Nucleic Acids Research, 2005, Vol. 33, No. 4 ***e34***
doi:10.1093/nar/gni032

A high-throughput screening of genes that encode proteins transported into the endoplasmic reticulum in mammalian cells

Takeaki Ozawa[1,2,3], Kengo Nishitani[1,2], Yusuke Sako[1,2] and Yoshio Umezawa[1,2,*]

[1]Department of Chemistry, School of Science, The University of Tokyo, Hongo, Bunkyo-ku, Tokyo 113-0033, Japan, [2]Japan Science and Technology Corporation, Tokyo, Japan and [3]PREST, Japan Science and Technology Agency, 4-1-8 Honcho Kawaguchi, Saitama, Japan

Received December 10, 2004; Revised and Accepted February 3, 2005

ABSTRACT

The compartments of eukaryotic cells maintain a distinct protein composition to perform a variety of specialized functions. We developed a new method for identifying the proteins that are transported to the endoplasmic reticulum (ER) in living mammalian cells. The principle is based on the reconstitution of two split fragments of enhanced green fluorescent protein (EGFP) by protein splicing with DnaE from *Synechocystis* PCC6803. Complementary DNA (cDNA) libraries fused to the N-terminal halves of DnaE and EGFP are introduced in mammalian cells with retroviruses. If an expressed protein is transported into the ER, the N-terminal half of EGFP meets its C-terminal half in the ER, and full-length EGFP is reconstituted by protein splicing. The fluorescent cells are isolated using fluorescence-activated cell sorting and the cDNAs are sequenced. The developed method was able to accurately identify cDNAs that encode proteins transported to the ER. We identified 27 novel proteins as the ER-targeting proteins. The present method overcomes the limitation of the previous GFP- or epitope-tagged methods, using which it was difficult to identify the ER-targeting proteins in a high-throughput manner.

INTRODUCTION

Protein transport from the cytosol to the endoplasmic reticulum (ER) in mammalian cells is an initial step in biogenesis (1–3), including not only secretory and plasma membrane proteins but also proteins that are destined to locate in endomembranes such as the ER, Golgi and lysosome. The identification of the proteins localized in such compartments is important for understanding the biological processes at the cellular levels (4,5). Protein transport to the ER occurs via a short N-terminal polypeptide known as the ER signal sequences (6–8). The signal sequences contain a positively charged N-terminal region, a hydrophobic central region and a polar C-terminal region that contains a signal peptide cleavage site. The signal sequences have a large variety in their amino acid sequences and their overall length, ranging from 15 to more than 50 amino acid residues. Although there are programs to predict the ER signal sequences (9–11), it is difficult to accurately identify ER-targeting proteins by analyzing the amino acid sequences with bioinformatics.

To date, the majority of localization studies have been undertaken in yeast (12–14), primarily due to the ease of generating proteins fused to the green fluorescent protein (GFP) or epitope for antibody. A list of protein localizations in yeast has thus been compiled and used for proteomic analysis. Now, identification of the protein localization in mammalian cells that varies widely between different tissues and stages is one of the most intriguing and rapidly advancing areas in cell biology (15–17). GFP- or epitope-tagged approach in combination with fluorescence microscopy is however often time consuming for systematically analyzing complementary DNAs (cDNAs) for mammalian cells (4). For high-throughput analysis of the protein localization inside the cells, automated fluorescence microscopy has been developed (18). Although such technological progress is important, image acquisition is slow and tedious. In particular, algorithms to automatically determine the protein localization in an acquired image still remain imprecise and slow.

To address difficulties encountered with this GFP- or epitope-tagged approach, we have developed a method with general applicability for high-throughput identification of the genes from large-scale cDNA libraries that encode proteins with the ER signal sequences. The principle is based on the reconstitution of split-enhanced GFP (EGFP) fragments by protein splicing. The protein splicing is a post-translational

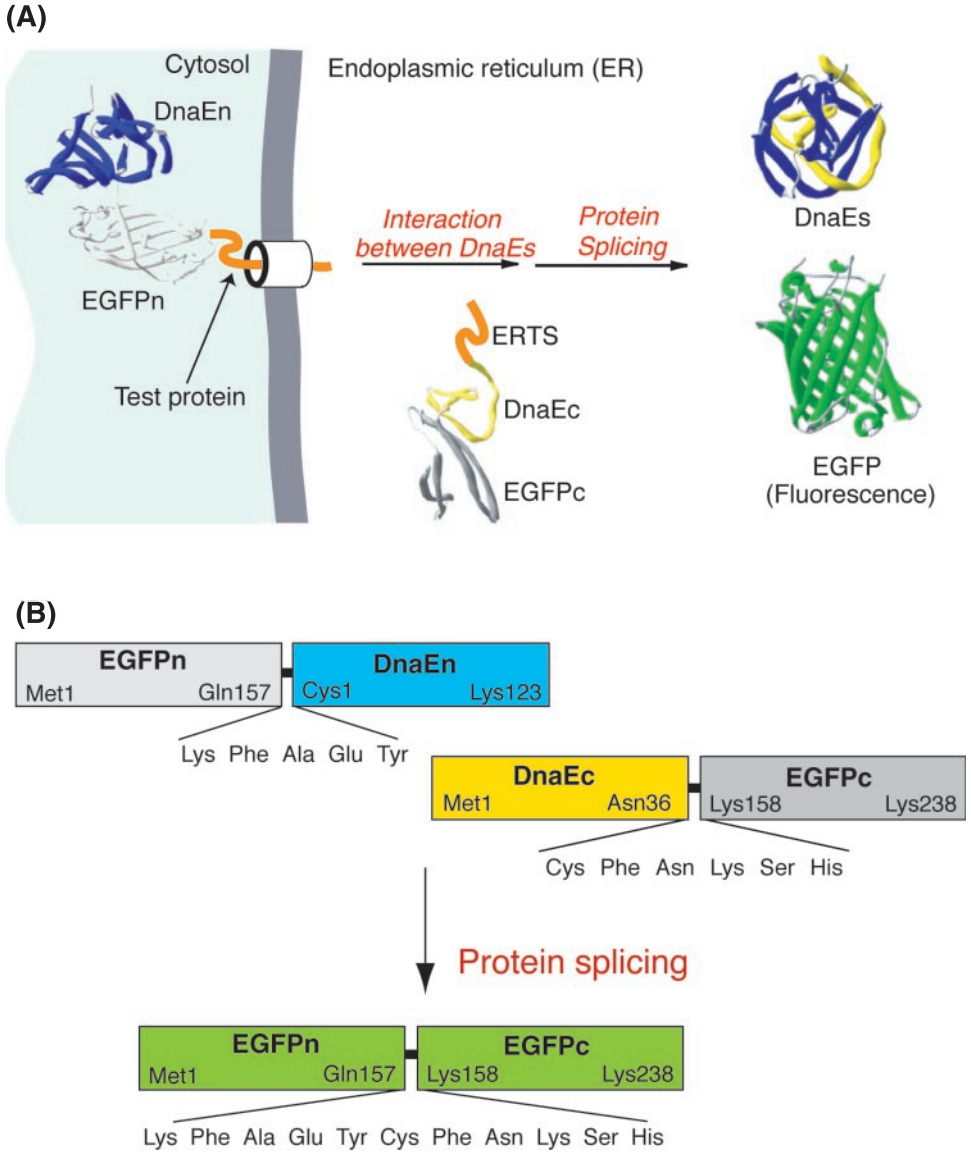
*To whom correspondence should be addressed. Tel: +81 3 5841 4351; Fax: +81 3 5841 8349; Email: umezawa@chem.s.u-tokyo.ac.jp

