Abstract

One reason that ovarian cancer is such a deadly disease is because it is not usually diagnosed until it has reached an advanced stage. In this study, we developed a novel algorithm for group biomarkers identification using gene expression data. Group biomarkers consist of coregulated genes across normal and different stage diseased tissues. Unlike prior sets of biomarkers identified by statistical methods, genes in group biomarkers are potentially involved in pathways related to different types of cancer development. They may serve as an alternative to the traditional single biomarkers or combination of biomarkers used for the diagnosis of early-stage and/or recurrent ovarian cancer. We extracted group biomarkers by applying biclustering algorithms that we recently developed on the gene expression data of over 400 normal, cancerous, and diseased tissues. We identified several groups of coregulated genes that encode for secreted proteins and exhibit expression levels in ovarian cancer that are at least 2-fold (in log₂ scale) higher than in normal ovary and nonovarian tissues. In particular, three candidate group biomarkers exhibited a conserved biological pattern that may be used for early detection or recurrence of ovarian cancer with specificity greater than 99% and sensitivity equal to 100%. We validated these group biomarkers using publicly available gene expression data sets downloaded from a NIH Web site (http://www.ncbi.nlm.nih.gov/geo). Statistical analysis showed that our methodology identified an optimum combination of genes that have the highest effect on the diagnosis of the disease compared with several computational techniques that we tested. Our study also suggests that single or group biomarkers correlate with the stage of the disease. [Mol Cancer Ther 2008;7(1):27–37]

Introduction

Epithelial ovarian cancer is the most lethal form of gynecologic cancer and the fourth leading cause of cancer death among women in developed countries, claiming about 15,000 lives in the United States each year (1). One reason it is so deadly is the fact that ovarian cancer is not usually diagnosed until it has reached an advanced stage. Early detection can help prolong or save lives, but clinicians currently have no specific and sensitive screening method and the disease displays very subtle symptoms (2). The well-known CA-125 blood test and other imaging techniques, such as ultrasound and computed tomographic scan, or the combination of the CA-125 blood test with one of the above imaging techniques, are useful for tracking patients already diagnosed with ovarian cancer but have not proven sensitive enough to be used as an early diagnostic test (3).

In recent years, large-scale gene expression analyses have been done to identify differentially expressed genes in ovarian carcinoma (4–11). A common goal of these studies was to identify potential tumor markers for the diagnosis of early-stage ovarian cancer as well as to use these markers as targets for improved therapy and treatment of the disease during all stages.

Numerous computational tools have been developed to analyze gene expression data for biomarker discovery (12–19). Most focus on differential gene expression, which is tested by a simple calculation of the fold changes by t test, F test, scoring methods (12), or cluster analysis (13). Many other computational techniques based on a supervised learning approach have also been developed (e.g. support vector machine; ref. 15, naive Bayes method, and Fisher discriminant analysis; refs. 16–19). Although most of these approaches have been successful in uncovering interesting patterns that can be used to discriminate between healthy and diseased tissues, computational tools for the identification of potential blood biomarkers are still not well developed or do not take into account all of the input variables. Most approaches only do a comparison between healthy and diseased tissues of the corresponding disease and do not take into consideration other tissues in the body that may produce the same protein as the diseased tissue. Therefore, potential biomarkers identified using these approaches may introduce false positives in a diagnostic blood test.

In this study, we first used a computational tool that we recently developed (20) to identify single biomarkers. Then, we used a second novel algorithm that we recently developed to identify group biomarkers (20–22). Group biomarkers correspond to a set of single biomarkers that exhibit coherent behavior across an ordered set of ovarian tissue samples. This approach is based on the observation that single or group biomarkers correlate with the stage of the disease.
cancer tissue samples, representing distinct stages of the disease. That is, their expression level increases or decreases coherently during the progression of the disease. This unique pattern shows a correlation or coregulation among the set of genes that belong to the same group biomarkers, suggesting that they respond similarly to the same environmental conditions. Prior studies on different organisms have examined several biclusters of coregulated genes and showed that the genes in a given bicluster typically participate in a single pathway (23).

Our methodology for identifying single or group biomarkers is based on unifying techniques that are well understood and developed in the literature: gene expression data analysis, biclustering algorithms, and receiver operating characteristics (ROC) curves. Furthermore, our approach for identifying blood biomarkers is based on the observation that, if we are looking for biomolecular patterns in the blood that are caused by ovarian cancer, those patterns should only be present in the gene expression data of ovarian cancer tissue samples compared with the gene expression data of normal ovary tissue samples or any nonovarian healthy or diseased tissue samples.

We implemented our approach using the computer program Matlab and applied it to a comprehensive set of well-defined gene expression data corresponding to normal ovary, ovarian cancer, and nonovarian tissue samples. We identified three candidate group biomarkers that encode for secreted proteins, membrane proteins, and/or extracellular matrix proteins. These three candidate group biomarkers clearly discriminate between the sample sets, and they are promising candidates to be used for early detection or recurrence of ovarian cancer using a blood test. Statistical analysis showed that these group biomarkers have a much better detection performance than single biomarkers and combinations of biomarkers identified using other computational approaches. Our data also suggest that single or group biomarkers correlate with the stage of the disease.

Materials and Methods

Tissue Samples

Table 1 lists all of the tissue samples used in this study. They can be classified into four different sample sets: normal ovary, ovarian cancer, normal nonovarian, and diseased nonovarian. Normal ovaries were obtained from 62 women. Seven borderline ovarian tumors were obtained; these tumors are considered to be of low malignant potential and were not staged. Next, we obtained tissue samples of stage III or IV serous epithelial ovarian cancer derived from two different sites: 22 from the ovary itself and 16 from the omentum. Tissues were ranked from normal to low malignancy to highly malignant as follows: normal ovaries, borderline ovarian tumors, primary serous epithelial ovarian tumors present in ovarian tissues, and serous epithelial ovarian tumors present in omental tissues. None of the patients had been treated with chemotherapy before surgical resection of the tissues (10, 11).

