

# Genome-wide analysis of mRNA translation profiles in *Saccharomyces cerevisiae*

Yoav Arava\*<sup>†</sup>, Yulei Wang\*, John D. Storey<sup>‡</sup>, Chih Long Liu\*, Patrick O. Brown\*<sup>†§</sup>, and Daniel Herschlag\*<sup>§</sup>

\*Department of Biochemistry, Stanford University, Stanford, CA 94305-5307; <sup>†</sup>The Howard Hughes Medical Institute, Stanford, CA 94305-5428; and <sup>‡</sup>Department of Statistics, Stanford University, Stanford, CA 94305-4065

Edited by Nicholas R. Cozzarelli, University of California, Berkeley, CA, and approved January 23, 2003 (received for review August 26, 2002)

We have analyzed the translational status of each mRNA in rapidly growing *Saccharomyces cerevisiae*. mRNAs were separated by velocity sedimentation on a sucrose gradient, and 14 fractions across the gradient were analyzed by quantitative microarray analysis, providing a profile of ribosome association with mRNAs for thousands of genes. For most genes, the majority of mRNA molecules were associated with ribosomes and presumably engaged in translation. This systematic approach enabled us to recognize genes with unusual behavior. For 43 genes, most mRNA molecules were not associated with ribosomes, suggesting that they may be translationally controlled. For 53 genes, including *GCN4*, *CPA1*, and *ICY2*, three genes for which translational control is known to play a key role in regulation, most mRNA molecules were associated with a single ribosome. The number of ribosomes associated with mRNAs increased with increasing length of the putative protein-coding sequence, consistent with longer transit times for ribosomes translating longer coding sequences. The density at which ribosomes were distributed on each mRNA (i.e., the number of ribosomes per unit ORF length) was well below the maximum packing density for nearly all mRNAs, consistent with initiation as the rate-limiting step in translation. Global analysis revealed an unexpected correlation: Ribosome density decreases with increasing ORF length. Models to account for this surprising observation are discussed.

Genome sequences and DNA microarray technology have given us the ability to analyze cellular mRNA levels globally. Indeed, measurements of overall mRNA levels via these approaches are providing an unprecedented picture of the control and integration of gene expression (e.g., refs. 1–6). Nevertheless, gene expression involves many steps, and understanding the mechanisms and control at each step and how these steps are integrated will require a quantitative description of each RNA at each of these steps as well as the changes in mRNA levels and transit times at each step in response to mutation and changes in growth conditions.

Previous studies have addressed global aspects of mRNA transcription (3, 4, 7), splicing (8), decay (9), localization (10, 11), protein binding (12), and translation (13–16). With respect to translation, several experiments were designed to identify candidates for translational regulation. Researchers have analyzed changes in the association of mRNAs with ribosomes after serum activation or viral infection of mammalian cell lines (13, 14), activation of T cells by anti-CD3 and anti-CD28 antibodies (16), change in carbon source for yeast cells (15), and the absence of the RNA-binding protein, fragile X mental retardation protein (17). These studies revealed changes in ribosome association of specific mRNAs, and further analysis confirmed several of these candidates for regulation.

Fundamental information about translation can also be obtained by a detailed and comprehensive analysis at a single, well defined condition. A wealth of information is available to provide an interpretative framework for such an analysis, including the sequence and abundance of each mRNA in yeast (3, 7, 18), the identity and binding interactions of translational factors (19, 20), and structural information about the ribosome (21). In addition, comprehensive information about the translational status of all mRNAs would provide a basis to interpret

global and specific changes in translation in response to mutations and changes in growth conditions.

We have carried out a comprehensive and detailed characterization of mRNA association with ribosomes in yeast cells growing in rich medium to probe the general features of translational behavior and identify mRNAs that behave distinctively. The results reveal an efficient recruitment of mRNAs to the translating pool, yet several mRNAs with low recruitment efficiency were identified. The data also suggest that initiation is the slowest step in translation for all or nearly all yeast mRNAs. Surprisingly, although longer mRNAs are associated with more ribosomes, the density of ribosomes per unit length is lower for the longer mRNAs.

## Materials and Methods

**Yeast Strains.** Polysomal RNA was isolated from *Saccharomyces cerevisiae* strain BY4741 (MATa his3Δ1 leu2Δ met15Δ ura3Δ) (5).

**Polysomal RNA Preparation.** Cells were grown in 1% yeast extract/2% peptone/2% dextrose (YPD) medium to midlog phase (OD<sub>600</sub> = 0.4–0.6) at 30°C, and polysomal extracts were prepared essentially as described (15). A detailed protocol is available at our web site ([http://genome-www.stanford.edu/yeast\\_translation/index.shtml](http://genome-www.stanford.edu/yeast_translation/index.shtml)).

**Microarray Analysis and Data Selection.** Microarray production, fluorescent labeling, hybridization, and washing were carried out according to published protocols (refs. 4 and 9; <http://cmgm.stanford.edu/pbrown/mguide/index.html>). Half of the RNA isolated from each fraction (equal to polysomes from 120 ml of cells) was used for preparation of Cy5-labeled cDNA by reverse transcription in the presence of Cy5-dUTP dye (Amersham Pharmacia) using oligo(dT) primers. Fifteen micrograms of RNA extracted by the hot-phenol procedure (22) from strain S288c were labeled by the same procedure with Cy3-dUTP. The two differentially labeled cDNA probes were mixed and hybridized to a DNA microarray representing all known and predicted genes of *S. cerevisiae* (*Saccharomyces* Genome Database). Microarrays were scanned by using an Axon Instruments (Foster City, CA) Scanner 4000, and data were collected using the GENEPIX 3.0 program (Axon Instruments). Spots with abnormal morphology were excluded from further analysis. Spots with signal indistinguishable from the local background (usually <1.5× difference) or with a large variation in the ratios of the pixels (correlation coefficient <0.6) were also excluded.

