Cancer Medicine

Open Access

METHOD

Identification of potential biomarkers from microarray experiments using multiple criteria optimization

Matilde L. Sánchez-Peña[1], Clara E. Isaza[1,2], Jaileene Pérez-Morales[1], Cristina Rodríguez-Padilla[2], José M. Castro[3] & Mauricio Cabrera-Ríos[1]

[1]Bio IE Lab, Industrial Engineering Department, University of Puerto Rico at Mayagüez, Mayagüez, Puerto Rico
[2]Immunology and Virology Laboratory, Universidad Autónoma de Nuevo León, Monterrey, México
[3]Integrated Systems Engineering, The Ohio State University, Columbus, Ohio

Keywords
Cancer biomarkers, cervical cancer, data envelopment analysis, microarray data analysis, multiple criteria optimization

Correspondence
Mauricio Cabrera-Ríos, Bio IE Lab, Industrial Engineering Department, University of Puerto Rico at Mayagüez, Mayaguez, Puerto Rico. Tel: 7878324040, ext 5964; Fax: 7872653820; E-mail: mauricio.cabrera1@upr.edu

Funding Information
This study is based on study supported by the National Science Foundation (NSF) under Grant HRD 0833112 (CREST program), as well as the National Institutes of Health (NIH) MARC Grant 5T36GM095335-02 "Bioinformatics Programs at Minority Institutions." UPRM BioSEI Grant 330103080301 awarded to M. Cabrera-Ríos and PROMEP Grant 103.5/07/2523 awarded to C. Isaza were also critical to the development of this study. M. Sánchez-Peña was supported through a research fellowship in the Department of Industrial Engineering at UPRM.

Received: 23 January 2012; Revised: 23 January 2013; Accepted: 24 January 2013

***Cancer Medicine* 2013; 2(2): 253–265**

doi: 10.1002/cam4.69

Abstract

Microarray experiments are capable of determining the relative expression of tens of thousands of genes simultaneously, thus resulting in very large databases. The analysis of these databases and the extraction of biologically relevant knowledge from them are challenging tasks. The identification of potential cancer biomarker genes is one of the most important aims for microarray analysis and, as such, has been widely targeted in the literature. However, identifying a set of these genes consistently across different experiments, researches, microarray platforms, or cancer types is still an elusive endeavor. Besides the inherent difficulty of the large and nonconstant variability in these experiments and the incommensurability between different microarray technologies, there is the issue of the users having to adjust a series of parameters that significantly affect the outcome of the analyses and that do not have a biological or medical meaning. In this study, the identification of potential cancer biomarkers from microarray data is casted as a multiple criteria optimization (MCO) problem. The efficient solutions to this problem, found here through data envelopment analysis (DEA), are associated to genes that are proposed as potential cancer biomarkers. The method does not require any parameter adjustment by the user, and thus fosters repeatability. The approach also allows the analysis of different microarray experiments, microarray platforms, and cancer types simultaneously. The results include the analysis of three publicly available microarray databases related to cervix cancer. This study points to the feasibility of modeling the selection of potential cancer biomarkers from microarray data as an MCO problem and solve it using DEA. Using MCO entails a new optic to the identification of potential cancer biomarkers as it does not require the definition of a threshold value to establish significance for a particular gene and the selection of a normalization procedure to compare different experiments is no longer necessary.

Introduction

Microarrays are frequently used to simultaneously analyze the expression level of tens of thousands of genes. Analysis of microarray data has become a useful tool for the study of different illnesses including all types of cancer [1–3]. Microarray analyses are carried out, essentially, with the objective to detect variation patterns of genetic expression. In cancer research, these patterns can be used for various purposes such as eliciting a diagnosis or prognosis, characterizing a particular illness stage, or detecting and proposing the role of specific genes in the development of cancer. In this last classification, lies the detection of cancer biomarkers. Because biomarker genes detected using only microarray data are not experimentally validated yet, at that point they are deemed potential biomarkers.

Microarray experiments generate large amounts of information whose analysis and interpretation are nontrivial [4]. Traditional statistical approaches are challenged by large variances, incommensurability, nonnormality, and the small number or replicates frequently present in these experiments. These challenges hamper finding consistent analysis results [5], thereby leading to a large number of potential biomarkers to be investigated, the research of which could prove lengthy and very expensive.

An example that illustrates the difficulties of obtaining cancer biomarkers consistently is the 70-gene signature for identification of patients with a high probability for breast cancer relapse after its eradication. The original results are reported previously [6]. A 76-gene signature is reported in Wang et al. [7] with the same purpose; however, there are only three genes that intersect with the original signature. This issue has been also reported for the specific case of breast cancer by Ein-Dor et al. [8].

It is also notorious that truly integrated work across disciplines is not frequent in most microarray analysis works. Biology and Medicine experts are usually left with the burden of using coded analysis tools with a series of parameters – of statistical, computational, or mathematical nature – that significantly affect the outcome of the software packages [4]. This leads to issues in results' reproducibility and comparability between studies.

These challenges motivate the search for microarray analysis techniques from which consistent results can be achieved across several experiments and researches, particularly for the identification of potential cancer biomarkers. In this study, a multiple criteria optimization (MCO) approach is proposed for the identification of potential cancer biomarkers from microarray data. An MCO problem aims to find the best compromises between two or more conflicting criteria [9]. The best compromises are located in the so-called Pareto-efficient frontier. It is proposed that the genes in the efficient frontier of the MCO problem, built with performance measures relating to the significant change in gene expression, are potential cancer biomarkers.

The potential of an MCO analysis for the identification of relevant genes has been recognized before [10] through the use of ranking methods. Here, the proposed MCO problem is solved through the use of data envelopment analysis (DEA) [11]. DEA has been used to find the convex efficient frontier of MCO problems [12]. DEA is a very computationally convenient technique that is capable to deal with multiple and incommensurable performance measures. A clear applicability to meta-analysis follows from these characteristics. Using MCO provides a new optic to the identification of potential cancer biomarkers as it does not require the definition of a threshold value to establish significance for a particular gene and the selection of a normalization procedure to compare different experiments is no longer necessary.

The proposed method is tested here through its initial application to a microarray database related to cervix cancer [13] and the results are successfully validated through the information available in the literature for the selected genes. Furthermore, two additional studies involving two independent experiments using the same microarray [14, 15] platform further corroborate the performance of the proposed method. Finally, the novelty of this approach is contrasted with the use of a single criterion – or performance measure – to find potential biomarkers.

Methods

Potential biomarkers through MCO

In microarray experiments, it is critical to be able to quantify changes in genetic expression. A series of measurements have been proposed in the literature that include variations of pure magnitude of relative change of expression versus a control [16] as well as *P*-values obtained from various statistical tests [17]. A *P*-value, in statistical comparison procedures, can be understood as the probability associated with finding – by pure chance – a difference in the populations being compared that is at least as large as the observed difference of the samples involved. Lower *P*-values indicate larger differences and therefore show stronger evidence favoring statistical significance. Due to their interpretation capabilities, *P*-values have been a favored performance measure in microarray experiments in recent years. Obtaining a *P*-value for a particular gene is illustrated in Figure 1.

A *P*-value, when obtained for a particular gene in a microarray experiment, can be thought of as a criterion to be minimized since the smaller the *P*-value the more important the change in expression of the gene under consideration. Now, if more than one *P*-value is available for a particular gene, then the task at hand is one of multiple criteria minimization. An illustrative example with a series of genes is shown in Figure 2. In this figure, each gene is represented by a pair of *P*-values. Because low *P*-values are attractive, the ideal gene would be found in the southwest corner of the graph. When no single gene is best in all criteria under consideration, a conflict exists.

The key idea in this study is that the potential biomarker genes can be identified as efficient solutions of the MCO problem that results from representing each gene under analysis through a series of associated *P*-values. In order to develop the idea, two issues must be addressed (i) how can one obtain several *P*-values for one gene?

