Comparative Gene Expression Analysis of Ovarian Carcinoma and Normal Ovarian Epithelium by Serial Analysis of Gene Expression

David G. Peters,[1,2,3] Donna M. Kudla,[1] Julie A. DeLoia,[3,4] Tian Jiao Chu,[5] Liane Fairfull,[1] Robert P. Edwards,[3,4] and Robert E. Ferrell[1,3]

[1]Department of Pharmacology and Therapeutics, School of Medicine, University of Liverpool, Liverpool, United Kingdom; [2]Department of Human Genetics, Graduate School of Public Health, [3]University of Pittsburgh Cancer Institute, School of Medicine, and [4]Department of Obstetrics and Gynecology, Magee-Women's Research Institute, University of Pittsburgh, Pittsburgh, Pennsylvania; and [5]Institute for Human and Machine Cognition, University of West Florida, Pensacola, Florida

Abstract

Despite the poor prognosis of ovarian cancer and the importance of early diagnosis, there are no reliable noninvasive biomarkers for detection in the early stages of disease. Therefore, to identify novel ovarian cancer markers with potential utility in early-stage screening protocols, we have undertaken an unbiased and comprehensive analysis of gene expression in primary ovarian tumors and normal human ovarian surface epithelium (HOSE) using Serial Analysis of Gene Expression (SAGE). Specifically, we have generated SAGE libraries from three serous adenocarcinomas of the ovary and, using novel statistical tools, have compared these to SAGE data derived from two pools of normal HOSE. Significantly, in contrast to previous SAGE-based studies, our normal SAGE libraries are not derived from cultured cell lines. We have also compared our data with publicly available SAGE data obtained from primary tumors and "normal" HOSE-derived cell lines. We have thus identified several known and novel genes whose expressions are elevated in ovarian cancer. These include but are not limited to *CLDN3, WFDC2, FOLR1, COL18A1, CCND1*, and *FLJ12988*. Furthermore, we found marked differences in gene expression patterns in primary HOSE tissue compared with cultured HOSE. The use of HOSE tissue as a control for these experiments, along with hierarchical clustering analysis, identified several potentially novel biomarkers of ovarian cancer, including *TACC3, CD9, GNAI2, AHCY, CCT3*, and *HMGA1*. In summary, these data identify several genes whose elevated expressions have not been observed previously in ovarian cancer, confirm the validity of several existing markers, and provide a foundation for future studies in the understanding and management of this disease. (Cancer Epidemiol Biomarkers Prev 2005;14(7):1717–23)

Introduction

Ovarian cancer is the fourth leading cause of death from cancer among women and the most fatal among gynecologic tumors (1). High levels of mortality from ovarian cancer are primarily due to the lack of reliable methods for early detection. Consequently, the vast majority of invasive epithelial ovarian cancer remains undetected until stage III or IV, by which time the prognosis is very poor. At this late stage of diagnosis, the 5-year survival rate is <25% and >75% of women will ultimately die of their disease (2, 3). In contrast, patients diagnosed with stage I epithelial ovarian cancers have a 90% survival rate (4).

Several biomarkers for early detection of ovarian cancer have been evaluated, the best described of which is the product of the mucin 16 gene, *CA125*. *CA125* is detectable in the serum of 80% of women with ovarian tumors (5) and has been used for monitoring of patients during chemotherapy and for the detection of relapse. However, the utility of CA125 as an early screening tool is somewhat limited due to the fact that it is also elevated in various benign diseases, including endometriosis, ovarian cysts, uterine fibroids, and chronic liver disease (6), and has been reported to be elevated only in 60% of stage I tumors (7).

To expand our knowledge of the molecular pathology of ovarian carcinoma and identify potential novel markers of diagnosis and prognosis, we have undertaken a large-scale gene expression analysis of primary ovarian tumors and normal surface ovarian epithelium using novel statistical tools. We have also done comparative analysis of our own Serial Analysis of Gene Expression (SAGE) data with publicly available data derived from primary tumors and tumor cell lines.

Materials and Methods

Tissue Acquisition and Preparation. Tissue archived from patients who gave written informed consent was obtained through Magee-Women's Tissue Procurement Program. Samples were collected in the operating room, immediately snap frozen on dry ice, and then transferred directly to a liquid nitrogen cooled freezer (−130°C) for storage. Ovarian surface epithelial cells were scraped from normal ovaries directly into 1 mL TRIzol (Invitrogen, Carlsbad, CA), snap frozen on dry ice, and stored at −80°C. Table 1 shows the pathologic findings for each tumor. RNA was isolated from both tumors and normal tissue following the standard TRIzol protocol according to the manufacturer's instructions.

SAGE Library Synthesis. Both human ovarian surface epithelial (HOSE) libraries were derived from isolated RNA

Received 9/22/04; revised 4/7/05; accepted 4/14/05.

Grant support: National Cancer Institute Early Detection Research Network U01CA84968 (R.E. Ferrell and R.P. Edwards), NASA (D.G. Peters), and Scaife Family Foundation (J.A. DeLoia and R.P. Edwards).

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 are available at Cancer Epidemiology Biomarkers & Prevention Online (http://cebp.aacrjournals.org/).

Requests for reprints: David G. Peters, Department of Pharmacology and Therapeutics, University of Liverpool, The Sherrington Buildings, Ashton Street, Liverpool L69 3GE, United Kingdom. Phone: 44-151-794-5477. E-mail: david.peters@liverpool.ac.uk

