Cancer Medicine

Open Access

ORIGINAL RESEARCH

Multiple criteria optimization joint analyses of microarray experiments in lung cancer: from existing microarray data to new knowledge

Katia I. Camacho-Cáceres[1], Juan C. Acevedo-Díaz[1], Lynn M. Pérez-Marty[1], Michael Ortiz[1], Juan Irizarry[1], Mauricio Cabrera-Ríos[1] & Clara E. Isaza[1,2]

[1]Bio IE Lab, The Applied Optimization Group, Industrial Engineering Department, University of Puerto Rico, Mayaguez, Puerto Rico
[2]Public Health Program, Ponce Health Sciences University, Ponce, Puerto Rico

Keywords
Biomarker, lung cancer, meta-analysis, multi-criteria optimization

Correspondence
Clara E. Isaza, Bio IE Lab, The Applied Optimization Group, Industrial Engineering Department, University of Puerto Rico, Mayaguez, Puerto Rico.
Tel: 1-787-832-4040 ext. 5964;
Fax: 1-787-265-3820;
E-mail: applied.optimization@gmail.com

Funding Information
This work was supported in part 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."

Received: 8 February 2015; Revised: 30 July 2015; Accepted: 14 August 2015

Cancer Medicine **2015; 4(12): 1884–1900**

doi: 10.1002/cam4.540

Abstract

Microarrays can provide large amounts of data for genetic relative expression in illnesses of interest such as cancer in short time. These data, however, are stored and often times abandoned when new experimental technologies arrive. This work reexamines lung cancer microarray data with a novel multiple criteria optimization-based strategy aiming to detect highly differentially expressed genes. This strategy does not require any adjustment of parameters by the user and is capable to handle multiple and incommensurate units across microarrays. In the analysis, groups of samples from patients with distinct smoking habits (never smoker, current smoker) and different gender are contrasted to elicit sets of highly differentially expressed genes, several of which are already associated to lung cancer and other types of cancer. The list of genes is provided with a discussion of their role in cancer, as well as the possible research directions for each of them.

Introduction

According to the International Agency for Research on Cancer, the world's most commonly diagnosed cancer is lung cancer, with 1.8 million cases or 13% of total in 2012. Additionally, lung cancer was the first cause of death in the world, with 1.6 million deaths or 19.4% in 2012 [1]. This analysis was conducted in 184 countries. This work intends to facilitate uncovering new information related to cancer using publicly available lung cancer microarray data. The aim is to find those genes that changed their relative expression the most in order to propose potential lung cancer biomarkers.

Microarray experiments quantify the relative expression of tens of thousands of genes. These experiments have been highly utilized in the past decade to study a number of health conditions, including cancer [2, 3]. These experiments, however, are often times measured in different units, thus making it difficult to analyze several of them simultaneously. Furthermore, because the measured level of expression is relative, a normalization process is commonly required. All of these have hampered the search for cancer biomarkers in the past.

The strategy to detect potential biomarkers utilized in this work is based on mathematical optimization.

Optimization can be defined as a decision-making process aimed to obtain the best possible values in a series of performance measures (PMs) of interest. The decision variables are habitually constrained to fall within specific ranges or to maintain mathematical relationships among them [4, 5]. Mathematical optimization (MO) has been widely used in many fields, including Economics and Engineering, and clearly it can be applied to biological analysis. MO can make a system or design effective, functional, or in its most basic form, possible [6, 7]. Multiple Criteria Optimization (MCO) is an optimization problem that finds a set of solutions corresponding to the best possible balances among two or more conflicting PMs under study [8]. These solutions are known as Pareto-Efficient solutions and are mathematically characterized by the well-established Pareto-optimality conditions. In general terms, then, the idea behind a MCO problem is to find the Pareto-Efficient Frontier formed by the Pareto-efficient solutions.

In this work the analysis of a publicly available microarray database for lung cancer is presented as a MCO problem. The genetic expression changes in this analysis were quantified using two metrics that do not have a perfect correlation and thus, are in conflict: difference of means and difference of medians. For the analysis, the MCO solutions will be those genes that have associated the greatest differences in the selected metrics. The solution genes are the ones that changed their expression the most between the compared conditions and could be potential biomarkers and can, after further study and confirmation, help with the diagnosis, prognosis, treatment, and recurrence prediction for the condition under analysis [9]. It can be appreciated that the method seeks to minimize the likelihood of false positives due to its focus on frontier analysis. This, expectedly, comes at the cost of false negatives in genes that might not appear in the Pareto-efficient frontier due to experimental error. A simple strategy of finding several consecutive frontiers is proposed to alleviate this issue.

The analysis strategy in this paper has, however, the advantage of providing objectivity, as it does not require the analyst to change or adjust any parameters, thereby fostering repeatability across analysts. It has also been shown to have a high discrimination power. The method used to solve the associated MCO problem is a full pairwise comparison scheme that effectively finds the genes that show high expression change across multiple PMs. This scheme is an improvement in terms of precision and convergence over the Data Envelopment Analysis approach presented by our group in [9]. The genes identified this way are located on the Pareto-efficient frontier of the MCO problem, that is, they are demonstrably Pareto-efficient [10] and are, in consequence, proposed as potential biomarkers.

Literature Review

Microarray experiments have been very popular among researchers [11]. In the Gene Expression Omnibus (GEO) as of May 2015 there are 3848 databases with 1,392,278 samples. Microarray experiments are sufficiently accepted as a reliable technology where the most common use is to find differentially expressed genes between two experimental conditions or samples [12]. Moreover, microarrays have been used to study how different biological processes or pathways work in several organisms [13]. To analyze the experimental data, statistics have been used for these types of studies [14, 15]. However, producing a standard method for analysis has never been accomplished.

In the literature there are many methods to find highly differentially expressed genes to characterize them as potential biomarkers. Most of them focus on statistical procedures [15, 16]. This research adopts multiple criteria optimization and Pareto conditions to find biomarkers following the direction of our research group [9, 17], and proposes extending the application to this end through simultaneous analysis of multiple independent experiments, that is carrying out meta-analysis. In 2010, in our group, Sanchez-Peña [9] used a combination of two performance measures (two *P*-values) obtained from a single-microarray database to cast the MCO problem and Data Enveloped Analysis (DEA) to solve it. The pairwise comparison scheme in the present work yields a more precise Pareto-efficient frontier than DEA, as it can deal with nonconvexity from the onset.

An important direction of this work is to use the proposed method for meta-analysis of high throughput biological experiments, starting with microarrays. Glasser and Duval [18] provide the definition: “Meta-analysis refers to methods for the systematic review of a set of individual studies or patients within each study, with the aim to quantitatively combine their results.” Meta-analysis is a method capable of taking independent, but associated studies to obtain a set of solutions through all studies. It is possible to find different applications and examples about meta-analysis. Li and his research group led a systematic review and meta-analysis to determine whether two polymorphisms (V89L and A49T) are associated with the risk of prostate cancer. They found 31 articles and reviews related to such risk [19]. On the other hand, R makes available a tool for microarray meta-analysis called MetaOmics. MetaOmics integrates Quality Control (Meta QC), Differentially Expressed (Meta DE), and Pathway (Meta pathway) [20]. Also, Zhuohui et al. (2014) research developed a tool, “MAAMD” [21]. They carried out meta-analysis using Affymetrix microarray data. The tool automates the process to analyze microarrays and requires normalization and several statistical methods to detect differential gene expression. To this end, they used Kepler, AltAnalyze and Bioconductor software packages. The parametric approaches in these works differ from our nonparametric approach. Therefore, it is clear that multiple criteria optimization differs from the reviewed approaches and constitute a novelty in meta-analysis. It must be emphasized, however that meta-analysis is a study that comprehends a larger area than afforded by the use of a single technique and that it requires a methodical design to be reliable. Especial care, for instance, must be given to the selection of studies to be included [22], as well as their heterogeneity [23]. Meta-analysis has become a cornerstone for evidence-based medicine [24] and follows widely accepted standards for its realization [25–27].

As noted earlier, the MCO problem has been approached in our group by Sánchez-Peña, et al. [9] through Data Envelopment Analysis (DEA). This work approaches the larger problem of analyzing multiple microarray databases simultaneously that is, to carry out meta-analysis, formulating the analysis as an MCO problem and solving it through a pairwise-comparison scheme that facilitates the evaluation of Pareto-efficiency conditions. In the literature, the authors of [28] have successfully applied Pareto – concepts for gene selection coupled with the use of a series of parametric statistical methods [28]. It is the intention of this work to keep the analysis strategy as nonparametric as possible, so as to not depend heavily on statistical assumptions or –in a different sense of nonparametric- the adjustment of parameters by the user that might bias the analysis results.

Method: MCO Problem Formulation

Figure 1 shows the elements of the graphical representation of the MCO problem. G denotes the universe of solutions that comprises the n genes to be analyzed with g_i representing each gene under analysis, (i = 1, 2, … n). Figure 2 shows the space defined by two criteria or PMs under analysis, m^1 and m^2. In the generalization of this figure, m_i^k is the value for the i-th gene in the k-th PM. Then k = 1, 2, … C, where C is the number of PMs considered in the analysis. The Pareto-efficient frontier in Figure 2 is formed by the genes g_i^*. These genes have indeed the best possible balances among the two PMs to be minimized and are the ones proposed as potential biomarkers.

When it comes to microarray analysis, the PMs of choice are usually related to the difference of gene expression measured in two distinct states for comparison purposes. Looking for the most differentially expressed genes is akin to looking for potential biomarkers, and it is a problem that can be casted as described up to this point.

According to Deb [29] and Ehrgott [30] the Pareto-efficient solutions must meet the Pareto-optimality conditions. In practical terms, this relates to finding nondominated solutions in the following sense: a solution $X^{(1)}$ is said to dominate the other solution $X^{(2)}$, if both conditions 1 and 2 are true:

1. The solution $X^{(1)}$ is no worse than $X^{(2)}$ in all PMs.
2. The solution $X^{(1)}$ is strictly better than $X^{(2)}$ in at least one PM.