**Normalization.** To control for variable losses of mRNA from fractions during purification and handling and to control for signal variation from differences in dye-labeling efficiency, hybridization, and scanning, a “normalization mix” of five different *in vitro*-transcribed *Bacillus subtilis* RNAs was added to each of the fractions immediately at collection. These mRNAs were transcribed *in vitro* from the following ATCC clones: LysA

This paper was submitted directly (Track II) to the PNAS office.

<sup>§</sup>To whom correspondence may be addressed. E-mail: [herschla@cmgm.stanford.edu](mailto:herschla@cmgm.stanford.edu) or [pbrown@cmgm.stanford.edu](mailto:pbrown@cmgm.stanford.edu).

(clone no. 87482), Phe (no. 87483), Thr (no. 87484), Trp (no. 87485), and Dap (no. 87486). All clones contained poly(A) sequence at the 3' end. To each polysomal fraction, 0.13, 0.2, 0.26, 0.33, and 0.4 ng of LysA, Dap, Thr, Trp, and Phe RNA, respectively, were added. Half of these amounts were added to the reference sample. These clones were also PCR-amplified and spotted onto the microarrays, each in at least five different locations. Typically, the maximal variation of the ratios obtained from these spots on each microarray was <2-fold, and the fractional standard error of the mean was 0.02 with a standard deviation for the 14 fractions of  $\pm 0.01$ . The normalization for variations between samples was applied by adjusting the signals of the Cy5 channel such that the average ratio for the *B. subtilis* spots was equal to 1 on each microarray. Control experiments to validate the normalization procedure are described at our web site.

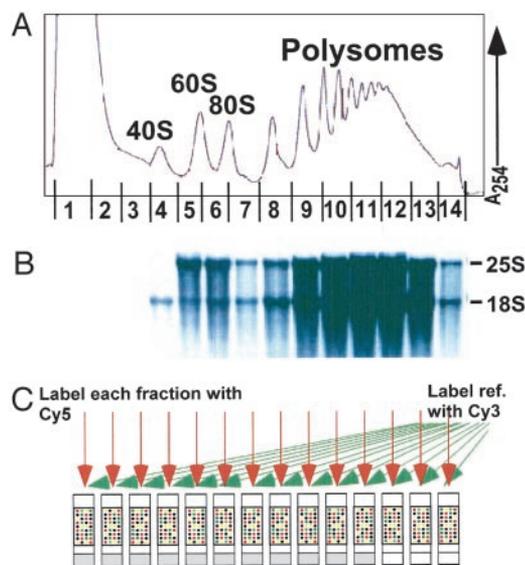
**Peak Fraction and Ribosomal Association Assignment.** There were three independent measurements for each fraction, derived by carrying out the entire protocol, from yeast culture through microarray hybridization, in triplicate. For each gene, the three values for each fraction (expressed as percentage of the total RNA for that gene found in a given fraction) were averaged, and the fraction (between fractions 6 and 14) with the highest average value was defined as the peak fraction.

A bootstrap method (23) was used to assess the accuracy in determination of the peak fraction for each gene. Accordingly, the residuals from each averaged value, from all 14 fractions, were pooled together and reassigned back to the 14 fractions at random to create a bootstrap data set. One thousand bootstrap data sets were made for each profile, and the peak fraction in each bootstrap data set was determined analogously to the initial data set. Taking the 2.5–97.5% quantiles from the 1,000 bootstrap peaks gives a 95% bootstrap confidence interval of the fraction with the most RNA. For 739 genes, the 95% bootstrap confidence interval was contained in a single fraction, indicating a highly precise assignment of peak fraction, and for 2,128 genes the 95% confidence interval constrained the peak to two adjacent fractions. The distributions of all bootstrap peaks for all genes are available at our web site.

The number of ribosomes per transcript in fractions 1–11 was deduced directly from the peaks in the OD<sub>254</sub> profile (Fig. 1A). For fractions 12–14, which corresponded to portions of the OD<sub>254</sub> profile that lacked single ribosome resolution, the number of ribosomes per transcript was estimated by a logarithmic extrapolation from the clearly defined peaks (24, 25). In all cases, the  $R^2$  value of the logarithmic curve over the defined region was >0.995. Estimates for the number of ribosomes in fraction 12 were confirmed by site-specific cleavage of several mRNA–polysome complexes derived from this fraction followed by Northern analysis to determine the number of ribosomes associated with each cleavage product; in all cases the sum of the ribosomes on the cleavage products was similar to the estimated number of ribosomes for the uncleaved transcript, thereby confirming the estimated number of ribosomes for the full-length mRNA (Y.A., unpublished data). Because none of the 739 genes described above (and only 26 of the 2,128-gene set) had polysome profiles with peaks in fractions 13 or 14, the conclusions are not affected significantly by error in estimating the number of ribosomes in those fractions.

## Results

**Whole-Genome Polysomal Profiles.** The standard approach used to assess translation *in vivo* has been the analysis of polysome profiles after treatment with cycloheximide to trap elongating ribosomes. Free mRNAs and those with varying numbers of bound ribosomes can be separated on sucrose gradients. The bound ribosomes represent a steady-state balance between the steps of initiation, elongation, and termination, i.e., faster initiation leads to more bound ribosomes, and faster elongation and termination lead to

