2011

A Comparative Analysis of Gene-Expression Data of Multiple Cancer Types

Kun Xu
University of Georgia

Juan Cui
University of Georgia, jcui@unl.edu

Victor Olman
University of Georgia

Qing Yang
Jilin University

David Puett
University of Georgia

See next page for additional authors

Follow this and additional works at: http://digitalcommons.unl.edu/csearticles

Xu, Kun; Cui, Juan; Olman, Victor; Yang, Qing; Puett, David; and Xu, Ying, "A Comparative Analysis of Gene-Expression Data of Multiple Cancer Types" (2011). CSE Journal Articles. 169.
http://digitalcommons.unl.edu/csearticles/169

This Article is brought to you for free and open access by the Computer Science and Engineering, Department of at DigitalCommons@University of Nebraska - Lincoln. It has been accepted for inclusion in CSE Journal Articles by an authorized administrator of DigitalCommons@University of Nebraska - Lincoln.
Authors
Kun Xu, Juan Cui, Victor Olman, Qing Yang, David Puett, and Ying Xu
A Comparative Analysis of Gene-Expression Data of Multiple Cancer Types

Kun Xu1,2,*, Juan Cui1,*, Victor Olman1, Qing Yang3, David Puett1, Ying Xu1,3,4*

1 Department of Biochemistry and Molecular Biology, Institute of Bioinformatics, University of Georgia, Athens, Georgia, United States of America, 2 Department of Statistics, University of Georgia, Athens, Georgia, United States of America, 3 Joint Center for Systems Biology, Jilin University, Changchun, China, 4 College of Computer Science and Technology, Jilin University, Changchun, China

Abstract

A comparative study of public gene-expression data of seven types of cancers (breast, colon, kidney, lung, pancreatic, prostate and stomach cancers) was conducted with the aim of deriving marker genes, along with associated pathways, that are either common to multiple types of cancers or specific to individual cancers. The analysis results indicate that (a) each of the seven cancer types can be distinguished from its corresponding control tissue based on the expression patterns of a small number of genes, e.g., 2, 3 or 4; (b) the expression patterns of some genes can distinguish multiple cancer types from their corresponding control tissues, potentially serving as general markers for all or some groups of cancers; (c) the proteins encoded by some of these genes are predicted to be blood secretory, thus providing potential cancer markers in blood; (d) the numbers of differentially expressed genes across different cancer types in comparison with their control tissues correlate well with the five-year survival rates associated with the individual cancers; and (e) some metabolic and signaling pathways are abnormally activated or deactivated across all cancer types, while other pathways are more specific to certain cancers or groups of cancers. The novel findings of this study offer considerable insight into these seven cancer types and have the potential to provide exciting new directions for diagnostic and therapeutic development.

Introduction

Cancer is a key threat to people’s health and life, accounting for ~13% of all disease-causing deaths in the world [1]. In 2007, 7.6 million people died of cancer worldwide. In the U.S, over 1.4 million new cancer cases were reported each year in the past few years, and cancer becomes the second leading cause of death following heart disease. Statistics from the SEER reports indicate that the mortality rate across all cancer types in the U.S. went from 195.4 per 100,000 cases in 1950, continued an upward trend till 1978 reaching 204.4, and then steadily decreased to 184.0 in 2005 [2]. This decreasing trend has been mostly due to the improved diagnostic techniques for detecting the early stage of cancer. General survival statistics of cancer indicate that early detection and treatment are the key to longer survival across all cancer types.

Challenges in early cancer detection arise mainly from the reality that most patients are asymptomatic in the early stages of cancer, and only a few effective cancer-screening tests are clinically available. While some tests have proved to be effective in detecting cancer at its early stage, they are often too invasive, such as colonoscopy, to be routinely used during regular physicals and are currently limited to only a small number of cancer types. Often a cancer is already in an advanced stage when diagnosed; clearly, more effective techniques for early cancer detection are needed.

A number of genetic markers have been proposed for various cancers, such as BRCA1 and BRCA2 for breast cancer and CDH1 for gastric cancer. In addition, a number of promising serum markers for cancer have been used clinically. Among them, PSA (prostate-specific antigen) is the most well known and has been widely used for diagnosing prostate cancer through blood tests [3]. However, its effectiveness of detection is far from adequate, widely considered as having a false positive rate that is too high to be a reliable cancer-indicator [4]. Similar observations have been made about other serum markers such as CA125 for ovarian cancer [5].