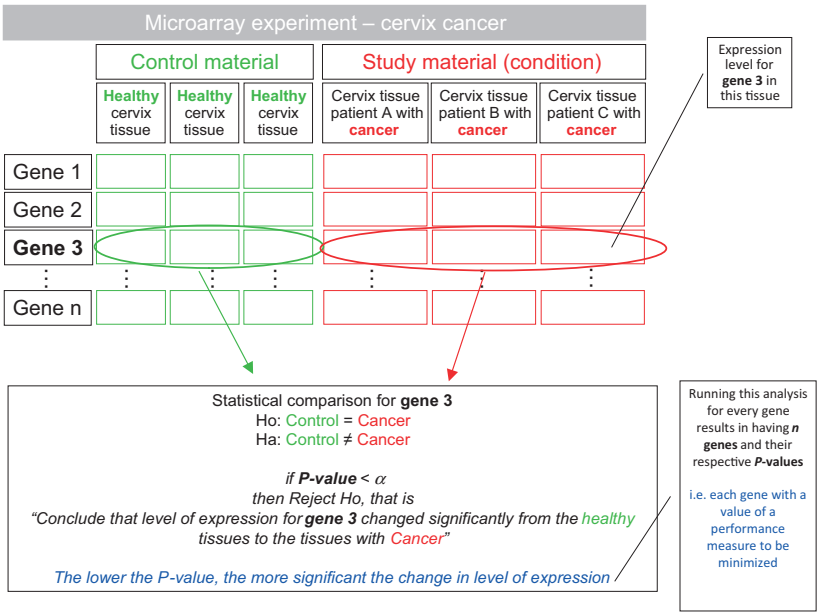

Figure 1. Schematic example of how to obtain a *P*-value. This is a schematic example of how to obtain one *P*-value for a particular gene in a microarray experiment with $l = 3$ healthy tissues as controls and $m = 3$ tissues with cancer. If statistical comparison is carried out for each gene, then at the end one has n genes each one with an associated *P*-value.

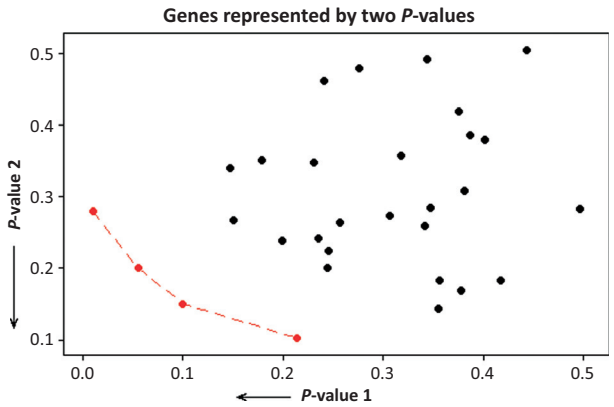

Figure 2. Pareto-efficient frontier. The existence of conflict causes that different genes be attractive when lying in the southwest envelope of the gene set. In general, in multiple criteria optimization (MCO), that envelope is called a Pareto-efficient frontier and it is conformed by Pareto-efficient solutions.

and (ii) which method can be used to solve the MCO problem.

Obtaining multiple *P*-values for a particular gene

Consider the results of a microarray experiment laid out on a table where the first column contains the names of the n genes under study; the columns to the right contain the measurements for l healthy tissues followed by m cancer tissues. Thus, for each gene, there are l replicated measurements of relative expression for state 1 (healthy) and m replicates for state 2 (cancer).

A statistical comparison procedure can be used to obtain a *P*-value when contrasting parameters from the two states – cancer and healthy – for a particular gene. A common interest is to compare the population centers, which are estimated either through sample means or sample medians. For MCO purposes, however, more than one *P*-value per gene is necessary. Two cases can be distinguished here: (c1) having a single microarray experiment to study one type of cancer and (c2) having several microarray experiments to study one type of cancer. In c1, if a leave-one-out strategy is applied to the tissues pertaining to one state, then it is possible to obtain several *P*-values. In c2, an additional *P*-value can be obtained for the genes that are common to both experiments. This study focuses on c1 to introduce the proposed analysis strategy, leaving c2 for future publication.

For c1, the leave-one-tissue-out strategy implies extracting a particular tissue associated with one state ("leaving one column out"). By removing a vector (column), a replicate is deleted from the set, thereby forcing a *P*-value that is different from the original one. Thus, two different *P*-values are effectively created. The selection of the tissue to be removed to create a distinct matrix is performed considering the variance of expression on each tissue (stored in each column). Then, a first matrix is

built leaving out the tissues (columns) with the highest variance for each state and the second matrix by leaving out the tissues with the lowest variance for each state. Through this strategy, the resulting matrices show extreme cases in terms of data variance. Any other combination of tissues to leave out would have statistical differences lying between these two "extreme" cases.

Thus, two extreme cases span all the possible cases in terms of variance for the leave-one-out cases. This fact can be used to avoid unnecessary computational effort and, by using just two dimensions, it is possible to illustrate the problem graphically.

c1 is important because the vast majority of published microarray experiments are instances of this type, and – as explained previously – it is the subject of study in this manuscript. c2 can be built from several c1 instances, however, it is envisioned that this case becomes an archetype for a study designed to keep the same genes throughout all microarrays experiments involved. c2 will also represent the case where meta-analysis must be addressed and will be approached in a future publication.

Solving the MCO problem

The decision that must result from the solution of the MCO problem can be stated as "a selection of those genes that show the highest possible expression change in all experimental instances when considered simultaneously." Due to the large variability encountered in microarray experiments, this is a nontrivial decision that will lead to a set of genes that will have very low *P*-values in certain instances, although not necessarily in all of them, that is, the genes that are Pareto-efficient as illustrated in Figure 2.

DEA is a technique that has been shown capable to identify the efficient solutions located in the convex hull of an MCO problem [11]. In its most popular form, DEA finds the Pareto-efficient solutions through the sequential solution of a series of linear optimization models. One of the most popular and effective DEA formulations is the Banker–Charnes–Cooper model (BCC), which is shown next in its two formulations (input oriented and output oriented):

$$
\begin{aligned}
&\text{Find } \boldsymbol{\mu}, \boldsymbol{\nu}, \mu_0^+, \mu_0^- \text{ to}\\
&\text{Maxmize } \boldsymbol{\mu}^T \mathbf{Y}_0^{\max} + \mu_0^+ - \mu_0^-\\
&\text{Subject to}\\
&\qquad \boldsymbol{\nu}^T \mathbf{Y}_0^{\min} = 1\\
&\boldsymbol{\mu}^T \mathbf{Y}_j^{\max} - \boldsymbol{\nu}^T \mathbf{Y}_j^{\min} + \mu_0^+ - \mu_0^- \le 0 \quad j = 1, \ldots, n\\
&\qquad \boldsymbol{\mu}^T \ge \varepsilon \cdot \mathbf{1}\\
&\qquad \boldsymbol{\nu}^T \ge \varepsilon \cdot \mathbf{1}\\
&\qquad \mu_0^+, \mu_0^- \ge 0
\end{aligned}
$$

$$
\begin{aligned}
&\text{Find } \boldsymbol{\nu}, \boldsymbol{\mu}, \nu_0^+, \nu_0^- \text{ to}\\
&\text{Minimize } \boldsymbol{\nu}^T \mathbf{Y}_0^{\min} + \nu_0^+ - \nu_0^-\\
&\text{Subject to}\\
&\qquad \boldsymbol{\mu}^T \mathbf{Y}_0^{\max} = 1\\
&\boldsymbol{\nu}^T \mathbf{Y}_j^{\min} - \boldsymbol{\mu}^T \mathbf{Y}_j^{\max} + \nu_0^+ - \nu_0^- \ge 0 \quad j = 1, \ldots, n\\
&\qquad \boldsymbol{\nu}^T \ge \varepsilon \cdot \mathbf{1}\\
&\qquad \boldsymbol{\mu}^T \ge \varepsilon \cdot \mathbf{1}\\
&\qquad \nu_0^+, \nu_0^- \ge 0
\end{aligned}
$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ are vectors containing nonnegative multipliers and $\mu_0^+, \mu_0^-, \nu_0^+$ and ν_0^- are scalar numbers to be determined optimally, $\mathbf{Y}_j^{\min}$ and $\mathbf{Y}_j^{\max}$ are vectors containing the values of performance measures to be minimized and maximized, respectively, for the *j*th solution. The subindex 0 is used to denote the solution currently under analysis, and ε is a small constant usually set to a value of 1×10^{-6}. The results of solving these two linear optimization problems, for the n genes in a set, are a series of hyperplanes that forms a convex envelope around this set, as depicted in Figure 2.