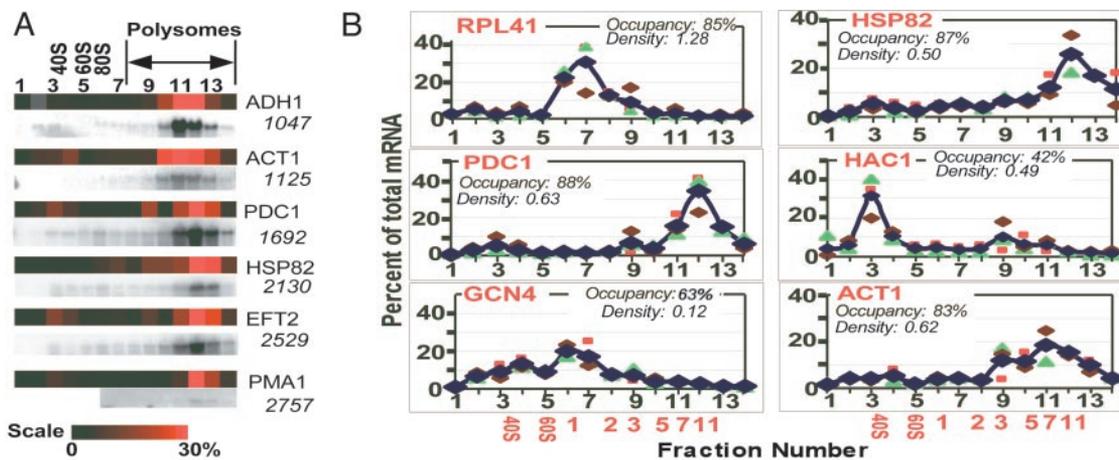


**Fig. 1.** Obtaining polysome profiles by microarray analysis. (A) Representative absorbance profile for RNA separated by velocity sedimentation through a 10–50% sucrose gradient (see *Materials and Methods*). The positions of the 40S, 60S, 80S, and polysomal peaks are indicated. The expected peaks for the tRNAs and other small RNAs in fractions 1 and 2 are obscured by the high absorbance, presumably from proteins and detergent used in the preparation. (B) RNA was extracted from each of the 14 fractions, and an aliquot from each fraction was subjected to electrophoresis through a formaldehyde gel, transferred to a nylon membrane, and stained with methylene blue. As expected, 25S and 18S ribosomal RNAs were the prominent species (markers not shown). (C) RNA aliquots from each fraction were reverse-transcribed to incorporate Cy5-dUTP, and a common reference (unfractionated RNA) was reverse-transcribed in the presence of Cy3-dUTP. The resulting labeled cDNAs from each fraction and the common reference were mixed and then hybridized to a DNA microarray containing all yeast ORFs, as predicted from the genome sequence (*Saccharomyces* Genome Database). Known concentrations of *B. subtilis* mRNAs were added to each sample immediately at collection to allow normalization of RNA levels and comparison between fractions (see *Materials and Methods*).

fewer bound ribosomes. The changes in overall and mRNA-specific polysome profiles observed for specific mutants and after changing growth conditions are typically consistent with models and expectations based on other metabolic data, suggesting that the profiles provide a reasonable picture of the translational status of the cell (e.g., refs. 26–29 and ¶).

We sought to obtain high-resolution polysome profiles for nearly all yeast mRNAs under a single condition, exponential growth on rich medium with glucose as the carbon source. RNA from BY4741 cells grown at 30°C was fractionated by sucrose-gradient sedimentation (Fig. 1). The assignment of OD<sub>254</sub> peaks corresponding to the 40S and 60S subunits and to intact ribosomes was confirmed by gel electrophoresis and staining of aliquots from each fraction to

¶It is not clear, however, whether there are secondary effects on translation after cycloheximide addition. To minimize the possible effects we immediately cooled and harvested the cells after cycloheximide addition. To investigate the possibility of additional initiation events after cycloheximide addition, we quantified [<sup>35</sup>S]Met incorporation prior to and subsequent to cycloheximide addition. Initiation was inhibited at least 99%, but determining a limit to the maximal number of initiation events after cycloheximide addition was not possible, because the intracellular pool size of unlabeled Met and background are not known. Further varying the cycloheximide concentrations (0.01–0.25 mg/ml) had no effect on the overall polysome profile, providing no indication of complicating secondary effects or additional initiation events under our preparation conditions. The conclusions described herein would be unaffected even in the extreme case that each mRNA was to have a ribosome added subsequent to cycloheximide treatment. Finally, effects of cycloheximide on translation termination have been reported in mammalian systems, which give considerable differences in the polysomes response to varying concentrations of cycloheximide (refs. 30–32 and data not shown). Mapping studies suggest that the termination effect is not significant for the results presented herein (Y.A., F. E. Boas, P.O.B., and D.H., unpublished data).



**Fig. 2.** (A) Polysomal profiles from DNA microarray analysis and Northern blots. (Upper) Profiles obtained from the microarray analysis. (Lower) Profiles from Northern blots. For each gene, the sample intensity in each fraction has been color-coded and aligned horizontally (color scale shown below). Gray squares represent values that did not pass the quality selection criteria; to allow normalization these fractions were assigned to have the averaged intensity of their neighbors. For PMA1, fractions 1–5 were not analyzed by Northern blot. ORF length is indicated below each gene. (B) Sample polysomal profiles from triplicate DNA microarray analyses. The smaller symbols represent data from individual experiments, and the blue diamonds connected by a line are the average values for each fraction. The percentage of total mRNA for the given gene in a given fraction is plotted. Black numbers on the x axis are the fraction numbers, and the red numbers indicate the number of bound ribosomes. Ribosome occupancy and ribosome density values (see *Discussion*) are indicated for each gene. Analogous plots for all genes are available at our web site.

identify 18S and 25S RNA from the 40S and 60S ribosomal subunits, respectively (Fig. 1B).

Fourteen fractions were collected from each sucrose gradient as depicted schematically in Fig. 1. Messenger RNA in each of the fractions was labeled with the fluorescent dye Cy5 via reverse transcription with oligo(dT) primers, to label poly(A)-containing RNA, and each sample was mixed with a standard reference sample prepared by reverse transcription of total RNA in the presence of Cy3-dUTP and then hybridized to a microarray containing all  $\approx 6,000$  yeast ORFs.