Herein we present a computational study on prediction of both genetic and serum markers for seven cancer types, based on public microarray gene-expression data and a computer program for prediction of blood-secretory proteins [6]. Compared to earlier studies on cancer marker identification, including meta-analyses on multi-types of cancers [7], the present study has the following unique features: (i) a focus on identification of multi-gene markers through exhaustive analysis of all possible combinations of genes, taking full advantage of the available high-level computing power, rather than using heuristic approaches that may not necessarily find the optimal markers; (ii) an attempt to find markers for groups of cancers in addition to those for individual cancers; (iii) an attempt to link the information derived from transcriptomic data of tissues to marker prediction in serum using the novel prediction...
program [6]; and (iv) identification of pathways that are abnormally regulated, either common across multiple cancer types or specific to individual cancer types. We believe that these novel data will prove highly valuable in elucidating the genetic alterations in various cancers, as well as offering potential directions for new approaches in diagnostics and therapeutics.

**Materials and Methods**

1. Microarray gene expression data for human cancers

   Microarray gene expression data were downloaded for seven cancer types, namely, breast, colon, kidney, lung, pancreatic, prostate and stomach cancer from the GEO database of NCBI [8]. To ensure that our prediction results can be generalized to different datasets, two independent test sets were used to evaluate the robustness of the predicted gene markers obtained from the training set. Detailed information of the data is listed in Table S1. In this study, we have chosen the largest available microarray datasets from each of the seven cancer types, where each dataset includes the (normalized) gene expression levels of each gene in both cancer and control tissues of each patient, along with the stage information for the majority of the cancer samples (some data does not have this information). Note that all the microarray datasets used are normalized using RMA, which has been reported to be more accurately reflective of biological changes compared to other methods like MAS5 (Affymetrix). The distributions of the fold-changes (FC) of individual genes across all genes between cancer and the corresponding control tissues for the seven types of cancers were checked and found to be highly similar. Figure S1 shows one such comparison of FC distributions between breast cancer and lung cancer; hence we believe that comparisons of fold-changes across different cancer datasets in our study are meaningful.

2. Identification of differentially expressed genes

   For datasets with unpaired cancer and control samples from the same patients, Mann-Whitney test was applied to identify genes that are differentially expressed in cancer versus control samples. For those datasets with paired information the test is as follows: Given the hypothesis $H_0$ that a particular gene is not differentially expressed in cancer versus the control group, the rejection of this hypothesis means that the gene is differentially expressed in cancer. Let $N[i]$ and $C[i]$, be the gene’s expression levels in control and cancer tissues of $i$-th patient, $i=1,…,m$ and $m$ be the number of patients. It is obvious that if the hypothesis $H_0$ is true, then the probability $P(N[i] > C[i]) = P(N[i] < C[i]) = 0.5$, assuming the gene’s expression is a continuous random variable. Let $K$ be the number of patients with $N[i] > C[i] > 0.5$, then the random variable $K/m$ approximately follows a normal distribution (according to the Central Limit Theorem or de Moivre-Laplace Theorem) with its mean $= 0.5$ and a standard variation $= \sqrt{0.5}/\sqrt{m}$, or $X = 2K / \sqrt{m}$ follows a normal distribution $\mathcal{N}(0,1)$. Thus the $p$-value can be estimated as $P(X > \sqrt{2K}/\sqrt{m})$, where $K_{exp}$ is the number of patients satisfying $P(N[i] < C[i])$. Overall, we consider a gene being differentially expressed if the statistic significance, $p$-value, is less than 0.05 and its fold-change is at least 2.

3. Prediction of blood secreted proteins

   All genes predicted to be differentially expressed between cancer and the corresponding control samples were analyzed to predict if their proteins are blood-secretory, using a program that our group recently developed [6]. The basic idea of the algorithm is to train a support vector machine (SVM)-based classifier to distinguish between the blood-secretory proteins and proteins that are not secreted, using various sequence-based features such as signal peptides, transmembrane domains, glycosylation sites and polarity measures. On a large independent test set containing 105 secretory proteins and 7,250 non-secretory proteins of humans, the classifier achieved ~94% prediction sensitivity and ~90% prediction specificity.

4. Prediction of marker genes for each cancer type

   For each $k$-gene combination out of the differentially expressed genes defined in the above section, an SVM-based classifier was trained to achieve the highest possible classification accuracy defined as

   $\text{Overall accuracy} = \frac{(TP + TN)}{N},$

   where $TP$ and $TN$ are the numbers of true positives and negatives, respectively, and $N$ is the total number of samples. A linear kernel function was used for training through LIBSVM [9]. For each cancer type, all markers were ranked according to the 5-fold cross-validation performance on the training dataset. In order to find markers that are generalized well to other datasets, we tested the predicted gene markers on two independent test datasets.