Because of the nature of DEA, the model needs at least one performance measure to be maximized. For the case under consideration, a transformation of at least one set of *P*-values is required. The following transformation is applied to switch from minimization to maximization in a set of n *P*-values:

$$
P\text{-value}_i^* = (\text{Max } P\text{-value} + \text{Min } P\text{-value}) - P\text{-value}_i, \quad i = 1 \text{ to } n \qquad (1)
$$

where the transformation is carried out for the *i*th gene. Maximizing the transformed performance measure is fully equivalent to minimizing the original *P*-value.

DEA has several advantages including (i) computational efficiency owing to its linear optimization structure; (ii) objectivity and consistency of results, which follows from not requiring the adjustment of parameters or assigning weights to the different performance measures; and (iii) capability of analyzing several microarray experiments with incommensurate units. Furthermore, linear optimization is – by far – the most coded type of optimization. Algorithms for linear programing (as this type of optimization is known as) are available in modules from the very common MS Excel package to the mathematically oriented software Matlab [18] and to the very specialized solvers like Lingo [19]. There are also DEA solvers like DEA Solver Pro [20] that make adopting the proposed approach even easier. So, in order to use the approach proposed here, all the user needs is a list of genes, with one *P*-value obtained as usual, and a

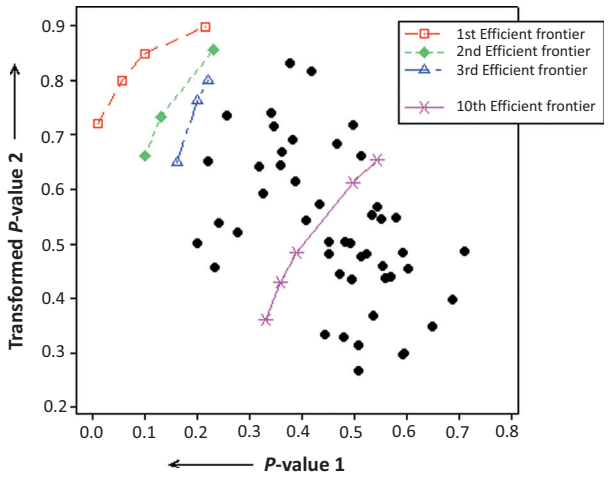

Figure 3. The two performance measures for each gene. This figure schematically shows a case with genes characterized by two performance measures: an untransformed *P*-value and a transformed one with equation (1). Referring to this figure, and following the proposed method, at this point it is recommended to identify the first 10 efficient frontiers. This can be easily done by identifying the genes in the first efficient frontier through data envelopment analysis (DEA), then removing them from the set and continuing with a second DEA iteration. This is repeated until the tenth frontier is identified. A method to determine the number of adequate frontiers to be analyzed is currently under development by our research group.

second *P*-value transformed using equation (1), and an optimization solver capable to deal with linear programming to use the DEA formulations outlined above.

One limitation of DEA is that of depending on a series of local linear approximations, as shown in Figure 2. Every time that a hyperplane is superimposed over the set under analysis, there are genes lying in the nonconvex part of the set frontier that escape detection. These genes could be potential biomarkers, however.

In order to circumvent this limitation, it is proposed that DEA be applied successively 10 times, each time removing the genes found in a particular iteration from the set for subsequent analyses. This strategy results in 10 frontiers, as seen in Figure 3.

Results

Analysis of a single microarray experiment to study one type of cancer

The first results on the application of the proposed method include the analysis of the microarray database used by Wong et al. [13] related to cervix cancer. The database consists of eight healthy tissues and 25 cervix cancer tissues, all of them with expression level readings for 10,692 genes from a cDNA microarray. The Mann–Whitney nonparametric two-sided test for comparison of medians was used to generate two different *P*-values per gene [21], following the leave-one-tissue-out strategy as outlined in the methods section. Both formulations were applied to each gene characterized by a *P*-value as an input and as a transformation of the other *P*-value as an output (equation 1). The first 10 frontiers were identified, and they contained 28 potential biomarkers. Numerically, reducing 10,692 genes to only 28 of them evidences the screening power of the proposed method. Table 1 outlines the genes identified in the analysis. These were then investigated in the literature to assess their cervix cancer biomarking potential as discussed next.

In the first efficient frontier there is only one gene: the *NAB1* gene that codes for EGR1-binding protein 1, which has been reported as a potential tumor suppressor in different cancer types including prostate cancer [22], breast cancer [23], esophageal cancer [24], hepatoma [25], and leukemia [26].

The LIM domain 7 (*LMO7*) gene was selected in the second frontier. The protein product of the *LMO7* belongs to the PDZ-LIM family. Regulation problems with these proteins can support the development of cancer [27].

Third frontier holds *DDR2*, *PPP1R1A*, *ARF4*, and *KPNA6*. Changes in expression of *DDR2* have been linked to several human cancers, for example, in non-small cell lung carcinoma (NSCLC) [28] and in nasopharyngeal carcinoma [29]. The *PPP1R1A* product is the protein phosphatase 1, regulatory (inhibitor) subunit 1A. In a recent study, the *PPP1R1A* expression in lung, colorectal, and gastric cancer cell lines was different from that of the normal tissues [30], as well as in some cell lines developed from different pediatric tumors [31]. The ADP-ribosylation factor 4 (*ARF4*) gene protein product interacts with epidermal growth factor receptor (EGFR) mediating the EGF-dependent cellular activation of phospholipase D2 (*PLD2*) [32]. An increased *PLD2* activity has been reported for human cancers including breast, colon, gastric, and kidney [33]. The *ARF4* has also been proposed as an antiapoptotic gene in human glioblastoma-derived U373MG cells [34]. The product of the *KPNA6* gene has been reported to play an important role in the antioxidant response and in keeping the redox homeostasis of the cell [35]. Its downregulation was reported to inhibit HeLa cell proliferation [36].

The fourth frontier holds *RAD52* along with an expressed sequence tag (EST). *RAD52* codes for a protein that is homolog to the *Saccharomyces cerevisiae* Rad52. The overexpression of *RAD52*, along with *RAD51* and *TOP2A*, all three DNA repair genes, has been reported to be predictive of poor relapse-free survival for melanoma [37].

The genes in the fifth frontier are *RBM25* and *UBE3A*. The product of the *RBM25* gene is an RNA-binding protein that acts as a splicing factor and has been shown to act on the alternative splicing of apoptotic factors [38].

Table 1. List of the 28 genes identified in the first 10 frontiers of the proposed multiple criteria optimization (MCO) problem.

Frontier	Accession number	Symbol	Name	Expression in cervix cancer (using data from Wong et al. [13])
1	AA488645	*NAB1*	NGFI-A-binding protein 1 (EGR1 binding protein 1)	Underexpressed
2	H22826	*LMO7*	LIM domain 7	Overexpressed
3	AI553969	*KPNA6*	Karyopherin α6 (importin α7)	Overexpressed
3	T71316	*ARF4*	ADP-ribosylation factor 4	Overexpressed
3	AA243749	*DDR2*	Discoidin domain receptor tyrosine kinase 2	Overexpressed
3	AA460827	*PPP1R1A*	Protein phosphatase 1, regulatory (inhibitor) subunit 1A	Underexpressed
4	AA454831		EST: zx79c10.s1	Overexpressed
4	AA913408, AA913864	*RAD52*	DNA damage repair and recombination protein RAD52 pseudogene	Overexpressed
5	AA487237	*UBE3A*	Ubiquitin protein ligase E3A	Underexpressed
5	AA446565	*RBM25*	RNA-binding motif protein 25	Overexpressed
6	H23187	*CA2*	Carbonic anhydrase II	Overexpressed
7	AI221445	*KCNE3*	Potassium voltage-gated channel, Isk-related family, member 3	Overexpressed
7	R36086		EST: yh88d01.s1	Underexpressed
7	AA282537	*LOC729991*	Hypothetical protein LOC729991	Overexpressed
8	N93686	*ALDH3B1*	Aldehyde dehydrogenase 3 family, member B1	Underexpressed
8	R91078	*CYP3A7*	Cytochrome P450, family 3, subfamily A, polypeptide 7	Overexpressed
8	R44822	*PRPSAP1*	Phosphoribosyl pyrophosphate synthetase-associated protein 1	Underexpressed
9	AI334914	*ITGA2B*	Integrin, alpha 2b (platelet glycoprotein IIb of IIb/IIIa complex, antigen CD41)	Overexpressed
9	R93394		Transcribed locus	Overexpressed
9	AA621155	*MSH5*	MutS homolog 5 (*Escherichia coli*)	Underexpressed
9	AA705112	*MOCS1*	Molybdenum cofactor synthesis 1	Overexpressed
9	R52794	*PTPRT*	Protein tyrosine phosphatase, receptor type, T	Underexpressed
10	AA424344	*UROD*	Uroporphyrinogen decarboxylase	Overexpressed
10	H69876	*LOC100132707*	Hypothetical LOC100132707	Underexpressed
10	H55909	*SRSF1*	Serine/arginine-rich splicing factor 1	Underexpressed
10	W74657	*KLF2*	Kruppel-like factor 2 (lung)	Overexpressed
10	AI017398	*ACCN2*	Amiloride-sensitive cation channel 2, neuronal	Overexpressed
10	H99699	*POLR3H*	Polymerase (RNA) III (DNA directed) polypeptide H (22.9 kD)	Overexpressed

