 from G2/M phase– arrested cells). Of these tags, 56,291 tags (93%) precisely matched the yeast genome, 88 tags matched the mitochondrial genome, and 91 tags matched the 2-micron plasmid.

The number of SAGE tags required to define a yeast transcriptome depends on the confidence level desired for detecting low abundance mRNA molecules. Assuming the previously derived estimate of 15,000 mRNA molecules per cell (Hereford and Rosbash, 1977), 20,000 tags would represent a 1.3-fold coverage even for mRNA molecules present at a single copy per cell and would provide a 72% probability of detecting such transcripts

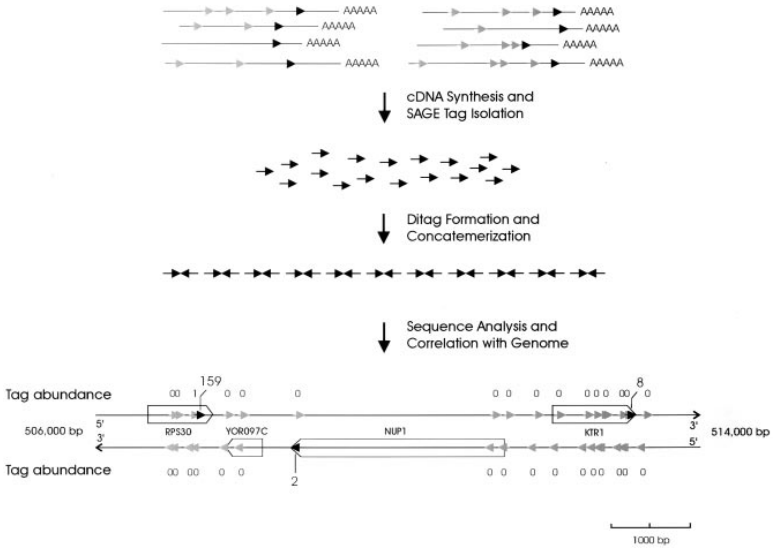


Figure 1. Schematic of SAGE Method and Genome Analysis

In applying SAGE to the analysis of yeast gene expression patterns, the 3′ most NlaIII site was used to define a unique position in each transcript and to provide a site for ligation of a linker with a BsmFI site. The type IIs enzyme BsmFI, which cleaves a defined distance from its nonpalindromic recognition site, was then used to generate a 15 bp SAGE tag (designated by the black arrows), which includes the NlaIII site. Automated sequencing of concatenated SAGE tags allowed the routine identification of ~1000 tags per sequencing gel. Once sequenced, the abundance of each SAGE tag was calculated, and each tag was used to search the entire yeast genome to identify its corresponding gene. The lower panel shows a small region of Chromosome 15. Gray arrows indicate all potential SAGE tags (NlaIII sites) and black arrows indicate 3′-most SAGE tags. The total number of tags observed for each potential tag is indicated above (+ strand) or below (− strand) the tag. As expected, the observed SAGE tags were associated with the 3′ end of expressed genes.

(as determined by Monte Carlo simulations). Analysis of 20,184 tags from log phase cells identified 3,298 unique genes. As an independent confirmation of mRNA copy number per cell, we compared the expression level of *SUP44*/*RPS4*, one of the few genes whose absolute mRNA levels have been reliably determined by quantitative hybridization experiments (Iyer and Struhl, 1996), with expression levels determined by SAGE. *SUP44*/*RPS4* was measured by hybridization at 75 ± 10 copies per cell (Iyer and Struhl, 1996), in good accord with the SAGE data of 63 copies/cell, suggesting that the estimate of 15,000 mRNA molecules per cell was reasonably accurate. Analysis of SAGE tags from S phase–arrested and G2/M phase–arrested cells revealed similar expression levels for this gene (range 52–55 copies per cell), as well as for the vast majority of expressed genes. Since less than 1% of the genes were expressed at

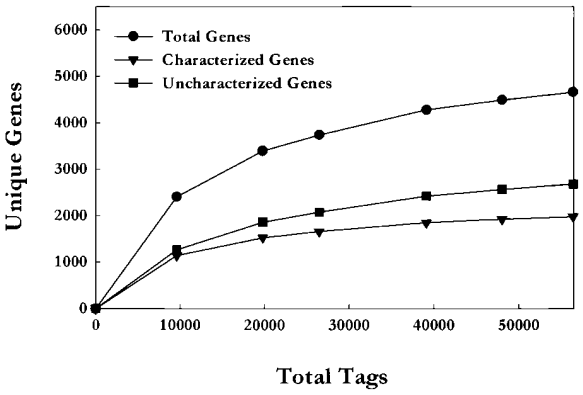
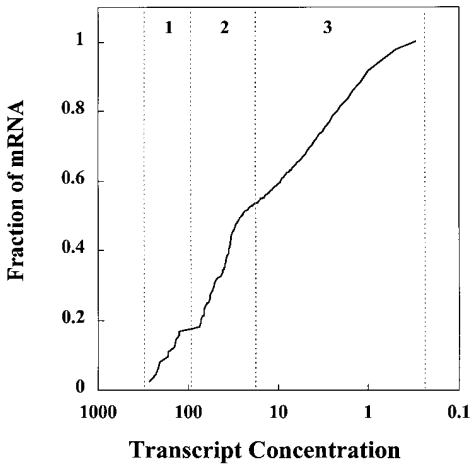


Figure 2. Sampling of Yeast Gene Expression

Analysis of increasing amounts of ascertained tags reveals a plateau in the number of unique expressed genes. Triangles represent genes with known functions, squares represent genes predicted on the basis of sequence information, and circles represent total genes.