Tissues were obtained from the University of Minnesota Cancer Center Tissue Procurement Facility on approval by the University of Minnesota Institutional Review Board. Tissue Procurement Facility employees obtained signed consent from each patient, allowing procurement of excess waste tissue and access to medical records. Bulk tumor and normal tissues were identified, dissected, and snap frozen in liquid nitrogen within 15 to 30 min of resection from the patient. Tissue sections were made from each sample, stained with H&E, and examined independently by two pathologists to confirm the pathological state of each sample. The integrity of the RNA was verified before use in gene array experiments (10, 11).

Gene Expression Matrix

The gene expression data were determined by Gene Logic using the Affymetrix GeneChip HG U95A, which contains 12,651 known genes and 48,000 expressed sequence tags. The gene expression data were normalized using Affymetrix M.A.S. 4.0.1, and the log-floor data transform with a floor value of 1 was done (24). After this process, the data ranged from 0 to 4. The data were then organized into three matrices defined as follows: matrix A is a 12,651 × 62 matrix that represents the gene expression of the 62 normal ovary tissue samples; matrix B = [B1 B2 B3] is a 12,651 × 45 matrix that represents the gene expression of the 45 ovarian cancer tissue samples; submatrix B1 is a 12,651 × 7 matrix representing the gene expression of the 7 borderline ovarian cancer tissues; submatrix B2 is a 12,651 × 22 matrix, which represents the gene expression of the 22 papillary serous adenocarcinoma tumors; submatrix B3 is a 12,651 × 16 matrix representing the gene expression of the 16 omentum papillary serous adenocarcinoma; and matrix C is a 12,651 × 319 matrix that represents the gene expression of the 319 nonovarian tissues.

Identification of Single Biomarkers in Ovarian Carcinoma

Biomarkers specific for ovarian cancer should be highly expressed in ovarian cancer samples and low or absent in other samples, including normal ovaries and nonovarian tissues. Mathematically, they should correspond to the set of genes that are up-regulated in ovarian cancer tissue samples compared with normal ovary tissue samples and each set of nonovarian tissue samples. In this study, we assume that for a given gene to correspond to a potential biomarker it should be at least 2-fold (log2 scale) up-regulated in ovarian cancer tissue samples compared with normal ovary tissue samples and each set of nonovarian tissue samples [that is, log2(y/x) ≥ 2 and log2(y/z) ≥ 2, where x, y, and z correspond to the expression level of a gene in the healthy ovary, ovarian cancer, and nonovarian data, respectively]. Also, the corresponding gene must exhibit a sensitivity of ≥90% for a specificity of ≥90%, with accuracy of ≥90%. Identification of such a pattern is done in this study using the combination of the Robust Biclustering Algorithm (20) and the ROC approach we will define here. To develop a diagnostic assay for the detection of ovarian cancer using a blood test, biomarkers should also correspond to genes that encode for predicted secreted
proteins, membrane proteins, and/or extracellular matrix proteins. These types of proteins are more likely to be present in the blood than proteins localized to the cell nucleus or cytoplasm.

**Biclustering Approach**

We used the Robust Biclustering Algorithm that we recently developed (20) to identify biclusters with constant values. Given the above gene expression matrices: \( A, B, \) and \( C \), with sets of rows or genes \( G = \{g_1, g_2, \ldots, g_N\} \) and set of conditions or tissue samples \( S_A = \{s_1^A, s_2^A, \ldots, s_{M_A}^A\}, S_B = \{s_1^B, s_2^B, \ldots, s_{M_B}^B\}, \) and \( S_C = \{s_1^C, s_2^C, \ldots, s_{M_C}^C\} \), respectively. We define a bicluster with constant values, that is, a subset of genes that the expression level stay constant across a subset of conditions or tissue samples as \( M_k = \{i_k^A, j_k^B\} \), \( M_k^B = \{i_k^B, j_k^B\}, \) and \( M_k^C = \{i_k^C, j_k^C\} \) or as submatrices \( M_k^A = [I_k^A(i,j)], M_k^B = [I_k^B(i,j)], \) and \( M_k^C = [I_k^C(i,j)] \) of \( A, B, \) and \( C, \) respectively. The \( I_k \) correspond to the subsets of genes \( G, \) the \( J_k \) correspond to the subsets of tissue samples \( S_A, S_B, \) or \( S_C, \) and \( M(i,j) \) corresponds to the expression level of gene \( i \) in condition \( j \)th under condition \( i \), with \( i = i_k^A \) and \( j = j_k^B \). Identification of potential biomarkers can be done using Eq. (1) below, with \( 1 \leq k \leq N_A, 1 \leq i \leq N_B, \) and \( 1 \leq m \leq N_C \)

\[
I = \cup (I_k^A \cap I_k^B \cap I_k^C).
\]

In Eq. (1), \( N_A \) corresponds to the number of biclusters \( M_k^A = \{i_k^A, j_k^A\} \) with constant values \( x \) in the healthy ovary tissue samples data set, \( N_B \) corresponds to the number of biclusters \( M_k^B = \{i_k^B, j_k^B\} \) with constant values \( y \) (\( y >> x \)) in the ovarian cancer tissue samples data set, and \( N_C \) corresponds to the number of biclusters \( M_k^C = \{i_k^C, j_k^C\} \) with constant values \( z \) (\( z << y \)) in the nonovarian tissue sample data set.

Here, we considered each set of ovarian cancer tissue data separately and the expression level of the gene considered in the ovarian cancer tissue samples should be at least 2-fold (log2 scale) greater than the expression level of the same gene in normal ovary tissue samples and in each set of nonovarian tissue samples. Ideally, when dealing with blood biomarkers, we would like \( x = z = 0 \).

The statistical performance of a given bicluster \( M = [M(i,j)] \) with constant value was then evaluated using the following equation: for all rows, \( M(i,:) \) of \( M = [M(i,j)] \),

\[
\text{Max}(M(i,:)) - \text{Min}(M(i,:)) \leq \delta
\]

with \( \delta \rightarrow 0 \) (that is, \( \delta \) is a real positive small number).