event involving precise excision of an internal sequence termed intein and ligation of the flanking sequences, termed N- and C-exteins by a peptide bond (19,20). While most of the inteins are composed of single polypeptides, a pair of functional and naturally split intein-coding sequences has been found from the split *dnaE* genes in the genome of *Synechocystis* sp. PCC6803 (21). With this split fragments of the DnaE intein, we have previously developed a new split-EGFP reporter for identifying the mitochondrial proteins (22). The fluorescence of the split-EGFP reporter can be recovered by protein splicing when the splicing protein of DnaE is assembled by protein transports into mitochondria. This basic concept was extended for designing a new indicator for identifying proteins transported into the ER. A tandem fusion protein containing an ER-targeting signal (ERTS) and the C-terminal fragments of DnaE and EGFP localizes in the lumens of the ER (Figure 1A). cDNA libraries are genetically fused to the sequences encoding the N-terminal fragments of DnaE and EGFP. If test proteins expressed from the cDNAs contain an ERTS, the fusion products translocate into the ER, in which the N- and C-terminal halves of DnaEs are brought close enough to fold correctly, thereby initiating protein splicing of the split EGFP to recover its EGFP fluorescence (Figure 1B). The fluorescent cells are collected by fluorescence-activated cell sorting (FACS), and from each clone, cDNA is retrieved and its sequence is analyzed (Figure 1C). Using this method, we were able to identify 109 non-redundant genes that encode proteins transported to the ER.

MATERIALS AND METHODS

Construction of expression vectors

A DNA fragment encoding the C-terminus of DnaE and EGFP (158–238 amino acids), termed DnaEc–EGFPc, was amplified by PCR as a template pMX–MTS/DEc(neo) (22), resulting in the addition of an ERTS of preprolactin (MNSQVSARKAGTLLLLMMSNLLFCQNVQTLP) (23) and ER retention signal (KDEL) (24) to the N- and C-terminus of



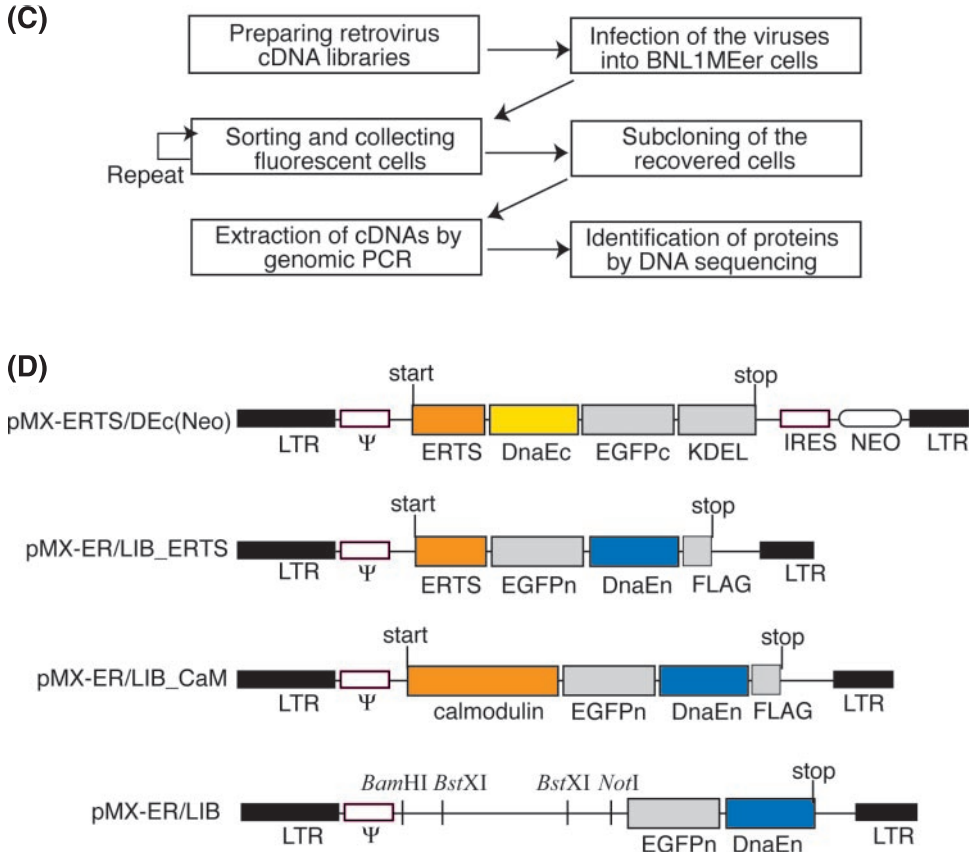


Figure 1. Scheme of the basic principle and the strategy for identifying ER-targeting proteins. (**A**) Principle for detecting translocation of a test protein into the ER using protein splicing of split EGFP. EGFPc is connected with DnaEc and ERTS, which is predominantly localized in the ER. A test protein is connected with EGFPn and DnaEn, which is expressed in the cytosol. When the test protein translocates into the ER, the DnaEn interacts with DnaEc resulting in protein splicing. The EGFPn and EGFPc are linked together by a peptide bond, and the reconstituted EGFP recovers its fluorescence. (**B**) Domain structures of DnaE-tagged EGFP before and after protein splicing, showing sequences of the boundaries between DnaEn and EGFPn, DnaEc and EGFPc, and EGFPn and EGFPc. (**C**) Strategy for identifying ER-targeting proteins. BNL1MEer cells were infected with retrovirus libraries with 20% infection efficiency. Fluorescent cells were sorted by FACS and collected on 48-well plates. cDNAs integrated in the genome were extracted by PCR and identified by DNA sequencing. (**D**) Schematic structures of major constructs. ERTS, an ER-targeting signal derived from mouse preprolactin; EGFPn, N-terminal half (1–157 amino acids) of EGFP; EGFPc, C-terminal half (158–238 amino acids) of EGFP; DnaEn and DnaEc, N- and C-terminal DnaEs; IRES, internal ribosome entry site; NEO, neomycin resistance; LTR, long terminal repeat; Ψ, retrovirus-packaging signal and FLAG, FLAG epitope (DYKDDDDK).

