OPTIMIZATION-DRIVEN META-ANALYSIS: THE SIMULTANEOUS SEARCH FOR CANCER BIOMARKERS WITH MULTIPLE MICROARRAY EXPERIMENTS

by

Katia Iris Camacho-Cáceres

A thesis submitted in partial fulfillment of the requirements for the degree of

MASTERS OF SCIENCE
in
INDUSTRIAL ENGINEERING

UNIVERSITY OF PUERTO RICO
MAYAGÜEZ CAMPUS
2014

Approved by:

________________________________ Saylisse Dávila, PhD Member, Graduate Committee	__________________ Date
________________________________ Jaime Seguel, PhD Co-President, Graduate Committee	__________________ Date
________________________________ Mauricio Cabrera-Ríos, PhD President, Graduate Committee	__________________ Date
________________________________ Pedro Torres, PhD Representative of Graduate Studies	__________________ Date
________________________________ Viviana Cesaní, PhD Chairperson of the Department	__________________ Date

ABSTRACT

In bioinformatics, it is possible to generate experimental data at a high pace. For example, microarrays can provide large amounts of data for genetic relative expression in illnesses of interest such as cancer. These data are stored and often times abandoned when new experimental technologies arrive. This work, re-examines lung cancer microarray data with a multiple criteria optimization-based strategy developed in our research group. This strategy does not require any adjustment of parameters by the user and is capable to converge consistently to important genes –potential biomarkers- even in the presence of multiple and incommensurate units across microarrays. In this thesis, three different cases were approached with the proposed method: lung cancer and leukemia, each using microarrays, and breast cancer with microRNA. The lists of resulting genes in the first two cases are provided with a discussion of their role in cancer, as well as the possible research directions for each of them. A list of microRNA sequences is also provided in the third case, emphasizing that this last case is to demonstrate the transferability of analysis ideas to other high throughput biological experiments. It is also recognized at this point that experimental validation is necessary to confirm the role of genes for which not enough evidence was found in the literature. Fundamentally, these genes with little reported information represent the best opportunities for biological discovery from existing data.

RESUMEN

En bioinformática es posible la generación de datos experimentales a un paso muy acelerado. Los microarreglos, por ejemplo, pueden proveer grandes cantidades de datos sobre expresión genética relativa en una enfermedad de interés como lo es el cáncer. Esta data es almacenada y en muchas ocasiones abandonada cuando nueva tecnología experimental es desarrollada. Este trabajo re examina la data proveniente de microarreglos de cáncer del pulmón utilizando una estrategia de optimización multicriterio que fue desarrollada previamente en nuestro grupo de investigación. Esta estrategia no requiere de ajustes en los parámetros por parte del usuario y es capaz de converger consistentemente a genes importantes considerados como biomarcadores potenciales, incluso en presencia de unidades múltiples e inconmensurables a través de los microarreglos. En esta tesis tres casos diferentes fueron abordados con el método propuesto: el cáncer de pulmón y leucemia, cada una usando microarreglos y el cáncer de seno con microRNA. La lista de los genes resultantes de los dos primeros casos, se provee con una discusión sobre su rol en el cáncer, al igual que las posibles direcciones de investigación para cada uno de ellos. Igualmente se provee la lista de secuencias de microRNA del tercer caso, enfatizando que este último puede ser transferido y aplicado a otros tipos de alto rendimiento de experimentos biológicos. Se reconoce en este punto que la validación experimental es necesaria para confirmar el rol de estos genes para los cuales no se encontró suficiente evidencia en la literatura. Fundamentalmente estos genes con poca información representan mejores oportunidades para descubrimientos biológicos en la data ya existente.

DEDICATION

This is to a very special person in my life, Benjamín Niño de Guzmán Camacho, my nephew. He captivated my heart with the intensity of his gaze. In his eyes I saw courage and a strong will to fight for his life. This gave the inspiration, and motivates to do my research. Also, he gave me strength and bravery that I needed during the Master Studies. Benjamín is a Leukemia survivor. Four days before he was one year old, he was detected with leukemia. In our family and his father’s side of the family never had any type of cancer before. However, my sweet Benjamín was the exception. Thanks to God, the doctors and science, after five years, Benjamin is a healthy, smart, good, and playful boy.

Katia Iris Camacho Cáceres

ACKNOWLEGMENTS

I am thankful to God for all the blessings during the Master Studies and for introducing me to wonderful people who are now my friends and family here in Puerto Rico. Also, for the love and encouragement I received from my family.

I want to give thanks with all my heart to my wonderful family. I thank you mom and dad, Teresa Cáceres and Mauro Camacho, for being my support, my role models, my strength and motivation in my life and throughout this journey. I give thanks to my sisters and brothers: Lisbeth Camacho who is an angel in my life; Erika Camacho, who is my example and courage. Both of you are my confidants, my allies, my best friends. Marco Camacho, Harry Camacho and Harold Camacho, who always give me support and refresh my moods with laughter and jokes. Lastly, my wonderful nephews: Sebastian Walker, Haylliq and Benjamin Niño de Guzman as well as my beautiful nieces: Qorichaska Walker, Abigail and Alba Camacho. Thanks to my beautiful and wonderful family for their love and truth that empowered me to achieve my dreams.

I am very thankful to my advisor, Dr. Mauricio Cabrera Ríos, for his advice, guidance, support and motivation during the Master Studies. Thanks for trusting in me and let me be part of this awesome research group. Furthermore, thank you for being my mentor and a friend.

I would like to express my sincere gratitude to a really special person, who is not my advisor. She is not working at UPRM. Also, she is not part of my committee members. However, she has an incredible dedication and love for science. Dra. Clara Isaza, thank you for your guidance and advice in the fields of Biology which were part of my thesis. Thank you for also being a friend.

I want to acknowledge the rest of my thesis committee, Dra. Saylisse Dávila and Dr. Jaime Seguel. Thank you for your collaborations and recommendations. For all of that excelled to a presentable and excellent thesis.

I am thankful to the University of Puerto Rico, Mayagüez Campus and the Industrial Engineering Department for all knowledge and for wonderful friends that I met there. I also want to acknowledge the Chemistry Department for sponsoring me although I was not their student. Dra. Ivelisse Padilla and Dra. Verónica Sanchez I am very grateful. Thank you for believing in me and for giving me the opportunity to continue my dreams. I am also thankful for CREST, Center of Research in Engineering, Science and Technology, for the opportunities and facilities.

I am very thankful to my wonderful research group, the Applied Optimization Group (AOG). I am really blessed to be a part of this group. Esmeralda Niño, Yaileen Méndez, Nitza García, Juan Rosas, Yaritza Santiago, Mary Carmen Acosta, Berenice Rodríguez, Jaileen Pérez, Enery Lorenzo, Yazeli Cruz and Hency Pérez. Thank you for your support, help, friendship, and for be my family in Puerto Rico.

Finally, I want to express my deepest gratitude to my wonderful friends who I met during my Master Studies. Thank you for taking part of my dreams, for bearing with my insane moments, my disappointments, my sadness and joys. Thank you Joan Asencio, Fanny Nina, Mariana Torres, Carlos Calero, Omar Lazo, Ulises Barajas, Moralba Domínguez, Danilo Rojas, Davis Chacón, and Fernando Mérida. Thank you for simply being there for me and sharing every moment in every step of my life and this journey. Thank you all for your friendship, companionship, your time, love and efforts. The Lord has truly blessed me in unimaginable and unexpected ways. Therefore, to Him I give all thanks, honor and glory. Through Him all of these encounters, events and circumstances have been possible.

TABLE OF CONTENTS

TABLE LIST

FIGURE LIST

CHAPTER 1: INTRODUCTION

1.1. Introduction

Cancer statistics for the US in 2014 include 1,665,540 new cancer cases and 585,720 cancer-related deaths [1]. Cancer can attack any organ or tissue of the body. There is no universal cure for cancer, in spite of the many discoveries made every day. This research intends to facilitate the discovering of new information related to cancer from the simultaneous analysis of multiple microarray experiments (meta-analysis) already available in specialized repositories [2–4]. It specifically targets the identification of potential cancer biomarker genes.

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 [5, 6]. However, these experiments are sometimes 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 meta-analysis search for cancer biomarkers in the past.

The process of putting together several analyses to obtain general conclusions across them is called, meta-analysis as mentioned before [7]. Meta-analysis is a valuable resource in the identification of biomarkers, as it allows adding up experimental evidence to strengthen biomarking signals in genes that might be otherwise overlooked. This work proposes the use of multi-criteria optimization to search for biomarkers in the presence of inconmensurate units and without the need to normalize data. To further preserve objectivity, the method will not require for the analyst to define thresholds, significance values, number of desired genes, or preference structures a priori.