There was one critical modification from previous microarray protocols: the inclusion of an external standard in both the experimental and control samples to allow quantitative comparison between different samples. These standards control for differences between fractions in recovery, or efficiency of the purification, labeling, hybridization, or scanning procedures. In microarray analysis, it is typically assumed that the overall amount of mRNA in the different samples is constant, and a normalization is carried out by setting the median fluorescence ratio to 1 for each microarray. However, each fraction across a sucrose gradient contains different amounts of mRNA. Because the basis of the translation experiment is to compare the amount of mRNA across these fractions, external standards are required. More generally, such normalization allows absolute comparison of changes in mRNA levels between experimental samples (9).

We added a normalization mixture to each fraction immediately after collection and to the control RNA before reverse transcription and dye labeling. This mixture contained five mRNAs derived from *B. subtilis* that do not cross-hybridize with yeast cDNA, and multiple copies of cDNAs corresponding to these RNAs were included on the arrays. As equal amounts of the control mRNAs were added to each fraction, the observed ratios from the control spots were normalized to 1 for each microarray, and this normalization factor was applied to the rest of the  $\approx 6,000$  spots on the microarray. This normalization protocol was shown previously to allow comparison between samples in global determination of mRNA-decay rates (9).

As a first test of the accuracy of this method in tracing mRNA abundances across the polysome profile, we compared the profiles obtained from our first set of 14 microarrays with profiles we

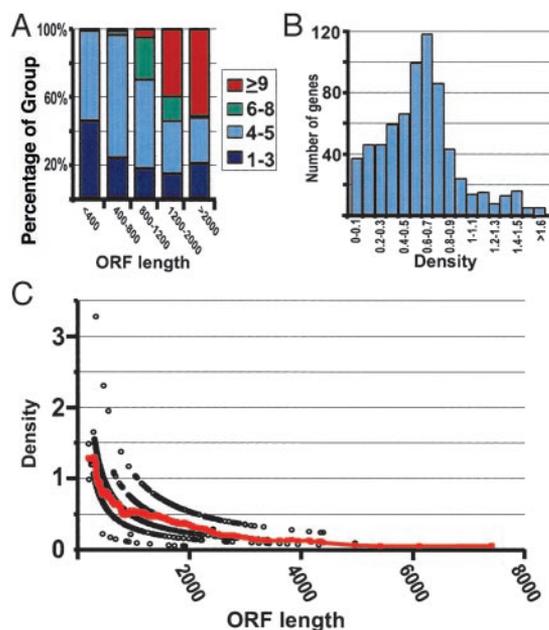
obtained by traditional Northern blot analysis (Fig. 2A). We also compared our microarray profiles with published profiles obtained under similar growth conditions for GCN4, RPL41, RPL30, PYK1, HSP26, RPL28, PGK1, HAC1, RPL1, RPL2, RPL19, and ICY2 (15, 27, 33–35). In all cases the Northern and microarray profiles gave the same peak fraction and had overall patterns that agreed well (Fig. 2A and data not shown).

To determine the reproducibility of our global polysomal analysis, the procedure was carried out three times, each with an independent RNA preparation and each analyzed with 14 microarrays. Spots with abnormal morphology or with fluorescent signals indistinguishable from background were excluded from the analysis, as is standard in microarray analysis (see *Materials and Methods* for details). For 5,701 genes profiles were obtained with data for at least 11 of the 14 fractions in two of the replicates, and for 4,283 genes data were obtained for at least 11 of the 14 fractions in all three replicates. Fig. 2B shows the triplicate microarray polysome profiles for some of the mRNAs that have been analyzed by Northern blot in this or previous studies (15, 33, 34). Polysome data for the remainder of the mRNAs are available at our web site. The overall pattern is reproduced well for most mRNAs; statistical measures of reproducibility are included with each mRNA and are available at our web site.

## Discussion

We have carried out a global analysis of the mRNA translational state in rapidly growing yeast. The reproducibility of the polysome profiles from our microarray data and the agreement with Northern blots suggest that these profiles can be used to ascertain global and specific properties of translation. These data increase by more than an order of magnitude the number of mRNAs species for which such detailed information about their translational status is available.

**Analysis of Ribosome Occupancy.** A key parameter of the translational status of each gene is the fraction of its transcripts engaged in translation. We refer to this as the “ribosome occupancy.” To obtain the ribosome occupancy for each gene, we summed the amount of that gene’s mRNA migrating in fractions containing



**Fig. 3.** Number and density of ribosomes associated with mRNAs. (A) The selected 739 genes (see *Results*) were grouped according to ORF length. Each group contained at least 85 genes. For each length group, the fraction of genes in that group containing one to three ribosomes (dark blue), four to five ribosomes (light blue), six to eight ribosomes (green), and nine or more ribosomes (red) is shown. (B) The number of genes as a function of ribosome density (average =  $0.64 \pm 0.31$  ribosomes per 100 nts). (C) Ribosome density as a function of ORF length (open circles). The red line indicates moving average density value (window of 50). The ordering of the density values (open circles) in apparent hyperbolic curves arises, because the ORF lengths are divided by quantized values (whole number of ribosomes).

ribosomes (fractions 6–14) and compared that to the total amount of that mRNA.<sup>||</sup>

The average ribosome occupancy for the 5,701 genes is 71% (SD 8.1%) (see Fig. 4, which is published as supporting information on the PNAS web site, [www.pnas.org](http://www.pnas.org), and our web site). This suggests that for most genes, most mRNA molecules are associated with ribosomes under conditions of rapid growth, and abundant stored pools of nontranslationally engaged, polyadenylated mRNAs are uncommon or absent. This could be because the recruitment process is highly efficient and/or because mRNAs that are not recruited are degraded rapidly. In addition, integration of the 254-nm absorbances of polysome profiles from six independent experiments (e.g., Fig. 1A) indicates that under these growth conditions most of the ribosomes ( $85 \pm 5\%$ ) are also engaged in translation. Efficient use of the ribosomes and mRNAs might be expected for yeast growing in rich medium; these cells grow rapidly under these conditions and are therefore programmed to synthesize new proteins rapidly and efficiently.

