Global Gene Expression Profiling of Circulating Tumor Cells

Denis A. Smirnov,1,2 Daniel R. Zweitzig,3,4 Bradley W. Foulk, M. Craig Miller,1 Gerald V. Doyle,1 Kenneth J. Pienta,5 Neal J. Meropol,6 Louis M. Weiner,7 Steven J. Cohen,3 Jose G. Moreno,1 Mark C. Connelly, Leon W.M.M. Terstappen,1 and S. Mark O’Hara1

1Immunicon Corporation, Huntingdon Valley, Pennsylvania; 2Fels Institute for Cancer Research, Temple University School of Medicine; 3Department of Medical Oncology, Fox Chase Cancer Center; 4Department of Urology, Thomas Jefferson University Hospital, Philadelphia, Pennsylvania; and 5Internal Medicine, University of Michigan Medical School, Ann Arbor, Michigan

Abstract

Metastases from primary tumors are responsible for most cancer deaths. It has been shown that circulating tumor cells (CTCs) can be detected in the peripheral blood of patients with a variety of metastatic cancers and that the presence of these cells is associated with poor clinical outcomes. Characterization of CTCs in metastatic cancer patients could provide additional information to augment management of the disease. Here, we describe a novel approach for the identification of molecular markers to detect and characterize CTCs in peripheral blood. Using an integrated platform to immunomagnetically isolate and immunofluorescently detect CTCs, we obtained blood containing ≥100 CTCs from one metastatic colorectal, one metastatic prostate, and one metastatic breast cancer patient. Using the RNA extracted from the CTC-enriched portion of the sample and comparing it with the RNA extracted from the corresponding CTC-depleted portion, for the first time, global gene expression profiles from CTCs were generated and a list of cancerspecific, CTC-specific genes was obtained. Subsequently, samples immunomagnetically enriched for CTCs from 74 metastatic cancer patients and 50 normal donors were used to confirm by quantitative real-time reverse transcription-PCR CTC-specific expression of selected genes and to show that gene expression profiles for CTCs may be used to distinguish normal donors from advanced cancer patients as well as to differentiate among the three different metastatic cancers. Genes such as AGR2, S100A14, S100A16, FABP1, and others were found useful for detection of CTCs in peripheral blood of advanced cancer patients. (Cancer Res 2005; 65(12): 4993-7)

Introduction

Metastatic lesions and not the primary tumors are the leading cause of death in patients with carcinomas (1). During the process of metastasis, cancerous cells detach from the primary tumor and migrate to secondary organs to form a metastatic lesion. The presence of circulating tumor cells (CTCs) has been associated with poor prognosis in patients with metastatic breast cancer (2). Similar conclusions are likely true for other types of cancer. Detection of CTCs in peripheral blood by PCR using markers identified through gene expression profiling of primary tumors is often hampered by expression of tumor-associated transcripts in leukocytes that contribute most of the total RNA mass extracted from the sample even following substantial enrichment for CTCs (3, 4). Our objective here was to identify a set of genes that can be used for detection and characterization of CTCs in peripheral blood from patients with colorectal, prostate, and breast cancers. We generated global gene expression profiles for CTCs using a method that compares the RNA extracted from the CTC-enriched fraction of a peripheral blood sample with the RNA extracted from the corresponding CTC-depleted (leukocytes only) fraction. A novel set of genes was identified and tested for its ability to detect and characterize CTCs in blood samples from carcinoma patients. We show that CTCs from different primary cancers possess unique gene expression signatures.

Materials and Methods

Patient samples, immunomagnetic sample preparation, and circulating tumor cell enumeration. Blood from cancer patients and healthy volunteers was drawn into one 10-mL EDTA-containing Vacutainer Tubes (Becton Dickinson, Franklin Lakes, NJ) of which 7.5 mL were used for RNA extraction and one 10-mL CellSave Preservative Tube (Immunicon Corp., Huntingdon Valley, PA) of which 7.5 mL were used for CTC enumeration. Samples were kept at room temperature and processed within 36 hours after collection. The detailed information about all samples used in this study is presented in Supplementary Tables 1A and 2B. All participants signed an Institutional Review Board–approved informed consent form before study participation. Immunomagnetic enrichment of CTC for RNA isolation and enumeration has been described in detail previously (4–6).