These conditions can be evaluated for every single pair of genes to find those that are not dominated by any other gene. These are the Pareto-efficient genes that form the Pareto-efficient frontier of the MCO problem at hand.

As stated previously, in the search for the most differentially expressed genes, the expressions of all candidate genes are measured in two states to be then further

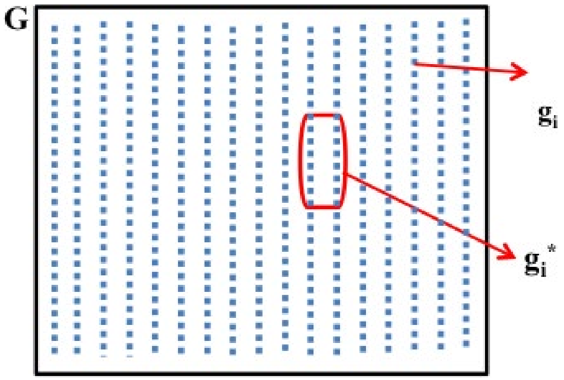


Figure 1. Problem representation where $G = \{g_i\}$, i = 1,2,3,…,n and $g_i^* \in G$.

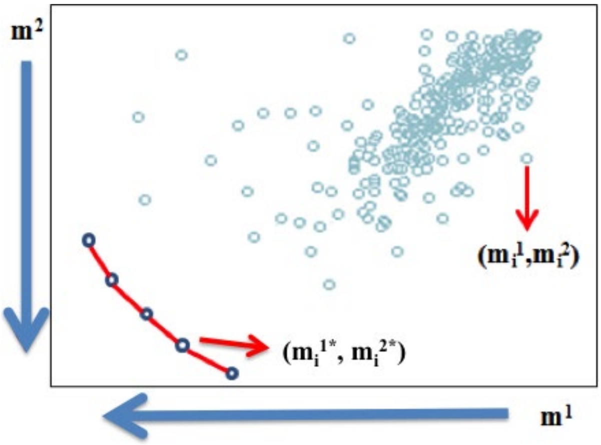


Figure 2. Representation of the Pareto-efficient frontier of the MCO problem.

compared. It is common, then, to use the difference of the means or the medians of the relative gene expression in these two states, for example. In this work, each of the C experiments will contribute one difference of medians between two states termed "control" and "cancer." This translates into each gene being evaluated through C PMs. The absolute value of these differences will then be transformed to follow a minimization direction to match the illustration in Figure 2, where the following notation is introduced:

Let us represent the i-th gene in terms of its values on each of the C PMs as $g_i \Leftrightarrow \left(m_i^1, m_i^2, \ldots, m_i^k, \ldots, m_i^C\right)$, for i = 1, 2, 3, …, n. Then, the objective of the analysis is to find the set of Pareto-efficient solutions: $g_i^* \Leftrightarrow \left(m_i^{1*}, m_i^{2*}, \ldots, m_i^{k*}, \ldots, m_i^{C*}\right)$. This is accomplished through a full pairwise comparison among the n genes as explained next.

First, a matrix δ^k is built for the k-th PM resulting in C squared matrices of size n built as follows:

$$\delta^k = \begin{array}{cccccc} & \mathbf{m}_1^k & \mathbf{m}_2^k & \cdots\ \mathbf{m}_j^k & \cdots & \mathbf{m}_n^k \\ \mathbf{m}_1^k & \delta_{11}^k & \delta_{12}^k & \cdots\ \delta_{1j}^k & \cdots & \delta_{1n}^k \\ \mathbf{m}_2^k & \delta_{21}^k & \delta_{22}^k & \cdots\ \delta_{2j}^k & \cdots & \delta_{2n}^k \\ \vdots & \vdots & \vdots & \cdots\ \vdots & \cdots & \vdots \\ \mathbf{m}_i^k & \delta_{i1}^k & \delta_{i2}^k & \cdots\ \delta_{1j}^k & \cdots & \delta_{in}^k \\ \vdots & \vdots & \vdots & \cdots\ \vdots & \cdots & \vdots \\ \mathbf{m}_n^k & \delta_{n1}^k & \delta_{n2}^k & \cdots\ \delta_{nj}^k & \cdots & \delta_{nn}^k \end{array}$$

where:

$$\delta_{ij}^k = \begin{cases} -1, \text{ if } m_i^k < m_j^k \\ 0, \text{ if } m_i^k = m_j^k \\ W, \text{ if } m_i^k < m_j^k \end{cases} \text{ for } i = 1, 2, \ldots, n\ j = 1, 2, \ldots, n; k = 1, 2, \ldots, C; \quad (1)$$

and W is defined as a large positive integer number used as a penalty. In this work, W = 1000 is used.

Next, a summation matrix is computed with elements $\alpha_{ij} = \sum_{k=1}^{C} \delta_{ij}^k$. This is exemplified in Table 1 when C = 2:

Table 1. All the possible combinations of a minimization problem for two criteria.

Outcome number	δ_{ij}^1	δ_{ij}^2	α_{ij}	Outcome
1	0	0	0	*X*^{i} is not worse and not better either in *m*^{1} or *m*^{2}
2	0	−1	−1	*X*^{i} is better in *m*^{2}
3	0	W	W	*X*^{i} is worse in *m*^{2}
4	−1	0	−1	*X*^{i} is better in *m*^{1}
5	−1	−1	−2	*X*^{i} is better in both *m*^{1} and *m*^{2}
6	−1	W	W-1	*X*^{i} is better in *m*^{1} and worse *m*^{2}
7	W	0	W	*X*^{i} is worse in *m*^{1}
8	W	−1	W-1	*X*^{i} is better in *m*^{2}
9	W	W	2W	*X*^{i} is worse in *m*^{1} and *m*^{2}

A new matrix γ is then build by assessing the values α_{ij}. For C=2, for example, the following assessment applies:

$$\gamma_{ij} = \begin{cases} W, \textit{ if } \alpha_{ij} \in \{0, W\} \\ 2W, \textit{ if } \alpha_{ij} = 2W \\ 0, \text{ otherwise} \end{cases}, \quad \begin{cases} i = 1,2, \ldots n \\ j = 1,2, \ldots n \end{cases}$$

In general for any value $C \geq 2$

$$\gamma_{ij} = \begin{cases} \frac{C}{2} W, \textit{ if } \alpha_{ij} \in \{0, W, \ldots, (C-1)\, W\} \\ CW, \textit{ if } \alpha_{ij} = CW \\ 0, \text{ otherwise} \end{cases}, \begin{cases} i = 1, 2, \ldots n \\ j = 1, 2, \ldots n \end{cases} \quad (2)$$

Thus, in summary, this process will result in the γ matrix:

$$\gamma = \begin{array}{cccccc} & m_1^{\cdot} & m_2^{\cdot} & \cdots\ m_j^{\cdot} & \cdots & m_n^{\cdot} \\ \mathbf{m}_1^{\mathbf{c}} & \gamma_{11} & \gamma_{12} & \cdots\ \gamma_{1j} & \cdots & \gamma_{1n} \\ \mathbf{m}_2^{\mathbf{c}} & \gamma_{21} & \gamma_{22} & \cdots\ \gamma_{2j} & \cdots & \gamma_{2n} \\ \vdots & \vdots & \vdots & \cdots\ \vdots & \cdots & \vdots \\ \mathbf{m}_i^c & \gamma_{i1} & \gamma_{i2} & \cdots\ \gamma_{ij} & \cdots & \gamma_{1n} \\ \vdots & \vdots & \vdots & \cdots\ \vdots & \cdots & \vdots \\ \mathbf{m}_n^c & \gamma_{n1} & \gamma_{n2} & \cdots\ \gamma_{nj} & \cdots & \gamma_{nn} \end{array}$$

In order to find g_i^*, a vector β is built containing the sums of each row of matrix γ as follows:

$$\beta_i = \sum_{j=1}^{n} \gamma_{ij},\ i = 1, 2, \ldots n \quad (3)$$

$$\beta = \begin{array}{cccccccc} \beta_1 & = & \gamma_{11} & +\gamma_{12} & + & \cdots & \gamma_{1j} + & \cdots\ \gamma_{1n} \\ \beta_2 & = & \gamma_{21} & +\gamma_{22} & + & \cdots & \gamma_{2j} + & \cdots\ \gamma_{2n} \\ \beta_3 & = & \gamma_{31} & +\gamma_{32} & + & \cdots & \gamma_{3j} + & \cdots\ \gamma_{3n} \\ \vdots & & \vdots & \vdots & & \cdots & \vdots & \cdots\ \vdots \\ \beta_i & = & \gamma_{i1} & +\gamma_{i2} & + & \cdots & \gamma_{ij} + & \cdots\ \gamma_{in} \\ \vdots & & \vdots & \vdots & & \cdots & \vdots & \cdots\ \vdots \\ \beta_n & = & \gamma_{n1} & +\gamma_{n2} & + & \cdots & \gamma_{nj} + & \cdots\ \gamma_{nn} \end{array}$$

The Pareto-efficient frontier will, then, contain all solutions that meet equation (4):

$$g_i^* = \left\{ g_i \middle| \beta_i < \mathrm{CW},\ i = 1, 2, \ldots n \right\} \quad (4)$$

With this last step, the Pareto-efficient solutions, $g_i^* = \left\{ m_i^{1*}\ , m_i^{2*}, \ldots m_i^{k*}, \ldots, m_i^{C*} \right\}$, are clearly identified.

This algorithm identifies all the solutions of the Pareto-efficient frontier. The maximum number proved and coded in this work is five PMs. The MatLab code is available in Appendix A1. In addition, Appendix A2 contrasts the proposed method with the use of a volcano plot to detect differentially expressed genes. Indeed, the mathematical description provided here is sufficient for the interested reader to code the method. The next illustration should help in this endeavor.

Implementation of method

The next example will explain the application of the method. The objective is to find the Pareto-efficient solutions g_i^* for the minimization of two PMs ($C = 2$).

