**Invasion-specific Genes in Malignancy: Serial Analysis of Gene Expression Comparisons of Primary and Passaged Cancers**

Byungwoo Ryu, Jessa Jones, Michael A. Hollingsworth, Ralph H. Hruban, and Scott E. Kern

*Departments of Oncology [B. R., J. J., S. E. K.] and Pathology [R. H. H., S. E. K.], The Johns Hopkins Medical Institutions, Baltimore, Maryland 21231, and The Eppley Institute, The University of Nebraska Medical Center, Omaha, Nebraska 68198 [M. A. H.]*

**Abstract**

The invasive growth of malignant cells induces an admixture of host reactions including desmoplasia, angiogenesis, and immune reactions. Pancreatic cancer has a prominent and characteristic host reaction at the site of primary invasion. To obtain new insights into the process of tumor invasion, we studied global patterns of gene expression using serial analysis of gene expression in pancreatic cancer, with extension to other tumor types. Here we report a cluster of invasion-specific genes in pancreatic and other cancers. This cluster contains genes that derive from distinct components of the host reaction, including some that may be useful as diagnostic markers and therapeutic targets.

**Introduction**

Invasion is central to the nature of malignancy. Distinguishable components of this process are known from histopathological studies (fibrosis, angiogenesis, and immune reactions) and from biochemical analyses that have revealed, for example, that elevated levels of proteases may mediate the migration of tumor cells through normal tissues. Each aspect of invasion could offer diagnostic or therapeutic targets; however, this information has not been fully used in the clinical setting. Moreover, the overall process has not been studied comprehensively. For example, *in vitro* studies of neoplastic cells in culture, which are commonly investigated, are inadequate experimental systems through which to address major features of tumor invasion.

Pancreatic cancer is an excellent example to illustrate these considerations. Its extremely high mortality and short survival period are largely attributable to a late clinical presentation, which generally occurs well after the establishment of the invasive focus. Furthermore, clinical options are limited after the development of advanced symptoms and subsequent diagnosis, and molecular targets for effective therapies remain unclear. There remains a major and unfortunate discrepancy: at a time when most patients come to medical attention, the clinical manifestations are usually well established, and yet, highly useful diagnostic laboratory markers that correlate with the invasive process remain lacking. This disparity is not because of the indolence of the invasive focus. Pancreatic cancer is remarkable for having a characteristic and exuberant host reaction, termed desmoplasia, at the site of primary invasion. This reaction can also be seen in other carcinomas, although less universally. These considerations suggest that the invasive process of pancreatic cancer should be carefully examined.

We studied global patterns of gene expression using SAGE (1) in pancreatic cancer (2, 3), with extension to other tumor types, to provide the following: (a) a comprehensive and unbiased study of the components of invasion; (b) new insights into the process of invasion; and (c) potential diagnostic markers and targets for imaging and therapy.

**Materials and Methods**

**SAGE Data.** For the intra-SAGE study of colorectal and pancreatic cancer, data from the initial report of gene expression profiles in normal and cancer cells by Zhang et al. (2) and Zhou et al. (3) were reanalyzed. The NlaIII restriction site used in library construction defined all SAGE tags. A total of 69,471 unique SAGE tags, which were identified among 300,706 tags sequenced from 6 colorectal and 4 pancreatic SAGE libraries, were used for the hierarchical clustering and PCA (4). These 10 SAGE libraries were generated from two each of normal colorectal mucosa, primary colorectal cancers, colorectal cancer cell lines (SW837 and Caco 2), primary pancreatic cancers, and pancreatic cancer cell lines (PL45 and AsPc1). For assessment of the contributions of acinar cells, a SAGE data set of normal pancreas generated previously by Zhang et al. (2) was used. For assessment of the contributions of ductal epithelium, short-term cultures of ductal cells from human pancreas processed for transplantation (5) were used to generate SAGE libraries as described (1); the library was sequenced, and the tags were extracted by the National Cancer Institute CGAP. Short-term culture BX is from a 32-year-old patient; H126 is from a 46-year-old patient. Extensive validation of the pancreatic cancer SAGE data used in this study was performed by Northern blot analysis as reported by Zhang et al. (2) and Zhou et al. (3). For the SAGE meta-analysis, SAGE data of ovarian, breast, and prostate were acquired from the CGAP database available in the NCBI SAGEmap database. New SAGE data of four additional pancreatic cell lines (CAPAN1, CAPAN2, Panc1, and Hs766T) were also obtained in collaboration with CGAP.

