Large-scale identification of secreted and membrane-associated gene products using DNA microarrays

Maximilian Diehn[1], Michael B. Eisen[2], David Botstein[2] & Patrick O. Brown[1,3]

Membrane-associated and secreted proteins are an important class of proteins and include receptors, transporters, adhesion molecules, hormones and cytokines. Although algorithms have been developed to recognize potential amino-terminal membrane-targeting signals or transmembrane domains in protein sequences, their accuracy is limited and they require knowledge of the entire coding sequence, including the N terminus[1], which is not currently available for most of the genes in most organisms, including human. Several experimental approaches for identifying secreted and membrane proteins have been described, but none have taken a comprehensive genomic approach[2–6]. Furthermore, none of these methods allow easy classification of clones from arrayed cDNA libraries, for which large-scale gene-expression data are now becoming available through the use of DNA microarrays. We describe here a rapid and efficient method for identifying genes that encode secreted or membrane proteins. mRNA species bound to membrane-associated polysomes were separated from other mRNAs by sedimentation equilibrium or sedimentation velocity. The distribution of individual transcripts in the 'membrane-bound' and 'cytosolic' fractions was quantitated for thousands of genes by hybridization to DNA microarrays. Transcripts known to encode secreted or membrane proteins were enriched in the membrane-bound fractions, whereas those known to encode cytoplasmic proteins were enriched in the fractions containing mRNAs associated with free and cytoplasmic ribosomes. On this basis, we identified over 275 human genes and 285 yeast genes that are likely to encode previously unrecognized secreted or membrane proteins.

We isolated RNA from the membrane or cytosolic fractions of cells using standard methods, synthesized fluorescently labelled cDNA from each fraction (Cy5-labelled for membrane-associated mRNA and Cy3-labelled for cytosolic mRNA) and hybridized these to microarrays (Fig. 1). Our results showed that hundreds of genes had Cy5/Cy3 fluorescent ratios that deviated from unity, reflecting enrichment of the corresponding mRNA in one of the two fractions.

To examine more closely the relationship between the observed mRNA distributions and the subcellular localization of the corresponding gene products, we assembled a list of the genes present on our arrays that encoded proteins whose subcellular localization (that is, bound to membranes, secreted, mitochondrial, cytosolic or nuclear) has been empirically determined. The fluorescence ratio distributions for mRNAs encoding products that have been found empirically to be membrane-associated or secreted (blue curve) and cytosolic or nuclear (red curve) are shown (Fig. 2). The distribution of fluorescence ratios for these characterized genes showed two overlapping populations. Similar distributions were obtained for two different organisms using two different fractionation procedures (sedimentation velocity and sedimentation equilibrium), suggesting that the observed distribution reflected the true partitioning of transcripts between the two subcellular compartments. Indeed, earlier studies have also found a reproducible overlap between poly(A)$^+$ RNA from free polysomes and poly(A)$^+$ RNA from membrane-bound polysomes that cannot be attributed to cross-contamination[7,8]. Thus, many mRNAs are not simply partitioned between membrane-associ-