Table 1. Pathologic findings for ovarian carcinomas

Tumor ID	Final diagnosis
OVCA 1102	Poorly differentiated adenocarcinoma of the ovary
OVCA 1214	Poorly differentiated papillary serous cystadenocarcinoma of the ovary
OVCA 1232	Moderately differentiated adenocarcinoma of the ovary

from several combined samples. HOSE1 consisted of a pool of 20 specimens (1 μg each) and HOSE2 consisted of a pool of 10 specimens (2 μg each). Tumor libraries were derived from single tumor samples. Total RNA (20 μg) was used to construct each SAGE library using the MicroSAGE protocol (8) with some minor modifications. In brief, double-stranded cDNA was synthesized from mRNA bound to oligo(dT) magnetic beads (Dynal Biotech, Lake Success, NY) using SuperScript II reverse transcriptase (Invitrogen). The cDNAs were cleaved with *Nla*III (anchoring enzyme) and the most 3′ terminal cDNA fragments were captured with magnetic beads and divided into two pools. Each pool was ligated to 5′ biotinylated linker A/B (8), containing recognition site for the tagging enzyme *Bsm*FI. After ligation, the beads were washed and the SAGE tags released from both pools by digestion with *Bsm*FI. Tags were blunted at their 3′ ends and combined to form the 104-bp ditags-linker products, which then were amplified by PCR. The amplified ditags-linkers were redigested with *Nla*III to remove the linkers and the ditags (26 bp) were isolated by gel electrophoresis and purified through Spin X tubes (VWR, West Chester, PA) and concatemerized by self-ligation. Concatemers with sizes between 500 and 2,500 bp were obtained by gel purification and cloned into the *Sph*I site of vector pZero (Invitrogen) and transformed into *Escherichia coli* strain DH10B (Invitrogen) by electroporation. For each library, ~1,200 colonies were random picked and plasmids with concatemer inserts were cycle sequenced with Big Dye terminator chemistry (Big Dye version 1, Applied Biosystems, Foster City, CA) and analyzed on a 3700 Applied Biosystems DNA sequencer.

SAGE Data Analysis. SAGE data were extracted using the SAGE 2000 software package (version 4.12; http://www.sagenet.org). The number of duplicate dimers for each library was <2% of the total tags for each library. A nonnormalized, side-by-side comparison was done with all five libraries in SAGE 2000 and these numbers were exported to Microsoft Access for further analysis. A query was run in Microsoft Access to link the UniGene identifier and gene description to each tag. The tag descriptions were downloaded from the National Institute for Biotechnology Information ftp server (ftp://ftpl.nci.nih.gov/pub/SAGE/HUMAN) and imported in Microsoft Access. The data were then exported to Microsoft Excel, where tag counts were normalized to counts per 30,000 tags and sorted based on average differences in expression between HOSE and tumor. Gene matches for significant tags were manually verified using both SAGEGenie (http://cgap.nci.nih.gov/SAGE/AnatomicViewer) and SAGEmap (http://www.ncbi.nlm.nih.gov/SAGE/).

In addition, the sequence files from four libraries on National Center for Biotechnology Information's public SAGE library database (http://www.ncbi.nlm.nih.gov/SAGE/) were downloaded. Table 2 shows the tissue source descriptions for each of the libraries. These sequence files were analyzed in the same manner as our own libraries.

Testing for Differentially Expressed Genes in SAGE Data. We have shown previously (9, 10) that the count of the tag corresponding to a gene (G) detected in library L, isolated from a tissue sample (S), follows a binomial distribution with variables (pS, N) where pS is the expression level of gene G in tissue S. Generally, the concentration level of a given gene is not the same for different tissues. In our analysis, we treat it as a random variable. Furthermore, for the sake of computational simplicity, we assume that, given that a tissue S is randomly picked from a population A, the concentration level pS of gene G in tissue S has a β distribution with variables (a_A, b_A). Consequently, if a library L of size N is generated from tissue S, then the count of the tags corresponding to gene G follows a β-binomial distribution with variables (a_A, b_A, N). For our analyses, we regarded our SAGE libraries as being from either a control (normal) population C or a target (cancer) population T and used the above-mentioned β-binomial distribution to model the tags corresponding to gene G. We estimated the variables (a_C, b_C) and (a_T, b_T) from the two sets of libraries and set the variable N to 30,000. A score, which is defined as the total variation distance between the two fitted models, was then assigned to gene G. A higher score indicates greater separation between the two fitted models and a greater difference in gene expression level of gene G between the two populations of samples (normal and cancer). This score can also be interpreted in the following way. Consider a randomly chosen sample tissue, which based on our prior information is equally likely to come from the control population or the target population. A SAGE library L of size 30,000 is generated from that tissue. If gene G is assigned a score of s, then based only on the count of the tags corresponding to gene G in library L we can correctly label the sample tissue with probability $(1 + s) / 2$.

In this study, for each tag, we compute the above-mentioned score for ovarian carcinoma versus normal HOSE, ovarian carcinoma (inclusive of publicly available tumor data) versus normal HOSE, and ovarian carcinoma (inclusive of publicly available tumor and HOSE cell line data) versus normal bulk HOSE tissue, respectively. To ensure a reasonable reliability, we only consider the tags with a minimum average concentration level of 100 per 1,000,000 tags. The tags with a score of at least 0.5 are reported.

Hierarchical Clustering Analysis. Differentially expressed tags (n = 192) identified by the methods described above were analyzed by hierarchical clustering with the GeneSpring package version 4.2 (Silicon Genetics, Redwood City, CA) using the Pearson correlation function. Tags were clustered by expression pattern and 12 major clusters were identified.

TaqMan Reverse Transcription-PCR. Total RNAs were purified by the RNeasy Mini Kit (Qiagen, Valencia, CA), cleared of residual genomic DNA by the DNA-free kit (Ambion, Austin, TX) according to the manufacturer's protocol, and quantified by spectrophotometry (Beckman DU 640). The optimal reverse transcription was carried out in 100 μL volumes as described (11) using two amounts of RNA template (100 and 400 ng). No reverse transcriptase controls were carried out with 400 ng total RNA. Quantitative PCR was done on this cDNA on

Table 2. Tissue type and pathologic descriptions for public libraries

Library	Sample type	Tissue description
HOSE 4	Cell line	Derived from ovary, normal surface epithelium
IOSE29_11	SV40	Transformed cell line derived from ovary, normal surface epithelium
OVT6	Bulk tissue	Primary ovarian tumor, serous adenocarcinoma
OVT7	Bulk tissue	Primary ovarian tumor, serous adenocarcinoma
OVT8	Bulk tissue	Primary ovarian tumor, serous adenocarcinoma

the ABI 7700 Sequence Detection Instrument (Applied Biosystems) using TaqMan MGB probes. PCR primers and probes for all genes analyzed were designed using the Primer Express software (Applied Biosystems). PCR amplification of cDNA was done in duplicate in 50 μL volumes as described (11) with the optimal primer and probe concentrations used for each gene (300 nmol/L for primer and 100 nmol/L for probe). Gene expressions were measured relative to the endogenous reference gene, human β-glucuronidase (*β-GUS*), using the comparative C_T method described previously (11). Standard *t* tests and the Wilcoxon two-sample rank sum test were used to generate *P*s reported in Table 3A and B, respectively.

