# Discovery of the breast cancer gene *BASE* using a molecular approach to enrich for genes encoding membrane and secreted proteins

**Kristi A. Egland*, James J. Vincent*, Robert Strausberg†, Byungkook Lee*, and Ira Pastan*‡**

*Laboratory of Molecular Biology and †Cancer Genomics Office, National Cancer Institute, National Institutes of Health, Bethesda, MD 20892

Contributed by Ira Pastan, December 5, 2002

**To identify unknown membrane proteins that could be used as targets for breast and prostate cancer immunotherapies and secreted proteins to be used as diagnostic markers, a cDNA library was generated from membrane-associated polyribosomal RNA derived from four breast cancer cell lines, one normal breast cell line, and a prostate cancer cell line. The membrane-associated polyribosomal cDNA library was subtracted with RNA from normal brain, liver, lung, kidney, and muscle. Of the 15,581 clones sequenced from the subtracted cDNA library, sequences from 10,506 clones map to known genes, but 5,075 sequences, representing 3,181 unique transcripts, are not associated with known genes. As one example, we experimentally investigated expression of a previously uncharacterized breast cancer gene that encodes a secreted protein designated *BASE* (*b*reast cancer *a*nd *s*alivary gland *e*xpression). *BASE* is expressed in many breast cancers but not in essential normal tissues including the five organs used for subtraction. Further analysis of this library should yield additional gene products of use in the diagnosis or treatment of breast or prostate cancer.**

In the United States, one in eight women will develop breast cancer during her lifetime (http://cis.nci.nih.gov/fact/5_6.htm). Recently, advances in the use of tumor-specific immunotherapies, such as the anti-ErbB2 monoclonal antibody, Herceptin (or TrastuzuMab), have shown clinical efficacy for the treatment of metastatic breast cancer with ErbB2 overexpression (1, 2). Because only 25–30% of human breast cancers overexpress ErbB2 (3, 4), there is a great need for the identification of more breast tumor-specific immunotherapy targets. One limitation is the availability of unique protein targets that are present on cancer cells but are not expressed in normal essential tissues such as brain, liver, or kidney.

To discover new antigens as targets for immunotherapies of breast cancers and secreted proteins for use as diagnostic markers, we have taken a molecular approach to identify membrane and secreted proteins that are present in breast cancers but are not expressed in normal essential tissues. Secretory and integral membrane proteins are translated from mRNA on membrane-bound ribosomes associated with the endoplasmic reticulum. Isolation of the membrane-associated polyribosomal RNA produces an enriched population of transcripts that encode membrane and secretory proteins. We generated a high-quality cDNA library that is enriched with genes that encode membrane and secreted proteins using membrane-associated RNA from six cell lines: four different breast cancer cell lines, one normal breast cell line, and a prostate cancer cell line. We subtracted this cDNA library with RNA from a pool of five libraries derived from liver, kidney, brain, lung, and muscle to enrich for differentially expressed genes in breast and prostate cancer while removing or reducing ubiquitously expressed genes. Subtraction of the membrane-associated polyribosomal library with libraries from normal tissues has a twofold consequence. First, it identifies known genes that are either activated or up-regulated in breast or prostate cancer cells and could facilitate identifying genes that play a role in carcinogenesis. Second, it identifies membrane proteins that are enriched in or specific for breast and prostate, which potentially could be used for targets of immunotherapies. In addition, secreted proteins could be used in breast cancer diagnostic tests.

We sequenced 943 random clones from the unsubtracted membrane-associated polyribosomal cDNA library (MAPcL) and 15,581 clones from the subtracted MAPcL. The subtracted MAPcL was enriched for genes that encode membrane and secreted proteins and numerous genes that are associated with or overexpressed in breast and prostate cancers. Of 15,581 clones sequenced from the subtracted MAPcL, 10,506 clones mapped to known genes, 4,074 mapped to UniGene clusters that are not associated with known genes, and 1,001 are comprised of unknown sequences. Here we describe a previously uncharacterized breast cancer gene from the subtracted MAPcL designated *BASE* (*b*reast cancer *a*nd *s*alivary gland *e*xpression). *BASE* encodes a secreted protein that is expressed in breast cancer but not in the tissues used for subtraction, verifying the effectiveness of our method.

## Materials and Methods

**Primers.** The primers used were KAE08h07-For (5′-CAAGCCCTTAATGATTTGACTC-3′), KAE08h07-Rev (5′-AGGTTTCTCTCTATGTTTGCCAC-3′), transferrin-For (5′-CATTCTCTAACTTGTTTGGTGG-3′), and transferrin-Rev (5′-CCAGGTAAACAAGTCTACCG-3′). The primers were synthesized by Lofstrand Laboratories (Gaithersburg, MD). The primers Actin-For (5′-GCATGGGTCAGAAGGAT-3′) and Actin-Rev (5′-CCAATGGTGATGACCTG-3′) were purchased from OriGene Technologies (Rockville, MD).

**Cell Culture.** MCF7, SK-BR-3, ZR-75-1, MDA-MB-231, and LNCaP cell lines were maintained as recommended by the American Type Culture Collection. The hTERT-HME1 cell line (CLONTECH) was maintained according to the manufacturer's instructions.

