[CANCER RESEARCH 64, 8682–8687, December 1, 2004]

Membrane-Associated and Secreted Genes in Breast Cancer

Nathan O. Stitziel,[1] Brenton G. Mar,[2] Jie Liang,[1] and Carol A. Westbrook[3]

Departments of [1]Bioengineering and [2]Biochemistry and Molecular Genetics, University of Illinois at Chicago, Chicago, Illinois; and [3]Department of Medicine, Boston University, Boston, Massachusetts

ABSTRACT

The identification of membrane-associated and secreted genes that are differentially expressed is a useful step in defining new targets for the diagnosis and treatment of cancer. Extracting information on the subcellular localization of genes represented on DNA microarrays is difficult and is limited by the incomplete sequence and annotation that is available in existing databases. Here we combine a biochemical and bioinformatic approach to identify membrane-associated and secreted genes expressed in the MCF-7 breast cancer cell line. Our approach is based on the analysis of differential hybridization levels of RNAs that have been physically separated by virtue of their association with polysomes on the endoplasmic reticulum. This approach is specifically applicable to oligonucleotide microarrays such as Affymetrix, which use single-color hybridization instead of dual-color competitive hybridizations. Assignment to membrane-associated and secreted class membership is based on both the differential hybridization levels and an expression threshold, which are calculated empirically from data collected on a reference set of known cytoplasmic and membrane proteins. This method enabled the identification of 755 membrane-associated and secreted probe sets expressed in MCF-7 cells for which this annotation did not previously exist. The data were used to filter a previously reported expression dataset to identify membrane-associated and secreted genes which are associated with poor prognosis in breast cancer and represent potential targets for diagnosis and treatment. The approach reported here should provide a useful tool for the analysis of gene expression patterns, identifying membrane-associated or secreted genes with biological relevance that have the potential for clinical applications in diagnosis or treatment.

INTRODUCTION

With the advent of high-throughput global genomic strategies, the potential exists for the identification of many novel genes that have a specific association with cancer, and the gene product of which has diagnostic or therapeutic implications. The task remains to determine which of these have the most immediate potential for clinical translation. Among the most useful proteins in the clinical setting are those that are associated with the cancer cell membrane, including those that are membrane-bound and those that are secreted extracellularly (referred to as membrane-associated and secreted or membrane-associated and secreted genes). Membrane-bound proteins include surface antigen targets for diagnosis or treatment, receptors for external factors that regulate cell growth, and proteins that regulate cell adhesion and metastases. Secreted proteins and peptides can be used as circulating tumor markers for diagnosis and monitoring.

The characterization of a novel gene as one that encodes a membrane-associated or secreted protein can be difficult. Although computational methods exist for predicting whether a protein is membrane-bound or secreted (1, 2), these methods cannot be applied to incomplete or poorly annotated gene sequence and are inexact even in the best setting.

An alternative method to identify membrane-associated and secreted genes experimentally based on differential hybridization to glass slide cDNA arrays was recently shown (3). This method takes advantage of the fact that proteins that function at the membrane surface or are immediately secreted are preferentially translated from ribosomes at the endoplasmic reticulum to which they are directed by their signal peptide. Because their association with the endoplasmic reticulum membrane makes them less dense, these membrane-bound polysomes can be separated from their heavier cytosolic counterparts by sucrose gradient centrifugation (4). RNA prepared from these two cellular subfractions is used for differential cDNA hybridization to identify those that are most highly associated with the membrane-bound polysomes.

Here we report the application of this method to the global analysis of the genes expressed in a breast cancer cell line, MCF-7, modifying it for the widely used Affymetrix chips. We develop a statistical approach to determine the membrane association of each Affymetrix probe set, as expressed in the cell line, by comparing the ratio on two chips to a reference set of known cytoplasmic and membrane proteins. The results of this study were then used to analyze the data from a previously reported differential expression study in breast cancer (5), to identify membrane-associated and secreted genes that are associated with poor prognosis, demonstrating the utility of this approach to identify potential targets for diagnosis and treatment from differential hybridization studies.

MATERIALS AND METHODS

Cell Line Preparation

MCF-7 cells were purchased from the American Type Culture Collection (Manassas, VA) and were cultured in Eagle's MEM supplemented with 0.01 mg/mL bovine insulin, 10% fetal bovine serum in 5% CO_2 at 37°C.

Polysome Fractionation

Polysomes were fractionated by sucrose density gradient centrifugation with a modification of the method described by Mechler (4). After treatment with cyclohexamide (10 μg/mL) for 10 minutes at 37°C, 3×10^8 MCF-7 cells in log growth were collected by scraping the dishes into cold PBS. The cells were then resuspended at a concentration of 2.5×10^8 cells/mL in a hypotonic lysis buffer [10 mmol/L KCl, 1.5 mmol/L $MgCl_2$, and 10 mmol/L Tris-Cl (pH 7.4)] and were allowed to rest on ice for 10 minutes. After lysing cells with a Dounce homogenizer, nuclear and cell debris were removed by centrifugation at $2,000 \times g$ (4°C) for 2.5 minutes. The supernatant was loaded on a discontinuous-step sucrose gradient (2.5 mol/L, 2.1 mol/L, 1.95 mol/L, and 1.3 mol/L sucrose) and centrifuged at $26,000 \times g$ for 5 hours. After centrifugation, successive 1.5 mL fractions were collected from the bottom of the centrifugation tube, and the $A_{260\,nm}$ was measured to estimate the RNA content.