Let $G = \{g_1, g_2, g_3, g_4, g_5, g_6\}$ be a set of $n = 6$ genes. The values for the PMs per gene are $g_1(1, 4)$; $g_2(3,4)$; $g_3(5,6)$; $g_4(7,5)$; $g_5(3,2)$; $g_6(4,1)$. This leads to having $\{m_1^1 m_2^1 m_3^1 m_4^1 m_5^1 m_6^1\} = \{1, 3, 5, 7, 3, 4\}$ and $\{m_1^2 m_2^2 m_3^2 m_4^2 m_5^2 m_6^2\} = \{4, 4, 6, 5, 2, 1\}$. Figure 3 shows the MCO problem for the case of minimization of both performance measures and its mathematical solution.

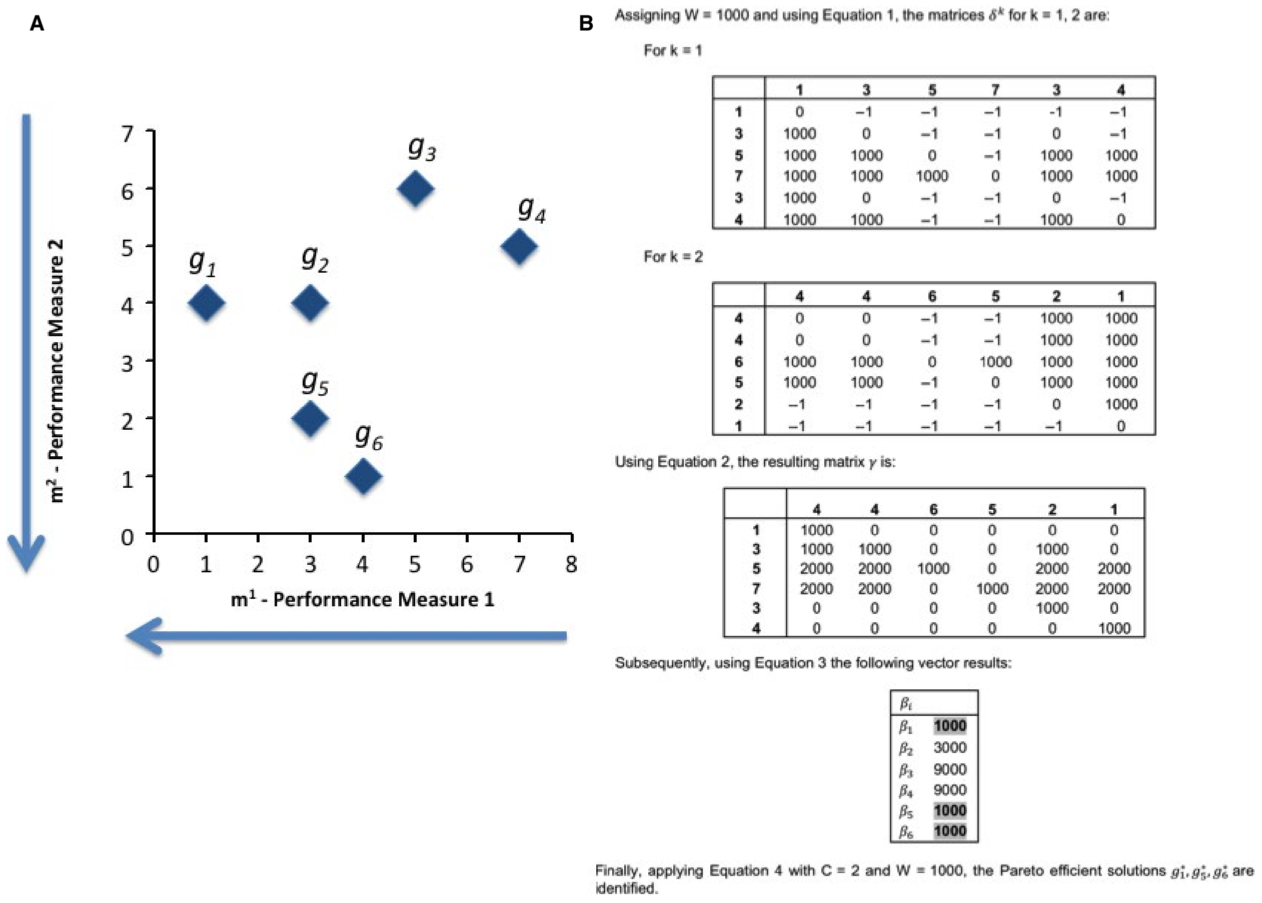
Finally, applying equation (4) the Pareto-efficient solutions implies comparing the beta values to a threshold of 2000. The solutions g_i^*, for this MCO problem are g_1^*, g_5^*, g_6^*. These solutions are graphically shown in Figure 4.

Analysis and Results of Lung cancer Microarray

In this analysis, the database with GEO identifier GDS3257 was used. This database was first reported by Landi MT and collaborators [31]. The database contains measures of relative expression for 22,283 genes from 107 samples: 49 control and 58 cancerous tissues. The age of the donors was between 44 and 79 years old. Samples were from never smokers, former smokers, and current smokers (See Fig. 5).

Case 1: Comparative analysis between different pairs of subgroups

For the first analysis the group of never smokers was considered and the comparison was between controls and cancer samples. There were fifteen controls (HNS) and sixteen



Assigning W = 1000 and using Equation 1, the matrices δ^k for k = 1, 2 are:

For k = 1

	1	3	5	7	3	4
1	0	−1	−1	−1	-1	−1
3	1000	0	−1	−1	0	−1
5	1000	1000	0	−1	1000	1000
7	1000	1000	1000	0	1000	1000
3	1000	0	−1	−1	0	−1
4	1000	1000	−1	−1	1000	0

For k = 2

	4	4	6	5	2	1
4	0	0	−1	−1	1000	1000
4	0	0	−1	−1	1000	1000
6	1000	1000	0	1000	1000	1000
5	1000	1000	−1	0	1000	1000
2	−1	−1	−1	−1	0	1000
1	−1	−1	−1	−1	−1	0

Using Equation 2, the resulting matrix γ is:

	4	4	6	5	2	1
1	1000	0	0	0	0	0
3	1000	1000	0	0	1000	0
5	2000	2000	1000	0	2000	2000
7	2000	2000	0	1000	2000	2000
3	0	0	0	0	1000	0
4	0	0	0	0	0	1000

Subsequently, using Equation 3 the following vector results:

β_i	
β_1	1000
β_2	3000
β_3	9000
β_4	9000
β_5	1000
β_6	1000

Finally, applying Equation 4 with C = 2 and W = 1000, the Pareto efficient solutions g_1^*, g_5^*, g_6^* are identified.

Figure 3. Graphical and Mathematical representation of the sample problem. (A) The six candidate solutions of the sample problem. (B) Mathematical formulation of the problem.

cancer (CNS) samples. The absolute value of the differences of means and medians for each gene were calculated. The analysis in MatLab tool was run in a computer with 4 GB of memory RAM and 2.66 GHz CPU. Due to this memory constraint, the Pareto-efficient frontier was found in a tournament fashion [32] as explained next. The 22,283 genes were divided into three groups: two groups of 7500 and one of 7283 genes. The MatLab tool was used to find the locally efficient frontier in each group. Finally, the genes in each one of the three efficient frontiers were analyzed together to find the global Pareto-efficient frontier. It is important to point out that the order of the partition and input of the data does not affect the final efficient frontier, as this is a case of explicit full comparison. In one criterion, the process would be similar to finding the tallest person in a room by picking the tallest one in different subgroups and comparing the local winners in the end to find the global winner. With enough computing memory, partitioning the data is not necessary. For each group, the locally nondominated subset was identified (Fig. 6). Then the locally nondominated subsets were used to obtain the globally optimal Pareto-efficient frontier, as seen in Figure 7. For this first analysis *RAGE* and *SPP1* are the genes in the global Pareto-efficient frontier. It is important to recall that the user does not need to normalize or use a threshold value to achieve this result.

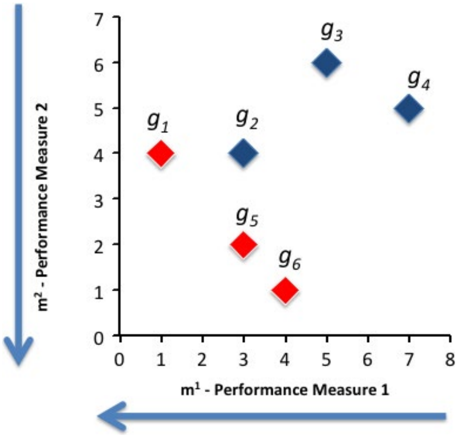


Figure 4. Pareto-efficient solutions for the sample problem.

For the second analysis the selected group was the one for the current smokers, and again the comparison was between control current smokers (HCS) and cancer current smokers (CCS). The group had 16 samples for HCS and 24 samples for CCS. The process was performed as in the previous analysis. In this case the global Pareto-efficient frontier had just one gene, the *SPP1*.