**ROC Approach**

Given the gene expression data as defined above, the ROC approach first assumes that all genes correspond to potential biomarkers. Then, it uses the following criterion based on the detection performance exhibited by their respective ROC curve to select the ones with high specificity corresponding to high sensitivity and high accuracy.

For a given screening cutoff point, let \( a \) be the number of healthy ovary and nonovarian tissue samples (healthy and diseased) that screen positive, \( b \) the number of ovarian cancer tissue samples that screen positive, \( c \) the number of healthy ovary and nonovarian tissue samples that screen negative, and \( d \) the number of ovarian cancer tissue samples that screen negative. The sensitivity of a potential blood biomarker (Se) is the number of ovarian cancer tissue samples that screen positive divided by the total number of ovarian cancer tissue samples: \( Se = b / (b + d) \). The specificity of a potential blood biomarker (Sp) is the number of healthy ovary and nonovarian tissue samples that screen negative divided by the total number of healthy ovary and nonovarian tissue samples: \( Sp = c / (c + a) \). Using these variables, we compute the ROC function of each potential blood biomarker using the following equation: \( Se = f(1 - Sp) \). Basically, \( f(1 - Sp) \) describes the relationship between the true-positive rate (sensitivity) and the false-positive rate (1 - specificity) for different screening cutoff points. Finally, the ROC methodology keeps all genes capable of achieving specificity corresponding to sensitivity as well as accuracy greater than the defined specified thresholds. The resultant family of genes will correspond to biomarkers that may then be evaluated for their use in the detection of ovarian cancer using a blood test.

The \( P \) of each identified single biomarker, that is, the probability of observing the given result, or one more extreme by chance if the null hypothesis is true, was estimated using a two-sided test.
Table 2. Fifty-four genes identified by the single biomarker algorithm to be up-regulated in ovarian cancer tissue samples compared with normal ovary tissue samples and nonovarian tissue samples

<table>
<thead>
<tr>
<th>Fragment name</th>
<th>Gene name</th>
<th>Known gene symbol</th>
<th>Ovarian cancer borderline</th>
<th>Ovarian cancer primary</th>
<th>Ovarian cancer omentum</th>
</tr>
</thead>
<tbody>
<tr>
<td>33454_at</td>
<td>Agrin</td>
<td>AGRN</td>
<td>2.2</td>
<td>2.4</td>
<td>2.3</td>
</tr>
<tr>
<td>757_at</td>
<td>Annexin A2, Annexin A2 pseudogene 2</td>
<td>ANXA2</td>
<td>3.0</td>
<td>3.1</td>
<td>2.1</td>
</tr>
<tr>
<td>35099_at</td>
<td>Apolipoprotein L1</td>
<td>APOL1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2011_s_at</td>
<td>BCL2-interacting killer (apoptosis-inducing)</td>
<td>(BIK, KIAA1654)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>35822_at</td>
<td>B-factor properdin</td>
<td>BF*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>41534_at</td>
<td>BH-protocadherin (brain-heart)</td>
<td>PCDH7</td>
<td>2.0</td>
<td>4.4</td>
<td>5.2</td>
</tr>
<tr>
<td>1620_at</td>
<td>Cadherin 6, type 2; K-cadherin (fetal kidney)</td>
<td>CDH6</td>
<td>2.6</td>
<td>6.2</td>
<td>5.2</td>
</tr>
<tr>
<td>41660_at</td>
<td>Cadherin, EGF LAG seven-pass G-type receptor 1 (flamingo homologue, Drosophila)</td>
<td>CELSR1*</td>
<td>3.5</td>
<td>3.8</td>
<td>3.8</td>
</tr>
<tr>
<td>36499_at</td>