DnaEc–EGFPc, respectively. Unique BamHI (5′) and SalI (3′) sites were introduced in both ends. The PCR product was subcloned into pMX at the restriction sites and the product was sequenced to ensure the fidelity of the DNA sequence. To express a DNA fragment encoding N-terminus of EGFP (1–157 amino acids) and DnaE, termed EGFPn–DnaEn, connected with ERTS or calmodulin, the cDNA was amplified by PCR to introduce BamHI (5′) and NotI(3′) restriction sites. The PCR products were inserted into pMX–Mito/LIB in frame (22) and their sequences were verified.

Selection of stable clones

The fusion gene of pMX–ERTS/DEc(Neo) was transfected into the packaging cell line PlatE with Lipofectamine 2000 (Invitrogen, Carlsbad, CA) (25). After 2 days of culture, high-titer retroviruses were collected and infected into BNL1ME cells. The cells stably expressing DnaEc–EGFPc were obtained after ~10 days of selection in G418 (Invitrogen) containing a growth medium. The cells were subcloned and a cell line that expressed DnaEc–EGFPc in the ER was obtained.

cDNA library

We used the cDNA libraries constructed in the previous study (22). Briefly, polyA(+)RNAs extracted from cultured BNL1ME cells derived from mouse normal liver were converted to cDNAs using random hexamer primers. The resulting cDNAs longer than 600 kb were ligated to BstXI adapters and inserted into two BstXI sites of pMX–ER/LIB vector (Figure 1D). Thus, constructed cDNAs were connected with the cDNAs of N-terminal halves of DnaE and EGFP (1–157 amino acids). The ligated cDNAs were transformed into competent DH10B cells (Invitrogen), and the plasmids were purified using a Qiaex kit (Quiagen, Hilden, Germany).

Sorting and identification of cDNA

The plasmids encoding cDNAs were transfected into the PlatE cells and high-titer retroviruses were collected after 2 days of culture. Using the retrovirus libraries, BNL1MEer cells were infected with an infection efficiency of <20% as estimated by a control experiment using pMX–EGFP (22). The cells were incubated for 5 days at 37°C and 24 h at 28°C before sorting.

The cells were stripped off with trypsin–EDTA and dissolved in phosphate-buffered saline. Fluorescent cells were sorted by FACS and collected on 48-well microtiter plates.

To recover cDNAs integrated into the genome, the genome was extracted from each BNL1MEer clone and subjected to PCR. The following PCR primers were used: AGGACCTTACACAGTCCTGCTGACC (forward) and GCCCTCGCCGGACACGCTGAACTTG (reverse). The PCR was run for 30 cycles (30 s at 98°C for denaturation, 30 s at 58°C for annealing and 2 min at 72°C for extension) using an LA *Taq* polymerase (Takara Shuzo, Kyoto, Japan). The resulting fragments were sequenced using a BigDyeTerminator Cycle Sequencing Kit and were analyzed by an automatic sequencer (310 Genetic Analyzer; Applied Biosystems, Foster City, CA).

Gene assignments and functional annotation of genes

Each cDNA sequence was compared with the sequences of DNAs in National Center for Biotechnology Information (NCBI) databases using BLASTn (http://www.ncbi.nlm.nih.gov/BLAST/) and FANTOM database (http://fantom2.gsc.riken.go.jp/db/search/). Expected values (*E*-values), which are defined as the statistical significance threshold for reporting matches against database DNA sequences (26), $<1 \times 10^{-10}$ were accepted. Sense or anti-sense orientation of each cDNA strand was determined based on the RIKEN clone sets (25).

Imaging cells

Each clone was spread on a glass-base dish and incubated for 24 h in the presence of a growth medium. The medium was replaced with Hanks' balanced salt solution and reconstituted EGFP was directly imaged using a confocal laser-scanning microscope (LSM510meta; Carl Zeiss, Jena, Germany). ER and lysosome were stained with 1.0 μM BODIPY-BFA and 0.5 μM LysoTracker Red DND-99 (Molecular Probes, Eugene, OR), respectively.

Antibodies and immunoblots

The cell lysate of BNL1ME cells infected with retroviruses was subjected to SDS–PAGE using 10% acrylamide gels and transferred to nitrocellulose membrane. The membranes were blotted with mouse anti-GFP antibody (Roche Applied Science, Mannheim, Germany) or anti-FLAG antibody (Sigma, St Louis, MO), and then with alkaline phosphatase-conjugated secondary antibody. The secondary antibody was visualized by chemiluminescence (New England Biolabs, Beverly, MA).

RESULTS

Construction of reporter proteins for identifying ER-signal sequences

The purpose of the present method is to identify all the proteins to be transported into the ER from large-scale cDNA libraries. To implement this, we placed a bait protein inside the ER; the coding sequence of ERTS derived from mouse preprolactin is fused to the C-terminal sequences of DnaE and EGFP (Figure 1) (23). For the fusion protein to retain in the ER, an amino-acid sequence of KDEL is connected with the C-terminal end of the fusion protein (24). The cDNA is introduced into BNL1ME cells, and a stable cell line expressing the ERTS–DnaEc–EGFPc protein in the ER is developed, which is named BNL1MEer. cDNA libraries generated from mRNAs are genetically fused to the sequences encoding the N-terminal halves of EGFP and of DnaE. The cDNA libraries are transformed into retrovirus libraries, which are used to infect the BNL1MEer cells. If a prey protein encoded in the cDNA contains a functional ERTS, it translocates into the ER, where EGFP is formed by protein splicing (Figure 1A). The cells harboring this reconstituted EGFP are screened rapidly by FACS and each cDNA is isolated from the cells and identified (Figure 1C).

To examine whether the splicing reaction with the DnaE intein occurs efficiently in the luminal space of the ER, the plasmids of pMX–ERTS/DEc(Neo) and pMX–ER/LIB–ERTS were converted into retroviruses and the proteins were expressed in the BNL1ME cells. Western blot analysis of the cell lysates is shown in Figure 2A. When the protein was expressed with pMX–ERTS/DEc(Neo), 15 kDa of the precursor protein was observed. In contrast, when the cell lysates harboring both pMX–ERTS/DEc(Neo) and pMX–ER/LIB–ERTS were analyzed, the splicing product was found, molecular size of which was consistent with that of EGFP. A band of 12 kDa was also obtained, similar to the size of the C-terminal half of EGFP. Such a C-terminal cleavage of EGFP from the C-terminal end of DnaE is known to occur in the presence of a reducing agent, such as thiol groups (27). The results indicate that protein splicing occurred exclusively in the cells expressing the plasmids of pMX–ERTS/DEc(Neo) and pMX–ER/LIB–ERTS.