A third analysis compared groups HNS and CCS. There were 15 HNS samples and 24 CCS samples and *RAGE*

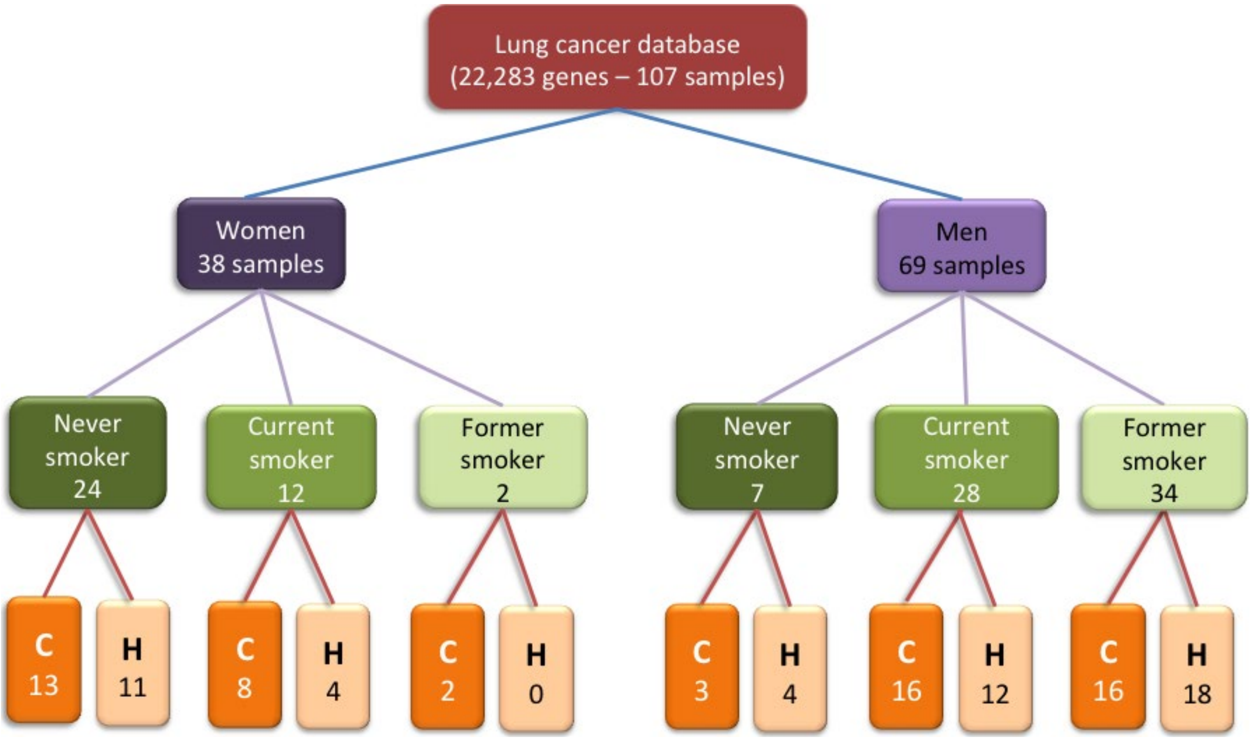


Figure 5. Organization of database GDS3257. "C" indicates cancer and "H" indicates controls.

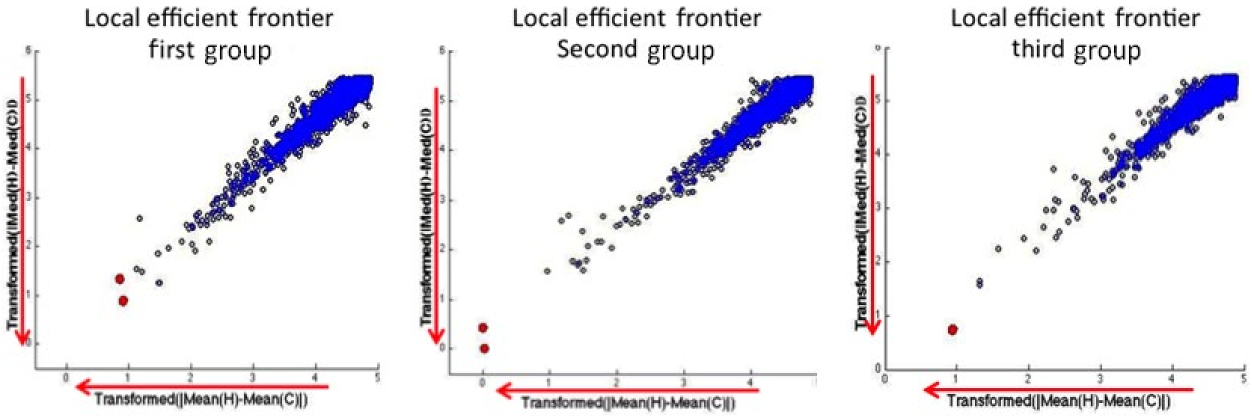


Figure 6. Local Pareto-efficient frontiers of all groups. For the first and second groups, two genes are at the local Pareto-efficient frontier, and only one gene for the third group.

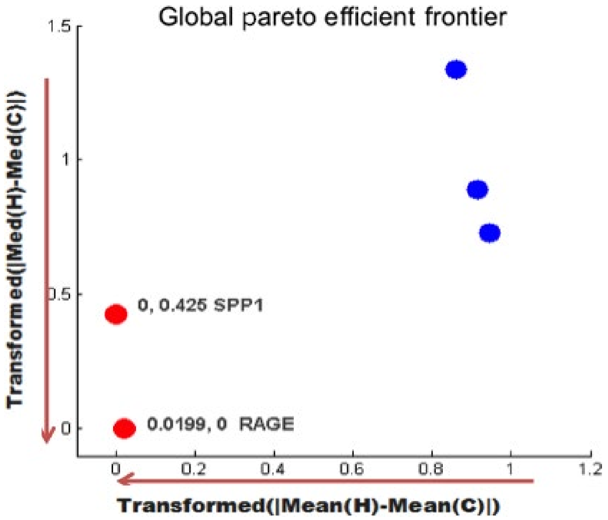


Figure 7. Globally-optimal Pareto-efficient frontier consisting of *RAGE* and *SPP1* genes.

was the only solution. In the fourth analysis comparing the 16 HCS samples and the 16 CNS samples the gene with the largest change is *SPP1*.

In a fifth analysis, the 15 HNS samples are compared with the 16 samples of HCS, resulting in three genes in the efficient frontier: *RPS4Y1*, *CYP1B1*, and *XIST*. When, in the sixth analysis, the comparison is done for the cancer group between nonsmokers (CNS, 16 samples) and current-smokers (CCS, 24 samples) there is only one gene present in the solution: *XIST*.

Figure 8 shows a summary of the six analyses between never smoker versus current smoker in cancer and control tissues. The circles on the left side represent the controls never smoker (HNS) and controls current smoker (HCS) tissues, while the circles on the right hand side represent the cancer never smoker (CNS) and cancer current smoker (CCS) tissues. Additionally, the upper circles represent never smoker tissues, whereas the lower circles symbolize current smoker tissues.

Case 2: Analysis of lung cancer in women: never smoker versus current smoker in cancer and control tissues

Figure 9 shows the result with the same analysis described before, but selecting only for women's tissues. For this representation, the only efficient solution is *RAGE*, which showed a large change when controls (HNS and HCS) were compared to cancer.

Case3: Analysis of lung cancer in men: never smoker versus current smoker

Figure 10 shows the results with an analysis similar to the one described before, but using only men samples. For this representation, as in previous cases, *RAGE* and *SPP1* showed significant changes when controls (HNS or HCS) were compared to cancer.

Table 2 shows the scientific names of genes obtained in the Pareto-efficient frontier from all previous analyses.

Case 4: The possibility of meta-analysis with four performance measures: a prototype for meta-analysis

In the previous analyses two PMs (absolute value of differences in means and absolute value of differences in medians) were used. In this analysis, MCO meta-analysis is carried out using four PMs, which were the absolute value of differences in medians for each group [16]. The medians were used for their nonparametric characteristics,

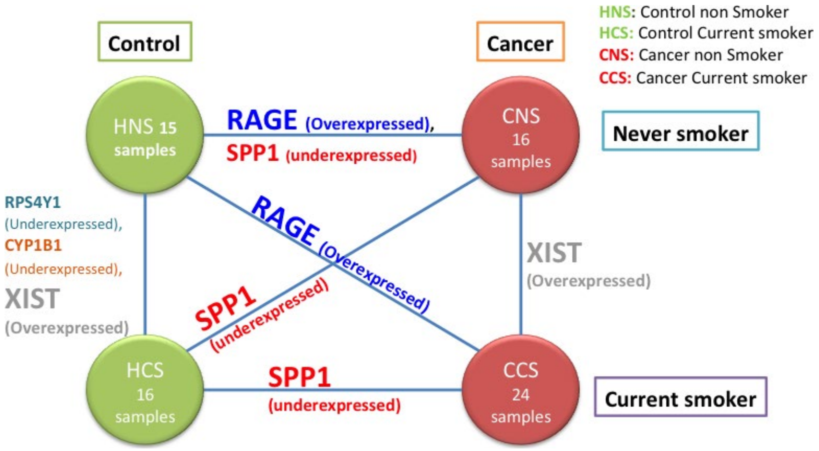


Figure 8. Diagram representing six analyses between four different conditions (HNS, HCS, CNS, CCS). The edges of the graph list the genes in the associated Pareto-efficient frontier.

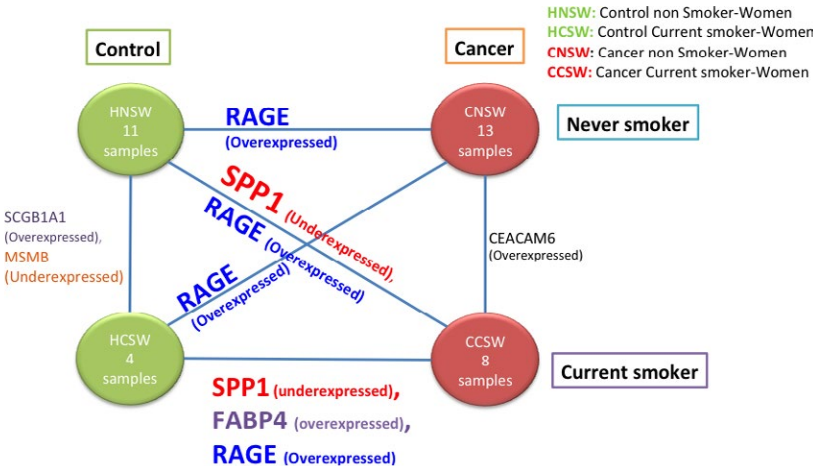


Figure 9. Diagram representing six analyses between four different conditions for women samples (HNSW, HCSW, CNSW, CCSW). The edges of the graph list the genes in the associated Pareto-efficient frontier.