**Biocomputational Tools.** The Cluster and TreeView computer programs were obtained from the online resource and used for hierarchical clustering analysis, PCA, and visualization of clustering trees and PCA (4). For the intra-SAGE study of the data from colorectal and pancreatic tissues, exclusion was applied to tags where fewer than two samples contained at least 5 tags in the raw data, and where minimum and maximum values among all samples differed by <4 tags. This produced a dataset of 1,620 tags from an original 69,471 unique tags. For the SAGE meta-analysis, the tags were normalized to tags/million, and exclusion was applied where fewer than two samples contained at least 200 tags each and where minimum and maximum values among all samples differed by <100 tags. This data filtering produced a dataset of 1,213 tags from an original 239,303 unique tags. The filtered data were log-transformed, after conversion of values of 0–0.5, prior to analysis. An online data analysis tool, xProfiler, was used for analysis of SAGE data of ovary, breast, and prostate by the described algorithm (6). The names of genes...
and ESTs were identified from the tag sequences using an online resource from the NCBI.  

Results and Discussion

Hierarchical Clustering. Hierarchical clustering was used to examine the relationship among the 10 samples. Tree diagrams based on degrees of similarities between samples were constructed. Within a single SAGE experiment that included the 10 original colorectal and pancreatic samples (2), multiple means of clustering, including variations of the Spearman ranked, centered or uncentered, and absolute centered or uncentered correlations, produced similar trees. Cell lines clustered together rather than with the primary cancers or normal mucosa of the matched organ (Fig. 1A). This approach was then expanded to include additional new SAGE data. These additional data from SAGE libraries, constructed from different laboratories of multiple institutions, included SAGE analysis of tissues from additional organ sites (ovary, breast, and prostate; SAGEmap). SAGE data from an independent analysis of four additional pancreatic cancer cell lines were also available; their use avoided a reflection of intra-experiment clustering (a potential bias of meta-analysis) in this more stringent test. Similar to the results of the intra-SAGE dataset analysis, application of multiple methods of clustering produced similar trees, as seen in a representative tree diagram (Fig. 1B). In the case of comparisons of pancreatic cancers to ovarian cancers (ovary was the organ best represented in SAGE data among the invasion samples), the invasive tumors from pancreas and ovary clustered together rather than with their organ-matched cancer cell lines. In both the intra-SAGE and SAGE meta-analyses, the invasive tumors sometimes clustered with normal tissues but always would cluster separately from the cell lines. This result could be rationalized because the cell lines represent pure epithelial components, whereas the normal tissues and invasion sites comprise both epithelial and stromal elements. These patterns permitted an in-depth analysis of the gene expression patterns of the invasion site and suggested that common patterns might be seen across organ systems.

PCA. The multivariate statistical method of PCA is a useful tool for obtaining two-dimensional views of a multidimensional dataset. We performed PCA to delineate the gene cluster that distinguished the sample categories in the intra-SAGE dataset of colorectal and pancreatic samples (Fig. 2). PCA groups and orders genes based on variables that show colinear variation. Because the nature of SAGE precludes the attainment of negative values, the gene clusters were identified that had a similar relatedness of high frequencies of expression. Clusters of genes specific for, and highly expressed in, normal colorectal mucosa, pancreatic cancer itself, primary sites of pancreatic cancer, primary invasion sites of colon cancer, and shared among pancreatic and colorectal cancer cell lines were identified (Fig. 2). The largest cluster was the one most closely identified with expression within the primary invasion sites of pancreatic cancers. Therefore, we directed additional attention to this cluster (Table 1).

