

In Vivo Functional Proteomics: Mammalian Genome Annotation Using CD-Tagging

BioTechniques 33: 852-866 (October 2002)

**J.W. Jarvik, G.W. Fisher,
C. Shi, L. Hennen, C. Hauser,
S. Adler, and P.B. Berget**
Carnegie Mellon University,
Pittsburgh, PA, USA

INTRODUCTION

The proteome of every cell comprises thousands of protein species with different and distinct locations, abundances, interactions, and biochemical activities. As the cell proceeds through the cell cycle, ages, differentiates, or responds to changing internal or external environments, the proteome changes with it. Clearly, the more completely we can describe the composition and dynamics of the proteome, the better we will understand the biology of the cell in health and disease.

Central Dogma (CD)-tagging (8) is a genetic technology that provides a novel opportunity to annotate protein functions in living cells and organisms. CD-tagging is performed by the insertion of a specially designed DNA sequence (the CD cassette) into genomic DNA. When the CD cassette is inserted in the proper orientation in an intron of a transcriptionally active gene, the cassette provides splicing signals that direct the inclusion of a new exon (the guest exon) into the transcript. Translation of the tagged transcript results in the incorporation of a unique peptide tag (the guest peptide) into the protein. Thus, a single DNA insertion event leads to the specific tagging of all three CD molecules—DNA, RNA, and protein. Significantly, there is no deletion of sequence from the tagged gene, transcript, or protein, just the addition of new tag sequences. As a result, the natural regulation of the gene and protein is expected to be maintained.

Previous work from our laboratories showed that the CD-tagging process was

effective for the functional analysis of cloned genes from algal, insect, and mammalian cells (8,19). The approach has also been used successfully to randomly tag *Drosophila* genes and proteins in live animals using an endogenous transposition system (8,12). Here we describe the delivery of CD tags directly to random sites in the genome of cultured mouse NIH 3T3 fibroblasts using a retroviral vector. Two tandem guest exons were delivered, one encoding an epitope tag and the other encoding a GFP tag. More than 300 GFP-expressing cell lines were isolated, and more than 60 of these were analyzed at the nucleic acid level to identify the genes and proteins that were tagged. Our results demonstrate that CD-tagging is an efficient means to functionally annotate large numbers of proteins by characterizing their natural abundances, localization patterns, and dynamics in living cells.

MATERIALS AND METHODS

Stealth 1.0 Vector

Stealth 1.0 was constructed in pDO Δ MP, a self-inactivating derivative of the Moloney murine leukemia virus (MMLV)-based retroviral vector pDON-AI (Takara Bio, Shiga, Japan) using a variety of standard methods. DNA sequences within the vector were derived from synthetic oligonucleotides or from PCR-derived fragments of the vector pJJ225 (8), the reading frame-independent vector described by Nelson et al. (13), and pEGFP-N1 (BD Biosciences Clontech, Palo Alto, CA,

ABSTRACT

A self-inactivating CD-tagging retroviral vector was used to introduce epitope and GFP tags into genes and proteins in NIH 3T3 cells. Several hundred cell clones, each expressing GFP fluorescence in a distinctive pattern, were isolated. Molecular analysis showed that a wide variety of genes and proteins, some known and some newly discovered, had been tagged. The analysis also revealed that, in the great majority of instances, the abundance and cellular location of the tagged protein mirrored that of its untagged counterpart. This approach provides a systematic means for the functional annotation of mammalian genomes and proteomes in living cells.

USA). Stealth 1.0 was designed to function when it is inserted into class 0 introns (8). The sequence and detailed annotation of the complete retroviral portion of Stealth 1.0 has been deposited in GenBank® (accession no. AF515704).

Retrovirus Preparation and Cell Infections

Standard medium was DMEM containing 10% FCS, 100 U/mL penicillin, and 100 µg/mL streptomycin. The standard incubation conditions were 37°C under 5% CO₂. Flasks (75 cm²) were inoculated with 4 × 10⁵ pStealth1.0 stable transfectants of Phoenix cells in 10 mL standard medium and grown to approximately 70% confluence (four days). Culture medium was replaced with 10 mL standard medium containing 2% FCS. Viral preparations were made by filtering the supernatant media from these flasks after an additional 48-h incubation through a Millex-HV 0.45-µm syringe filter (Millipore, Bedford, MA, USA).

To generate infected cells, 25-cm² flasks were inoculated with 4 × 10⁴ NIH 3T3 cells in 5 mL standard medium and grown for 48 h. Polybrene (hexadimethrine bromide; Sigma, St. Louis, MO, USA) was added to viral preparations to a final concentration of 8 µg/mL. The culture medium was removed from the NIH 3T3 cells and replaced with 5 mL viral preparation for 4 h. Culture supernatant was removed and replaced with 10 mL standard medium. After 24 h additional incubation, the cells were trypsinized, and 2 × 10⁴ cells were re-plated onto 40-mm round coverslips in 5 mL standard medium in 60-mm Petri dishes and incubated until microscopic examination.

Purification of Tagged Clones

Tagged NIH 3T3 cell lines were identified by fluorescence microscopy 5–6 days after infection as small clusters (8–32 cells) of GFP-positive cells. The number of such positive clusters averaged four per coverslip. Labeled cells were subcloned on the microscope stage using 4.7 × 8 mm cloning cylinders (Scienceware; Fisher Scientific, Pittsburgh, PA, USA) and re-plated at

high dilution. After several rounds of subcloning, most of the cell lines were sorted for GFP fluorescence using a Coulter Elite Fluorescence-Activated Cell Sorter (Beckman Coulter, Miami, FL, USA). For cell lines whose fluorescence intensity was barely above autofluorescence levels, subcloning had to be done entirely by cloning cylinders. In these cases, subcloning was facilitated by the removal of unlabeled cells by 532-nm laser ablation. The laser ablation system (2) was built around an Axiovert® 135 microscope (Carl Zeiss, Thornwood, NY, USA) and a 300-mJ Surlite II pulsed laser (Continuum, Santa Clara, CA, USA). All subsequent studies were performed on the isolated cloned cell lines. We found the clones to be quite stable; however, re-sorting (or subcloning) was occasionally performed after several months of cell growth to eliminate cells that did not express the GFP-tagged protein at the original level.

Fluorescence Microscopy

Cell screening by fluorescence microscopy was performed using an Axiovert IM35 microscope (Carl Zeiss) equipped with GFP filters (Chroma Technology, Brattleboro, VT, USA). Candidate cells were photographed using a 12-bit cooled Photometrics 250™ charge-coupled device (CCD) camera (Photometrics, Tucson, AZ, USA) using BDS software (BioDetection Systems, Pittsburgh, PA, USA).