The ideas presented in this thesis are developed in the context of cancer, capitalizing in previous work within our research group on multiple criteria optimization and data envelopment analysis for the detection of biomarkers using a single microarray database [8]. To this end, the case studies set forth involve lung cancer, leukemia, and breast cancer.

Lung cancer is one of the first causes of cancer related death across age and gender, while leukemia is the first cause of cancer-related deaths in people younger than 20 [1]. For the study of both illnesses, publicly available microarray databases contrasting measures in cancer-ridden tissue and healthy tissue are used. Aforementioned Multiple Criteria Optimization (MCO) [9] will be the tool to test these data through the direct application of the Pareto-optimality conditions. The idea is to represent each gene found in every distinct microarray database through multiple performance measures (PM) to be either minimized or maximized. Initially, one tries to maximize the absolute value of the difference of means or medians between two different conditions (healthy and cancer). Those genes that are found to change their relative expression the most across these states and across the different microarray experiments are proposed as potential cancer biomarkers. See Figure 1-1.

The capabilities offered by MCO in the analysis of microarray experiments can be extended to others - omics - experiments related to cancer, such as microRNA experiments, as approached in Chapter 6 of this thesis.

MicroRNAs (miRNA), are molecules resulting from large segments of RNA that are found in all diverse multicellular organisms [10]. miRNAs have been involved in a wide range of biological processes such as cell cycle control, apoptosis and several developmental and physiological processes including stem cell differentiation, cardiac and skeletal muscle development, immune responses, viral replication, among others [11]. Numerous types of cancer, heart diseases and neurological diseases have been associated to changes with microRNAs. These could represent important challenges for early cancer detection and diagnosis [12, 13].

1.2. Problem Description

Microarrays form databases of tens of thousands of measures of relative expression levels of genes from samples in two different conditions. The analysis of these measurements drives the identification of genes that are important in diseases such as cancer. Genes, whose relative expression is proven to be consistently different in, for example in healthy tissues and cancer tissues, are considered cancer biomarkers. Biomarkers have multiple uses in the fight against cancer [10], including early detection, diagnosis, prognosis, and - sometimes - treatment.

When multiple microarray databases are analyzed simultaneously in search of features that apply generally across all experiments, then one speaks of meta-analysis. In its more general form, and as the name suggests, meta-analysis uses a series of multiple analyses. Potential cancer biomarkers identified through meta-analysis are considered to have stronger evidence on their capabilities.

Detecting cancer biomarkers through meta-analysis of microarrays is complicated due to inconsistencies in the units of measurements across the different publicly-available microarray platforms, as well as the use of different normalization schemes. In this case, then the problem at hand is the identification of potential cancer biomarkers from the meta-analysis of microarray experiments in the presence of incommensurate units.

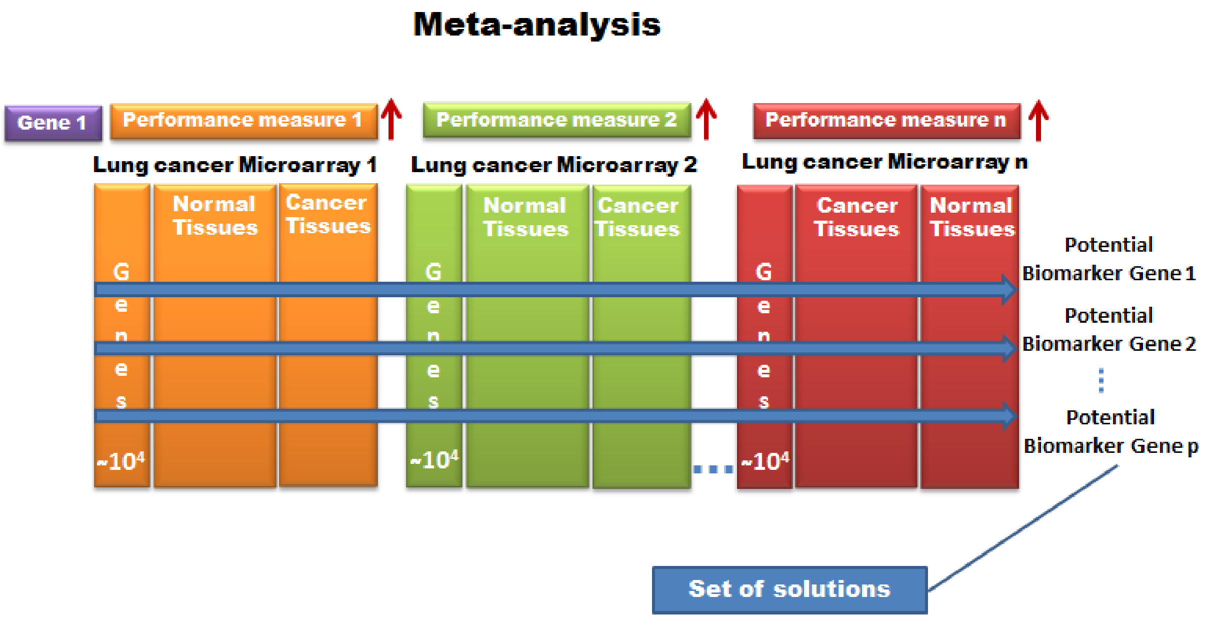


Figure 1-1: Graphic description of the meta-analysis problem with different Lung cancer microarrays. The results of this analysis will be a set of solutions in the form of potential biomarkers.

1.3. Objective

The objective of this thesis is to approach the meta-analysis of multiple and potentially incommensurate microarray experiments with multiple criteria optimization (MCO) aiming to detect potential cancer biomarkers. MCO, as proposed in this work, does not require either data normalization or the adjustment of parameters by the user. MCO is expected to converge to potential cancer biomarkers with objectivity and consistency.

1.4. Contribution of this Thesis

There are significant beneficial aspects derived from the proposed methodology when compared to the traditional bioinformatics techniques:

I. Data normalization is not required.

II. Preference among different measures or different genes is not required a priori.

III. There is no need for parameter adjustment by the user.

IV. There is no need for the definition of a threshold value to establish relevance of a gene.

V. Meta-analysis is achieved practically in an automatic fashion.

Previous results by our research group have shown that MCO has a high discriminatory power and effective detection rate [8]. This holds true even when the previous results were achieved with Data Envelopment Analysis, which identifies only the convex portions of the Pareto-efficient frontier. Novel to this work is the capability to identify both the convex and nonconvex portions of the efficient frontier, which is a clear analysis enhancement.

1.5. Contrast of this work with previous work in the Bio IE Lab

Uribe-Mastache [14] and Pérez-Vicente [15] started the work in microarray analysis within the Bio IE Lab, under the advice of Dr. Clara Isaza and Dr. Mauricio Cabrera-Ríos in 2008. Uribe-Mastache wrote a thesis entitled "*Metodología para el análisis de datos de Microarreglos para la detección de Cáncer"* (Methodology for cancer detection based on microarray data analysis) and Pérez-Vicente wrote a thesis entitled "*Diagnóstico de cáncer a partir de datos de Microarreglos"* (Cancer diagnosis with microarray data). Their joint work proposed statistical testing strategies for microarray analysis for the diagnosis of different types of cancer [14, 15]. No optimization procedure was explored in these theses. Mainly, the proposed strategies used in this thesis were focused on the extensive use the non-parametric Mann-Whitney test for differences of medians.

Sanchez-Peña [16] in 2010, continued the biologically-related research with the thesis *Identification of Potential Cancer Biomarkers through Multiple Criteria Optimization Using Microarray Data* [16]. The identification of potential cancer biomarkers from microarray data was solved as a MCO problem. Sanchez-Peña used a combination of two performance measurements (both p-values) obtained from a single microarray database. The efficient solution to this problem was found through Data Enveloped Analysis (DEA), where genes with lower p-values indicate stronger statistical significance [8]. This thesis proposes an improvement over DEA and generalizes the MCO formulation for meta-analysis.

Rodriguez-Yañez [17] wrote the thesis *Process Windows Considering two conflicting criteria: The injection molding case* in 2011 [17]. In this work, Rodriguez-Yañez developed a method of building process windows under multiple and conflicting criteria to aid in setting the processing conditions in injection molding operations. She developed the pairwise comparison scheme that enabled the application of the Pareto-optimality conditions adopted in this work. Improvement and automation of such scheme is achieved in this thesis.

1.6. Thesis Organization

This thesis is structured as follows: The second chapter is a review of the most relevant literature regarding cancer, microarrays, biomarkers, and gene selection across multiple high-throughput biological experiments. The third chapter presents a Multiple Criteria Optimization as a competitive approach to meta-analysis, the main part of this thesis. The fourth and fifth chapters show case studies in lung cancer and leukemia, respectively. The sixth chapter extends the application of this work to micro-RNAs, a newer and highly relevant class of high throughput biological experiments. The seventh chapter sets forth the comparison of the proposed method with another method in the literature: Volcano plot. The eighth chapter lays down the general conclusions of this work and establishes directions for future research.