The table shows complete list of genes identified in the first 10 efficient frontiers. In the last column, the expression change from the normal state to the cancer state is shown.

The product of the *UBE3A* gene is an E3 ubiquitin protein ligase, the E6-associated protein (E6AP). This protein is used by the E6 oncoprotein, from high-risk human papillomavirus (HPV) types, to produce the proteolysis of the tumor suppressor p53 [39]. The E6AP is also used by E6 to stimulate the telomerase activity, generally present in cancer cell lines [40].

CA II, the gene in the sixth frontier, has been reported to be expressed in the neovessel endothelium and the tumor cell cytoplasm of medulloblastomas and primitive neuroectodermal tumors [41] and has been proposed as a biomarker gene for gastrointestinal stromal tumors [42].

In the seventh frontier *KCNE3*, the uncharacterized conserved protein LOC729991, and the EST yh88d01.s1 were selected. The *KCNE3* gene codes for the potassium voltage-gated channel, Isk-related family, member 3. An increase in the activity of plasma membrane voltage-gated potassium channels promote neuronal cell death by apoptosis [43].

The genes in the eighth frontier are *ALDH3B1*, *CYP3A7*, and *PRPSAP1*. In a recent study, the expression of *ALDH3B1* was found to be tissue dependent, being upregulated in a high percentage of tumors used in the study (lung > breast = ovarian > colon) [44]. *CYP3A7* codes for a protein from the cytochrome P450 superfamily of enzymes. Proteins of this family play an important role in carcinogenesis because they metabolically activate precarcinogens and can metabolize anticancer drugs. The product of the *PRPSAP1* gene has been suggested to play a negative regulatory role in 5-phosphoribose 1-diphosphate synthesis and to bind to

PRPS1 and *PRPS2* [45], enzymes involved in the synthesis of purine and pyrimidine nucleotides.

The genes in the ninth frontier are *ITGA2B*, *MSH5*, *MOCS1*, and *PTPRT*. The *ITGA2B* gene codes for the integrin alpha chain 2b. Integrins can activate protein kinases involved in the regulation of cell growth, division, survival, differentiation, migration, and apoptosis. The *MSH5* gene codes for a member of the mutS family of proteins. These proteins are involved in promoting ionizing radiation-induced apoptosis [46]. A recent study found that the level of mRNA for genes involved in mismatching repair, including *MSH5*, was lower in colorectal cancer samples than in normal tissues [47]. The product of the *MOCS1* gene is involved in the molybdenum cofactor biosynthesis. Deficiency in molybdenum cofactor produces deficiency in the sulfite oxidase, xanthine dehydrogenase, and aldehyde oxidase [48]. Xanthine oxidoreductase has been associated with various forms of cancers as well as other human diseases (reviewed in [49]). The *PTPRT* gene codes for a tyrosine phosphatase protein, receptor type T, and has been suggested that its product has tumor suppression functions [50].

In the 10th frontier, the genes selected by the analysis method used in this study are *UROD*, *LOC100132707*, *SRSF1*, *KLF2*, *ACCN2*, and *POLR3H*. The *UROD* gene has been reported to be overexpressed in biopsies from patients with head and neck cancer [51]. *LOC100132707* is a hypothetical gene, the product of which is uncharacterized. The *SRSF1* gene codes for a member of the arginine/serine-rich splicing factor protein family, its product works activating or repressing splicing of pre-mRNA [52]. It has been proposed that *KLF2* could have a tumor suppressor activity in the MCF-7 mammary carcinoma cells [53]. Also, the expression of *KLF2* has been reported to inhibit Jurkat T leukemia cell growth [54]. The *ACCN2* product is an acid-sensing ion channel (ASIC) shown to have higher expression in human glioblastoma multiforme cells as compared with primary human astrocytes [55]. The *POLR3H* gene codes for the polymerase (RNA) III (DNA-directed) polypeptide H. RNA polymerase (pol) III synthesizes several products required for protein synthesis, and there have been detected high rates of pol III transcription in several cancers (reviewed in [56]).

As it can be seen, the literature marshaled about the genes detected by the proposed method evidences the biological relevance of the analysis output. The following section presents cross-validation studies that support analysis consistency.

Cross-validation studies of results in cervix cancer

The proposed method is capable to importantly accelerate the detection of potential cancer biomarkers, as shown in the previous study. In the following studies, the objective was to cross-validate the use of the method following (1) a genetic signature approach and (2) a statistical classification procedure.

Two independent cervix cancer databases using the same microarray platform, the Affymetrix U133A (with 22,283 probe set), were identified [14, 15]. Using the proposed method as in the previous study, and considering only the healthy and cancer data, a series of potential biomarkers was selected using solely database 1 [14]. These genes were then identified in database 2 [15] and the change in expression was compared between the datasets. Table 2 shows the overlap between the reference signature behavior from database 1 and the behavior of genes in database 2. The overlap amounts to 28 genes (29 probes with two probes for gene *SMC4*), which is 71.8% of the original signature, evidencing the effectiveness of the method. Table 2 also summarizes evidence found in the literature to support the genes' potential biomarking role in cervix cancer or in other types of cancer.

An important fact to emphasize in this study is, also, that of the discrimination power of the tool. The microarray platform used by both databases involved in the validation study contained 22,283 probes set. The fact that a signature of 39 genes was feasible to be built and tested evidences the advantage of using the proposed method.

A second cross-validation study entailed building a linear classifier with the set of genes identified as potential biomarkers in database 1, but applying it to classify the tissues in database 2. The classification rate in the 56 tissues of database 2 (24 healthy tissues and 32 cancer tissues) was 100%. The classifier was built with linear discriminant analysis and the results imply that the selection of potential biomarkers in database 1 achieved perfect linear separability in database 2. This provides solid evidence on the competitiveness of the proposed method.

Contrast with the single performance measure strategy

The single performance measure strategy is prevalent in the literature for the selection of genes that change their expression significantly between the conditions under comparison. It generally involves defining a threshold to select a number of potential biomarkers based on a single measurable criterion. The definition of such threshold may vary from experimenter to experimenter, however.

In this section, a multiple simultaneous hypothesis testing approach with a Bonferroni correction by Holms [101] was used to contrast a single performance measure strategy with the multiple performance measure strategy proposed here. For each gene in database 1 [14], a *P*-value was obtained based upon the Mann–Whitney non-

Table 2. List of genes from the cross-validation study.