RNA isolation. For gene expression studies following immunomagnetic enrichment, CTCs were lysed by adding 100 μL of Trizol reagent (Invitrogen, Carlsbad, CA). For all CTC samples, a CTC-depleted blood fraction was also saved by withdrawing 100 μL of whole blood recovered after the CTCs had been captured and removed. This fraction was then placed into a tube containing 900 μL of Trizol reagent. RNA from all samples was isolated using Trizol reagent according to manufacturer's instructions, DNase I treated, and Trizol repurified.

Target preparation, microarray hybridization, and microarray data analysis. Ten nanograms of total RNA from both the CTC-enriched and corresponding CTC-depleted fractions from each of the three patients with CTC counts of >100 CTCs/7.5 mL (one metastatic prostate, one metastatic breast, and one metastatic colorectal cancer patient) were used to prepare biotinylated hybridization targets with Affymetrix’s eukaryotic small sample target labeling assay, version II (http://www.wi.mit.edu/CMT/Protocols/AffySmSamplProto.pdf). Briefly, this protocol is designed to reproducibly amplify 10 to 100 ng of total RNA and is based on the principal of performing two cycles of double-stranded cDNA synthesis and in vitro transcription reactions using T7 RNA polymerase. Biotinylated target cRNA was then hybridized to an Affymetrix Focus array according to manufacturer's instructions and gene expression data was obtained using the Affymetrix Microarray Analysis Suite, version 5.0. A global scaling normalization procedure to normalize the expression data to the target value of 150 was done. This procedure uses a constant scaling factor for
every gene on an array, where the scaling factor is obtained from a trimmed average signal of the array after excluding the 2% of the probe sets with the highest and the lowest values. After normalization, the expression profiles were imported into a Microsoft Access 2000 database. Candidate genes for real-time reverse transcription-PCR (RT-PCR) verification studies were selected by comparing the corresponding depleted and enriched fractions for minimal expression in the CTC-depleted fraction and significant expression in the CTC-enriched fraction. We focused on genes for which the minimal expression in leukocytes (i.e., the CTC-depleted fraction) seen in the microarrays was corroborated by expression data published in the Cancer Gene Anatomy Project SAGE database (http://cgap.nci.nih.gov/SAGE/AnatomicViewer). More detailed information about the microarray experiments is available online at http://www.ebi.ac.uk/miameexpress/ (submission no. MIAMEEXPRESS#2137).

Multigene quantitative real-time reverse transcription-PCR analysis. The genes selected from the microarray analyses were evaluated in a separate set of metastatic cancer patients and a control group of healthy volunteers. To ensure that a sufficient amount of cDNA was available for multigene analysis, the RNA extracted from the CTC-enriched fraction of each blood sample was subjected to one round of amplification using the MessageAmp aRNA Kit (Ambion, Austin, TX) according to the manufacturer’s instructions. A total of 25 ng of the resulting aRNA was reverse-transcribed in the presence of 1 µL of a random 9-mer primer (50 ng/µL) to produce cDNA. The cDNA was diluted 30-fold with distilled water, and a volume of 10 µL of the cDNA samples was used in each RT-PCR reaction. Where possible, primer sequences that amplified a product of about 100 bp within 300 bases of the 3’ end of the transcript were selected (see Supplementary Table 2A). Quantitative real-time RT-PCR was done using the SYBR Green PCR Master Mix and an ABI Prism 7000 Sequence Detection System (Applied Biosystems, Foster City, CA). Gene expression levels were determined using a standard calibration curve prepared from gene-specific RT-PCR products with known concentrations. Gene expression levels between samples were normalized using the expression levels of the ribosomal protein RPS27A gene, which was shown recently to exhibit the least amount of variability in expression levels among different tissues (7).