Genes Characteristic of Invasive Pancreatic Cancer. Table 1 shows the identities of gene transcripts and their frequency of appearance in the SAGE data of this cluster, with extension to other tumor types. Among 90 tags, 74 tags matched known transcripts, and 16 were poorly characterized tags that may represent novel genes. In instances where a SAGE tag was matched to more than one Unigene identifier, all were reported. None of the 32 most frequent tags from normal pancreas, defined as all tags of a frequency $>0.5\%$, were represented in this cluster. Thus, this cluster identified the “invasion-specific genes” as seen in comparison with organ-matched, purified, passed tumor cells. This terminology can be useful if five clarifications are kept in mind: (a) this cluster can be distinguished from other basic types of expression clusters, such as the tumor-specific genes (present in both the invasive site and in passed cancer cells but not in microdissected samples of matching epithelium of a normal organ) and the tissue-specific genes (referring to the tissue type of the carcinoma cells, genes found in common among the primary invasion site, passaged cancer cells, and in microdissected matching normal samples). Some of these other cluster types are exposed through the PCA as seen in Fig. 2; (b) as expected for the 74 known genes, they were also variably expressed in other organs, tumor cells, and conditions that were not part of the initial SAGE comparison (data from NCBI SAGEmap); (c) the “invasion-specific” quality of gene expression in an organ is a quantitative, and not necessarily qualitative, feature. It can represent the re-expression of a normally silent gene, the up-regulation of an expressed gene, or the increased representation of a given cell type within the invasive focus; (d) comparisons between the invasive site and the other forms of a given tumor type (i.e., cell lines) rely on data obtained within individual experiments and should reflect fairly reliable forms of comparison, e.g., the pancreatic and colorectal data derived from a single SAGE experiment wherein the samples were analyzed concurrently. Comparisons with other tumor types, however, involve data from SAGE libraries constructed in different laboratories and may involve some systematic biases that are not easily identified. Given these considerations, the similarity of findings among differing tumor types is nonetheless quite impressive and may confirm the utility of SAGE data for use in meta-analysis (7); and (e) expression of some genes, such as insulin, may represent the residual parenchymal cells of the invaded organ. Normal ductal epithelium can express genes that are not expressed in cultured cells of ductal carcinoma (Table 1); these genes can be used to mark the neighboring ductal structures invaded by the malignancy. Glucagon and pancreatic polypeptide, which had been excluded by the initial filtering, also followed the general expression pattern of this cluster (Table 1).

Invasion-specific Genes in Other Cancer Types. Significant numbers of expressed genes that are invasion-specific for pancreatic cancer are also invasion-specific in the colorectum; however, they have relatively higher gene expression in pancreatic cancer. This may be a simple reflection of the fact that pancreatic cancer, as suspected from the characteristic histological patterns that distinguish the two
tumor types, has a more prominent desmoplastic host reaction (Fig. 3). Comparison with nongastrointestinal tumor types, using the NCBI SAGEmap data, demonstrated that the invasive primary cancers of multiple organs often share overall patterns of gene expression. As seen in Fig. 1B, the invasion-specific patterns are distinct from the tumor-specific and tissue-specific patterns. Therefore, a wider examination of differentially expressed genes was conducted using the xProfiler program (6). Primary invasive cancers were compared with enriched (cultured or microdissected neoplastic and nonneoplastic) epithelial samples from prostate, breast, and ovary. Genes that were statistically likely to have a 2-fold greater invasion-specific expression pattern were identified, and their combined SAGE data are depicted in Table 1. This analysis shows that many of the 90 primary invasive cancer genes identified in the analysis of pancreatic colorectal cancers are also expressed in an invasion-specific pattern in other tumor types. PCA was also performed, using SAGE data from online data from all libraries containing at least 15,000 tags on October 15, 2000 (NCBI SAGEmap). This meta-analysis dataset included invasive cancers and samples not representing an invasion site (neoplastic and nonneoplastic cell lines and microdissected cancers) from prostate, colorectum, pancreas, breast, and ovary. In such large analyses, tissue-specific genes often predominated in the hierarchical clustering methods. Invasion-specific genes could best be identified by the method of PCA. A highly interesting gene of this cluster is prostaglandin D2 synthase (SAGEtag, ACGGAACAAT), which, although not highly expressed, was statistically associated with invasion samples rather than normal tissue or cell lines (median of 118 tags/100,000 across nine invasive samples versus 28 tags/100,000 for 22 comparison samples; \( P = < 0.001 \), Mann-Whitney rank-sum test). In this large PCA, invasion-specific genes were split among a number of categories (or clusters), according to whether they were generally shared or were tissue restricted (data not shown). It was thus often more useful to compare a limited number of organ sites directly. For example, comparisons that included ovarian and pancreatic cancer samples identified an invasion-specific cluster of over 50 genes (data not shown). This cluster closely resembled the pattern seen with the above analysis of colorectal and pancreatic cancers. In addition to this overlap, additional genes become evident that were obscured in the prior analysis (data not shown). Analysis expression of these additional genes is beyond the scope of this paper.