A



B

Component	Virtual Rot (SAGE) %mRNA	Virtual Rot (SAGE) Copies/cell	Rot (Reassociation) %mRNA	Rot (Reassociation) Copies/cell
1	17	180	23	200
2	38	40	51	30
3	45	2.5	26	1.5

Figure 3. Virtual Rot

(A) Abundance Classes in the Yeast Transcriptome. The transcript abundance is plotted in reverse order on the abscissa, whereas the fraction of total transcripts with at least that abundance is plotted on the ordinate. The dotted lines identify the three components of the curve, 1, 2, and 3. This is analogous to a Rot curve derived from reassociation kinetics where the product of initial RNA concentration and time is plotted on the abscissa, and the percent of labeled cDNA that hybridizes to excess mRNA is plotted on the ordinate.

(B) Comparison of Virtual Rot and Rot Components. Transitions and data from virtual Rot components were calculated from the data in Figure 3A, while data for Rot components were obtained from Hereford and Rosbash, 1977.

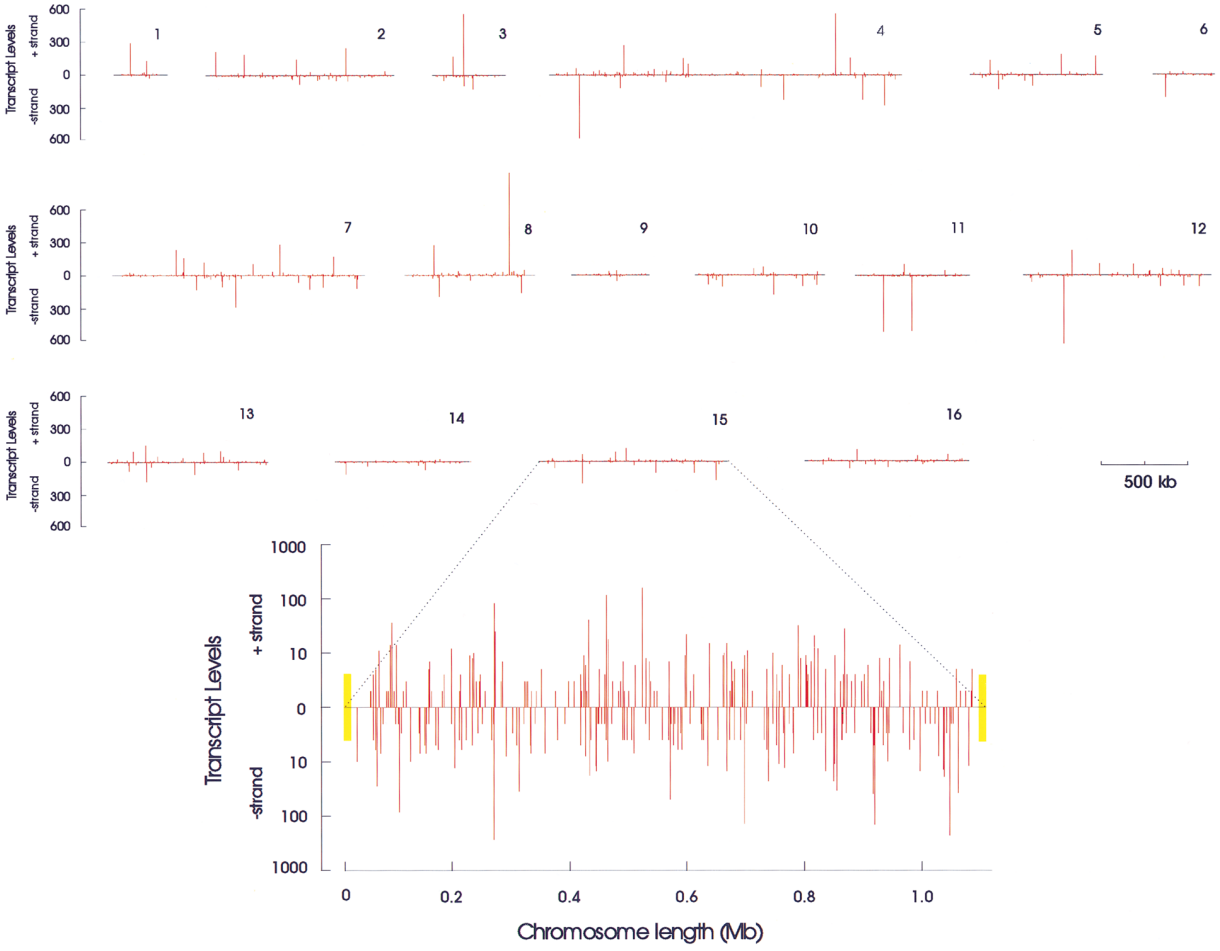


Figure 4. Chromosomal Expression Map for S. cerevisiae

Individual yeast genes were positioned on each chromosome according to their open reading frame (ORF) start coordinates. Abundance levels of tags corresponding to each gene are displayed on the vertical axis, with transcription from the + strand indicated above the abscissa and that from the − strand indicated below. Yellow bands at ends of the expanded chromosome represent telomeric regions that are undertranscribed (see text for details).

dramatically different levels among these three states (see below), SAGE tags obtained from all libraries were combined and used to analyze global patterns of gene expression.

Analysis of ascertained tags at increasing increments revealed that the number of unique transcripts plateaued at ~60,000 tags (Figure 2). This suggested that generation of further SAGE tags would yield few additional genes, consistent with the fact that 60,000 transcripts represented a 4-fold redundancy for genes expressed as low as one transcript per cell. Likewise, Monte Carlo simulations indicated that analysis of 60,000 tags would identify at least one tag for a given transcript 97% of the time if its expression level was one copy per cell.