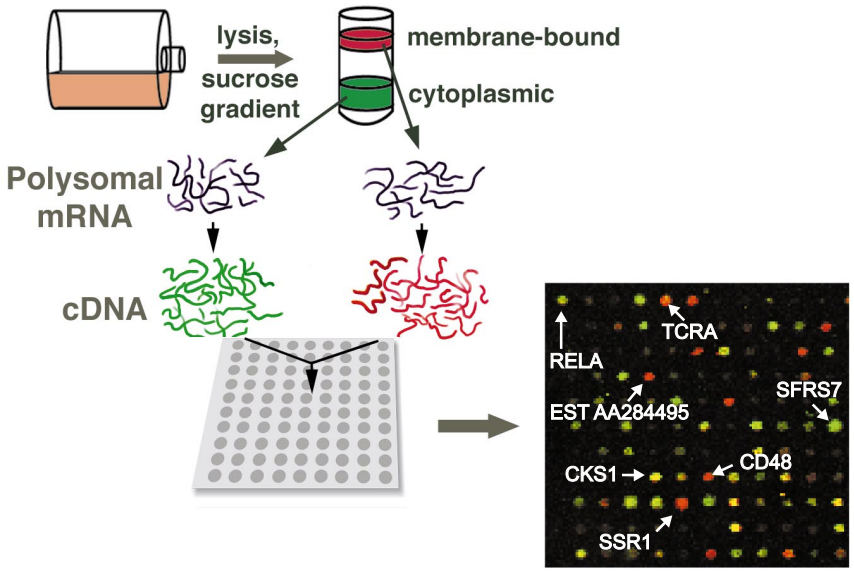

Fig. 1 Procedure for isolating membrane-bound polysomes from cell lines. Jurkat cells were hypotonically lysed, and membrane-bound RNA was separated from free RNA by equilibrium density centrifugation in a sucrose gradient. Total RNA was isolated separately from the fractions containing membrane-bound or free RNA. After linear amplification of mRNA (ref. 20), cDNA was synthesized from membrane-bound and free mRNA and labelled with the fluorescent dyes Cy5 and Cy3, respectively. The cDNAs were hybridized to a DNA microarray and analysed using standard methodology. The subsection of an array pictured shows the identity of some of the spots from a representative experiment.

Departments of [1]Biochemistry and [2]Genetics, and the [3]Howard Hughes Medical Institute, Stanford University School of Medicine, Stanford, California, USA. Correspondence should be addressed to P.O.B. (e-mail: pbrown@cmgm.stanford.edu).

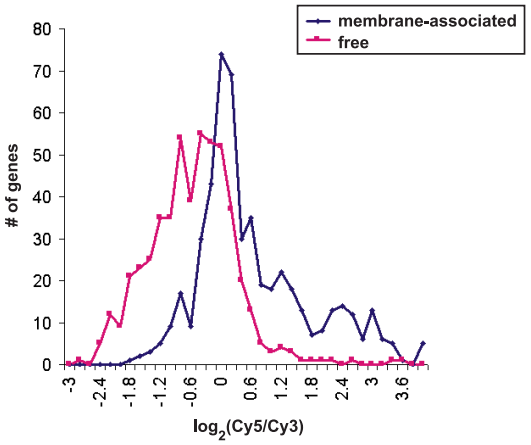

Fig. 2 Distribution of Jurkat T-cell mRNAs encoding proteins with characterized subcellular localization. Genes were classified into two categories: those whose protein products are membrane-associated (transmembrane, secreted or ER/Golgi/vesicle resident; blue curve) and those whose products are cytosolic (or nuclear; red curve). The graphs show the number of genes in each class plotted against the log-transformed (base 2) Cy5/Cy3 ratios in bins of 0.2.

ated and free polysomes. There are likely to be biological explanations (for example, translational regulation or ribosome-independent association of mRNAs with the membrane fraction) for the variability of the association of mRNAs with membranes.

The genes that showed the highest Cy5/Cy3 or Cy3/Cy5 ratios were highly enriched for those encoding membrane/secreted proteins or cytosolic/nuclear proteins, respectively (Fig. 3). A few mRNAs showed enrichment opposite to what would be expected based on the subcellular localization of the encoded protein. Although some of these anomalous observations at the extremes of the Cy5/Cy3 continuum may reflect experimental artefact, we believe that many correctly reflect the subcellular localization of the corresponding transcripts.

An advantage of our approach was that the characterized genes encoding proteins of known subcellular localization provided internal controls that we used to calibrate the relationship between the measured fluorescence ratio and the probability that any gene encodes a secreted or membrane protein. We calculated, using a moving average algorithm, the local percentage of characterized mRNAs encoding membrane-associated proteins as a function of the Cy5/Cy3 ratio (Fig. 4*a*,*b*). The probability that an uncharacterized gene with a given Cy5/Cy3 ratio encodes a membrane or secreted protein can be estimated from the fraction of the characterized genes with similar fluorescence ratios that encode membrane-associated proteins. (This assumes that the prior probability that any uncharacterized gene in the set encodes a membrane-associated or secreted protein is equivalent to the fraction of characterized proteins that have been assigned to this class, an assumption that is supported, at least for yeast, by the very similar frequency of computationally predicted signal peptides and transmembrane domains in the characterized and uncharacter-