Fluorescence confocal images were taken with a spinning confocal imaging system (Solamere Technology Group) mounted onto the side port of an Olympus IX50 microscope. The system consists of a Yokogawa CSU10 Confocal Scanner Unit, illuminated via optical fiber from a LaserPhysics Reliant 100s-488 Argon laser. Images were captured by a Roper Scientific/Photometrics CoolSnap HQ Cooled CCD Camera. Image processing and capturing were controlled by QED software (QED Imaging, Pittsburgh, PA, USA) on an Apple® Power Mac G4 computer.

RNA Isolation

Tagged NIH 3T3 cells (0.5–1.0 × 10⁶) were harvested by centrifugation

at 10 000× *g* and washed once in PBS. RNA was isolated from the cell pellet using the RNeasy® Mini Kit (Qiagen, Valencia, CA, USA), according to the manufacturer's protocol. Total RNA was obtained in a final volume of 35 µL. One microliter (20–40 U) RNasin® Ribonuclease Inhibitor (Promega, Madison, WI, USA) was added to the RNA that was then stored at -80°C.

cDNA Synthesis

Primer annealing was accomplished by incubating a mixture of 16 µL RNA, 1 µL 10 mM dNTP solution, and 1 µL 50 µM oligo(dT)-containing, 3'-RACE primer (5'-GCTGTCAACGATACGC-TACGTAACGGCATGACAGTGT₁₈-3') at 65°C for 5 min, followed by quick chilling on ice. First-strand cDNA was synthesized by the addition of 5 µL 5× first-strand buffer (250 mM Tris-HCl, pH 8.3, 375 mM KCl, 15 mM MgCl₂), 1 µL 100 mM DTT, and 1 µL SuperScript™ II Reverse Transcriptase (Invitrogen, Carlsbad, CA, USA) to the annealed mixture and incubation at 42°C for 50 min. Reverse transcriptase was inactivated by incubation at 70°C for 15 min. The RNA was removed by digestion with 2 U *E. coli* RNase H (Promega) at 37°C for 20 min. The cDNA was stored at -20°C.

Nested 3'-RACE PCR

First-stage 3'-RACE reaction (50 µL) contained 1 µL cDNA, 200 µM dNTPs, 1.25 U *Taq* DNA polymerase (Brinkmann Instruments, Westbury, NY, USA), 400 nM 3' primer GR3'P (5'-GCTGTCAACGATACGCTACGT-AACG-3'), 400 nM 5' GFP primer 1H (5'-GCAGAAGAACGGCATCAAGG-TGAAC-3') in 1× *Taq* DNA polymerase buffer (50 mM KCl, 10 mM Tris-HCl, pH 8.3, 1.5 mM MgCl₂, 0.1% Triton® X-100). Twenty cycles of a "semi-touchdown" program were used after a 2-min start incubation at 94°C. The conditions were as follows: five cycles of 94°C for 45 s and 72°C for 3 min; 5 cycles of 94°C for 45 s, 70°C for 45 s, and 72°C for 3 min; and 10 cycles of 94°C for 45 s, 68°C for 45 s, and 72°C for 3 min. This program was followed by an 8-min incubation at 72°C, and the reactions were then held at 12°C.

Research Report

Second-stage 3' RACE reaction (50 μ L) contained 1 μ L of the first-stage product, 200 μ M dNTPs, 1.25 U *Taq* DNA polymerase, 400 nM 3' primer GR3'NP (5'-CGCTACGTAACGGCAGTGACAGTG-3'), 400 nM 5' GFP primer 2H (5'-ACTACCAGCAGAA-CACCCCATC-3') in 1 \times *Taq* DNA polymerase buffer (50 mM KCl, 10 mM Tris-HCl, pH 8.3, 1.5 mM MgCl₂, 0.1% Triton X-100). Thirty-five cycles of 94°C for 45 s, 71.1°C for 45 s, and 72°C for 3 min were used after a 2-min start incubation at 94°C. After cycling, the reaction was incubated at 72°C for 8 min and then held at 12°C.

DNA Sequencing

Second-stage 3'-RACE PCRs that demonstrated a single band when analyzed by gel electrophoresis were sequenced directly. A 40- μ L portion of the reaction was purified using the QIAquick™ PCR Purification Kit (Qiagen) according to the manufacturer's protocol. An 8- μ L portion of the purified PCR product was mixed with 3.4 pmol of the GFP sequencing primer 1S (5'-GATCACATGGTCCTGCTGG-3') and sequenced in the University of Pittsburgh DNA Sequencing Core Facility using an ABI PRISM® 3100 Genetic Analyzer (Applied Biosystems, Foster City, CA, USA).

Preparative gel electrophoresis was performed on 40 μ L of the second-stage 3'-RACE PCRs that contained more than a single product. Individual bands were excised and purified with the QIAquick Gel Extraction Kit. A 2- μ L portion was cloned into the pCR®2.1-TOPO vector with TOPO® TA Cloning Kit (Invitrogen). Plasmid DNA was extracted from 4 mL of the overnight culture of individual TOPO clones with Wizard® Plus Miniprep DNA Purification System (Promega). An 8- μ L portion of the 60- μ L plasmid preparation was sequenced as described above.

RESULTS AND DISCUSSION

Stealth 1.0 CD-Tagging Vector

Figure 1 shows the retroviral-tagging vector Stealth 1.0. The vector contains two CD cassettes, one with a guest exon