Next we analyzed the localization of the spliced product of EGFP with the fluorescence imaging technique (Figure 2B). First, BNL1ME cells were transiently infected with the plasmids of both pMX–ERTS/DEc(Neo) and pMX–ER/LIB–ERTS, and were incubated for 2 days at 37°C. The results showed that the cells exhibited no fluorescence (data not shown). However, when the cells were incubated at 28°C under otherwise identical conditions, the cells were found to fluoresce, indicating that folding of the reconstituted EGFP and the fluorophore formation have a critical threshold temperature. The subcellular localization of the EGFP was merged with the ER stained with BODIPY-BFA. From the results, it was concluded that the N- and C-terminal halves of EGFP were reconstituted by protein splicing, and the full-length EGFP thus formed was correctly refolded in the ER at 28°C.

We next examined the fluorescence intensity of the EGFP reconstituted in the ER by FACS. BNL1MEer cells were infected with retroviruses harboring pMX–ER/LIB–ERTS at single multiplicity of infection (MOI), defined by the number of cDNAs per cell. After incubation at 28°C for 2 days, the cells were subjected to FACS analysis. Upon spreading viruses harboring the cDNA of ERTS connected with EGFPn–DnaEn onto the BNL1MEer, the population of fluorescent cells was observed (Figure 2C). The observed fluorescence intensity was found strong enough to be discriminated from the background fluorescence originating from the population of non-infected cells. In contrast, the fluorescence intensity of the cells harboring the cDNA of calmodulin connected with EGFPn–DnaEn was almost the same as the background fluorescence, demonstrating that the expression of the cytosolic protein

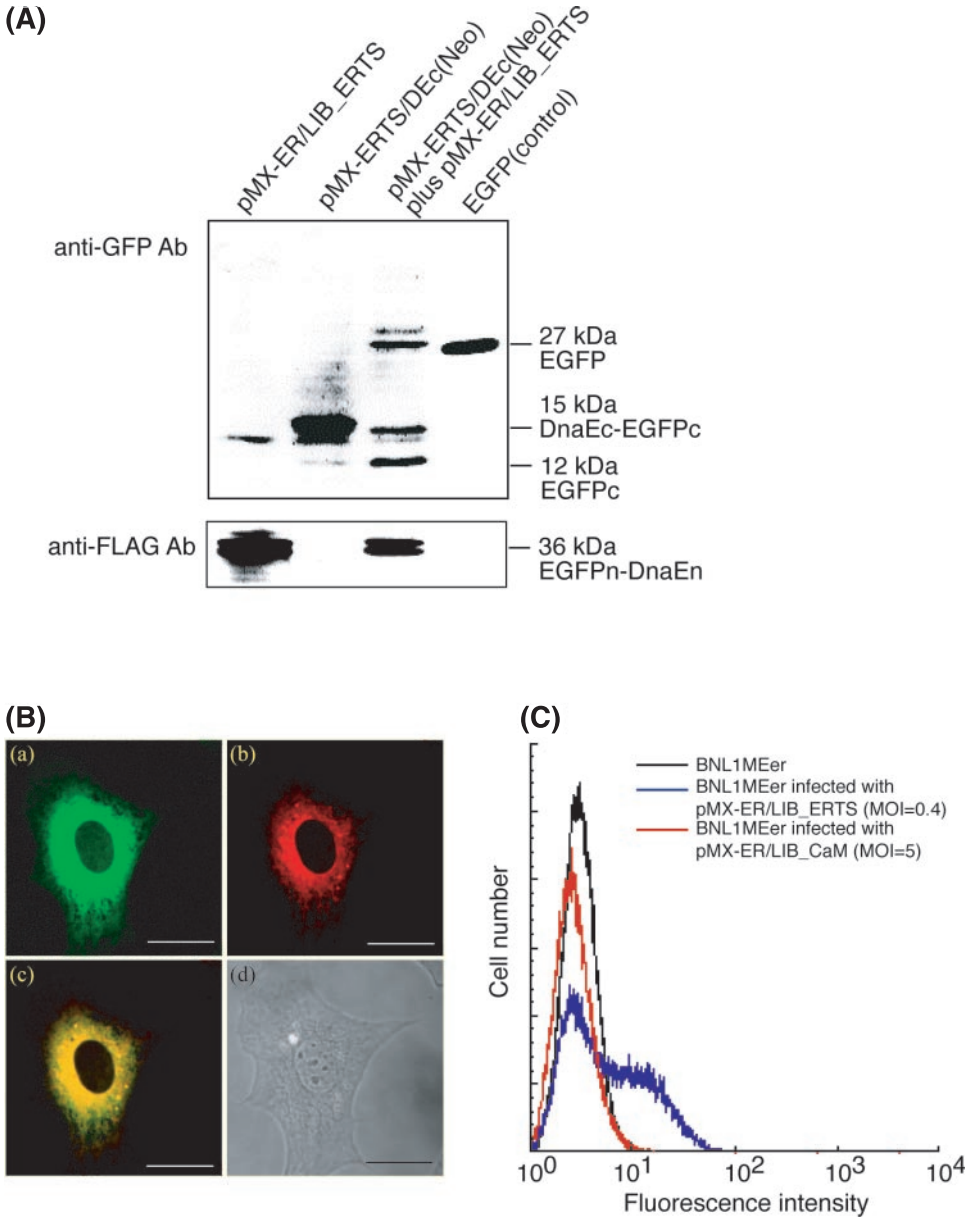


Figure 2. Selective and sensitive detection of ER-targeting proteins. (**A**) Western blot analysis of the whole cell lysates from BNL1ME cells expressing either EGFPn–DnaEn-tagged ERTS or DnaEc–EGFPc-tagged ERTS, or expressing both the fusion proteins. Blots were performed with a monoclonal anti-GFP antibody (upper panel) specific to the C-terminal half of EGFP and a monoclonal anti-FLAG antibody (lower panel). (**B**) Fluorescence images of the reconstituted EGFP localization. BNL1MEer cells were infected with retroviruses of pMX–ER/LIB–ERTS and cultured for 2 days at 28°C. The cells were spread on a glass dish and the fluorescent image was taken (a). After procedure (a), ER was stained with BODIPY-BFA, for which the fluorescence was recorded (b). Superimposed image (c) shows EGFP localization specific to ER. Image (d) is of transmission. Bar, 10 μm. (**C**) FACS profiles of BNL1MEer cells infected with either pMX–ER/LIB–ERTS or pMX–ER/LIB–CaM. For comparison, uninfected cells are also shown.

connected with EGFPn–DnaEn in the BNL1MEer cells did not induce the reconstitution of the split EGFP. The above results together with those of western blot and fluorescence imaging indicate that the cells that express proteins with ERTS are separated from the ones without ERTS by using the present N- and C-terminal probes.