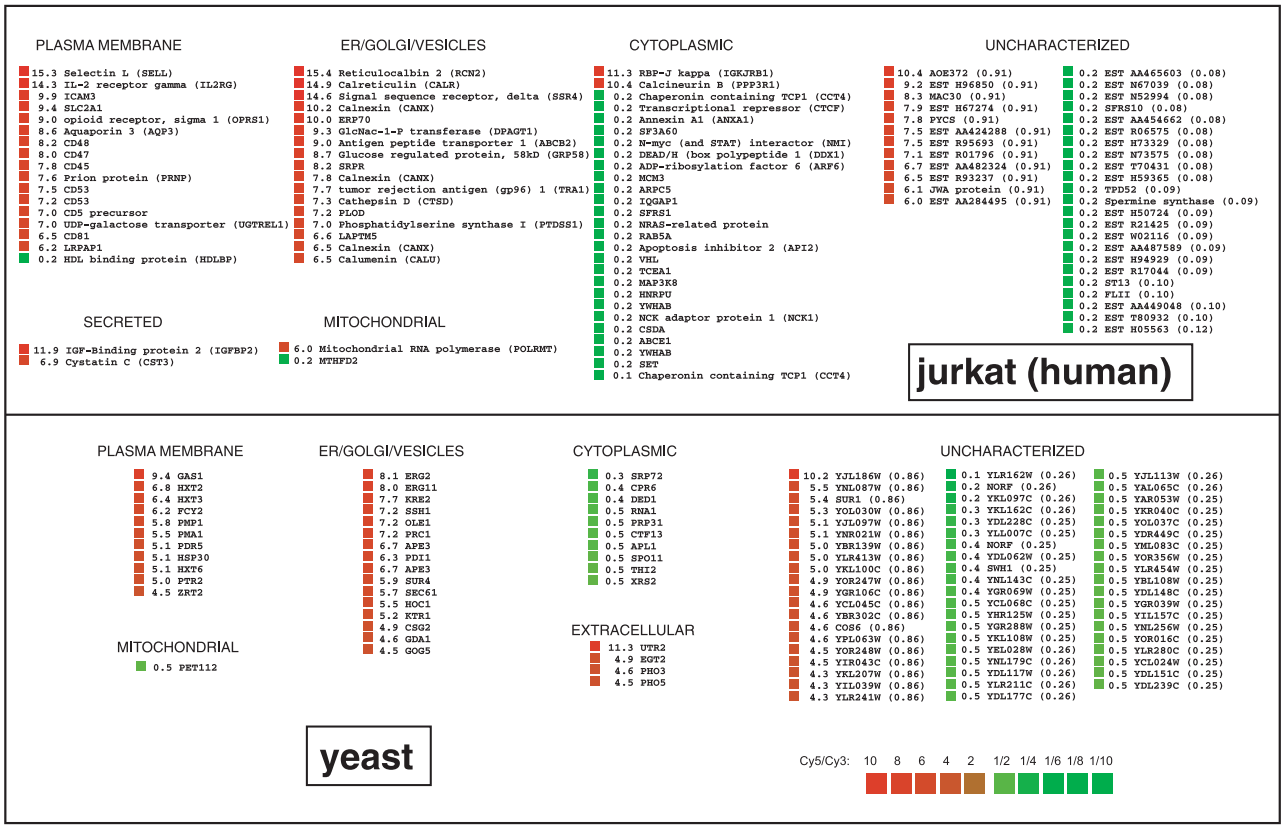

Fig. 3 Subcellular localization of the 50 mRNAs most enriched in the membrane-associated or cytosolic mRNA fractions from Jurkat T cells (top) and yeast (bottom). Subcellular localization data for human and yeast proteins were collated as in Fig. 2. Genes were then divided into the six categories indicated based on the reported localization of their protein products. The coloured squares represent the magnitude of the Cy5/Cy3 ratios, with red and green squares representing strong enrichment in the membrane-associated or cytosolic fractions, respectively. The actual Cy5/Cy3 ratio is listed to the left of the gene name. For uncharacterized genes, the value in parentheses to the right of the gene name represents the estimated likelihood that the gene encodes a secreted or membrane-associated protein