<td>Cadherin, EGF LAG seven-pass G-type receptor 2 (flamingo homologue, Drosophila)</td>
<td>CELSR2*</td>
<td>2.2</td>
<td>2.0</td>
<td>2.0</td>
</tr>
<tr>
<td>37890_at</td>
<td>CD47 antigen (Rh-related antigen, integrin-associated signal transducer) (cartilage glycoprotein-39)</td>
<td>CD47</td>
<td>2.1</td>
<td>2.4</td>
<td>2.7</td>
</tr>
<tr>
<td>39008_at</td>
<td>Ceruloplasmin (ferroxidase)</td>
<td>CP</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>431_at</td>
<td>Chemokine (C-X-C motif) ligand 10</td>
<td>CXCL10</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>36197_at</td>
<td>Chitinase 3-like 1 (cartilage glycoprotein-39)</td>
<td>CHI3L1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>33904_at</td>
<td>Claudin 3</td>
<td>CLDN3</td>
<td>6.5</td>
<td>6.6</td>
<td>1.0</td>
</tr>
<tr>
<td>35276_at</td>
<td>Claudin 4</td>
<td>CLDN4</td>
<td>5.9</td>
<td>5.5</td>
<td>1.6</td>
</tr>
<tr>
<td>38482_at</td>
<td>Claudin 7</td>
<td>CLDN7</td>
<td>5.1</td>
<td>5.5</td>
<td>1.6</td>
</tr>
<tr>
<td>37534_at</td>
<td>Coxsackie virus and adenovirus receptor</td>
<td>CXADR</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>35453_at</td>
<td>Dermatan sulfate proteoglycan 3</td>
<td>DSPG3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>36643_at</td>
<td>Discoidin domain receptor family, member 1</td>
<td>DDR1</td>
<td>2.1</td>
<td>2.4</td>
<td>2.2</td>
</tr>
<tr>
<td>1007_s_at</td>
<td>Discoidin domain receptor family, member 1</td>
<td>DDR1</td>
<td>2.1</td>
<td>2.1</td>
<td>2.1</td>
</tr>
<tr>
<td>41586_at</td>
<td>Fibroblast growth factor 18</td>
<td>FGF18</td>
<td>2.9</td>
<td>2.9</td>
<td>4.9</td>
</tr>
<tr>
<td>41587_g_at</td>
<td>Fibroblast growth factor 18</td>
<td>FGF18</td>
<td>3.7</td>
<td>1.6</td>
<td>3.6</td>
</tr>
<tr>
<td>534_s_at</td>
<td>Folate receptor 1 (adult)</td>
<td>FOLR1</td>
<td>2.5</td>
<td>2.6</td>
<td>4.9</td>
</tr>
<tr>
<td>821_s_at</td>
<td>Folate receptor 1 (adult)</td>
<td>FOLR1</td>
<td>2.4</td>
<td>8.0</td>
<td>7.6</td>
</tr>
<tr>
<td>38749_at</td>
<td>G protein-coupled receptor 39, LY6/PLAUR domain containing 1</td>
<td>GPR39, LYPDC1*</td>
<td>6.0</td>
<td>7.3</td>
<td>8.6</td>
</tr>
<tr>
<td>406_at</td>
<td>Integrin β3</td>
<td>ITGB4, (A)</td>
<td>3.2</td>
<td>2.9</td>
<td>2.4</td>
</tr>
<tr>
<td>3754_at</td>
<td>Kallikrein 6 (neurosin, zyme)</td>
<td>KLK6</td>
<td>3.3</td>
<td>4.4</td>
<td>4.7</td>
</tr>
<tr>
<td>38143_at</td>
<td>Kallikrein 7 (chymotryptic, stratum corneum)</td>
<td>KLK7, (C)</td>
<td>3.3</td>
<td>4.4</td>
<td>4.7</td>
</tr>
<tr>
<td>37131_at</td>
<td>Kallikrein 8 (neuropsin/ovasin)</td>
<td>KLK8, (C)</td>
<td>5.4</td>
<td>6.1</td>
<td>1.5</td>
</tr>
<tr>
<td>36838_at</td>
<td>Kallikrein 10</td>
<td>KLK10</td>
<td>6.1</td>
<td>3.6</td>
<td>3.5</td>
</tr>
<tr>
<td>40035_at</td>
<td>Kallikrein 11</td>
<td>KLK11</td>
<td>4.7</td>
<td>9.9</td>
<td>5.2</td>
</tr>
<tr>
<td>36929_at</td>
<td>Laminin β1</td>
<td>LAMB3</td>
<td>3.7</td>
<td>1.9</td>
<td></td>
</tr>
<tr>
<td>35280_at</td>
<td>Laminin γ2</td>
<td>LAMC2*</td>
<td>5.3</td>
<td>1.6</td>
<td>2.6</td>
</tr>
<tr>
<td>39583_at</td>
<td>Leucine-rich repeat neuronal 5</td>
<td>LRRN5</td>
<td>2.4</td>
<td>1.1</td>
<td></td>
</tr>
<tr>
<td>32821_at</td>
<td>Lipocalin 2 (oncogene 24p3)</td>
<td>LCN2*, (A)</td>
<td>6.3</td>
<td>5.1</td>
<td>4.9</td>
</tr>
<tr>
<td>40093_at</td>
<td>Lutheran blood group (Auberger b antigen included)</td>
<td>LU</td>
<td>2.6</td>
<td>3.5</td>
<td>2.8</td>
</tr>
</tbody>
</table>