Gene probe	Gene name	Sign of expression change from healthy tissues to cancer tissues		Efficient frontier in which it was identified	Examples of cancer types where the gene is involved	Reference
		Database 1 [14]	Database 2 [15]			
202575_at	*CRABP2*	–	–	3	Head and neck, breast	[57, 58]
205402_x_at	*PRSS2*	–	–	10	Colorectal, gastric tumorigenesis	[59, 60]
218677_at	*S100A14*	–	–	9	Esophageal squamous cell carcinoma cells, oral squamous cell carcinoma	[61, 62]
202096_s_at	*TSPO*	–	–	7	Thyroid, breast	[63, 64]
212249_at	*PIK3R1*	–	–	7	Endometrial, colorectal	[65, 66]
212567_s_at	*MAP4*	–	–	6	Breast, non-small cell lung carcinomas	[67, 68]
211366_x_at	*CASP1*	–	–	9	Cervical squamous carcinoma cells	[69]
213449_at	*POP 1*	–	+	3	Esophageal adenocarcinoma	[70]
214933_at	*CACNA1A*	–	+	5	Lung cancer cell lines	[71]
212889_x_at	*GADD45GIP1*	–	–	6	SKOV3 and HeLa cell lines	[72]
217912_at	*DUS1L*	–	+	7		
206626_x_at	*SSX1*	–	–	1	Prostate, multiple myeloma	[73, 74]
213450_s_at	*ICOSLG*	–	–	8	Metastatic melanoma, ductal pancreatic adenocarcinoma	[75, 76]
220405_at	*LOC100127998*	–	–	5		
208032_s_at	*GRIA3*	–	–	2	Pancreatic	[77]
205690_s_at	*BUD31*	–	–	4		
206543_at	*SMARCA2*	–	–	7	Prostate, skin	[78, 79]
203716_s_at	*DPP4*	+	–	1		
212291_at	*HIPK1*	+	+	3	Acute myeloid leukemia	[80, 81]
221632_s_at	*WDR4*	+	–	10		
66053_at	*HNRNPUL2*	+	–	3		
207142_at	*KCNJ3*	+	–	3	Pancreas, breast, lung	[57,58]
207742_s_at	*NR6A1*	+	–	3	Germ cell tumors of the testis	[82]
211615_s_at	*LRPPRC*	+	+	1	Lung adenocarcinoma cell lines, esophageal squamous cell carcinoma, stomach, colon, mammary and endometrial adenocarcinoma, and lymphoma	[83]
209245_s_at	*KIF1C*	+	–	3	Breast, non-small cell lung cancer metastatic spread to the brain	[84, 85]
213694_at	*RSBN1*	+	–	6		
222027_at	*NUCKS1*	+	+	7	Breast	[86]
205362_s_at	*PFDN4*	+	+	6	Colorectal	[87]
208706_s_at	*EIF5*	+	–	4	Chronic myeloid leukemia	[88]
211929_at	*HNRNPA3*	+	+	7	Non-small cell lung cancer	[89]
203738_at	*C5orf22*	+	+	3		
201794_s_at	*SMG7*	+	+	2		
200607_s_at	*RAD21*	+	+	5	Breast	[90]
201011_at	*RPN1*	+	+	9	Hematologic malignancies	[91]
201761_at	*MTHFD2*	+	+	8	Bladder, breast	[92, 93]
203880_at	*COX17*	+	+	1	Non-small cell lung cancer	[94]
212255_s_at	*ATP2C*	+	+	9	Breast, cervical	[95, 96]
205112_at	*PLCE1*	+	+	8	Gastric adenocarcinoma, colorectal	[97, 98]
201663_s_at	*SMC4*	+	+	9	Breast, cervical	[14, 99, 100]
201664_at	*SMC4*	+	+	7	Breast, cervical	[14, 99, 100]

The table shows genetic signature obtained in the cross-validation study. Both the matching and the nonmatching genes (shaded) are provided in this list along with evidence of their roles in cervix and other types of cancer.

parametric test for difference of medians between two groups. All genes and their associated *P*-values were sorted in increasing order in terms of *P*-value. To decide whether a gene in the (i)th place of the ordered sequence shows significantly different relative expression levels with the presence of cancer, the following criterion is evaluated:

$$P(i) \leq \frac{\alpha}{q - i + 1} \tag{2}$$

where α is the family-wise error rate and q is the number of total hypothesis tests being carried out, which in this instance, corresponds to the number of genes under evaluation.

The choice of the value of α is habitually left to the user. With database 1, when $\alpha < 0.1280$, no gene is deemed to change its relative expression significantly. At $\alpha = 0.1280$, a total of 86 genes are deemed to have changed their relative expression significantly. The number of genes in this category goes up to 116 at $\alpha = 0.1530$. The choice of α by the user, as it can be seen, greatly affects the number of genes that are considered important.

To make a fair comparison with the proposed multiple criteria method in this study, only the top 39 genes were chosen to build a linear classifier to be applied to database 2 [15] as in the previous section. The classification rate was also of 100% in both healthy tissues and cancer tissues. It is important to notice that although both methods achieved 100% classification rate in an independent database, the proposed multiple criteria method did not require for the user to set any parameter.

Conclusions

The search for potential cancer biomarkers can be greatly enhanced through the use of optimization techniques. In this study, a multiple criteria representation of the gene expression changes identification problem using microarray data is proposed. As a first case, the analysis of a single microarray experiment has been used to extract biologically relevant information in terms of potential biomarkers. The methodology can be extended to find the best compromises between data from different experiments for the same cancer type.

DEA is shown as a promising first approach to characterize the convex-efficient frontier of the MCO problem, and therefore to point toward potential biomarkers in a parameter-free and consistent fashion.

The proposed method, when applied to a publicly available microarray database from cervix cancer, identified genes already reported as relevant for different cancer types or cellular processes related to cancer. When the behavior of a selected gene was contrary to what was expected (*NAB1* [AA488645], *RBM25* [AA446565], *UBE3A* [AA487237], *ALDH3B1* [N93686], *PRPSAP1* [R44822]), the original data were reexamined. For those genes the readings showed great dispersion, from one run to the next, making the signal very noisy, which can explain the odd observed behavior. Genes without previous report of their relevance can be proposed for further *in vitro* validation.

Similarly, in the cross-validation studies, 39 genes were identified as potential cervix cancer biomarkers in a database. Of these genes, there was an overlap of 29 genes with similar behavior in a second database using the same microarray platform. These genes are proposed in this study as potential cervix biomarkers. A second cross-validation study showed that the proposed selection of potential biomarkers achieved perfect linear separability in an independent database, adding evidence in favor of the performance of the proposed approach. Furthermore, the convenience of not requiring the user to set parameters that affect the output of the analysis was demonstrated through a comparison with a commonly used strategy based on a single performance measure.

New methodologies for biological characterization have emerged after microarrays. The issues in handling large amounts of data, analysis reproducibility, and consistency, as well as computational convenience will continue to be challenges. This situates the proposed approach as a promising tool capable to accelerate biological discovery and to facilitate meta-analysis.

Acknowledgments

This study is based on study supported by the National Science Foundation (NSF) under Grant HRD 0833112 (CREST program), as well as the National Institutes of Health (NIH) MARC Grant 5T36GM095335-02 "Bioinformatics Programs at Minority Institutions." UPRM BioSEI Grant 330103080301 awarded to M. Cabrera-Ríos and PROMEP Grant 103.5/07/2523 awarded to C. Isaza were also critical to the development of this study. M. Sanchez Peña was supported through a research fellowship in the Department of Industrial Engineering at UPRM.

Conflict of Interest

None declared.