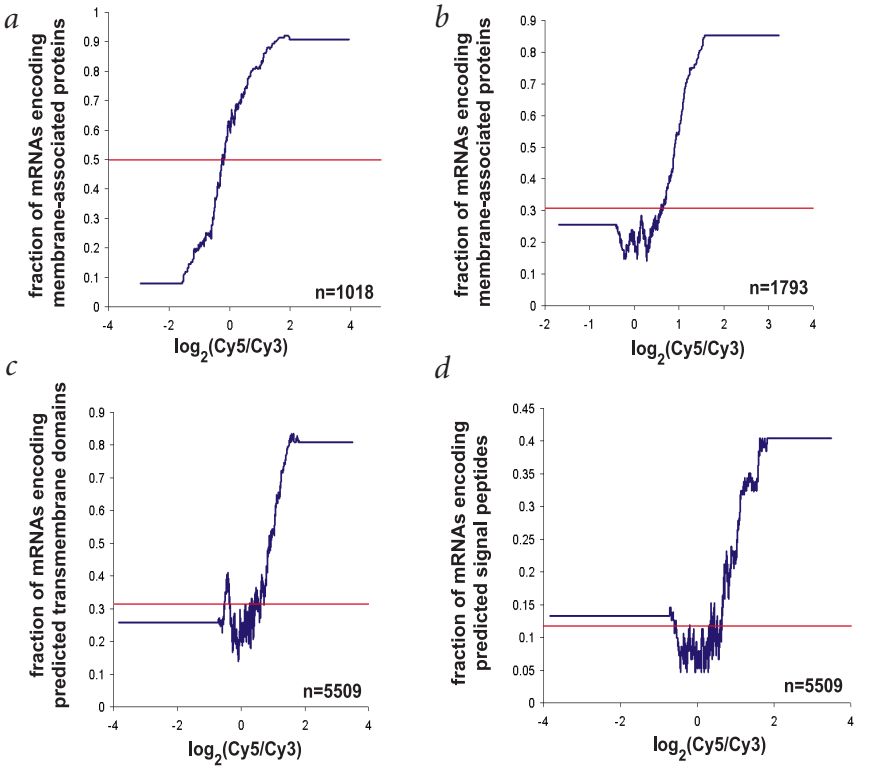

Fig. 4 Moving average analyses of mRNAs encoding experimentally determined or predicted membrane-associated proteins. Subcellular localization data for human and yeast proteins were collected as in Fig. 3. Genes whose mRNAs were well measured in representative experiments in human Jurkat (***a***) or yeast (***b***) cells were classified into two categories: those that encoded membrane-associated (transmembrane, secreted or ER/Golgi/vesicle resident) proteins and those that encoded free (cytosolic or nuclear) proteins. A moving average algorithm with a window size of 151 genes (Jurkat) or 175 genes (yeast) was applied to the data and the percentage of membrane-associated genes in each window was plotted against the log-transformed (base 2) Cy5/Cy3 ratio of the central gene. The horizontal line represents the overall percentage of membrane-associated genes on the microarrays used in the experiments. ***c***, Moving-average analysis as in (*b*) of yeast mRNAs encoding proteins containing one or more predicted transmembrane domains. The prediction algorithm used has been described[9]. The horizontal line represents the overall percentage of genes encoding proteins with predicted transmembrane domains in the set of genes that was assayed. ***d***, Moving-average analysis as in (*b*) of yeast mRNAs encoding proteins containing putative signal peptides. The prediction algorithm used was SignalP (ref. 10) The horizontal line represents the overall percentage of genes encoding proteins with predicted signal peptides in the set of genes that was assayed. n, number of genes in each data set.

ized gene products (see http://genome-www.stanford.edu/mbp/). On the basis of this analysis, more than 85% of the 277 uncharacterized genes (491, including characterized genes) whose transcripts were most highly enriched in the membrane fraction from Jurkat cells and the 289 uncharacterized genes (453, including characterized genes) whose transcripts were most highly enriched in the membrane fraction from yeast cells are expected to encode membrane-associated proteins. Assignments and estimated confidences for all previously uncharacterized genes are available (see http://genome-www.stanford.edu/mbp/). We have independently examined the subcellular localization of five of the unknown gene products by immunohistochemistry using antibodies raised against predicted peptides from these proteins, and confirmed the predicted localization of four of these (for results, see http://genome-www.stanford.edu/mbp/).