(Continued on the following page)
Identification of Group Biomarkers in Ovarian Carcinoma

Identification of group biomarkers was done using a randomly selected set of 40 of the 62 normal ovary tissues and 30 of the 45 ovarian cancer tissues (5 of the 7 borderline ovarian cancer tissues, 15 of the 22 papillary serous adenocarcinoma tumors, and 10 of the 16 omentum papillary serous adenocarcinoma metastases). Briefly, we applied Eqs. (1) and (2) and the ROC approach on the randomly selected set of data to uncover potential single biomarkers. Then, the gene expression data of the single biomarkers identified were sorted according to the progression of the disease. Given that we only had three different stages (normal ovary, borderline ovarian cancer, and ovarian cancer omentum).

Table 2. Fifty-four genes identified by the single biomarker algorithm to be up-regulated in ovarian cancer tissue samples compared with normal ovary tissue samples and nonovarian tissue samples (Cont’d)

<table>
<thead>
<tr>
<th>Fragment name</th>
<th>Gene name</th>
<th>Known gene symbol</th>
<th>Ovarian cancer borderline</th>
<th>Ovarian cancer primary</th>
<th>Ovarian cancer omentum</th>
</tr>
</thead>
<tbody>
<tr>
<td>32072_at</td>
<td>Mesothelin</td>
<td>MSLN, (B), (C)</td>
<td>3.4 2.5e-05</td>
<td>4.1 2.9e-19</td>
<td>4.5 1.5e-17</td>
</tr>
<tr>
<td>38784_g_at</td>
<td>Mucin 1, transmembrane</td>
<td>MUC1, (B)</td>
<td>4.9 4.7e-13</td>
<td>4.8 2.8e-26</td>
<td>4.6 1.4e-16</td>
</tr>
<tr>
<td>927_s_at</td>
<td>Mucin 1, transmembrane</td>
<td>MUC1</td>
<td>4.5 5.7e-07</td>
<td>4.9 7.5e-18</td>
<td>4.1 1.5e-12</td>
</tr>
<tr>
<td>38783_at</td>
<td>Mucin 1, transmembrane</td>
<td>MUC1</td>
<td>6.2 7.5e-07</td>
<td>6.5 6.6e-16</td>
<td>5.9 5.5e-12</td>
</tr>
<tr>
<td>1083_s_at</td>
<td>Mucin 1, transmembrane</td>
<td>MUC1</td>
<td>3.6 3.6e-06</td>
<td>4.1 2.3e-17</td>
<td>3.9 5.9e-13</td>
</tr>
<tr>
<td>39512_at</td>
<td>Mucin 4, tracheobronchial</td>
<td>MUC4</td>
<td>3.1 7.9e-05</td>
<td></td>
<td></td>
</tr>
<tr>
<td>32625_at</td>
<td>Natriuretic peptide receptor A/guanylate cyclase A (atrionatriuretic peptide receptor A)</td>
<td>NPR1</td>
<td></td>
<td></td>
<td>2.4 1.4e-06</td>
</tr>
<tr>
<td>33483_at</td>
<td>Neuraminidase U</td>
<td>NMX</td>
<td></td>
<td></td>
<td>4.3 3.2e-19</td>
</tr>
<tr>
<td>35663_at</td>
<td>Neuronal pentraxin II</td>
<td>NPTX2</td>
<td>2.0 1.9e-07</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1985_s_at</td>
<td>Nonmetastatic cells 1, protein (NM23A) expressed in nonmetastatic cells 2, protein (NM23B)</td>
<td>NME1, NME2</td>
<td>2.1 1.6e-11</td>
<td></td>
<td></td>
</tr>
<tr>
<td>33783_at</td>
<td>Plexin B1</td>
<td>PLXNB1*</td>
<td>2.3 1.7e-04</td>
<td>2.9 2.2e-12</td>
<td>2.8 6.9e-10</td>
</tr>
<tr>
<td>34780_at</td>
<td>Plexin B2</td>
<td>PLXNB2*</td>
<td>2.1 9.4e-08</td>
<td></td>
<td></td>
</tr>
<tr>
<td>41106_at</td>
<td>Potassium intermediate/small conductance calcium-activated channel, subfamily N, member 4</td>
<td>KCNN4</td>
<td>4.7 9.3e-07</td>
<td></td>
<td></td>
</tr>
<tr>
<td>41470_at</td>
<td>Prominin 1</td>
<td>PROM1</td>
<td>3.8 7.2e-05</td>
<td>3.4 3.0e-04</td>
<td>4.1 2.3e-08</td>
</tr>
<tr>
<td>32275_at</td>
<td>Secretory leukocyte protease inhibitor (antileukoproteinase)</td>
<td>SLPI*</td>
<td>4.2 1.3e-04</td>
<td>4.2 7.6e-11</td>
<td>4.1 2.3e-08</td>
</tr>
<tr>
<td>39075_at</td>
<td>Sialidase 1 (lysosomal sialidase)</td>
<td>NEU1</td>
<td>5.6 3.7e-06</td>
<td>6.0 2.9e-17</td>
<td>6.2 6.7e-14</td>
</tr>
<tr>
<td>35207_at</td>
<td>Sodium channel, non-voltage-gated 1x</td>
<td>SCN1A*</td>
<td>3.7 1.0e-06</td>
<td>1.2 2.4e-08</td>
<td>2.2 6.6e-08</td>
</tr>
<tr>
<td>36609_at</td>
<td>Solute carrier family 1 (glial high-affinity glutamate transporter), member 3</td>
<td>SLC1A3</td>
<td>5.4 4.4e-05</td>
<td>5.6 3.5e-13</td>
<td>5.4 1.1e-09</td>
</tr>
<tr>
<td>35277_at</td>
<td>Spondin 1, extracellular matrix protein</td>
<td>SPON1</td>
<td>2.2 1.6e-07</td>
<td></td>
<td></td>
</tr>
<tr>
<td>575_s_at</td>
<td>Tumor-associated calcium signal transducer 1</td>
<td>TACSTD1*</td>
<td>5.4 4.4e-05</td>
<td>5.6 3.5e-13</td>
<td>5.4 1.1e-09</td>
</tr>
<tr>
<td>291_s_at</td>
<td>Tumor-associated calcium signal transducer 2</td>
<td>TACSTD2*</td>
<td>4.7 2.4e-06</td>
<td></td>
<td></td>
</tr>
<tr>
<td>33218_at</td>
<td>V-erb-b2 erythroblastic leukemia viral oncogene homologue 2, neuro/glioblastoma-derived oncogene homologue (avian)</td>
<td>ERBB2</td>
<td>2.0 2.7e-05</td>
<td>2.1 8.5e-16</td>
<td></td>
</tr>
<tr>
<td>33933_at</td>
<td>WAP four-disulfide core domain 2</td>
<td>WFDC2, (B)</td>
<td>4.8 7.9e-06</td>
<td>5.3 1.5e-17</td>
<td>5.2 2.1e-13</td>
</tr>
<tr>
<td>1887_g_at</td>
<td>Wingless-type MMTV integration site family, member 7A</td>
<td>WNT7A*, (A)</td>
<td>3.0 7.8e-22</td>
<td>2.3 7.1e-22</td>
<td>3.4 2.2e-33</td>
</tr>
</tbody>
</table>

NOTE: “(A),” “(B),” and “(C)” are genes that belong to group biomarkers “A,” “B,” and “C,” respectively. Fold change relative to normal ovary tissues; $P$ values relative to normal ovary tissues and nonovarian tissues. Selection criteria: Up-regulated in ovarian cancer tissue samples at least 2-fold (in log2 scale) compared with normal ovary tissue samples and each set of nonovarian tissue samples. Specificity greater or equal to 90% corresponding to sensitivity greater or equal to 90%. Genes code for proteins that are secreted, extracellular, or membranous.

*Genes not previously linked in the literature to ovarian cancer.
primary ovarian cancer) and two different sites of ovarian cancer (ovary and the omentum), the stages were repeated periodically every three samples. Data were organized as 

\[ D = [D_1 \ D_2 \ D_3 \ D_1 \ D_2 \ D_3 \ \ldots \ D_1 \ D_2 \ D_3] \]

where \( D_1 \) is a column vector representing the expression level of one of the 40 randomly selected normal ovary tissues, \( D_2 \) is a column vector representing the expression level of one of the 5 randomly selected borderline ovarian tissues, and \( D_3 \) is one of the 15 randomly selected papillary serous adenocarcinoma or one of the 10 randomly selected omentum papillary serous adenocarcinoma metastases. Also, because we only had 5 randomly selected borderline samples, the maximum number of columns or tissue samples in \( D \) that we could have was 15. We therefore produced several \( D \) matrices, which used the same borderline data. Different matrices had different combinations of randomly selected normal ovary tissues and papillary serous adenocarcinoma of ovarian tumors or omentum papillary serous adenocarcinoma ovarian tumors.