References

1. Ho, L., N. Sharma, L. Blackman, E. Festa, G. Reddy, and G. M. Pasinetti. 2005. From proteomics to biomarker discovery in Alzheimer's disease. Brain Res. Rev. 48:360–369.
2. Riker, A. I., S. A. Enkemann, O. Fodstad, S. Liu, S. Ren, C. Morris, et al. 2008. The gene expression profiles of

primary and metastatic melanoma yields a transition point of tumor progression and metastasis. BMC Med. Genomics 1:13.
3. Di Valentin, E., C. Crahay, N. Garbacki, B. Hennuy, M. Guéders, A. Noël, et al. 2009. New asthma biomarkers: lessons from murine models of acute and chronic asthma. Am. J. Physiol. Lung Cell. Mol. Physiol. 296:L185–L197.
4. Olson, N. E. 2006. The microarray data analysis process: from raw data to biological significance. NeuroRX 3:373–383.
5. Ioannidis, J. P., D. B. Allison, C. A. Ball, I. Coulibaly, X. Cui, A. C. Culhane, et al. 2009. Repeatability of published microarray gene expression analyses. Nat. Genet. 41:149–155.
6. van 't Veer, L. J., H. Dai, M. J. van de Vijver, Y. D. He, A. A. Hart, M. Mao, et al. 2002. Gene expression profiling predicts clinical outcome of breast cancer. Nature 415:530–536.
7. Wang, Y., J. G. Klijn, Y. Zhang, A. M. Sieuwerts, M. P. Look, F. Yang, et al. 2005. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. Lancet 365:671–679.
8. Ein-Dor, L., I. Kela, G. Getz, D. Givol, and E. Domany. 2005. Outcome signature genes in breast cancer: is there a unique set? Bioinformatics 21:171–178.
9. Ehrgott, M. 2005. Multicriteria optimization. Springer, Heidelberg, New York.
10. Hero, A. O., and G. Fleury. 2004. Pareto-optimal methods for gene ranking. J. VLSI Signal Process. Syst. 38:259–275.
11. Charnes, A., W. Cooper, A. Lewin, and L. Seiford. 1995. Data envelopment analysis theory, methodology and applications. Kluwer Academic Publishers, Norwell, MA.
12. Marroquín, M. G. V., M. L. S. Peña, C. E. Castro, J. M. Castro, and M. Cabrera-Ríos. 2008. Use of data envelopment analysis and clustering in multiple criteria optimization. Intell. Data Anal. 12:89–101.
13. Wong, Y. F., Z. E. Selvanayagam, N. Wei, J. Porter, R. Vittal, R. Hu, et al. 2003. Expression genomics of cervical cancer. Molecular classification and prediction of radiotherapy response by DNA microarray. Clin. Cancer Res. 9:5486–5492.
14. Zhai, Y., R. Kuick, B. Nan, I. Ota, S. J. Weiss, C. L. Trimble, et al. 2007. Gene expression analysis of preinvasive and invasive cervical squamous cell carcinomas identifies HOXC10 as a key mediator of invasion. Cancer Res. 67:10163–10172.
15. Scotto, L., G. Narayan, S. V. Nandula, H. Arias-Pulido, S. Subramaniyam, A. Schneider, et al. 2008. Identification of copy number gain and overexpressed genes on chromosome arm 20q by an integrative genomic approach in cervical cancer: potential role in progression. Genes Chromosom. Cancer 47:755–765.
16. Chen, Y., E. R. Dougherty, and M. L. Bittner. 1997. Ratio-based decisions and the quantitative analysis of cDNA microarray images. J. Biomed. Opt. 2:364–374.
17. Pan, W. 2002. A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. Bioinformatics 18:546–554.
18. Matlab Software. Available at http://www.mathworks.com/ (accessed 23 February 2013).
19. Lingo Software. Available at http://www.lindo.com (accessed 23 February 2013).
20. DEA Solver Pro Software. Available at http://www.saitech-inc.com (accessed 23 February 2013).
21. Hollander, M., and D. A. Wolfe. 1999. Nonparametric statistical methods. Wiley-Interscience, John Wiley & Sons, Inc., New York, NY.
22. Virolle, T., A. Krones-Herzig, V. Baron, G. De Gregorio, E. D. Adamson, and D. Mercola. 2003. Egr1 promotes growth and survival of prostate cancer cells. J. Biol. Chem. 278:11802–11810.
23. Ronski, K., M. Sanders, J. A. Burleson, V. Moyo, P. Benn, and M. Fang. 2005. Early growth response gene 1 (EGR1) is deleted in estrogen receptor-negative human breast carcinoma. Cancer 104:925–930.
24. Wu, M. Y., M. H. Chen, Y. R. Liang, G. Z. Meng, H. X. Yang, and C. X. Zhuang. 2001. Experimental and clinicopathologic study on the relationship between transcription factor Egr-1 and esophageal carcinoma. World J. Gastroenterol. 7:490–495.
25. Hao, M. W., Y. R. Liang, Y. F. Liu, L. Liu, M. Y. Wu, and H. X. Yang. 2002. Transcription factor EGR-1 inhibits growth of hepatocellular carcinoma and esophageal carcinoma cell lines. World J. Gastroenterol. 8:203–207.
26. Shafarenko, M., D. A. Liebermann, and B. Hoffman. 2005. Egr-1 abrogates the block imparted by c-Myc on terminal M1 myeloid differentiation. Blood 106:871–878.
27. Krcmery, J., T. Camarata, A. Kulisz, and H. G. Simon. 2010. Nucleocytoplasmic functions of the PDZ-LIM protein family: new insights into organ development. BioEssays 32:100–108.
28. Ford, C. E., S. K. Lau, C. Q. Zhu, T. Andersson, M. S. Tsao, and W. F. Vogel. 2007. Expression and mutation analysis of the discoidin domain receptors 1 and 2 in non-small cell lung carcinoma. Br. J. Cancer 96:808–814.
29. Chua, H. H., T. H. Yeh, Y. P. Wang, Y. T. Huang, T. S. Sheen, Y. C. Lo, et al. 2008. Upregulation of discoidin domain receptor 2 in nasopharyngeal carcinoma. Head Neck 30:427–436.
30. Takakura, S., T. Kohno, R. Manda, A. Okamoto, T. Tanaka, and J. Yokota. 2001. Genetic alterations and expression of the protein phosphatase 1 genes in human cancers. Int. J. Oncol. 18:817–824.
31. Wai, D. H., K. L. Schaefer, A. Schramm, E. Korsching, F. Van Valen, T. Ozaki, et al. 2002. Expression analysis of pediatric solid tumor cell lines using oligonucleotide microarrays. Int. J. Oncol. 20:441–451.