Table 1. CD-Tagged Genes and Tag Locations Within Gene, Transcript, and Protein

Gene ^a	mRNA Insertion Point ^b	Protein Insertion Point ^c	Intron Tagged ^d
Actn1*	105 (2679) [13591901]*	35 (892) [13591902]*	Unknown
Adrp	30 (1278) [6680649]	10 (425) [12846852]	2 (7) [191730]
Anxa5	3 (960) [6753059]	1 (319) [6753060]	1 (12) [4007577]
Atf1	93 (810) [6680737]	31 (269) [6680738]	1 (5) [NT_009782]*
Atp5a1	60 (1662) [6680747]	20 (553) [6680748]	1 (11) [NT_028380]*
Atx1	6 (207) [6753135]	2 (68) [6753136]	1 (2) [NT_023152]*
Bat1	753 (1287) [4235115]	251 (428) [4235116]	EXON 6 [4809329] ^e
Beta15*	57 (1338) [57428]*	19 (445) [57429]*	Unknown
Cald1	597 (1593) [18043855]	199 (530) [18043856]	Unknown
Calm1	3 (450) [6753243]	1 (149) [12836015]	Unknown
Cav	30 (537) [6705976]	10 (178) [6705977]	1 (2) [6705976]
Clim1	96 (984) [13435938]	32 (327) [13435939]	1 (6) [NT_029381]*
Cnn2	63 (918) [6680951]	21 (305) [6680952]	Unknown
Col1a2	1215 (4122) [6680979]	405 (1373) [6680980]	Unknown
Dia1	18 (906) [14193687]	6 (301) [14193688]	Unknown
Ddx3	45 (1989) [6753619]	15 (662) [6753620]	1 (16) [NT_011793]*
Dxlmx46e	21 (1128) [7673615]	7 (375) [7673616]	Unknown
ETFB	48 (759) [12832366]	16 (252) [12832367]	1 (5) [NT_011091]*
FLJ10849*	27 (1290) [14727615]*	9 (429) [13630152]*	1 (9) [NT_006088]*
Fnbp3	165 (2862) [9055218]	55 (953) [9055218]	Unknown
Fxh	27 (1134) [16716398]	9 (377) [16716399]	Unknown
Grp58	1404 (1518) [13096983]	468 (505) [13096984]	12 (12) [NT_010194]*
Glut1	18 (1479) [193704]	6 (492) [309280]	1 (9) [220414, NT_004852]
Hmg17	15 (273) [8393533]	5 (90) [8393534]	Unknown
Hmga1	186 (291) [14198133]	62 (96) [14198134]	2 (3) [NT_007592]*
Hnrpa1	15 (963) [6754219]	5 (320) [6754220]	Unknown
Kif4	2901 (3696) [6680567]	967 (1231) [6680568]	2 (7) [NT_019696]*
Lasp1	249 (792) [6754507]	83 (263) [6754508]	3 (6) [NT_010647]*
Lgals1	9 (408) [6678681]	3 (135) [6678682]	1 (3) [NT_011520]*
Lmna	20 (1389) [9506842]	6 (462) [9506843]	Unknown ^f
Mrps18b	72 (765) [13620908]	24 (254) [13620909]	1 (6) [NT_007592]*
Msh2	1386 (2808) [726085]	462 (935) [726086]	Unknown
Ncl	18 (2124) [12802526]	6 (707) [12802527]	1 (13) [53453]
Nedd4a*	45 (2661) [1293646]*	15 (887) [1293647]*	Unknown
Nfix	27 (1203) [6754837]	9 (400) [6754838]	Unknown
Oxct	1248 (1563) [4557816]*	416 (520) [4557817]*	Unknown
Pes1	24 (1755) [11875634]	8 (584) [11875635]	1 (14) [NT_011520]*
Prim2	1299 (1518) [6679460]	433 (505) [6679461]	Unknown
Rab21	69 (147) [200622]	23 (49) [13177608]	Unknown
Rpl22	12 (387) [6677774]	4 (128) [6677775]	Unknown
Rpl29	102 (483) [12805206]	34 (160) [12805207]	2 (2) [7800211]
Rpl32*	96 (408) [6981481]*	32 (135) [6981482]*	1 (2) [NT_005718]*
Rpl36	93 (318) [9055321]	31 (105) [9055322]	1 (2) [NT_011169]*
Rps11	15 (477) [1938405]	5 (158) [1938406]	1 (4) [6552367]
Rps17	261 (408) [6677800]	87 (135) [6677801]	3 (4) [NT_029446]*
Rps4x	3 (792) [6677804]	1 (263) [6677805]	1 (6) [NT_011594]*
Sdc4	75 (597) [6755441]	25 (198) [6755442]	1 (4) [2373478]
Sdpr	483 (1257) [455718]	161 (418) [455719]	Unknown
Sep15	243 (489) [11139619]	81 (162) [11139620]	2 (4) [NT_004380]
Siahbp	78 (1698) [5524726]*	26 (565) [5524727]*	Unknown
Tmpo	660 (1359) [1335840]	220 (452) [1335841]	Unknown
Tpm4*	132 (747) [6981671]*	44 (248) [6981672]*	1 (7) [57371]*
Trx	24 (318) [6755910]	8 (105) [6755911]	1 (4) [517128]
Tuba1	3 (1356) [12805486]	1 (451) [12805487]	1 (3) [NT_005289]*
U17HG	27 (336) [3282045]	9 (112) [3282045]	2 (2) [3236112]
ZnBP*	42 (309) [14010874]*	14 (102) [14010875]*	Unknown

Research Report

Table 1. CD-Tagged Genes and Tag Locations Within Gene, Transcript, and Protein (Continued)

Gene ^a	mRNA Insertion Point ^b	Protein Insertion Point ^c	Intron Tagged ^d
Riken cDNA 1810057F21	143 (741) [12841726]	41 (246) [12841727]	Unknown
Riken cDNA 2610301D06	12 (1314) [16506250]	4 (437) [16506251]	Unknown
Riken cDNA 2610301D06	18 (954) [13385511]	6 (317) [13385512]	1 (6) [NT_009296]*
Riken cDNA 2700092018	270 (537) [12849451]	90 (178) [12849452]	Unknown
Riken cDNA 2810036K01	66 (1326) [13385537]	22 (441) [13385538]	Unknown

^aThe gene is identified by the highest BLAST score of the cDNA sequence immediately 3' to the GFP guest exon sequence. Asterisks mark the genes that were identified as homologs in species other than *Mus musculus*.

^bThe first number of each entry indicates the position in the coding sequence after which the guest exons are inserted. The number of nucleotides in the complete coding sequence in the mRNA (from start codon to stop codon) is shown in parentheses. The GenBank entry for the mRNA sequence is shown in brackets. Asterisks mark the genes that were identified as homologs in species other than *M. musculus*, and the insertion point is given with respect to that mRNA homolog.

^cThe first number of each entry indicates the amino acid position in the protein after which the guest peptides are inserted. The number of amino acids in the complete protein is shown in parentheses, and the GenBank entry for the protein is shown in brackets. Asterisks mark the genes that were identified as homologs in species other than *M. musculus*, and the insertion point is given with respect to that protein homolog.

^dThe first number of each entry indicates the intron into which the Stealth 1.0 vector was inserted. The number of introns in the gene is shown in parentheses. GenBank entries are shown in brackets. Unknown indicates the intron/exon structure of the gene is unavailable.