5. Prediction of markers for multiple cancer types

   To identify $k$-gene discriminators for multiple cancer types, all genes that consistently exhibit differential expressions in at least two cancer types were considered. For each $k$-gene combination among these genes, its classification accuracy between each cancer type and the corresponding control tissues was calculated. Then, the $k$-gene combinations exhibiting discerning power across multiple cancer types were determined. The top discriminators for multi-cancer types were selected by using a fixed cut-off on classification accuracies. Throughout the remainder of this paper, $k$-gene groups refer to combinations of $k$-genes for $k = 1, 2, 3, 4$ unless stated otherwise.

6. Pathway enrichment analysis of differentially expressed genes

   Functional analysis and pathway enrichment analysis were conducted using DAVID [10], where the pathway information is based on the annotation from KEGG, BBID and BIOCARTA. A $p$-value < 0.05 was used to guarantee the significance level of an enriched pathway.

**Results**

This study is focused on seven of the most prevalent cancer types in the world, which also have large sets of microarray gene-expression data available in the public domain, collected at a genome scale from tissues of each cancer type as well as from their corresponding noncancerous control tissues. By working on multiple cancer types simultaneously, we can derive potential markers either specific to individual cancer types or general to all or groups of cancers, as well as to identify abnormally activated or deactivated pathways.

1. Predicted marker genes for individual cancer types

   We have searched for individual genes and gene combinations whose expression patterns can best distinguish between cancer and associated control tissues for each cancer types. Specifically, all 1-, 2-, 3- and 4-gene combinations encoded in the human genome were ranked in terms of their discriminating power in distinguishing the cancer samples from the corresponding control samples for
each cancer type. In addition, we have also ranked $k$-gene combinations, based on their discerning power between early cancer samples and control samples if the relevant data are available and sufficiently large.

**A. Breast cancer.** The analysis was done on a gene-expression dataset consisting of 43 paired breast cancer and cancer-adjacent control tissues from the same patients [11]. Of the 43 samples, 32 were early-stage cancers (stages I and II). 294 genes were found to be consistently and abnormally expressed with at least a 2-fold change in their expression across the cancer and the control tissues, 81 of which were up-regulated and 213 were down-regulated in the cancer tissues. Among the differentially expressed genes, 69 of their encoded proteins are predicted to be blood secretory by our prediction program [6], and could thus serve as potential serum biomarkers (Supplementary Information File S1).

Classification analysis was then conducted (see Materials and Methods), with the goal of identifying $k$-gene combinations whose expression patterns can accurately distinguish between the cancer and the control samples. Figure 1 (A) and (D) show the classification accuracies of the best 100 $k$-gene combinations on the whole training set and on the training set containing only early stage samples, respectively. Two independent evaluation sets are used to assess the generality of the identified gene markers, which consist of 31 and 68 breast cancer, and 27 and 61 control samples [12], respectively. Figure 1 (B) and (C) show the classification performance by the trained classifiers on the two evaluation sets. The detailed list of these 100 $k$-gene combinations is given in Supplementary Information S1.

As shown in Figure 1, the majority of the top $k$-gene combinations, particularly for $k \geq 1$, perform well on both training
and the independent test sets with overall accuracy better than 85% although their ranking orders on the two datasets may not be well preserved. The fluctuations in their classification accuracies are believed to be due to the small size of the training data. Similar observations were made on all the predicted top markers across the seven cancer types.

The best three single gene discriminators are PCOLCE2, ANGPTL4 and LEP, having 86.4%, 80.4% and 87.2% classification accuracy on the training set and 94.8% and 84.1%, 84.5% and 79.5% and 96.6% and 98.1% on the two test sets, respectively. The top three 2-, 3- and 4-gene combinations are \{TACSTD2+CHRDL1, TACSTD2+Cavin1, PPARG+TMEM97\}, \{RM2+COL1A1+PPARG, RM2+COL1A1+PCOLCE2, RM2+GPR109B+SPINT2\}, and \{RM2+COL1A1+GPR109B+SPINT2, RM2+GPR109B+INHBA+SPINT2, TACSTD2+HGF+BPI+GF1+TF\}, respectively. Similarly, for early breast cancer, the best three k-gene discriminators are \{GPR109B, PCOLCE2, PCSK5\}, \{PCSK5+COL10A1, FERMT2+SPINT2, MAOA+HGF\}, \{COL1A1+PCSK5+TF, GPX3+COL1A1+SPINT2, GPX3+FAP+TMEM97\}, and \{RM2+COL1A1+GPR109B+HGF, RM2+COL1A1+GPR109B+HGF\}, respectively.