The 56,291 tags that precisely matched the yeast genome represented 4,665 different genes. This number is in agreement with the estimate of 3,000–4,000 expressed genes obtained by RNA-DNA reassociation kinetics (Hereford and Rosbash, 1977). These expressed genes included 85% of the genes with characterized functions (1981 of 2340) and 76% of the total genes predicted from analysis of the yeast genome (4665 of 6121). These numbers are consistent with a relatively complete sampling of the yeast transcriptome, given the limited number of physiological states examined and the large number of genes predicted solely on the basis of genomic sequence analysis.

The transcript expression per gene was observed to vary from 0.3 to over 200 copies per cell. Analysis of the distribution of gene expression levels revealed several abundance classes that were similar to those observed in previous studies using reassociation kinetics. A "virtual Rot" of the genes observed by SAGE (Figure 3A) identified three main components of the transcriptome with abundances ranging over three orders of magnitude. A Rot curve derived from RNA–cDNA reassociation kinetics also contained three main components distributed over a similar range of abundances (Hereford and Rosbash, 1977). Although the kinetics of reassociation of a particular class of RNA and cDNA may be affected by numerous experimental variables, there were striking similarities between Rot and virtual Rot analyses (Figure 3B). Because Rot analysis may not detect all transcripts

Table 1. Highly Expressed Genes

Tag	Gene	Locus	Copies/Cell	Description
GGTGTTAACG	*TDH2/TDH3*	YJR009C/YGR192C	425	Glyceraldehyde-3-phosphate dehydrogenase 2 & 3
AGACAAACTG	*TEF1/TEF2*	YPR080W/YBR118W	248	Cytosolic elongation factor eEF-1 alpha-A chain
TACCACTCCT	*ENO2*	YHR174W	229	2-Phosphoglycerate dehydratase
GGTTTCGGTT	*RPLA1, A2, A3, 10E*	YDL081C/YOL039W/YDL 130W/YLR340W	207	Acidic ribosomal protein a1/P2.beta/L44prime/L10
TTGCCAGTCT	*PDC1*	YLR044C	207	Pyruvate decarboxylase isozyme 1
GGTGAAAACG	*ADH1, ADH2*	YOL086C/YMR303C	182	Alcohol dehydrogenase I/II
ATCGCCGCTC	*GPM1*	YKL152C	168	Phosphoglycerate mutase
GGTGCTAAGA	*FBA1*	YKL060C	166	Fructose-bisphosphate aldolase II
TTAGTTTCTA	*RPL47A*	YDL184C	143	Ribosomal protein
TCTCTACTGG	*PGK1*	YCR012W	139	Phosphoglycerate kinase
GGTTTTGGTT	*RPLA4*	YDR382W	138	Acidic ribosomal protein L45
GGTCCAGCTT	*SSM1A/SSM1B*	YPL220W/YGL135W	128	Ribosomal protein
AATCCAGTTG	*RPL5A/RPL5B*	YIL018W/YFR031AC	102	Ribosomal protein
TTCGTTCACT		NORF1	94	Nonannotated ORF
AACAGACCAG	*RPL16A/RPL16B*	YPR102C/YGR085C	83	Ribosomal protein
CTGCTCTGGG	*CUP1A/CUP1B*	YHR053C/YHR055C	75	Metallothionein
GCAATACTAC		YOR293W/YMR230W	73	Ribosomal protein S10/similarity to ribosomal protein S10
GCTCTCCCCC		NORF2	73	Nonannotated ORF
AAAGACAGAG	*RPS31A*	YGR027C	72	Ribosomal protein
TGTCGTGGTG	*RPL2A/RPL2B*	YBR031W/YDR012W	70	Ribosomal protein
CCAAGGGTAT	*RPS28A*	YGR118W	69	Ribosomal protein
TCTCCAGAAG	*RPL35B*	YDR500C	69	Ribosomal protein
GTTTTTCTTT	*PYK1*	YAL038W	69	Pyruvate kinase
ATCACTGGTG	*RPL9A/RPL9B*	YGL147C/YNL067W	68	Ribosomal protein L9
ATGAAGGTTC	*RPL27A*	YHR010W	68	Ribosomal protein L27
GTAGAGCCGG	*RPS21*	YOL040C	67	Ribosomal protein
GGTACTGATG	*RPL43A*	YDL075W	67	Ribosomal protein L31
CCAGATTTGT	*NAB1A/NAB1B*	YGR214W/YLR048W	67	40S ribosomal protein p40 homolog A
GTGCCGTCCA	*URP1A*	YBR191W	62	Ribosomal protein L21
CAAAACCCAA	*RPS18EB*	YML026C	60	Ribosomal protein S18

Tag represents the 10 bp SAGE tag adjacent to the NlaIII site; Gene represents the gene or genes corresponding to a particular tag (multiple genes that match unique tags are from related families, with an average identity of 93%); Locus and Description denote the locus name and functional description of each ORF, respectively; Copies/cell represents the abundance of each transcript in the SAGE library, assuming 15,000 total transcripts per cell and 60,633 ascertained transcripts.

of low abundance (Lewin, 1980), it is not surprising that SAGE revealed both a larger total number of expressed genes and a higher fraction of the transcriptome belonging to the low abundance transcript class.

Integration of Expression Information with the Genomic Map

The SAGE expression data were integrated with existing positional information to generate chromosomal expression maps (Figure 4). These maps were generated using the sequence of the yeast genome and the position coordinates of ORFs obtained from the Saccharomyces Genome Database. Although there were a few genes that were noted to be physically proximal and have similarly high levels of expression, there did not appear to be any clusters of particularly high or low expression on any chromosome. Genes like histones H3 and H4, which are known to have coregulated divergent promoters and are immediately adjacent on chromosome 14 (Smith and Murray, 1983), had very similar expression levels (five and six copies per cell, respectively). The distribution of transcripts among the chromosomes suggested that overall transcription was evenly dispersed, with total transcript levels being roughly linearly related to chromosome size ($r^2 = 0.85$, data not shown). Regions within 10 kb of telomeres, however, appeared to be uniformly undertranscribed, containing on average 3.2 tags per gene as compared with 12.4 tags per gene for nontelomeric regions (Figure 4). This is consistent with the previously described observations of telomeric silencing in yeast (Gottschling et al., 1990). Recent studies have reported telomeric position effects as far as 4 kb from telomere ends (Renauld et al., 1993).