CHAPTER 2: LITERATURE REVIEW

2.1. Cancer

According to the American Cancer Society (ACS), cancer is a set of illnesses characterized by uncontrolled cell replication caused by diverse reasons, driving to serious problems in our body including death [18]. In 2014, an estimated of 1,665,540 new cancer cases and 585,720 cancer deaths in United States will occur. Lung cancer, which is of interest to this work, is the primary cause of death by cancer in women and men. Lung cancer represents 28% of cancer deaths in men and 26% in women in the US [1]. Research to advance treatment against cancer is ubiquitous as there are many scientists dedicated to solve this health problem [19].

Our body is made up of trillions of cells [20]. Each cell has its own replication system. When this system loses control or presents a failure, the abnormal cell could grow uncontrollably. Also, there is a possibility to extend the proliferation across the system and cause cancer in different places of the body (a phenomenon called metastasis) [20]. At the present time, around 100 types of cancer are known.

The human being possesses different sets of systems in the body; each system is constituted by organs, and organs by tissues, which, in turn, have sets of cells with specific functions. Each cell contains deoxyribonucleic acid (DNA), which has the chemical machinery to produce proteins [18].

According to National Cancer Institute [18], DNA exists as two long, paired strands spiraled into the double helix. Each strand is made up of millions of chemical building blocks called bases. There are only four different chemical bases in DNA: cytosine, guanine, adenine, and thymine, but they can be arranged and rearranged in large numbers of ways. The order in which the bases occur determines the messages to be conveyed, much as specific letters of the alphabet combine to form words and sentences [21].

The human cell has 46 molecules of double-stranded DNA. Each DNA molecule is made up of 50 to 250 million bases stored in a chromosome[18].

A DNA molecule is composed by several working units called genes. A gene is a segment of DNA containing a specific set of instructions, which are used by cells to create a specific product (proteins among others). Every gene consists of thousand, even hundreds of thousands of chemical bases [18]. "For a cell to make a protein, the information from a gene is copied, base by base, from a strand of DNA into a strand of messenger RNA"[18]. Each of these cells contains the most of the complete genetic information to produce all types of proteins that the body need.

In some cases, our genes can be changed or altered in its molecular base, this process is called mutation. One example could be when there is a base insertion in the DNA sequence that make a change, or a disparity exists in some sections of DNA caused by that insertion. In consequence, this mutated gene could change the function of a protein or other molecule. Some gene mutations can produce cancer [18]. When a gene produces more protein than normal, it is called *overexpressed*, if, in the contrary, it produces less protein than normal, it is called *underexpressed*.

2.2. Microarray experiments

Microarray experiments produce a set of measurements of relative expression of genes from tissues or cell lines on an artificial platform called a microarray chip. This chip uses fluorescence detection to this end [22].

Microarray experiments are important because they foster the knowledge about the behavior of certain illnesses, including all forms of cancer. In cancer research, the main advantage of microarray experiments is the variety of analysis allowed by them and the possibility of comparison with other databases.

Microarray experiments have been very popular among researchers [23]. At some point, low cost, accessible DNA chips, and valuable information about the study were deemed advantages of this technology [3]. Also, 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 [3]. Moreover, in an attempt to study how different biological processes or pathway work in several organism, microarrays have been used as a powerful tool [24]. To analyze the obtained data, statistics have been used for these type of studies [23,24]. However, producing a standard analysis method has never been accomplished, even when it comes to normalization procedures to account for variation among or within microarrays [27].

2.3. Biomarkers

The Biomarkers Definition Working Group defines a biomarker as: "A characteristic that is objectively measured and evaluated as an indicator of normal biologic processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention" [28]. A biomarker could be used for diagnosis, prognosis, or discovery of genetic or non - genetic illnesses.

Currently, there are several types of biomarkers: genes, proteins, DNA, mitochondrial DNA, mitochondrial mutations, RNA, and microRNAs, among others. This thesis will focus on genes, RNA, and microRNA biomarkers.

RNA is a nucleic acid similar to DNA, stranded and contains information to regulate how genes are expressed to make proteins. MicroRNAs (miRNA) are small, mostly non-coding RNA gene products. They are molecules derived from larger segments of “precursor” RNA. miRNA are found in all diverse organisms. miRNAs regulates activity at multiple levels; specifically transcription, translation, and protein degradation [10]

In the literature, there are many methods to find potential biomarkers. Most of them are focused in statistical procedures including but not limited to t-test, Wilcoxon rank sum, Fisher exact test, Likelihood ratio test, edgeR, DESeq, baySeq, BBSeq, and Two-stage poisson model, among others

[23], [29]. This research adopts multiple criteria optimization and Pareto conditions to find biomarkers following the direction of our research group [8, 28], and proposes extending the application to this end through simultaneous analysis of multiple independent experiments, this is carrying out meta-analysis. As such, the perspective of this thesis is deterministic and departs from many of the considerations directly applicable to statistical procedures.

2.4. Gene Expression

Gene expression can be defined as a process where genes use information for the synthesis of a functional product [31]. Some cases, when a gene produces more than normal levels of a certain product, such as proteins, it can be described as “overexpressed” (See Figure 2-1). In the case that a gene produces less than normal levels it can be described as “under-expressed” (See Figure 2-2). In the analysis of each case study the sign of the difference of means or medians among two conditions (healthy and not-healthy) is used to assess if a biological entity is overexpress or underexpressed in the non-healthy material.

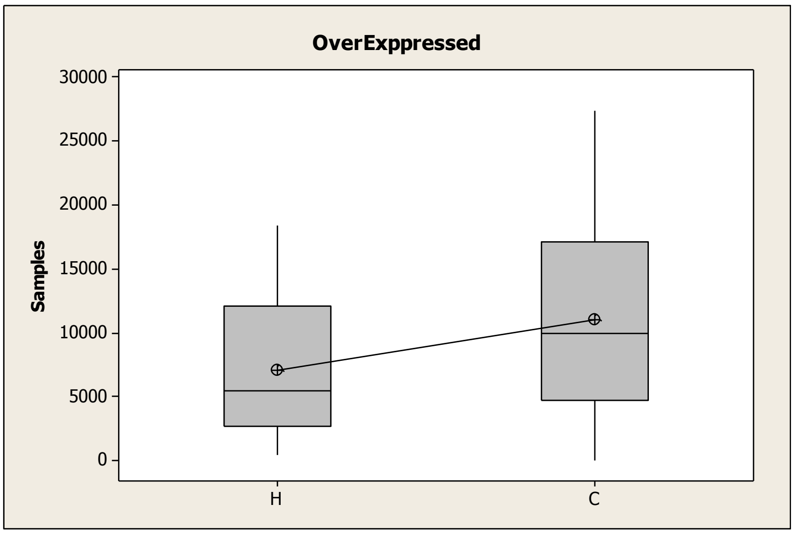


Figure 2-1: Example of overexpression of a gene.

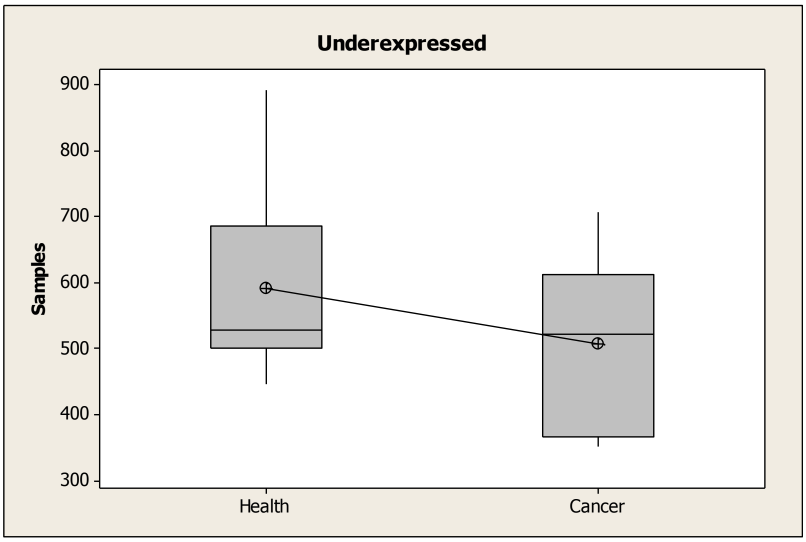


Figure 2-2: Example of underexpression of a gene

2.5. Meta-analysis

This thesis will focus on carrying out meta-analysis to analyze different microarrays with their own measurements, number of samples, and units. Microarray experiments involve hundreds of thousands of data. The data will be analyzed with a MCO formulation, applying Pareto conditions and resulting in potential biomarkers, as explained previously.