^ecDNA sequence indicated the Stealth 1.0 vector insertion is in exon 6 (see text for details).

encoding an epitope tag and the other with a guest exon encoding GFP. Both guest exons begin and end at codon boundaries and thus will successfully tag genes when inserted into class 0 introns (8). The vector contains a promoter/enhancer deletion in the U3 element

of its 3' long terminal repeat (LTR). This self-inactivation feature (22) was included in the Stealth 1.0 design to ensure that there would be no initiation of RNA synthesis from within the insert. Thus, any transcript containing the guest exon should originate from the

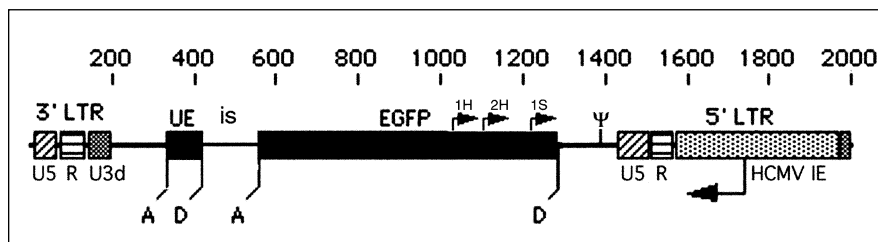


Figure 1. The structure of the self-inactivating Stealth 1.0 retroviral vector. The retroviral portion of the vector based on the MMLV vector pDON-AI is shown. The complete vector plasmid also contains a puromycin resistance gene outside of the retroviral portion that can be used to select stable transfectants of the vector in packaging cells. The universal epitope exon (UE) and the enhanced GFP exon (EGFP) are flanked by consensus splice acceptor (A) and donor (D) sites as previously described (8,19). Promoter and enhancer sequences in the U3 element of the 3' LTR have been deleted by in vitro manipulation, resulting in the transcriptionally-defective U3d element. The promoter and enhancer sequences in the U3 element of the 5' LTR have been replaced by the immediate early promoter of the human cytomegalovirus. The direction of transcription from this promoter is indicated by the large arrow pointing left. The U5 and R elements of the LTRs and the retroviral packaging signal (Ψ) are from MMLV. After the infection of the cells with retroviruses made by this vector, reverse transcription of the packaged viral RNA results in the transfer of the U3d element from the 3' LTR to the 5' LTR in the provirus integrated into target DNA. This results in the transcriptional inactivation of the 5' LTR. The location and orientation of the 3'-RACE PCR primers (1H and 2H) and sequencing primer (1S) are indicated as small arrows pointing right in the EGFP exon. Numbering is from the end of the U5 element in the 3' LTR.

natural promoter of the tagged gene.

Figure 2 is a schematic diagram of one of the tagged genes that we identified, the mouse Rpl29 gene in which Stealth 1.0 is resident in the 5'-end of intron 2. Transcription and splicing of the gene results in the addition of the epitope-tagged and GFP exons to the transcript between the second and third Rpl29 exons, which yields a tagged mRNA. The translation of the tagged mRNA yields a tagged protein.

Isolation of CD-Tagged Cells and Subcellular Localization of CD-Tagged Proteins

NIH 3T3 cells were infected with the Stealth 1.0 retrovirus, and several hundred cell clones expressing the GFP tag were isolated as described in the Materials and Methods section. The clones varied widely with respect to the intensity and subcellular location of their GFP fluorescence. The fluorescence patterns could be roughly grouped or categorized as follows: 49% were cytoplasmic or plasma membrane; 34% were nuclear (nucleoplasm, chromatin, nucleolus, or nuclear membrane); 6% were mitochondrial, endoplasmic reticulum, or golgi; 5% were cytoskeletal; 3% were cytoplasmic vesicles; and in 3% the fluorescent material appeared to be secreted. Figure 3 shows a panel of micrographs of representative tagged cells. Additional micrographs and cell line information are provided online at <http://www.andrew.cmu.edu/~berget/CDTagdatabase.html>.

Identification and Analysis of Tagged Genes

3' RACE was used to obtain several hundred bases of nucleotide sequence immediately 3' to the enhanced GFP (EGFP) guest exon for 61 independent, tagged cell clones. BLAST analysis of the sequences against the available databases at the National Center for Biotechnology Information (NCBI) revealed an exact match to a known mouse gene, mouse cDNA, or mouse expressed sequence tag (EST) in 53 cases. In the other eight cases, there was a close match to a gene, cDNA, or EST from another animal species. Table 1 lists these genes and shows the position of the insertion

Research Report

within the coding sequence of each tagged mRNA and within the amino acid sequence of each tagged protein. Some tags are near the N-terminus, some are in the middle, and some are near the C-terminus, with a bias toward locations close to the N-terminus. The basis for this bias is unknown.

The CD-tagging model (see Figure 2) makes two specific predictions with respect to the structure of the tagged genes and transcripts. (i) In each mRNA (or cDNA derived from the mRNA), the sequence immediately 3' of the EGFP guest exon should be the exact 5'-end of an existing exon in the target gene. (ii) Because the Stealth 1.0 vector was designed to functionally tag only class 0 introns, the intron that contains the tagging vector should be class 0. These predictions were satisfied by our data, both for the 10 cases for which the sequence of the mouse gene was available and for the 20 additional cases for which mouse genomic sequence was unavailable, but the sequence of a closely related mammalian gene was available. For the remaining 31 cases for which no NCBI database information existed for the tagged gene's exon/intron structure, our insertion locations predicted the location of a type 0 intron in the gene. In our data set, there was just one exception to the above expectations: the Bat1 gene insertion was inside exon 6 rather than in an intron. Preliminary genomic DNA sequence information from this cell line suggests that a fortuitous splice acceptor sequence may have been created at the downstream retroviral LRT-exon junction, allowing aberrant splicing to include the distal portion of exon 6 in the mRNA.

For the analysis of the cell lines reported here, we have no evidence to suggest that any of these cell lines contain more than one retroviral insert. We did not observe sequence heterogeneity in the cDNA sequence adjacent to the GFP exon in any of the cell lines. Thus, while we have not directly tested for multiple vector insertions in the lines (e.g., by Southern blot analysis of genomic DNA), we do not think it is likely that any of them is tagged in more than one expressed gene. Furthermore, none of the subcellular localization information presented below indicates multiple, overlapping fluorescent protein patterns.