We compared our results with the predictions of computational models that attempt to recognize transmembrane domains or signal peptides in protein sequences. Because full coding sequences were not available for most of the human genes represented on our microarrays (which contained mainly cDNAs defined only by ESTs), this analysis was restricted to yeast. We used described algorithms to predict transmembrane domains (H19 method; data from YPD; ref. 9) and signal peptides[10]. mRNAs preferentially recovered in the membrane fraction were enriched for those encoding proteins with predicted transmembrane domains or signal peptides (Fig. 4*c*,*d*). On the other hand, mRNAs preferentially recovered in the cytosolic fraction encoded proteins with fewer predicted transmembrane domains or signal peptides than the average for all the known and hypothetical proteins encoded by the yeast genome. For proteins known to be secreted or membrane-associated, the average prediction score for transmembrane domains or signal peptides was higher for those whose transcripts were highly enriched in the membrane fraction, as assayed using the microarray method (Fig. 5*a*,*b*). Our empirical approach, however, also identified many genes that encoded bona fide membrane-associated or secreted proteins that were not predicted by either computational method (Fig. 5*c*). Thus, our empirical method and the computational methods identified partially overlapping sets of membrane-associated proteins.

The apparent misclassifications of some mRNAs by the microarray method may reflect features of subcellular compartmentalization. For example, of the yeast mRNAs that have been found to encode cytosolic proteins, the two that were most enriched in our membrane fraction were the *ASH1* and *HAC1* transcripts. *ASH1* mRNA has been shown to be asymmetrically localized to the distal tip of daughter buds and to be attached to the cell cortex[11]. *HAC1* mRNA is known to be spliced in the cytoplasm by interaction with Ire1p, an endoplasmic reticular transmembrane protein[12]. *In situ* hybridization shows *HAC1* mRNA in punctate structures throughout the cytoplasm[13], consistent with the distribution of rough endoplasmic reticulum (rER).

Other biological mechanisms might also account for the association of mRNAs encoding cytosolic proteins with cellular membranes. For example, if the nascent N terminus of a cytosolic protein contains a membrane-binding domain (for example, a PH domain) or post-translational modification site (for example, myristoylation), polyribosomes translating this mRNA may be recruited to membranes. Calcineurin B has been shown to associate with the cytoplasmic surface of the plasma membrane[14], and we found that, in Jurkat cells, the mRNA encoding calcineurin B was the most enriched in the membrane fraction of all mRNAs known to encode cytoplasmic proteins. Finally, some mRNAs that encode proteins characterized as cytosolic have alternatively spliced forms that are membrane-associated, as is the case for NADH-cytochrome b5 reductase[15]. Because the microarrays used here do not allow alternative splice forms to be distinguished, such mRNAs may appear enriched in the membrane-associated fraction for biological reasons.