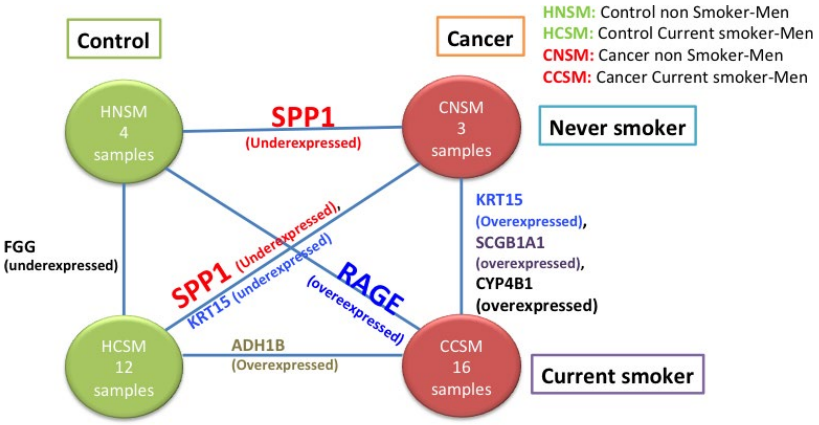


Figure 10. Diagram representing six analyses between four different conditions for men samples (HNSM, HCSM, CNSM, CCSM). The edges of the graph list the genes in the associated Pareto-efficient frontier.

Table 2. Scientific names the genes identified in the analyses of this work.

Official symbol	Official name
RAGE	Receptor for Advanced Glycosylation End Products
SPP1	Secreted PhosphoProtein 1
XIST	X Inactive Specific Transcript (nonprotein coding)
RPS4Y1	Ribosomal Protein S4, Y-linked 1
CYP1B1	Cytochrome P450, family 1, subfamily B, polypeptide 1
FABP4	Fatty Acid Binding Protein 4, adipocyte
CEACAM6	Carcinoembryonic Antigen-related Cell Adhesion Molecule 6 (nonspecific cross reacting antigen)
MSMB	Microseminoprotein, beta
SCGB1A1	Secretoglobin, family 1A, member 1 (uteroglobin)
ADH1B	Alcohol Dehydrogenase 1B (class I), beta polypeptide
CYP4B1	Cytochrome P450, family 4, subfamily B, polypeptide 1
KRT15	Keratin 15
FGG	Fibrinogen Gamma chain

as it has been habitual in analyses previously carried out by our group. Continuing with the case, the difference in medians between the groups of cancer and control tissues is calculated for each one of the 22,283 genes in the database. These groups are: HNS (15 samples) versus CNS (16 samples), HNS (15 samples) versus CCS (24 samples), HCS (16 samples) versus CNS (16 samples), HCS (16 samples) versus CCS (24 samples) as seen in Figure 11.

In this way, the four PMs were calculated and MCO was applied to find the genes with high variation levels of the relative expressions throughout all PMs. Among all the 22,283 genes and using four PMs, the genes with high variation were *RAGE* and *SPP1*. This analysis supports the potential of the proposed method for meta-analysis.

Table 3. Summary from Pareto-efficient frontier genes and their related cancer.

Gene name	Examples of cancer types where the gene is involved	Reference
RAGE	Pancreas, colon and prostate, colorectal, gastric, liver, lung	[42–47]
SPP1	Oral, lung, bone, bladder, prostate, cervical, breast, head and neck, liver	[36, 48–54]
XIST	Meninges, breast, ovarian	[55–57]
RPS4Y1	Meninges	[55]
CYP1B1	Lung, cervical, head and neck, prostate	[58], [59, p. 1], [60], [61, p. 1]
Genes from the analysis with data pertaining only to Women		
FABP4	Prostate and breast, ovarian	[62, 63]
MSMB	Prostate	[64]
CEACAM6	Head and neck, breast, colon, lung	[65–68]
SCGB1A1*	Lung	[69]
Genes from the analysis with data pertaining only to Men		
FGG	Liver	[70]
KRT15	Lung, ovarian	[71, 72]
ADH1B	Esophageal, colorectal, head and neck	[73–75]
CYP4B1	Bladder	[76]
SCGB1A1*	Lung	[69]

Discussion

Table 3 presents the summary of genes obtained from eighteen analyses of the lung cancer database. The first group consists of the genes obtained from an analysis

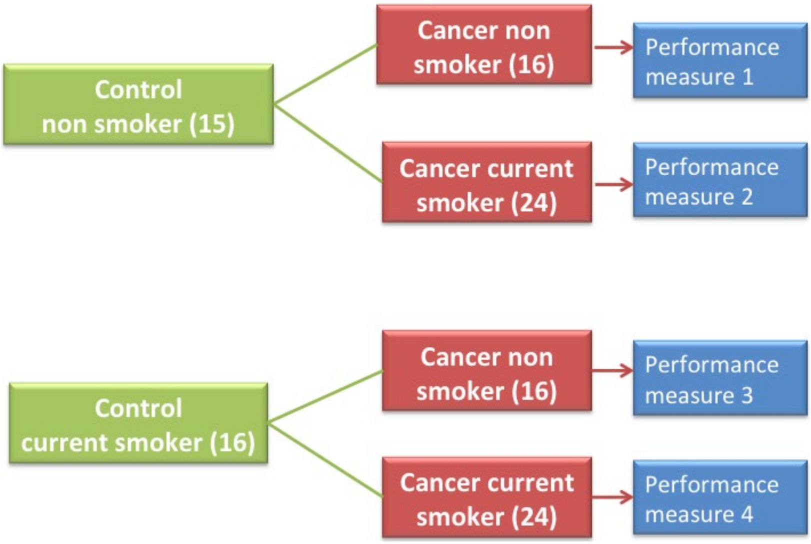


Figure 11. Groups for meta-analysis with four PMs.

from both women and men. The second group is obtained from a group analysis of only women, and the last group is the results of a group analysis of only men. The common genes for all groups are *RAGE*, *SPP1*, and *SCGB1A1*. The products of these three genes are associated with inflammatory processes and different cancer types including lung [23, 33–38]. From this table, three important conclusions are obtained. First, those genes found in the literature as biomarkers such as *CYPIB1* [39] and FABP4 [40] validate our method. Secondly, those genes found in the literature as associated with other types of cancer, such as, *XIST* (a nonprotein coding gene) [41], among others, could eventually be validated and proposed as lung cancer biomarkers with the precursor that they are important genes for other types of cancer and could uncover relations between different cancer types. Also, these genes could possibly have a relation with lung cancer biomarkers in a pathway to be researched. Third, the genes that do not have any evidence found in literature indicating or any identification as biomarkers in other types of cancer, are the opportunities for discovery and thus, offer the potential for a larger contribution.

Conclusions

The method applied in this study could be used to analyze data related to biological health care research where microarrays and other –omics are the driving experiments for exploration.

The tool coded in MatLab can currently analyze five criteria, that is, it can be used to meta-analyze up to five different datasets in one run. The discrimination rate makes the analysis very manageable. Also, the results will be friendly and conveniently available to physicians or biological researchers, as this analysis does not require normalization, preference of objectives, parameter adjustments by user, or the definition of a threshold value. Importantly, the mathematical treatment is easy to translate into a functional code of the analyst's choice.

In the case study in lung cancer the general conclusions are: *RAGE* and *SPP1* showed large change between controls and cancer. Moreover, *SPP1* showed a large change between the Control Current Smoker and the Cancer Nonsmoker, and *RAGE* showed large change between Control Never Smoker and Cancer Current Smoker. Also, *XIST* showed a large difference when comparing Never Smoker and Current Smoker (both in control and cancer tissues). The fact that these genes have already been related to cancer, indicate the capability of the proposed method.

It should be taken into consideration that *SCGB1A1* was found in this study to have an over expression in both Cancer Never Smoker and Cancer Current Smoker. However, *SCGB1A1* expression has been found to be reduced in current smokers [60]**.** Further biological studies should be performed to validate the results obtained by the methodology applied in this study.

Currently we are working on improving the usability of the code to make the method more amicable to the users. Future work should include further investigation of the potential biomarkers proposed in this document and experimental validation. It is certainly also envisioned the future tests of the proposed method with different –omics.

Acknowledgments

This work was supported in part 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."

Conflict of interest

None declared.

References

1. The International Agency for Research on Cancer (IARC), "Global battle against cancer won't be won with treatment alone Effective prevention measures urgently needed to prevent cancer crisis," Mar. 2014.
2. Mohammadi, A., M. H. Saraee, and M. Salehi. 2011. Identification of disease-causing genes using microarray data mining and Gene Ontology. BMC Med. Genomics 4:12.
3. Research and Markets Offers Report: Microarray Markets. 2013. *Prof. Serv. Close - Up*.
4. Velazquez, M. A., D. Claudio, and A. R. Ravindran. 2010. Experiments in multiple criteria selection problems with multiple decision makers. International Journal of Operation Research 7:413–428.
5. Augusto, O. B., F. Bennis, and S. Caro. 2012. A new method for decision making in multi-objective optimization problems. Pesqui. Oper. 32:331–369.
6. Statnikov, R. B., A. Bordetsky, and A. Statnikov. 2005. Multicriteria analysis of real-life engineering optimization problems: statement and solution. Nonlinear Anal. Theory Methods Appl. 63:e685–e696.
7. Huang, P.-H., J.-S. Tsai, and W.-T. Lin. 2010. Using multiple-criteria decision-making techniques for eco-environmental vulnerability assessment: a case study on the Chi-Jia-Wan Stream watershed, Taiwan. Environ. Monit. Assess. 168:141–158.
8. Statnikov, R., J. Matusov, and A. Statnikov. 2012. Multicriteria engineering optimization problems:

statement, solution and applications. J. Optim. Theory Appl. 155:355–375.

