Generation of Expression Clone Set for Functional Proteomics of Human Gastric and Liver Cancers

Nang-Soo Oh,* Ji-Seon Park,* Yeo-Jin Jeon,* Jung-Hwa Oh,* So-Young Jeong,* Jin-Ok Yang,† Yong-won Park,‡ Hyang-Sook Yoo,* and Nam-Soon Kim*,[1]

*Laboratory of Human Genomics, Genome Research Center, and †Korean Bioinformation Center (KOBIC), KRIBB, Daejeon 305-806, Korea; and ‡CoreBioSystem Co., Ltd, Seoul 150-073, Korea

Two thousand sixty-eight multi-purpose expression clones for the 326 candidate genes related to gastric or liver cancers were constructed using the Gateway system. These clones can be expressed as His, Glutathione-S-transferase (GST) or Enhanced version of the green fluorescent protein (EGFP) fusion proteins in *E. coli*, insect cells or mammalian cells. For the 246 *E. coli* expression clones, the GST fusion proteins had greater expression efficiency and solubility than the His fusion proteins. Approximately 20% of the expressed proteins had unexpected molecular weights. A detailed sequence analysis of these clones revealed frameshift mutations resulting from insertion, deletion or substitution of nucleotides. The results indicate that these changes in the candidate genes may affect the occurrence of gastric or liver cancers. In addition, when 105 proteins, which were expressed in *E. coli* at very low or undetectable levels, were expressed in insect cells, 76% of the proteins were expressed very well and most were soluble. We also found that most of the 30 proteins prepared using EGFP mammalian expression clones were localized to cellular compartments expected by Gene ontology (GO) and this localization was unaffected if the EGFP-fusion was at the N-terminal or C-terminal region of the protein. Antibody production and subcellular localization analysis of the candidate genes as well as a screen of genes involved in carcinogenesis pathways are currently in progress using these expression clones. These studies provide a valuable resource for developing a better understanding of the molecular mechanism of carcinogenesis in both gastric and liver cancer and would be very helpful in diagnosis and therapeutic predictions. Exp Biol Med 234:1220–1229, 2009

Key words: high-throughput expression; gastric cancer; liver cancer; Gateway system; functional study

Introduction

Gastric and liver cancers are among the leading causes of cancer deaths worldwide (1). They are particularly prevalent in Korea where they are the leading causes of cancer-associated death (2). While advances in diagnostic and treatment technologies have enabled us to offer excellent long-term survival for early-diagnosed gastric or liver cancer, the prognosis of advanced cancers still remains poor (3). Molecular analyses of gastric cancer have revealed a close relationship with genetic alterations in several genes, including *p53*, *APC*, *E-cadherin*, *beta-catenin*, *TGF-alpha*, *c-met*, *trefoil factor 1* and *Runx3* (4–7). Additionally, for liver cancer, reports indicate that mutations in p53, Rb and β-catenin, as well as the over-expression of c-myc and cyclin D1, are related to hepatocellular carcinoma (HCC), the most common primary cancer of the liver (8–11). Several growth factors, such as TGF α and β, also are implicated in the development of HCC (12, 13). Many gene expression profile analyses of gastric or liver cancer using microarrays have also been reported (14–17). However, the common carcinogenesis pathways and the subsequent progression of these cancers are not understood.

To elucidate the molecular mechanisms associated with gastric or liver carcinogenesis, we isolated gastric or liver cancer-related genes by the three approaches in the first phase of the "Project for Functional Analysis of Human Genome, Korea." These approaches are as follows: i) analysis of gene expression by examining Expressed sequence tags (ESTs) via single-pass sequencing of clones randomly selected from cDNA libraries of human gastric or liver cancer (18–20); ii) analysis of the expression profile of

This work was supported by the 21C Frontier Functional Human Genome Project from the Ministry of Science and Technology, by the technology-innovation-development project of the Small and Medium Business Administration, and by KRIBB Research Initiative Program in Korea.

[1] To whom correspondence should be addressed at Laboratory of Human Genomics, Genome Research Center, KRIBB, Daejeon 305–806, Korea. E-mail: nskim37@kribb.re.kr

Received December 21, 2008.
Accepted June 16, 2009.

DOI: 10.3181/0812-RM-371
1535-3702/09/23410-1220$15.00
Copyright © 2009 by the Society for Experimental Biology and Medicine

Downloaded from ebm.sagepub.com at SAGE Publications on November 3, 2014

genes expressed in human gastric or liver cancer via high-density cDNA microarray containing 38,000 unigene clusters (21, 22); and iii) analysis of the proteome expressed in human gastric or liver cancer via Two-dimensional electrophoresis (2-DE) (23, 24). Using these approaches, 883 genes were identified as candidate diagnosis markers and/or therapeutic targets of gastric or liver cancer. The focus of the next phase is to validate the cellular function of these candidate genes and to develop new diagnostics and therapies for gastric and liver cancer. To accomplish this, material resources such as expression clones, purified proteins and monoclonal antibodies for these candidate genes are essential. The first step in the production of such materials is to construct an expression system for the candidate genes in which proteins could be expressed under different conditions and in various hosts. This system requires a high-throughput-based DNA cloning technology enabling a large number of protein-encoding Open reading frames (ORFs) to be cloned precisely into different expression vectors.

The Gateway cloning system is a genome-wide recombinant protein expression system. It does not involve standard nucleic acid digestion and ligation but instead permits in vitro site-specific recombination using a universal system that is independent of expression vector function, host background or candidate target genes. This technique enables highly efficient and accurate directional cloning (25–28). Once Gateway tags are added to a gene, its subsequent sub-cloning into an expression system requires only a simple in vitro reaction that is amenable to high-throughput operation. The technique has been used to generate functional ORFeomes, such as *C. elegans* ORFeome version 1.1, human ORFeome version 1.1 and *P. falciparum* functional molecules, as well as to study protein-protein interactions and protein subcellular localization (29–33).

To generate resources for the functional study of candidate genes in human gastric or liver cancer, the Gateway cloning system was applied to construct expression clones of the candidate genes described above. Among the 883 genes of version 1, expression clones for 326 candidate genes, which have already been obtained as full-length cDNA by our EST collection project (18), were constructed. These expression clones can be expressed as His, GST or GFP fusion proteins in *E. coli*, insect cells or mammalian cells. The clones can be directly used in a mass scale cell-based assay for the functional study of candidate genes as well as in the mass scale preparation of proteins for antibody production. Here, we describe the system for constructing multi-purpose expression clones of the candidate genes using Gateway recombinational cloning and how this system is being used for the integrated study of functional proteomics of human gastric and liver cancers.

Materials and Methods

ORF Source of Cancer Candidate Genes. We used cDNAs constructed from 18 gastric samples and 15 liver samples as ORF sources of cancer candidate genes. The origin and nature of the gastric samples have been described previously (18–20).

Construction of the Entry Clones by PCR and BP Reaction. The forward primers were designed to have 18~22 bases on the ATG start initiator codon of the candidate genes, while the reverse primers were designed to have 18~22 bases not with the termination codon, in order to allow for the expression of N- and C-terminal fusion proteins. Using these primers, the ORFs of the candidate genes were first amplified using *Pfu* DNA polymerase (Promega, WI, USA) in 96-well plates and these amplified mixtures were then used as templates for secondary PCR reactions containing the *attB* adapter primers.