Gene Expression Patterns

Table 1 lists the 30 most highly expressed genes, all of which were expressed at greater than 60 mRNA copies per cell. As expected, these genes mostly corresponded to well-characterized enzymes involved in energy metabolism and protein synthesis and were expressed at similar levels in all three growth states (examples in Figure 5). Some of these genes, including *ENO2* (McAlister and Holland, 1982), *PDC1* (Schmitt et al., 1983), *PGK1* (Chambers et al., 1989), *PYK1* (Nishizawa et al., 1989), and *ADH1* (Denis et al., 1983), are known to be dramatically induced in the glucose-rich growth conditions used in this study. In contrast, glucose–repressible genes such as the *GAL1/GAL7/GAL10* cluster (St. John and Davis, 1979) and *GAL3* (Bajwa et al., 1988) were observed to be expressed at very low levels (0.3 or fewer copies per cell). As expected for the yeast strain used in this study, mating type **a** specific genes, such as the **a** factor genes (*MFA1, MFA2*) (Michaelis and Herskowitz, 1988), and alpha factor receptor (*STE2*) (Burkholder and

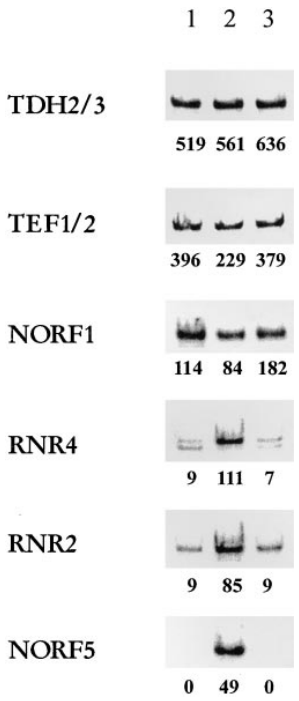


Figure 5. Northern Blot Analysis of Representative Genes

TDH2/3, *TEF1/2*, and *NORF1* are expressed relatively equally in all three states (lane 1, G2/M arrested; lane 2, S phase–arrested; lane 3, log phase), while *RNR4*, *RNR2*, and *NORF5* are highly expressed in S phase–arrested cells. The expression level observed by SAGE (number of tags) is noted below each lane and is highly correlated with quantitation of the Northern blot by PhosphorImager analysis ($r^2 = 0.97$).

Hartwell, 1985) were all observed to be expressed at significant levels (range 2–10 copies per cell), while mating type alpha specific genes (*MFα1, MFα2, STE3*) (Kurjan and Herskowitz, 1982; Singh et al., 1983; Hagen et al., 1986) were observed to be expressed at very low levels ($<$0.3 copies per cell).

Three of the highly expressed genes in Table 1 had not been previously characterized. One gene contained an ORF with predicted ribosomal function, previously identified only by genomic sequence analysis. Analyses of all SAGE data suggested that there were 2684 such genes corresponding to uncharacterized ORFs that were transcribed at detectable levels. The 30 most abundant of these transcripts were observed more than 30 times, corresponding to at least eight transcripts per cell (Table 2). The other two highly expressed uncharacterized genes corresponded to ORFs not predicted by analysis of the yeast genome sequence. Analyses of SAGE data suggested that there were approximately 160 *Nonannotated ORF* (*NORF*) genes transcribed at detectable levels. The 30 most abundant of these transcripts were observed at least nine times (Table 3 and examples in Figure 5).

Interestingly, one of the *NORF* genes (*NORF5*) was only expressed in S phase–arrested cells and corresponded to the transcript whose abundance varied the most in the three states analyzed ($>$49-fold, Figure 5). Comparison of S phase–arrested cells to the other states also identified greater than 9-fold elevation of the *RNR2* and *RNR4* transcripts (Figure 5). Induction of these ribonucleoside reductase genes is likely to be due to the hydroxyurea treatment used to arrest cells in S phase (Elledge and Davis, 1989). Likewise, comparison of G2/M–arrested cells identified elevation of RBL2 and dynein light chain, both microtubule-associated proteins (Archer et al., 1995; Dick et al., 1996). As with the RNR inductions, these elevated levels seem likely to be related to the nocodazole treatment used to arrest cells in the G2/M phase. While there were many relatively small differences between the states (for example, *NORF1*, Figure 5), overall comparison of the three states revealed surprisingly few dramatic differences; there were only 29 transcripts whose abundance varied more than 10-fold among the three different states analyzed. Tables including all SAGE tags and expression levels are available from the authors upon request.

Discussion

Analysis of a yeast transcriptome affords a unique view of the RNA components defining cellular life. We observed gene expression levels to vary over three orders of magnitude, with the transcripts involved in energy metabolism and protein synthesis the most highly expressed. Key transcripts, such as those encoding enzymes required for DNA replication (e.g. *POL1* and *POL3*), kinetochore proteins (*NDC10* and *SKP1*), and many other interesting proteins, were present at one or fewer copies per cell on average. These abundances are consistent with previous qualitative data from reassociation kinetics, which suggested that the largest number of expressed genes were present at one or two copies per cell. These observations indicate that low transcript copy numbers are sufficient for gene expression in yeast and suggest that yeast possess a mechanism for rigid control of RNA abundance.