**Distinct Components of Invasive Primary Cancers.** The interpretation of gene expression patterns of the host-tumor interaction requires consideration of the known histological and molecular biological features of invasive cancers. Admixed areas of normal parenchyma at the advancing edge of pancreatic cancers will often include acinar units, duct structures, and neuroendocrine islets. At this advancing edge, there is often a pronounced atrophy of exocrine structures, leaving residual neuroendocrine islets as the most prominent parenchymal component. RNases are a major secretory product of pancreatic acinar cells. A relative paucity of acinar cells within the mass formed by the cancer itself thus is not only common, it facilitates the study of mRNA expression patterns within the core regions of the cancer. Even within this core, most of the cells in a primary pancreatic cancer are not neoplastic epithelial cells. A semiquantitative study found that even the most cellular regions of most (75%) primary pancreatic cancers have <50% neoplastic cells. The balance of the cells in a mass formed by a primary pancreatic cancer are from the host reaction of nonneoplastic fibroblasts, endothelial-lined vessels, and leukocytes including plasma cells (8). At the molecular biological level, studies of the host-tumor reaction have demonstrated that ligands secreted by the tumor cells may induce responses in the stroma, and vice versa. For example, human pancreatic cancers overexpress a number of important growth factor receptors and their ligands. These include the epidermal growth factor receptor and related receptors,

---

**Fig. 2.** Gene expression clustered by principal components. PCA produced these blocks of clustered genes, sections of which were labeled here after visual interpretation. The intra-SAGE study is reported with false colors, applied in a continuous scale according to the normalized tag counts; discrete levels are illustrated in the color legend. Top left, sample identities are aligned with respective columns using a visual guide.
Table 1  The pancreatic cancer invasion cluster: genes and frequencies among tissue types

<table>
<thead>
<tr>
<th>Gene</th>
<th>Known gene</th>
<th>Involved in SAGE data</th>
<th>XPosGen SAGE data</th>
</tr>
</thead>
<tbody>
<tr>
<td>A.1836</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Invasiomes</td>
<td>1836</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Pancreatic cancer</td>
<td>1836</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Invasion-specific</td>
<td>1836</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Genes</td>
<td>1836</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Invasive cancer</td>
<td>1836</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Invasion-specific</td>
<td>1836</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Genes</td>
<td>1836</td>
<td>3</td>
<td>3</td>
</tr>
</tbody>
</table>

Note: The table lists genes associated with pancreatic cancer invasion, their known gene names, and their involvement in SAGE data analysis. The XPosGen SAGE data column indicates the frequency of these genes in the SAGE dataset.

References:

License: Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (CC BY-NC-ND 4.0)
bias toward secreted genes that may reflect the transcriptional demands within tissues exhibiting extensive architectural remodeling. Microarray data to date have addressed a more limited scope of genes but can also be used to support the concept of major categories of genes within complex tumor samples, specifically segregating the stromal, epithelial, leukocytic, and proliferation-associated genes (10).