32. Kim, S. W., M. Hayashi, J. F. Lo, Y. Yang, J. S. Yoo, and J. D. Lee. 2003. ADP-ribosylation factor 4 small GTPase mediates epidermal growth factor receptor-dependent phospholipase D2 activation. J. Biol. Chem. 278:2661–2668.
33. Su, W., and Q. Chen, M. A. Frohman. 2009. Targeting phospholipase D with small-molecule inhibitors as a potential therapeutic approach for cancer metastasis. Future Oncol. 5:1477–1486.
34. Woo, I. S., S. Y. Eun, H. S. Jang, E. S. Kang, G. H. Kim, H. J. Kim, et al. 2009. Identification of ADP-ribosylation factor 4 as a suppressor of N-(4-hydroxyphenyl) retinamide-induced cell death. Cancer Lett. 276:53–60.
35. Sun, Z., T. Wu, F. Zhao, A. Lau, C. M. Birch, and D. D. Zhang. 2011. KPNA6 (importin {alpha}7)-mediated nuclear import of KEAP1 represses the NRF2-dependent antioxidant response. Mol. Cell. Biol. 31:1800–1811.
36. Quensel, C., B. Friedrich, T. Sommer, E. Hartmann, and M. Kohler. 2004. In vivo analysis of importin alpha proteins reveals cellular proliferation inhibition and substrate specificity. Mol. Cell. Biol. 24:10246–10255.
37. Jewell, R., C. Conway, A. Mitra, J. Randerson-Moor, S. Lobo, J. Nsengimana, et al. 2010. Patterns of expression of DNA repair genes and relapse from melanoma. Clin. Cancer Res. 16:5211–5221.
38. Zhou, A., A. C. Ou, A. Cho, E. J. Benz Jr., and S. C. Huang. 2008. Novel splicing factor RBM25 modulates Bcl-x pre-mRNA 5′ splice site selection. Mol. Cell. Biol. 28:5924–5936.
39. Scheffner, M., B. A. Werness, J. M. Huibregtse, A. J. Levine, and P. M. Howley. 1990. The E6 oncoprotein encoded by human papillomavirus types 16 and 18 promotes the degradation of p53. Cell 63:1129–1136.
40. Kelley, M. L., K. E. Keiger, C. J. Lee, and J. M. Huibregtse. 2005. The global transcriptional effects of the human papillomavirus E6 protein in cervical carcinoma cell lines are mediated by the E6AP ubiquitin ligase. J. Virol. 79:3737–3747.
41. Nordfors, K., J. Haapasalo, M. Korja, A. Niemelä, J. Laine, A. K. Parkkila, et al. 2010. The tumour-associated carbonic anhydrases CA II, CA IX and CA XII in a group of medulloblastomas and supratentorial primitive neuroectodermal tumours: an association of CA IX with poor prognosis. BMC Cancer 10:148.
42. Parkkila, S., J. Lasota, J. A. Fletcher, W. B. Ou, A. J. Kivelä, K. Nuorva, et al. 2010. Carbonic anhydrase II. A novel biomarker for gastrointestinal stromal tumors. Mod. Pathol. 23:743–750.
43. Lauritzen, I., M. Zanzouri, E. Honoré, F. Duprat, M. U. Ehrengruber, M. Lazdunski, et al. 2003. K+-dependent cerebellar granule neuron apoptosis. J. Biol. Chem. 278:32068–32076.
44. Marchitti, S. A., D. J. Orlicky, C. Brocker, and V. Vasiliou. 2010. Aldehyde dehydrogenase 3B1 (ALDH3B1): immunohistochemical tissue distribution and cellular-specific localization in normal and cancerous human tissues. J. Histochem. Cytochem. 58:765–783.
45. Uniprot. Available at http://www.uniprot.org/ (accessed 23 February 2013).
46. Tompkins, J. D., X. Wu, Y. L. Chu, and C. Her. 2009. Evidence for a direct involvement of hMSH5 in promoting ionizing radiation induced apoptosis. Exp. Cell Res. 315:2420–2432.
47. Ioana, M., C. Angelescu, F. Burada, F. Mixich, A. Riza, T. Dumitrescu, et al. 2010. Gene expression pattern in sporadic colorectal cancer. J. Gastrointestin. Liver Dis. 19:155–159.
48. Reiss, J., and J. L. Johnson. 2003. Mutations in the molybdenum cofactor biosynthetic genes MOCS1, MOCS2, and GEPH. Hum. Mutat. 21:569–576.
49. Lin, J., P. Xu, P. LaVallee, and J. R. Hoidal. 2008. Identification of proteins binding to E-Box/Ku86 sites and function of the tumor suppressor SAFB1 in transcriptional regulation of the human xanthine oxidoreductase gene. J. Biol. Chem. 283:29681–29689.
50. Scott, A., and Z. Wang. 2011. Tumour suppressor function of protein tyrosine phosphatase receptor-T. Biosci. Rep. 31:303–307.
51. Ito, E., S. Yue, E. H. Moriyama, A. B. Hui, I. Kim, W. Shi, et al. 2011. Uroporphyrinogen decarboxylase is a radiosensitizing target for head and neck cancer. Sci. Transl. Med. 3:67ra7.
52. NCBI gene SRSF1 serine/arginine-rich splicing factor 1 [*Homo sapiens*]. Available at http://www.ncbi.nlm.nih.gov/pubmed?Db=gene&Cmd=retrieve&dopt=full_report&list_uids=6426 (accessed 23 February 2013).
53. Kannan-Thulasiraman, P., D. D. Seachrist, G. H. Mahabeleshwar, M. K. Jain, and N. Noy. 2010. Fatty acid-binding protein 5 and PPARβ/δ are critical mediators of epidermal growth factor receptor-induced carcinoma cell growth. J. Biol. Chem. 285:19106–19115.
54. Wu, J., and J. B. Lingrel. 2004. KLF2 inhibits Jurkat T leukemia cell growth via upregulation of cyclin-dependent kinase inhibitor p21WAF1/CIP1. Oncogene 23:8088–8096.
55. Kapoor, N., R. Bartoszewski, Y. J. Qadri, Z. Bebok, J. K. Bubien, C. M. Fuller, et al. 2009. Knockdown of ASIC1 and epithelial sodium channel subunits inhibits glioblastoma whole cell current and cell migration. J. Biol. Chem. 284:24526–24541.
56. White, R. J. 2004. RNA polymerase III transcription and cancer. Oncogene 23:3208–3216.
57. Calmon, M. F., R. V. Rodrigues, C. M. Kaneto, R. P. Moura, S. D. Silva, L. D. Mota, et al. 2009. Epigenetic silencing of CRABP2 and MX1 in head and neck tumors. Neoplasia 11:1329–1339.
58. Geiger, T., S. F. Madden, W. M. Gallagher, J. Cox, and M. Mann. 2012. Proteomic portrait of human breast

cancer progression identifies novel prognostic markers. Cancer Res. 72:2428–2439.
59. Williams, S. J., D. C. Gotley, and T. M. Antalis. 2001. Human trypsinogen in colorectal cancer. Int. J. Cancer 93:67–73.
60. Rajkumar, T., N. Vijayalakshmi, G. Gopal, K. Sabitha, S. Shirley, U. M. Raja, et al. 2010. Identification and validation of genes involved in gastric tumorigenesis. Cancer Cell Int. 10:45.
61. Chen, H., Y. Yuan, C. Zhang, A. Luo, F. Ding, J. Ma, et al. 2012. Involvement of S100A14 protein in cell invasion by affecting expression and function of matrix metalloproteinase (MMP)-2 via p53-dependent transcriptional regulation. J. Biol. Chem. 287:17109–17119.
62. Sapkota, D., O. Bruland, D. E. Costea, H. Haugen, E. N. Vasstrand, and S. O. Ibrahim. 2011. S100A14 regulates the invasive potential of oral squamous cell carcinoma derived cell-lines in vitro by modulating expression of matrix metalloproteinases, MMP1 and MMP9. Eur. J. Cancer 47:600–610.
63. Klubo-Gwiezdzinska, J., K. Jensen, A. Bauer, A. Patel, J. Costello, K. Burman, et al. 2012. The expression of translocator protein in human thyroid cancer and its role in the response of thyroid cancer cells to oxidative stress. J. Endocrinol. 214:207–216.
64. Mukherjee, S., and S. K. Das. 2012. Translocator protein (TSPO) in breast cancer. Curr. Mol. Med. 12:443–457.
65. Cheung, L. W., B. T. Hennessy, J. Li, S. Yu, A. P. Myers, B. Djordjevic, et al. 2011. High frequency of PIK3R1 and PIK3R2 mutations in endometrial cancer elucidates a novel mechanism for regulation of PTEN protein stability. Cancer Discov. 1:170–185.
66. Nowakowska-Zajdel, E., U. Mazurek, E. Ziółko, E. Niedworok, E. Fatyga, T. Kokot, et al. 2011. Analysis of expression profile of gene encoding proteins of signal cascades activated by insulin-like growth factors in colorectal cancer. Int. J. Immunopathol. Pharmacol. 24:781–787.
67. Chen, X., J. Wu, H. Lu, O. Huang, and K. Shen. 2012. Measuring β-tubulin III, Bcl-2, and ERCC1 improves pathological complete remission predictive accuracy in breast cancer. Cancer Sci. 103:262–268.
68. Cucchiarelli, V., L. Hiser, H. Smith, A. Frankfurter, A. Spano, J. J. Correia, et al. 2008. Beta-tubulin isotype classes II and V expression patterns in nonsmall cell lung carcinomas. Cell Motil. Cytoskeleton 65:675–685.
69. Arany, I., I. A. Ember, and S. K. Tyring. 2003. All-trans-retinoic acid activates caspase-1 in a dose-dependent manner in cervical squamous carcinoma cells. Anticancer Res. 23:471–473.
70. Liu, C. Y., M. C. Wu, F. Chen, M. Ter-Minassian, K. Asomaning, R. Zhai, et al. 2010. A large-scale genetic association study of esophageal adenocarcinoma risk. Carcinogenesis 31:1259–1263.
71. Castro, M., L. Grau, P. Puerta, L. Gimenez, J. Venditti, S. Quadrelli, et al. 2010. Multiplexed methylation profiles of tumor suppressor genes and clinical outcome in lung cancer. J. Transl. Med. 8:86.
72. Nakayama, K., N. Nakayama, T. L. Wang, and IeM Shih. 2007. NAC-1 controls cell growth and survival by repressing transcription of Gadd45GIP1, a candidate tumor suppressor. Cancer Res. 67:8058–8064.
73. Smith, H. A., R. J. Cronk, J. M. Lang, andD. G. McNeel. 2011. Expression and immunotherapeutic targeting of the SSX family of cancer-testis antigens in prostate cancer. Cancer Res. 71:6785–6795.
74. Van Duin, M., A. Broyl, Y. de Knegt, H. Goldschmidt, P. G. Richardson, W. C. Hop, et al. 2011. Cancer testis antigens in newly diagnosed and relapse multiple myeloma: prognostic markers and potential targets for immunotherapy. Haematologica 96:1662–1669.
75. Fu, T., Q. He, and P. Sharma. 2011. The ICOS/ICOSL pathway is required for optimal antitumor responses mediated by anti-CTLA-4 therapy. Cancer Res. 71:5445–5454.
76. Tjomsland, V., A. Spångeus, P. Sandström, K. Borch, D. Messmer, and M. Larsson. 2010. Semi mature blood dendritic cells exist in patients with ductal pancreatic adenocarcinoma owing to inflammatory factors released from the tumor. PLoS ONE 5:e13441.
77. Ripka, S., J. Riedel, A. Neesse, H. Griesmann, M. Buchholz, V. Ellenrieder, et al. 2010. Glutamate receptor GRIA3–target of CUX1 and mediator of tumor progression in pancreatic cancer. Neoplasia 12:659–667.
78. Sun, A., O. Tawfik, B. Gayed, J. B. Thrasher, S. Hoestje, C. Li, et al. 2007. Aberrant expression of SWI/SNF catalytic subunits BRG1/BRM is associated with tumor development and increased invasiveness in prostate cancers. Prostate 67:203–213.
79. Moloney, F. J., J. G. Lyons, V. L. Bock, X. X. Huang, M. J. Bugeja, and G. M. Halliday. 2009. Hotspot mutation of Brahma in non-melanoma skin cancer. J. Invest. Dermatol. 129:1012–1015.
80. Mougeot, J. L., F. K. Bahrani-Mougeot, P. B. Lockhart, and M. T. Brennan. 2011. Microarray analyses of oral punch biopsies from acute myeloid leukemia (AML) patients treated with chemotherapy. Oral Surg. Oral Med. Oral Pathol. Oral Radiol. Endod. 112:446–452.
81. Aikawa, Y., L. A. Nguyen, K. Isono, N. Takakura, Y. Tagata, M. L. Schmitz, et al. 2006. Roles of HIPK1 and HIPK2 in AML1- and p300-dependent transcription, hematopoiesis and blood vessel formation. EMBO J. 25:3955–3965.
82. Juric, D., S. Sale, R. A. Hromas, R. Yu, Y. Wang, G. E. Duran, et al. 2005. Gene expression profiling differentiates germ cell tumors from other cancers and defines subtype-specific signatures. Proc. Natl. Acad. Sci. USA 102:17763–17768.