For >99% of the 5,701 mRNA species analyzed, the majority of mRNA molecules were associated with one or more ribosomes (see Fig. 4 and our web site). A small group of genes have their transcripts predominantly unassociated with ribosomes (listed in Table 1, which is published as supporting information on the PNAS

<sup>||</sup>mRNAs potentially sedimenting with the 40S subunit (fractions 3 and 4) were not considered to be translationally active, because they cannot be distinguished from mRNAs associated with other large RNA–protein complexes. Thus, the ribosome occupancy values may underestimate the number of molecules engaged in some step of translation. On the other hand, some of the mRNAs associated with a single ribosome could have accumulated after the cycloheximide treatment; rapid coding was employed to minimize such effects. We estimate, based on microarray polysome profiles and measurements of mRNA abundance (9), that 14% of all mRNA molecules are associated with a single ribosome.

web site, and our web site). We searched for possible sequence motifs upstream or downstream to the coding regions that may be common regulators for the low ribosomal occupancy of these mRNAs, but we failed to identify any motifs that were significant compared with random sets with the same number of genes. Functional analysis of this group did not reveal overrepresentation of any class. However, this group includes HAC1, which is known to be translationally controlled (15, 36), supporting the possibility that for at least some of these mRNAs the relatively low ribosomal occupancy may be a manifestation of translational regulation by yet-to-be-identified mechanisms.

**Analysis of the Number of Bound Ribosomes.** In the simplest model, the number of ribosomes associated with an mRNA molecule would be directly proportional to the length of the coding sequence, because as the length of the coding sequence increases each ribosome that begins translating will have a correspondingly longer transit time until it reaches the termination codon. To test this model we determined for each mRNA the polysomal fraction of the sucrose gradient that contained the most mRNA molecules: the “peak fraction.” For most of the fractions, the number of associated ribosomes was readily determined by simply counting the absorbance peaks in the polysome profile (Fig. 1A). For the fastest-sedimenting fractions, for which single ribosome resolution was not obtained, the value was estimated by extrapolation from the well resolved region of the gradient. Independent assays support the assignments of the number of bound ribosomes (see *Materials and Methods*).

To limit noise and possible artifacts, we focused this analysis on only those mRNAs that could be assigned a peak fraction with high confidence using a bootstrap analysis. For 739 mRNAs, the peak-fraction assignments (with  $\geq 95\%$  bootstrap confidence interval) was confined to one fraction. These mRNAs were therefore selected for the analyses presented below. \*\* Analyzing larger sets of 2,128 genes, for which the 95% confidence interval for the peak fraction included two adjacent fractions, and of all 5,701 genes did not change the conclusions presented below (see Fig. 5, which is published as supporting information on the PNAS web site, and our web site).

The number of ribosomes bound to each transcript indeed tends to increase as the length of the translated sequence increases (Fig. 3A). The 739 different mRNAs were grouped by the length of the coding sequence; the distribution with respect to the number of bound ribosomes is shown for each group. The shortest set of mRNAs (<400 nts) was predominantly associated with five or fewer ribosomes. As the length of coding sequence increased, the fraction of genes associated with six or more ribosomes increased. Most of the mRNAs with the longest coding sequences (>2,000 nts) were associated with nine or more ribosomes. Surprisingly, none of these 739 mRNAs appeared to be associated with >15 ribosomes (i.e., none peaked in fractions 13 or 14), although many are long enough to accommodate many more than 15 ribosomes (but see *Analysis of Ribosome Density* below).

Transcripts for GCN4, ICY2, and CPA1, which are translationally controlled mRNAs, peaked in the monosome fraction (15, 33, 37). We therefore paid special attention to other mRNAs that peak in the monosome fraction, identifying from the 2,128-gene set a list of 53 genes with transcripts that were predominantly monosome-associated (see Table 2, which is published as supporting information on the PNAS web site, and our web site). Except for RPL41, which has a coding sequence so short (78 nts) that it can only accommodate a single ribosome (34), the remainder may be

\*\*This selection biases the sample toward mRNAs with prominent peaks rather than broad distributions across the gradient (e.g., compare ACT1 and PDC1 in Fig. 2B). The distribution width may reflect underlying differences in translational usage. For example, mRNAs with varying translational rates through the cell cycle would appear more broadly distributed in polysome profiles of mRNA obtained from unsynchronized cells.

translationally regulated or may have unusual properties that lead to constitutively low ribosome association. Because both GCN4 and CPA1 are regulated through an upstream ORF, we screened these genes for enrichment of upstream ORFs in their 5'-noncoding region. A similar number of AUG codons was found in the 140 nts upstream to the start codon for these genes and for three random lists of genes. A search for other sequence or functional properties did not yield any significant pattern, suggesting that other properties such as structural elements may be involved in their regulation.

**Analysis of Ribosome Density.** Two mRNAs that differ in length but have identical initiation, elongation, and termination rates would have different numbers of bound ribosomes but the same ribosome density (number of ribosomes per unit ORF length).<sup>††</sup> To allow direct comparison of the steady-state outcome of the three translational steps for different mRNAs we normalized the number of bound ribosomes by the ORF length of the mRNA. As described above, the peak fraction was used to obtain the number of bound ribosomes.