Although the best three discriminators represent novel discoveries, we noticed some lower-ranked genes have been classified as breast cancer markers by previous studies. For example, ADIPQ (adiponectin) is found to be closely associated with a breast-cancer risk [13]. The SPINT2, an inhibitor of HGF activator, was reported to have higher expression levels in early stage breast cancer and associated with a poor prognosis [14], consistent with our findings. Some others are involved in the activities of cancer cells in general. For example, CAV1, down-regulated in the cancer samples, was found to inhibit breast cancer growth and metastasis [15]; the down-regulation of PPARG is associated with local recurrence and metastasis in breast cancer [16]; and ANGPTL4 may act as a regulator of angiogenesis [17]. To the best of our knowledge, all the 2-, 3- and 4-gene discriminators represent novel discoveries.

Similar analyses have been carried out on six other cancer types. The key findings on each of these six cancer types are highlighted below, with the summary being given in Table S2 and gene names listed in Supplementary Information File S1. In addition, Supplementary Information File S2 show the classification accuracies by the best 100 k-gene discriminators on both the training and the testing sets for each cancer type, respectively.

### B. Colon cancer

Our analysis was done on a microarray dataset consisting of 53 colon cancer and 28 cancer-adjacent control tissues from the same patients (some of the cancer samples have no reference samples) [18]. 247 genes were found to be consistently and abnormally expressed with at least a 2-fold change in their expression across the cancer and the control tissues in our training data, 56 of which are up-regulated and 191 are down-regulated in colon cancer tissues. Two independent test sets, consisting of 24 and 22 colon cancer and 24 and 20 cancer-adjacent control samples from the same patients [19], respectively, were used to assess the generality of the predicted markers.

We found the best three single-gene discriminators for colon cancer are MMP7, DPT and MMP1 having 97.5%, 96.3% and 95.1% classification accuracy on the training set, and 97.9% and 90.9%, 97.9% and 74.6%, and 91.7% and 84.1% on the two test sets, respectively. The top three 2-gene discriminators are SLIT3+MMP7, MATN2+MMP7, and MMP7+PTGS1. Some of our top discriminators have been previously studied in the context of colorectal cancer. For example, MMP1 is an invasion-promoting factor, and its up-regulation, as observed in our data, is associated with the invasiveness of the cancer [20]. MMP7 is known to play an important role in cancer growth, and its up-regulation could be a key mechanism for cancer cells’ escape from the immune surveillance [21].

### C. Kidney cancer

The analysis was carried on a microarray gene-expression dataset consisting of 49 kidney cancer and 23 cancer-adjacent control tissue samples from the same patients [22]. 231 genes were found to be consistently and abnormally expressed with at least a 2-fold change in their expression across the cancer and control tissues in our training data, 129 of which are up-regulated and 102 are down-regulated in cancer. Two independent evaluation sets, consisting of 35 and 36 kidney cancer samples and 12 and 9 cancer-adjacent control samples from the same patients, respectively, were used to assess the generality of the predicted markers [23,24]. The best three single gene discriminators are found to be UMOD, ACPP and CCL18 for kidney cancer, having the same classification accuracy, 98.6% on the training set and 100% and 94.4%, 95.7% and 86.11% and 89.4% and 68.1% on the two test sets, respectively. The top three 2-gene combinations are EGF+ALB, ACPP+UMOD, and UMOD+ALB. Among the top discriminators, UMOD has been reported to be related to kidney disease [25]. SERPINA5, down-regulated in the cancer, regulates the invasive potential of renal cancer growth and invasion. Other top discriminators represent new discoveries. For example, AFM has not been reported to be related to cancer, and C6orf155 does not have a characterized function.

### D. Lung cancer

The analysis was done on a microarray dataset consisting of 58 lung cancer tissue and 49 cancer-adjacent control tissue samples from the same patients [26]. 683 genes were found to be consistently and abnormally expressed with at least a 2-fold change in their expression across the cancer and control tissues in our training data, 255 of which are up-regulated and 428 are down-regulated in lung cancer tissues. Two independent sets, consisting of 27 and 20 lung cancer and 27 and 19 cancer-adjacent control samples from the same patients [27], was used to assess the generality of the predicted markers.