Cloning of the proteins transported to the ER

We used the BNL1MEer cells and pMX–ER/LIB for screening and isolating the proteins transported into the ER from large-scale cDNA libraries. The cDNA libraries were converted into retrovirus libraries and the BNL1MEer cells were infected with the retroviruses at an MOI of 0.1–0.2 (cDNAs/cell). The cells were incubated for 5 days and the fluorescence intensity of each cell was analyzed by FACS (Figure 3A and B). The percentage of the fluorescent cells was $(4.26 \pm 0.09) \times 10^{-2}$% of the total analyzed cells. The cells were collected by FACS and analyzed again to increase the number of collected fluorescent cells. The ratio of the number of fluorescent cells to that of the total number of analyzed cells increased up to 30.3 ± 1.1% (red line).

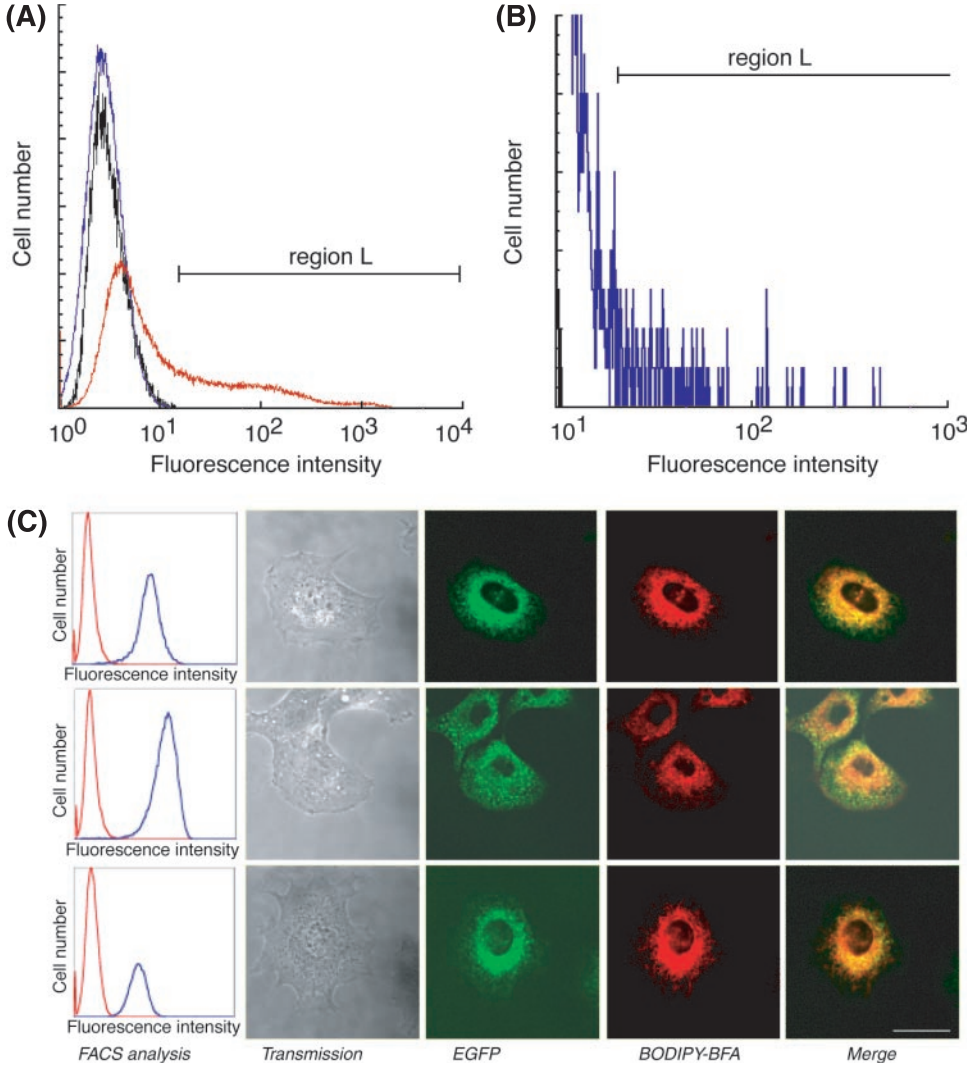


Figure 3. FACS profiles of BNL1MEer cells infected with retrovirus cDNA libraries. (**A**) BNL1MEer cells were infected with the cDNA retrovirus libraries with an infection efficiency of 20%. Five days after incubation, the cells were stripped off and sorted by FACS (blue). The enlarged FACS profiles of the cells in region L are shown in (**B**). The fluorescent cells within region L were collected and again sorted by FACS (red). Uninfected cells were inserted to show the background fluorescence (black). (**C**) FACS profiles and fluorescent images of three representative clones. Left, fluorescence intensity of cloned cells (blue) and uninfected BNL1MEer cells (red) were analyzed by FACS. Total cell counts analyzed were 10^5 cells. Right, each cloned cell was cultured on the glass slide and a confocal image of the live cells harboring reconstituted EGFP was taken. Cells were stained with BODIPY-BFA to show the ER localization of individual cells.

The fluorescent cells thus analyzed were then isolated on the 48-well plates by FACS.

After each cell thus collected was cultured on the plates, their fluorescence intensities and subcellular localizations of the reconstituted EGFP were analyzed (Figure 3C). Randomly selected 10 clones showed strong fluorescence and their localizations were found to be exclusively in the ER.

Next we collected 1500 clones by FACS, extracted their genome and then analyzed the cDNAs. Of the 1500 clones, 1300 cDNAs were obtained by the genomic PCR and their sequences were analyzed. From the rest 200 clones, no cDNA was recovered, probably because the concentration of the extracted genome was low or the cDNA was not integrated in the genome. The obtained cDNAs have a wide range of lengths ranging from 300 to 2000 bp. The sequences of such cDNAs were searched in the NCBI nucleotide and genome sequence databases. We used a criterion for accepting a database hit: only a cDNA sequence corresponding to the specificity of *E*-values lower than 1×10^{-10} was accepted (see Materials and Methods). Using this criterion, we succeeded in identifying a set of 1104 mouse cDNAs, which are listed in Supplementary Table 1. In this list, alternative splicing variants of mRNAs were included within the same gene. The list contained a high degree of redundancy. The reason for this may be that some mRNAs were expressed in large quantities in the cells, and strongly fluoresced cells were collected with higher probability than the weakly fluoresced ones. Consequently, we identified 149 non-redundant cDNAs, 109 unique proteins of which were encoded. The rest 40 cDNAs encoded wrong-reading frames or the sequences of wrong-initiated ATG. While many redundant proteins were included in the list, 38 unique proteins (35% of the total) were identified by a single-hit cDNA, indicating that the present method enabled to obtain comprehensive cDNAs that encode the proteins transported to the ER.