RNA Preparation

Total RNA was isolated from the pooled sucrose gradient fractions by mixing with TRIzol LS reagent (Invitrogen, Carlsbad, CA) at a 3:1 ratio

Received 5/16/04; revised 8/5/04; accepted 9/15/04.

Grant support: Department of Defense (DAMD17-02-1-0683, BC011054), the Whitaker Foundation (RG-00-0085), the University of Illinois Research Initiative in Biotechnology, a NIH/National Institutes of Diabetes, Digestive and Kidney-funded predoctoral training program (T32 DK007739) in Signal Transduction and Cellular Endocrinology (N. Stitziel), and a Department of Defense-funded breast cancer predoctoral training grant (BC011054; B. Mar).

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Note: Supplementary data for this article can be found at Cancer Research Online at http://cancerres.aacrjournals.org.

Requests for reprints: Carol A. Westbrook, 650 Albany Street, Evans Biomedical Research Center 405, Boston, MA 02118. E-mail: cwcw@bu.edu.

(3 parts TRIzol to 1 part sucrose), and extracting was done according to manufacturer's instructions, followed by two additional salt precipitations [0.3 mol/L sodium acetate (pH 5.2) with 3 volumes of EtOH].

Real-time Quantitative Reverse-transcriptase PCR

First-Strand cDNA Synthesis. For generation of first-strand cDNA, ~1 μg of RNA was reverse-transcribed with Superscript II Reverse Transcriptase Kit (Invitrogen) in the presence of oligodeoxythymidylic acid (12–18) in a final 20-μL reaction volume with reverse transcriptase per manufacturer's recommended protocol followed by RNase H treatment.

Real-time Reverse Transcriptase PCR Setup. Real-time PCR reactions were done with DNase-free cDNA templates generated above and SYBR Green PCR Core Reagents (Applied Biosystems, Branchburg, NJ) following manufacturer's protocol with the following modifications: a 25 μL reaction was used, which contained 1× SYBR Green PCR mix, 3 mmol/L $MgCl_2$, 1× deoxynucleoside triphosphate blend (0.2 mmol/L of dATP, dCTP, dGTP, and 0.4 mmol/L of dUTP), 0.625 units of AmpliTaq Gold, 0.125 units of AmpErase UNG, 50 nmol/L each of forward and reverse primer, and 1 μL of cDNA template. Default PCR amplification cycles were used as specified by the ABI Prism 7700 Sequence Detection System (Applied Biosystems): 50°C for 2 minutes, 95°C for 10 minutes, 40 cycles of 15 seconds at 95°C, and 60°C for 1 minute. PCR amplification was followed by melting curve analysis with the following 3 hold cycles: 95°C for 15 seconds, 60°C for 20 seconds, and 95°C for 15 seconds, with the ramping time at maximum value 19:59 minutes set at the last hold cycle. PCR amplification analysis was done on Sequence Detector v.1.7a, and melting curves were analyzed on Dissociation Curves v.1 according to Applied Biosystems guidelines.

MCF-7 cDNA template was used to generate a relative standard curve for either the endogenous control or the target gene. MCF-7 cDNA was serially diluted at 1:10 dilution factors starting with the highest arbitrary concentration of 50 ng/μL (taken from one twentieth of a reverse transcriptase reaction of 1 μg of starting total RNA). The sample templates were diluted at 1:5. All samples were done in triplicate. The sequence of the primers used in real time reverse transcriptase-PCR is as follows: endogenous control 18S rRNA: F:5′-GTAACCCGTTGAACCCCATT-3′, R:5′-CCATCCAATCGGTAGTAGCG-3′ (6) with the expected size of 150 bp; and junctional adhesion molecule (JAM1): F:5′-CCCTCTTGGCTTGATTTTGC-3′, R:5′-TGACCTTGACTGATGGCTTC-3′ with the expected size of 115 bp. The glyceraldehyde-3-phosphate dehydrogenase (GAPDH) primers were obtained from Applied Biosystems (Foster City, CA). The quantity of target RNA (either JAM1 or GAPDH) was calculated by interpolating from the standard curve generated for that specific target. Both JAM1 and GAPDH quantities were normalized to 18S rRNA to calculate the relative quantity.

Microarray Hybridization and Data Processing

To minimize the effects of technical variability, membrane-associated and secreted and cytoplasmic RNA pools from one fractionation were processed in triplicate as follows. In three parallel reactions, 10 μg of total RNA from each pool was labeled, hybridized to the Affymetrix U133A microarray, processed, and scanned according to standard Affymetrix protocols. The six resulting CEL files (three membrane-associated and secreted and three cytoplasmic) were processed with the Bioconductor software suite (a set of libraries for R; ref. 7). The robust multiarray average algorithm (8–10) was used for normalization, background correction, and expression value calculation. Membrane-associated and secreted/cytoplasmic ratios for each microarray element were calculated by taking the ratios of the average membrane-associated and secreted and cytoplasmic expression values for that microarray element. Log_2 transformed data (as returned by the robust multiarray average algorithm) were used for the ratio calculation.