**Isolation of Membrane-Associated Polyribosomal RNA.** MCF7, SK-BR-3, ZR-75-1, MDA-MB-231, hTERT-HME1, and LNCaP cells ($\approx 1 \times 10^8$ cells per prep) were individually treated with 50 $\mu$M cycloheximide (Sigma) for 10 min at 37°C. The cells were washed twice with ice-cold PBS solution and scraped from the dish into a 50-ml conical tube. The cells were centrifuged and resuspended to $1.25 \times 10^8$ cells per ml in hypotonic buffer [10 mM KCl/1.5 mM $MgCl_2$/10 mM Tris·HCl, pH 7.4/200 units/ml RNase inhibitor (Roche, Indianapolis)]. The cells were placed

GENETICS

Abbreviations: MAPcL, membrane-associated polyribosomal cDNA library; *BASE*, breast cancer and salivary gland expression gene; dbEST, EST database; GO, Gene Ontology Consortium; ER, estrogen receptor; EEF1A1, eukaryotic elongation factor 1$\alpha$1.

Data deposition: The sequence reported in this paper has been deposited in the GenBank database (accession no. AY180924).

‡To whom correspondence should be addressed at: Laboratory of Molecular Biology, 37 Convent Drive, Room 5106, National Cancer Institute, Bethesda, MD 20892. E-mail: pasta@helix.nih.gov.

on ice to swell for 10 min and ruptured with a Dounce by using pestle B (Kontes). The membrane-associated polyribosomes and cytosolic polyribosomes were separated by isopycnic centrifugation in a discontinuous sucrose density gradient at 90,000 × *g* for 15 h at 4°C (5). The total RNA was isolated from the membrane polyribosomal fraction by using Trizol LS reagent (Invitrogen Life Technologies, Carlsbad, CA). The quality of the total RNA was verified by using the Agilent 2100 bioanalyzer. Individual preps of membrane-associated polyribosomal RNA from each cell line were pooled as follows: 300 $\mu$g each isolated from MCF7, SK-BR-3, ZR-75-1, MDA-MB-231, and hTERT-HME1 cells and 200 $\mu$g from LNCaP cells. Of the pooled RNA, 100 $\mu$g was saved for future analysis, and 1.6 mg was given to Invitrogen Life Technologies for library construction.

**Generation of the MAPcL.** mRNA was isolated from the pooled total membrane-associated polyribosomal RNA, and cDNA was generated by using an oligo(dT) primer by Invitrogen Life Technologies. The cDNA fragments were cloned directionally into the *Eco*RV and *Not*I sites of pCMVSport6.0 (Invitrogen Life Technologies), resulting in the destruction of the *Eco*RV site. The library was electroporated into *Escherichia coli* EMDH10B cells, and the titer of the library was determined. Twenty-three clones were picked randomly to determine the average insert size of the library.

**Subtraction of the MAPcL.** Clones ($5 \times 10^6$) of the MAPcL were amplified 26,000-fold by Invitrogen Life Technologies by using their semisolid agarose procedure, which minimizes clone bias that normally occurs during liquid amplification. A driver library was created by pooling Invitrogen Life Technologies's premade liver, brain, kidney, lung, and skeletal muscle libraries in equimolar amounts. The amplified MAPcL was subtracted with the driver library (6). The subtracted library contains $1.3 \times 10^7$ colony-forming units total with an average insert size of 1,800 bp.

**Sequencing of the Subtracted and Unsubtracted MAPcL.** The 5′ sequencing reactions were performed at the Advanced Technology Center (Gaithersburg, MD) by using the M13 reverse primer.

**Human Multiple Tissue Expression Array and Northern Blot Hybridization.** The human multiple tissue expression array was purchased from CLONTECH. The 1.4-kb *BASE* probe used for hybridization was generated by digesting the MAPcL clone pKAE08h07 with *Eco*RI and *Not*I and purifying the cDNA insert by agarose gel electrophoresis. The cDNA insert was labeled with $^{32}$P by random primer extension (Lofstrand Laboratories), and the hybridization conditions were performed as described (7). The membrane was exposed to film for 2 days.

Samples for Northern blot hybridization, 2 $\mu$g of poly(A) RNA per lane, were separated by using a 1.25% agarose gel containing 2% formaldehyde. Salivary gland poly(A) RNA was purchased from CLONTECH, and ZR-75-1 poly(A) RNA was generated by using the FastTrack 2.0 mRNA-isolation system from Invitrogen. Generation of the *BASE* probe and hybridization and washing conditions were performed as described above. The 0.24- to 9.5-kb RNA ladder was purchased from Invitrogen Life Technologies. The blot was exposed to film for 1 day.

**RT-PCR and Rapid-Scan Gene Expression Panel Analysis.** Total RNA was isolated from frozen breast tumor samples acquired from the Cooperative Human Tissue Network and tissue culture cell lines by using the StrataPrep Total RNA miniprep kit (Stratagene) according to manufacturer instructions. To generate single-stranded cDNAs, total RNA (5 $\mu$g) was used with the First-Strand cDNA synthesis kit by using random hexamer priming according to manufacturer instructions (Amersham Pharmacia). PCRs were performed by using the following protocol: initial denaturation at 94°C for 3 min, 35 cycles of denaturation at 94°C for 1 min, annealing at 60°C for 1 min, elongation at 72°C for 1 min, and a final 5-min extension at 72°C. Similar PCR conditions were used with the Rapid-Scan gene expression panel except elongation at 72°C was performed for 2 min (OriGene Technologies).

**Sequence Identification.** Sequences of clones from the MAPcL were identified initially by comparison to the NCBI RefSeq, GenBank, and expressed sequence tags databases (dbEST) using BLAST (8, 9). Full-length clones were identified as MAPcL sequences with a hit to a RefSeq protein at 70% identity or better and an alignment starting at amino acid 1 of the RefSeq protein. Membrane and secreted proteins were identified by using Gene Ontology Consortium (GO) classifications associated with RefSeq genes. The NIH_MGC_87 cDNA library from the NIH Mammalian Gene Collection was used as a control for membrane and secreted proteins (10). This library contains >19,000 ESTs and was made from an adenocarcinoma breast tissue-derived cell line. MAPcL sequences representing unknown genes were classified by tissue expression by using EST sequences from the dbEST. (We are in the process of depositing the 10,500 sequences representing known genes in GenBank under the library name MAPcL.)