Table 2. Cellular Location of CD-Tagged Proteins for which Localization Data for the Native Protein Is Available

Gene	Protein	Observed Protein Location	Expected Protein Location Based on the Literature
Actn1	Non-Muscle α -Actinin 1	Cytoplasm	Cytoskeleton
Adrp	Adipose Differentiation Related Protein	Vesicles	Vesicles
Anx5	Annexin V	Nucleus, Cytoplasm	Nucleus, Cytoplasm
Atf1	Activating Transcription Factor 1	Nucleus	Nucleus
Atp5a1	ATP Synthase, isoform 1	Mitochondria	Mitochondria
Atx1	Antioxidant Protein 1	Cytoplasm	Cytoplasm
Bat1	Nuclear RNA Helicase	Nucleus	Nucleus
Beta15	β -tubulin isoform	Cytoskeleton	Cytoskeleton
Cald1	Caldesmon 1	Cytoskeleton	Cytoskeleton
Calm1	Calmodulin	Cytoplasm	Cytoplasm
Cav	Caveolin-1	Cell Surface	Cell Surface
Clim1	C-terminal LIM domain containing protein, Elfin	Cytoskeleton	Cytoskeleton
Cnn2	Calponin 2	Cytoplasm	Cytoskeleton
Col1a2	Procollagen Type I α 2	Cytoplasm, Vesicles	Endoplasmic Reticulum
Ctsj	Cathepsin J	Vesicles	Vesicles
Ddx3	DEAD Box Polypeptide 3	Nucleus, Cytoplasm	Nucleus, Cytoplasm
Dia1	Cytochrome b-5 reductase	Cytoplasm	Cytoplasm
Grp58	ER60 Protease/Glucose Regulated Protein	Endoplasmic Reticulum	Endoplasmic Reticulum
Glut1	Type 1 Glucose Transporter	Plasma Membrane, Vesicles	Plasma Membrane, Vesicles
Hmg17	High Mobility Group Protein 17	Nucleus	Nucleus
Hmga1	High Mobility Group AT-hook 1	Nucleus	Nucleus
Hnrpa1	Heterogeneous Nuclear Ribonucleoprotein A1	Nucleus	Nucleus
Kif4	Kinesin Heavy Chain Member 4	Nucleus, Cytoplasm	Nucleus, Cytoplasm
Lasp1	LIM and SH3 domain containing protein	Cytoplasm	Cytoplasm
LgalS1	Galectin-1	Cytoplasm, Nucleus	Cytoplasm, Nucleus
Lmna	Lamin A	Nucleus	Nuclear Membrane, Nucleus
Mrps18b	Mitochondrial Ribosomal Protein S18B	Mitochondria	Mitochondria
Msh2	MutS Homolog 2	Nucleus, Cytoplasm	Nucleus, Cytoplasm
Ncl	Nucleolin	Nucleolus	Nucleolus
Nedd4a	Ubiquitin Protein Ligase	Cytoplasm	Cytoplasm
Nfix	Nuclear Factor I-X	Nucleus	Nucleus
Oxct	Succinyl CoA Transferase	Mitochondria	Mitochondria
Pes1	Pescadillo homolog 1	Nucleolus	Nucleolus
Rab21	GTP-binding protein Rab21	Endoplasmic Reticulum-like	Endoplasmic Reticulum-like
Rpl22	Ribosomal Protein L22	Nucleolus, Cytoplasm	Nucleolus, Cytoplasm
Rpl29	Ribosomal Protein L29	Nucleolus	Nucleolus, Cytoplasm
Rpl32	Ribosomal Protein L32	Nucleolus	Nucleolus, Cytoplasm
Rpl36	Ribosomal Protein L36	Nucleolus, Cytoplasm	Nucleolus, Cytoplasm
Rps11	Ribosomal Protein S11	Nucleolus, Cytoplasm	Nucleolus, Cytoplasm
Rps17	Ribosomal Protein S17	Nucleolus, Cytoplasm	Nucleolus, Cytoplasm

Table 2. Cellular Location of CD-Tagged Proteins for which Localization Data for the Native Protein Is Available (Continued)

Gene	Protein	Observed Protein Location	Expected Protein Location Based on the Literature
Rps4x	Ribosomal Protein S4	Nucleolus	Nucleolus, Cytoplasm
Sdc4	Syndecan 4	Plasma Membrane	Plasma Membrane
Sdrp	Serum Deprivation Response Protein	Plasma Membrane	Plasma Membrane
Sep15	15-kDa Selenoprotein	Endoplasmic Reticulum	Endoplasmic Reticulum
Tmpo	Thymopoietin β	Nuclear Membrane	Nuclear Membrane
Tpm4	Tropomyosin 4	Cytoplasm	Cytoskeleton
Trx	Thioredoxin	Nuclear Membrane, Nucleus, Cytoplasm	Nuclear Membrane, Nucleus, Cytoplasm
Tuba1	α -Tubulin	Cytoskeleton	Cytoskeleton
ZnBP	Parathyrosin	Nucleus	Nucleus

CD-Tagged Proteins Generally Exhibit Normal Subcellular Localization and Retain Normal Biological Function

Tables 2 and 3 list the observed subcellular locations of the tagged proteins in broad categories for all cases in which the tagged gene was identified. For a large number of the genes (49 of 61; 80%), we found at least one published paper describing subcellular localization of the gene product or a close relative of it in a mammalian cell type. In the great majority of these cases (42 of 49; 86%), the published localization pattern was concordant with our observations. We conclude that these CD-tagged proteins

retain normal function with respect to protein localization within the cell.

The tagged proteins were not only generally localized to the expected cellular compartments but also were associated in most cases with the expected structural elements within those compartments. Specific examples include tagged caldesmon (associated with actin filaments), tagged α -tubulin (associated with microtubules), tagged nucleolin (associated with nucleoli), tagged hmgal and hmg17 proteins (associated with chromosomes), and tagged ribosomal proteins Rpl22, Rpl36, Rps11, and Rps17 (associated with nucleoli and cytoplasm). In these cases, we can conclude that the tags do

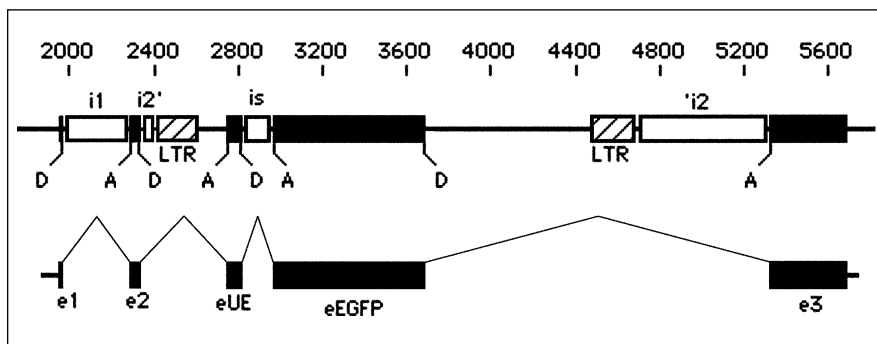


Figure 2. The top line is a diagram of the Stealth 1.0 vector insertion into the 3-exon *M. musculus* large ribosomal protein 29 gene (Rp129). Splice acceptor (A) and donor (D) sites in both the Rp129 gene and the Stealth 1.0 vector are shown. Exons (e) and introns (i) are depicted as black and white rectangles, respectively. The transcriptionally inactive proviral LTRs are shown as hatched rectangles. The precise location of the Stealth 1.0 vector insertion in the Rp129 gene was determined by DNA sequencing across both LTR/intron 2 junctions (data not shown). The bottom line is the putative mRNA splicing pattern of the tagged Rp129 gene. The correct splicing of the tagged Rp129 gene requires the removal of Rp129 intron 1 (i1), the synthetic intron (is) between the UE and EGFP exons (eUE and eEGFP) from the Stealth 1.0 provirus, and the two chimeric introns created from the Rp129 intron 2 by the integration of the Stealth 1.0 vector (i2' and i2). The Stealth 1.0 insertion occurs immediately after nucleotide 2412, splitting intron 2 into two parts. The numbering system up to nucleotide 2412 is from the GenBank sequence for the Rp129 gene [accession no. AF236069 (9)].