Membrane and Cytoplasmic Gene Reference Set

A reference set was developed containing genes that are known to have either membrane or cytoplasmic location and are represented on the Affymetrix U133A microarray with an automatic database search based on Swiss-Prot (11) release 44. Each Affymetrix microarray element has a unique Affymetrix Probe ID that can be mapped to at least one Swiss-Prot accession number.[4] Each Swiss-Prot entry was then searched for "cellular location" comment tags. Proteins were considered to have membrane-associated and secreted localization if the Swiss-Prot cellular location tag contained one of the following identifiers: "secreted," "Golgi," "vesicular," "membrane," "lysosome," or "peroxisome." Entries were considered tentative if they contained "probable," "possible," "potential," and "by similarity" and considered unambiguous if not. Entries that contained "nuclear," "nucleus," and "mitochondrial" were removed as there is some evidence that nuclear and mitochondrial proteins can be synthesized in either pathway (12, 13). This resulting list was then hand edited to remove entries containing multiple isoforms targeted for different subcellular compartments. Proteins are considered to have cytoplasmic localization if the Swiss-Prot cellular location tag contained "cytoplasmic" or "cytoplasm." Again, entries with probable, possible, potential, and by similarity were considered tentative, and entries containing nuclear, nucleus, and mitochondrial were removed. This list was hand edited to remove any entries with multiple isoforms as well as entries that contained any references to membrane association or organelles.

Statistical Calculations

At a given membrane-associated and secreted/cytoplasmic expression ratio r, the probability of belonging to the membrane-associated and secreted class for probe sets with ratios above r [$p(m|R > r)$] is calculated by using Bayes' rule as shown in Equation 1.

$$p(m|R > r) = \frac{p(R > r|m)\,P_m}{p(R > r|m)\,P_m + p(R > r|c)\,P_c} \quad (1)$$

where $p(R > r|a)$ is the proportion of class a above ratio r. We calculate this probability for the entire range of membrane-associated and secreted/cytoplasmic ratios at intervals of 0.01 and choose the ratio that corresponds to the maximum ratio (we choose the lowest ratio for which the posterior probability rises to within 10% of the maximum probability). The P_a factor corresponds to the prior probability of belonging to class a.

Because of a lack of previous data on which to base our prior probabilities, we estimate these prior probabilities by determining the contributions of the two known distributions to the distribution of probe sets with unknown localization as follows. We assume the distribution of membrane-associated and secreted/cytoplasmic ratios for the unknown set will be approximated by a linear combination of the membrane-associated and secreted/cytoplasmic ratio distributions for the known membrane-associated and secreted and known cytoplasmic distributions, as shown in Equation 2.

$$f_{unknown} = \alpha f_{MS} + \beta f_{CYT} \quad (2)$$

To estimate the contributions of the two known distributions, we find α and β that minimize the sum of the squared errors between these two quantities as shown in Equation 3.

$$\min_{\alpha,\beta} \sum [f_{unknown} - (\alpha f_{MS} + \beta f_{CYT})]^2 \quad (3)$$

For this calculation, we use discretized data (bins of width 0.01) and scale the original membrane-associated and secreted and cytoplasmic distributions to a maximum of one.

Sensitivity is defined as TP/(TP+FP), specificity is defined as TN/(FN+TN), and positive predictive value is defined as TP/(TP+FN), where TP (true positives) are the number of membrane-associated and secreted genes that are labeled correctly, FN (false negatives) are the number of membrane-associated and secreted genes that are labeled incorrectly, TN (true negatives) are the number of cytoplasmic genes that are labeled correctly, and FP (false positives) are the number of cytoplasmic genes that are incorrectly labeled. Sensitivity is a measure of the portion of membrane-associated and secreted genes we can detect, specificity is a measure of the portion of cytoplasmic genes we can detect, and positive predictive value is a measure of how many predicted membrane-associated and secreted genes are truly membrane-bound or secreted.

[4] Web address: www.affymetrix.com/analysis/.

RESULTS

RNA Fractionation and Verification. The fractionation of polysomes by sucrose density gradient centrifugation was first described by Mechler (4), and it was based on the observation that genes encoding proteins that are membrane-associated or secreted are translated by ribosomes bound to the endoplasmic reticulum (membrane bound polysomes), whereas genes encoding proteins that are cytosolic are translated by ribosomes free in the cytosol (free polysomes). Membrane-bound polysomes, being less dense, rise to the top of the gradient, whereas free polysomes remain near the bottom of the gradient.

Here, the method was used to separate intact polysomes prepared from the MCF-7 breast cancer cell line. In a typical fractionation, the separation results in two distinct peaks of $A_{260\ nm}$, with the lower peak representing free polysomes and the upper peak containing the less dense membrane-bound polysomes. To prepare sufficient RNA for Affymetrix hybridizations, it was necessary to fractionate polysomes from 3×10^8 cells; the results of this procedure are shown in Fig. 1. Fractions from each peak were pooled, and the fractions in the peak nearest the bottom (2 to 15) of the gradient are designated cytoplasmic, whereas fractions in the peak nearest the top of the gradient (20 though 26) are designated membrane and secreted. Sucrose fractions with near-baseline absorption (16 though 19) in between the cytoplasmic and membrane-associated and secreted pools were saved as negative controls. The peak at the surface of the gradient (the top 1.5 mL fraction) was discarded.