**Web Sites.** UniGene, RefSeq, dbEST, and GenBank sequence databases, the BLAST program, LocusLink, OMIM, and the CGAP project can be accessed from www.ncbi.nlm.nih.gov. The GO database can be obtained from ftp://ftp.geneontology.org/pub/go. The NIH_MGC_87 cDNA library can be obtained from http://mgc.nci.nih.gov/. The GoldenPath genome build and annotation databases can be accessed from http://genome.ucsc.edu. All database versions except GoldenPath were taken as a snapshot from public releases available as of March 14, 2002. Genome sequences and annotations were taken from the December 2001 build of GoldenPath.

## Results

**Generation of a MAPcL.** To generate a breast and prostate cancer cDNA library enriched with genes that encode membrane and secreted proteins, membrane-associated polyribosomal RNA was isolated from four breast cancer cell lines (MCF7, ZR-75-1, SK-BR-3, and MDA-MB-231), one telomerase immortalized normal breast cell line (hTERT-HME1), and the prostate cancer cell line (LNCaP) that produces prostate-specific antigen. The addition of the prostate cancer cell line RNA served two purposes. It served as a test of our approach because if our hypothesis was correct, we expected our library to be enriched in prostate-specific antigen. Also, it gives us the opportunity to discover unknown genes expressed in prostate cancers. It was shown previously by using cDNA microarray analysis that there are two main subgroups of breast tumors based on their gene-expression profiles: estrogen receptor (ER)-positive and ER-negative (11). In addition, the overexpression of ErbB2 correlated with low levels of the ER. Because there are numerous breast cancer cell lines available, we chose four to represent the recognized range of phenotypic diversity of breast tumors. MCF7 and ZR-75-1 both express the ER and express ErbB2 at low levels. SK-BR-3 and MDA-MB-231 do not express the ER. SK-BR-3 contains gene amplifications of ErbB2, whereas MDA-MB-231 expresses ErbB2 at low levels. Membrane-associated polyribosomal RNA was isolated individually from the six cell lines, and the RNA was pooled. A cDNA library was generated from the pooled membrane-associated polyribosomal RNA as described in *Materials and Methods*. The library contains $2.01 \times 10^7$ colony-forming units total with an average insert size of 2 kb.

**Subtraction of the Initial MAPcL.** To remove ubiquitously expressed genes and enrich for genes specifically expressed in breast and prostate cancers, the initial MAPcL was subtracted by using

**Table 1. Initial library list of the most abundant MAPcL genes**

| Count* | Symbol | Name | Cellular location† |
|---|---|---|---|
| 9 | FN1 | Fibronectin 1 | Secreted |
| 7 | GAPDH | Glyceraldehyde-3-phosphate dehydrogenase | Cytoplasm |
| 7 | KRT8 | Keratin 8 | Cytoskeletal |
| 6 | EEF1A1 | Eukaryotic translation elongation factor 1$\alpha$1 | Cytoplasm |
| 6 | GRP58 | Glucose-regulated protein, 58 kDa | Endo retic‡ |
| 6 | KRT18 | Keratin 18 | Cytoskeletal |
| 5 | SSR2 | Signal sequence receptor, $\beta$ | Endo retic‡ |
| 5 | TRA1 | Tumor rejection antigen (gp96) 1 | Membrane |
| 4 | EIF4A2 | Eukaryotic translation initiation factor 4A, isoform 2 | Cytoplasm |
| 4 | ANXA2 | Annexin A2 | Membrane |
| 4 | RPL4 | Ribosomal protein L4 | Cytoplasm |
| 4 | PPIB | Peptidylprolyl isomerase B (cyclophilin B) | Endo retic‡ |
| 4 | CD151 | CD151 antigen | Membrane |
| 4 | P4HB | Procollagen-proline, 2-oxoglutarate 4-dioxygenase, $\beta$ | Endo retic‡ |
| 3 | — | DKFZP566C243 protein | Mitochondria |

*Of 943 sequences from the initial library.
†Determined from GO, LocusLink, and OMIM.
‡Endoplasmic reticulum.

biotinylated RNA generated from five normal libraries: brain, liver, lung, kidney, and skeletal muscle as described in *Materials and Methods*. The efficiency of subtraction was determined by measuring the level of a housekeeping gene, eukaryotic elongation factor 1$\alpha$1, or *EEF1A1*. The *EEF1A1* gene was reduced by 85-fold in the subtracted library (data not shown).

To determine which genes are represented in the initial and subtracted MAPcLs, one sequencing reaction was performed on the 5′ end of 943 unsubtracted clones and 15,581 subtracted clones that were chosen randomly. The sequences from the two libraries were compared with known genes in RefSeq, a public database of curated genes (8). The most abundant known genes in the initial and subtracted libraries are shown in Tables 1 and 2. Although the initial library was made from enriched membrane-associated polyribosomal RNA, it still contained cDNAs derived from highly expressed genes that encode soluble, housekeeping proteins such as glyceraldehyde-3-phosphate dehydrogenase and EEF1A1 (Table 1). Subtraction of the initial library successfully removed these contaminating sequences such that of the 15,581 sequenced clones, glyceraldehyde-3-phosphate dehydrogenase was not present and only one clone encoded EEF1A1. Furthermore, the most abundant gene was prostate-specific antigen, a secreted protein highly expressed in the LNCaP cell line (Table 2).