Research Report

Table 3. Cellular Location of CD-Tagged Proteins for which Localization Data for the Native Protein Is Unavailable

Gene	Protein or Description	Observed Protein Location
Riken cDNA 1810057F21	Unknown	Nucleolus
Riken cDNA 2610301D06	Homology to human translation elongation factor 1 gamma	Cytoplasm
Riken cDNA 2700092O18	Unknown	Cytoplasm
Riken cDNA 2810036K01	Unknown	Nucleolus
DXlmx46e	Unknown	Nucleus
FLJ10849	Unknown	Cytoskeleton
Fnbp3	Formin Binding Protein 3	Nucleus
Fxh	Fox-1 Homolog (putative RNA-binding protein)	Nucleus
Prim2	DNA Primase, p58 Subunit	Cytoplasm
RP1-302G2	Unknown	Cytoplasm
RP23-278K23	Unknown	Cell Surface
Siahbp1	Binding protein to homolog of Drosophila seven in absentia	Nucleus
U17HG	None	Cytoplasm

not interrupt regions of the proteins that are required for interaction with the appropriate structural elements.

In some cases (5 of 49; 10%), the tagged protein did not show the expected localization pattern. For example, tagged calponin showed diffuse cytoplasmic localization, whereas calponin itself is known to be associated with the actin cytoskeleton. We presume that the tag interferes, directly or indirectly, with the proper localization of the tagged protein to the corresponding structure.

In a few other cases, only a partial loss of interaction or partial interference with localization was observed. As described earlier, four of the seven tagged ribosomal proteins showed the expected nucleolar and cytoplasmic localization. The other three (Rpl29, Rpl32, and Rps4x) showed only nucleolar localization. This indicates that while the translation of these proteins and their transport and localization to the nucleoli appear normal, some step in their subsequent assembly into ribosomes and their transport out of the nucleoli into the cytoplasm is impaired.

For a significant fraction of the cases (12 of 61; 20%), we could find no publication in the literature that described the location of the protein within the cell; indeed, for most of these,

only a cDNA sequence was available. For these cases, presented in Table 3, our data represent the first information about the cellular locations of the proteins encoded by the respective genes.

The average number of introns in a human gene is between six and eight (10). With this many potential tagging sites available, we expect that it will be possible to CD-tag a large fraction of the proteome in a way that does not interfere with the function of the individual tagged proteins.

CD-Tagged Protein and Transcript Abundance

Our CD-tagged proteins retain their entire polypeptide sequences, so they should retain any sites for proteolysis or posttranslational modification, such as ubiquitination or phosphorylation, which could affect the *in vivo* stability of the protein. Therefore, we expect the cellular levels of the tagged proteins to be equivalent, for the most part, to those of their untagged counterparts. Exceptions could occur if the tag interrupts sites of modification or proteolysis or if the tagged protein fails to fold properly and is prematurely degraded. Our data generally support the expectation of normal protein abundance for

Table 4. CD-Tagged Gene Representation in NIH 3T3 SAGE Data

Gene or Gene Designation	Hits in SAGE Dataset	Percent of Entries
LgalS1	674	2.597
Col1a2	131	0.505
Tuba1	108	0.416
Rps11	49	0.189
Rps6	48	0.185
Rpl36	46	0.177
Rpl29	39	0.150
RIKEN cDNA 2610301D06	38	0.146
Grp58	13	0.050
Calm1	13	0.050
Cnn2	9	0.035
Nfix	8	0.031
Cox8a	7	0.027
Anxa5	7	0.027
Pes1	6	0.023
Ncl	6	0.023
Hnrpa1	6	0.023
Fxh	6	0.023
Rps4x	5	0.019
Bat1	5	0.019
Lasp1	4	0.015
Nedd4a	3	0.012
Hmg17	3	0.012
Kif4	2	0.008
Glut1	2	0.008
Tmpo	1	0.004
Ddx3	1	0.004
Cav	1	0.004
Atp5a1	1	0.004
Adfp	1	0.004

the tagged proteins, in that the GFP intensity was generally in accord with the expected abundance of the protein, based on the available literature.

Because CD-tagged genes retain their normal promoters and other transcriptional regulatory elements, we also expect cellular levels to be normal for the tagged transcripts. Although we did not measure directly tagged transcript levels, we did examine a serial analysis of gene expression (SAGE) database (20) of about 25 000 sequence tags that represented about 3000 expressed genes in NIH 3T3 cells (<http://www.sagenet.org/SAGEData/3T3.htm>) to determine whether our tagged genes were represented. Of the 61 tagged genes for which we obtained sequence data, 30 transcripts were present among the approximately 3000 transcripts in the SAGE dataset. Several of these tran-

Research Report

scripts were abundant [e.g., transcripts for galectin (Lgals1), α -tubulin (Tuba1), and the ribosomal protein genes Rps11, Rps6, Rpl36, and Rpl29, but others were represented only once or a few times in the dataset (Table 4). Thirty-one of the tagged genes were not listed in the SAGE dataset at all, indicating that their transcripts are of relatively low abundance. We conclude that visualization of CD-tagged proteins in live cells is not severely biased in favor of those encoded in high-abundance transcripts.

Comparison to Other In Vivo Approaches for Annotating Protein Function

Several studies have annotated gene and protein functions by constructing 5' or 3' fusions between cDNA and GFP sequences, expressing the fusion genes in cells or whole organisms, and observing the GFP-tagged fusion proteins in live cells by fluorescence microscopy (4,11,15,16). This approach differs from CD-tagging in that the fu-

sion protein is expressed from a heterologous promoter (usually a strong promoter of viral origin) rather than from the promoter that is associated with the natural gene. An advantage of gene overexpression is that the fusion protein can be visualized in the cell even when it would normally be present at too low a level to be seen or not be present at all. However, overexpression comes with a risk because it may lead to protein mislocalization or produce unexpected effects on the cell as a whole, thereby yielding misleading results. CD-tagging, in contrast, does not alter natural regulation so that it provides the opportunity to study authentic cell-type-specific and cell-stage-specific protein expression in vivo.