To confirm that the membrane-associated and secreted and cytoplasmic pools were enriched for membrane-associated and secreted-associated and cytoplasmic-associated mRNA, respectively, real-time quantitative reverse-transcriptase PCR was done with two primer pairs expected to amplify coding sequences specific for each population. We reverse transcribed 1 μg of total RNA each from the membrane-associated and secreted and cytoplasmic pools and labeled the resulting cDNA in three separate reactions for each pool. JAM1 is primarily cell-surface associated, whereas GAPDH is a protein found free in the cytoplasm. Because of their different biological sequestering, we expected JAM1 to be more highly represented in the membrane-bound polysome RNA, whereas the opposite will be true for GAPDH. To confirm the physical separation of these two RNAs, the membrane-associated and secreted/cytoplasmic expression ratio was calculated (see Table 1) by taking the ratios of the averages from the triplicates. As seen in Table 1, the membrane-associated and secreted/cytoplasmic expression ratio is about 1,000-fold greater for JAM1 than for GAPDH, demonstrating a marked enrichment in the membrane-associated and secreted pool for JAM1.

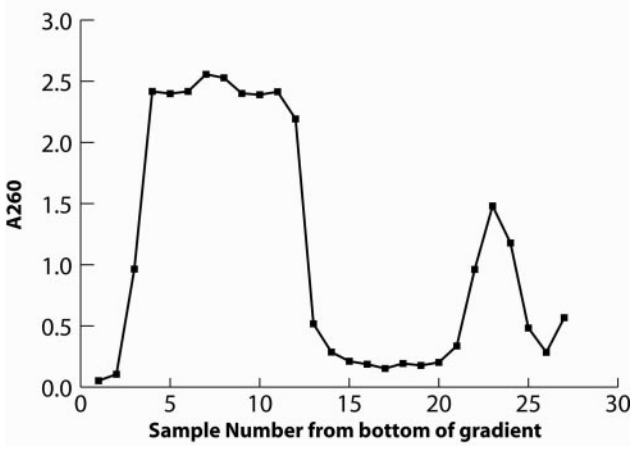

Fig. 1. RNA content of fractions taken from the sucrose gradient. *Vertical axis* shows the $A_{260\ nm}$, and the *horizontal axis* gives the fraction number from the bottom of the gradient.

Table 1 *Difference in RNA expression ratios for a membrane-associated and cytoplasmic gene*

Target	MS/CYT ratio, as measured by reverse transcriptase-PCR	MS/CYT ratio, as measured by Affymetrix U133A microarray
GAPDH	0.00064	0.986
JAM1	0.387	1.173

Abbreviation: MS/CYT, membrane-associated and secreted/cytoplasmic.

The RNA pools were then labeled and hybridized to Affymetrix U133A microarrays. Membrane-associated and secreted expression and cytoplasmic expression are the values returned by the robust multiarray average calculation of expression measured on the Affymetrix array hybridized to membrane-associated and secreted and cytoplasmic RNA, respectively, and were calculated for each microarray element by averaging the expression value across the appropriate triplicate (supplementary data). The membrane-associated and secreted/cytoplasmic ratio for GAPDH (AFFX-HUMGAPDH/M33197_M_at) and JAM1 (221664_s_at) was then calculated, as shown in Table 1. As expected, the membrane-associated and secreted pool shows an enrichment for JAM1 as compared with GAPDH, whereas the cytoplasmic pool shows an enrichment for GAPDH. All microarray data are available at the Gene Expression Omnibus (14) as accession number GSE1400.

Reference Set Construction. Because the distribution of membrane-associated and secreted/cytoplasmic ratios for either class is not known *a priori*, it was necessary to train a classifier with a reference set of genes with known subcellular localization. Of all 22,283 elements on the Affymetrix U133A array, subcellular location annotation, as described in Materials and Methods, was available for 9,851 elements. Unambiguous membrane-associated and secreted annotation was found for 3,188 of these, whereas unambiguous cytoplasmic annotation was found for 798 elements. These elements with unambiguous annotation represent the reference set.

Expression Threshold Calculation. It is likely that only a subset of the elements on the U133A microarray will be expressed in MCF-7 cells at a level great enough for meaningful measurement. To determine that level, we evaluated our ability to distinguish known membrane-associated and secreted genes from known cytosolic genes in the reference set at various total expression (E_T) levels, where $E_T \equiv$ membrane-associated and secreted expression + cytoplasmic expression, where membrane-associated and secreted and cytoplasmic expression are the average exponentiated expression values for the membrane-associated and secreted and cytoplasmic microarrays, respectively. A 10-fold cross validation was done at increasing threshold levels of E_T, including only training set members with an E_T value $\geq$ threshold. Briefly, for each E_T level, the data were randomly partitioned into 10 groups, 9 of which were used as a "training" set, and the remaining group was designated as a "testing" set. At each E_T level, the membrane-associated and secreted/cytoplasmic ratio threshold was calculated (as described in Materials and Methods) for the training set at that E_T level. The positive predictive value, sensitivity, and specificity were calculated by examining the performance of predicting the testing set for that E_T level, and averages over the 10 groups were recorded. The results of these calculations are shown in Fig. 2. As shown in Fig. 2*A*, the performance of prediction for E_T thresholds ranging from 22,283 (100% of the microarray elements) to 1,106 (4.9% of the microarray elements) was examined. The E_T level that corresponded to the highest sensitivity without a significant drop in positive predictive value or specificity was 738. At this level only 24.6% of probe sets with the highest E_T are included, resulting in a