In the literature, it is possible to find different applications and examples of meta-analysis. Glasser and Duval, for example, in the book *Essentials of Clinical Research,* Chapter 10 indicate, "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" [7]. Meta-analysis is a method capable of taking independent, but associated studies to obtain a set of solutions through all studies.

Li and his research group led a systematic review and meta-analysis of different papers to determine whether two polymorphisms (V89L and A49T) are associated with the risk of prostate cancer. After the searching they analyzed 27 articles and reviews related to such risk [32] using a website tool called HUGE review. This tool identifies human genetics variations at one or more genetic localization (loci) in the chromosomes [33]. The review of these papers comprehends from the 1997 to 2007 years. The authors performed meta-analysis between two conditions (healthy and cancer tissues) using the Begg Mazumdar rank correlation test. The result of this meta-analysis was that prostate cancer was not associated with V89L and was probably not associated with A49T. On the other hand, R developed the software for microarray meta-analysis called MetaOmics. MetaOmics integrates Quality Control (Meta QC), Differentially Expressed (Meta DE), and Pathway (Meta pathway) [34]. MetaDE was designed to find candidate marker or genes biomarkes. MetaDE implements 12 different statistically based methods to carry out meta-analysis such: Fisher (Rhodes, 2012), Stouffer (1949), adaptively weighted Fisher (Aw), minimum p-value (minP), maximum p-value (maxP), among others. The application of these methods depends of the expertise of the users and the type of data for analysis. Often times, however, the final users –medical doctors, biologists- will find it difficult to properly select statistical methods and their parameters.

George Tseng and others wrote a review of microarray meta-analysis [35]. They reviewed 620 genomic meta-analysis papers. The authors show a general categorization on meta-analysis: descriptive review (2%), target gene meta-analysis (13%), and genome-wide meta-analysis (85%). Purposes of meta-analysis are: detection of differentially expressed (DE) genes or signaling pathway detection (66%), network or gene co-expression analysis (10%), classification analysis (8%), reproducibility or bias analysis (6%), and others (10%). Types of papers include: review paper (3%), biological application (60%), novel methodology (25%), and database/software (12%). Type of meta-analysis methods in DE gene detection are: combined p-values (42%), combined effect sizes (22%), combined ranks (9%), and direct merge (27%). According to this analysis, the major uses of microarray meta-analysis are for DE gene detection, which is akin to the detection of biomarkers.

Different ways to carry out microarray meta-analysis to discover gene expression [36] can be found in the literature. Many of these methods are used to find differential expressions of genes and biomarkers [37]. However, none of the described methods use multiple criteria optimization to analyze microarray, aside from our research group [8].

Zhuohui et al. (2014) research developed a tool called “MAAMD” [38]. They carry out meta-analysis using different Affymetrix microarrays data using the tool. The MAAMD tool automates the process to analyze microarrays and requires normalization and several statistical methods to detect DE genes. MAAMD is aimed to summarize all steps to analyze the Affymetrix microarrays existing in GEO repository. The tool automates multiple dataset downloading, data organization, normalization, and employs a series of statistics procedures for the determination of differential gene expression, multiple testing adjustments, clustering, and GO-Elite pathways all of that in one tool. To this end, the authors used Kepler, AltAnalyze and Bioconductor software packages. The parametric approach of these authors differs from our nonparametric approach. It is clear that multiple criteria optimization would differ from the reviewed approaches and would constitute a novelty in meta-analysis.

2.6. Multiple criteria Optimization

Optimization is very valuable to decision making and design processes [35, 36]. Optimization can make a system or design effective, functional, or in its most basic form, possible [9, 37] . This research will focus in the MCO problems to find a set of solutions with the best possible balance among multiple conflicting performance measures (PM) [42]. This thesis will use the cone of dominance formed by the linear convex combinations of the desired directions to find the best possible compromise, a method explored by Rodriguez in 2012 [17] in manufacturing applications. To find the best solutions between multiple criteria, Pareto optimality conditions are used.

CHAPTER 3: MULTIPLE CRITERIA OPTIMIZATION

3.1. Multiple Criteria Optimization (MCO)

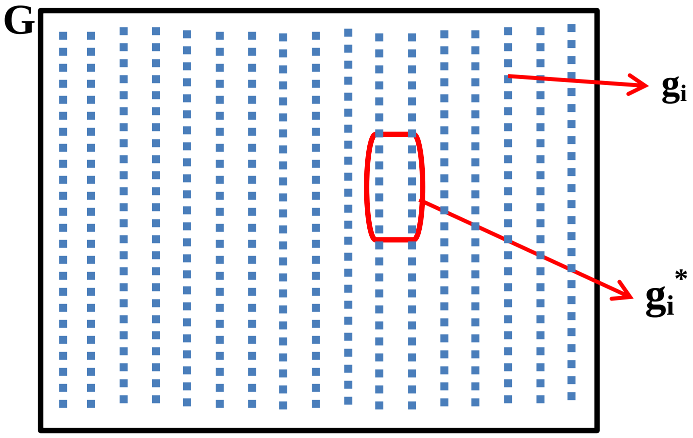
The MCO problem aims to choose the best compromising solutions among a set of candidate solutions assessed through at least two performance measures in conflict. The solutions to the MCO problem are called Pareto-efficient and determine the efficient frontier of the problem. The MCO problem has been approached in our group by Sánchez-Peña, et al (2013) [37] through Data Envelopment Analysis (DEA), and by Rodríguez-Yañez, et al (2014) [38] through full pairwise comparison. Sánchez-Peña, et al (2013) used DEA to detect potential cancer biomarkers using a single microarray database and multiple performance measures. This thesis approaches the larger problem of analyzing multiple microarray databases simultaneously, that is, to carry out meta-analysis. In Rodríguez-Yañez, et al (2014), the said full pairwise comparison scheme was developed to improve upon DEA's constraint of finding only the convex portion of the efficient frontier in the context of manufacturing. This full pairwise comparison finds the entire efficient frontier, both the convex and non-convex parts. Thus, in this thesis, the latter scheme is adopted and proposed to carry out meta-analysis in the context of –omics, in particular with microarrays and microRNAs.

In the literature, one particularly interesting paper used Pareto – concepts for gene selection: Rajapakse and Mundra applied F-scores and KW-scores to determine the Pareto-Frontier in the selection of genes. In this case, they divided the genes into different fronts (groups) by using parametric statistical methods (combine p-values, combine ranks). Moreover, they exclude some genes before applying Pareto conditions [43]. Do notice that using F-scores and KW-scores impose different assumptions a priori that are not necessary in our proposed approach.

3.2. MCO Problem Formulation

First, to explain the problem formulation, Figure 1-1 shows the elements of the graphical representation of the MCO problem. G denotes the universe, comprised of the n genes to be analyzed and g_i represents each gene of the problem, where i = 1, 2, …n. Figure 1-2 shows the space defined by two criteria 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 criterion or performance measure (PM). Then k = 1, 2, … C, and C is the number of criteria considered in the analysis. The Pareto efficient frontier in Figure 1-2 is formed by the genes g_i^*. These genes have indeed the best possible balances among the two criteria to be minimized. These genes are the ones proposed as potential biomarkers, as they dominate the rest of the genes.

When it comes to microarray analysis, the PMs of choice are usually related to gene expression. Looking for the most differentially expressed genes is akin to look for potential biomarkers, and it is a problem that can be casted as described up to this point.



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

Figure 3-1: Problem representation

According to K. Deb [44] and M. Ehrgott [45] the Pareto efficient solutions must meet the Pareto-optimality conditions for minimization instances. 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 objectives.
2. The solution $X^{(1)}$ is strictly better than $X^{(2)}$ in at least one objective.

These conditions can be evaluated for every single pair of genes to find those that are not dominated. These are the Pareto-efficient genes that form the Pareto-efficient frontier of the MCO problem at hand. The idea in this thesis is, then, to automate this full pairwise comparison to detect the Pareto-efficient genes.

In the search for the most differentially expressed genes, the expressions of all candidate genes are measured in two states to be then further 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 ‘healthy’ and ‘cancer’. The absolute value of these differences will then be transformed to follow a minimization direction to match the illustration in Figure 1-2.