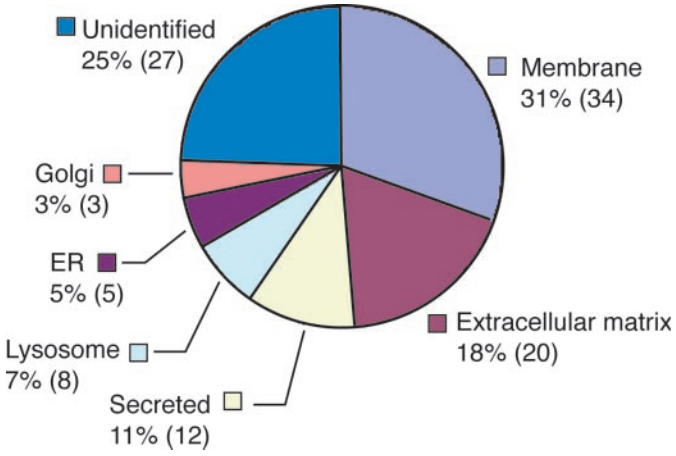


Figure 4. Category of the non-redundant proteins according to their subcellular localization. The pie chart shows the percentage of the proteins localized in each organelle. The number of proteins in each category is in the parenthesis.

Of the 109 non-redundant proteins with reading frames and start codon completely identical to those of RIKEN full-length cDNA clones (28,29), 82 proteins were found to be known as to their localization (Figure 4 and Supplementary Table 2). Each protein was found to respectively contain an N-terminal signal sequence with a typical hydrophobic core region consisting of 6–15 amino acid residues, the most essential part required for transport to the ER (7). From the results of the cDNA analysis together with the cloning of the fluorescent cells described above, we concluded that the present method is accurate for identifying the cDNAs that encode proteins transported to the ER.

The 27 proteins of the 109 proteins were not known as to their localization, and therefore, we regarded these as novel proteins. It was found that nine proteins of the novel proteins were homologous to *Homo sapiens* or *Rattus norvegicus* proteins (Supplementary Table 3). Of the nine proteins, we randomly selected six proteins and determined their subcellular localization of the EGFP-tagged proteins using confocal fluorescence microscopy (Figure 5). The EGFP-tagged protein

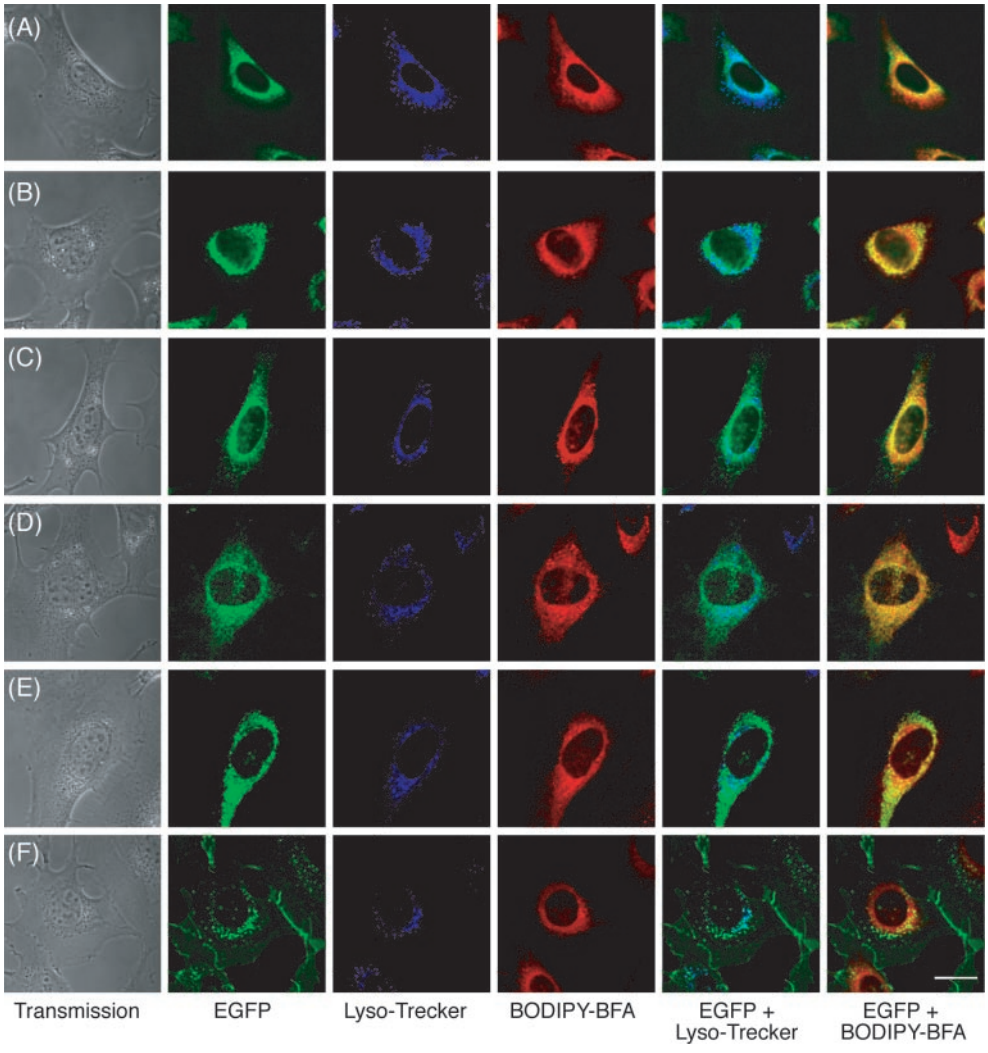


Figure 5. Subcellular localization of the proteins expressed from EGFP-tagged cDNA clones. The cDNAs connected with the full-length EGFP were converted to retroviruses, and introduced into the BNL1ME cells. The cells were cultured for 2 days and the images of the live cells were taken using confocal microscopy. Lysosome and ER were stained with Lyso-Trecker and BODIPY-BFA, respectively. The images were merged with those of EGFP. (**A**) Clone GI26335108, (**B**) clone GI12852288, (**C**) clone GI26347862, (**D**) clone GI12852495, (**E**) clone GI12851248 and (**F**) clone GI26344595.