In all, we examined 8 such matrices using the order preserving biclustering algorithm of Tchagang and Tewfik (21, 22) and retained the genes that appeared as many times as possible in the same bicluster. The order preserving biclustering algorithm has the advantage of being insensitive to the relative position of each tissue of a given kind. That is, it produces the same output under any permutation of the positions of the normal ovary tissues, papillary serous adenocarcinoma of ovarian tumors, or omentum papillary serous adenocarcinoma of ovarian tumors.

In our problem, the bicluster identification step of refs. 21, 22 consists of two substeps. In the first substep, the procedure enumerates all combinations of \( K \) tissues, where \( K \geq K_{\text{min}} \) the prespecified minimum number of tissues in a valid bicluster, from the given \( M_D \) tissues in matrix \( D \) that could potentially appear in a valid bicluster. For each subset of \( K \) tissues, it then uses a row sort procedure that allows us to focus on the coherent evolutions of gene expression levels rather than the raw or processed expression levels. The output of this step is a matrix that contains the rank of each of the \( K \) tissues for each row (gene) when the expression level at each tissue for the given gene are ordered in a nondecreasing manner. This matrix is referred to as the “tissue rank matrix” and used as the input to the main bicluster identification routine (22).

In the second substep, the main bicluster identification routine identifies all valid coherent evolution patterns involving all genes and a set of \( K \) tissues “simultaneously” through a fast row sorting procedure. Note that this allows the algorithm to identify all the possible valid biclusters “without” an exhaustive enumeration of all possible \( K! \) permutations of the \( K \) tissues. The procedure will also yield biclusters of genes where a subset of genes are coherently up-regulated and another subset coherently down-regulated across the \( K \) tissues. A final pruning step eliminates all biclusters that are completely included in larger ones (22).

The statistical significance of each identified group biomarker with \( G \) genes is assessed using Eq. (3), that is, the upper bound of the tail probability that a random data set of size \( I \times J \) will contain an order preserving bicluster with \( G \) or more genes in it (21).

\[
Z(J, G) = J! \sum_{i=0}^{J} \binom{J}{i} \left( \frac{1}{J} \right)^i \left( 1 - \frac{1}{J} \right)^{J-i}
\]

As long as that upper bound probability is smaller than any desired significance level, the identified group biomarker will be statistically significant.

**Results**

**Single Biomarker Algorithm**

Using Eqs. (1) and (2) with \( \delta < 1 \) and the ROC approach, we identified 54 genes that are up-regulated in ovarian cancer tissue samples at least 2-fold (\( \log_2 \) scale) compared with normal ovary tissue samples and each set of nonovarian healthy and diseased tissue samples used in this study (Table 2). The 54 genes achieved a specificity greater or equal to 90% corresponding to a sensitivity greater or equal to 90% using the ROC approach. The 54 genes encode for predicted secreted proteins, membrane proteins, and/or extracellular proteins. Therefore, they represent proteins that have the potential to be present in the blood of ovarian cancer patients and may prove useful in an ovarian cancer diagnostic blood test.

We analyzed each stage of ovarian cancer separately. From our analysis and as shown in the following sections, we found that many of the ovarian cancer biomarkers correlated with the stage of ovarian cancer. That is, some biomarkers did very well on some stages of ovarian cancer but not as well on others. For example, chemokine (C-X-C motif) ligand 10 and chitinase 3-like 1 were found to be up-regulated in the omental metastases but were not up-regulated in the borderline ovarian cancer or primary ovarian cancer tissue relative to normal ovary (Table 2).

The genes listed in Table 2 include most of the potential biomarkers uncovered by previous studies (4–11) as well as an additional 13 that do not appear to have been mentioned in the literature before as potential ovarian cancer biomarkers (indicated by an asterisk). For example, the G protein-coupled receptor 39, LY6/PLAUR domain containing 1 (GPR39, LYPDC1) corresponds to a gene that has not been mentioned in the literature before as a potential ovarian cancer biomarker. GPR39, LYPDC1 is at least 6-fold (\( \log_2 \) scale) up-regulated in ovarian cancer tissue samples compared with normal ovary tissue samples and each set of nonovarian tissue samples used in this study (Fig. 1A). By ROC analysis, GPR39, LYPDC1 achieved a specificity greater than 90% for a sensitivity greater than 90% when used to detect ovarian cancer at each stage (Fig. 1B).

**Group Biomarkers**

Using the order preserving technique as described above on the gene expression data of the set of single biomarkers listed in Table 2, we identified three potential group biomarkers that exhibited unique and conserved biological

---

3 http://ca.expasy.org/sprot/
patterns across the ranked data sets that we randomly generated. Because we were looking for the group of genes that exhibited coherent behavior across the largest number of ranked tissue samples in each one of the eight matrices that we randomly generated and because each random matrix had 54 rows (genes) and 15 columns (ranked tissue samples), we fixed \( K_{\text{min}} = 15 \).

The three genes (ITGB4, LCN2, and WNT7A) identified with “A” in Table 2 represent the set of genes that belong to group biomarker “A,” \( Z = 3.7 \times 10^{-92} \). They exhibit a coherent behavior across the following ranked conditions: normal ovary, borderline ovarian cancer, and primary ovarian cancer. Integrin \( \beta_4 \) (ITGB4) encodes for a membrane protein. It is a receptor for laminin and it plays a critical structural role in the hemidesmosomes of epithelial cells (25). It has been shown previously to be up-regulated in ovarian cancer (10). Lipocalin 2 (oncogene 24p3; LCN2) encodes for a secreted protein. It transports small lipophilic substances and forms a heterodimer with type V collagenase (MMP-9). Although LCN2 has been shown to be up-regulated in patients with renal cell carcinoma (26), little is mentioned in the literature about LCN2 and its role in ovarian cancer. Wingless-type MMTV integration site family, member 7A (WNT7A) encodes for a secreted protein that is present in the extracellular matrix. It is a ligand for members of the frizzled family of seven transmembrane receptors (27). It is a developmental protein; signaling by WNT7A allows sexual dimorphic development of the Mullerian ducts (27). WNT7A has been

Figure 1. A, mean expression level of GPR39, LYPD1 in various tissues. GPR39, LYPD1 is at least 6-fold (log2 scale) up-regulated in ovarian cancer tissue samples compared with normal ovary tissue samples and each set of nonovarian tissue samples used in this study. B, ROC analysis of GPR39, LYPD1 on each stage of ovarian cancer (omentum papillary serous adenocarcinomas (●), papillary serous adenocarcinomas of the ovary (▲), and borderline ovarian cancer (▼)) shows that with a specificity greater than 90%, GPR39, LYPD1 achieves a sensitivity greater than 90% when used to detect ovarian cancer at each stage.
shown to be up-regulated in lung cancer patients (28), but no one has shown a role for WNT7A in ovarian cancer.