Gene trapping (5,6) is another gene annotation approach with similarities to CD-tagging. Gene trapping, like CD-tagging, depends on the insertion of a

foreign DNA into an intron and the splicing of a reporter-encoding guest exon into the transcript. However, in gene trapping, the DNA insert includes transcriptional and translational termination sequences so that the guest exon becomes the 3'-terminal exon of the transcript. Translation of the tagged message yields a deletion/substitution fusion protein, with its N-terminal portion derived from the target gene and its C-terminal portion derived from the guest exon. Because such deletion/substitution fusion proteins generally lack normal biological activity, gene trapping has been used extensively to create gene knockouts (1,23). Gene trapping has also been used to examine the subcellular location of the trapped gene products, using β -galactosidase or GFP reporters encoded in the guest exon (17,18). The approach works in cases in which the retained N-terminal portion

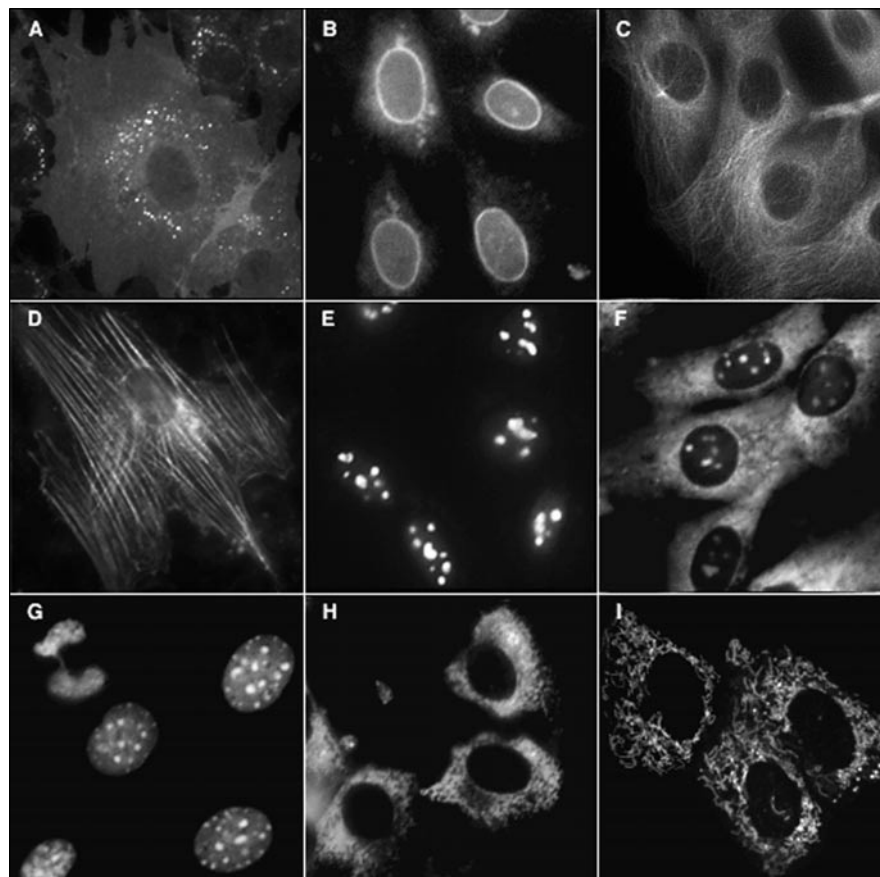


Figure 3. Fluorescence micrographs of live NIH 3T3 cell lines CD-tagged in specific genes with a compound epitope and GFP tag. (A) Glut1 gene (type 1 glucose transporter). (B) Tmpo gene (thymopoietin β). (C) Tuba1 gene (α -tubulin). (D) Cald gene (caldesmon 1). (E) Ncl gene (nucleolin). (F) Rps11 gene (ribosomal protein S11). (G) Hmgal1 gene (high mobility group AT-hook 1). (H) Col1a2 gene (procollagen type I α 2). (I) Atp5a1 gene (ATP synthase isoform 1).

of the fusion protein is sufficient for correct localization, but it fails whenever correct localization depends on the deleted C-terminal portion of the protein. In contrast, CD-tagging leaves the entire amino acid sequence of the target protein intact, significantly increasing the likelihood that normal biological function is retained.

Further Studies and Applications

The GFP tags in our cell clones allow for the direct observation of protein dynamics in living cells through, for example, the generation of real-time and time-lapse movies. The opportunity to track the spatial and temporal comings and goings of naturally regulated individual proteins in living cells should make CD-tagged cell lines valuable for use in a wide variety of cell-based analyses and assays.

In addition to the GFP tag, the Stealth 1.0 vector encodes an epitope tag that is incorporated into each tagged protein. The epitope tag has a number of virtues. The tag can be used in conjunction with its monoclonal antibody to visualize tagged proteins in Western blots of cellular proteins, thereby confirming or determining the molecular weight of the tagged protein. It can also serve as an affinity tag whereby tagged proteins can be purified for biochemical or physical analysis (13). For the latter purpose, mass spectrometry can play a major role because it can be used to identify tagged proteins, detect post-translational modifications, and detect and identify other proteins that are physically associated with the tagged protein. The epitope tag can also be used to visualize tagged proteins at the ultrastructure level using immuno-electron microscopy.

The fluorescence images that we obtained using the methods described here contain a wealth of morphological detail. These images are well suited to analysis using automated methods for image classification and comparison—methods that dramatically extend the reach of our approach. Application of such methods to our analysis will be described elsewhere.

The results reported here demonstrate the effectiveness of random CD-tagging in NIH 3T3 cells. However, we know that with our present methods, many tagged cells are missed because the tagged protein is not readily detected above background cell autofluorescence, it is present for a limited portion of the cell cycle, or because the tagged gene is not expressed in NIH 3T3 cells. To improve the versatility of our approach, a variety of improvements to the system are under exploration. These

Research Report

include the incorporation of elements in the vectors that allow for the positive selection of intron insertions (7), the use of fluorescent proteins of greater intensity and/or of different spectral properties than EGFP (14), and the use of delivery vectors that allow for tagging of non-proliferating cells (21).

While NIH 3T3 cells have provided an excellent test bed for examining the effectiveness of CD-tagging in mammalian cells, the fact that these cells are aneuploid, with chromosome number and composition in flux, presents problems in comparing different tagged lines and in comparing different subclones within the same line. Furthermore, many cell-type-specific proteins are simply not expressed in NIH 3T3 cells so that they cannot be studied at all in these cells using our approach. Fortunately, a wide variety of other cell types including primary cultured cells and stem cells may also be tagged using the approach described here. A particularly attractive possibility is the tagging of embryonic stem cells, since they can be converted into transgenic animals in which tagged gene expression can be observed and analyzed in any cell type or tissue.