**Percentage of cDNA Inserts Representing the Entire Coding Region.** Because all sequencing reactions were performed from the 5′ end, it is possible to determine what percentage of the cDNA inserts encode full-length transcripts of known genes. Using BLASTX to compare translated MAPcL sequences to the RefSeq protein database, we determined that 30% of the MAPcL sequences contain the 5′ end of the encoded proteins (9).

**Quantitation of Genes Encoding Membrane and Secreted Proteins.** To quantify the enrichment of clones encoding membrane and secreted proteins, MAPcL sequences representing known genes were assessed by using the GO database (12). According to the cellular location classification from the GO database, 49% of the known genes in the subtracted MAPcL encode membrane or secreted proteins. In contrast, only 14% of known genes encode membrane or secreted proteins from a control library derived from unfractionated mRNA from an adenocarcinoma breast tissue (see *Materials and Methods*). Cellular locations of the most abundant genes in the initial and subtracted libraries are listed in Tables 1 and 2.

**Table 2. Subtracted MAPcL list of the most abundant genes**

| Count* | Symbol | Name | Cellular location† |
|---|---|---|---|
| 87 | KLK3 | Kallikrein 3, (prostate-specific antigen) | Secreted |
| 86 | SLC7A5 | Solute carrier family 7 (cationic amino acid transporter, y+ system), member 5 | Membrane |
| 59 | KRT18 | Keratin 18 | Cytoskeletal |
| 58 | ITGA3 | Integrin, $\alpha$3 ($\alpha$3 subunit VLA-3 receptor) | Membrane |
| 56 | ITGB4 | Integrin, $\beta$4 | Membrane |
| 48 | LENG4 | Leukocyte receptor cluster (LRC) member 4 | Membrane |
| 43 | LAMC2 | Laminin, $\gamma$2 | Secreted |
| 42 | ERBB2 | v-erb-b2 erythroblastic leukemia viral oncogene homolog 2 | Membrane |
| 41 | KRT14 | Keratin 14 | Cytoskeletal |
| 40 | SPINT2 | Serine protease inhibitor, Kunitz type, 2 | Membrane |
| 40 | BSG | Basigin (OK blood group) | Membrane |
| 35 | LGALS3BP | Lectin, galactoside-binding, soluble 3-binding protein | Secreted |
| 34 | KRT17 | Keratin 17 | Cytoskeletal |
| 34 | SLC16A3 | Solute carrier family 16 (monocarboxylic acid transporters), member 3 | Membrane |
| 30 | GRN | Granulin | Secreted |

*Of 15,581 sequences from the subtracted library.
†Determined from GO, LocusLink, and OMIM.

**Table 3. Classification of the subtracted MAPcL sequences**

| | No. of sequences | No. of genes represented | %* |
|---|---|---|---|
| RefSeq genes† | 10,506 | 3,814‡ | 54.5 |
| Non-RefSeq UniGene clusters§ | 4,074 | 2,382‡ | 34.0 |
| Non-UniGene EST hits¶ | 354 | 342‖ | 5.0 |
| No EST hits | 647 | 457‖ | 6.5 |
| Total | 15,581 | 6,995 | 100 |

*Percentage of total genes represented.
†Determined by BLAST against the RefSeq database.
‡Determined by counting unique RefSeq genes or UniGene clusters.
§Determined by BLAST against NT database and dbEST.
¶Determined by BLAST against the dbEST.
‖Determined by BLAST of the MAPcL sequences against themselves.

**Unknown Sequences.** We classified the 15,581 sequences from clones of the subtracted MAPcL as either known or unknown based on a BLAST analysis (Table 3; ref. 9). Sequences were labeled as known if they aligned to a gene sequence in the RefSeq database; otherwise, they were labeled unknown. Of 15,581 MAPcL sequences, 10,506 sequences aligned with 3,814 RefSeq genes. We further divided the remaining 5,075 unknown sequences into three groups: (*i*) 4,074 sequences, aligned with 2,382 UniGene clusters that were not associated with known genes, (*ii*) 354 sequences, representing 342 unique transcripts, overlapped with ESTs that were not part of any UniGene clusters, and (*iii*) 647 sequences, representing 457 unique transcripts, which did not overlap any known sequences (13). Numbers of sequences with each classification are given in Table 3.

The 5,075 sequences from the subtracted MAPcL that are not associated with known genes were examined to narrow the search for genes encoding potential immunotherapy targets. Candidate sequences chosen for further study either align to EST sequences derived only from nonessential tissue libraries, have alternative splice forms different from ESTs derived from essential tissues, or do not align with any ESTs. We aligned the sequences to the human genome using BLAST from the GoldenPath project (December 2001 build) and surveyed the genomic region around these sequences for evidence of gene structure based on other ESTs that were also aligned to the genome (14–16). MAPcL sequences that seemed to represent the 5′ end of genes containing ESTs from excluded tissues were eliminated. In addition, all candidate MAPcL sequences contain a predicted ORF based on the sequence obtained from one reaction from the 5′ end.

**Characterization of *BASE*.** As an example, one previously uncharacterized sequence from the subtracted MAPcL that fits the above criteria was experimentally characterized and designated *BASE*. Fig. 1 shows the cDNA sequence of *BASE* (KAE08h07) aligned to chromosome 20. Initially, a single 5′ sequencing reaction was performed, and the sequence was aligned with the human genome (Fig. 1, pink boxes). Completion of the full-length sequence (Fig. 1, blue boxes) shows that *BASE* has an ORF encoding a 19.5-kDa protein. Analysis of the amino acid sequence of BASE by using the PSORT program predicts it to be a secreted protein (17). Three additional MAPcL cDNA sequences align with the KAE08h07 sequence; however, no ESTs in the dbEST (13) align with *BASE* (Fig. 1). Because coverage of the dbEST is incomplete (18), expression specificity of *BASE* had to be verified experimentally.