83. Tian, T., J. I. Ikeda, Y. Wang, S. Mamat, W. Luo, K. Aozasa, et al. 2012. Role of leucine-rich pentatricopeptide repeat motif-containing protein (LRPPRC) for anti-apoptosis and tumourigenesis in cancers. Eur. J. Cancer 48:2462–2473.
84. De, S., R. Cipriano, M. W. Jackson, and G. R. Stark. 2009. Overexpression of kinesins mediates docetaxel resistance in breast cancer cells. Cancer Res. 69:8035–8042.
85. Grinberg-Rashi, H., E. Ofek, M. Perelman, J. Skarda, P. Yaron, M. Hajdúch, et al. 2009. The expression of three genes in primary non-small cell lung cancer is associated with metastatic spread to the brain. Clin. Cancer Res. 15:1755–1761.
86. Ziółkowski, P., E. Gamian, B. Osiecka, A. Zougman, and J. R. Wiśniewski. 2009. Immunohistochemical and proteomic evaluation of nuclear ubiquitous casein and cyclin-dependent kinases substrate in invasive ductal carcinoma of the breast. J. Biomed. Biotechnol. 2009:919645.
87. Miyoshi, N., H. Ishii, K. Mimori, N. Nishida, M. Tokuoka, H. Akita, et al. 2010. Abnormal expression of PFDN4 in colorectal cancer: a novel marker for prognosis. Ann. Surg. Oncol. 17:3030–3036.
88. Balabanov, S., A. Gontarewicz, P. Ziegler, U. Hartmann, W. Kammer, M. Copland, et al. 2007. Hypusination of eukaryotic initiation factor 5A (eIF5A): a novel therapeutic target in BCR-ABL-positive leukemias identified by a proteomics approach. Blood 109:1701–1711.
89. Boukakis, G., M. Patrinou-Georgoula, M. Lekarakou, C. Valavanis, and A. Guialis. 2010. Deregulated expression of hnRNP A/B proteins in human non-small cell lung cancer: parallel assessment of protein and mRNA levels in paired tumour/non-tumour tissues. BMC Cancer 10:434.
90. Atienza, J. M., R. B. Roth, C. Rosette, K. J. Smylie, S. Kammerer, J. Rehbock, et al. 2005. Suppression of RAD21 gene expression decreases cell growth and enhances cytotoxicity of etoposide and bleomycin in human breast cancer cells. Mol. Cancer Ther. 4:361–368.
91. Shimizu, S., K. Suzukawa, T. Kodera, T. Nagasawa, T. Abe, M. Taniwaki, et al. 2000. Identification of breakpoint cluster regions at 1p36.3 and 3q21 in hematologic malignancies with t(1;3)(p36;q21). Genes Chromosom. Cancer 27:229–238.
92. Andrew, A. S., J. Gui, A. C. Sanderson, R. A. Mason, E. V. Morlock, A. R. Schned, et al. 2009. Bladder cancer SNP panel predicts susceptibility and survival. Hum. Genet. 125:527–539.
93. Xu, X., M. Qiao, Y. Zhang, Y. Jiang, P. Wei, J. Yao, et al. 2010. Quantitative proteomics study of breast cancer cell lines isolated from a single patient: discovery of TIMM17A as a marker for breast cancer. Proteomics 10:1374–1390.
94. Suzuki, C., Y. Daigo, T. Kikuchi, T. Katagiri, and Y. Nakamura. 2003. Identification of COX17 as a therapeutic target for non-small cell lung cancer. Cancer Res. 63:7038–7041.
95. Grice, D. M., I. Vetter, H. M. Faddy, P. A. Kenny, S. J. Roberts-Thomson, and G. R. Monteith. 2010. Golgi calcium pump secretory pathway calcium ATPase 1 (SPCA1) is a key regulator of insulin-like growth factor receptor (IGF1R) processing in the basal-like breast cancer cell line MDA-MB-231. J. Biol. Chem. 285:37458–3766.
96. Wilting, S. M., J. de Wilde, C. J. Meijer, J. Berkhof, Y. Yi, W. N. van Wieringen, et al. 2008. Integrated genomic and transcriptional profiling identifies chromosomal loci with altered gene expression in cervical cancer. Genes Chromosom. Cancer 47:890–8905.
97. Wang, M., R. Zhang, J. He, L. Qiu, J. Li, Y. Wang, et al. 2012. Potentially functional variants of PLCE1 identified by GWASs contribute to gastric adenocarcinoma susceptibility in an eastern Chinese population. PLoS ONE 7:e31932.
98. Danielsen, S. A., L. Cekaite, T. H. Ågesen, A. Sveen, A. Nesbakken, E. Thiis-Evensen, et al. 2011. Phospholipase C isozymes are deregulated in colorectal cancer – insights gained from gene set enrichment analysis of the transcriptome. PLoS ONE 6:e24419.
99. Chang, H., H. C. Jeung, J. J. Jung, T. S. Kim, S. Y. Rha, and H. C. Chung. 2011. Identification of genes associated with chemosensitivity to SAHA/taxane combination treatment in taxane-resistant breast cancer cells. Breast Cancer Res. Treat. 125:55–63.
100. Kulawiec, M., A. Safina, M. M. Desouki, I. Still, S. Matsui, A. Bakin, et al. 2008. Tumorigenic transformation of human breast epithelial cells induced by mitochondrial DNA depletion. Cancer Biol. Ther. 7:1732–1743.
101. Holm, S. 1979. A simple sequentially rejective multiple test procedure. Scand. J. Stat. 6:65–70.