Results

We sequenced a total of 84,810 tags from the three tumor libraries and 60,562 tags from the two HOSE libraries after excluding duplicate dimers. Analysis of these 145,372 tags generated 41,892 unique tags. Table 4 shows the tag analysis information for each of the five libraries. Duplicate dimers ranged from 0.40% to 1.78% of the total tags for each library, with an average of 1.1% duplicate dimers.

SAGE data were analyzed using novel statistical methods (9, 10), and genes whose expressions varied significantly between normal HOSE and the three tumor libraries (OVCA 1102, OVCA 1214, and OVCA 1232) were identified. The top 25 scoring overexpressed and underexpressed tumor tags identified by these analyses are listed in Table 5A and B, respectively. The full list of statistically significant tags is shown in Supplementary Table S2. Complete data sets can be downloaded at (http://www.phil.cmu.edu/projects/genegroup/papers.html).

Validation of Ovarian Tumor Markers in a Wider Sample Set. A subset of high-scoring differentially expressed genes that we considered to be putative ovarian tumor markers were selected for further analysis by real-time quantitative reverse transcription-PCR (qRT-PCR). The clinical characteristics of the specimens used for these analyses are included in Table 3A and B. qRT-PCR analysis was done for eight genes, including claudin 3 (*CLDN3*), WAP four-disulfide core domain 2 (*WFDC2*, also known as HE4), folate receptor 1 (*FOLR1*), collagen type XVIII α1 (*COL18A1*), carbamoyl-phosphate synthetase 2, aspartate transcarbamylase, and dihydroorotase (*CAD*), cyclin D1 (*CCND1*), *FLJ12988*, and

Table 3.

A. Initial qRT-PCR analysis of putative tumor markers identified by SAGE*

Samples	Stage/grade	Age	Histology type	*CLDN3*	*WFDC2*	*FOLR1*	*COL18A1*	*FLJ12988*	*CAD*	*FLJ22795*	*CCND1*
TP99-250	S2/G3	51	Endometrial/serous	332	109.11	685	9.46	2.1	7.64	6.77	20
TP99-265	S3/G3	57	Clear cell	126	6.29	156	3.96	6.8	1.14	5.92	1.5
TP99-445	S3/G3	69	Serous	426	1,916.28	570	1.37	5.8	2.65	6.45	4.5
TP00-331	S3/G3	71	Papillary serous	118	656.68	659	7.17	2.3	12.16	2.76	15
TP00-363	S3/G3	63	Papillary serous	312	815.6	221	1	11	2.72	3.08	2
TP00-423	S2/G2	43	Papillary serous	378	667.86	220	1.44	8.7	4.6	2.48	5
TP00-729	S3/G2	84	Clear cell	734	177.04	49.7	1.67	3.8	4.5	2.72	5
			Diagnosis								
TP01-104	Normal	47	Fibroids	25	1	1.41	1.61	2.2	2.1	2.15	1.2
TP01-322	Normal	45	Fibroids	1.2	2.98	1.15	1.89	4.5	1.74	1.74	1
TP01-364	Normal	49	Menorrhagia	5.5	26.71	7.86	4.84	19	5.29	3.29	1.6
TP01-400	Normal	39	Menorrhagia	1.1	4.25	4.6	1.71	1	1	1.05	1
TP01-417	Normal	47	Fibroids	1	3.63	1	1.37	7.4	2.83	1	1.8
P				0.0048	0.0473	0.0125	0.3918	0.7435	0.2018	0.0272	0.0801

B. Further qRT-PCR analysis of putative tumor markers identified by SAGE*

Samples	Stage/grade	Age	Histology type	*CLDN3*	*WFDC2*	*FOLR1*	*COL18A1*	*FLJ12988*
TP02-075	S1/G1	52	Metastasis from endometrial	884	6,398	58,117	6	6
TP02-657	S1/G1	49	Endometrial	2,110	7,452	5,336	5	37
TP02-222	S1/G1	33	Papillary serous, mucinous, endometrial	851	3,323	1,378	87	31
TP02-252	S1/G1	55	Endometrial	1,612	849	109	100	9
TP02-429	S/G1	53	Endometrial	473	374	288	100	11
TP02-480	S1/G1	34	Mucinous	5,746	26	172	17	24
TP03-186	S1/G2	53	Papillary serous	2,380	970	12,781	32	9
TP03-212	S1/G2	45	Endometrial	1,336	13	316	27	6
TP02-163	NA	53	Metastasis from gall bladder	7,512	10	1,183	15	169
TP02-724	S2/G2	46	Endometrial	4,738	2,288	16,365	4	31
TP02-203	S2/G3	84	Papillary serous	2,601	2,135	244,589	51	14
TP02-635	NA	45	Bladder metastasis	3	0.42	26	9	2
TP02-349	S3/G3	63	Papillary serous	692	284	2,180	8	3
TP02-628	S3/G2	53	Papillary serous	6,039	2,435	2,721	22	4
TP02-637	3C/G2	53	Papillary serous	10,587	1,172	9,508	22	6
TP02-794	3C/G3	52	Endometrial	3,115	348	2,656	18	3
TP02-500	S3/G2	39	Papillary serous	5,143	1,452	17,338	8	60
TP02-539	S3/G3	47	Papillary serous	89,732	3,611	386,918	197	753
TP02-545	S3/G3	75	Other	3,133	909	33,149	11	22
TP02-774	S4/G3	45	Papillary serous	11,049	1,289	37,554	40	25
TP02-559	LMP	73	Serous	269	26	862	21	22
TP03-137	LMP	43	Endometrial	8,345	1,215	7,231	47	50
TP03-062	Benign	75	Cystadenofibroma	3,692	165	4,182	25	26
TP03-424	Normal	76	—	1	4	6	1	2
TP03-661	Normal	44	—	1	4	1	2	1
TP03-665	Normal	39	—	4	1	3	3	2
P				0.004037	0.006408	0.000385	0.003165	0.003978

*Values are displayed as fold changes in expression relative to the sample (normal or tumor) with the lowest expression (= 1).