The best three single gene discriminators are CAV1, SFTPC and VWF for lung cancer, having the same classification accuracy, 99.1% on the training set and 98.2% and 100%, 96.3% and 82.5%, and 88.9% and 100% on the two test sets, respectively. The top three 2-gene combinations are FERMT2+GREM1, TEK+FASC, CAV1+MMP1. Among the top discriminators, CAV1 has been found to be down-regulated in breast cancer [28], and has been reported to be associated with metastasis in lung cancer [29]. SFTPC has been reported to be associated with interstitial lung disease [30]. FAM107A, which suppresses cell growth, may play a role in cancer development [31]. Other top discriminators represent new observations. For examples, TNXB, SPP1 and EMCN have not previously been reported as cancer-related.

### E. Pancreatic cancer

The analysis was done on a microarray dataset consisting of 39 paired pancreatic cancer and cancer-adjacent control tissue samples from the same patients [32]. 883 genes were found to be consistently and abnormally expressed with at least a 2-fold change in their expression across the cancer and control tissues in the training data, 616 of which are up-regulated and 269 are down-regulated in pancreatic cancer. Two independent sets, consisting of 36 and 29 pancreatic cancer samples and 16 and 5 cancer-adjacent control samples from the same patients [33], was used to assess the generality of the predicted markers.

The best three single-gene discriminators are KRT17, COL10A1 and CTHRC1 for pancreatic cancer, having the same classification accuracy, 93.6% on the training set and 88.5% and
80.4%, 84.6% and 73.2%, and 84.6% and 85.7% on the two test sets, respectively. The top three 2- and 3-gene discriminators are \{MMP7+AZGPI; MMP+FGP1; MMP+PLA2G18\} and \{THRC1+SGPP2+CCL18; TNFRSF21+EGF; CTHRC1; COL10A1+S100A6+RSAD2\}, respectively. Among the top discriminators, KRT17 is known to be involved in tissue repair [34]. AZGPI has been reported to cause extensive loss of fat, often associated with advanced cancers [35]. Other top discriminators represent new findings. For example, RSAD2, involved in antiviral defense, has not been reported as being related to cancer, as well as SGPP2, known to be involved in pro-inflammatory signaling [36], and CST1.

### F. Prostate cancer

The analysis was done on a microarray dataset consisting of 65 prostate cancer and 63 cancer-adjacent control tissue samples from the same patients [37]. 118 genes were found to be consistently and abnormally expressed with at least a 2-fold change in their expression across the cancer and control tissues in our training data, of which 23 are up-regulated and 95 are down-regulated in lung cancer tissues. Two independent sets, consisting of 62 and 53 prostate cancer samples and 47 and 14 cancer-adjacent control samples from the same patients [38], was used to assess the generality of the predicted markers.

The best three single gene discriminators are MYLK, PALLD and CAV1 for prostate cancer, having 73.4%, 71.9% and 71.1% classification accuracy on the training set and 85.3% and 62.3%, 69.6% and 72.6%, and 94.2% and 75.5% on the two test sets, respectively. The top three 2- and 3-gene discriminators are \{LTF+IGF1; LTF+SPARC11; SMTN+CCK\}, \{SMTN+CCK+CCL2; SMTN+CCK+COMP; SMTN+CCK+PLA2G7\}, respectively. Among the top discriminators, LTF is known to inhibit the growth of tumors [39]. IGF1, a growth factor, plays a role in the development of prostate cancer [40] and has been shown to be associated with advanced prostate cancer [41]. Other top discriminators represent new discoveries. For example, CHG1 may play a role in regulating angiogenesis [42] but has not been reported to be related to cancer. The same is with SMTN.

### G. Stomach cancer

The analysis was done on a microarray dataset consisting of 89 stomach cancer and 23 cancer-adjacent control tissue samples from the same patients [43]. Out of the 89 cancer tissue samples, 31 are early-stage cancers. 311 genes were found to be consistently and abnormally expressed with at least a 2-fold change in their expression across the cancer and control tissues in our training data, 166 of which are up-regulated and 145 are down-regulated in lung cancer tissues. Two independent sets, consisting of 38 and 16 stomach cancer samples and 31 and 13 cancer-adjacent control samples from the same patients [44,45] was used to assess the generality of the predicted markers, of which 12 are early stage samples partially paired with 10 control samples.