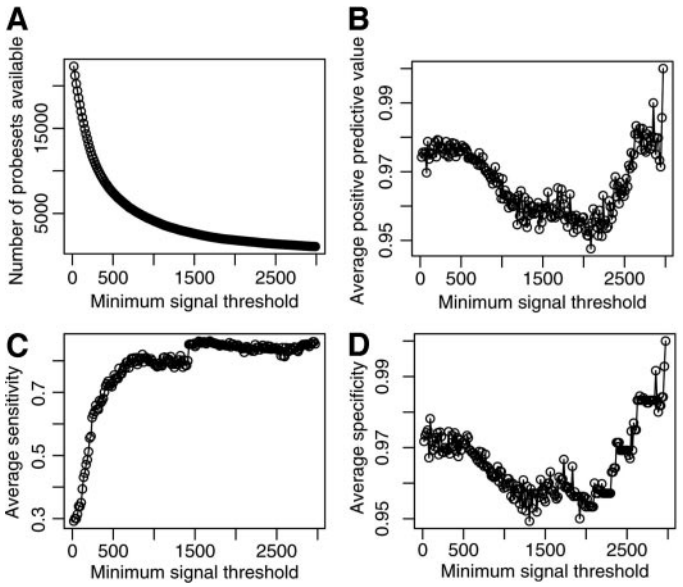

Fig. 2. Identifying the optimal E_T threshold for predicting membrane-associated and secreted genes. *A*. The number of probes with total expression level E_T above specific threshold values of E_T. *B*. The percentage of correctly labeled membrane-associated and secreted and cytoplasmic genes at differing E_T thresholds. *C*. The number of known membrane-associated and secreted genes that are correctly predicted at differing E_T thresholds. *D*. The number of known cytoplasmic genes that are correctly predicted at differing E_T thresholds. Averages of the 10-fold cross validation results are plotted in *B–D*.

final dataset of 5,483 probe sets that pass this threshold filtering. At this E_T level, our membrane-associated and secreted prior probability estimate is 7.2%, with a corresponding cytoplasmic prior probability of 92.8%. Of those probe sets above this E_T, 538 have unambiguous membrane-associated and secreted annotation and 305 have unambiguous cytoplasmic annotation. Additionally, at this level, our 10-fold cross-validation yields a 97.5% positive predictive value with 80.7% sensitivity and 96.9% specificity.

Membrane-Associated and Secreted/Cytoplasmic Ratio Threshold Calculation. All of the 843 probe sets in the reference set (with an E_T above the threshold of 723) were used to determine the membrane-associated and secreted/cytoplasmic ratio that corresponds to the maximum posterior probability of belonging to the membrane-associated and secreted class. The distribution of membrane-associated and secreted/cytoplasmic ratios for genes with known localizations was examined (Fig. 3). It is interesting to note that the cytoplasmic genes show a discrete peak, whereas the membrane-associated and secreted genes show a bimodal distribution with a smaller peak that associates with the cytoplasmic genes. The membrane-associated and secreted/cytoplasmic ratio of 1.08 was calculated as giving the maximum posterior probability. Note that above this level, the majority of known cytoplasmic genes are excluded (only 3.2% are above this level), and a sizeable fraction of the known membrane-associated and secreted genes (22%) show a lower membrane-associated and secreted/cytoplasmic ratio. Thus, genes with a ratio below 1.08 cannot be designated with certainty as either membrane-associated and secreted or cytoplasmic.

The distribution of membrane-associated and secreted/cytoplasmic ratios for the remaining probe sets (genes of unknown cellular localization) is plotted in Fig. 4. Of these, 755 probe sets fall above the expression threshold and above the membrane-associated and secreted/cytoplasmic ratio of 1.08. These 755 probe sets are labeled as "predicted membrane-associated and secreted." The remaining 3,885 probe sets found above the expression threshold and below the membrane-associated and secreted/cytoplasmic ratio of 1.08 are labeled as "indeterminate," because we expect a mixture of cytoplasmic and membrane-associated and secreted genes in this range of membrane-associated and secreted/cytoplasmic ratios. Of the predicted membrane-associated and secreted probe sets, 323 were found to have a tentative subcellular annotation but did not meet the criteria previously established for the reference set. The remaining 432 probe sets have no subcellular annotation. A similar percentage of indeterminate probe sets were found to have some tentative subcellular annotation (1516 of 3885).