9. Sánchez-Peña, M. L., C. E. Isaza, J. Pérez-Morales, C. Rodríguez-Padilla, J. M. Castro, and M. Cabrera-Ríos. 2013. Identification of potential biomarkers from microarray experiments using multiple criteria optimization. Cancer Med. 2:253–265.
10. Rodríguez, A. B., E. Niño, J. M. Castro, M. Suarez, and M. Cabrera. 2012. Injection molding process windows considering two conflicting criteria: simulation results. ASME Proceedings 3:1447–1453.
11. Hurd, P. J., and C. J. Nelson. 2009. Advantages of next-generation sequencing versus the microarray in epigenetic research. Brief. Funct. Genomic. Proteomic. 8:174–183.
12. McCall, M. N., P. N. Murakami, M. Lukk, W. Huber, and R. A. Irizarry. 2011. Assessing affymetrix GeneChip microarray quality. BMC Bioinformatics 12:137.
13. Mohamed Salleh, A. H., M. S. Mohamad, S. Deris, and R. M. Illias. 2013. A review on pathway analysis software based on microarray data interpretation. Int. J. Bio-Sci. Bio-Technol. 5:149–157.
14. Tyagi, V., and A. Mishra. 2013. A survey on different feature selection methods for microarray data analysis. Int. J. Comput. Appl. 67:36–40.
15. Lyttleton, O., A. Wright, D. Treanor, P. Quirke, and P. Lewis. 2011. Extending the tissue microarray data exchange specification for inclusion of data analysis results. J. Pathol. Inform. 2:17.
16. Biomarkers Definitions Working Group. 2001. Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. Clin. Pharmacol. Ther. 69:89–95.
17. Isaza, C. E.., L. Uribe, H. A. Pérez, C. Rodríguez, and M. Cabrera-Ríos. 2009. Cancer diagnosis through microarray analysis using the Mann-Whitney statistical test. Proceedings of the 2009 Industrial Engineering Research Conference, 736–741.
18. Glasser, S. P., and S. Duval. 2008. Meta-analysis. Pp. 159–177 *in* S. P. Glasser, ed. Essentials of clinical research. Springer, The Netherlands.
19. Li, J., R. J. Coates, M. Gwinn, and M. J. Khoury. 2010. Steroid 5-{alpha}-reductase Type 2 (SRD5a2) gene polymorphisms and risk of prostate cancer: a HuGE review. Am. J. Epidemiol. 171:1–13.
20. Wang, X., D. D. Kang, K. Shen, C. Song, S. Lu, L.-C. Chang, et al. 2012. An R package suite for microarray meta-analysis in quality control, differentially expressed gene analysis and pathway enrichment detection. Bioinforma. Oxf. Engl. 28:2534–2536.
21. Gan, Z., J. Wang, N. Salomonis, J. C. Stowe, G. G. Haddad, A. D. McCulloch, et al. 2014. MAAMD: a workflow to standardize meta-analyses and comparison of affymetrix microarray data. BMC Bioinformatics 15:69.
22. Page, M. J., J. E. McKensie, M. Chau, S. E. Green, and A. Forbes. 2015. Methods to select results to include in meta-analyses deserve more consideration in systematic reviews. J. Clin. Epidemio. S0895–4356: 00105–5.
23. Begg, C. B. 2008. Meta-analysis methods for diagnostic accuracy. J. Clin. Epidemio. 61:1081–1082.
24. Manchikanti, L., S. Datta, H. S. Smith, and J. A. Hirsch. 2009. Evidence-based medicine, systematic reviews, and guidelines in interventional pain management: part 6. Systematic reviews and meta-analyses of observational studies. Pain Physician 12:819–850.
25. Haidich, A. B. 2010. Meta-analysis in medical research. Hippokratia 14(Suppl 1):29–37.
26. Moher, D., D. J. Cook, S. Eastwood, I. Olkin, D. Rennie, and D. F. Stroup. 1999. Improving the quality of reports of meta-analysis of randomised controlled trials: the QUOROM statement. Quality of Reporting of Meta-Analyses. Lancet 354:1896–1900.
27. Hutton, B., G. Salanti, D. M. Caldwell, A. Chaimani, C. H. Schmid, C. Cameron, et al. 2015. The PRISMA extension statement for reporting of systematic reviews incorporating network meta-analyses of health care interventions" checklist and explanations. Ann. Intern. Med. 162:777–784.
28. Rajapakse, J. C., and P. A. Mundra. 2013. Multiclass gene selection using pareto-fronts. IEEEACM Trans. Comput. Biol. Bioinforma. 10:87–97.
29. Deb, K. 2001. Multi-objective optimization using evolutionary algorithms. Wiley, New York, NY, USA.
30. Greco, S., and M. Ehrgott. 2005. Multiple criteria decision analysis: state of the art surveys. Springer, Boston, MA, USA.
31. Landi, M. T., T. Dracheva, M. Rotunno, J. D. Figueroa, H. Liu, A. Dasgupta, et al. 2008. Gene expression signature of cigarette smoking and its role in lung adenocarcinoma development and survival. PLoS ONE 3:e1651.
32. Cabrera-Rios, M. Using data clustering to aid the solution of multiple criteria optimization problems through data envelopment analysis. [Online]. Available: http://www.academia.edu/2852113/Using_Data_Clustering_to_Aid_the_Solution_of_Multiple_Criteria_Optimization_Problems_through_Data_Envelopment_Analysis (accessed 20 March 2014).
33. Sims, G. P., D. C. Rowe, S. T. Rietdijk, R. Herbst, and A. J. Coyle. 2010. HMGB1 and RAGE in inflammation and cancer. Annu. Rev. Immunol. 28:367–388.
34. Marinakis, E., G. Bagkos, C. Piperi, P. Roussou, and E. Diamanti-Kandarakis. 2014. Critical role of RAGE in lung physiology and tumorigenesis: a potential target of therapeutic intervention? Clin. Chem. Lab. Med.: 52:189–200.

35. Riccardo, F., M. Arigoni, G. Buson, E. Zago, M. Iezzi, D. Longo, et al. 2014. Characterization of a genetic mouse model of lung cancer: a promise to identify Non-Small Cell Lung Cancer therapeutic targets and biomarkers. BMC Genom. 15 (Suppl. 3):S1.
36. Zhang, H., H. B. Liu, D. M. Yuan, Z. F. Wang, Y. F. Wang, and Y. Song. 2014. Prognostic value of secreted phosphoprotein-1 in pleural effusion associated with non-small cell lung cancer. BMC Cancer 14:280.
37. Lakind, J. S., S. T. Holgate, D. R. Ownby, A. H. Mansur, P. J. Helms, D. Pyatt, et al. 2007. A critical review of the use of Clara cell secretory protein (CC16) as a biomarker of acute or chronic pulmonary effects. Biomarkers 12:445–467.
38. Guerra, S., M. M. Vasquez, A. Spangenberg, M. Halonen, and F. D. Martinez. 2013. Serum concentrations of club cell secretory protein (Clara) and cancer mortality in adults: a population-based, prospective cohort study. Lancet Respir. Med. 1:779–785.
39. Li, M. Y., Y. Liu, L. Z. Liu, A. W. Kong, Z. Zhao, B. Wu, et al. 2015. Estrogen receptor alpha promotes smoking-carcinogen-induced lung carcinogenesis via cytochrome P450 1B1. J. Mol. Med. (Berl). [Epub ahead of print].
40. Hancke, K., D. Grubeck, N. Hauser, R. Kreienberg, and J. M. Weiss. 2010. Adipocyte fatty acid-binding protein as a novel prognostic factor in obese breast cancer patients. Breast Cancer Res. Treat. 119:367–377.
41. Chaligné, R., and E. Heard. 2014. X-chromosome inactivation in development and cancer. FEBS Lett. 588:2514–2522.
42. Kang, R., D. Tang, N. Schapiro, T. Loux, K. Livesey, T. Billiar, et al. 2014. The HMGB1/RAGE inflammatory pathway promotes pancreatic tumor growth by regulating mitochondrial bioenergetics. Oncogene 33:567–577.
43. Sparvero, L. J., D. Asafu-Adjei, R. Kang, D. Tang, N. Amin, J. Im, et al. 2009. RAGE (Receptor for Advanced Glycation Endproducts), RAGE ligands, and their role in cancer and inflammation. J. Transl. Med. 7:17.
44. Dahlmann, M., A. Okhrimenko, P. Marcinkowski, M. Osterland, P. Herrmann, J. Smith, et al. 2014. RAGE mediates S100A4-induced cell motility via MAPK/ERK and hypoxia signaling and is a prognostic biomarker for human colorectal cancer metastasis. Oncotarget 5:3220–3233.
45. Xu, X. C., X. Abuduhadeer, W. B. Zhang, T. Li, H. Gao, and Y. H. Wang. 2013. Knockdown of RAGE inhibits growth and invasion of gastric cancer cells. Eur. J. Histochem. 57:e36.
46. Yaser, A.-M., Y. Huang, R.-R. Zhou, G.-S. Hu, M.-F. Xiao, Z.-B. Huang, et al. 2012. The Role of Receptor for Advanced Glycation End Products (RAGE) in the proliferation of hepatocellular carcinoma. Int. J. Mol. Sci. 13:5982–5997.
47. Buckley, S. T., and C. Ehrhardt. 2010. The Receptor for Advanced Glycation End Products (RAGE) and the Lung. BioMed Res. Int. 2010.
48. Mardani, M., A. Andisheh-Tadbir, B. Khademi, M. J. Fattahi, S. Shafiee, and M. Asad-Zadeh. 2014. Serum levels of osteopontin as a prognostic factor in patients with oral squamous cell carcinoma. Tumour Biol. J. Int. Soc. Oncodevelopmental Biol. Med. 35:3827–3829.
49. Dalla-Torre, C. A., M. Yoshimoto, C.-H. Lee, A. M. Joshua, S. R. de Toledo, A. S. Petrilli, et al. 2006. Effects of THBS3, SPARC and SPP1 expression on biological behavior and survival in patients with osteosarcoma. BMC Cancer 6:237.
50. Zaravinos, A., G. I. Lambrou, D. Volanis, D. Delakas, and D. A. Spandidos. 2011. Spotlight on differentially expressed genes in urinary bladder cancer. PLoS ONE, 6:e18255.
51. Ding, Z., C.-J. Wu, G. C. Chu, Y. Xiao, D. Ho, J. Zhang, et al. 2011. SMAD4-dependent barrier constrains prostate cancer growth and metastatic progression. Nature 470:269–273.
52. Thomas, A., U. Mahantshetty, S. Kannan, K. Deodhar, S. K. Shrivastava, C. Kumar-Sinha, et al. 2013. Expression profiling of cervical cancers in Indian women at different stages to identify gene signatures during progression of the disease. Cancer Med. 2:836–848.
53. Das, S., R. S. Samant, and L. A. Shevde. 2011. Hedgehog signaling induced by breast cancer cells promotes osteoclastogenesis and osteolysis. J. Biol. Chem. 286:9612–9622.
54. Weber, G. F., G. S. Lett, and N. C. Haubein. 2010. Osteopontin is a marker for cancer aggressiveness and patient survival. Br. J. Cancer 103:861–869.
55. Tabernero, M. D., A. B. Espinosa, A. Maillo, O. Rebelo, J. F. Vera, J. M. Sayagues, et al. 2007. Patient gender is associated with distinct patterns of chromosomal abnormalities and sex chromosome linked gene-expression profiles in meningiomas. Oncologist 12:1225–1236.
56. Sirchia, S. M., S. Tabano, L. Monti, M. P. Recalcati, M. Gariboldi, F. R. Grati, et al. 2009. Misbehaviour of XIST RNA in breast cancer cells. PLoS ONE 4:e5559.
57. Huang, K.-C., P. H. Rao, C. C. Lau, E. Heard, S.-K. Ng, C. Brown, et al. 2002. Relationship of XIST expression and responses of ovarian cancer to chemotherapy 1 this work was partly supported by NIH Grants CA70216 and GM 59920 (to S-W. N.). 1. Mol. Cancer Ther. 1:769–776.