The synthesis of chromosomal expression maps presents a cataloging of the expression level of genes, organized by their genomic positions. It is not surprising that gene expression is well-balanced throughout the 16 chromosomes of S. cerevisiae. Since most genes have independent regulatory elements, it would have been surprising to find a large number of physically adjacent genes that had similar high levels of expression. Of the few genes that were known to have coregulated divergent promoters, like the H3/H4 pair, SAGE data confirmed concordant levels of expression. For areas like telomere ends that are known to be transcriptionally suppressed, SAGE data corroborated low levels of expression. Other expected expression patterns were observed, such as high levels of glucose-induced glycolytic enzymes, low levels of glucose- repressed *GAL* genes, expression of mating type **a** specific genes, and low of expression of mating type alpha genes. Finally, identification of tags corresponding to *NORF* genes suggests that there is a significant number of small proteins encoded by the yeast genome that were undetected by the criteria used for systematic sequence analysis. The yeast genome sequence has been annotated for all ORFs larger than 300 bp (encoding proteins 100 amino

Table 2. Expression of Putative Coding Sequences

SAGE Tag	Locus	Copies/Cell	Description
TTGAACTACC	YKL056C	58	Strong similarity to human IgE-dependent histamine-releasing factor (21 kDa tumor protein)
TTCGGGTCAC	YDR276C	56	Strong similarity to Hordeum vulgare blt101 protein
CCAGATATGA	YIL093C	41	Hypothetical protein
TTTAAAATGG	YMR116C	38	Similarity to N. crassa CPC2 protein
GGTGTCGTTG	YBR078W	34	Strong similarity to sporulation specific Sps2p
TACTCTTCGC	YEL033W	33	Hypothetical protein
TGTAATTAAA	YOR182C	26	Homology to human ubiquitin-like protein/ribosomal protein S30
GGAGATCTTG	YCR013C	24	Weak similarity to M. lepra B1496_F1_41 protein
TCAAGAAGTT	YER056AC	20	Strong similarity to ribosomal protein L34
AAAAACTTTG	YIL051C	18	Strong similarity to YER057c
AAGTTGAACA	YPR043W	17	Ribosomal protein L37
GGGTGCGGGT	YDR032C	16	Strong similarity to YCR004c and S. pombe obr1
TGACTCTTTG	YLR390W	14	Hypothetical protein
GGTCAATGGC	YJR105W	11	Hypothetical protein
TAAGAATTCT	YJL158C	11	Member of the Pir1p/Hsp150p/Pir3p family
TCAATTATGT	YDR033W	11	Strong similarity to putative heat-shock protein YRO2
ACGGCCAAGA	YBR162C	10	Similarity to YJL171p
TTGGGCTAGT	YJL171C	10	Similarity to YBR162c
CCTTCCAGGT	YJR085C	10	Hypothetical protein
CCTCTCTTGT	YOR310C	10	Homology to SIK1 protein
CCCAAAACTT	YEL018W	9	Weak similarity to Rad50p
AACAAGTACT	YGL037C	9	Similarity to E. coli hypothetical 23K protein
AACAATAAAA	YER072W	8	Similarity to YFL004w
CAAAAGACCG	YML056C	8	Homology to human IMP dehydrogenase I
GGTTTTTGAT	YOR182C	8	Homology to human ubiquitin-like protein/ribosomal protein S30
CAATCCATTT	YBR106W	8	Hypothetical protein
TTTTGGGTCT	YMR318C	8	Putative alcohol dehydrogenase
AACTGTCCAT	YDR429C	8	Similarity to nuclear RNA binding proteins
CCAAGGTTAA	YAR002AC	8	Strong similarity to YGL002w
GGTTTTTGAA	YOR273C	8	Putative resistance protein

Table columns are the same as for Table 1.

acids or greater). Genes encoding proteins below this cutoff are therefore commonly unannotated. This class of genes might also be underrepresented in mutational collections because of the small target size for mutagenesis and, given their small size, may encode proteins with novel functions. The systematic knockout of these *NORF* genes will therefore be of great interest.

Comparison of gene expression patterns from altered physiologic states can provide insight into genes that are important in a variety of processes. Comparison of transcriptomes from a variety of physiologic states should provide a minimum set of genes whose expression is required for normal vegetative growth and another set composed of genes that will be expressed only in response to specific environmental stimuli or during specialized processes. For example, recent work has defined a minimal set of 250 genes required for prokaryotic cellular life (Mushegian and Koonin, 1996). Examination of the yeast genome readily identified homologous genes for 196 of these, over 90% of which were observed to be expressed in the SAGE analysis. Detailed analyses of yeast transcriptomes, as well as transcriptomes from other organisms, should ultimately allow the generation of a minimal set of genes required for eukaryotic life.

Like other genome-wide analyses, SAGE analysis of yeast transcriptomes has several potential limitations. First, a small number of transcripts would be expected to lack an NlaIII site and therefore would not be detected by our analysis. Second, our analysis was limited to transcripts found at least as frequently as 0.3 copies per cell. Transcripts expressed in only a minute fraction of the cell cycle, or transcripts expressed in only a fraction of the cell population, would not be reliably detected by our analysis. Finally, mRNA sequence data are practically unavailable for yeast. Consequently, some SAGE tags cannot be unambiguously matched to corresponding genes. Tags that were derived from overlapping genes or genes that have unusually long 3′ untranslated regions may be misassigned. Increased availability of 3′ UTR sequences in yeast mRNA molecules should help to resolve the ambiguities.