The best three single-gene discriminators are SERPINH1, BGN and COL12A1 for stomach cancer, having 99.1%, 98.2% and 98.2% classification accuracy on the training set and 94.2% and 96.7%, 88.4% and 93.3%, and 84.1% and 75.8% on the two test sets, respectively. The top three 2-gene combinations are \{CHGA+SERPINH1; TGFB1+CHGA and PGC+SERPINH1\}, respectively. For early stomach cancer, the best three 3-gene discriminators are also SERPINH1, BGN and COL12A1, respectively. Among the top discriminators, BGN is known to have a role in controlling cell growth in cancer [46]. The abnormal expression of CTHRC1, a regulator of matrix deposition, has been widely found across different solid cancers and is considered to be associated with cancer invasion and metastasis [34]. Of particular interest is that PGC has been proposed as an indicator of gastric cancer [47], and the serum level of PGC was used as a biomarker for precancerous lesions of the stomach [48]. Other top discriminators represent new discoveries. For example, ABCA5, ADAMTS12 and CLEC3B have not been reported to be cancer related.

Interestingly, the number of differentially expressed genes across different cancer types has a wide spread, ranging from 118 (prostate), 231 (kidney), 247 (colon), 294 (breast), 311 (stomach) to 683 (lung) and 885 (pancreatic). One possible explanation is that these numbers may reflect the aggressiveness of the corresponding cancers. We did notice that there is strong correlation between the number of differentially expressed genes in a given cancer type and the five-year survival rate of patients with that cancer [49] (Figure 2). The detailed statistics is given in Table S3. Another interesting observation is that, while the majority of the differentially expressed genes with at least a 2-fold change in five cancer types (breast, colon, lung, prostate, stomach) are down-regulated, in kidney and pancreatic cancers, the majority of such genes are up-regulated, possibly suggesting unique characteristics of these two cancer types.

### 2. Markers for multiple cancer types

We have also sought to identify genes that could be used as indicators for cancer in general or for a group of cancers. It is possible to find common gene ‘‘markers’’ across different cancer types because of the observation that the majority of the cancers, if not all, undergo a common set of alterations [50] during oncogenesis, such as self-sufficiency in growth signals, insensitivity to antigrowth signals, evasion of apoptosis, and tissue invasion and metastasis. Some of these biological processes may be executed by the same groups of proteins during the formation and progression of different cancers, hence possibly giving rise to common markers for different cancer types.

#### A. Identification of genes differentially expressed across multiple cancer types

We have examined differentially expressed genes with at least 2-fold changes between cancer and corresponding control tissues across all seven cancer types and attempted to find those genes common to multiple cancer types. The key findings are summarized in Table 1.

85 genes are found to be differentially expressed across at least three cancer types (Table S4), among which 19 genes are across at least four cancer types, and five genes (ABCA8, DPT, FHL,CDC2 and TOP2A) across five cancer types. The differences in the gene expression across different cancer types may indicate either a general or specific relevance of the gene to the corresponding

![Figure 2. Correlation between 5-year survival rate and the number of differentially expressed genes in each cancer type.](https://example.com/figure2.png)

doi:10.1371/journal.pone.0013696.g002
Table 1. A list of 19 genes that are differentially expressed in more than 4 cancer types and their relevance to different cancer types.

<table>
<thead>
<tr>
<th>Gene ID</th>
<th>Direction of regulation</th>
<th>Reported to be related to cancers</th>
</tr>
</thead>
<tbody>
<tr>
<td>CDC2</td>
<td>↑</td>
<td>↑</td>
</tr>
<tr>
<td>AURKA</td>
<td>↑</td>
<td>↑</td>
</tr>
<tr>
<td>ABCA8</td>
<td>↓</td>
<td>↓</td>
</tr>
<tr>
<td>DPT</td>
<td>↓</td>
<td>↓</td>
</tr>
<tr>
<td>TOP2A</td>
<td>↑</td>
<td>↑</td>
</tr>
<tr>
<td>MMP7</td>
<td>↑</td>
<td>↑</td>
</tr>
<tr>
<td>MAD2L1</td>
<td>↑</td>
<td>↑</td>
</tr>
<tr>
<td>KLF4</td>
<td>↓</td>
<td>↓</td>
</tr>
<tr>
<td>MELK</td>
<td>↑</td>
<td>↑</td>
</tr>
<tr>
<td>C7</td>
<td>↓</td>
<td>↓</td>
</tr>
<tr>
<td>ECT2</td>
<td>↑</td>
<td>↑</td>
</tr>
<tr>
<td>PRC1</td>
<td>↑</td>
<td>↑</td>
</tr>
<tr>
<td>RRM2</td>
<td>↑</td>
<td>↑</td>
</tr>
<tr>
<td>ALDH1A1</td>
<td>↓</td>
<td>↓</td>
</tr>
<tr>
<td>PMAIP1</td>
<td>↑</td>
<td>↑</td>
</tr>
<tr>
<td>FABP4</td>
<td>↓</td>
<td>↓</td>
</tr>
<tr>
<td>COL11A1</td>
<td>↑</td>
<td>↑</td>
</tr>
<tr>
<td>TTK</td>
<td>↑</td>
<td>↑</td>
</tr>
<tr>
<td>CENPF</td>
<td>↑</td>
<td>↑</td>
</tr>
</tbody>
</table>