The Swiss-Prot annotations were searched for terms that might

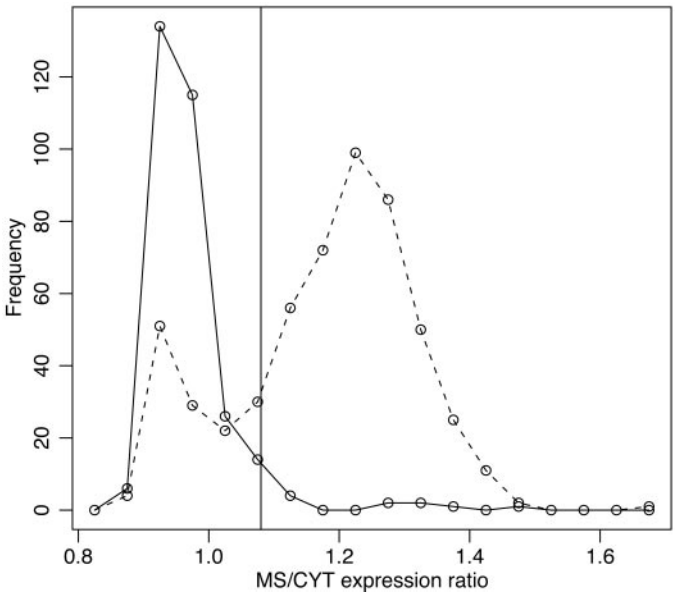

Fig. 3. Distribution of membrane-associated and secreted/cytoplasmic ratios for all of the genes in the reference set expressing above the E_T. The midpoints of bins from frequency histograms are plotted (for visual clarity, bins are 0.05 units wide). The *vertical line* indicates a membrane-associated and secreted/cytoplasmic ratio of 1.08. The distribution of membrane-associated and secreted genes is plotted with *dashed lines*, whereas the *solid line* indicates the distribution of cytoplasmic genes. (*MS/CYT*, membrane-associated and secreted/cytoplasmic)

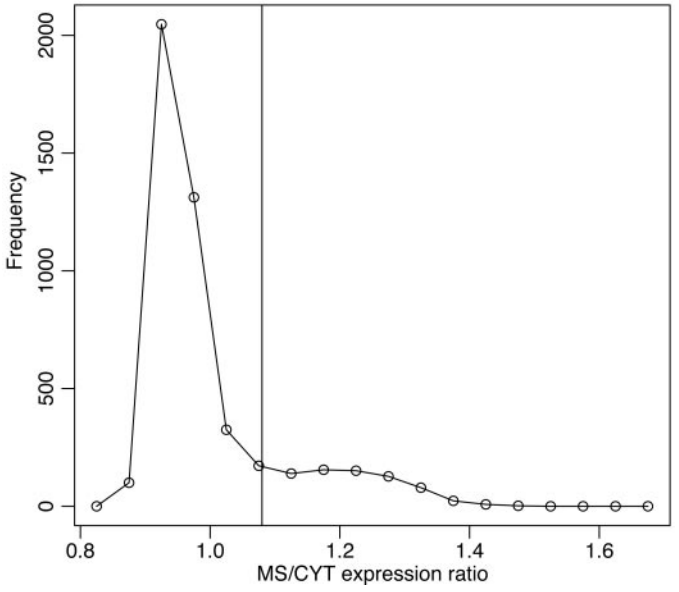

Fig. 4. Distribution of membrane-associated and secreted/cytoplasmic ratios for genes that are not in the reference set expressing above the E_T. The midpoints of bins from frequency histograms are plotted (for visual clarity, bins are 0.05 units wide). The *vertical line* indicates a membrane-associated and secreted/cytoplasmic ratio of 1.08. (*MS/CYT*, membrane-associated and secreted/cytoplasmic)

Table 2 *Tentative subcellular annotation for probe sets with predicted localization*

Predicted location	Total probe sets	Probe sets with tentative subcellular annotation	Tentative MS annotation	Tentative cytoplasmic annotation	Tentative nuclear or mitochondrial annotation	Conflicting annotation	Other
MS	755	323	214 (69.8%)	6 (1.4%)	56 (15.9%)	15 (3.6%)	32 (9.2%)
Indeterminate	3885	1516	113 (7.5%)	189 (12.5%)	961 (63.4%)	219 (14.4%)	34 (2.2%)

Abbreviation: MS, membrane-associated and secreted.

indicate a tentative assignment to a cellular fraction (*e.g.*, "membrane by similarity" or "nuclear"). Table 2 summarizes the tentative localization annotations for the predicted membrane-associated and secreted and indeterminate groups. Seventy percent (214 of 323) of the predicted membrane-associated and secreted probe sets with tentative annotations indicate a membrane-associated and secreted subcellular location. The binomial probability (with the prior probability of membrane-associated and secreted class membership as calculated in Statistical Calculations) of obtaining this number of membrane-associated and secreted probe sets by chance is very low ($P <<< 0.005$), indicating that the probe set population with membrane-associated and secreted/cytoplasmic ratios ≥ 1.08 is significantly enriched for membrane-associated and secreted genes. Less than 2% (6 of 323) of the predicted membrane-associated and secreted probe sets with tentative annotation are thought to be cytoplasmic. The remaining probe sets have either conflicting annotation or are thought to be localized to the nucleus, the endoplasmic reticulum, mitochondria, or other intracellular locations. Biochemical process annotation was available for 224 of these 323 probe sets in Gene Ontology (15). Over half of these seem to be involved in metabolism, whereas one third are involved in cell growth. Almost 25% of the predicted membrane-associated and secreted class are involved in cell communication. (Although these annotations seem to comprise a greater number than the actual number of annotated probe sets, Gene Ontology is organized in a way such that multiple annotations can correspond to a single probe set.)