For construction of entry clones for candidate genes, each purified PCR product was incubated with the donor vector pDONOR201 in reaction solution containing BP clonase enzyme mix (Invitrogen Life Technologies, Carlsbad, CA) and were then transformed into *E. coli DH5*α in 96-well plates. The 5′ and 3′ regions of the positive entry clones were sequenced in order to confirm frameshift mutations of the ORF.

Construction of Expression Clones by LR Reaction. DNA from the entry clones was incubated with the appropriate expression vector in LR clonase enzyme mix (Invitrogen Life Technologies, Carlsbad, CA) at 25°C for 4 h. The used expression vector system is as follows: pDEST15 (N-GST) and pDEST17 (N-His) for *E. coli*, BacHTS (N-GST, Neurogenx, Daejeon, Korea) for insect cells, and pDEST-27 (N-GST), pcDNA-DEST53E (N-EGFP) and pcDNA-DEST47E (C-EGFP) for mammalian cells. The reactions, except for the insect vector, were transformed into *E. coli DH5*α in 96-well plates. The expression clones for gene were analyzed for the presence of the ORF by the colony PCR. The reaction conditions followed the manufacturer's instructions.

Recombinant Protein Expression in *E. coli*. *E. coli* expression clones were inoculated in M9ZB minimal medium of a 96-well plate at 37°C, treated with 0.1 mM IPTG, and then the cells were suspended in lysis buffer using general procedures. The used sample is as follows: the lysis solution as the total fraction, supernatants of lysis solution as the soluble fraction and resuspended solution of the remaining cell pellets as insoluble fraction. The samples were electrophoresed on 12% polyacrylamide SDS gels and their molecular weight (MW) calculated. If MW of the protein was in the range of ±10% compared to the expected MW, the protein was considered as a protein with the corrected MW.

Recombinant Protein Expression in Insect Cells. After infection of insect expression clones with Sf21 cells at a 20 MOI, the cells were suspended in lysis

Downloaded from ebm.sagepub.com at SAGE Publications on November 3, 2014

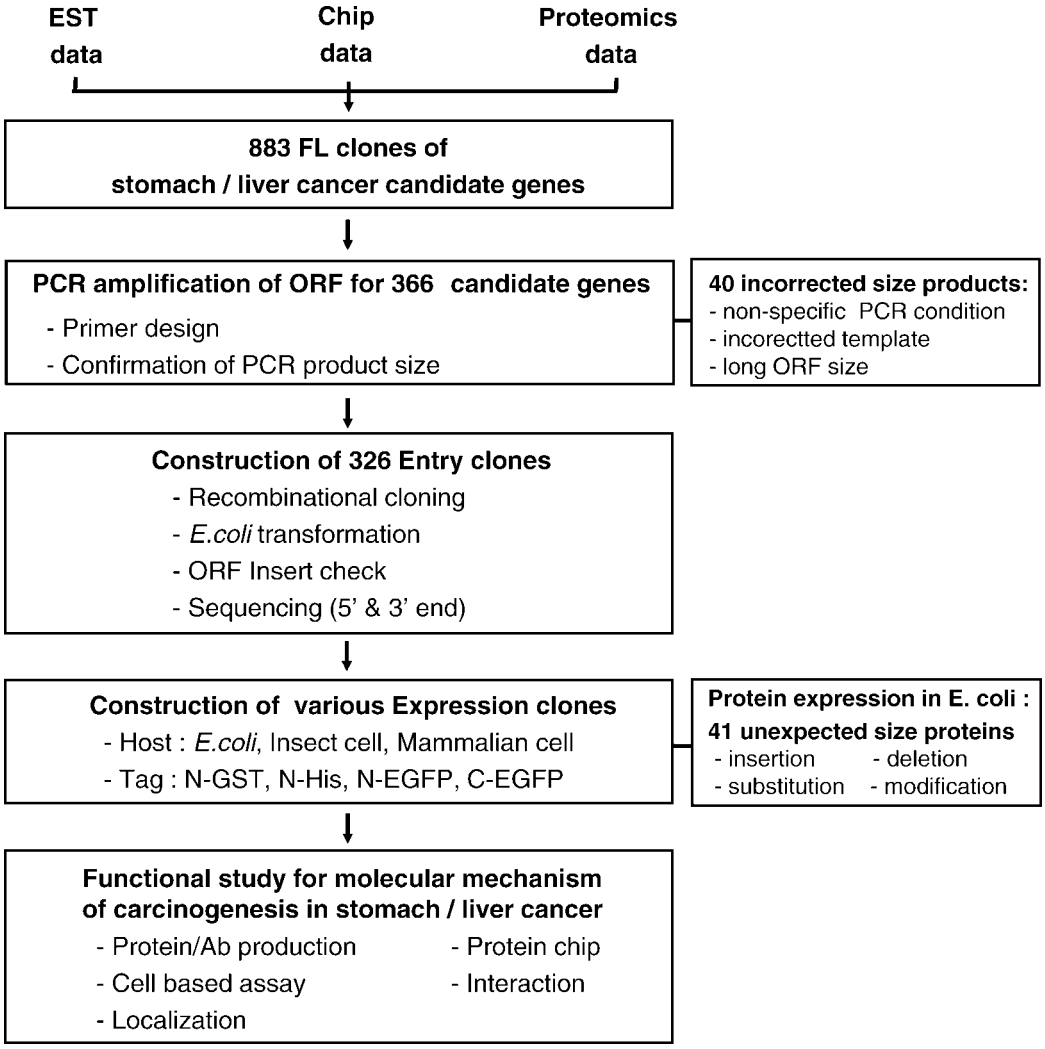

Figure 1. Overall scheme for the Gateway cloning of human candidate genes related to gastric and liver cancers.

buffer by general procedures. Total and soluble protein fractions were separated on a 12% SDS-polyacrylamide gel and the Western blotting was performed with anti-GST-HRP. Analysis of MW of the proteins was performed as described in *E. coli* expression.

Recombinant Protein Expression in Mammalian Cells and Image Analysis. The NIH3T3 and HEK293 cells transfected with EGFP mammalian expression clones were grown on glass cover slips for 48 h in DMEM, and then fixed in 4% paraformaldehyde. The fixed cells were imaged using a Nikon ECLIPSE (TS100) microscope and the localization of the protein was determined.

Bioinformatic Analysis. The Molecular weight of the protein encoded in the candidate genes was predicted from the RefSeq data in public database (as of July 12, 2004). The classification of the subcellular localization of the proteins was performed based on the cellular components category of GO database (http://www.geneontology.org, as of March 2005).

Results

Construction of Multi-Purpose Expression Clones for Candidate Genes. Figure 1 shows the outlines for constructing high-throughput-based expression clones for the candidate genes. We have identified 883 gastric or liver cancer-related genes using the transcriptome data, which includes EST frequency, cDNA microarray, and proteomic data all generated by our project team. Among these candidate genes, the full-length cDNAs of 366 genes were previously obtained from the 21C Frontier Human cDNA Bank (http://kugi.kribb.re.kr:8080/genbank/) and were used as templates for generating multi-purpose expression clones. The GO analysis of these 366 genes revealed that 288 showed catalytic activity, mainly hydrolase and oxidoreductase activity, 73 genes also showed transferase activity, and 48 genes showed signal transducer activity (data not shown). In addition, 59 genes were involved in nucleic acid binding, and 91 genes were involved in protein binding.