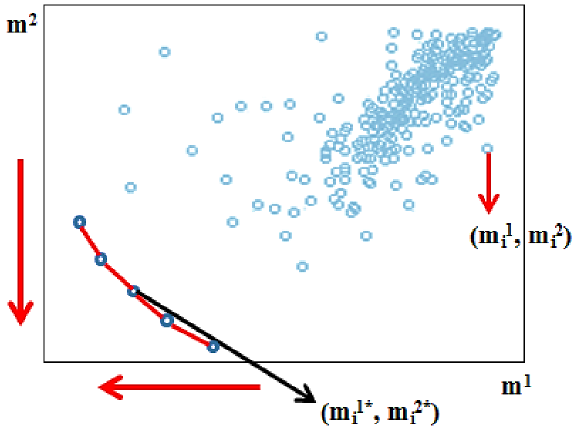


Figure 3-2: Represent the Pareto efficient frontier of the problem

Representation: $g_i => (m_i^1, m_i^2, \ldots, m_i^k, \ldots, m_i^C)$, For i = 1, 2, 3,…, n and k = 1, 2, …,C

Set of solutions: $g_i^* => (m_i^{1*}, m_i^{2*}, \ldots, m_i^{k*}, \ldots, m_i^{C*})$

To allow for the full pairwise comparison, a matrix δ^k. is developed for the k-th criterion. These will be C squared matrices size n built as follows:

For k = 1 => $\delta^1 =$

$$\begin{array}{ccccccc} & \boldsymbol{m_1^1} & \boldsymbol{m_2^1} & \cdots & \boldsymbol{m_j^1} & \cdots & \boldsymbol{m_n^1} \\ \boldsymbol{m_1^1} & \delta_{11}^1 & \delta_{12}^1 & \cdots & \delta_{1j}^1 & \cdots & \delta_{1n}^1 \\ \boldsymbol{m_2^1} & \delta_{21}^1 & \delta_{22}^1 & \cdots & \delta_{2j}^1 & \cdots & \delta_{2n}^1 \\ \vdots & \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ \boldsymbol{m_i^1} & \delta_{i1}^1 & \delta_{i2}^1 & \cdots & \delta_{ij}^1 & \cdots & \delta_{in}^1 \\ \vdots & \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ \boldsymbol{m_n^1} & \delta_{n1}^1 & \delta_{n2}^1 & \cdots & \delta_{nj}^1 & \cdots & \delta_{nn}^1 \end{array}$$

For the k-th criterion, then:

$$\delta^k = \begin{array}{ccccccc} & \boldsymbol{m_1^k} & \boldsymbol{m_2^k} & \cdots & \boldsymbol{m_j^k} & \cdots & \boldsymbol{m_n^k} \\ \boldsymbol{m_1^k} & \delta_{11}^k & \delta_{12}^k & \cdots & \delta_{1j}^k & \cdots & \delta_{1n}^k \\ \boldsymbol{m_2^k} & \delta_{21}^k & \delta_{22}^k & \cdots & \delta_{2j}^k & \cdots & \delta_{2n}^k \\ \vdots & \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ \boldsymbol{m_i^k} & \delta_{i1}^k & \delta_{i2}^k & \cdots & \delta_{ij}^k & \cdots & \delta_{in}^k \\ \vdots & \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ \boldsymbol{m_n^k} & \delta_{n1}^k & \delta_{n2}^k & \cdots & \delta_{nj}^k & \cdots & \delta_{nn}^k \end{array}$$

where:

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

W is defined as a large positive integer number used as a penalty. For this thesis, W=1000 is used.

Next, the matrix γ is defined as the sum of all matrices δ^k (k = 1, 2,…, C). First, however, the following formula is applied” $\alpha_{ij} = \sum_{k=1}^{C} \delta_{ij}^k$. For example Table 3-1 shows the results for α_{ij} when C=2:

Table 3-1: All the possible combinations of a minimization problem for two criteria.

Outcome number	$\boldsymbol{\delta_{ij}^1}$	$\boldsymbol{\delta_{ij}^2}$	$\boldsymbol{\alpha_{ij}}$	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*

For C=2, the following assessment applies

$$C = 2, \qquad \gamma_{ij} = \begin{cases} W, & if\ \alpha_{ij} \in \{0, W\} \\ 2W, & if\ \alpha_{ij} = 2W \\ 0, & otherwise \end{cases} \qquad ,\begin{cases} i = 1,2, \dots n \\ j = 1,2, \dots n \end{cases}$$

In general for any value C≥2

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

Then, in general, one can build the matrix γ:

$$\gamma = \begin{array}{ccccccc} & \boldsymbol{m_1} & \boldsymbol{m_2} & \cdots & \boldsymbol{m_j} & \cdots & \boldsymbol{m_n} \\ \boldsymbol{m_1^C} & \gamma_{11} & \gamma_{12} & \cdots & \gamma_{1j} & \cdots & \gamma_{1n} \\ \boldsymbol{m_2^C} & \gamma_{21} & \gamma_{22} & \cdots & \gamma_{2j} & \cdots & \gamma_{2n} \\ \vdots & \vdots & \vdots & \dots & \vdots & \dots & \vdots \\ \boldsymbol{m_i^C} & \gamma_{i1} & \gamma_{i2} & \cdots & \gamma_{ij} & \cdots & \gamma_{1n} \\ \vdots & \vdots & \vdots & \dots & \vdots & \cdots & \vdots \\ \boldsymbol{m_n^C} & \gamma_{n1} & \gamma_{n2} & \cdots & \gamma_{nj} & \cdots & \gamma_{nn} \end{array}$$

To find g_i^*, a vector β is built containing the sums of each row of matrix γ (Equation 3)

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

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

The Pareto efficient frontier g_i^* will contain all solutions that meet the Equation 4

$$g_i^* = \{g_i \mid \beta_i < CW, \; i = 1,2, \ldots n\} \qquad \text{(Equation 4)}$$

In other words: $g_i^* = \{m_i^{1*}, m_i^{2*}, \ldots m_i^{k*}, \ldots, m_i^{C*}\}$

This algorithm identifies all the solutions of the Pareto efficient frontier. The maximum number proved and coded in this thesis is five criteria. The MatLab code is available in Appendix 1.

3.3. Implementation of Method

The next example will explain the application of the method. Find the m_i^{k*} for minimization problem and two criteria.

Let G = $\{g_1, g_2, g_3, g_4, g_5, g_6\}$ a set of data. The PMs per gene are m_1^k (1, 4); m_2^k (3,4); m_3^k (5,6); m_4^k (7,5); m_5^k (3,2); m_6^k (4,1). Then m_i^1 = {1, 3, 5, 7, 3, 4} and m_i^2 = {4, 4, 6, 5, 2, 1}, where i = 6 and C = 2. The Figure 1-3 shows the MCO problem for the minimization case:

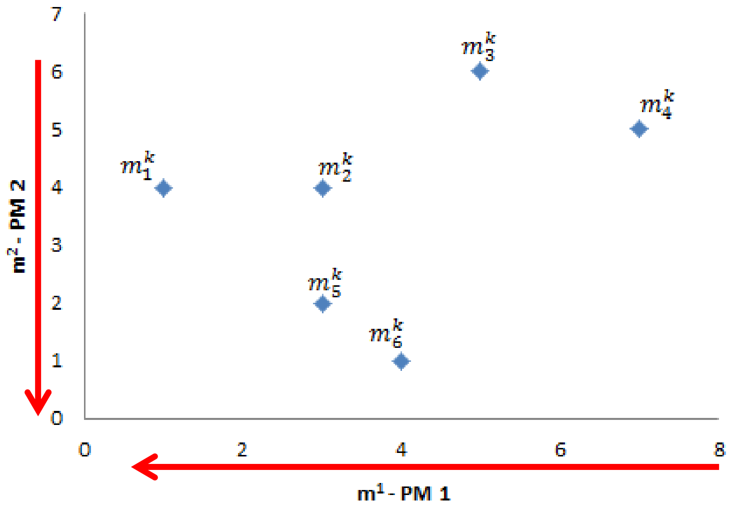


Figure 3-3: Represent the elements of the example

Assuming that W = 1000 and using the 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

Afterwards, using Equation 2, the resulting matrix γ is:

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

Subsequently, using Equation 3 and obtained the vector β:

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

Finally, using Equation 4 the Pareto efficient solutions, g_i^*, for this problem will be:

$g_i^* = \{(1,4),(3,2),(4,1)\}$, as graphically shown below.