Matrix-remodeling Genes. A dense desmoplastic reaction is a prominent component of >90% of pancreatic carcinomas, many breast and ovarian cancers, and a variable component of many other cancer types (Fig. 3). Matrix remodeling has distinct components, such as protease activity (MMPs and complement proteases) and collagen genes (Table 1). Proteases such as the MMPs can be experimentally associated with enhancement of the invasive process (11). Specific collagens, such as collagen Iα1, are expressed at many-fold higher levels in invasive pancreatic cancer tissues than in pure populations of fibroblasts (Table 1 and data from NCBI SAGEmap), suggestive of a tumor-induced stimulus to collagen synthesis. In addition, fibroblasts express a less familiar complement of genes that can be seen within the invasion-specific cluster, such as CGI-101 and DKFPZ586B0621 protein (data from NCBI SAGEmap). It is of interest that MMP11, cartilage matrix protein, and DKFPZ586B0621 protein are more highly expressed in primary pancreatic cancers than in any other sample type examined by SAGE to date (data from NCBI SAGEmap; Table 1).

Angiogenesis. Pan-endothelial cell genes associated with angiogenesis were reported by St. Croix et al. (12). These genes include hevin, collagen types Vα1 and Vα2, PEM1 (similar to collagens), insulin-like growth factor binding protein-7 (angiomodulin), SPARC (osteonectin), and matrix Gla protein. The same genes were found in the invasion-specific cluster, but some also are found in precrisis fibroblasts (data of NCBI SAGEmap). Also reported by St. Croix were genes specifically elevated in tumor endothelium as compared with normal endothelium, including MMP2 and MMP11. Thy-1 cell surface antigen, and collagen types Iα1, Iα2, and SPARC, connective tissue growth factor, tubulins, and others (12) and data from NCBI SAGEmap. It can be argued that a component of these transcripts is of probable endothelial origin, although the individual genes (such as hevin, angiomodulin, and thrombospondin) are not endothelial specific (data of NCBI SAGEmap). Some of these are also expressed in other primary invasive cancers (Table 1).

Immune Response. The presence of mRNA for immunoglobulin heavy and light in the cluster reflects presence of plasma cells in the tissues. These are overrepresented within the SAGE data because of the focused transcriptional program of plasma cells. The Thy-1 gene can mark T lymphocytes, but here may largely reflect its high expression in endothelial cells (12). The immune reaction-associated genes of the cluster are also expressed in some other tumor types (Table 1).

Parenchymal Changes. The presence of few digestive enzymes in the cluster reflected residual atrophic acinar units in the pancreatic cancers. Insulin gene transcription was overrepresented in comparison with other islet cell products. This may reflect a preferential destruct-

---

* All numbers have been converted to tags per 100,000. Data for pancreas and colorectum are from Zhang et al. (2) or our new data. Other samples are from the SAGEmap site as of 7/20/2000. Complete colorectal and pancreas data are shown for all SAGE tags identified in the pancreatic cancer invasion cluster by PCA (see Fig. 2) and for two genes not defined within the cluster. Ovarian, breast, and prostate data are provided only where the gene was identified by SProfiler as being among the 100 genes most likely to represent at least a 2-fold difference between sample groups (this analysis minimizes the identification of spurious data variations). Prov. Assign. a provisional assignment of gene classification relative to invasive tumors is suggested for many genes based on known functions or distribution among gene expression databases: M, matrix remodeling; A, angiogenesis; I, immune reaction; P, parenchymal cells of the normal organ. CR, colorectal samples. Pa, pancreatic samples. Bt, breast samples. M1 and M2, two independent normal mucosa samples. H and I, independent sample of invasive primary cancer bulk tissue from colorectum and pancreas. C11 and C12, independent cancer cell lines of colorectum and pancreas. I, grouped data of invasive primary cancers of the respective organ sites (for ovary: OVT-6, OVT-7, and OVT-8; for breast, Duke 96-349, for prostate, Chen Tumor Pr.), D1 and D2, independent cultures of normal ductal epithelium from pancreas. C1, grouped data of cell lines and microdissected epithelial samples (for ovary: ESV-1, OVA343, OVA100-3, HOSE 4, OVP 5, A2780-9, ML10-10, and IOSE29-11; for breast: HMEC-B41, MDA-453, SKBR3, mammary epithelium, MCF7 control 3h, MCF7 control 0h, MCF7 estradiol 3h, and MCF7 estradiol 10h; for prostate: LNCaP, A, TSU, Chen LNCaP, Chen LNCaP no-DHT, CPDR LNCaP-C, and CPDR LNCaP-T). EST, a match to a Unigene cluster defined by expressed sequence tags.
tion of non-β islet cells by the cancer or an increased physiological stimulus for insulin secretion.