To construct the ORFs containing the Gateway

Downloaded from ebm.sagepub.com at SAGE Publications on November 3, 2014

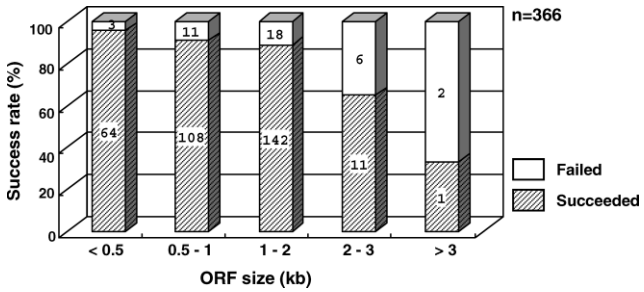

Figure 2. Success rate of PCR amplification for constructing entry clones using the Gateway system.

recombination site, the full-length cDNA for each of the 366 candidate genes was amplified with gene-specific primers by a second PCR in 96-well plates as described in the Materials and Methods section. When the amplified PCR products were subjected to electrophoresis and the length of the observed products were compared to their expected length, as calculated from RefSeq data in the public database, 326 PCR products (89%) were found to be of the correct size, while 40 PCR products had the incorrect size from no products, an unexpected size, multiple or smear banded products (data not shown). This failure may be mainly due to the non-specific PCR conditions or the use of incorrect genes as a template in the course of the colony re-array. Of the 366 genes, ORFs up to 2 kb were successfully amplified with a rate higher than 89%, whereas the successful amplification of ORFs larger than 2 kb was dramatically decreased (Fig. 2). These results are consistent with previous reports of amplification of PCR products dependent on ORF length.

To prepare entry clones for the candidate genes, the 326 PCR products with the correct size were cloned into a Gateway donor vector, pDONR201. The identities of the entry cloned ORFs were examined by PCR and DNA sequencing. The 326 PCR products were successfully constructed into entry clones and to have the correct frame sequence; however, a few clones were shown to have frameshift mutations in the ORF due to incorrect gene specific primers. To generate expression clones which can be expressed in various hosts with various protein tags, the 326 entry clones were recombined with four Gateway destination vectors: the pDEST-15, pDEST-17, BacHTS and pDEST-27, and two modified GFP expression vectors, pcDNA-DEST-53E and pcDNA-DEST-47E. PCR analysis of the constructed expression clones showed that almost all of the entry clones were successfully recloned into six types of Gateway destination vectors. In total, from 366 amplified PCR products, we successfully constructed 326 entry clones and 2,068 expression clones including the 244 GST fusion or 208 His_6 fusion expression clones for *E. coli* using the LR recombinational cloning.

Expression and Analysis of Recombinant Fusion Protein in *E. coli*. We transformed 244 GST and 208 His_6 fusion expression clones to competent BL21 (DE3) cells, and the expressed proteins were examined by SDS-PAGE analysis. A summary of protein expression is shown in Table 1. Of the 244 GST fusion clones, 175 (72%) were expressed successfully in *E. coli*. Among them, 27 clones (15%) were expressed as soluble proteins, while 148 clones (85%) were expressed as insoluble proteins. In addition, 134 clones (77%) expressed proteins with molecular weight predicted by RefSeq, of which 18 clones (10%) expressed soluble proteins. In contrast, His_6 fusion protein expression was reduced (61%) compared to GST fusion protein expression. Of the 126 expressed His_6 fusion proteins, 98 clones (78%) expressed proteins with the expected MW and one clone was expressed as a soluble protein of an unexpected size. When we performed a GO analysis for the 61 clones that expressed only the GST fusion protein and not the His fusion protein, genes showed catalytic activity containing hydrolase family proteins such as *PGC* (progastricin) and *PLG* (plasminogen); also, signal transducer activity such as *CD14* and *F3* (coagulation factor III) were highly involved. Figure 3-I-A shows the expression yield for the 244 GST or 208 His_6 fusion proteins according to their molecular weight. The overlapped proteins which are expressed both as a GST fusion and a His_6 fusion protein were calculated as one protein. GST fusion proteins up to 60 kDa were expressed with success rate of 70–80%, and for proteins larger than 60 kDa, the success rate was decreased to only 38%. This observation demonstrates that the expression of recombinant proteins can be affected by their size (34). In addition, the His_6 fusion proteins were expressed with low yield of about 60% compared to the GST fusion proteins, and proteins smaller than 10 kDa were not expressed at all. It is possible that, when a smaller protein is mixed with the CBB dye, it is not detected. These findings are consistent with a previous report that GST fusion has a higher success rate for expression than His_6 fusion (35). However, the expression of proteins larger than 50 kDa were higher for the His fusion proteins than for the GST fusion proteins. On the other hand, when 246 proteins were analyzed according to the subcellular localization based on GO, 92 proteins (37%)

Table 1. Expression of N-GST Fusion or N-His_6-Fusion Protein

	E. coli		Insect
	N-GST	N-His_6	N-GST
No. of used clones	244	208	105
No. of expressed clones	175 (72)	126 (61)	80 (76)
Soluble protein	27 (15)	1 (1)	74 (93)
Insoluble protein	148 (85)	125 (99)	6 (8)
Expected size protein*	134 (77)	98 (78)	72 (90)
Soluble/Expected size protein	18 (10)	0 (0)	66 (83)

*, M.W. was estimated by RefSeq data in public database
( ), indicates % of the success rate