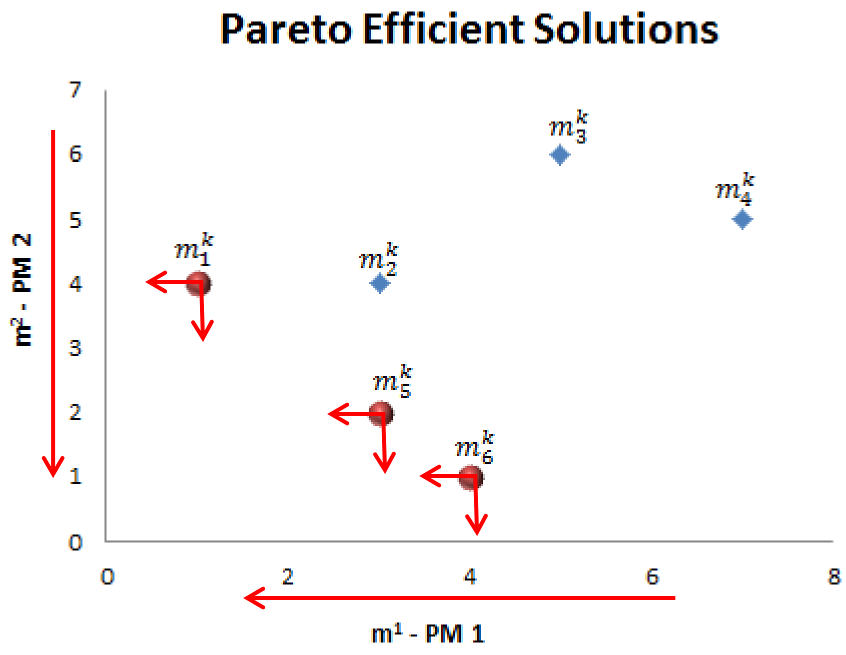


Figure 3-4: Represent the Pareto efficient solutions for the problem

3.3.1. An example with microarrays

Following the methodology and the previous example, the microarray database used for implementation of method was the GDS3257, first reported by Landi MT and collaborators. It measured the relative expression for 22,283 genes from 107 samples: 49 healthy and 58 lung cancer tissues. To formulate the MCO problem, first the means and medians for the control and case samples were calculated for each gene. Then, also for each gene, the absolute value of the differences between the two groups: means and medians were computed. These values correspond to the PMs of the problem. An important gene would display a large absolute difference between group medians or group means so this is the focus of the method: finding genes with large absolute differences in the said measures. It can be verified that when the groups under comparison show asymmetrical distributions, an ordering of genes based on difference of medians would be different than an ordering of genes based on difference of means, thereby imposing a conflict. An MCO problem in this context identifies the genes with the best possible balance between both performance measures.

For convenience of our method as coded in Matlab, the resulting optimization problem must be stated as series of minimization cases. A linear transformation is used for this purpose as described in [16] and Equation 5 using the notation defined in the previous sections

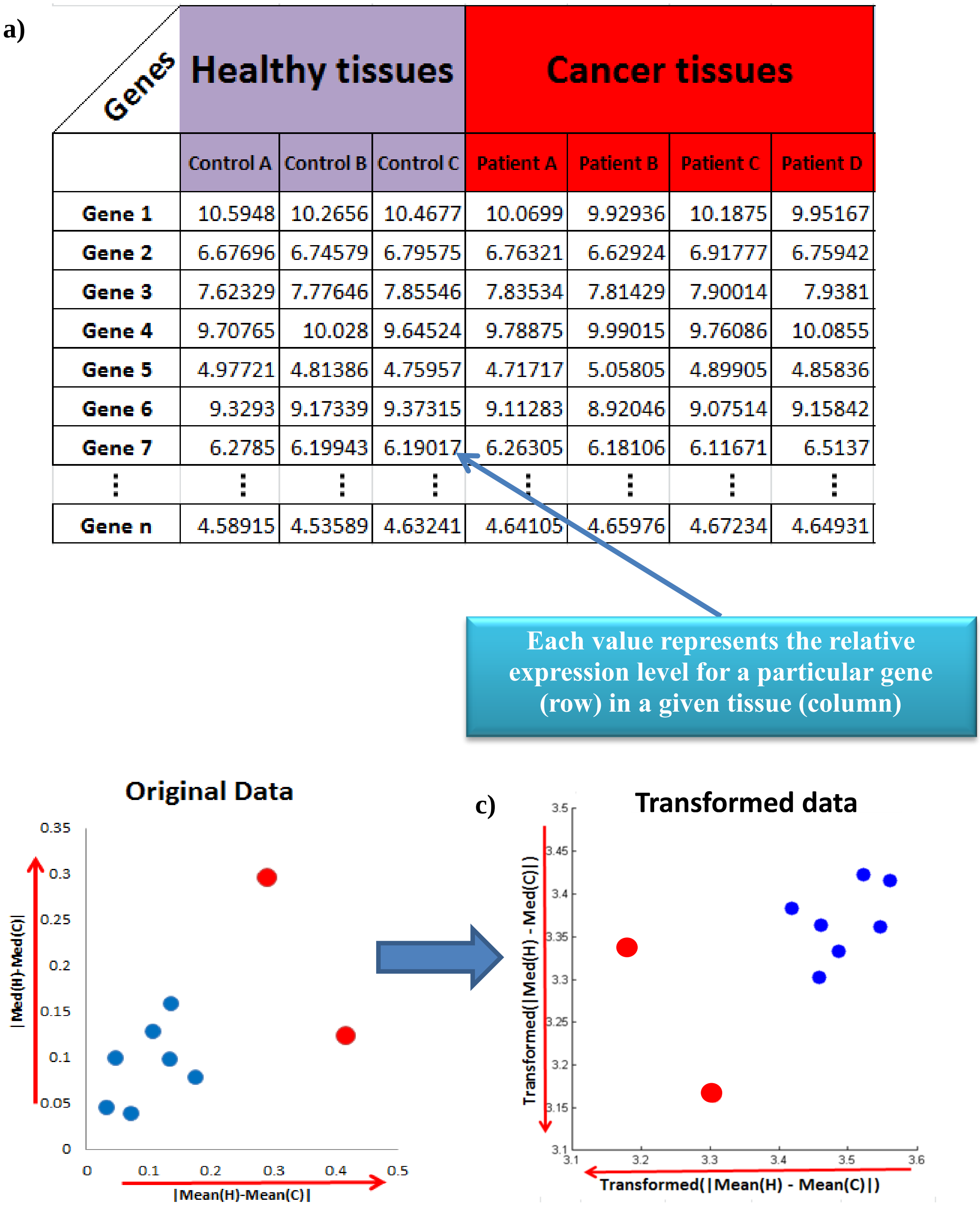
Transformed(m_i^k) $= \left[Max\{m_i^k\} + Min\{m_i^k\}\right] - m_i^k$, i = 1, 2, …n, k = 1, 2,… C

For this specific implementation case, in one PM and using H to symbolize the control material (healthy) and C to symbolize the treatment material (cancer):

$$\begin{aligned} &Transformed[|Mean(H) - Mean(C)|]_i \\ &\qquad = [Max(|Mean(H) - Mean(C)|) + Min(|Mean(H) - Mean(C)|)] \\ &\qquad - [|Mean(H) - Mean(C)|]_i \end{aligned}$$

i = 1, 2, … n

Using such transformation to arrive to an MCO instance of minimization of criteria, the result is depicted in Figure 3-5.



Genes	Healthy tissues			Cancer tissues			
	Control A	Control B	Control C	Patient A	Patient B	Patient C	Patient D
Gene 1	10.5948	10.2656	10.4677	10.0699	9.92936	10.1875	9.95167
Gene 2	6.67696	6.74579	6.79575	6.76321	6.62924	6.91777	6.75942
Gene 3	7.62329	7.77646	7.85546	7.83534	7.81429	7.90014	7.9381
Gene 4	9.70765	10.028	9.64524	9.78875	9.99015	9.76086	10.0855
Gene 5	4.97721	4.81386	4.75957	4.71717	5.05805	4.89905	4.85836
Gene 6	9.3293	9.17339	9.37315	9.11283	8.92046	9.07514	9.15842
Gene 7	6.2785	6.19943	6.19017	6.26305	6.18106	6.11671	6.5137
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
Gene n	4.58915	4.53589	4.63241	4.64105	4.65976	4.67234	4.64931

Each value represents the relative expression level for a particular gene (row) in a given tissue (column)

Figure 3-5: a) General representation of multiple criteria optimization problem using microarrays with two performance measures. b) The important genes would display a large absolute difference between group medians or group means. c) Method coded in MatLab, the resulting optimization problem must be stated as a series of minimization cases.

CHAPTER 4: LUNG CANCER GENETIC BIOMARKERS

4.1. First case study: Lung Cancer

Cancer starts in cells, when normal cells in tissues grow and divide to form new cells where they are needed, old cells die. During this process, some cells do not die as they should. The excessive growth of cells form a mass of tissue called a tumor.

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 cancer cases in 2012. Additionally, lung cancer was the first cause of death in the world, with 1.6 million deaths or 19.4% of all cancer related deaths in 2012 [46]. This analysis was conducted in 184 countries. Specifically, in the United States, lung cancer is expected to be the highest cause of death in 2014, for both men and women [1].