The average ribosome density for the 739 mRNAs with precisely defined peak fractions was 0.64 per 100 nts (SD = 0.31) (Fig. 3B). This corresponds to one ribosome per 156 nts. If each eukaryotic ribosome spans  $\approx 35$  nts of the mRNA (38), the average density that we observed would be  $\approx 1/5$  of the maximal packing density. Thus, ribosomes are far from maximally packed along mRNAs for the vast majority of mRNAs, and elongation and termination occur faster than initiation. A similar ribosome density distribution was observed in the analyses of two larger, less stringently filtered data sets representing 2,128 or 5,701 genes (see Fig. 5 and our web site). Overall, the results strongly suggest that translation of all or nearly all mRNAs under conditions of rapid growth is limited by initiation or the recycling of ribosomal subunits in preparation for initiation.

**An Underlying Relationship Between the Length of a Coding Sequence and Ribosome Density.** The 739 mRNAs with precisely defined peak fractions had ribosome densities varying from 0.03 to 3.3 ribosomes per 100 nts. We searched for features of these mRNAs that might correlate with the variation in ribosome density, and thus provide clues to its origins. No significant correlation was found between ribosome density and mRNA abundance, codon usage, or half-life (see Fig. 6, which is published as supporting information on the PNAS web site, and our web site). A strong inverse correlation, however, was observed between ribosome density and ORF length (Spearman rank correlation of  $-0.78$ ). As the ORF length increased, the ribosome density dropped from an average of 1.2 ribosomes per 100 nts for ORFs shorter than 400 nts to an average of 0.14 ribosomes per 100 nts for ORFs longer than 3,600 nts (Fig. 3C). A similar trend, with an even stronger correlation, was observed when the density values were calculated based on weighted average of the signal in fractions 6–14 rather than a peak fraction (see Fig. 7, which is published as supporting information on the PNAS web site, and our web site).

We searched first for possible spurious origins for the observed correlation. To exclude possible systematic error of underestimating the number of ribosomes in the fastest migrating fractions (fractions 12–14), we redid the analysis using extreme error limits for the relationship between peak fractions and the number of bound ribosomes (see Fig. 7 and our web site). Doubling the estimated number of ribosomes for fraction 12, tripling the estimate for fraction 13, and quadrupling the esti-

mate for fraction 14 did not vitiate the observed correlation (see Fig. 7 and our web site). Additionally, analyzing only those mRNAs sedimenting in the well resolved fractions of the gradient (peak at fractions 6–11) or only the highly abundant ribosomal protein mRNAs revealed the same correlation (see Fig. 7 and our web site).

Nonspecific degradation of mRNA could also produce an inverse correlation between apparent ribosome density and predicted ORF length. Such an artifactual correlation due to RNA degradation should be recognizable as a difference between the sedimentation profile of the full-length mRNA, observed by Northern analysis, and the sedimentation profile of the putative cleavage products, detected by the microarray analysis. We performed Northern analysis for six mRNAs varying in ORF length from 1,047 to 2,757 nts and compared the sedimentation profile of the full-length mRNA to the profile obtained by the microarray analysis (Fig. 2A). For each of these mRNAs, the sedimentation profiles from the Northern and microarray analyses were similar, and the peak fraction was the same. Hence, the correlation apparent in Fig. 3C cannot be attributed to an artifact of mRNA degradation.

We considered the possibility that the inverse correlation might be a consequence of a nonrandom association between coding sequence lengths and physiological functions of the encoded product, resulting in a biased relationship between expression and ORF length. For example, mRNAs encoding ribosomal proteins are relatively short, thus preferential translation of these mRNAs might produce a spurious correlation with ORF length. However, exclusion of this subset of mRNAs from our analysis did not affect the overall correlation. Furthermore, the inverse correlation is observed within functional subsets of mRNAs such as those encoding ribosomal proteins or proteins targeted to membranes or mitochondria (see Fig. 8, which is published as supporting information on the PNAS web site, and our web site).

The analyses and control described above suggest that mechanistic origins for the inverse correlation between ORF length and ribosome density must be sought. We suggest three general classes of mechanisms: initiation, elongation, and termination effects. Translation initiation is stimulated by the presence of the 3'-poly(A) tail, apparently from an interaction between the poly(A)-binding protein bound to the 3'-poly(A) tail and eIF4G bound at the 5' UTR (39). This stimulatory interaction may be less efficient with longer coding sequences because the probability of 5'- and 3'-end association decreases. Alternatively or in addition, mRNAs with more structure may have lower initiation rates (20); if longer mRNAs are more prone to or happen to have self-structure, they may initiate translation less efficiently.

A second potential mechanism is limited processivity of ribosomes during elongation. Incomplete processivity has been observed in *Escherichia coli*, where it is estimated that  $<75\%$  of translation initiation events ultimately result in a complete protein (40–42). Similar attrition during translation has been proposed to occur in mammalian cells (43). Modeling limited processivity with uniform initiation, elongation, and termination rates revealed that processivity of 0.994 per elongation step (i.e., 0.6% probability of dissociation at each elongation step) best fit our data (not shown).

Finally, limiting termination rates would lead to accumulation of ribosomes at the 3' end of the mRNA. This accumulation of ribosomes could preferentially increase the overall density for mRNAs with short ORFs (see also ††).

These three distinct models make different, testable predictions for distribution of ribosomes along mRNAs. We have used this different physical property to test these models for the basis for the inverse correlation (Y.A., F. E. Boas, P.O.B., and D.H., unpublished data).

<sup>††</sup>This holds for most models. However, one can envision extreme cases in which densities will vary between two mRNAs with similar initiation, elongation, and termination rates. For example, if termination were severely rate-limiting, then ribosomes would tend to accumulate at the 3' end of the message, and mRNAs with shorter coding sequences would tend to have higher ribosome densities than mRNAs with longer coding sequences even if the rates of initiation, elongation, and termination were the same for the two.