Figure 2A shows the expression profile of the three genes that belong to group biomarker “A” across one of the 8 randomly generated matrices. The three genes that belong to this group behave coherently across these ranked conditions. This pattern is conserved across most of the 8 matrices that we randomly generated (Supplementary Fig. S1). This correlation may mean that they respond similarly to the same environmental conditions.

Three other genes (WFDC2, MUC1, and MSLN) labeled as “B” in Table 2 belong to group biomarker “B,” \( Z = 3.7e^{-92} \). They exhibit a coherent behavior across the following ranked conditions: normal ovary, borderline ovarian cancer, and primary ovarian cancer. WAP four-disulfide core domain 2 (WFDC2) encodes for a secreted protein that is expressed in several tumor cells, such as ovarian, colon, breast, lung, and renal (29). WFDC2, also known as HE4, has been shown to be highly up-regulated in ovarian cancer (30, 31). Mucin 1 (MUC1) encodes for a membrane protein that is also secreted. It may play a role in adhesive functions and in cell-cell interactions, metastasis, and signaling (32). MUC1 may provide a protective layer on epithelial surfaces (32). MUC1 has been shown to be highly up-regulated in ovarian cancer (33). Mesothelin (MSLN) encodes for a membrane protein. Its function is unknown, but it may play a role in cell adhesion. It has multiple transcripts due to alternative splicing. MSLN has been shown to be highly up-regulated in ovarian cancer (11, 34–36). Figure 2B shows the expression profile of the three genes that belong to group biomarker “B” across one of the 8 randomly generated matrices. This pattern is conserved across most of the 8 matrices that we randomly generated (data not shown). The genes in this group behave coherently across these ranked conditions.

Finally, the three genes (MSLN, KLK8, and KLK7) labeled as “C” in Table 2 belong to group biomarker “C,” \( Z = 3.7e^{-92} \). They exhibit a coherent behavior across the following ranked conditions: normal ovary, borderline ovarian cancer, and secondary ovarian cancer of the omentum. Kallikrein 8 (neuropsin/ovasin; KLK8) encodes for a secreted protein. KLK8 may be involved in epileptogenesis and hippocampal plasticity. KLK8 has been shown to be highly up-regulated in ovarian cancer (10, 11, 37, 38). Kallikrein 7 (chymotryptic, stratum corneum; KLK7) encodes for a secreted protein. KLK7 is highly up-regulated in ovarian cancer (10, 11) and is present at the apical membrane and in the cytoplasm at the invasive front.

Figure 2C shows the expression profile of the three genes that belong to group biomarker “C” across one of the 8 randomly generated matrices. This pattern is conserved across the 8 matrices that we randomly generated (data not shown).

**Comparison of Group Biomarkers with Other Sets of Biomarkers Obtained with Alternative Statistical Approaches**

We next did statistical analysis and validation of our three group biomarkers on the entire set of gene expression data for the ovary tissue samples. Thus, we analyzed data from the 62 normal ovaries, 7 borderline ovarian cancers, 22 papillary serous adenocarcinomas, and 16 omentum papillary serous adenocarcinoma metastases. We also compared the performance of our three group biomarkers with that of the combinations of the best biomarkers identified using other computational approaches: \( F \) test, ROC approach, and clustering.

ROC plots of group biomarkers “A” (Fig. 3A), “B” (Fig. 3B), and “C” (Fig. 3C) were compared with the

---

Three other genes (WFDC2, MUC1, and MSLN) labeled as “B” in Table 2 belong to group biomarker “B,” \( Z = 3.7e^{-92} \). They exhibit a coherent behavior across the following ranked conditions: normal ovary, borderline ovarian cancer, and primary ovarian cancer. WAP four-disulfide core domain 2 (WFDC2) encodes for a secreted protein that is expressed in several tumor cells, such as ovarian, colon, breast, lung, and renal (29). WFDC2, also known as HE4, has been shown to be highly up-regulated in ovarian cancer (30, 31). Mucin 1 (MUC1) encodes for a membrane protein that is also secreted. It may play a role in adhesive functions and in cell-cell interactions, metastasis, and signaling (32). MUC1 may provide a protective layer on epithelial surfaces (32). MUC1 has been shown to be highly up-regulated in ovarian cancer (33). Mesothelin (MSLN) encodes for a membrane protein. Its function is unknown, but it may play a role in cell adhesion. It has multiple transcripts due to alternative splicing. MSLN has been shown to be highly up-regulated in ovarian cancer (11, 34–36). Figure 2B shows the expression profile of the three genes that belong to group biomarker “B” across one of the 8 randomly generated matrices. This pattern is conserved across most of the 8 matrices that we randomly generated (data not shown). The genes in this group behave coherently across these ranked conditions.

Finally, the three genes (MSLN, KLK8, and KLK7) labeled as “C” in Table 2 belong to group biomarker “C,” \( Z = 3.7e^{-92} \). They exhibit a coherent behavior across the following ranked conditions: normal ovary, borderline ovarian cancer, and secondary ovarian cancer of the omentum. Kallikrein 8 (neuropsin/ovasin; KLK8) encodes for a secreted protein. KLK8 may be involved in epileptogenesis and hippocampal plasticity. KLK8 has been shown to be highly up-regulated in ovarian cancer (10, 11, 37, 38). Kallikrein 7 (chymotryptic, stratum corneum; KLK7) encodes for a secreted protein. KLK7 is highly up-regulated in ovarian cancer (10, 11) and is present at the apical membrane and in the cytoplasm at the invasive front.