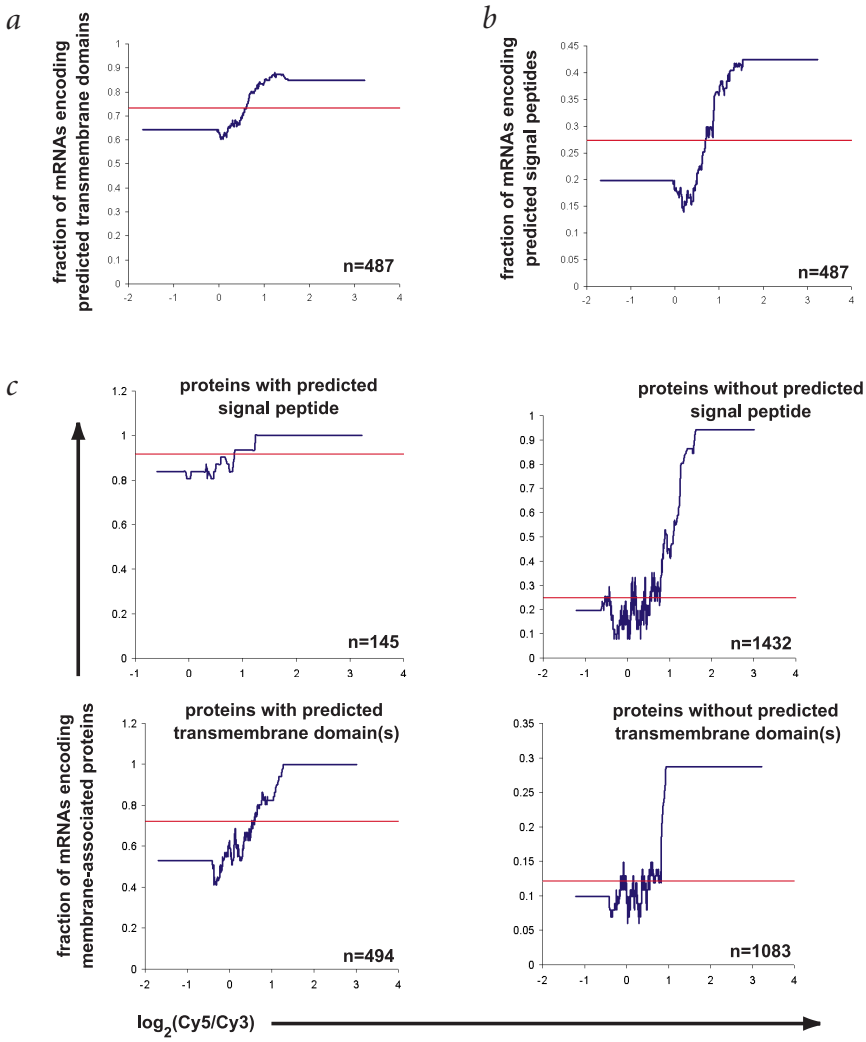

Fig. 5 Moving average analysis of yeast mRNAs encoding empirically determined membrane-associated or cytosolic proteins. Moving average analysis was performed as in Fig. 4, but only on yeast mRNAs with experimentally documented subcellular localization. For (***a***) and (***b***), only genes encoding membrane-associated proteins were considered; the window size was 151 genes. The red lines represent the overall percentage of mRNAs in this set with transmembrane domains or signal peptides. In (***c***), empirically determined membrane-associated and cytosolic proteins were considered together. The data were broken down into the indicated subsets and the percentage of mRNAs encoding membrane-associated proteins was calculated using a moving average algorithm. The window sizes used were 31, 51, 51 and 101 genes, from left to right, top to bottom. n, number of genes in each data set.

We were able to identify hundreds of previously uncharacterized mRNAs that are likely to encode membrane-associated proteins. In both cases, these results are from the analysis of a single cell type under a single condition—many genes represented on our microarrays could not be assessed because they were not expressed at a sufficient level in these cell populations. Because the method is cumulative, however, fractionation and analysis of other cell types and conditions should allow the rapid classification of thousands of unknown genes. Furthermore, our data suggest that modifications to the centrifugation separation methods used here, or antibody-based purifications of various subcellular compartments, will allow the development of a comprehensive and detailed picture of the distribution of each mRNA in the cell. This information should make an important contribution to our understanding of the regulation and functional roles of the proteins they encode.

Methods

Subcellular fractionation and RNA isolation. We used equilibrium density gradient centrifugation to separate free and rER-bound polysomes from human Jurkat cells, and a differential precipitation procedure to separate free and rER-bound polysomes from an exponentially growing culture of *Saccharomyces cerevisiae*. Briefly, 5×10^8 Jurkat cells were treated with cycloheximide (50 μM; Sigma) for 10 min at 37 °C, lysed hypotonically using a ball-bearing homogenizer and fractionated by sedimentation equilibrium as described[16]. The membrane-associated and cytosolic ribosomes appeared well-separated, based on OD260 profiles. For yeast cultures, yeast strain DBY7286 (MATa *GAL2 ura3*; ref. 17) was grown to exponential phase in YEP medium supplemented with glucose[18], disrupted on liquid nitrogen using a mortar and pestle, and fractionated by sedimentation velocity as described[19]. In both cases, we isolated total RNA from the membrane and cytoplasmic fractions using Trizol (Life Technologies). For Jurkat cells, poly(A)$^+$ RNA was isolated from total RNA using an Oligotex mRNA kit (Qiagen) The resulting products were then amplified using a linear, *in vitro* transcription-based, antisense RNA amplification[20].