Because membrane-associated RNAs derived from diverse cell lines were used to make the MAPcL, we determined which of the cell lines express *BASE* using RT-PCR analysis of the membrane-associated polyribosomal RNA (Fig. 2*a*). The specific primers used for PCR are located in separate exons of *BASE* and amplify a 464-bp fragment (Fig. 1). *BASE* had strong expression in the breast cancer cell line, ZR-75-1, and low expression in SK-BR-3 and MCF7. No expression was detected in the normal breast cell line, hTERT-HME, and weak expression was observed in LNCaP (Fig. 2*a*). As a control for the quality of the generated cDNA, separate PCRs were performed by using primers to the *transferrin receptor* (Fig. 2*a*).

For the potential proteins encoded by the MAPcL genes to be used as therapeutic targets or diagnostic markers, the genes must be expressed in breast cancers. To examine expression levels of *BASE* in breast cancers, an RT-PCR analysis was performed (Fig. 2*b*). Total RNA was isolated from eight primary and three metastatic frozen breast cancer samples from patients, and the RNA was used as a template to generate cDNA. The *BASE*-specific primers used for PCR are shown in Fig. 1. *BASE* was expressed in five primary breast cancer samples (Fig. 2*b*, lanes 2, 5, 7, 8, and 11) and one metastatic sample (Fig. 2*b*, lane 10). As a positive control for the PCRs, pKAE08h07 was used as a template (Fig. 2*b*, lane 12). Separate PCRs were performed by using actin primers to verify the quality of the generated cDNA

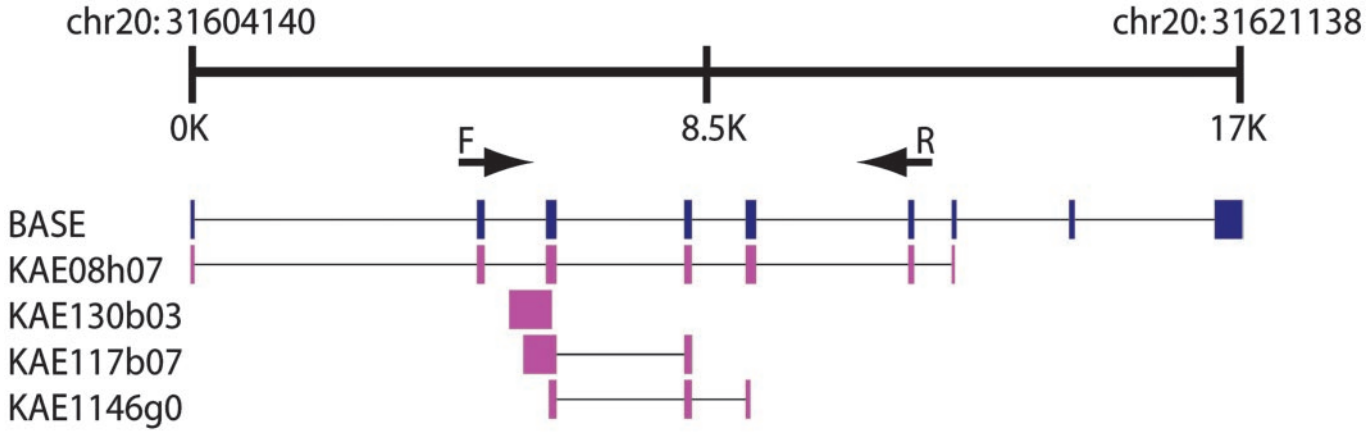


**Fig. 1.** Alignment of the *BASE* cDNA sequence with the human genome. The thick black line represents DNA sequence from the human genome from chromosome 20. The numbers above the black line indicate the location of the sequence on chromosome 20, and the numbers below the black line are relative locations in base pairs (GoldenPath, December 2001). The blue boxes represent exons from the full-length sequence of the *BASE* cDNA clone (KAE08h07). The pink boxes represent the exons from the single 5′ sequencing reaction for the MAPcL cDNA clones listed (*Left*). Connecting thin black lines represent introns. The locations of primers used to amplify *BASE* are indicated as arrows labeled F (forward) and R (reverse).