## Summary and Perspective

We used DNA microarrays to investigate the process of protein synthesis on a genome-wide scale in yeast. Critical for our analysis was the use of external normalization controls to allow fractions across polysomal gradients to be compared and profiles for individual mRNAs to be constructed. Comparison to Northern blots and the reproducibility of the triplicate data suggest that the translational behavior of most if not all mRNAs can be ascertained by this approach.

Under these conditions of rapid growth in rich medium, the vast majority of mRNAs were in the translational pool. This is consistent with the gene expression program of the cell tuned for rapid synthesis of proteins as necessary for rapid growth.

Our analysis suggests that ribosomes are well spaced on nearly all mRNAs, suggesting that initiation is the limiting step in translation of all or nearly all yeast mRNAs. The “outliers,” i.e., genes with properties that fell well outside of the norm, are good candidates for distinct mechanisms or extreme variations in regulation. For translation, mRNAs that are predominantly in the untranslated pool or in the monosome fraction provide such candidates (15, 33, 36). By using a systematic, global, and quantitative approach to analysis of the translational behavior of mRNAs, we could recognize trends that would not be apparent from studies of a limited set of genes. In this way we found an unexpected inverse correlation between the length of the coding region in an mRNA and the density of ribosomes associated with it. The reciprocal relationship between coding sequence length and ribosome density may prove to be a general characteristic of eukaryotic translation. Although a systematic analysis of this kind has not yet been carried out for other eukaryotes, circumstantial evidence is consistent with such a relationship. For example, Cataldo *et al.* (44) analyzed the polysome profiles of 18 mouse mRNAs by using Northern blots. Plotting their data reveals an inverse correlation (Spearman rank correlation of  $-0.84$ ) between ORF length and density (see Fig. 9, which is published as supporting information on the PNAS web site, and our web site).

One of the goals of global translational analysis was to provide a clearer quantitative picture of the link between mRNA and protein levels (45). We have used our translational profiles, together with measurements of mRNA copy number, to obtain estimates of protein-synthesis rates (see Table 3, which is published as supporting information on the PNAS web site, and our web site). There are several limitations, however, in obtaining such global estimates. First, there is considerable uncertainty from accumulated errors in copy numbers and polysomal positions, especially for low abundance mRNAs. Second, this estimate requires the assumption that elongation rates and termination rates are the same for all mRNAs, such that protein-synthesis rates are directly proportional to the number of bound ribosomes. Finally, there are only limited global data for protein levels (46, 47) and no extensive information directly about protein-synthesis rates. Future work is needed to assess and improve the quantitative accuracy of microarray data to determine the extent and character of variation in elongation and termination for different mRNAs and to provide direct quantitative measures of protein synthesis for large numbers of proteins.

The current study provides background and methodologies for future studies of cellular translation at the global and individual level. Many studies have provided evidence for translation regulation in response to environmental changes (e.g., reviews in refs. 48–51). Global analyses of different growth conditions will allow exploration of the generality of these translational responses and the discovery of novel regulated mRNAs. In addition, exploring the role of specific translation factors by carrying out a similar analysis in yeast strains with mutations in the genes that encode them will expand our knowledge of their global and gene-specific effects on translation.

We thank Ken Kuhn for advice and protocols and Matt Peck, Jimmy Yi-Ming Huang, Arkady Khodursky, Paul Fawcett, and Ed Boas for advice and help with analyses. This work was supported by grants from the National Human Genome Research Institute (to P.O.B.), the American Heart Association and the National Institutes of Health (to D.H.), and The Howard Hughes Medical Institute (HHMI). P.O.B. is an HHMI investigator, and Y.A. is an HHMI research associate.