Microarray manufacture and hybridizations. DNA microarrays were produced and hybridized as described[21] (http://cmgm.stanford.edu/pbrown). To quantitate the representations of mRNAs in each fraction, we prepared Cy5-labelled cDNA from mRNA extracted from the rER fractions and Cy3-labelled cDNA from mRNA extracted from the cytoplasmic complement. The yeast genome microarray contained essentially all of the ORFs of *S. cerevisiae*[22] and the human cDNA microarray contained a set of ~9,000 sequence-confirmed cDNA clones, representing both characterized and uncharacterized genes[23,24]. The microarrays were scanned on a ScanArray3000 (General Scanning) and analysed using ScanAlyze (available at http://rana.stanford.edu/software/). Raw images and data from the experiments shown here are available (http://genome-www.stanford.edu/mbp). The data shown are representative experiments from multiple, independent subcellular fractionations and microarray hybridizations. The fractionation was highly reproducible and supplemental data from independent fractionations and analyses of reproducibility is available (see http://genome-www.stanford.edu/mbp/). Only genes for which the fluorescence signal in each channel exceeded a value corresponding to roughly 0.5% of the dynamic range above background were considered in this analysis.

Identification of empirically determined membrane-associated proteins. We collected experimentally determined subcellular localization information of protein products for as many genes as possible for both microarrays. The sources for this information included literature searches and publicly available databases (SWISS-PROT, http://www.expasy.ch/sprot/; SGD, http://genome-www.stanford.edu/Saccharomyces/; YPD, http://www.proteome.com/YPDhome.html). Proteins documented to be secreted, or to be localized to the ER, golgi, vesicles or plasma membrane (all of which were expected to have been bound to rER-associated polysomes) were grouped together as 'membrane-associated' gene products. Genes

encoding cytosolic proteins were designated as 'free'. Nuclear genes encoding mitochondrial proteins were excluded from this analysis because there is evidence for two classes of such proteins: those that are translated freely in the cytosol and post-translationally transported into the mitochondria, and those whose mRNA is associated with the cytoplasmic surface of mitochondria[25,26]. In our experiments, transcripts of nuclear genes encoding mitochondrial proteins were observed to be enriched in both fractions.

Moving average analyses. We used a similar moving average calculation for Figs 4 and 5. As an example, in Fig. 4*a* only genes of known subcellular localization were considered. To calculate a moving average of known membrane-associated proteins using a window size of 151, the fraction of membrane-associated proteins for 151 adjacent genes in Cy5/Cy3 ratio space was computed and plotted as a function of the central gene in the window. The 151 gene window was then moved by one gene on the Cy5/Cy3 axis and the fraction was re-calculated. This process was reiterated until the end of the Cy5/Cy3 distribution was reached.

Acknowledgements:
We thank the members of the Brown and Botstein labs for assistance and discussions, and M. Niwa, J. Peters and P. Walter for helpful advice and assistance. This work was supported by the Howard Hughes Medical Institute and by grants from the NHGRI (HG00983) and the NCI (CA77097). M.D. was supported by an MSTP fellowship. M.B.E. was supported by a DOE/NSF Sloan Fellowship. P.O.B. is an associate investigator of the Howard Hughes Medical Institute.

Received 7 October 1999; accepted 15 March 2000.