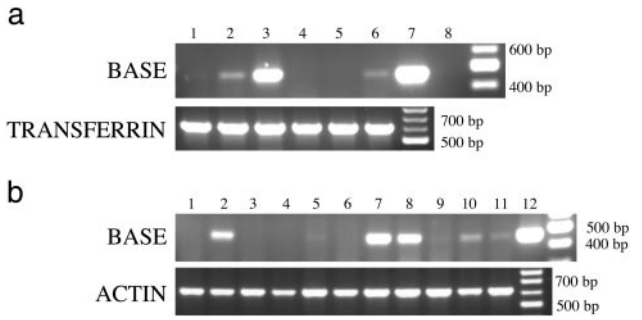


**Fig. 2.** Expression of *BASE* in the MAPcL cell lines and breast cancers. (*a*) Expression of *BASE* in the MAPcL cell lines. Expression levels were determined by RT-PCR using membrane-associated polyribosomal RNA isolated from the library cell lines as a template for cDNA synthesis. PCR was performed by using primers to *BASE* (Fig. 1). The PCR products were analyzed on a 1.5% agarose gel with ethidium bromide staining as follows: lane 1, MCF7; lane 2, SK-BR-3; lane 3, ZR-75-1; lane 4, MDA-MB-231; lane 5, hTERT-HME1; lane 6, LNCaP; lane 7, pKAE08h07 (*BASE*); and lane 8, no template. Separate PCRs were done by using *transferrin receptor* primers, which amplify a 615-bp fragment, to verify the quality of the generated cDNA. The DNA ladder in base pairs is indicated on the right. (*b*) Expression of *BASE* in breast cancer samples. RT-PCR analysis was performed by using 11 breast cancer total RNA samples as templates for cDNA synthesis. The *BASE* PCR primers, shown in Fig. 1, amplify a 464-bp fragment. The PCR products were analyzed on a 1.5% agarose gel with ethidium bromide staining as follows: lanes 1–11, breast tumors; and lane 12, pKAE08h07 (*BASE*). PCRs using actin primers were performed separately and produced a 640-bp product. The DNA ladder in base pairs is indicated on the right.

(Fig. 2*b*). Expression of *BASE* in breast cancers was confirmed by using *in situ* hybridization analysis (data not shown).

The expression profile of *BASE* was analyzed by using a human multiple tissue expression array containing mRNA from 61 different normal tissues. When the cDNA insert of the *BASE* clone pKAE08h07 was used as a probe, it only reacted with salivary gland mRNA (Fig. 3*a*, E9). Expression of *BASE* was not detected in the mammary gland mRNA sample (Fig. 3*a*, F9). Next, PCR analysis was performed by using a rapid-scan panel containing cDNA samples derived from 24 normal tissues (Fig. 3*b*). The rapid-scan panel revealed an abundant 464-bp PCR product from cDNA derived from salivary gland (Fig. 3*b*, lane 13), confirming the dot blot result. To verify the quality of the cDNA templates, separate PCRs were performed by using actin primers, and bands of equal intensity were observed (Fig. 3*b*).

To determine the transcript length of *BASE* and verify that the cDNA clone represents the full-length transcript, Northern blot analysis was performed. Salivary gland and ZR-75-1 mRNA were probed with the 1.4-kb cDNA insert of KAE08h07 (Fig. 3*c*). Two bands were observed with the most abundant transcript at ≈2.3 kb and a less abundant transcript at ≈1.7 kb. The insert size for the KAE08h07 *BASE* clone is 1.4 kb, which corresponds to the 1.7-kb transcript with the addition of the poly(A) tail. The 2.3-kb transcript indicates that there may be alternatively spliced forms of *BASE*, which is consistent with the presence of two MAPcL clones, KAE130b03 and KAE117b07, that overlap KAE08h07 but have an extended exon (Fig. 1). A strong actin band was observed with all the samples.

*BASE* is an example of a previously uncharacterized breast cancer gene from the subtracted MAPcL that does not overlap any other sequences in the human database. The identification of *BASE*, which encodes a predicted secreted protein that is expressed in breast cancers but not in the organs used for subtraction, demonstrates that our method was successful. If BASE is present in human blood, it could be of use in the diagnosis of breast cancer.