Despite these potential limitations, it is clear that the analyses described here furnish both global and local pictures of gene expression precisely defined at the nucleotide level. These data, like the sequence of the yeast genome itself, provide simple, basic information integral to the interpretation of many experiments in the future. The availability of mRNA sequence information from EST sequencing, as well as various genome projects, will soon allow definition of transcriptomes from a variety of organisms, including human. The data recorded here suggest that a reasonably complete picture of a human cell transcriptome will require only about 10- to 20-fold more tags than evaluated here, a number well within the practical realm achievable with a small number of automated sequencers. The analysis of global expression patterns in higher eukaryotes is expected, in general, to be similar to those reported here for S. cerevisiae. However, the analysis of the transcriptome

Table 3. Expression of *NORF* Genes

SAGE Tag	Locus	Copies/Cell	Chr	Tag Pos	ORF Size (bp)
TTCGTTCACT	NORF1	94	4	1489450	198
GCTCTCCCCC	NORF2	73	16	75633	243
TGTACGCATT	NORF3	16	15	301251	189
TTTTATTATC	NORF4	15	6	223182	177
CTTCTCTTTT	NORF5	12	13	158973	204
TTTCCTATAA	NORF6	11	13	511754	252
TCTAGTCGCC	NORF7	10	12	669659	192
ATCGTTTTAT	NORF8	8	15	877140	174
GGCCAATGGT	NORF9	8	4	1202289	267
ACCCTGTCAT	NORF10	7	2	418633	255
AAAAGATCAT	NORF11	7	4	1489453	87
CAGAAAATGG	NORF12	6	8	115655	279
TGACATTCTT	NORF13	6	16	883669	183
TAGACATCTA	NORF14	6	2	491117	141
TGCCCTGGCC	NORF15	5	5	166452	216
GGTTTTGGCG	NORF16	4	3	24169	291
CCATACAGGT	NORF17	4	12	673851	114
CCAAATCAAA	NORF18	3	4	229494	258
AAGCGGTACT	NORF19	3	9	47889	399
AACGCTTTTC	NORF20	3	2	351456	198
GAGGATAGAG	NORF21	3	2	356201	240
CAATGAACCG	NORF22	3	16	75541	243
TCTTTATATA	NORF23	3	1	73363	90
CGCCTCCAGT	NORF24	3	7	485774	108
TACGTAAGTT	NORF25	3	10	156139	81
GATTTAAACT	NORF26	3	15	254749	93
GCGCCTCCAA	NORF27	2	5	42622	222
CAATGGCCCA	NORF28	2	13	511751	78
TTGAGGAACG	NORF29	2	3	154681	264
GCTAAGAACC	NORF30	2	4	302607	204

SAGE Tag, Locus, and Copies/cell are the same as for Table 1; Chr and Tag Pos denote the chromosome and position of each tag; ORF Size denotes the size of the ORF corresponding to the indicated tag. In each case, the tag was located within or less than 250 bp 3′ of the *NORF*.

in different cells and from different individuals should yield a wealth of information regarding gene function in normal, developmental, and disease states.

Experimental Procedures

Yeast Cell Culture

The source of transcripts for all experiments was S. cerevisiae strain YPH499 (*MATa ura3–52 lys2–801 ade2–101 leu2–Δ1 his3–Δ200 trp1–Δ63*) (Sikorski and Hieter, 1989). Logarithmically growing cells were obtained by growing yeast cells to early log phase (3×10^6 cells/ml) in YPD (Rose et al., 1990) rich medium (YPD supplemented with 6 mM uracil, 4.8 mM adenine, and 24 mM tryptophan) at 30°C. For arrest in the G1/S phase of the cell cycle, hydroxyurea (0.1 M) was added to early log phase cells, and the culture was incubated an additional 3.5 hr at 30°C. 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 for an additional 100 min at 30°C. Harvested cells were washed once with water prior to freezing at −70°C. The growth states of the harvested cells were confirmed by microscopic and flow-cytometric analyses (Basrai et al., 1996).

RNA Isolation and Northern Blot Analysis

Total yeast RNA was prepared using the hot phenol method as described (Leeds et al., 1991). mRNA was obtained using the MessageMaker Kit (GIBCO/BRL) following the manufacturer's protocol. Northern blot analysis was performed as described (El-Deiry et al., 1993) using probes PCR amplified from yeast genomic DNA.

SAGE Protocol

The SAGE method was performed as previously described (Velculescu et al., 1995), with exceptions noted below. PolyA RNA was converted to double-stranded cDNA with a BRL synthesis kit using the manufacturer's protocol except for the inclusion of primer biotin-5′-T_{18}-3′. The cDNA was cleaved with NlaIII (Anchoring Enzyme). Since NlaIII sites were observed to occur once every 309 bp in three arbitrarily chosen yeast chromosomes (1, 5, 10), 95% of yeast transcripts were predicted to be detectable with a NlaIII-based SAGE approach. After capture of the 3′ cDNA fragments on streptavidin-coated magnetic beads (Dynal), the bound cDNA was divided into two pools, and one of the following linkers containing recognition sites for BsmFI was ligated to each pool: linker 1, 5′-TTTGGATTT GCTGGTGCAGTACAACTAGGCTTAATAGGGACATG- 3′, 5′-TCCCT ATTAAGCCTAGTTGTACTGCACCAGCAAATCC[amino mod. C7]-3′; linker 2, 5′-TTTCTGCTCGAATTCAAGCTTCTAACGATGTACGGGG ACATG-3′ 5′-TCCCCGTACATCGTTAGAAGCTTGAATTCGAGCAG [amino mod. C7]-3′.

Since BsmFI (Tagging Enzyme) cleaves 14 bp away from its recognition site, and the NlaIII site overlaps the BsmFI site by 1 bp, a 15 bp SAGE tag was released with BsmFI. SAGE tag overhangs were filled in with Klenow, and tags from the two pools were combined and ligated to each other. The ligation product was diluted and then amplified with PCR for 28 cycles with 5′-GGATTTGCTGGTGCAGT ACA-3′ and 5′-CTGCTCGAATTCAAGCTTCT-3′ as primers. The PCR product was analyzed by polyacrylamide gel electrophoresis (PAGE), and the PCR product containing two tags ligated tail to tail (ditag) was excised. The PCR product was then cleaved with NlaIII, and the band containing the ditags was excised and self-ligated. After ligation, the concatenated products were separated by PAGE and products between 500 bp and 2 kb were excised. These products were cloned into the SphI site of pZero (Invitrogen). Colonies were screened for inserts by PCR with M13 forward and M13 reverse sequences located outside the cloning site as primers.