In contrast, 7.5% (113 of 1516) of the indeterminate probe sets with tentative annotation indicate a membrane-associated and secreted localization. Although only 12.5% (189 of 1516) of these are thought to be cytoplasmic, a significant fraction of the probe sets with conflicting annotation indicate a possible cytoplasmic localization. Interestingly, >60% of the indeterminate probe sets contain nuclear or mitochondrial annotation.

Analysis of a Gene Expression Study for Membrane-Associated and Secreted Gene Content. The MCF-7 membrane-associated and secreted gene dataset was used to filter a differential gene expression study in breast cancer, which compared tumors with good *versus* poor 5-year outcome (5). We asked whether the membrane-associated and secreted localization provided by our study might give additional insight into the interpretation of the results and facilitate the selection of target genes for additional evaluation.

In the van't Veer *et al.* (5) study, RNA from 98 primary breast tumors was hybridized to cDNA microarrays, and the resultant analysis led to a 231-gene expression profile associated with poor prognosis. The original study was preformed on cDNA glass slide microarrays; therefore, we needed to find which elements of the Affymetrix U133A microarray corresponded to the 231 genes from the original study. It was possible to map 166 of these 231 genes to 269 probe sets on the Affymetrix microarray. Of these 269 probe sets, 20 were found in our predicted membrane-associated and secreted database representing 15 unique genes (see Table 3); an additional 52 were found in our training set of previously known membrane-associated and secreted genes. Of the genes not in the training set, almost half (7 of 15) had no subcellular location annotation in Gene Ontology or Swiss-Prot, although one had a published characterization. Of the 9 genes with functional annotation, 5 are involved in metabolism, along with one each involved in signal transduction, cell-cycle regulation, proteolysis, and calcium binding. It is interesting to note that of the genes without functional annotation, HCCR1 is a putative proto-oncogene, fucosyltransferase 8 is thought to contribute to malignancy, "G protein-coupled receptor 126" contains a "protein tyrosine phosphatase-like protein" domain, and "hypothetical protein FLJ22341" contains a rhomboid domain, thought to regulate epidermal growth factor receptor expression. Any of these proteins, the up-regulation of which is associated with poor prognosis in breast cancer, merit additional investigation as potential treatment targets.

DISCUSSION

We describe here a novel set of membrane-associated and secreted genes expressed in MCF-7 cells. We are able to annotate 755 probe sets as membrane-associated or secreted, 432 of which had no previous subcellular location annotation. Two levels of validation strengthen our location predictions. First, we did 10-fold cross validation on the set of genes with annotated localization, which is a robust method for estimating performance on future datasets with similar characteristics. On the basis of the results of the 10-fold cross validation, it is likely that a great number of the predicted membrane-associated and secreted genes will have membrane-associated and secreted localization. This is reflected by the average 97% positive predictive value observed in the 10-fold cross validation. Second, we examined the tentative annotations of genes in the set that were not used in the cross validation test and for which we predicted subcellular localization. Many of these have some tentative annotation, which we do not consider definitive. Nevertheless, our membrane-associated and secreted predictions coincide with these tentative annotations 70% of the time.

Here we describe a general method of applying density gradient fractionation of RNA to the Affymetrix platform, including a robust statistical analysis. Furthermore, we have described an approach that can easily be modified for other tissues or states for comparative studies.

To minimize technical variability, hybridization data were collected in triplicate, with three independent labeling experiments on RNA collected from one fractionation experiment. It was not possible to compare the results obtained from multiple fractionations of different cell cultures because of the prohibitive cost of processing these large volumes of cells and of Affymetrix hybridization. Thus, the results shown here represent a "snapshot" of a cell line at a single point in time; it is possible that the representation of some genes, and even their membrane-associated and secreted/cytoplasmic distribution, will vary with different culture conditions. Indeed, this approach might be used to investigate global changes in subcellular distribution of proteins under various biological conditions, which to our knowledge has not been addressed previously.

Our Bayesian analysis may be over- or underestimating membrane-associated and secreted localization because of some violations of the equation assumptions. The localization of different genes are not entirely independent observations. For instance, there are clearly genes that colocalize because of genetic interactions. In addition, we

Table 3 *Predicted membrane-associated and secreted genes in a breast cancer expression dataset (see text for details)*