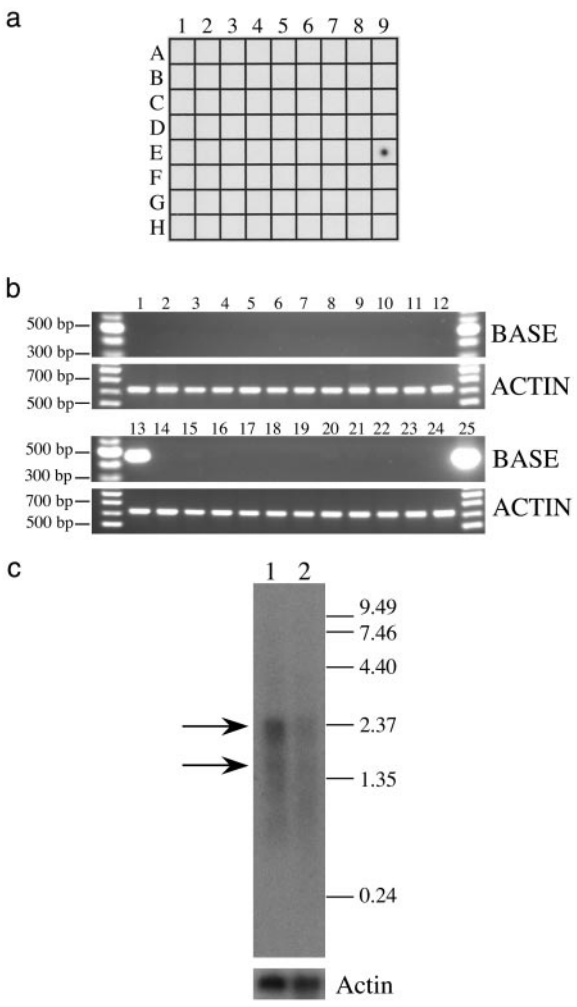


**Fig. 3.** Analysis of *BASE* expression in normal tissues and transcript size. (*a*) Expression of *BASE* in normal tissues. A human multiple tissue expression array containing 61 tissue-specific mRNA samples was hybridized with the cDNA insert of *BASE* (KAE08h07). The array is as follows: A1, whole brain; B1, cerebral cortex; C1, frontal lobe; D1, parietal lobe; E1, occipital lobe; F1, temporal lobe; G1, paracentral gyrus of cerebral cortex; H1, pons; A2, cerebellum left; B2, cerebellum right; C2, corpus callosum; D2, amygdala; E2, caudate nucleus; F2, hippocampus; G2, medulla oblongata; H2, putamen; A3, substantia nigra; B3, nucleus accumbens; C3, thalamus; D3, pituitary gland; E3, spinal cord; A4, heart; B4, aorta; C4, atrium left; D4, atrium right; E4, ventricle left; F4, ventricle right; G4, interventricular septum; G4, apex of the heart; A5, esophagus; B5, stomach; C5, duodenum; D5, jejunum; E5, ileum; F5, ileocecum; G5, appendix; H5, colon ascending; A6, colon transverse; B6, colon descending; C6, rectum; A7, kidney; B7, skeletal muscle; C7, spleen; D7, thymus; E7, peripheral blood leukocytes; F7, lymph node; G7, bone marrow; H7, trachea; A8, lung; B8, placenta; C8, bladder; D8, uterus; E8, prostate; F8, testis; G8, ovary; A9, liver; B9, pancreas; C9, adrenal gland; D9, thyroid gland; E9, salivary gland; and F9, mammary gland. (*b*) RT-PCR analysis of *BASE* expression in 24 normal tissues. PCRs were performed by using a rapid-scan gene expression panel containing cDNA samples from 24 different normal tissues: 1, brain; 2, heart; 3, kidney; 4, spleen; 5, liver; 6, colon; 7, lung; 8, small intestine; 9, muscle; 10, stomach; 11, testis; 12, placenta; 13, salivary gland; 14, thyroid; 15, adrenal gland; 16, pancreas; 17, ovary; 18, uterus; 19, prostate; 20, skin; 21, peripheral blood lymphocyte; 22, bone marrow; 23, fetal brain; 24, fetal liver; and 25, pKAE08h07. PCR primers for *BASE* are shown in Fig. 1. These primers are located in different exons and amplify a 464-bp *BASE* fragment. As a positive control, pKAE08h07 (*BASE*) was used as a template for the PCR (lane 25). PCRs using actin primers were performed separately and produced a 640-bp product. The PCR products were analyzed on a 1.5% agarose gel with ethidium bromide staining. (*c*) Northern blot analysis of *BASE* transcripts. Each lane contains poly(A) RNA (2 $\mu$g) from salivary gland (lane 1) and ZR-75-1 cells (lane 2). The membrane was probed with the 1.4-kb cDNA insert of KAE08h07. The membrane was stripped and analyzed with the $\beta$-actin probe to verify equal loading (*Lower*). RNA size markers in kilobases are indicated on the right.

## Discussion

A molecular approach was used to identify previously uncharacterized genes encoding membrane and secreted proteins expressed in breast cancers that have limited expression in normal tissues for use as immunotherapy targets and diagnostic markers. To increase our chances of finding these membrane and secreted proteins, membrane-associated polyribosomal mRNA, which encodes membrane and secreted proteins, was isolated from four breast cancer cell lines, one normal breast cell line, and a prostate cancer cell line. A cDNA library was generated and subsequently subtracted with five different libraries made from normal tissues to reduce ubiquitously expressed genes and enrich for genes expressed in breast and prostate cancers. This approach was especially feasible for breast cancer because numerous cell lines are available with a diverse range of phenotypes.

To determine what genes are represented in the subtracted MAPcL, we used an unbiased method of randomly sequencing 15,581 clones. We found that 45% of the nonredundant MAPcL sequences did not align with any known genes as determined by using BLAST against the RefSeq database (Table 3), indicating that this approach was very successful in identifying previously unknown genes. Of the 2,382 UniGene clusters represented by the MAPcL sequences not associated with known genes, 27 align with ESTs derived only from libraries made from nonessential tissues such as breast, ovary, and testis. Of the 342 unique sequences that align with non-UniGene ESTs, 50 sequences have alignment restricted to ESTs derived from nonessential tissues. Lastly, 457 unique sequences align with no ESTs, and consequently information about tissue specificity of expression is not available for these clones. Most of these sequences probably represent transcripts that previously have not been detected. However, this number may be an overestimate, because the MAPcL clones have an average insert size of 1,800 bp, whereas the sequences are 578 bp on average and are generated from the 5′ end of the cDNA clones. Although we have estimated that 30% of the MAPcL cDNA inserts contain the entire ORF of the encoded protein, UniGene EST sequences in the database most often consist of the 3′ end of transcripts. Consequently, the 5′ sequences from the full-length MAPcL clones may not overlap with the corresponding 3′ EST clusters. Ongoing full-length sequencing of the clones will allow identification of the corresponding UniGene EST clusters and thus provide information on tissue expression.

As an example, we analyzed the expression pattern of a breast cancer gene that we designated *BASE*. *BASE* does not overlap any ESTs in the dbEST, and *BASE* was not expressed in the organs used for subtraction, indicating that our method was successful. *BASE* was expressed only in salivary gland (Fig. 3 *a* and *b*). Most importantly, although the gene was identified by using tissue-culture cell lines, *BASE* also was expressed in both primary and metastatic human breast cancers as determined by RT-PCR (Fig. 2*b*) and *in situ* hybridization (data not shown). The PSORT program predicts that BASE is a secreted protein (17). Results of a protein BLAST search indicate that BASE shares sequence similarity with Latherin, a 228-aa protein that is a major component of horse sweat and is responsible for rendering hydrophobic surfaces wettable by water (19). BASE, a 179-aa protein, is 42% identical and 63% similar to the first 178 aa of Latherin. Antibodies will be generated against BASE to clarify its cellular location.