Statistical analysis of gene expression data. Gene expression levels in the CTC-enriched fraction of the blood samples from the set of metastatic cancers and the normal donors were compared using the Kruskal-Wallis test to identify genes with significantly different expression levels (see Supplementary Table 2C). The estimation of the ability of the CTC-related gene expression data to discriminate between the various patient groups was done using a support vector machine (SVM)–supervised learning algorithm (8). Briefly, the SVM algorithm tries to find a hyperplane that provides optimal separation between the different classes of data so that there is maximal distance between the hyperplane and the nearest point of any of the classes. We used the SVM classification tool developed by National Cancer Institute of Spain as part of the Gene Expression Pattern Analysis Suit (http://gepas.bioinfo.cnb.es/tools.html). Detailed description of the SVM tool can be found at the following web site: http://tnasas.bioinfo.cnb.es/help/tnasas-help.html#methods. To provide a more realistic prediction model, the SVM program performs a 10-fold cross-validation process which results in the selection of the set of predictor genes that leads to the smallest error rate (9). A bootstrap analysis of the selected predictor genes using the SVM software was done with 15 randomly selected training (90% of the patient sample) and test (remaining 10% of the patient sample) sets. The selected set of predictor genes were used with each training set to generate a classification model, which was then applied to the respective test set for determination of its classification accuracy. The range, average, and SD of the classification accuracies were determined for the selected set of predictor genes using the results from the 15 different bootstrap models.

Results

Generation of global expression profiles for circulating tumor cells from patients with metastatic cancers. An integrated sample preparation system was used to enrich CTCs from 7.5 mL of blood using magnetic nanoparticles conjugated to...
monoclonal antibodies against epithelial cell adhesion molecule (EpCAM; refs. 4, 5). Despite a 10,000-fold enrichment, CTCs are still outnumbered by “nonspecifically” carried-over leukocytes. Initial experiments were done using blood added to tumor cell lines to estimate the number of CTCs required in a 7.5-mL blood sample to detect differentially expressed genes using the Affymetrix GeneChip system. It was determined that this could be done if ≥100 CTCs were present in a background of ~1,000 to 10,000 leukocytes (data not shown). Using these criteria, we selected a metastatic colorectal, prostate, and a breast cancer patient who each had high CTC counts (105, 647, and 3700 CTCs/7.5 mL, respectively; see Supplementary Table 1A for additional information).

To generate global gene expression profiles of CTCs from the selected cancer patients, RNA was extracted from the CTC-enriched fraction and was compared with the RNA from the corresponding CTC-depleted (leukocytes only) fraction of each patient’s blood sample by using the Affymetrix GeneChip platform. For this comparison each corresponding fraction was hybridized to a separate GeneChip (see Supplementary Table 1B). After a global scaling procedure was used to normalize the expression data between experiments, we selected sets of candidate marker genes that exhibited minimal expression in the CTC-depleted fraction and significant expression in the corresponding CTC-enriched fraction that were common to all three cancers, or specific to either the metastatic breast, prostate, or colorectal cancer patient (Fig. 1; also see Supplementary Table 1C). As EpCAM was the target for CTC enrichment, it was reassuring that two members of the EpCAM family (TACSTD1 and TACSTD2) were among the genes up-regulated in the CTC-enriched samples. In addition, keratin 19 (KRT19), a gene frequently used to identify CTCs of epithelial origin, was in the marker gene set common to the three cancer types. Among the genes with expression patterns specific for the metastatic breast cancer patient was the well-characterized marker mammaglobin 1 (MGB1/SCGB2A2). The prostate-specific antigen gene (PSA/KLK3) was specifically up-regulated in the metastatic prostate cancer patient whereas the carcinoembryonic antigen gene (CEA/CEACAM5) was specifically up-regulated in the metastatic colorectal cancer patient. These findings indicate the efficiency of the leukocyte background subtraction approach used to deduce CTC-specific gene expression profiles.

Verification of circulating tumor cell–specific expression of the candidate genes by quantitative real-time reverse transcription-PCR. It was crucial to confirm that the global expression profiles generated from CTCs enriched from peripheral blood of the three cancer patients are reflective of CTC gene expression signatures in a larger population of patients. Therefore, expression of the candidate genes selected from the microarray analyses were measured in the CTC-enriched blood fractions from 30 metastatic colorectal cancer patients, 31 metastatic prostate cancer patients, 13 metastatic breast cancer patients, and a control group of 50