Table 4. SAGE library statistics

Library total tags	Total files	Sequenced	Duplicate dimers, *n* (%)
OVCA 1102	34,751	1,152	267 (0.77)
OVCA 1214	27,566	1,496	109 (0.40)
OVCA 1232	22,493	1,056	334 (1.48)
HOSE1	25,893	652	460 (1.78)
HOSE2	34,669	2,152	378 (1.09)

FLJ22795. Note that *FLJ12988* and *FLJ22795* were found to match the same SAGE tag (TGCTCTGAAT). We initially compared the expression of these genes in eight tumor samples and five normal HOSE specimens. These data are presented in Table 3A. Of the eight genes assayed by qRT-PCR, folate receptor 1 adult (*FOLR1A*; $P = 0.01252$), *WFDC2* ($P = 0.04735$), *FLJ22795* ($P = 0.02723$), and *CLDN3* ($P = 0.00486$) were significantly overexpressed in the ovarian carcinoma samples. *COL18A1, FLJ12988*, and *CAD* also gave promising results but did not reach statistical significance (Table 3A).

We then expanded our analyses to include a further 22 tumor samples and 3 normal HOSE specimens and again determined levels of gene expression by qRT-PCR. As shown in Table 3B, the genes, *FOLR1A* ($P = 0.000385$), *WFDC2* ($P = 0.006408$), *CLDN3* ($P = 0.004037$), *COL18A1* ($P = 0.003165$), and *FLJ12988* ($P = 0.003978$) were markedly and consistently overexpressed in all the tumor samples relative to normal controls, confirming their potential utility as markers of ovarian carcinoma. Overexpression of all of these genes was detectable in all tumor stages analyzed, including stage 1A, suggesting that overexpression of these genes may be useful for the detection of early-stage ovarian tumors. Furthermore, expressions of *FOLR1A, CLDN3*, and *WFDC2* by qRT-PCR in a metastatic bladder tumor (TP02-635) were equivalent to levels found in normal HOSE, suggesting that these markers may be tumor type specific. However, high expressions of all genes tested by qRT-PCR were observed in a metastatic gall bladder tumor (TP02-163). There was a trend toward greater expression in higher stages for *CLDN3* and *FLJ12988* and for more aggressive grade for *CLDN3, FOLR1*, and *FLJ12988*.

Comparison to Publicly Available SAGE Data. To take advantage of the fact that the SAGE technique generates immortal data that can be readily compared with other SAGE data sets generated in different laboratories (12), we directly compared our own results firstly with publicly available SAGE libraries generated from both bulk ovarian tumors (OVC14, OVT6, OVT7, and OVT8) and secondly with these plus two normal cell lines (IOSE29_11 and HOSE4) derived from HOSE. The results of these comparisons are shown in Supplementary Tables S3 and S4, respectively. We found that the genes identified by this approach were generally similar to those identified in our own data (Table 5A and B). Specifically, 16 (64%) of the genes listed in Table 5A were also found to be differentially expressed when the public tumor data were included in the analysis. These genes are marked with an asterisk in Table 5A. However, only 5 (20%) were retained in the top 25 high-scoring genes. These are rhophilin, Rho GTPase-binding protein 1 (*RHPN1*; CTGGAGGCTG), CD24 antigen (small cell lung carcinoma cluster 4 antigen; *CD24*; GGAACAAACA), *CLDN3* (CTCGCGCTGG), high mobility group AT-hook 1 (*HMGA1*; ATTTGTCCCA), and CD9 antigen (*CD9*; AAGATTGGTG). Similarly, when we also included publicly available SAGE data from two normal HOSE-derived cell lines, 13 (52%) of the top 25 genes identified in our own data remained overexpressed. These genes are marked with a hash (#) in Table 5A. However, only 4 (16%) of these were ranked in the top 25 high-scoring genes. These are *FOLR1A* (GTCGGGCCTC), *RHPN1* (CTGGAGGCTG), *CLDN3* (CTCGCGCTGG), and *FLJ20297* hypothetical protein (*FLJ20297*; TCCTTGCTTC). The reasonably good correlation between our own and publicly available data is corroborative evidence that the genes identified as overexpressed in ovarian carcinoma are generally robust.

Clustering Analysis of Differentially Expressed Genes. One clearly important requirement of a tumor biomarker is that its expression be easily detectable and highly specific for disease state. Therefore, the focus of the approaches described above was to identify genes whose overexpression correlates strongly with ovarian cancer. However, we also sought to gain insight into the biological features of the samples assayed by performing clustering analysis of differentially expressed genes. Our aim was to identify coexpressed genes that might reveal information about the biological basis of ovarian tumors and also reveal potential tumor markers that were missed by the analyses described thus far. Therefore, we subjected the differentially expressed tags identified when all of our own and the publicly available data were analyzed by hierarchical clustering analysis. We identified 12 distinct clusters of coexpressed genes that are shown in Supplementary Table S1.

There are some notable features of our data that are revealed by clustering analysis. For example, it is clear that tumors OVCA 1232 and OVT7 express high levels of genes associated with an immune response, suggesting infiltration of leukocytes in those tissue samples. These genes include immunoglobulin heavy constant γ3 (*IGHG3*), immunoglobulin heavy constant μ (*IGHM*; cluster 2), MHC class I A, B, and C (*HLA-A, HLA-B*, and *HLA-C*, respectively), immunoglobulin κ constant (*IGKC*), immunoglobulin λ joining 3 (*IGLJ3*), MHC class II DP α1 (*HLA-DPA1*), and MHC class II DP β1 (*HLA-DPB1*; cluster 5). Significantly, one of the putative tumor markers identified by our SAGE analysis (*WFDC2*) is co-expressed with these genes, suggesting the possibility that *WFDC2* is a marker of leukocyte infiltration. This observation reduces the potential of *WFDC2* as a useful tumor marker in peripheral blood.