Affymetrix ID	Original accession no.	Gene name	Description	Localization annotation (GO and Swiss-Prot)	MS/CYT ratio
212640_at	AF052159		*Homo sapiens* clone 24416 mRNA sequence	None	1.294
212248_at	AK000745		*Homo sapiens* cDNA FLJ20738 fis, clone HEP08257	None	1.261
212250_at	AK000745		FLJ20738 fis, clone HEP08257	None	1.232
212251_at	AK000745		*Homo sapiens* cDNA FLJ20738 fis, clone HEP08257	None	1.217
201818_at	AF052162	FLJ12443	Hypothetical protein FLJ12443	None	1.205
218686_s_at	Contig55188_RC	FLJ22341	Hypothetical protein FLJ22341	None	1.116
219202_at	Contig55188_RC	FLJ22341	Hypothetical protein FLJ22341	None	1.133
207170_s_at	NM_015416	HCCR1	Cervical cancer 1 Proto-oncogene	None	1.080
201037_at	D25328	PFKP	Phosphofructokinase, platelet	None	1.115
219197_s_at	NM_020974	CEGP1	CEGP1 protein protein disulfide isomerase related protein (calcium-binding protein, intestinal-related)	Not annotated, but literature suggests secreted protein	1.327
208658_at	NM_004911	ERP70	Protein disulfide isomerase related protein (calcium-binding protein, intestinal-related)	Endoplasmic reticulum	1.221
211048_s_at	NM_004911	ERP70		Endoplasmic reticulum	1.263
210074_at	NM_001333	CTSL2	Cathepsin L2	Lysosome	1.310
			Homo sapiens mRNA; cDNA DKFZp564D016 (from clone DKFZp564D016)	Membrane protein	1.212
212290_at	AL050021		*Homo sapiens* mRNA; cDNA DKFZp564D016 (from clone DKFZp564D016)	Membrane protein	1.223
212295_s_at	AL050021		Hypothetical protein		
213094_at	AL080079	DKFZP564D0462	DKFZp564D0462	Membrane protein	1.345
219410_at	NM_018004	FLJ10134	Hypothetical protein FLJ10134	Membrane protein	1.210
221675_s_at	NM_020244	LOC56994	Cholinephosphotransferase 1	Membrane protein	1.356
203988_s_at	NM_004480	FUT8	Fucosyltransferase 8 (alpha (1,6) fucosyltransferase)	Membrane protein (by similarity).	1.206
203362_s_at	NM_002358	MAD2L1	MAD2 (mitotic arrest deficient, yeast, homolog)-like 1	Nucleus	1.112

Abbreviations: GO, Gene Ontology; MS/CYT, membrane-associated and secreted/cytoplasmic.

make the assumption that these two classes are mutually exclusive, which may not be true for a small fraction of genes. The robust multiarray average algorithm might be a different source of underestimation for membrane-associated and secreted prediction, as it uses quantile normalization and might be overcorrecting for underrepresented membrane-associated and secreted genes. It is possible that alternative microarray processing algorithms may yield additional predicted membrane-associated and secreted genes. Despite these drawbacks, we believe this will be a useful tool for investigators wishing to filter existing or future breast cancer Affymetrix datasets to look for membrane-associated and secreted genes. Alternative statistical methods may be useful for additional analysis and confirmation of our results.

There are a significant number of genes with unambiguous membrane-associated and secreted annotation that fall below our membrane-associated and secreted/cytoplasmic threshold. It is unclear if this is because of a real biological process (some of those membrane-associated and secreted genes are not membrane-associated and secreted localized in MCF-7 cells, for instance) or a processing artifact. Additional experimental analysis is needed to elucidate the mechanism in action. Additional study is also needed to determine whether the protein localization we discovered for MCF-7 cells holds true when analyzing other breast cancer cells.

ACKNOWLEDGMENTS

The authors wish to acknowledge the helpful equipment support of Dr. Alexander Mankin from University of Illinois at Chicago Pharmaceutical Biotechnology and the technical assistance of Mr. Doron Galili.

REFERENCES

1. Nielsen H, Brunak S, von Heijne G. Machine learning approaches for the prediction of signal peptides and other protein sorting signals. Protein Eng 1999;12:3–9.
2. Emanuelsson O, Nielsen H, Brunak S, von Heijne G. Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. J Mol Biol 2000; 300:1005–16.
3. Diehn M, Eisen MB, Botstein D, Brown PO. Large-scale identification of secreted and membrane-associated gene products using DNA microarrays. Nat Genet 2000; 25:58–62.
4. Mechler BM. Isolation of messenger RNA from membrane-bound polysomes. Methods Enzymol 1987;152:241–8.
5. van't Veer LJ, Dai H, van de Vijver MJ, et al. Gene expression profiling predicts clinical outcome of breast cancer. Nature (Lond) 2002;415:530–6.
6. Schmittgen TD, Zakrajsek BA. Effect of experimental treatment on housekeeping gene expression: validation by real-time, quantitative RT-PCR. J Biochem Biophys Methods 2000;46:69–81.
7. Ihaka R, Gentleman RR. A language for data analysis and graphics. J Comput Graph Stat 1996;5:299–314.
8. Bolstad BM, Irizarry RA, Astrand M, Speed TP. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics (Oxford) 2003;19:185–93.
9. Irizarry RA, Bolstad BM, Collin F, et al. Summaries of Affymetrix GeneChip probe level data. Nucleic Acids Res 2003;31:e15.
10. Irizarry RA, Hobbs B, Collin F, et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics 2003;4:249–64.
11. Boeckmann B, Bairoch A, Apweiler R, et al. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res 2003;31:365–70.
12. Egea G, Izquierdo JM, Ricart J, San Martin C, Cuezva JM. mRNA encoding the beta-subunit of the mitochondrial F1-ATPase complex is a localized mRNA in rat hepatocytes. Biochem J 1997;322(Pt 2):557–65.
13. Lightowlers RN, Sang AE, Preiss T, Chrzanowska-Lightowlers ZM. Targeting proteins to mitochondria: is there a role for mRNA localization? Biochem Soc Trans 1996;24:527–31.
14. Edgar R, Domrachev M, Lash AE. Gene expression omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res 2002;30:207–10.
15. Harris MA, Clark J, Ireland A, et al. The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res 2004;32(Database issue):D258–61.