Downloaded from ebm.sagepub.com at SAGE Publications on November 3, 2014

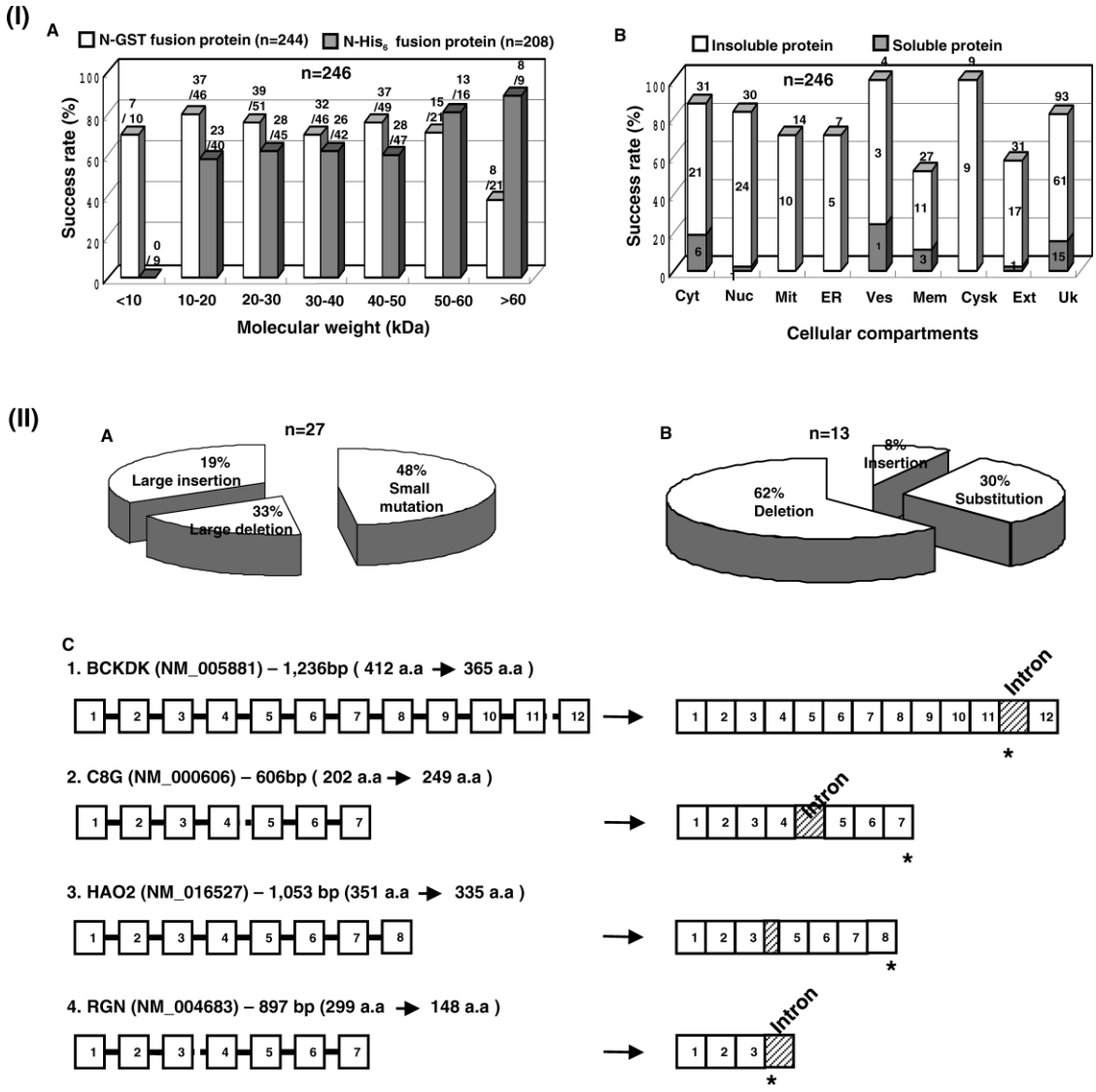

Figure 3. Analysis of recombinant proteins expressed in *E. coli.* (I) For 246 expression clones, 206 clones included two different tags, GST and His_6 while 38 clones were tagged only with GST and 2 clones only with His_6. The success rate was calculated by dividing the number of proteins in these categories by the number of expressed proteins. (A) Size distribution of expressed proteins. Proteins were divided into seven categories according to their predicted molecular weights. For the fraction on the top of bar, the lower number means the number of used total clones, the upper number for the number of the expressed clones. (B) Correlation of expression success with the predicted subcellular localization of the proteins. The localization of proteins was estimated based on GO term. Cytosol (Cyt), nucleus (Nuc), mitochondrion (Mit), endoplasmic reticulum (ER), vesicle (Ves), membrane (Mem), cytoskeleton (Cysk), extracellular (Ext), unknown (Uk). The overlapped proteins which are expressed both with GST fusion recombinant protein and with His_6 fusion recombinant protein were calculated as one protein. The number in the upper of the bar represents the number of used total clones. (II) Analysis of incorrect-sized proteins in *E. coli* expression. 41 clones encoding proteins with an unexpected molecular weight in *E. coli* were selected and the full-region of the cDNA sequenced. The obtained sequence data were compared to the RefSeq data in the public database. (A) Analysis of incorrect sized proteins. Small mutation indicates change by less than 10 bases, large mutation for change by more than 40 bases. (B) Detailed analysis of the small mutation. (C) Examples of the incorrect sized proteins. 1 and 2, insertion of intron; 3, deletion of exon; 4, insertion of intron and deletion of exons simultaneously; *, stop codon.

were expected to be localized to the cytosol, nucleus, and extracellular compartments, and 61 clones (25%) to the mitochondrion, ER (Endoplasmic Reticulum), vesicle, membrane and cytoskeleton. The subcellular localization of the remaining (38%) was not assigned. The 188 expressed proteins, which should be localized to the cytosol, nucleus, vesicle and cytoskeleton, were highly expressed with a success rate of more than 80%, although there were few proteins in the vesicle and cytoskeleton (Fig. 3-I-B). Most soluble proteins were localized to the cytosol, vesicle and membrane compartments.

To determine why proteins of unexpected size were produced, we randomly selected 41 clones which expressed GST or His fusion proteins of unexpected molecular weight and analyzed their ORF region by full-sequencing. Of the 41 genes, 27 genes (66%) were found to have various mutations. 14 genes (52%) contained an insertion or deletion mutation of more than 40 bases, while 13 genes

Downloaded from ebm.sagepub.com at SAGE Publications on November 3, 2014

(48%) had insertion, deletion, or substitution mutations of less than 10 bases (Fig. 3-II-A & B). These results indicate that the proteins with an unexpected MW were a result of frameshift mutations. The remaining clones had no mutation. From these clones, the production of proteins with an unexpected MW may be due to the modification or the degradation of the proteins. The insertions of >40 bases were shown to occur primarily in intron regions, whereas the deletions occurred primarily at exon regions. Examples are shown in Figure 3-II-C. For *BCKDK* (branched chain alpha-ketoacid dehydrogenase kinase), our transcript contains the insertion of intron 11 (92 bp) between exon 11 and exon 12, a mutation predicted to generate a 365 amino acid protein rather than the 412 amino acid protein predicted by RefSeq. In contrast, the *C8G* (complement component 8, gamma polypeptide) clone produced a 249 amino acid protein due to the insertion of intron 4 (239 bp) between exon 4 and exon 5 of the gene. Additionally, the *HAO2* (hydroxyacid oxidase 2) clone, produced a smaller transcript resulting from the partial deletion of exon 4, and the *RGN* (regucalcin) clone produced a new transcript containing an insertion of intron 3 and a deletion from exons 4, 5, 6 and 7.

To determine whether such mutations were present in the original tumor, the mutations were examined by RT-PCR using the original total RNA. The results revealed that large mutations such as *BCKDK, C8G* and *GK* (glycerol kinase) were present in the original tumors, while *HAO2* and *RGN* occurred in normal tissue as an alternative splicing form (data not shown). However, small mutations such as *UBE2C* (ubiquitin-conjugating enzyme E2C), *LENG5* (tRNA splicing endonuclease 34 homolog) and *SGK* (serum/glucocorticoid regulated kinase) were not detected in the original DNA. This indicates that single mutations occurred during the creation of the clone. This observation can be compared to a previous report concerning the mutation rate for the creation of clones by PCR amplification (29). These genes which have different alternative splicing forms or large mutations could be provide valuable clues toward understanding the molecular mechanism associated with the tumorigenesis of gastric or liver carcinogenesis.

Expression of Recombinant Fusion Protein in Insect Cells. To produce antibodies specific for the proteins encoded by the candidate genes, we identified 105 clones which were poorly expressed or not expressed in *E. coli*. The insect expression clones corresponding to these genes were transfected to insect Sf21 cells in 96-well plates. The expression of the GST fusion recombinant protein was analyzed by Western blot analysis using a GST antibody. Most of the proteins were expressed well in insect cells, although a large number of clones also expressed 27 kDa GST proteins along with their target proteins. In some clones, degraded forms of the target proteins were also produced (data not shown). The expression of the proteins is summarized in Table 1. Of the 105 GST fusion clones, 80 (76%) were successfully expressed. Among them, 72 clones (90%) expressed proteins with the expected MW and 66 (92%) were determined to express soluble proteins. Among the 74 soluble proteins, 53 (72%) were successfully purified on a small scale. As expected, these results indicate that the insect cell system significantly increased the expression and solubility of proteins in comparison with the *E. coli* system.