We also found coexpression of genes encoding ribosomal proteins S3, S9, S13, S23, L5, L10, L17, L32, and X4 (*RPS3, RPS9, RPS13, RPS23, RPL5, RPL10, RPL17, RPL32*, and *RPSX4*, respectively) in cluster 9, reflecting moderately elevated expression of these genes in normal HOSE samples (HOSE2, HOSE4, and IOSE29_11) relative to tumor samples. Also of interest is the coexpression in cluster 8 of several structural genes of the extracellular matrix in cancer cells. These include collagen type I α1 (*COL1A1*), collagen type I α2 (*COL1A2*), collagen type I α3 (*COL3A1*), lumican (*LUM*), and biglycan (*BGN*). Cluster 8 also revealed coexpression of the calcium signal transducers tumor-associated calcium signal transducer 1 and 2 (*TACSTD1* and *TACSTD2*), which are widely expressed in human cancers (13).

Primary and Cultured HOSE Are Distinguishable by Comparison of SAGE Data. It is notable that the tumor suppressor gene junB proto-oncogene (*JUNB*) is highly expressed in the primary HOSE samples (HOSE1 and HOSE2) relative to all the tumor samples yet undetectable in the HOSE cell lines (HOSE4 and IOSE29_11). Coexpressed with *JUNB* is the negative regulator of cell cycle progression, cyclin-dependent kinase inhibitor 1A (*CDKN1A*). Similarly, the cell cycle regulator *CCND1* is overexpressed (cluster 3) in all the tumor samples analyzed by SAGE and most of those assessed by qRT-PCR (Table 3A and B) relative to normal HOSE, yet its expression levels were also found to be very high in the ''normal'' HOSE cell line IOSE29_11. Notably, *CCND1* is coexpressed in cluster 3 with *TACC3*, which is involved in driving cell cycle progression via a mechanism that involves interaction with the histone acetyltransferases (14, 15). Taken together, these observations suggest that the process of cell

Table 5.

Tag	L1102	L1214	L1232	HOSE1	HOSE2	Scores	Hs.*	Description
A. Tags whose expressions are elevated in ovarian carcinoma relative to normal HOSE †								
*#GTCGGGCCTC	71.65	39.18	22.67	1.16	0	1	73,769	*FOLR1A*
*#CTGGAGGCTG	9.5	8.71	10.67	0	0	1	149,152	*RHPN1*
GCAACTGTGA	7.77	8.71	6.67	0	0	1	169,476	*GAPDH* (glyceraldehyde-3-phosphate dehydrogenase)
*#ATTTGTCCCA	14.68	7.62	5.33	0	0	1	57,301	*HMGA1*
*#ATGACTCAAG	37.12	54.41	24.01	5.79	0.87	0.99	239,752	*NR2F6* (nuclear receptor subfamily 2, group F, member 6)
*#GGAACAAACA	8.63	3.26	18.67	0	0	0.99	75,108	*CD24*
*#TTTGTGTCAC	13.81	9.79	9.34	0	1.73	0.98	15,093	*CXXC5* (CXXC finger 5)
GGAGCACACA	8.63	2.18	4	0	0	0.98	193,490	*FLJ31952* (hypothetical protein FLJ31952) /*FLJ34922* (hypothetical protein FLJ34922)
*#CTCGCGCTGG	50.07	8.71	32.01	1.16	0	0.98	25,640	*CLDN3*
*TGCTGAATCA	14.68	8.71	18.67	2.32	1.73	0.98	327,068	*CCDC6* (coiled-coil domain containing 6)
*#CTTGAGCAAT	13.81	10.88	16	2.32	1.73	0.98	848	*FKBP4* (FK506-binding protein 4, 59 kDa)
TTAAAGGCCG	2.59	9.79	2.67	0	0	0.97	79,086	*MRPL3* (mitochondrial ribosomal protein L3)
*#TAATCCTCAA	25.04	39.18	13.34	0	3.46	0.96	78,409	*COL18A1*
AGGGGATTCC	9.5	3.26	1.33	0	0	0.95	75,412	*ARMET* (arginine-rich, mutated in early-stage tumors)
*#TGGAACTGTA	8.63	9.79	5.33	1.16	0.87	0.94	132,262	*C10orf4* (chromosome 10 open reading frame 4)
ATGTAGTAGT	15.54	13.06	16	2.32	5.19	0.94	406,404	*HNRPD* [heterogeneous nuclear ribonucleoprotein D (AU-rich element RNA-binding protein 1, 37 kDa)]
GGGGTGGGGC	5.18	8.71	5.33	0	0.87	0.94	154,868	*CAD*
*#AAAGTCTAGA	8.63	6.53	2.67	1.16	0	0.94	82,932	*CCND1* [*PRAD1* (parathyroid adenomatosis 1)]
#TTGATGTACA	10.36	15.24	9.34	3.48	1.73	0.93	433,581	*SFRS11* (splicing factor, arginine/serine-rich 11)
*CCCCAGTTGC	29.35	26.12	30.68	16.22	9.52	0.93	74,451	*CAPNS1* (calpain, small subunit 1)
*#TCCTTGCTTC	10.36	22.85	12	1.16	4.33	0.93	94,491	*FLJ20297*
TAGGCCCAAG	12.09	8.71	9.34	2.32	1.73	0.93	78,880	*ILVBL* [ilvB (bacterial acetolactate synthase)-like]
GAGAAATATC	1.73	11.97	2.67	0	0	0.92	169,984	*ZFN638* (zinc finger protein 638)
*#ATCGCTTTCT	14.68	19.59	14.67	3.48	6.92	0.91	177,486	*APP* [amyloid β (A4) precursor protein (protease nexin-II, Alzheimer disease)]
*#AAGATTGGTG	7.77	6.53	12	2.32	0	0.91	1,244	*CD9* (p24)
B. Tags whose expressions are lower in ovarian carcinoma relative to normal HOSE ‡								
CCCAACGCGC	0	0	0	83.42	47.59	1	347,939	*HBA2* (hemoglobin, α2)
GCAAGAAAGT	0	0	0	26.65	39.81	1	155,376	*HBB* (hemoglobin, β)
CTTCTTGCCC	0	0	1.33	47.5	36.34	1	347,939	*HBA2* (hemoglobin, α2)
ACACAGCAAG	0	0	0	23.17	15.58	1	—	—
CCCTTGTCCG	0.86	0	0	26.65	20.77	1	127,824	na LOC349752
TCTCCATACC	0.86	1.09	0	23.17	25.09	1	—	—
ACCCACGTCA	0.86	0	1.33	27.81	20.77	1	400,124	*JUNB*
AGCTTCCACC	0	0	0	11.59	7.79	1	490,252	Transcribed locus, strongly similar to XP_530188.1 LOC457315
CTGTACTTGT	0.86	0	0	63.72	29.42	1	75,678	*FOSB* (FBJ murine osteosarcoma viral oncogene homologue B)
GAGTGGCTAC	0	0	0	9.27	6.92	1	—	—
TTGGTGAAGG	10.36	18.5	17.34	61.41	49.32	1	426,138	*TMSB4X* (thymosin, β4, X chromosome)
ATGGTGGGGG	0	0	0	8.11	22.5	1	343,586	*ZFP36* [zinc finger protein 36, C3H type, homologue (mouse)]
AGATCCCAAG	0	0	0	5.79	8.65	1	50,813	*ITLN1* [intelectin 1 (galactofuranose binding)]
TGGAAGGAGG	0	0	0	8.11	6.06	1	—	—
TAGCCGGGAC	0	0	0	5.79	7.79	1	107,740	*KLF2* [Kruppel-like factor 2 (lung)]
TGTGGATGTG	0	0	0	4.63	12.11	1	180,878	*LPL* (lipoprotein lipase)
GGGTAGGGGG	0	0	0	34.76	9.52	1	13,323	*FOSB* [FBJ murine osteosarcoma viral oncogene homologue B (internal tag)]
AGGGCTTCCA	56.98	37	69.35	134.4	141.05	1	458,148	*RPL10* (ribosomal protein L10)
ATTCTCCAGT	35.39	39.18	44.01	85.74	89.13	1	406,300	*RPL23* (ribosomal protein L23)
CTGCTATACG	11.22	14.15	9.34	41.71	38.94	1	180,946	*RPL5* (ribosomal protein L5)
GCCGTGTCCG	21.58	9.79	8	54.45	58.84	1	380,843	*RPS6* (ribosomal protein S6)
GAGGGAGTTT	117.41	138.21	110.7	200.44	182.58	0.99	76,064	*RPL27A* (ribosomal protein L27a)
TAGTTGGAAC	0	0	1.33	6.95	13.85	0.99	1,119	*NR4A1* (nuclear receptor subfamily 4, group A, member 1)
GGGCAGGCGT	1.73	2.18	0	18.54	11.25	0.99	501,629	*IER2* (immediate-early response 2)
GGCCCCTCAC	0	1.09	1.33	31.28	11.25	0.99	274,313	*IGFBP6* (insulin-like growth factor–binding protein 6)