ACKNOWLEDGMENTS

The authors acknowledge Jeff Bray, Bruce Taillon, and Cheryl Telmer for insightful commentary and for performing initial experiments exploring CD-tagging with retroviral delivery vectors, and Song-Hee Kim for assistance in the construction of the Stealth 1.0 vector. We also thank Bob Murphy, Paul Robbins, Byron Ballou, and Chris Szent-Gyorgyi for additional insightful advice and commentary. This work was supported by the National Institutes of Health grant no. CA83219 to J.W.J. from the National Cancer Institute and funds from the National Science Foundation Science and Technology grant no. MCB-8920118 to Alan Waggoner.

REFERENCES

1. Amsterdam, A., S. Burgess, G. Golling, W. Chen, Z. Sun, K. Townsend, S. Farrington, M. Haldi, and N. Hopkins. 1999. A large-

scale insertional mutagenesis screen in zebrafish. *Genes Dev.* 13:2713-2724.

2. Berns, M.W., Y. Tadir, H. Liang, and B. Tromberg. 1998. Laser scissors and tweezers. *Methods Cell Biol.* 55:71-98.

3. Boland, M.V. and R.F. Murphy. 2001. A neural network classifier capable of recognizing the patterns of all major subcellular structures in fluorescence microscope images of HeLa cells. *Bioinformatics* 17:1213-1223.

4. Cutler, S.R., D.W. Ehrhardt, J.S. Griffiths, and C.R. Somerville. 2000. Random GFP::cDNA fusions enable visualization of subcellular structures in cells of Arabidopsis at a high frequency. *Proc. Natl. Acad. Sci. USA* 97:3718-3723.

5. Friedrich, G. and P. Soriano. 1991. Promoter traps in embryonic stem cells: a genetic screen to identify and mutate developmental genes in mice. *Genes Dev.* 5:1513-1523.

6. Gossler, A., A.L. Joyner, J. Rossant, and W.C. Skarnes. 1989. Mouse embryonic stem cells and reporter constructs to detect developmentally regulated genes. *Science*. 244:463-465.

7. Ishida, Y. and P. Leder. 1999. RET: a poly A-trap retrovirus vector for reversible disruption and expression monitoring of genes in living cells. *Nucleic Acids Res.* 27:e35.

8. Jarvik, J.W., S.A. Adler, C.A. Telmer, V. Subramaniam, and A.J. Lopez. 1996. CD-tagging: a new approach to gene and protein discovery and analysis. *BioTechniques*. 20:896-904.

9. Kirn-Safran, C.B., S. Dayal, P.A. Martin-DeLeon, and D.D. Carson. 2000. Cloning, expression, and chromosome mapping of the murine Hip/Rpl29 gene. *Genomics* 68:210-219.

10. Lander, E.S., L.M. Linton, B. Birren, C. Nusbaum, M.C. Zody, J. Baldwin, K. Devon, K. Dewar, et al. 2001. Initial sequencing and analysis of the human genome. *Nature* 409:860-921.

11. Misawa, K., T. Nosaka, S. Morita, A. Kaneko, T. Nakahata, S. Asano, and T. Kitamura. 2000. A method to identify cDNAs based on localization of green fluorescent protein fusion products. *Proc. Natl. Acad. Sci. USA* 97:3062-3066.

12. Morin, X., R. Daneman, M. Zavortink, and W. Chia. 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.

13. Nelson, R.W., J.W. Jarvik, B.E. Taillon, and K.A. Tubbs. 1999. BIA/MS of epitope-tagged peptides directly from *E. coli* lysate: multiplex detection and protein identification at low-femtomole to subfemtomole levels. *Anal. Chem.* 71:2858-2865.

14. Peelle, B., T.L. Gururaja, D.G. Payan, and D.C. Anderson. 2001. Characterization and use of green fluorescent proteins from *Renilla mulleri* and *Ptilosarcus guernei* for the human cell display of functional peptides. *J. Protein Chem.* 20:507-519.

15. Rolls, M.M., P.A. Stein, S.S. Taylor, E. Ha, F. McKeon, and T.A. Rapoport. 1999. A visual screen of a GFP-fusion library identifies a new type of nuclear envelope membrane protein. *J. Cell Biol.* 146:29-44.

16. Simpson, J.C., R. Wellenreuther, A. Poustka, R. Pepperkok, and S. Wiemann. 2000. Systematic subcellular localization of novel proteins identified by large-scale cDNA sequencing. *EMBO Rep.* 1:287-292.

17. Sutherland, H.G., G.K. Mumford, K. Newton, L.V. Ford, R. Farrall, G. Delleire, J.F. Caceres, and W.A. Bickmore. 2001. Large-scale identification of mammalian proteins localized to nuclear sub-compartments. *Hum. Mol. Genet.* 10:1995-2011.

18. Tate, P., M. Lee, S. Tweedie, W.C. Skarnes, and W.A. Bickmore. 1998. Capturing novel mouse genes encoding chromosomal and other nuclear proteins. *J. Cell Sci.* 111:2575-2585.

19. Telmer, C.A., P.B. Berget, B. Ballou, R.F. Murphy, and J.W. Jarvik. 2002. Epitope tagging genomic DNA using a CD-tagging Tn10 minitransposon. *BioTechniques* 32:422-430.

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

21. Xu, K., H. Ma, T.J. McCown, I.M. Verma, and T. Kafri. 2001. Generation of a stable cell line producing high-titer self-inactivating lentiviral vectors. *Mol. Ther.* 3:97-104.

22. Yu, S.F., T. von Ruden, P.W. Kantoff, C. Garber, M. Seiberg, U. Ruther, W.F. Anderson, E.F. Wagner, and E. Gilboa. 1986. Self-inactivating retroviral vectors designed for transfer of whole genes into mammalian cells. *Proc. Natl. Acad. Sci. USA* 83:3194-3198.

23. Zambrowicz, B.P., G.A. Friedrich, E.C. Buxton, S.L. Lilleberg, C. Person, and A.T. Sands. 1998. Disruption and sequence identification of 2000 genes in mouse embryonic stem cells. *Nature* 392:608-611.

Received 22 April 2002; accepted 29 May 2002.

Address correspondence to:

Drs. Peter B. Berget and Jonathan W. Jarvik
Department of Biological Sciences
4400 Fifth Avenue
Carnegie Mellon University
Pittsburgh, PA 15213, USA
e-mail: berget@andrew.cmu.edu

For reprints of this or
any other article, contact
Reprints@BioTechniques.com