* ↑ indicates up-regulated gene expression in the corresponding cancer type while ↓ is down-regulation. ** indicates that a gene has been reported as relevant to the corresponding cancer type. "B." for breast cancer; "C." for colon cancer; "K." for kidney cancer; "L." for lung cancer; "Pa." for pancreatic cancer; "Pr." for prostate cancer and "S." for stomach cancer.

doi:10.1371/journal.pone.0013696.t001

B. Pathway enrichment analysis of differentially expressed genes. We have carried out a pathway enrichment analysis on genes that are differentially expressed in any of the seven cancer types. Overall, a number of signaling pathways are consistently and highly enriched across all seven types of cancers, such as Wnt, p33 and integrin signaling pathways, as well as a few other processes like phospho-APC/C-mediated degradation of cyclin A and inflammation determined by chemokine and cytokine signaling pathways (in addition to the general cellular processes such as cell cycle, DNA replication and repair, apoptosis and various metabolic pathways). Notably, these pathways are mostly enriched with up-regulated genes in cancer, suggesting a possible activation of these processes. In addition, a few metabolic pathways such as tyrosine, histidine, phenylalanine, butanoate and 5-hydroxytryptamine pathways are enriched only with down-regulated genes across all cancers. This may indicate a possible deficiency of the relevant metabolic enzymes in cancer, which could, for example, arise from loss-of-function mutations in their genes. These observations may suggest the essential roles played by these processes in cancer formation and progression.

Other than the above processes common to all cancers, a few pathways are enriched only in specific cancers. For example, arginine, proline, glutamate and riboflavin (vitamin B2) metabolisms are enriched with up-regulated genes only in lung cancer; folate biosynthesis and nitrogen metabolism pathways are...
enriched in breast cancer; formyltetrahydroformate biosynthesis in stomach cancer; and NF-kB activation and Csk activation by cAMP-dependent protein kinase inhibits signaling through T-cell receptor in kidney cancer. Of particular interest is the finding that pancreatic cancer has the greatest number of differentially expressed genes, compared to other cancer types, that are involved in a complex network consisting of the EGF signaling pathway, purine and aminosugar metabolism, PKC-catalyzed phosphorylation of inhibitory phosphoprotein of myosin phosphatase, metabotropic glutamate receptor group II pathway, Fc epsilon receptor I signaling and the BCR and IL 4 signaling pathways. This suggests a highly active state of the underlying pathways. This suggests a highly active state of the underlying 

carcinogenesis of different cancers.

Table 2 gives the top two 2-gene combinations across at least three cancer types. CDC2+DPT and CDC2+TOP2A are found to be good discriminators for five cancer types, namely breast, colon, lung, prostate and stomach cancers. Similarly, ABCA8+ALD-H1A1+DPT and ABCA8+AURKA+DPT are good 3-gene discriminators for four cancer types with higher classification accuracies than the top 2-gene discriminators, as shown in Table S6.

As noted, CDC has been reported to play a key role in cell proliferation and apoptosis [59], and DPT is suggested to have a possible role in carcinogenesis through its interaction with a known oncoprotein, TGFβ1. Moreover, some of the top discriminator genes have been reported to be cancer relevant. For example, ECT2 is reported to be involved in cancer development, influencing processes such as the cell cycle, apoptosis and cell division [60]; FABP4 is involved in the activation of the immune response and is reported to be related to breast cancer [61] and bladder cancer [62]; and TOP2A is involved in stomach cancer [51]. These independent observations confirm that the findings herein are meaningful.

D. Top k-gene markers that are blood secretory. By combining our blood-secretion prediction capability [6] with the above top gene discriminators, we have predicted proteins that may be secreted into circulation, thus providing candidate serum marker proteins for cancer detection. Table 3 summarizes the top k-gene discriminators that are predicted to have their proteins secreted into blood. Some genes involved in these top candidate discriminators have been previously reported to be cancer related, e.g. MMP7 [63]. Other predicted blood-secretory marker proteins such as PAICS, CHRDL1, KLFL2, COL10A1 and MYL9 have not heretofore been reported to be cancer related.