NOTE: The tag CTGGAGGCTG matching *RHPN1* is listed as an internal tag match in SAGEGenie (http://cgap.nci.nih.gov/SAGE/AnatomicViewer).
*Hs. corresponds to UniGene no. Tag counts are expressed as tags per 30,000.
†Tags also identified as being statistically significant when publicly available tumor and HOSE data were included are identified by * and #, respectively (see text).
‡Blank cells indicate no database match for that SAGE tag. Tag counts are expressed as tags per 30,000.

culture is associated with alterations in cell cycle regulation in the normal HOSE cell lines.

These analyses also identify several potential novel ovarian tumor markers in our data. For example, coexpressed with *CCND1* are *CD9*, lysophospholipase II (*LYPLA2*), and G protein α−inhibiting activity polypeptide 2 (*GNAI2*). *CD9* is involved in cell proliferation (16). Its overexpression has not been previously associated with ovarian carcinoma, although it has been described as a possible marker for gastric cancer (17). Notably, *CD9* underexpression has been associated with ovarian tumor progression (18). To our knowledge, neither *LYPLA2* nor *GNAI2* overexpression have been previously associated with ovarian cancer. Therefore, these genes, along with *TACC3*, may be novel ovarian tumor markers.

We also found strong coexpression in cluster 4 of genes associated with response to cellular stress. These are glutathione peroxidase 1 (*GPX1*), chaperonin containing TCP1, subunit 3 (*CCT3*), and 27-kDa heat shock protein 1 (*HSPB1*). Coexpressed with these genes is the gene encoding *S*-adenosylhomocysteine hydrolase (*AHCY*). These genes are overexpressed in ovarian tumors relative to primary normal HOSE (HOSE1 and HOSE2) but not relative to cultured HOSE (HOSE 4 and IOSE29_11) in which displayed levels of expression of these genes that were comparable with the primary tumors. In cluster 4, we also found the *HMGA1* gene, the overexpression of which has been previously associated with ovarian carcinoma (19). The biological significance of these observations is unclear.

Discussion

The objective of this study was to identify potential markers of ovarian carcinoma and provide a snapshot of the molecular pathology of this disease at the level of the transcriptome. We used SAGE to analyze gene expression in three ovarian serous adenocarcinomas and two pools of normal HOSE. Furthermore, we compared our own SAGE data with publicly available similar data sets for ovarian cancers and epithelial cell lines cultured from normal HOSE. To perform these comparisons, we used novel statistical tools designed specifically for these purposes (9, 10).

We identified several potential biomarkers of ovarian cancer, five of which (*FOLR1A, WFDC2, CLDN3, COL18A3*, and *FLJ12988*) were further analyzed and their expression changes were confirmed by qRT-PCR in a larger sample set. High levels of expression of three of these markers (*FOLR1A, WFDC2*, and *CLDN3*) have previously been associated with ovarian tumors. In particular, the role of *FOLR1A* has been extensively studied in the context of ovarian cancer. *FOLR1A* expression has been reported at moderate levels in the normal epithelia of kidney, lung, and breast and high levels in placental tissue (20). However, its expression is absent in normal ovarian epithelium (21) and elevated in the majority of nonmucinous ovarian carcinomas (22). *CLDN3* and *WFDC2* have also been associated with elevated expression in ovarian cancer. For example, Hough et al. (23) reported overexpression of both *CLDN3* and *WFDC2* by SAGE analysis in ovarian tumors. Similarly, microarray approaches were used to identify *WFDC2* overexpression in ovarian tumors (24, 25).