A number of invasion-associated genes currently remain unassigned to one of the four major component processes listed above. Many of these are novel and currently of limited functional assignment (Table 1). For example, the tag, tcccttctaa, appears remarkably specific for pancreatic invasion and as yet has no EST or gene assignment. Even for those genes of known function, we have at best some hints of their role in the invasion process. For example, the presence of apolipoproteins C1 and D, α2 macroglobulin, and low-density lipoprotein-related protein 1 (an α2 macroglobulin receptor) could support a prior proposal that fatty acid synthesis plays an essential role in malignant behavior (13). Heat shock protein 70 is abundantly expressed in malignant human tumors of various organs and complexes with tumor antigens, whereas in normal cells, its expression is mainly stress inducible (14). The presence of hexokinase in the cluster can be rationalized in that tumor cells are known to have high levels of glycolytic activity, with a corresponding increase in glycolytic enzymes, including hexokinase. This has been detected in resected lung, gastrointestinal, and breast cancer (15). This accelerated glucose metabolism in cancers is being applied to image various cancers, including those of the pancreas, by positron emission tomography using 18F-fluorodeoxyglucose (16).

Future Directions. The laboratory-based monitoring and/or clinical imaging of desmoplasia (such as the detection of propeptide fragments of collagen synthesis, matrix protein production, and proteases) could be used as a diagnostic and prognostic markers (11, 17). Genes that are expressed in normal tissues may still offer promise as diagnostic markers if they are not normally present in the circulation.

The extracellular matrix of invasive cancers offers multiple potential therapeutic targets. For examples, drugs that chelate the zinc atom of metalloproteases are now under development. The variety of complement proteases and other proteases within tumors might explain in part the failure of a potent inhibitor of metalloproteases in clinical trials (18). A comprehensive overview of proteases expressed at the host/tumor interface, such as that provided by SAGE, offers a sobering picture of the range of proteases that must be targeted for effective therapy. Complement proteases have received less attention to date but are potential targets of drug inhibition.

The apparent consistency of the gene categories identified here in turn supports the need for an intensified exploration of the fundamental nature of host-tumor interactions. We need to understand better the novel invasion-specific genes identified in this study (Table 1), including their cells of origin and their functions. A deeper exploration of SAGE data, through the construction of larger data sets, would allow the examination of the cell types that are underrepresented in the current transcriptional data set. An example is that lymphocyte-specific genes are not reflected here in high numbers, although considerable numbers of lymphocytes infiltrate the primary site of pancreatic and other cancers.

The parenchymal changes seen in cancers have received little attention in the study of tumors. Parenchymal changes should be studied systematically because it is likely that they will provide a diagnostic utility or enable better understanding of the physiological effect of tumors on the organ of origin. The preferential expression of insulin within pancreatic cancer, for example, may provide clues to an altered balance of growth factors at the invasion site that may affect the invasive process itself.

Acknowledgments

We thank Dr. Greg J. Riggins and CGAP for the sequencing of new SAGE libraries from pancreatic cancer cell lines.

References

Invasion-specific Genes in Malignancy: Serial Analysis of Gene Expression Comparisons of Primary and Passaged Cancers


*Cancer Res* 2001;61:1833-1838.

Updated version

Access the most recent version of this article at:
http://cancerres.aacrjournals.org/content/61/5/1833

Cited articles

This article cites 18 articles, 8 of which you can access for free at:
http://cancerres.aacrjournals.org/content/61/5/1833.full#ref-list-1

Citing articles

This article has been cited by 28 HighWire-hosted articles. Access the articles at:
http://cancerres.aacrjournals.org/content/61/5/1833.full#related-urls

E-mail alerts

Sign up to receive free email-alerts related to this article or journal.

Reprints and Subscriptions

To order reprints of this article or to subscribe to the journal, contact the AACR Publications Department at pubs@aacr.org.

Permissions

To request permission to re-use all or part of this article, use this link:
http://cancerres.aacrjournals.org/content/61/5/1833.
Click on "Request Permissions" which will take you to the Copyright Clearance Center's (CCC) Rightslink site.