We also examined the relationship between the success rate of expression and the molecular size and subcellular localization of the proteins in insect cells. As shown in Figure 4A, unlike the *E. coli* system, the expression of the GST fusion proteins was remarkably unaffected by size. They were expressed with a high success rate of 60–87%, regardless of protein size. The analysis of the subcellular localization of 80 expressed proteins revealed that most of the proteins, except for those that should be localized either extracellular or to a mitochondrion, were highly expressed with a success rate of more than 70% (Fig. 4B). Interestingly, fusion proteins which would be predicted to be localized to membranes were expressed at a significant rate, compared to the *E. coli* system. Most of the proteins were expressed as soluble proteins.

Expression and Subcellular Localization of EGFP Recombinant Fusion Proteins in Mammalian Cells. To examine the expression and localization of proteins in mammalian cells, we used two types of new EGFP fusion expression systems instead of the GFP signal which is difficult to detect. Using the genes expressed with the expected MW in the *E. coli* system, the subcellular localization of their proteins was analyzed in mammalian cells based on the GO classification. The 30 proteins were classified into nine categories (extracellular, membrane, ER, mitochondrion, vesicle, nucleus, cytosol, cytoskeleton and unknown locations), and 60 expression clones of N-EGFP fusion or C-EGFP fusion of these proteins were then transfected into NIH3T3 and HEK293 cells. The localization of the EGFP-fusion protein was analyzed by fluorescent microscopy. The localization of the proteins was similar for both the NIH3T3 and HEK293 cell lines (data not shown) as well as between N-terminal and C-terminal EGFP-fusion proteins (Fig. 5). In addition, most of the proteins were localized to cellular components predicted by GO, although the localization of EGFP-fusion protein to membrane, ER and vesicle was not easy to distinguish from other localizations (Fig. 5A). Some proteins, for example *FABP1* (fatty acid binding protein 1), were localized in both the cytosol and nucleus but one of them corresponded to the GO database. *TCEA1* (transcription elongation factor A, 1) was localized to the nucleus, regardless of N-terminal and C-terminal EGFP-fusion. *HSPCB* (heat shock 90 kDa protein 1, beta) was detected as a diffuse cytoplasmic fluorescence, encompassing the entire cell. Figure 5B depicts proteins whose localization could not be determined. *GKN1* (gastrokine 1) and *LGALS1* (lectin, galactoside-binding, soluble, 1) were found to be localized to both the nucleus and cytosol. Like the *TCEA1*, *E2-EPF* (ubiquitin carrier protein) was localized to the nucleus, regardless of

Downloaded from ebm.sagepub.com at SAGE Publications on November 3, 2014

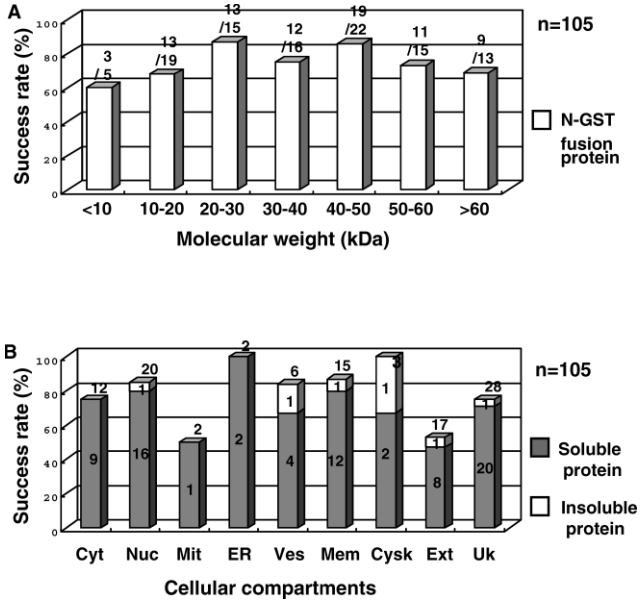

Figure 4. Analysis of recombinant proteins expressed in insect cells. (A) Size distribution of the expressed proteins. (B) Correlation of expression success to the subcellular localization of the proteins. The calculation of the success rate and details were the same as described in Figure 2.

the position of the EGFP tagged to target protein. These results indicate that the proteins expressed in mammalian cells had the correct folding conformation. These data for the protein localization can serve as clues for the function of unknown candidate genes.

Discussion

High-throughput expression systems are essential tools for large scale functional genomics. By constructing a multi-purpose expression system, we were able to develop a method for expressing large numbers of proteins for large-scale structural and functional studies. All of the procedures used in this study, including DNA cloning, plasmid preparation, protein induction and cell lysis, were preformed using a standard 96-well format in an efficient and reproducible manner, thus making these procedures suitable for automation. In the present study, entry clones for 326 candidate genes were successfully constructed, and 2,068 expression clones were then produced by cloning the ORFs of the entry clones into six different expression vectors. These expression clones of GST-, His_6- and EGFP-fusion proteins will provide valuable resources for antibody production and genome-wide protein analysis as well as high-throughput protein screening.

To ensure accurate nucleotides incorporation during the

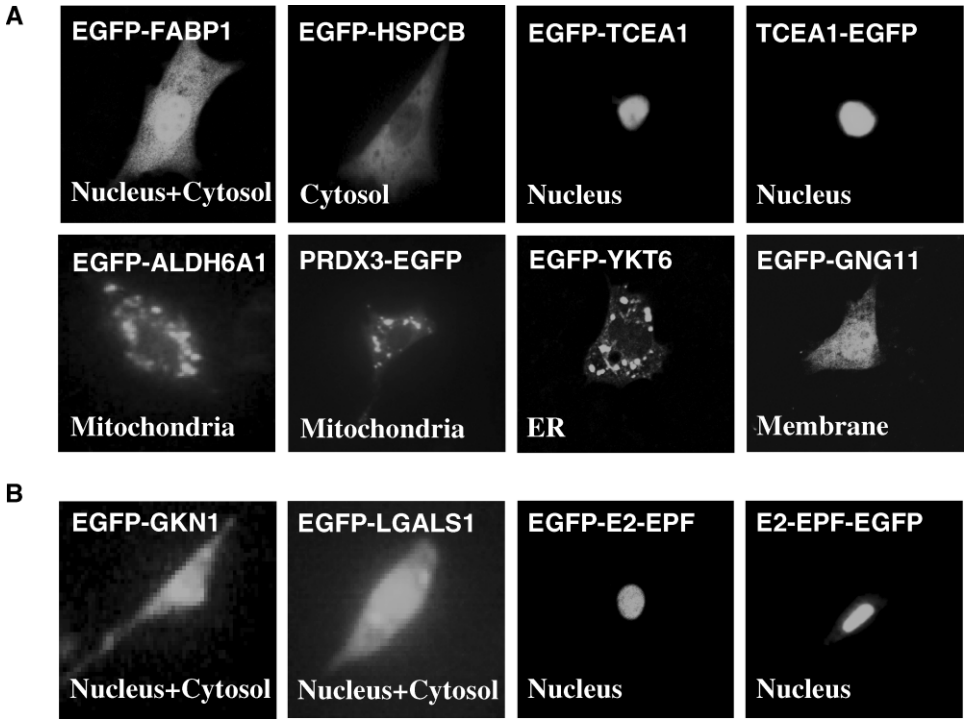

Figure 5. Localization of the proteins encoded by the candidate genes in NIH3T3 cells. Localization of the target protein was estimated by GO classification. The position of the EGFP represents the position of EGFP tagged to target protein. For example, EGFP-FABP1 means that EGFP is tagged at the N-terminus of FABP1. Twelve examples were found to be localized in the respective cellular components in the cells. The subcellular compartments in the bottom indicate the localization for protein which is observed under a fluorescent microscope. (A) Examples of the localization for known proteins which are estimated by GO classification. (B) Examples of the localization for unknown proteins for which could not be estimated their localization by GO. A color version of this figure is available in the online journal.