A COOH-terminal fragment of the *COL18A1* gene product corresponds to the antiangiogenic factor endostatin and overexpression of endostatin has been correlated with ovarian cancer (26). Because of the central involvement of endostatin in angiogenesis and its role in tumor growth (27), *COL18A1* overexpression is a promising biomarker for ovarian cancer. However, in a previous study, no correlation was observed between serum levels of endostatin and incidence of ovarian cancer (28).

Previous searches for ovarian tumor markers by SAGE have only considered normal samples that have been cultured *ex vivo* (HOSE4) or are SV40 transformed (IOSE29_11; refs. 23, 29). As noted above, our analysis of primary ovarian epithelium samples (HOSE1 and HOSE2) revealed altered expression of several genes not reported by previous SAGE studies. These are most evident in clusters 3 and 4 and include *TACC3, CD9, CCND1, LYPLA2, GNAI2, GPX1, AHCY, CCT3, HSPB1*, and *HMGA1*. Corroborative evidence that some of these genes are indeed potentially useful biomarkers for ovarian cancer is derived from the fact that overexpression of a subset of these genes, *HMGA1* (19), *CCND1* (30), *GPX1* (23), and *HSPB1* (31), have all been associated with ovarian cancer.

Interestingly, although *TACC3* has not, to our knowledge, been associated with ovarian carcinoma, it is highly expressed during oogenesis (32). *CD9* is associated with reduced tumor progression but is not a biomarker for OVCA (18). *LYPLA2* was also overexpressed in ovarian carcinoma. High levels of lysophosphatidic acid, a product of lysophospholipase catalytic activity, have been reported as a potential biomarker of ovarian cancer (33). However, lysophospholipase activity levels in serum do not seem to be associated with ovarian carcinoma (34). To our knowledge, *GNAI2, GPX1*, and *CCT3* are not known to be overexpressed in ovarian cancer and may be entirely novel markers for this disease.

The fact that our analysis of primary HOSE tissue leads to the identification of potentially novel tumor markers underlies the importance of avoiding cultured cells as normal controls for biomarker discovery. Our data suggest the activation of gene expression cascades in cultured HOSE that are involved in cell proliferation. Clearly, this is an undesirable control phenotype when performing biomarker screens in cancer. Therefore, comparison of gene expression patterns in cultured cells with those obtained from bulk tissue must be treated with caution. It should also be noted however that the collection of primary HOSE tissue might result in the sampling of contaminating stromal cells.

Clearly, our study has several limitations. One drawback is the use of bulk tumor samples for our analysis. As we have shown, these samples may contain multiple cell types whose distinct transcriptomic signature can create problems at the data analysis stage. One way to overcome this would be the use of technologies for analyzing gene expression in very small samples of laser-captured tissue of interest (35).

One disadvantage of using SAGE for gene expression analysis is that sample throughput is low due to the fact that the procedure is highly labor intensive. Furthermore, despite our efforts to comprehensively identify differentially expressed genes using novel statistical tools, it may be that we have missed important markers of disease. Similarly, several genes were identified by our analyses that we have not pursued by qRT-PCR in a wider sample set and there is much work to be done in confirming the utility of these novel markers that we have identified here. This will require extensive follow-up in a gene and/or protein−directed fashion involving further analysis of gene expression alterations in a wide variety of tumor samples, particularly those that are classified pathologically as stage 1.

The ultimate goal is to identify robust targets for the development of serum-based diagnostic tools. Clearly, this will require significant progress in translational research to develop mRNA tumor markers into reliable serum-based assays. One important consideration when selecting gene products for further analysis at the protein level is predicting the magnitude of altered expression at the mRNA level required to produce a detectable protein change. The combination of mRNA data sets with results from emerging proteomic efforts will likely accelerate biomarker identification

and development in this context. Despite these challenges, genome-wide data sets, such as ours, that can be readily shared between investigators will provide a vital foundation for development in this field. The use of an open platform tool, such as SAGE, is an advantage in this context in that it does not rely on any prior knowledge of genes of interest.

In conclusion, we have undertaken a genome-wide screen by SAGE for putative mRNA markers of ovarian cancer in bulk tissue obtained from three adenocarcinomas and two pools of normal HOSE. We further analyzed our data in comparison with publicly available ovarian cancer and HOSE SAGE libraries. The overexpression of a subset of genes was confirmed in a wider sample set of tumors and normal tissue. These data provide an immortal gene expression catalogue for public utility in the identification of potential markers for diagnosis and characterization of ovarian cancer.