PubMed is one resource for the National Center for Biotechnology Information (NCBI). PubMed has the major database for electronically searching and recovering biomedical literature from MedLine and other life science journals [47]. Dr. Isaza, from Ponce School of Medicine, recommended the database GDS3257 to apply the proposed methods.

For this case, the database GDS3257 was used it was first reported by Landi MT and collaborators [48]. Initially, this study analyzed 180 tissues from cancerous and healthy samples, but after several revisions 107 samples were used for database. Although curation is often times mandatory in microarray repositories, the rationale for this one in particular was not discussed in the original paper. These samples showed the measure of relative expression for 22,283 genes from 107 samples: 49

healthy and 58 cancerous tissues. Additionally, this microarray experiment analyzed samples from never smokers, former smokers, and current smokers (See Figure 4-1). For this study, the patients were between 44 and 79 years old and had stage I thru IV lung cancer.

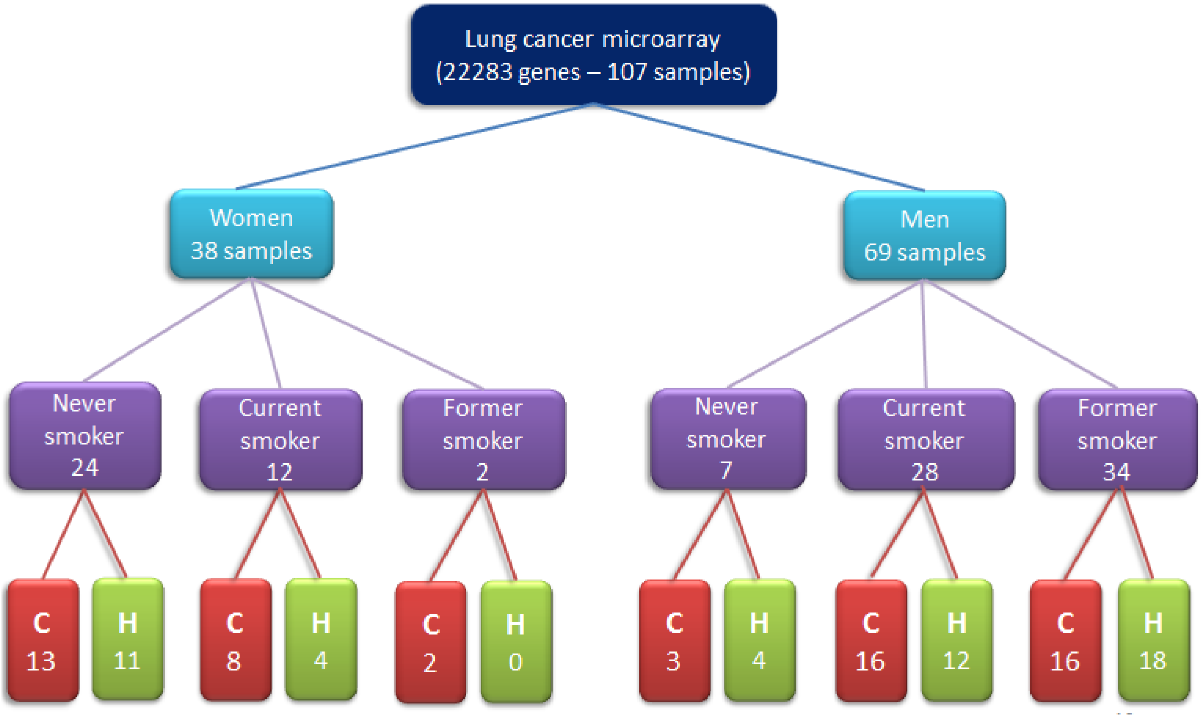


Figure 4-1: Represents the organization of the database used (GDS3257) for this research. "C" indicates cancer samples, and "H" indicates healthy samples.

4.1.1. First Analyses of a Healthy Never Smoker vs Cancer Never Smoker in Lung Cancer

For this first analysis, we used the fifteen healthy never smoker (HNS) tissues from men and women, and sixteen cancer never smoker (CNS) tissues from men and women. With these samples the absolute value of the differences of means and medians of healthy and cancer tissues for each gene were calculated. The analysis in MatLab tool was run in computer with 4 GB of memory RAM and 2.66 GHz CPU. Due to the memory restriction, Pareto efficient frontier was found in a tournament fashion. The two PMs calculated for 22,283 genes was divided in three groups: two groups of 7500 and one of 7283 genes. This data was run in MatLab tool to find the locally efficient frontier per group. Finally, the resulting genes from previous analysis were run again and found the global Pareto efficient frontier [49]. 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 akin to finding the tallest person in a room by picking the tallest in different subgroups and comparing the local winners in the end. With enough computing memory, partitioning the data is not necessary.

For each group, the locally non-dominated subset was identified (Figure 4-2), and the total number of selected genes for the three groups included five genes: WIF1, FCN3, SPP1, RAGE, and TMEM100.

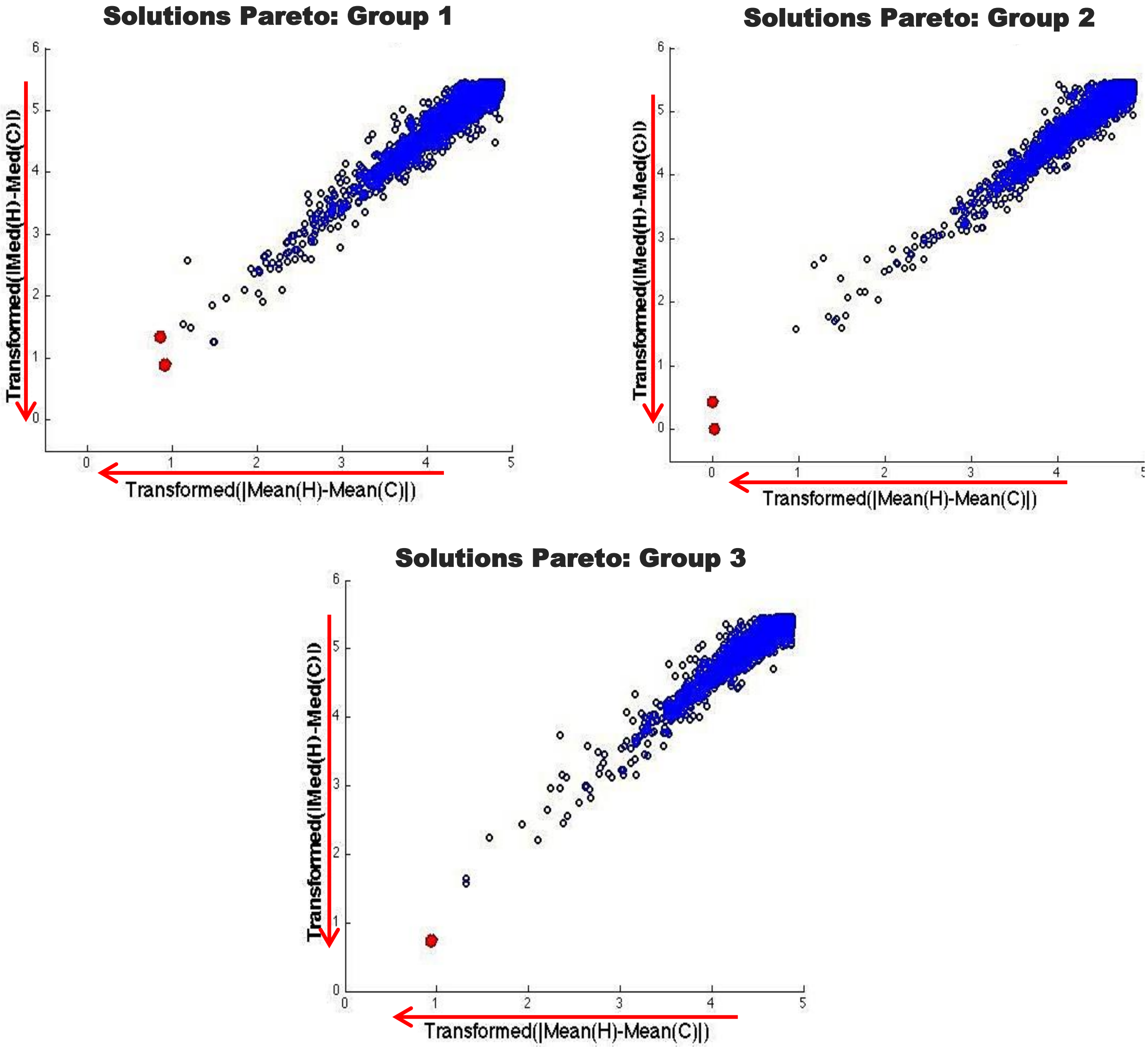


Figure 4-2: Local Pareto-Efficient frontiers of all groups. For the first and second groups, two genes are at the local Pareto-Efficient frontier, and one gene for the third group.