- 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) *Cell* **102**, 109–126.
- Spellman, P. T., Sherlock, G., Zhang, M. O., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D. & Futcher, B. (1998) *Mol. Biol. Cell* **9**, 3273–3297.
- Holstege, F. C., Jennings, E. G., Wyrick, J. J., Lee, T. I., Hengartner, C. J., Green, M. R., Golub, T. R., Lander, E. S. & Young, R. A. (1998) *Cell* **95**, 717–728.
- DeRisi, J. L., Iyer, V. R. & Brown, P. O. (1997) *Science* **278**, 680–686.
- Winzler, E. A., Shoemaker, D. D., Astromoff, A., Liang, H., Anderson, K., Andre, B., Bangham, R., Benito, R., Boeke, J. D., Bussey, H., *et al.* (1999) *Science* **285**, 901–906.
- Brown, P. O. & Botstein, D. (1999) *Nat. Genet.* **21**, 33–37.
- Velculescu, V. E., Zhang, L., Zhou, W., Vogelstein, J., Basrai, M. A., Bassett, D. E., Jr., Hieter, P., Vogelstein, B. & Kinzler, K. W. (1997) *Cell* **88**, 243–251.
- Clark, T. A., Sugnet, C. W. & Ares, M., Jr. (2002) *Science* **296**, 907–910.
- Wang, Y., Liu, C. L., Storey, J. D., Tibshirani, R. J., Herschlag, D. & Brown, P. O. (2002) *Proc. Natl. Acad. Sci. USA* **99**, 5860–5865.
- Diehn, M., Eisen, M. B., Botstein, D. & Brown, P. O. (2000) *Nat. Genet.* **25**, 58–62.
- Marc, P., Margeot, A., Devaux, F., Blugeon, C., Corral-Debrinski, M. & Jacq, C. (2002) *EMBO Rep.* **3**, 159–164.
- Tenenbaum, S. A., Carson, C. C., Lager, P. J. & Keene, J. D. (2000) *Proc. Natl. Acad. Sci. USA* **97**, 14085–14090.
- Johannes, G., Carter, M. S., Eisen, M. B., Brown, P. O. & Sarnow, P. (1999) *Proc. Natl. Acad. Sci. USA* **96**, 13118–13123.
- Zong, Q., Schummer, M., Hood, L. & Morris, D. R. (1999) *Proc. Natl. Acad. Sci. USA* **96**, 10632–10636.
- Kuhn, K. M., DeRisi, J. L., Brown, P. O. & Sarnow, P. (2001) *Mol. Cell. Biol.* **21**, 916–927.
- Mikulitis, W., Pradet-Balade, B., Habermann, B., Beug, H., Garcia-Sanz, J. A. & Mullner, E. W. (2000) *FASEB J.* **14**, 1641–1652.
- Brown, V., Jin, P., Ceman, S., Darnell, J. C., O'Donnell, W. T., Tenenbaum, S. A., Jin, X., Feng, Y., Wilkinson, K. D., Keene, J. D., *et al.* (2001) *Cell* **107**, 477–487.
- 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) *Science* **274**, 546–567.
- Hershey, J. W. B. & Merrick, W. C. (2000) in *Translational Control of Gene Expression*, eds. Sonenberg, N., Hershey, J. W. B. & Mathews, M. B. (Cold Spring Harbor Lab. Press, Plainview, NY), pp. 33–88.
- McCarthy, J. E. G. (1998) *Microbiol. Mol. Biol. Rev.* **62**, 1492–1553.
- Spahn, C. M., Beckmann, R., Eswar, N., Penczek, P. A., Sali, A., Blobel, G. & Frank, J. (2001) *Cell* **107**, 373–386.
- Ausubel, F. M., Brent, R., Kingston, R. E., Moore, D. D., Seidman, J. G., Smith, J. A. & Struhl, K. (1994) *Current Protocols in Molecular Biology* (Greene, New York).
- Efron, B. & Tibshirani, R. J. (1993) *An Introduction to the Bootstrap* (Chapman & Hall, New York).
- Noll, H. (1967) *Nature* **215**, 360–363.
- Rickwood, D. (1992) *Preparative Centrifugation: A Practical Approach* (Oxford Univ. Press, New York).
- Ashe, M. P., De Long, S. K. & Sachs, A. B. (2000) *Mol. Biol. Cell* **11**, 833–848.
- Dickson, L. M. & Brown, A. J. (1998) *Mol. Gen. Genet.* **259**, 282–293.
- Meyuhas, O., Bibberman, Y., Pierandrei-Amaldi, P. & Amaldi, F. (1996) in *A Laboratory Guide to RNA: Isolation, Analysis and Synthesis*, ed. Kreig, P. A. (Wiley-Liss, New York).
- Mathews, M. B., Sonenberg, N. & Hershey, J. W. B. (2000) in *Translational Control of Gene Expression*, eds. Sonenberg, N., Hershey, J. W. B. & Mathews, M. B. (Cold Spring Harbor Lab. Press, Plainview, NY), pp. 1–32.
- Godchaux, W., III, Adamson, S. D. & Herbert, E. (1967) *J. Mol. Biol.* **27**, 57–72.
- Rajalakshmi, S., Liang, H., Sarma, D. S., Kisilevsky, R. & Farber, E. (1971) *Biochem. Biophys. Res. Commun.* **42**, 259–265.
- Oleinick, N. L. (1977) *Arch. Biochem. Biophys.* **182**, 171–180.
- Tzamaras, D., Roussou, I. & Thireos, G. (1989) *Cell* **57**, 947–954.
- Yu, X. & Warner, J. R. (2001) *J. Biol. Chem.* **276**, 33821–33825.
- LaGrandeur, T. & Parker, R. (1999) *RNA* **5**, 420–433.
- Rueggsegger, U., Leber, J. H. & Walter, P. (2001) *Cell* **107**, 103–114.
- Werner, M., Feller, A., Messenguy, F. & Pierard, A. (1987) *Cell* **49**, 805–813.
- Wolin, S. L. & Walter, P. (1988) *EMBO J.* **7**, 3559–3569.
- Sachs, A. B. (2000) in *Translational Control of Gene Expression*, eds. Sonenberg, N., Hershey, J. W. B. & Mathews, M. B. (Cold Spring Harbor Lab. Press, Plainview, NY), pp. 447–466.
- Jorgensen, F. & Kurland, C. G. (1990) *J. Mol. Biol.* **215**, 511–521.
- Manley, J. L. (1978) *J. Mol. Biol.* **125**, 407–432.
- Tsung, K., Inouye, S. & Inouye, M. (1989) *J. Biol. Chem.* **264**, 4428–4433.
- Yewdell, J. W., Anton, L. C. & Bennink, J. R. (1996) *J. Immunol.* **157**, 1823–1826.
- Cataldo, L., Mastrangelo, M. A. & Kleene, K. C. (1999) *Mol. Hum. Reprod.* **5**, 206–213.
- Pradet-Balade, B., Boulme, F., Beug, H., Mullner, E. W. & Garcia-Sanz, J. A. (2001) *Trends Biochem. Sci.* **26**, 225–229.
- Gygi, S. P., Rochon, Y., Franza, B. R. & Aebersold, R. (1999) *Mol. Cell. Biol.* **19**, 1720–1730.
- Futcher, B., Latter, G. I., Monardo, P., McLaughlin, C. S. & Garrels, J. I. (1999) *Mol. Cell. Biol.* **19**, 7357–7368.
- Pyronnet, S. & Sonenberg, N. (2001) *Curr. Opin. Genet. Dev.* **11**, 13–18.
- Sierra, J. M. & Zapata, J. M. (1994) *Mol. Biol. Rep.* **19**, 211–220.
- Meyuhas, O. (2000) *Eur. J. Biochem.* **267**, 6321–6330.
- Sheikh, M. S. & Fornace, A. J., Jr. (1999) *Oncogene* **18**, 6121–6128.