PCR products from selected clones were sequenced with the TaqFS DyePrimer kits (Perkin Elmer) and analyzed using a 377 ABI automated sequencer (Perkin Elmer), following the manufacturer's protocol. Each successful sequencing reaction identified an average

of 26 tags; given a 90% sequencing reaction success rate, this corresponded to an average of about 850 tags per sequencing gel.

SAGE Data Analysis

Sequence files were analyzed by means of the SAGE program group (Velculescu et al., 1995), which identifies the anchoring enzyme site with the proper spacing, extracts the two intervening tags, and records them in a database. The 68,691 tags obtained contained 62,965 tags from unique ditags and 5,726 tags from repeated ditags. The latter tags were counted only once to eliminate potential PCR bias of the quantitation, as described (Velculescu et al., 1995). Of 62,965 tags, 2,332 tags corresponded to linker sequences, and were excluded from further analysis. Of the remaining tags, 4,342 tags could not be assigned and were likely due to sequencing errors in the tags or in the yeast genomic sequence. If the inability to assign these tags was only due to tag sequencing errors, this would correspond to a sequencing error rate of about 0.7% per base pair (for a 10 bp tag), not far from what we would have expected under our automated sequencing conditions. However, some unassigned tags had a much higher than expected frequency of A's as the last 5 bp of the tag (5 of the 52 most abundant unassigned tags), suggesting that these tags were derived from transcripts containing anchoring enzyme sites within several base pairs from their polyA tails. Given the frequency of NlaIII sites in the genome (1 in 309 bp), approximately 3% of transcripts were predicted to contain NlaIII sites within 10 bp of their polyA tails.

Since very sparse data are available for yeast mRNA sequences, and efforts to date have not been able to identify a highly conserved polyadenylation signal (Irniger and Braus, 1994; Zaret and Sherman, 1982), we used 14 bp of SAGE tags (i.e. the NlaIII site plus the adjacent 10 bp) to search the yeast genome directly (yeast genome sequence obtained from the Saccharomyces Genome Database ftp site [genome-ftp.stanford.edu] on August 7, 1996). Because only coding regions are annotated in the yeast genome, and SAGE tags can be derived from 3′ untranslated regions of genes, a SAGE tag was considered to correspond to a particular gene if it matched the ORF or the region 500 bp 3′ of the ORF (locus names, gene names and ORF chromosomal coordinates were obtained from Saccharomyces Genome Database ftp site, and ORF descriptions were obtained from MIPS www site [http://www.mips.biochem.mpg.de/] on August 14, 1996). ORFs were considered genes with known functions if they were associated with a three-letter gene name, while ORFs without such designations were considered uncharacterized.

As expected, SAGE tags matched transcribed portions of the genome in a highly nonrandom fashion, with 88% matching ORFs or their adjacent 3′ regions in the correct orientation (chi-squared P value $< 10^{-30}$). In instances when more than one tag matched a particular ORF in the correct orientation, the abundance was calculated to be the sum of the matched tags (for Figures 2–4). Tags that matched ORFs in the incorrect orientation were not used in abundance calculations. In instances when a tag matched more than one region of the genome (for example, an ORF and non-ORF region), only the matched ORF was considered. In some cases, the 15th base of the tag could also be used to resolve ambiguities. For Figure 4, only tags that matched the genome once were used.

For the identification of *NORF* genes, only tags were considered that matched portions of the genome that were further than 500 bp 3′ of a previously identified ORF and were observed at least two times in the SAGE libraries.

Acknowledgments

This work was supported by grants CA57345, CA35494, GM07309, CA16519, and HG00971. B. V. is an ACS Research Professor and an Investigator of the Howard Hughes Medical Institute. We thank members of our laboratories for helpful discussions and critical reading of the manuscript.

Received October 17, 1996; revised December 4, 1996.

References

Archer, J.E., Vega, L.R., and Solomon, F. (1995). Rbl2p, a yeast protein that binds to beta-tubulin and participates in microtubule function in vivo. Cell *82*, 425–434.

Bajwa, W., Torchia, T.E., and Hopper, J.E. (1988). Yeast regulatory gene GAL3: carbon regulation; UASGal elements in common with GAL1, GAL2, GAL7, GAL10, GAL80, and MEL1; encoded protein strikingly similar to yeast and Escherichia coli galactokinases. Mol. Cell. Biol. *8*, 3439–3447.

Basrai, M.A., Kingsbury, J., Koshland, D., Spencer, F., and Hieter, P. (1996). Faithful chromosome transmission requires Spt4p, a putative regulator of chromatin structure in Saccharomyces cerevisiae. Mol. Cell. Biol. *16*, 2838–2847.

Bishop, J.O., Morton, J.G., Rosbash, M., and Richardson, M. (1974). Three abundance classes in HeLa cell messenger RNA. Nature *250*, 199–204.

Burkholder, A.C., and Hartwell, L.H. (1985). The yeast alpha-factor receptor: structural properties deduced from the sequence of the STE2 gene. Nucleic Acids Res. *13*, 8463–8475.

Chambers, A., Tsang, J.S., Stanway, C., Kingsman, A.J., and Kingsman, S.M. (1989). Transcriptional control of the Saccharomyces cerevisiae PGK gene by RAP1. Mol. Cell. Biol. *9*, 5516–5524.