Subtraction of the initial library using biotinylated RNA derived from normal tissues enriched for transcripts that encode membrane and secreted proteins and genes that are up-regulated in breast and prostate cancers (Table 2). Based on the cellular location of proteins encoded by the known genes in the library, 49% of the MAPcL clones encode membrane or secreted proteins. Furthermore, 12 of the 15 most abundant genes represented in the subtracted MAPcL encode either secreted or membrane proteins (Table 2). The most abundant gene from the subtracted MAPcL is *kallikrein 3* (Table 2). The abundance of this gene is a good verification for subtraction of the library because it is expressed by the prostate cancer cell line LNCaP, encodes a secreted protein, and has expression associated with prostate tissue (Table 2).

Some of the most abundant genes represented in the subtracted MAPcL are associated with or up-regulated in breast cancer. The eighth most abundant gene from the subtracted library is *ErbB2*, which encodes a membrane protein. All the breast cancer cell lines used to make the MAPcL express *ErbB2* with SK-BR-3 containing gene amplification of *ErbB2*. Keratin proteins are used as markers for the detection of epithelial cells (20). The breast cancer cell lines MCF7, ZR-75-1, SK-BR-3, and MDA-MB-231 have been shown to produce large amounts of keratin 8, keratin 18, and keratin 19 (21). Furthermore, expression of keratin 8 and keratin 18 is normally maintained in carcinomas, whereas expression of other keratin family members is frequently lost (22). Keratin 18 is the third most abundant gene represented in the subtracted MAPcL, and there are 20 MAPcL clones that encode keratin 8 and 20 clones that encode keratin 19 (Table 2). In addition, MUC1, which is up-regulated ≈10-fold in 90% of breast tumors (23), is encoded by 12 cDNA clones.

Finally, with >3,000 unknown nonredundant transcripts and 49% of the MAPcL clones encoding membrane or secreted proteins, this library may contain numerous genes encoding therapeutic targets or diagnostic proteins for breast cancer.

We thank Drs. Paul Egland, Roberto Di Lauro, and Tapan Bera for helpful suggestions and critical reading of the manuscript, Verity Fogg for outstanding tissue-culture support, Stephen Squires for technical assistance, and Anna Mazzuca for excellent editorial assistance. This work was supported in part by a Collaborative Research and Development Agreement with the IDEC Pharmaceuticals Corporation. K.A.E. is a fellow in the National Institute of General Medical Sciences Pharmacology Research Associate Program.

1. Bange, J., Zwick, E. & Ullrich, A. (2001) *Nat. Med.* **7,** 548–552.
2. Shak, S. (1999) *Semin. Oncol.* **26,** 71–77.
3. Slamon, D. J., Clark, G. M., Wong, S. G., Levin, W. J., Ullrich, A. & McGuire, W. L. (1987) *Science* **235,** 177–182.
4. Slamon, D. J., Godolphin, W., Jones, L. A., Holt, J. A., Wong, S. G., Keith, D. E., Levin, W. J., Stuart, S. G., Udove, J., Ullrich, A. & Press, M. F. (1989) *Science* **244,** 707–712.
5. Mechler, B. M. (1987) *Methods Enzymol.* **152,** 241–248.
6. Li, W. B., Gruber, C. E., Lin, J. J., Lim, R., D'Alessio, J. M. & Jessee, J. A. (1994) *BioTechniques* **16,** 722–729.
7. Essand, M., Vasmatzis, G., Brinkmann, U., Duray, P., Lee, B. & Pastan, I. (1999) *Proc. Natl. Acad. Sci. USA* **96,** 9287–9292.
8. Pruitt, K. D. & Maglott, D. R. (2001) *Nucleic Acids Res.* **29,** 137–140.
9. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990) *J. Mol. Biol.* **215,** 403–410.
10. Strausberg, R. L., Feingold, E. A., Klausner, R. D. & Collins, F. S. (1999) *Science* **286,** 455–457.
11. Perou, C. M., Sorlie, T., Eisen, M. B., Van De Rijn, M., Jeffrey, S. S., Rees, C. A., Pollack, J. R., Ross, D. T., Johnsen, H., Akslen, L. A., *et al.* (2000) *Nature* **406,** 747–752.
12. The Gene Ontology Consortium (2001) *Genome Res.* **11,** 1425–1433.
13. Boguski, M. S., Lowe, T. M. & Tolstoshev, C. M. (1993) *Nat. Genet.* **4,** 332–333.
14. Kent, W. J. & Haussler, D. (2001) *Genome Res.* **11,** 1541–1548.
15. Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M. & Haussler, D. (2002) *Genome Res.* **12,** 996–1006.
16. Kent, W. J. (2002) *Genome Res.* **12,** 656–664.
17. Nakai, K. & Horton, P. (1999) *Trends Biochem. Sci.* **24,** 34–36.
18. Zhang, L., Zhou, W., Velculescu, V. E., Kern, S. E., Hruban, R. H., Hamilton, S. R., Vogelstein, B. & Kinzler, K. W. (1997) *Science* **276,** 1268–1272.
19. Beeley, J. G., Eason, R. & Snow, D. H. (1986) *Biochem. J.* **235,** 645–650.
20. Ronnov-Jessen, L., Petersen, O. W. & Bissell, M. J. (1996) *Physiol. Rev.* **76,** 69–125.
21. Trask, D. K., Band, V., Zajchowski, D. A., Yaswen, P., Suh, T. & Sager, R. (1990) *Proc. Natl. Acad. Sci. USA* **87,** 2319–2323.
22. Oshima, R. G., Baribault, H. & Caulin, C. (1996) *Cancer Metastasis Rev.* **15,** 445–471.
23. Hadden, J. W. (1999) *Int. J. Immunopharmacol.* **21,** 79–101.