58. Morissette, M. C., M. Lamontagne, J.-C. Berube, G. Gaschler, A. Williams, C. Yauk, et al. 2014. Impact of cigarette smoke on the human and mouse lungs: a gene-expression comparison study. PLoS ONE 9:e92498.
59. Li, Y., S.-Q. Tan, Q.-H. Ma, L. Li, Z.-Y. Huang, Y. Wang, et al. 2013. CYP1B1 C4326G polymorphism and susceptibility to cervical cancer in Chinese Han women. Tumour Biol. J. Int. Soc. Oncodevelopmental Biol. Med. 34:3561–3567.
60. Shatalova, E. G., A. J. P. Klein-Szanto, K. Devarajan, E. Cukierman, and M. L. Clapper. 2011. Estrogen and cytochrome P450 1B1 contribute to both early- and late-stage head and neck carcinogenesis. Cancer Prev. Res. Phila. Pa. 4:107–115.
61. Beuten, J., J. A. L. Gelfond, J. J. Byrne, I. Balic, A. C. Crandall, T. L. Johnson-Pais, et al. 2008. CYP1B1 variants are associated with prostate cancer in non-Hispanic and Hispanic Caucasians. Carcinogenesis 29:1751–1757.
62. Nieman, K. M., H. A. Kenny, C. V. Penicka, A. Ladanyi, R. Buell-Gutbrod, M. R. Zillhardt, et al. 2011. Adipocytes promote ovarian cancer metastasis and provide energy for rapid tumor growth. Nat. Med. 17:1498–1503.
63. Herroon, M. K., E. Rajagurubandara, A. L. Hardaway, K. Powell, A. Turchick, D. Feldmann, et al. 2013. Bone marrow adipocytes promote tumor growth in bone via FABP4-dependent mechanisms. Oncotarget 4:2108–2123.
64. Lou, H., H. Li, M. Yeager, K. Im, B. Gold, T. D. Schneider, et al. 2012. Promoter variants in the MSMB gene associated with prostate cancer regulate MSMB/NCOA4 fusion transcripts. Hum. Genet. 131:1453–1466.
65. Cameron, S., L. M. de Long, M. Hazar-Rethinam, E. Topkas, L. Endo-Munoz, A. Cumming, et al. 2012. Focal overexpression of CEACAM6 contributes to enhanced tumourigenesis in head and neck cancer via suppression of apoptosis. Mol. Cancer. 11:74.
66. Mukhopadhyay, A., T. Khoury, L. Stein, P. Shrikant, and A. K. Sood. 2013. Prostate derived Ets transcription factor and Carcinoembryonic antigen related cell adhesion molecule 6 constitute a highly active oncogenic axis in breast cancer. Oncotarget 4:610–621.
67. Ilantzis, C., L. Demarte, R. A. Screaton, and C. P. Stanners. 2002. Deregulated expression of the human tumor marker CEA and CEA family member CEACAM6 disrupts tissue architecture and blocks colonocyte differentiation. Neoplasia (New York, N.Y.) 4:151–163.
68. Singer, B. B., I. Scheffrahn, R. Kammerer, N. Suttorp, S. Ergun, and H. Slevogt. 2010. Deregulation of the CEACAM expression pattern causes undifferentiated cell growth in human lung adenocarcinoma cells. PLoS ONE 5:e8747.
69. Chang, S. H., S. G. Mirabolfathinejad, H. Katta, A. M. Cumpian, L. Gong, M. S. Caetano, et al. 2014. T helper 17 cells play a critical pathogenic role in lung cancer. Proc. Natl Acad. Sci. USA 111:5664–5669.
70. Zhu, W.-L., B.-L. Fan, D.-L. Liu, and W.-X. Zhu. 2009. Abnormal Expression of Fibrinogen Gamma (FGG) and plasma level of fibrinogen in patients with hepatocellular carcinoma. Anticancer Res. 29:2531–2534.
71. Boyero, L., A. Sanchez-Palencia, M. T. Miranda-Leon, F. Hernandez-Escobar, J. A. Gomez-Capilla, and M. E. Farez-Vidal. 2013. Survival, classifications, and desmosomal plaque genes in non-small cell lung cancer. Int. J. Med. Sci. 10:1166–1173.
72. Jiang, H., X. Lin, Y. Liu, W. Gong, X. Ma, Y. Yu, et al. 2012. Transformation of epithelial ovarian cancer stemlike cells into mesenchymal lineage via EMT results in cellular heterogeneity and supports tumor engraftment. Mol. Med. 18:1197–1208.
73. Yang, S.-J., A. Yokoyama, T. Yokoyama, Y.-C. Huang, S.-Y. Wu, Y. Shao, et al. 2010. Relationship between genetic polymorphisms of ALDH2 and ADH1B and esophageal cancer risk: a meta-analysis. World J. Gastroenterol. 16:4210–4220.
74. Crous-Bou, M., G. Rennert, D. Cuadras, R. Salazar, D. Cordero, H. Saltz Rennert, et al. 2013. Polymorphisms in alcohol metabolism genes ADH1B and ALDH2, alcohol consumption and colorectal cancer. PLoS ONE 8:e80158.
75. Hakenewerth, A. M., R. C. Millikan, I. Rusyn, A. H. Herring, K. E. North, J. S. Barnholtz-Sloan, et al. 2011. Joint effects of alcohol consumption and polymorphisms in alcohol and oxidative stress metabolism genes on risk of head and neck cancer. Cancer Epidemiol. Biomark. Prev. Publ. Am. Assoc. Cancer Res. Cosponsored Am. Soc. Prev. Oncol. 20:2438–2449.
76. Sasaki, T., M. Horikawa, K. Orikasa, M. Sato, Y. Arai, Y. Mitachi, et al. 2008. Possible relationship between the risk of Japanese bladder cancer cases and the CYP4B1 genotype. Jpn. J. Clin. Oncol. 38:634–640.
77. Li, W. 2012. Volcano plots in analyzing differential expressions with mRNA microarrays. J. Bioinform. Comput. Biol. 10:1231003.
78. Cui, X., and G. A. Churchill. 2003. Statistical tests for differential expression in cDNA microarray experiments. Genome Biol. 4:210.

Appendix A1 Matlab code to run the proposed MCO strategy

```
%Program: Analysis of Pareto frontier for five criteria
%Author: Katia I Camacho-Caceres
dataT = load(&‘data5Criteria.txt&#x2019;); %Load data
[x,y] = size(dataT);
data = dataT(:,2:end);
[n,m]=size(data);

c1 = 1000*ones(n,n,m); % matrix for first condition
for j=1:m
     for a=1:n
          for b=1:n
               if data(a,j) == data(b,j)
                    c1(a,b,j)=0;
               elseif data(a,j)<data(b,j)
                    c1(a,b,j)=-1;
               end
          end
     end
end

% Sum two matrices
c2=zeros(n,n);
for a=1:n
     for b=1:n
          if c1(a,b,1)+c1(a,b,2)+c1(a,b,3)+c1(a,b,4)+c1(a,b,5)==0
               c2(a,b)=2500;
          elseif c1(a,b,1)+c1(a,b,2)+c1(a,b,3)+c1(a,b,4)+c1(a,b,5)==1000
               c2(a,b)=2500;
          elseif c1(a,b,1)+c1(a,b,2)+c1(a,b,3)+c1(a,b,4)+c1(a,b,5)==2000
               c2(a,b)=2500;
          elseif (c1(a,b,1)+c1(a,b,2)+c1(a,b,3)+c1(a,b,4)+c1(a,b,5))==3000
               c2(a,b)=2500;
          elseif (c1(a,b,1)+c1(a,b,2)+c1(a,b,3)+c1(a,b,4)+c1(a,b,5))==4000
               c2(a,b)=2500;
          elseif (c1(a,b,1)+c1(a,b,2)+c1(a,b,3)+c1(a,b,4)+c1(a,b,5))==5000
               c2(a,b)=5000;
          end
     end
end

% Find non-dominated set (ncd), and dominated set (cd)
cnd = zeros(x,y);
cd = zeros(x,y);

i=0;
j=0;
for a=1:x
     sumfila=sum(c2(a,:));
      if sumfila>=5000; % condition for dominated set
          i=i+1;
          cd(i,:)=dataT(a,:);
     else  %  non-dominated  set
          j=j+1;
          cnd(j,:)=dataT(a,:);
     end
end

index  =  1:x;
disp([round(index&#x2019;)  cd]);
disp([round(index&#x2019;)  cnd]);

%Show  non-dominated  set  in  notepad  file
%Position  of  gene,f1  ,f2  ,f3  ,f4  ,f5  for  each  criteria
disp(‘      Non-dominated  set        ’);
cnd=cnd(1:j,:);
filecnd  =  fopen(‘cnd5CriteriaBio.txt&#x2019;,&#x2019;w&#x2019;);
fprintf(filecnd,&#x2019;%6s      %12s      %12s      %12s      %12s        %12s\r\n&#x2019;,&
#x2019;Posicion&#x2019;,&#x2019;F1&#x2019;,&#x2019;F2&#x2019;,&#x2019;F3&#x2019;,
‘F4&#x2019;,  ‘F5&#x2019;);
fprintf(filecnd,&#x2019;%6.4f      %12.4f      %12.4f      %12.4f      %12.4f
%12.4f\r\n&#x2019;,cnd&#x2019;);
fclose(filecnd);
```

Appendix A2 MCO compared to the use of a volcano plot

Volcano plot

In the literature there are many methods to detect DE genes from microarrays comparing two states. One of those methods is the Volcano Plot, which is a graphic method widely used by scientists and biologists [77]. This method is implemented in different software packages. The MCO method proposed in this research is here compared to the volcano plot in a series of analysis.