While Table 3 gives a detailed list of all the gene combinations whose proteins are predicted to be blood secretory, with
concerning potential serum biomarkers of specific cancers. Our analysis
of abnormally activated or deactivated pathways across multiple
cancer types, as well as expediting the search for
ing pathways in cancer, as well as expediting the search for
potential serum biomarkers of specific cancers.

Concluding Remarks
A computational protocol for predicting gene markers in cancer
tissues and protein markers in sera was developed for simultaneous
analyses of multiple cancer types. In addition to individual gene
markers, we have focused on gene combinations that can be used
to distinguish multiple cancer types from their corresponding
control tissues. The pathway enrichment analysis among the
differentially expressed genes across multiple cancer types, as well
as those specific to individual cancer types, has identified a number
of abnormally activated or deactivated pathways across multiple
cancers and for specific cancers. The information provided on
individual genes and pathways, along with potential serum
biomarkers, should provide highly useful information for elucidat-
ing pathways in cancer, as well as expediting the search for
potential serum biomarkers of specific cancers.

Supporting Information
File S1 Links to the supplementary tables for the top k-gene
markers for 7 individual cancer types.

Table 3. Top k-gene discriminators with their proteins being blood secretory.

<table>
<thead>
<tr>
<th>Markers</th>
<th>Breast</th>
<th>Colon</th>
<th>Kidney</th>
<th>Lung</th>
<th>Pancreas</th>
<th>Prostate</th>
<th>Stomach</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>train</td>
<td>test2</td>
<td>train</td>
<td>test2</td>
<td>train</td>
<td>test2</td>
<td>train</td>
</tr>
<tr>
<td>3 GREM1+MMP7</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>88.9</td>
<td>79.6</td>
<td>64.1</td>
</tr>
<tr>
<td>3 MMP7+MMP9</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>75.9</td>
<td>79.6</td>
<td>87.2</td>
</tr>
<tr>
<td>3 MMP11+MMP7+MMP9+RRM2</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>85.2</td>
<td>96.3</td>
<td>87.2</td>
</tr>
<tr>
<td>3 CCL18+TGFBI</td>
<td>-</td>
<td>-</td>
<td>74.5</td>
<td>80.9</td>
<td>80.0</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>3 DPT+MMP9</td>
<td>-</td>
<td>-</td>
<td>85.2</td>
<td>88.9</td>
<td>87.2</td>
<td>88.9</td>
<td>87.2</td>
</tr>
<tr>
<td>3 FAM107A+KLF4</td>
<td>-</td>
<td>-</td>
<td>97.9</td>
<td>89.6</td>
<td>76.2</td>
<td>84.2</td>
<td>81.2</td>
</tr>
<tr>
<td>3 FAM107A+KLF4+MMP7+PAICS</td>
<td>-</td>
<td>100</td>
<td>100</td>
<td>97.9</td>
<td>89.6</td>
<td>87.2</td>
<td>96.2</td>
</tr>
<tr>
<td>3 INHBA+RRM2</td>
<td>74.1</td>
<td>100</td>
<td>98.4</td>
<td>-</td>
<td>94.4</td>
<td>92.6</td>
<td>94.9</td>
</tr>
<tr>
<td>3 GPX3+RRM2</td>
<td>81.0</td>
<td>96.6</td>
<td>96.1</td>
<td>-</td>
<td>88.9</td>
<td>94.4</td>
<td>94.9</td>
</tr>
<tr>
<td>3 COL11A1+DPT</td>
<td>72.4</td>
<td>96.6</td>
<td>96.9</td>
<td>97.9</td>
<td>89.6</td>
<td>57.1</td>
<td>86.5</td>
</tr>
<tr>
<td>3 MMP11+RRM2</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>88.9</td>
<td>90.7</td>
<td>69.2</td>
</tr>
</tbody>
</table>

Each numerical value represents the classification accuracy (by percentage) between cancer tissues and their corresponding reference tissues.

doi:10.1371/journal.pone.0013696.t003

Figure S1 Comparison of the gene expression fold changes (A)
between breast cancer training and testing datasets, and (B)
between breast cancer and lung cancer.

Found at: doi:10.1371/journal.pone.0013696.s009 (0.30 MB TIF)
References