Downloaded from ebm.sagepub.com at SAGE Publications on November 3, 2014

construction of the entry clones, PCR was performed using high-fidelity *Pfu* DNA polymerase and less than 10 cycles of amplification were used. Under these conditions, a similar PCR success rate (89%) was observed in the large-scale cloning of the ORFs from *P. falciparum* (30) and from *C. elegans* (29). A quality assessment of the PCR products was performed by i) checking the length of the PCR product by electrophoresis, ii) sequencing the 5′ and 3′ ends of the PCR products, and iii) calculating the predicted molecular weight of the encoded protein. Using these criteria, 326 ORF-PCR products with corrected product size and sequence at the 5′ and 3′ ends were obtained from the 366 candidate genes. Among the proteins encoded by these ORF, 77% of the *E. coli*-expressed proteins and 90% of the insect cell-expressed proteins had the expected molecular weight. This observation is similar to a previous report showing that among the 68 clones of *C. elegans*, 61% clones expressed N-GST fusion proteins with an appropriate size in *E. coli* (29). In addition, it has been reported that 80% of the N-GST fusion proteins were of the correct size for 192 human proteins expressed in *E. coli* (36).

We used an N-terminus tagging system for the construction of expression clones in the *E. coli*. Because the amplified ORF was not confirmed by full-sequencing, a C-terminus tagged protein may not have been expressed properly. Interestingly, our His_6 fusion protein expression level in the *E. coli* expression system was higher than the 55% expression rate observed for human proteins reported by Rual *et al*. (33). In addition, our data revealed that the His_6 fusion proteins larger than 50 kDa showed an increased expression yield of 89% compared to the expression of the GST fusion proteins.

Most proteins expressed in *E. coli* were insoluble. Several different fusion partners, such as the maltose-binding protein (MBP), NusA, ZZ, Gb1 and thioredoxin, have been reported to increase the production of soluble proteins (37). Among these, MBP-tagged proteins are reported to have the greatest solubility (38). Temperature has also been reported to have a minor effect on solubility. To increase the solubility of proteins in *E. coli*, the protein expression was performed at 25°C. This change resulted in some insoluble proteins becoming soluble proteins (data not shown). These data indicate that a different tag, such as MBP, or a low temperature will increase the solubility of the proteins. In our high-throughput expression in the *E. coli* expression system, the success rate for expression was similar to that of previous reports but protein solubility was relatively low (35, 39). This may be due to the expression in 96-well plate formats under limited conditions such as one expression vector and one *E. coli* strain, since protein solubility is affected by the conditions used for protein expression, as well as the biological properties of each protein (39).

Mutations in oncogenes or suppressor genes are known to be associated with carcinogenesis (40, 41). Alternatively spliced transcripts are also observed more frequently in tumor tissues compared to normal tissues (42). Our expression data showed that 10–20% of the expressed proteins had an unexpected molecular weight. A detailed sequence analysis of these clones revealed frameshift mutations resulting from the insertion, deletion or substitution of nucleotides. Large mutations (>40 bp) were found at the same frequency as small mutations (<10 bp). Deletion mutations, regardless of the size, were more frequently observed than insertion mutations. This indicates that candidate genes related to gastric or liver cancers have a high frequency of mutations, and that these mutations may have an effect on the occurrence of gastric or liver cancers. This prediction is supported by reports that mutations which delete exons within the E-cadherin gene, a tumor suppressor gene, were found in diffuse type-gastric carcinomas (43). In addition, the relation between disease and mutation is supported by the data depicted in Figure 3-II-C. *C8G* is known to participate in the formation of the Membrane Attack Complex (MAC). Patients with a deficiency in *C8G* have been reported to be vulnerable to certain bacteria infections (44). Our variant, a longer transcript which was generated by the insertion of an intron, was found in a gastric cancer. To date, our new variant has not appeared in the UCSC database, although another 6 variants for *C8G* have been reported. It is proposed that *C8G* deficiency may be associated with susceptibility of *H. pylori* infection. *RGN* encodes a highly conserved, calcium-binding protein that may play an important role in calcium homeostasis (45). A decrease in the expression of this protein has been reported to lead to the dysregulation of calcium signaling in the aged liver (46). Seven forms of *RGN* transcript variants are reported in the UCSC database. Among them is our variant, which encodes a 148 amino acid rather than the 299 amino acid normal protein. This truncation is a result of an insertion of an intron and the deletion of exons. Our variant form was found only in purified pancreatic islets of the database. These findings suggest that a decrease or mutation in *RGN* may be associated with liver cancer. A further analysis of these mutations using DNAs from the original source showed that mutations occurred frequently in cancer cells compared to normal tissues as an alternative splicing form and would be related to carcinogenesis.

The insect expression system is a well known expression system for human proteins. Unlike the *E. coli* system, most of the expressed proteins were soluble and the fusion proteins which should be localized to the mitochondrion, ER and cytoskeleton, were almost all expressed as soluble proteins. The expression efficiency was not dramatically affected by molecular weight, unlike *E. coli*. In addition, Western blot analysis revealed that a large number of clones expressed 26 kDa GST proteins along with the target proteins. As described in a previous report, GST fusion proteins are prone to proteolysis in vivo (36). Our findings indicate that the insect cell system is useful for improving the expression and solubility of proteins in comparison with the *E. coli* system.

Downloaded from ebm.sagepub.com at SAGE Publications on November 3, 2014

Subcellular localization is an important parameter in examining protein functions. Waldo and co-workers reported that GFP is a useful marker for assessing the subcellular localization of proteins and correct protein folding (47, 48). Localization data for 30 EGFP fusion proteins demonstrated that most were localized to the cellular components predicted by GO, and both N-terminal and C-terminal EGFP-fusion proteins were localized in same subcellular compartments in NIH3T3 and HEK293 cell lines (data not shown). These results are supported by our preliminary data showing that >90% of the 110 EGFP fusion proteins were localized in cellular components predicted by GO, and in >80%, no significant difference was found between N-terminal and C-terminal EGFP-fusion as well as in various cell lines such as NIH3T3, HEK293 cell lines and a gastric cell line, SNU638 (data not shown). However, it has been reported that proteins localized in mitochondria or secretary proteins are localized in different subcellular compartments according to N-terminal and C-terminal GFP-fusion (49). This may be due to the N-terminal fusion blocking signal sequences associated with import into the mitochondria. Our preliminary data also showed similar results for these proteins, but these findings were frequently observed in ER proteins rather than mitochondrial proteins. On the other hand, proteins localized to membranes, the ER and vesicles were detected only with considerable difficulty. To clearly observe these proteins, organelle markers and confocal microscopy should be used. The constitution of the localization data set using EGFP clones should provide insights into the function of numerous uncharacterized genes.