References

1. Poveda A. Ovarian cancer treatment: what is new. Int J Gynecol Cancer 2003;13 Suppl 2:241–50.
2. Whittemore AS. Characteristics relating to ovarian cancer risk: implications for prevention and detection. Gynecol Oncol 1994;55:S15–9.
3. McGuire WP, Hoskins WJ, Brady MF, et al. Cyclophosphamide and cisplatin compared with paclitaxel and cisplatin in patients with stage III and stage IV ovarian cancer. N Engl J Med 1996;334:1–6.
4. Oram DH, Jacobs, IJ, Brady JL, Prys-Davies A. Early diagnosis of ovarian cancer. Br J Hosp Med 1990;44:320, 322, 324.
5. Bast RC Jr, Klug TL, St John E, et al. A radioimmunoassay using a monoclonal antibody to monitor the course of epithelial ovarian cancer. N Engl J Med 1983;309:883–7.
6. Mann WJ, Patsner B, Cohen H, Loesch M. Preoperative serum CA-125 levels in patients with surgical stage I invasive ovarian adenocarcinoma. J Natl Cancer Inst 1988;80:208–9.
7. Jacobs I, Bast RC Jr. The CA 125 tumour-associated antigen: a review of the literature. Hum Reprod 1989;4:1–12.
8. St Croix B, Rago C, Velculescu V, et al. Genes expressed in human tumor endothelium. Science 2000;289:1197–202.
9. Chu TJ. Learning from SAGE data. Ph.D. dissertation. Philosophy Department, Carnegie Mellon University; 2003. Available from: http://www.phil.cmu.edu/projects/genegroup/papers.html.
10. Chu TJ. A statistical analysis of SAGE data; 2002. Available from: http://www.phil.cmu.edu/projects/genegroup/papers.html.
11. Collins C, Rommens JM, Kowbel D, et al. Positional cloning of ZNF217 and NABC1: genes amplified at 20q13.2 and overexpressed in breast carcinoma. Proc Natl Acad Sci U S A 1998;95:8703–8.
12. Stein WD, Litman T, Fojo T, Bates SE. A Serial Analysis of Gene Expression (SAGE) database analysis of chemosensitivity: comparing solid tumors with cell lines and comparing solid tumors from different tissue origins. Cancer Res 2004;64:2805–16.
13. Szala S, Froehlich M, Scollon M, et al. Molecular cloning of cDNA for the carcinoma-associated antigen GA733–2. Proc Natl Acad Sci U S A 1990;87:3542–6.
14. Gergely F, Karlsson C, Still I, Cowell J, Kilmartin J, Raff JW. The TACC domain identifies a family of centrosomal proteins that can interact with microtubules. Proc Natl Acad Sci U S A 2000;97:14352–7.
15. Gangisetty O, Lauffart B, Sondarva GV, Chelsea DM, Still IH. The transforming acidic coiled coil proteins interact with nuclear histone acetyltransferases. Oncogene 2004;23:2559–63.
16. Higashiyama S, Iwamoto R, Goishi K, et al. The membrane protein CD9/DRAP 27 potentiates the juxtacrine growth factor activity of the membrane-anchored heparin-binding EGF-like growth factor. J Cell Biol 1995;128:929–38.
17. Hori H, Yano S, Koufuji K, Takeda J, Shirouzu K. CD9 expression in gastric cancer and its significance. J Surg Res 2004;117:208–15.
18. Houle CD, Ding XY, Foley JF, Afshari CA, Barrett JC, Davis BJ. Loss of expression and altered localization of KAI1 and CD9 protein are associated with epithelial ovarian cancer progression. Gynecol Oncol 2002;86:69–78.
19. Masciullo V, Baldassarre G, Pentimalli F, et al. HMGA1 protein overexpression is a frequent feature of epithelial ovarian carcinomas. Carcinogenesis 2003;24:1191–8.
20. Mantovani LT, Miotti S, Menard S, et al. Folate binding protein distribution in normal tissues and biological fluids from ovarian carcinoma patients as detected by the monoclonal antibodies MOv18 and MOv19. Eur J Cancer 1994;30A:363–9.
21. Bagnoli M, Tomassetti A, Figini M, et al. Downmodulation of caveolin-1 expression in human ovarian carcinoma is directly related to α-folate receptor overexpression. Oncogene 2000;19:4754–63.
22. Toffoli G, Cernigoi C, Russo A, Gallo A, Bagnoli M, Boiocchi M. Overexpression of folate binding protein in ovarian cancers. Int J Cancer 1997;74:193–8.
23. Hough CD, Sherman-Baust CA, Pizer ES, et al. Large-scale Serial Analysis of Gene Expression reveals genes differentially expressed in ovarian cancer. Cancer Res 2000;60:6281–7.
24. Ono K, Tanaka T, Tsunoda T, et al. Identification by cDNA microarray of genes involved in ovarian carcinogenesis. Cancer Res 2000;60:5007–11.
25. Welsh JB, Zarrinkar PP, Sapinoso LM, et al. Analysis of gene expression profiles in normal and neoplastic ovarian tissue samples identifies candidate molecular markers of epithelial ovarian cancer. Proc Natl Acad Sci U S A 2001;98:1176–81.
26. Hata K, Fujiwaki R, Nakayama K, Miyazaki K. Expression of the endostatin gene in epithelial ovarian cancer. Clin Cancer Res 2001;7:2405–9.
27. O'Reilly MS, Boehm T, Shing Y, et al. Endostatin: an endogenous inhibitor of angiogenesis and tumor growth. Cell 1997;88:277–85.
28. Hata K, Dhar DK, Kanasaki H, et al. Serum endostatin levels in patients with epithelial ovarian cancer. Anticancer Res 2003;23:1907–12.
29. Schummer M, Ng WV, Bumgarner RE, et al. Comparative hybridization of an array of 21,500 ovarian cDNAs for the discovery of genes overexpressed in ovarian carcinomas. Gene 1999;238:375–85.
30. Dhar KK, Branigan K, Parkes J, et al. Expression and subcellular localization of cyclin D1 protein in epithelial ovarian tumour cells. Br J Cancer 1999;81:1174–81.
31. Geisler JP, Tammela JE, Manahan KJ, et al. HSP27 in patients with ovarian carcinoma: still an independent prognostic indicator at 60 months follow-up. Eur J Gynaecol Oncol 2004;25:165–8.
32. Aitola M, Sadek CM, Gustafsson JA, Pelto-Huikko M. Aint/Tacc3 is highly expressed in proliferating mouse tissues during development, spermatogenesis, and oogenesis. J Histochem Cytochem 2003;51:455–69.
33. Xu Y, Shen Z, Wiper DW, et al. Lysophosphatidic acid as a potential biomarker for ovarian and other gynecologic cancers. JAMA 1998;280:719–23.
34. Tokumura A, Tominaga K, Yasuda K, Kanzaki H, Kogure K, Fukuzawa K. Lack of significant differences in the corrected activity of lysophospholipase D, producer of phospholipid mediator lysophosphatidic acid, in incubated serum from women with and without ovarian tumors. Cancer 2002;94:141–51.
35. Peters DG, Kassam AB, Yonas H, O'Hare EH, Ferrell RE, Brufsky AM. Comprehensive transcript analysis in small quantities of mRNA by SAGE-lite. Nucleic Acids Res 1999;27:e39.

Comparative Gene Expression Analysis of Ovarian Carcinoma and Normal Ovarian Epithelium by Serial Analysis of Gene Expression

David G. Peters, Donna M. Kudla, Julie A. DeLoia, et al.

Cancer Epidemiol Biomarkers Prev 2005;14:1717-1723.