Figure 2C shows the expression profile of the three genes that belong to group biomarker “C” across one of the 8 randomly generated matrices. This pattern is conserved across the 8 matrices that we randomly generated (data not shown).

**Comparison of Group Biomarkers with Other Sets of Biomarkers Obtained with Alternative Statistical Approaches**

We next did statistical analysis and validation of our three group biomarkers on the entire set of gene expression data for the ovary tissue samples. Thus, we analyzed data from the 62 normal ovaries, 7 borderline ovarian cancers, 22 papillary serous adenocarcinomas, and 16 omentum papillary serous adenocarcinoma metastases. We also compared the performance of our three group biomarkers with that of the combinations of the best biomarkers identified using other computational approaches: \( F \) test, ROC approach, and clustering.

ROC plots of group biomarkers “A” (Fig. 3A), “B” (Fig. 3B), and “C” (Fig. 3C) were compared with the
A combination of the six best biomarkers identified using the F test: GPR39, KLK8, LAMC2, LCN2, SCNN1A, and TACSTD1 from the borderline data set; BF, CLDN3, CLDN4, GPR39, KLK8, and SCNN1A from the papillary serous adenocarcinoma data set; and CLDN3, CLDN4, GPR39, KLK8, SCNN1A, and WFDC2 from the omentum papillary serous adenocarcinoma data set. The five best genes identified using the Eisen clustering approach were CDH6, DDR1, GPR39, KLK8, LAMC2, and LCN2 from the borderline dataset; CDH6, CLDN3, GPR39, KLK8, MUC1, and WFDC2 from the papillary serous adenocarcinoma data set; and CLDN3, CLDN4, GPR39, KLK8, SCNN1A, and WFDC2 from the omentum papillary serous adenocarcinoma data set. The ROC approach was DDR1, ITGB4, KLK8, LCN2, and WNT7A from the borderline data set; DDR1, KLK8, MSLN, MUC1, and WFDC2 from the papillary serous adenocarcinoma data set; and CD47, CLDN3, KLK7, KLK8, and MSLN from the omentum papillary serous adenocarcinoma data set.

With specificity greater than 99%, each of our three group biomarkers achieved 100% sensitivity, with accuracy greater than 99% (Fig. 3). Thus, our group biomarkers outperformed the combination of the best biomarkers identified using the other computational techniques.

**Conclusion**

In this study, we applied a novel set of biclustering algorithms and a ROC approach on well-defined gene expression data representing ovarian cancer, normal ovary, and nonovarian healthy and diseased tissues samples. We identified many significant patterns that encode for secreted proteins, membrane proteins, and/or extracellular matrix proteins that clearly discriminate between the gene expression data of ovarian cancer, normal ovary, and nonovarian tissues.

The advantage of using a biclustering algorithm is that it allows grouping together of subsets of genes that exhibit the same behavior across subsets of tissue samples. Therefore, the genes that belong to the same bicluster likely have similar responses to the same environmental condition. Thus, a biclustering algorithm approach will give more clinical and biological insight into the tissue samples analyzed and potential biomarkers uncovered.

A major difference between our ROC approach and other computational techniques based on ROC curves is that our

---

**Figure 3.** ROC curves comparison of group biomarkers “A” (A), “B” (B), and “C” (C), with the best genes uncovered using other computational techniques: F test (●), ROC approach (◆), and Eisen clustering (▲).

---


---

Mol Cancer Ther 2008;7(1). January 2008

Downloaded from mct.aacrjournals.org on June 21, 2017. © 2008 American Association for Cancer Research.
Another computational techniques only do a classification based on a comparison between healthy ovary and ovarian cancer tissue samples and do not account for other tissues in the body that may produce the same protein as the ovarian cancer tissue. Therefore, these approaches will result in less specific biomarkers than ours in a diagnostic blood test. The advantages of our ROC approach are 2-fold. A given gene with a high specificity corresponding to a high sensitivity will not only indicate that it is minimally or not expressed in normal ovary tissues and nonovarian tissues but also indicate that it is highly expressed in ovarian cancer tissues. Therefore, it will represent a highly specific and sensitive single biomarker for ovarian cancer detection using a blood test.

This study used the novel approach of group biomarkers as an alternative to the traditional single biomarkers or other combinations of biomarkers used to date for the detection of ovarian cancer using blood tests. Statistical analysis of the potential group biomarkers identified in this study showed that they outperform the combination of the best biomarkers identified using other computational approaches. We believe that our approach outperforms other computational techniques because there exists a correlation or coregulation among the genes that belong to the group biomarkers that we identified. In contrast, other techniques combined potential biomarkers without checking to see whether they are correlated or not.

Interestingly, our group biomarkers contain fewer genes (that is, maximum of three genes per group) and they do better than the combination of the best biomarkers identified by previous approaches, which contain more genes (that is, a minimum of five genes per group). Thus, our methodology identifies an optimum combination of genes that have the highest effect on the diagnosis of a disease. This suggests that the number of genes in a group biomarker is irrelevant, but how they behave together as a group is very important.

We statistically validated the group biomarkers identified in this study using publicly available gene expression data downloaded from a NIH Web site. Because the genes that we identified in this study encode for secreted proteins, they have the potential to be used as tumor markers for the detection of ovarian cancer in a diagnostic blood test. However, additional clinical studies assessing serum levels of the identified putative biomarkers are required to confirm their usefulness in the diagnosis and/or monitoring of ovarian cancer.

Acknowledgments

We thank Gene Logic for providing the gene expression data and Diane Rauch and Sarah Bowell for procuring the tissue samples (University of Minnesota Cancer Center Tissue Procurement Facility).

References

Early detection of ovarian cancer using group biomarkers


Updated version Access the most recent version of this article at:
doi:10.1158/1535-7163.MCT-07-0565

Cited articles This article cites 34 articles, 18 of which you can access for free at:
http://mct.aacrjournals.org/content/7/1/27.full.html#ref-list-1

Citing articles This article has been cited by 1 HighWire-hosted articles. Access the articles at:
/content/7/1/27.full.html#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, contact the AACR Publications Department at permissions@aacr.org.