We plan to construct multi-purpose expression clones of 1,000 candidate genes associated with gastric and liver cancers using high-throughput-based Gateway cloning. With these clones, the generation of antibodies specific for the candidate genes will be performed using soluble protein purified from *E. coli*. In the case of an insoluble protein, the material will be used to screen for an antibody prepared from a DNA vaccine. The proteins that could not be expressed in *E. coli* were applied to the insect cell system to express for monoclonal antibody production. Finally, purified proteins that were over 90% pure could be used as antigens for antibody production in mice. In addition, by using mammalian expression clones, an analysis of the localization of the candidate genes and the screening of genes involved in carcinogenesis pathways such as *E2F*, *p53* and *Myc* by means of a cell-based assay system are now underway. The purified proteins could also be used for the identification of interacting proteins with them using a protein chip as well as the structural elucidation of target proteins. Although cancer tissues include not only cancer cells but also other cellular components, these studies using multi-purpose expression clones promise to provide a valuable resource for understanding the molecular mechanism of carcinogenesis in gastric and liver cancers and would be very helpful in diagnosis and therapeutic prediction.

We thank Ms. Hee-Young Ahn and Ms. Sang-Soon Byun, for technical support during the construction of the expression clones and their expression in *E. coli*.

1. Parkin DM, Bray F, Ferlay J, Pisani P. Estimating the world cancer burden: Globocan 2000. Int J Cancer 94:153–156, 2001.
2. Neugut AI, Hayek M, Howe G. Epidemiology of gastric cancer. Semin Oncol 23:281–291, 1996.
3. Roukos DH. Current status and future perspectives in gastric cancer management. Cancer Treat Rev 26:243–255, 2000.
4. Park WS, Oh RR, Park JY, Lee JH, Shin MS, Kim HS, Lee HK, Kim YS, Kim SY, Lee SH, Yoo NJ, Lee JY. Somatic mutations of the trefoil factor family 1 gene in gastric cancer. Gastroenterology 119:691–698, 2000.
5. Yokozaki H, Yasui W, Tahara E. Genetic and epigenetic changes in stomach cancer. Int Rev Cytol 204:49–95, 2001.
6. Werner M, Becker KF, Keller G, Hofler H. Gastric adenocarcinoma: pathomorphology and molecular pathology. J Cancer Res Clin Oncol 127:207–216, 2001.
7. Li QL, Ito K, Sakakura C, Fukamachi H, Inoue K, Chi XZ, Lee KY, Nomura S, Lee CW, Han SB, Kim HM, Kim WJ, Yamamoto H, Yamashita N, Yano T, Ikeda T, Itohara S, Inazawa J, Abe T, Hagiwara A, Yamagishi H, Ooe A, Kaneda A, Sugimura T, Ushijima T, Bae SC, Ito Y. Causal relationship between the loss of RUNX3 expression and gastric cancer. Cell 109:113–124, 2002.
8. Murakami Y, Hayashi K, Hirohashi S, Sekiya T. Aberrations of the tumor suppressor p53 and retinoblastoma genes in human hepatocellular carcinomas. Cancer Res 51:5520–5525, 1991.
9. Nhieu JT, Renard CA, Wei Y, Cherqui D, Zafrani ES, Buendia MA. Nuclear accumulation of mutated beta-catenin in hepatocellular carcinoma is associated with increased cell proliferation. Am J Pathol 155:703–710, 1999.
10. Peng SY, Lai PL, Hsu HC. Amplification of the c-myc gene in human hepatocellular carcinoma: biologic significance. J Formos Med Assoc 92:866–870, 1993.
11. Nishida N, Fukuda Y, Komeda T, Kita R, Sando T, Furukawa M, Amenomori M, Shibagaki I, Nakao K, Ikenaga M, Ishizaki K. Amplification and overexpression of the cyclin D1 gene in aggressive human hepatocellular carcinoma. Cancer Res 54:3107–3110, 1994.
12. Ito N, Kawata S, Tamura S, Takaishi K, Shirai Y, Kiso S, Yabuuchi I, Matsuda Y, Nishioka M, Tarui S. Elevated levels of transforming growth factor beta messenger RNA and its polypeptide in human hepatocellular carcinoma. Cancer Res 51:4080–4083, 1991.
13. Cavin LG, Romieu-Mourez R, Panta GR, Sun J, Factor VM, Thorgeirsson SS, Sonenshein GE, Arsura M. Inhibition of CK2 activity by TGF-beta1 promotes IkappaB-alpha protein stabilization and apoptosis of immortalized hepatocytes. Hepatology 38:1540–1551, 2003.
14. Shirota Y, Kaneko S, Honda M, Kawai HF, Kobayashi K. Identification of differentially expressed genes in hepatocellular carcinoma with cDNA microarrays. Hepatology 33:832–840, 2001.
15. Xu XR, Huang J, Xu ZG, Qian BZ, Zhu ZD, Yan Q, Cai T, Zhang X, Xiao HS, Qu J, Liu F, Huang QH, Cheng ZH, Li NG, Du JJ, Hu W, Shen KT, Lu G, Fu G, Zhong M, Xu SH, Gu WY, Huang W, Zhao XT, Hu GX, Gu JR, Chen Z, Han ZG. Insight into hepatocellular carcinogenesis at transcriptome level by comparing gene expression profiles of hepatocellular carcinoma with those of corresponding noncancerous liver. Proc Natl Acad Sci U S A 98:15089–15094, 2001.
16. Suriawinata A, Xu R. An update on the molecular genetics of hepatocellular carcinoma. Semin Liver Dis 24:77–88, 2004.