Denis, C.L., Ferguson, J., and Young, E.T. (1983). mRNA levels for the fermentative alcohol dehydrogenase of Saccharomyces cerevisiae decrease upon growth on a nonfermentable carbon source. J. Biol. Chem. *258*, 1165–1171.

Dick, T., Surana, U., and Chia, W. (1996). Molecular and genetic characterization of SLC1, a putative Saccharomyces cerevisiae homolog of the metazoan cytoplasmic dynein light chain 1. Mol. Gen. Genet. *251*, 38–43.

El-Deiry, W.S., Tokino, T., Velculescu, V.E., Levy, D.B., Parsons, R., Trent, J.M., Lin, D., Mercer, W.E., Kinzler, K.W., and Vogelstein, B. (1993). WAF1, a potential mediator of p53 tumor suppression. Cell *75*, 817–825.

Elledge, S.J., and Davis, R.W. (1989). DNA damage induction of ribonucleotide reductase. Mol. Cell. Biol. *9*, 4932–4940.

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

Gottschling, D.E., Aparicio, O.M., Billington, B.L., and Zakian, V.A. (1990). Position effect at S. cerevisiae telomeres: reversible repression of Pol II transcription. Cell *63*, 751–762.

Hagen, D.C., McCaffrey, G., and Sprague, G.F., Jr. (1986). Evidence the yeast STE3 gene encodes a receptor for the peptide pheromone a factor: gene sequence and implications for the structure of the presumed receptor. Proc. Natl. Acad. Sci. USA *83*, 1418–1422.

Hereford, L.M., and Rosbash, M. (1977). Number and distribution of polyadenylated RNA sequences in yeast. Cell *10*, 453–462.

Irniger, S., and Braus, G.H. (1994). Saturation mutagenesis of a polyadenylation signal reveals a hexanucleotide element essential for mRNA 3′ end formation in Saccharomyces cerevisiae. Proc. Natl. Acad. Sci. USA *91*, 257–261.

Iyer, V., and Struhl, K. (1996). Absolute mRNA levels and transcriptional initiation rates in Saccharomyces cerevisiae. Proc. Natl. Acad. Sci. USA *93*, 5208–5212.

Kurjan, J., and Herskowitz, I. (1982). Structure of a yeast pheromone gene (MF alpha): a putative alpha-factor precursor contains four tandem copies of mature alpha-factor. Cell *30*, 933–943.

Leeds, P., Peltz, S.W., Jacobson, A., and Culbertson, M.R. (1991). The product of the yeast UPF1 gene is required for rapid turnover of mRNAs containing a premature translational termination codon. Genes Dev. *5*, 2303–2314.

Lewin, B. (1980). Gene Expression *2*, (New York: John Wiley and Sons), pp. 694–727.

McAlister, L., and Holland, M.J. (1982). Targeted deletion of a yeast enolase structural gene. Identification and isolation of yeast enolase isozymes. J. Biol. Chem. *257*, 7181–7188.

Michaelis, S., and Herskowitz, I. (1988). The a-factor pheromone of Saccharomyces cerevisiae is essential for mating. Mol. Cell. Biol. *8*, 1309–1318.

Mushegian, A.R., and Koonin, E.V. (1996). A minimal gene set for

cellular life derived by comparison of complete bacterial genomes. Proc. Natl. Acad. Sci. USA *93*, 10268–10273.

Nguyen, C., Rocha, D., Granjeaud, S., Baldit, M., Bernard, K., Naquet, P., and Jordan, B.R. (1995). Differential gene expression in the murine thymus assayed by quantitative hybridization of arrayed cDNA clones. Genomics *29*, 207–216.

Nishizawa, M., Araki, R., and Teranishi, Y. (1989). Identification of an upstream activating sequence and an upstream repressible sequence of the pyruvate kinase gene of the yeast Saccharomyces cerevisiae. Mol. Cell. Biol. *9*, 442–451.

Renauld, H., Aparicio, O.M., Zierath, P.D., Billington, B.L., Chhablani, S.K., and Gottschling, D.E. (1993). Silent domains are assembled continuously from the telomere and are defined by promoter distance and strength, and by SIR3 dosage. Genes Dev. *7*, 1133–1145.

Rose, M.D., Winston, F., and Hieter, P. (1990). Methods in Yeast Genetics (Cold Spring Harbor, New York: Cold Spring Harbor Laboratory Press), pp. 177.

Schena, M., Shalon, D., Davis, R.W., and Brown, P.O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science *270*, 467–470.

Schmitt, H.D., Ciriacy, M., and Zimmermann, F.K. (1983). The synthesis of yeast pyruvate decarboxylase is regulated by large variations in the messenger RNA level. Mol. Gen. Genet. *192*, 247–252.

Sikorski, R.S., and Hieter, P. (1989). A system of shuttle vectors and yeast host strains designed for efficient manipulation of DNA in Saccharomyces cerevisiae. Genetics *122*, 19–27.

Singh, A., Chen, E.Y., Lugovoy, J.M., Chang, C.N., Hitzeman, R.A., and Seeburg, P.H. (1983). Saccharomyces cerevisiae contains two discrete genes coding for the alpha-factor pheromone. Nucleic Acids Res. *11*, 4049–4063.

Smith, M.M., and Murray, K. (1983). Yeast H3 and H4 histone messenger RNAs are transcribed from two non-allelic gene sets. J. Mol. Biol. *169*, 641–661.

St. John, T.P., and Davis, R.W. (1979). Isolation of galactose-inducible DNA sequences from Saccharomyces cerevisiae by differential plaque filter hybridization. Cell *16*, 443–452.

Velculescu, V.E., Zhang, L., Vogelstein, B., and Kinzler, K.W. (1995). Serial analysis of gene expression. Science *270*, 484–487.

Zaret, K.S., and Sherman, F. (1982). DNA sequence required for efficient transcription termination in yeast. Cell *28*, 563–573.