1. Nielsen, H., Brunak, S. & von Heijne, G. Machine learning approaches for the prediction of signal peptides and other protein sorting signals. *Protein Eng.* **12**, 3–9 (1999).
2. Tashiro, K. *et al.* Signal sequence trap: a cloning strategy for secreted proteins and type I membrane proteins. *Science* **261**, 600–603 (1993).
3. Klein, R.D., Gu, Q., Goddard, A. & Rosenthal, A. Selection for genes encoding secreted proteins and receptors. *Proc. Natl Acad. Sci. USA* **93**, 7108–7113 (1996).
4. Zannettino, A.C., Rayner, J.R., Ashman, L.K., Gonda, T.J. & Simmons, P.J. A powerful new technique for isolating genes encoding cell surface antigens using retroviral expression cloning. *J. Immunol.* **156**, 611–620 (1996).
5. Kopczynski, C.C. *et al.* A high throughput screen to identify secreted and transmembrane proteins involved in Drosophila embryogenesis. *Proc. Natl Acad. Sci. USA* **95**, 9973–9978 (1998).
6. Scherer, P.E., Bickel, P.E., Kotler, M. & Lodish, H.F. Cloning of cell-specific secreted and surface proteins by subtractive antibody screening. *Nature Biotechnol.* **16**, 581–586 (1998).
7. Mechler, B. & Rabbitts, T.H. Membrane-bound ribosomes of myeloma cells. IV. mRNA complexity of free and membrane-bound polysomes. *J. Cell. Biol.* **88**, 29–36 (1981).
8. Mueckler, M.M. & Pitot, H.C. Structure and function of rat liver polysome populations. I. Complexity, frequency distribution, and degree of uniqueness of free and membrane-bound polysomal polyadenylate-containing RNA populations. *J. Cell. Biol.* **90**, 495–506 (1981).
9. Kyte, J. & Doolittle, R.F. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* **157**, 105–132 (1982).
10. Nielsen, H., Engelbrecht, J., Brunak, S. & von Heijne, G. Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng.* **10**, 1–6 (1997).
11. Takizawa, P.A., Sil, A., Swedlow, J.R., Herskowitz, I. & Vale, R.D. Actin-dependent localization of an RNA encoding a cell-fate determinant in yeast. *Nature* **389**, 90–93 (1997).
12. Sidrauski, C. & Walter, P. The transmembrane kinase Ire1p is a site-specific endonuclease that initiates mRNA splicing in the unfolded protein response. *Cell* **90**, 1031–1039 (1997).
13. Chapman, R.E. & Walter, P. Translational attenuation mediated by an mRNA intron. *Curr. Biol.* **7**, 850–859 (1997).
14. Lukyanetz, E.A. Evidence for colocalization of calcineurin and calcium channels in dorsal root ganglion neurons. *Neuroscience* **78**, 625–628 (1997).
15. Pietrini, G. *et al.* A single mRNA, transcribed from an alternative, erythroid-specific, promoter, codes for two non-myristylated forms of NADH-cytochrome b5 reductase. *J. Cell. Biol.* **117**, 975–986 (1992).
16. Mechler, B.M. Isolation of messenger RNA from membrane-bound polysomes. *Methods Enzymol.* **152**, 241–248 (1987).
17. Spellman, P.T. *et al.* Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. *Mol. Biol. Cell* **9**, 3273–3297 (1998).
18. Sherman, F. Getting started with yeast. *Methods Enzymol.* **194**, 3–21 (1991).
19. Stoltenburg, R., Wartmann, T., Kunze, I. & Kunze, G. Reliable method to prepare RNA from free and membrane-bound polysomes from different yeast species. *Biotechniques* **18**, 564–566, 568 (1995).
20. Kacharmina, J.E., Crino, P.B. & Eberwine, J. Preparation of cDNA from single cells and subcellular regions. *Methods Enzymol.* **303**, 3–18 (1999).
21. Eisen, M.B. & Brown, P.O. DNA arrays for analysis of gene expression. *Methods Enzymol.* **303**, 179–205 (1999).
22. DeRisi, J.L., Iyer, V.R. & Brown, P.O. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* **278**, 680–686 (1997).
23. Iyer, V.R. *et al.* The transcriptional program in the response of human fibroblasts to serum. *Science* **283**, 83–87 (1999).
24. Perou, C.M. *et al.* Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. *Proc. Natl Acad. Sci. USA* **96**, 9212–9217 (1999).
25. Egea, G., Izquierdo, J.M., Ricart, J., San Martin, C. & Cuezva, J.M. mRNA encoding the beta-subunit of the mitochondrial F1-ATPase complex is a localized mRNA in rat hepatocytes. *Biochem. J.* **322**, 557–565 (1997).
26. Lightowlers, R.N., Sang, A.E., Preiss, T. & Chrzanowska-Lightowlers, Z.M. Targeting proteins to mitochondria: is there a role for mRNA localization? *Biochem. Soc. Trans.* **24**, 527–531 (1996).