Volcano plot is a scatter plot built using p-values versus gene expression ratios of fold change (FC). This scatter plot used the negative log10-transformed p-values from the gene specific t-test against the log2 fold change. Genes with statistically significant differential expression according to the gene-specific t-test will lie above a horizontal threshold line. Genes with large fold-change values will lie outside a pair of vertical threshold lines [78].

P-values were calculated by unpaired t-test using the gene expression values from two experimental conditions: healthy and cancer tissues.

Fold Change is calculated as the ratio of the mean control and mean treatment observations. This is the extension of the difference of the logarithm of the control mean (y_1) and the logarithm of the control treatment (y_2):

$$\mathrm{FC} = \log(\overline{y_1}) - \log((\overline{y_2})$$

The ordinary *t*-statistic selects genes with low standard deviations while the fold-changes select genes with large shifts between control and treatment. Since the fold-changes and the ordinary *t*-statistic select different sets of genes, a researcher must decide whether a gene's importance is best quantified by the shift in expression or by the shift relative to the standard deviation.

According to the literature on the use of volcano plot, a researcher should choose the measure of differential expression based on the biological system of interest. On the one hand, if large absolute changes in expression are relevant to the system, then fold-change should be used; on the other hand, if changes in expression relative to the underlying noise are important, then a modified t-statistic is preferable. This, however, is the point of view from which this thesis wants to depart: the choice of ad-hoc threshold values to select genes.

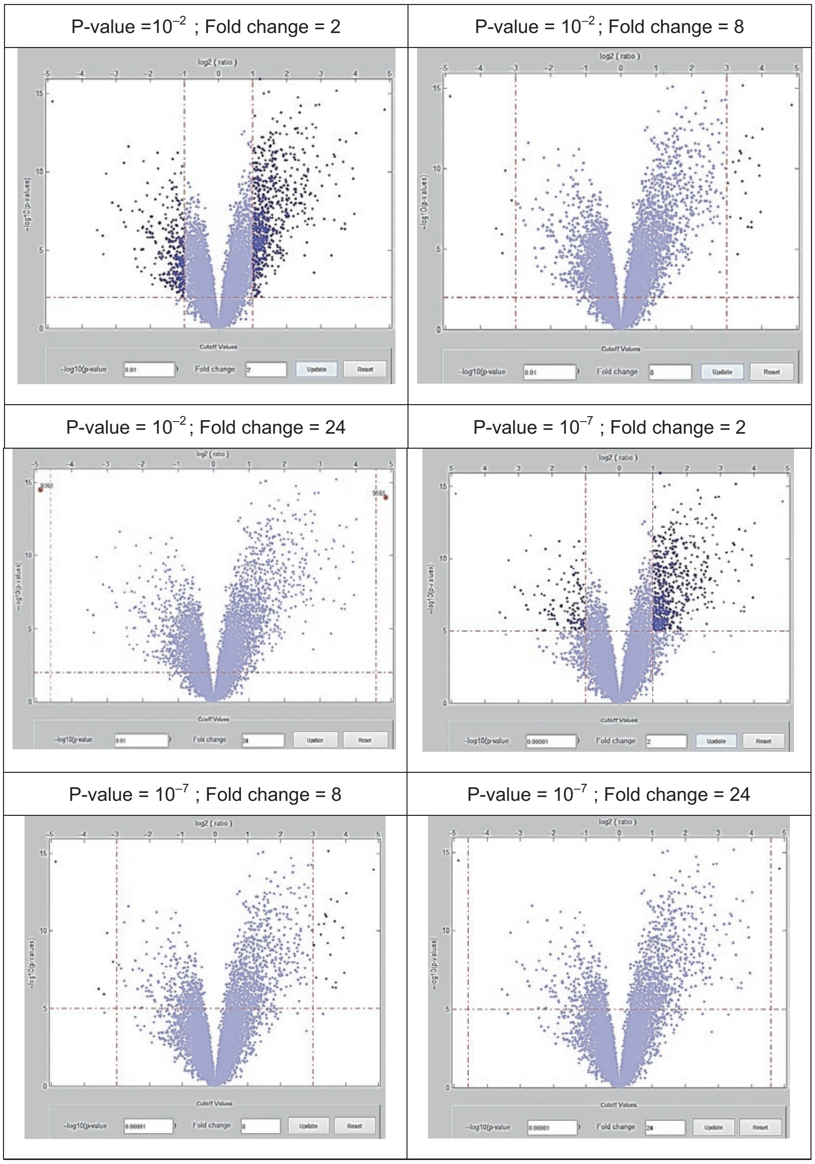
The analysis is required to choose threshold values for both measures to select important genes. The volcano plot is available in the bioinformatics toolbox for MatLab.

Given a particular microarray set with genetic expression levels measured. In two distinct states, the tool in MatLab obtains a p-value per gene using a t-test, and measures the FC in a logarithmic scale with base 2.

Comparison of volcano method using lung cancer microarray

The original database GDS3257 of lung cancer was used for this analysis. The samples HNS and CNS were used to build the Volcano plot. As mentioned previously, the Volcano plot requires the user to define thresholds for two parameters: p-value and FC to select genes. A 3^2 factorial experiment was used to explore these parameters as shown in Figure A1. The results are shown in Table A1.

From Table A1, it can be seen how the results depend highly in the user's selection of thresholds. The combinations that exactly match the output of MCO in this instance are the ones with FC = 24 at any of the values chosen for *P*-value in this experiment.



P-value =10^{-2} ; Fold change = 2

P-value = 10^{-2}; Fold change = 8

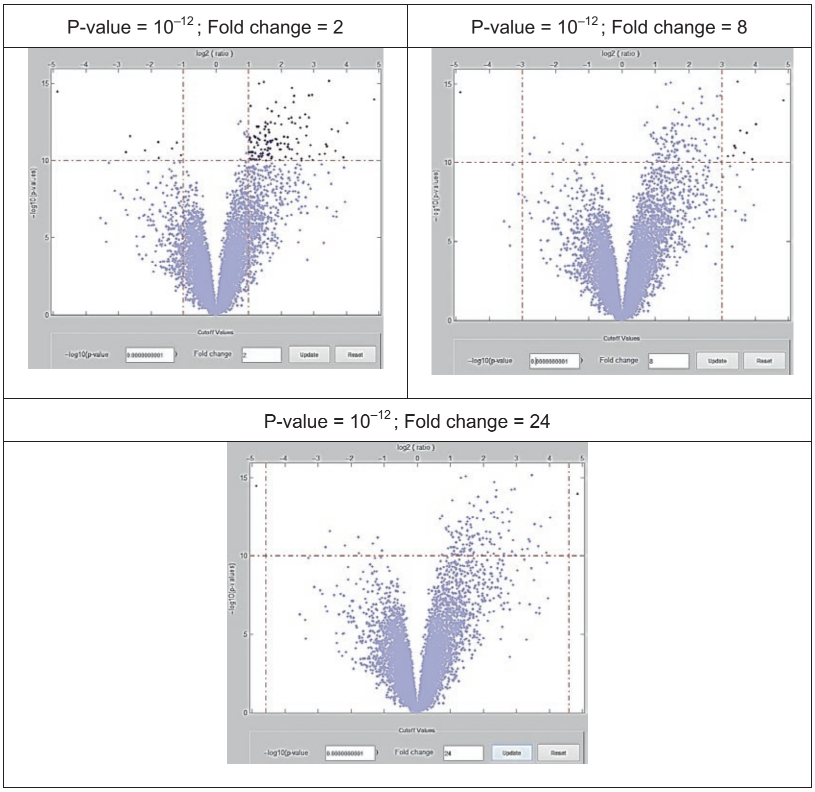
P-value = 10^{-2}; Fold change = 24

P-value = 10^{-7} ; Fold change = 2

P-value = 10^{-7} ; Fold change = 8

P-value = 10^{-7} ; Fold change = 24

Figure A1. Figures represent the results of genes with high DE using Volcano plot and varying the *P*-values and FC. The dark blue color indicates the genes most differentially expressed.



P-value = 10^{-12}; Fold change = 2

P-value = 10^{-12}; Fold change = 8

P-value = 10^{-12}; Fold change = 24

Figure A1. Continued

Table A1. Summary of important genes expressed using volcano plot.

P-value	Fold change	Differential expression	Overexpressed	Underexpressed
10^{-2}	2	934	645	289
10^{-2}	8	29	23	6
10^{-2}	24	2	1	1
10^{-7}	2	649	516	133
10^{-7}	8	27	22	5
10^{-7}	24	2	1	1
10^{-12}	2	130	121	9
10^{-12}	8	12	11	1
10^{-12}	24	2	1	1