Figure 2. Results of confirmatory gene expression profiling of candidates by quantitative real-time RT-PCR. Expression levels of the 25 candidate genes were calculated as log2 (transcript copy number + 1) and are presented in the form of a heatmap for the CTC-enriched blood fractions from 30 metastatic colorectal cancer patients, 31 metastatic prostate cancer patients, 13 metastatic breast cancer patients, and a control group of 50 apparently healthy normal donors. The number of CTCs detected per 7.5 mL of peripheral blood is also presented for each cancer patient. Based on previous studies, the blood from the control group of healthy donors was assumed to have no CTCs (2, 5). Genes that show an expression pattern specific to a particular sample type(s) are marked as pan cancer, breast cancer, colorectal cancer, and prostate cancer. Samples 1P-6P from the metastatic prostate cancer patients were from men with originally pathologically organ-confined disease or PSA recurrence only (stages A, B, and D1.5). Samples 7P-31P from the metastatic prostate cancer patients were from men with bone scan–positive or hormone-refractory disease (stages D2 and D3). For complete information, see Supplementary Table 2B.
apparently healthy normal donors using quantitative real-time RT-PCR (see Supplementary Table 2B). The range of CTCs per 7.5 mL of peripheral blood detected was 1 to 105 in the colorectal cancer samples (mean = 8, median = 3), 3 to 121 in the prostate cancer samples (mean = 21, median = 12), and 3 to 784 in the breast cancer samples (mean = 104, median = 4). Based on previous studies, the blood from the control group of healthy donors was assumed to have no CTCs (2, 5).

For quantitative real-time RT-PCR studies, 35 candidate genes were selected that represent known and novel markers genes common for CTCs from all cancer types as well as specific for CTCs from colorectal, prostate or breast cancer. Well-known CTC markers such as KRT19, PSA/KLK3, MGB1/SCGB2A2, and CEAC/CEACAM were carried through the verification process as indicators of the performance for the novel candidate CTC-specific markers. Two additional genes that were not represented on microarrays, keratin 20 (KRT20) and S100A16, were also tested. KRT20 is a known colorectal cancer marker. S100A16 is a newly discovered member of a large S100 family of Ca2+-binding proteins (10). It shares homology with S100A7, S100A13, and S100A14 genes that had been identified on the microarrays as potential CTC markers and has minimal expression in leukocytes.

Of the 35 candidate genes tested, 25 showed a statistically significant difference in expression among the four groups of samples tested (P < 0.01; see Supplementary Table 2C) and up-regulation in at least one of the cancer groups relative to the control group (Fig. 2; also see Supplementary Table 2B and C). The presence of transcripts for 9 of the 25 genes (TST, ASGR2, MARCO, TFF3, S100A13, MAOB, SLC2A10, and VIL1) that were up-regulated in the metastatic cancer patients was also detected in the majority of the normal donors (Fig. 2). Many of these nine genes are suspected to be involved in the processes of cellular proliferation, cell migration, and oncogenesis. The remaining 16 genes showed no significant expression in the majority of the normal donors and exhibited expression patterns associated with a particular cancer type (Fig. 2). The KRT19 and AGR2 (hAG-2) genes were expressed in the majority of the metastatic samples, regardless of the cancer type, whereas S100A14, S100A16, and CEACAM5 genes showed expression restricted to the metastatic colorectal and breast cancer samples. EABP1 and KRT20 genes showed expression patterns associated with colorectal cancer. The expression patterns of the KRT2, MSMB, DDC, AR, HPN, and KRT3 genes were associated with prostate cancer, whereas those of the SCGB2A1, SCGB2A2, and PIP genes were associated with breast cancer. Of interest was the fact that if a difference in gene expression was observed within the group of prostate cancer samples (Fig. 2). Samples 1P to 6P were obtained from patients with organ-confined disease or PSA recurrence only (stages A, B, and D1.5), whereas samples 7P to 31P originated from patients with bone scan–positive or hormone-refractory disease (stages D2 and D3; see Supplementary Table 2B). This observation suggests that gene expression signatures of CTCs may change as the disease progresses.