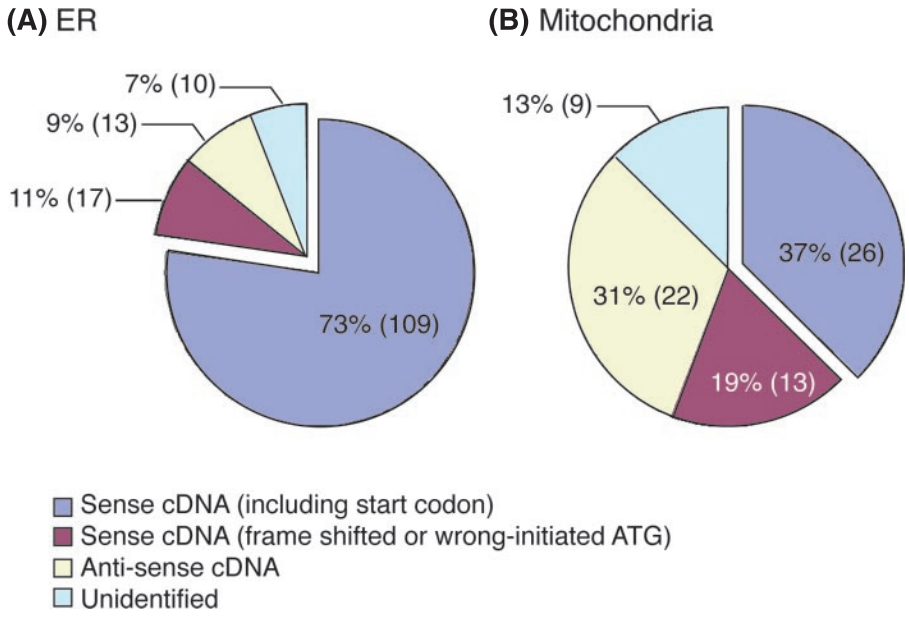


Figure 6. Comparison of the signal sequences included in the cDNA libraries between ER and mitochondria. The pie charts show the percentage of the non-redundant proteins identified from mouse cDNA libraries by RING technologies. Correctly annotated sense cDNAs are shown in purple, the other sense cDNAs including a wrong initiator ATG are shown in magenta, anti-sense cDNAs are shown in yellow and unidentified cDNAs are shown in pale blue.

GI26347862 was localized in the ER, the protein GI12852288 in lysosomes and the ER, and the protein GI26344595 on the extracellular matrix. The proteins GI12852495 and GI26335108 were found to localize both in the ER and the intracellular vesicles. These proteins are homologous to known secreted proteins, indicating that both proteins GI12852495 and GI26335108 were secreted proteins. The protein GI12851248 revealed that it localized exclusively in the ER. Taken all together, we summarized in Supplementary Table 3 the results of the newly identified (novel) nine ER-targeting proteins as to their final organelle localizations.

Seventy-three percent of the 149 non-redundant cDNAs encoding the signal sequences were correctly annotated sense cDNAs (Figure 6A) based on the NCBI DNA databases. Frame-shifted or wrong-initiated sense cDNAs were 11% and anti-sense cDNAs were 9%. The frame-shifted or wrong-initiated cDNAs may be caused by incomplete reverse transcription, while the anti-sense cDNAs may be caused by pMX-vector ligation with the cDNAs at the BstXI sites during the construction of the cDNA libraries (see Materials and Methods). Because such cDNA fragments have a wrong initiator ATG encoding methionine and a stretch encoding hydrophobic amino acids, the translated peptide fragment acted as a signal sequence transporting to the ER.

It is known that the N-terminal location of signal sequences is not only for ER-targeting proteins but also for the mitochondrial and peroxisomal proteins, although their respective signal sequences are different (30). The genes encoding ER- and peroxisome-targeting proteins are originated from the genome of ancient eukaryotic cells, while the genes encoding mitochondrial proteins are from the eubacterial genome, transferred to the eukaryotic genome in evolution (31). Thus, the origin of the signal sequences transporting to the ER is different from those of mitochondria. It is therefore of interest to compare the differences in the signal sequences between the ER and mitochondria. We examined the differences between the ER and the mitochondrial signal sequences as for the percentage of cDNAs containing correctly initiated ATG in the total cDNAs. The results are shown in Figure 6, in which the percentage for mitochondria was obtained from our previous study (22). In the cDNA libraries, 73% of correctly annotated sense cDNAs were included that encode the ER-targeting proteins. In the case of mitochondria, the cDNAs having the correct reading frames and the start codon were only 37% of the total, and the rest were associated with frame-shifted, wrong-initiated or reverse-oriented cDNAs (Figure 6B). Therefore, we found that the incorrectly annotated reading frames for mitochondrial-targeting signals were more frequent than those for ERTS.

DISCUSSION

Up to now, it has been possible to predict signal sequences targeted to the ER with computer programs such as SignalP or PSORT (9–11). However, the prediction of the signal sequences requires that precise information of the cDNAs or amino acid sequences be known and deposited in the databases. GFP-tagged approaches also need a great number of cDNAs for comprehensive analysis of subcellular localizations. On the contrary, the present method has a remarkable advantage of the use of cDNA libraries. In general, cDNA libraries prepared from mRNAs include many unknown cDNAs, and therefore, it is possible with this method to identify novel cDNAs from the cDNA libraries. One of the practical utilities of this method that uses cDNA libraries would be capable of seeking novel proteins that regulate cell-to-cell communications. Secreted and cell-surface proteins are known to play important roles in cellular interactions and to be potential therapeutic agents or targets for antagonistic

or agonistic therapy. The present method should be useful to identify such novel secreted and cell-surface proteins. The limitation of the present method is that it is difficult to discover rarely expressed genes in standard cDNA libraries, because cDNAs from highly-expressed mRNAs are preferentially identified. This problem could possibly be overcome by employing the subtraction methods, which were generally used in the comprehensive collection of cDNAs from animals and plants (28,29,32).

In conclusion, a new method was developed to accurately identify the cDNAs that encode proteins transported to the ER. We determined 109 non-redundant proteins, of which 82 proteins were known as ER-targeting proteins localized in the ER, Golgi, lysosome, plasma membrane, extracellular matrix and secreted. We identified 27 novel proteins as the ER-targeting proteins. This method is potentially very useful to identify the proteins localized in the endomembranes such as the ER, Golgi and lysosome as well as membrane and secreted proteins in a high-throughput manner.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at NAR Online.

ACKNOWLEDGEMENTS

This work was supported by Japan Science and Technology Agency (JST) and Japan Society for the Promotion Science (JSPS). Also acknowledged is the support to T.O. from the TAKEDA science foundation, Japan. Funding to pay the Open Access publication charges for this article was provided by JSPS.