Downloaded from ebm.sagepub.com at SAGE Publications on November 3, 2014

17. Okabe H, Satoh S, Kato T, Kitahara O, Yanagawa R, Yamaoka Y, Tsunoda T, Furukawa Y, Nakamura Y. Genome-wide analysis of gene expression in human hepatocellular carcinomas using cDNA microarray: identification of genes involved in viral carcinogenesis and tumor progression. Cancer Res 61:2129–2137, 2001.
18. Kim NS, Hahn Y, Oh JH, Lee JY, Oh KJ, Kim JM, Park HS, Kim S, Song KS, Rho SM, Yoo HS, Kim YS. Gene cataloging and expression profiling in human gastric cancer cells by expressed sequence tags. Genomics 83:1024–1045, 2004.
19. Yoon SY, Kim JM, Oh JH, Jeon YJ, Lee DS, Kim JH, Choi JY, Ahn BM, Kim S, Yoo HS, Kim YS, Kim NS. Gene expression profiling of human HBV- and/or HCV-associated hepatocellular carcinoma cells using expressed sequence tags. Int J Oncol 29:315–327, 2006.
20. Wang AG, Yoon SY, Oh JH, Jeon YJ, Kim M, Kim JM, Byun SS, Yang JO, Kim JH, Kim DG, Yeom YI, Yoo HS, Kim YS, Kim NS. Identification of intrahepatic cholangiocarcinoma related genes by comparison with normal liver tissues using expressed sequence tags. Biochem Biophys Res Commun 345:1022–1032, 2006.
21. Kim B, Bang S, Lee S, Kim S, Jung Y, Lee C, Choi K, Lee SG, Lee K, Lee Y, Kim SS, Yeom YI, Kim YS, Yoo HS, Song K, Lee I. Expression profiling and subtype-specific expression of stomach cancer. Cancer Res 63:8248–8255, 2003.
22. Choi JK, Choi JY, Kim DG, Choi DW, Kim BY, Lee KH, Yeom YI, Yoo HS, Yoo OJ, Kim S. Integrative analysis of multiple gene expression profiles applied to liver cancer study. FEBS Lett 565:93–100, 2004.
23. Ha GH, Lee SU, Kang DG, Ha NY, Kim SH, Kim J, Bae JM, Kim JW, Lee CW. Proteome analysis of human stomach tissue: separation of soluble proteins by two-dimensional polyacrylamide gel electrophoresis and identification by mass spectrometry. Electrophoresis 23:2513–2524, 2002.
24. Kim J, Kim SH, Lee SU, Ha GH, Kang DG, Ha NY, Ahn JS, Cho HY, Kang SJ, Lee YJ, Hong SC, Ha WS, Bae JM, Lee CW, Kim JW. Proteome analysis of human liver tumor tissue by two-dimensional gel electrophoresis and matrix assisted laser desorption/ionization-mass spectrometry for identification of disease-related proteins. Electrophoresis 23:4142–4156, 2002.
25. Hartley JL, Temple GF, Brasch MA. DNA cloning using in vitro site-specific recombination. Genome Res 10:1788–1795, 2000.
26. Van Mullem V, Wery M, De Bolle X, Vandenhaute J. Construction of a set of Saccharomyces cerevisiae vectors designed for recombinational cloning. Yeast 20:739–746, 2003.
27. Cheo DL, Titus SA, Byrd DR, Hartley JL, Temple GF, Brasch MA. Concerted assembly and cloning of multiple DNA segments using in vitro site-specific recombination: functional analysis of multi-segment expression clones. Genome Res 14:2111–2120, 2004.
28. Sasaki Y, Sone T, Yoshida S, Yahata K, Hotta J, Chesnut JD, Honda T, Imamoto F. Evidence for high specificity and efficiency of multiple recombination signals in mixed DNA cloning by the Multisite Gateway system. J Biotechnol 107:233–243, 2004.
29. Reboul J, Vaglio P, Rual JF, Lamesch P, Martinez M, Armstrong CM, Li S, Jacotot L, Bertin N, Janky R, Moore T, Hudson JR Jr, Hartley JL, Brasch MA, Vandenhaute J, Boulton S, Endress GA, Jenna S, Chevet E, Papasotiropoulos V, Tolias PP, Ptacek J, Snyder M, Huang R, Chance MR, Lee H, Doucette-Stamm L, Hill DE, Vidal M. C. elegans ORFeome version 1.1: experimental verification of the genome annotation and resource for proteome-scale protein expression. Nat Genet 34:35–41, 2003.
30. Aguiar JC, LaBaer J, Blair PL, Shamailova VY, Koundinya M, Russell JA, Huang F, Mar W, Anthony RM, Witney A, Caruana SR, Brizuela L, Sacci JB Jr, Hoffman SL, Carucci DJ. High-throughput generation of P. falciparum functional molecules by recombinational cloning. Genome Res 14:2076–2082, 2004.
31. Wiemann S, Arlt D, Huber W, Wellenreuther R, Schleeger S, Mehrle A, Bechtel S, Sauermann M, Korf U, Pepperkok R, Sultmann H, Poustka A. From ORFeome to biology: a functional genomics pipeline. Genome Res 14:2136–2144, 2004.
32. Dupuy D, Li QR, Deplancke B, Boxem M, Hao T, Lamesch P, Sequerra R, Bosak S, Doucette-Stamm L, Hope IA, Hill DE, Walhout AJ, Vidal M. A first version of the Caenorhabditis elegans Promoterome. Genome Res 14:2169–2175, 2004.
33. Rual JF, Hirozane-Kishikawa T, Hao T, Bertin N, Li S, Dricot A, Li N, Rosenberg J, Lamesch P, Vidalain PO, Clingingsmith TR, Hartley JL, Esposito D, Cheo D, Moore T, Simmons B, Sequerra R, Bosak S, Doucette-Stamm L, Le Peuch C, Vandenhaute J, Cusick ME, Albala JS, Hill DE, Vidal M. Human ORFeome version 1.1: a platform for reverse proteomics. Genome Res 14:2128–2135, 2004.
34. Smith DB. Generating fusions to glutathione S-transferase for protein studies. Methods Enzymol 326:254–270, 2000.
35. Shih YP, Kung WM, Chen JC, Yeh CH, Wang AH, Wang TF. High-throughput screening of soluble recombinant proteins. Protein Sci 11:1714–1719, 2002.
36. Braun P, Hu Y, Shen B, Halleck A, Koundinya M, Harlow E, LaBaer J. Proteome-scale purification of human proteins from bacteria. Proc Natl Acad Sci U S A 99:2654–2659, 2002.
37. Hammarstrom M, Hellgren N, van Den Berg S, Berglund H, Hard T. Rapid screening for improved solubility of small human proteins produced as fusion proteins in Escherichia coli. Protein Sci 11:313–321, 2002.
38. Kapust RB, Waugh DS. Escherichia coli maltose-binding protein is uncommonly effective at promoting the solubility of polypeptides to which it is fused. Protein Sci 8:1668–1674, 1999.
39. Luan CH, Qiu S, Finley JB, Carson M, Gray RJ, Huang W, Johnson D, Tsao J, Reboul J, Vaglio P, Hill DE, Vidal M, Delucas LJ, Luo M. High-throughput expression of C. elegans proteins. Genome Res 14:2102–2110, 2004.
40. Benvenuti S, Arena S, Bardelli A. Identification of cancer genes by mutational profiling of tumor genomes. FEBS Lett 579:1884–1890, 2005.
41. Bardelli A, Velculescu VE. Mutational analysis of gene families in human cancer. Curr Opin Genet Dev 15:5–12, 2005.
42. Wang Z, Lo HS, Yang H, Gere S, Hu Y, Buetow KH, Lee MP. Computational analysis and experimental validation of tumor-associated alternative RNA splicing in human cancer. Cancer Res 63:655–657, 2003.
43. Handschuh G, Candidus S, Luber B, Reich U, Schott C, Oswald S, Becke H, Hutzler P, Birchmeier W, Hofler H, Becker KF. Tumour-associated E-cadherin mutations alter cellular morphology, decrease cellular adhesion and increase cellular motility. Oncogene 18:4301–4312, 1999.
44. Del Rio C, Stephens DS, Knapp JS, Rice RJ, Schalla WO. Comparison of isolates of Neisseria gonorrhoeae causing meningitis and report of gonococcal meningitis in a patient with C8 deficiency. J Clin Microbiol 27:1045–1049, 1989.
45. Yamaguchi M. Role of regucalcin in maintaining cell homeostasis and function (review). Int J Mol Med 15:371–389, 2005.
46. Yamaguchi M. Role of regucalcin in calcium signaling. Life Sci 66:1769–1780, 2000.
47. Waldo GS, Standish BM, Berendzen J, Terwilliger TC. Rapid protein-folding assay using green fluorescent protein. Nat Biotechnol 17:691–695, 1999.
48. Pedelacq JD, Piltch E, Liong EC, Berendzen J, Kim CY, Rho BS, Park MS, Terwilliger TC, Waldo GS. Engineering soluble proteins for structural genomics. Nat Biotechnol 20:927–932, 2002.
49. Simpson JC, Wellenreuther R, Poustka A, Pepperkok R, Wiemann S. Systematic subcellular localization of novel proteins identified by large-scale cDNA sequencing. EMBO Rep 1:287–292, 2000.

Downloaded from ebm.sagepub.com at SAGE Publications on November 3, 2014