**Table 1. Results of classification of CTC samples based on their multigene expression profiling by quantitative real-time RT-PCR**

<table>
<thead>
<tr>
<th>Disease status</th>
<th>Accuracy ( n = 15 ), random training/testing sets</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Average ± SD (%)</td>
</tr>
<tr>
<td>A. Discrimination between cancer patient’s samples and samples obtained from “healthy” donors.</td>
<td></td>
</tr>
<tr>
<td>Metastatic cancers ( n = 74 )</td>
<td>83.2 ± 2.65</td>
</tr>
<tr>
<td>“Healthy” controls ( n = 50 )</td>
<td>80.1 ± 2.33</td>
</tr>
<tr>
<td>Overall</td>
<td>81.9 ± 1.21</td>
</tr>
<tr>
<td>Predictor genes: KRT19, AGR2, S100A13, ASGR2, and TST</td>
<td></td>
</tr>
<tr>
<td>B. Discrimination between patients with metastatic colorectal, prostate, and breast cancer and “healthy” donors.</td>
<td></td>
</tr>
<tr>
<td>Colorectal cancers ( n = 30 )</td>
<td>80.9 ± 5.23</td>
</tr>
<tr>
<td>Prostate cancers ( n = 31 )</td>
<td>71 ± 4.1</td>
</tr>
<tr>
<td>Breast cancers ( n = 13 )</td>
<td>62.1 ± 7.14</td>
</tr>
<tr>
<td>“Healthy” controls ( n = 50 )</td>
<td>88 ± 2.63</td>
</tr>
<tr>
<td>Overall</td>
<td>79.3 ± 1.71</td>
</tr>
<tr>
<td>Predictor genes: KL3, S100A14, MSMB, CEACAM5, S100A13, KL2, SCGB2A2, TST, S100A16, and KRT20</td>
<td></td>
</tr>
</tbody>
</table>

Discrimination between metastatic cancers and normals using real-time reverse transcription-PCR gene expression data. Using a SVM algorithm with a 10-fold cross-validation procedure (8, 9, 11), we evaluated the ability of gene expression profiling of the CTC-enriched fractions from peripheral blood samples to discriminate between metastatic cancers and normals as well as to discriminate between the different metastatic cancer types. A binary classification model combining the expression of the KRT19, AGR2, S100A13, ASGR2, and TST genes provided the best discriminative power between the combined metastatic cancer and control samples, with a mean overall classification accuracy of 81.9% (Table 1A). Receiver operating characteristic curves was also used to represent the accuracy of the system for the chosen five-gene predictor set that distinguishes cancer patients from “normal” donors. The area under the curve (AUC) for this model was 94.3% (see Supplementary Fig. 1). A categorical classification model combining the expressions of the S100A14, KL3, S100A13, CEACAM5, SCGB2A2, MSMB, KL2, S100A16, KRT20, and TST genes provided the best discrimination of the three metastatic cancers and control group. The mean overall classification accuracy for this model was 79.3% (Table 1B). For all of the models generated, similar average accuracies were estimated when the K nearest neighbor method was used for data classification (data not shown).

**Discussion**

Little is known about the molecular characteristics of CTCs that are detected in peripheral blood of cancer patients. To identify a
novel set of genes that can be used for detection and characterization of CTCs in peripheral blood, we generated the first global gene expression profiling of CTCs isolated from the blood of patients with metastatic colorectal, prostate, and breast cancer. To find candidate genes with expression specific to CTCs, we extracted RNA from the CTC-enriched and the corresponding CTC-depleted fractions from each patient’s blood sample and then generated separate gene expression profiles using the Affymetrix GeneChip platform. From generated profiles, candidate genes were selected that exhibited CTC-specific expression that was common to all three cancers or specific to either the metastatic breast, prostate, or colorectal cancer patient. CTC-specific expression of 35 candidate genes was then verified by quantitative real-time RT-PCR in a larger sample set using blood samples from 74 metastatic cancer patients containing varying numbers of CTCs and 50 healthy controls.

We identified novel CTC-associated genes such as AGR2, FABP1, S100A13, S100A14, S100A16, and others that can be used for CTC monitoring in peripheral blood. The role of these genes in cancer progression is not known. We found that CTCs from patients with different metastatic cancers possess unique gene expression signatures. Observed overall accuracy of our tissue of origin classification model calculated from the gene expression profiles of CTCs was 79.3%. This is on par with the tissue of origin classification accuracies of 70% to 80% calculated using the gene expression profiles of primary tumor samples (12, 13). We believe that global expression profiles of CTCs may provide insights that could improve our understanding of cancer and could lead to the development of both novel noninvasive diagnostic tools as well as novel therapeutic targets.

Acknowledgments


Grant support: Immunicon Corporation.

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.

References