REFERENCES

1. Bankaitis,V.A. and Morris,A.J. (2003) Lipids and the exocytotic machinery of eukaryotic cells. *Curr. Opin. Cell Biol*., **15**, 389–395.
2. Paulsson,K. and Wang,P. (2003) Chaperones and folding of MHC class I molecules in the endoplasmic reticulum. *Biochim. Biophys. Acta*, **1641**, 1–12.
3. Sitia,R. and Braakman,I. (2003) Quality control in the endoplasmic reticulum protein factory. *Nature*, **426**, 891–894.
4. Davis,T.N. (2004) Protein localization in proteomics. *Curr. Opin. Chem. Biol*., **8**, 49–53.
5. Phizicky,E., Bastiaens,P.I., Zhu,H., Snyder,M. and Fields,S. (2003) Protein analysis on a proteomic scale. *Nature*, **422**, 208–215.
6. von Heijne,G. (1985) Signal sequences. The limits of variation. *J. Mol. Biol*., **184**, 99–105.
7. Martoglio,B. and Dobberstein,B. (1998) Signal sequences: more than just greasy peptides. *Trends Cell Biol*., **8**, 410–415.
8. Zheng,N. and Gierasch,L.M. (1996) Signal sequences: the same yet different. *Cell*, **86**, 849–852.
9. Horton,P. and Nakai,K. (1997) Better prediction of protein cellular localization sites with the k nearest neighbors classifier. *Proc. Int. Conf. Intell. Syst. Mol. Biol*., **5**, 147–152.
10. Nielsen,H., Engelbrecht,J., Brunak,S. and von Heijne,G. (1997) Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng*., **10**, 1–6.
11. Bendtsen,J.D., Nielsen,H., von Heijne,G. and Brunak,S. (2004) Improved prediction of signal peptides: SignalP 3.0. *J. Mol. Biol*., **340**, 783–795.
12. Huh,W.K., Falvo,J.V., Gerke,L.C., Carroll,A.S., Howson,R.W., Weissman,J.S. and O'Shea,E.K. (2003) Global analysis of protein localization in budding yeast. *Nature*, **425**, 686–691.
13. Kumar,A., Agarwal,S., Heyman,J.A., Matson,S., Heidtman,M., Piccirillo,S., Umansky,L., Drawid,A., Jansen,R., Liu,Y. *et al.* (2002) Subcellular localization of the yeast proteome. *Genes Dev*., **16**, 707–719.
14. Ding,D.Q., Tomita,Y., Yamamoto,A., Chikashige,Y., Haraguchi,T. and Hiraoka,Y. (2000) Large-scale screening of intracellular protein localization in living fission yeast cells by the use of a GFP-fusion genomic DNA library. *Genes Cells*, **5**, 169–190.
15. Escobar,N.M., Haupt,S., Thow,G., Boevink,P., Chapman,S. and Oparka,K. (2003) High-throughput viral expression of cDNA-green fluorescent protein fusions reveals novel subcellular addresses and identifies unique proteins that interact with plasmodesmata. *Plant Cell*, **15**, 1507–1523.
16. Misawa,K., Nosaka,T., Morita,S., Kaneko,A., Nakahata,T., Asano,S. and Kitamura,T. (2000) A method to identify cDNAs based on localization of green fluorescent protein fusion products. *Proc. Natl Acad. Sci. USA*, **97**, 3062–3066.
17. Morin,X., Daneman,R., Zavortink,M. and Chia,W. (2001) A protein trap strategy to detect GFP-tagged proteins expressed from their endogenous loci in *Drosophila*. *Proc. Natl Acad. Sci. USA*, **98**, 15050–15055.
18. Kau,T.R., Way,J.C. and Silver,P.A. (2004) Nuclear transport and cancer: from mechanism to intervention. *Nature Rev. Cancer*, **4**, 106–117.
19. Paulus,H. (2000) Protein splicing and related forms of protein autoprocessing. *Annu. Rev. Biochem*., **69**, 447–496.
20. Evans,T.J.T. and Xu,M.Q. (2002) Mechanistic and kinetic considerations of protein splicing. *Chem. Rev*., **102**, 4869–4884.
21. Wu,H., Hu,Z. and Liu,X.Q. (1998) Protein trans-splicing by a split intein encoded in a split DnaE gene of *Synechocystis* sp. PCC6803. *Proc. Natl Acad. Sci. USA*, **95**, 9226–9231.
22. Ozawa,T., Sako,Y., Sato,M., Kitamura,T. and Umezawa,Y. (2003) A genetic approach to identifying mitochondrial proteins. *Nat. Biotechnol*., **21**, 287–293.
23. Harigaya,T., Nakayama,K., Ohkubo,H., Nakanishi,S., Seo,H. and Hoshino,K. (1986) Cloning and sequence analysis of cDNA for mouse prolactin. *Biochim. Biophys. Acta*, **868**, 30–38.
24. Munro,S. and Pelham,H.R. (1987) A C-terminal signal prevents secretion of luminal ER proteins. *Cell*, **48**, 899–907.
25. Morita,S., Kojima,T. and Kitamura,T. (2000) Plat-E: an efficient and stable system for transient packaging of retroviruses. *Gene Ther*., **7**, 1063–1066.
26. Karlin,S. and Altschul,S.F. (1990) Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. Natl Acad. Sci. USA*, **87**, 2264–2268.
27. Martin,D.D., Xu,M.Q. and Evans,T.C.,Jr (2001) Characterization of a naturally occurring trans-splicing intein from *Synechocystis* sp. PCC6803. *Biochemistry*, **40**, 1393–1402.
28. Kawai,J., Shinagawa,A., Shibata,K., Yoshino,M., Itoh,M., Ishii,Y., Arakawa,T., Hara,A., Fukunishi,Y., Konno,H. *et al.* (2001) Functional annotation of a full-length mouse cDNA collection. *Nature*, **409**, 685–690.
29. Okazaki,Y., Furuno,M., Kasukawa,T., Adachi,J., Bono,H., Kondo,S., Nikaido,I., Osato,N., Saito,R., Suzuki,H. *et al.* (2002) Analysis of the mouse transcriptome based on functional annotation of 60 770 full-length cDNAs. *Nature*, **420**, 563–573.
30. Paetzel,M., Karla,A., Strynadka,N.C. and Dalbey,R.E. (2002) Signal peptidases. *Chem. Rev*., **102**, 4549–4580.
31. Dyall,S.D., Brown,M.T. and Johnson,P.J. (2004) Ancient invasions: from endosymbionts to organelles. *Science*, **304**, 253–257.
32. Seki,M., Narusaka,M., Kamiya,A., Ishida,J., Satou,M., Sakurai,T., Nakajima,M., Enju,A., Akiyama,K., Oono,Y. *et al.* (2002) Functional annotation of a full-length *Arabidopsis* cDNA collection. *Science*, **296**, 141–145.