Then the locally non-dominated subsets (i.e., the five genes obtained before) were used to obtain the globally-optimal Pareto Efficient Frontier, as seen in Figure 4-3.

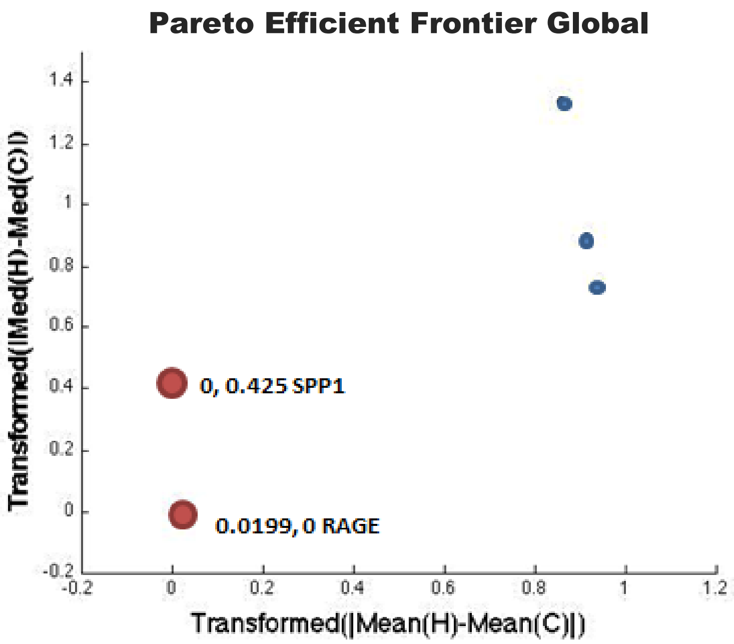


Figure 4-3: Represents the final globally-optimal Pareto-Efficient Frontier, which consists of Rage and SPP1 genes.

In figure 4-3, RAGE and SPP1 are the genes in the global Pareto-efficient frontier. It is important to mention that to achieve this result the user does not need to normalize or use a threshold value.

Receptor for Advanced Glycosylation End Products, RAGE, is a multiligand receptor involved with the regulation of multiple cell processes, such as homeostasis, development, and inflammation [50]. In the literature, RAGE is proposed as lung cancer biomarker, by R. Jing et al, *"Receptor for advanced glycation end products (RAGE) soluble form (sRAGE): a new biomarker for lung cancer"*.

Secreted PhosphoProtein 1, SPP1, is a protein-coding gene. Diseases associated with SPP1 include ossifying fibroma, and papillary cystadenocarcinoma [51].The changes in its gene expression implies alterations in cell properties involved in malignancy such as adhesion, migration, invasion, enhanced tumor survival, tumour angiogenesis, and metastasis [52]. This gene is proposed to as biomarker in V. Lazar et al, "*Integrated molecular portrait of non-small cell lung cancers*".

Similar to this analysis, eighteen additional analyses related to never smoker (Healthy and cancer tissues) were carried out. For the first 6 analyses, samples from men and women were used together. In the next six analyses, only samples of women were used. In the last six analyses, only men tissues were used. All of these groups were obtained from database GDS3257 (See Figure 4-1).

4.1.2. Analysis of Lung Cancer: Never smoker vs Current Smoker in Cancer and Healthy tissues

Figure 4-4 shows a summary of this case study. The circles on the left side represent the healthy never smoker (HNS) and healthy current smoker (HCS) tissues, while the circles on the right 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.

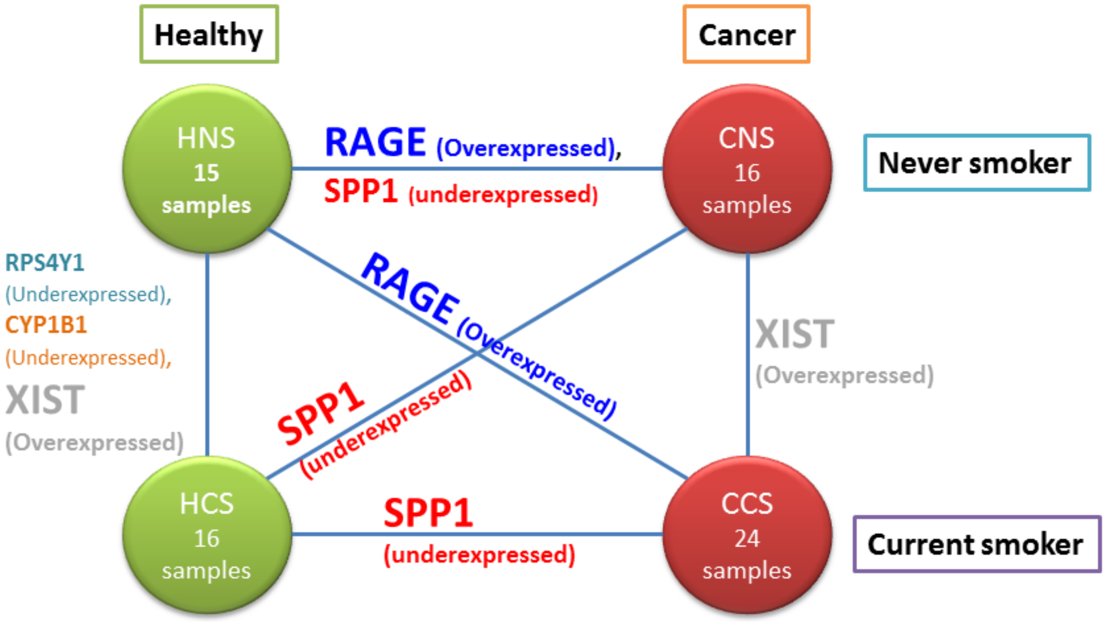


Figure 4-4: Diagram representing six analyses between four different conditions of microarray (HNS vs HCS vs CNS vs CCS). The edges of the graph represent genes of the Pareto Efficient Frontier. For each case is overexpressed or underexpressed in cancer when compared to healthy tissues.

The first analyses show that, between HNS and CNS, the MCO solution consisted of just two genes as the genes that changed their expression the most: RAGE and SPP1. The result of MCO for the second analysis between 16 samples of HCS and 24 samples of CCS is just the SPP1 gene. In the analysis concerning 15 samples between HNS and 24 samples of CCS, RAGE was the only solution, while in the analysis between 16 samples of HCS and 16 samples of CNS the gene with the largest change is SPP1. With the results from these four analyses, we can conclude that RAGE and SPP1 showed significant changes between a state of health and a state of cancer. Also SPP1 showed a large change between HCS and CNS. In addition, RAGE showed significant change between HNS and CCS.

The other two analyses were carried out between 15 samples of HNS versus 16 samples of HCS and 16 samples of CNS with 24 samples of CCS. The result of these analyses present three genes: RPS4Y1, CYP1B1 and XIST in healthy tissues, while in cancer, there is just XIST. In this case, we could conclude that XIST showed an important difference when comparing NS and CS (both in healthy and cancer tissues). See Table 4-1.

Table 4-1 shows the genes that form the Pareto Efficient Frontier and the calculated values in order to apply the MCO method. If a gene is overexpressed or under-expressed when comparing healthy samples to cancer samples, it is determined if the values of the differences in mean and median are positive or negative respectively. For example if the signs of both the differences in mean and median are positive, the gene can be considered as overexpressed in cancer, in other way it could be under-expressed in cancer.

Table 4-1: Represent the summarizing of genes from Pareto-Efficient Frontier in the analysis for never and current smoker and their expressions.

Analysis 1

ID_Ref	Identifier	Mean (HNS)	Mean (CNS)	Median (HNS)	Median (CNS)	Mean(HNS)-Mean(CNS)	Med(HNS)-Med(CNS)	Expression in cancer
210081_at	AGER	12.3777	7.5234	12.7605	7.3027	4.8543	5.4578	Overexpressed
209875_s_at	SPP1	7.1348	12.0090	7.0223	12.0551	-4.8742	-5.0328	Underexpressed

Analysis 2

ID_Ref	Identifier	Mean (HCS)	Mean (CCS)	Median (HCS)	Median (CCS)	Mean(HCS)-Mean(CCS)	Med(HCS)-Med(CCS)	Expressed in cancer
209875_s_at	SPP1	7.4312	11.8356	7.3415	11.9560	-4.4045	-4.6144	Underexpressed

Analysis 3

ID_Ref	Identifier	Mean (HNS)	Mean (CCS)	Median (HNS)	Median (CCS)	Mean(HNS)-Mean(CCS)	Med(HNS)-Med(CCS)	Expressed in cancer
210081_at	AGER	12.3777	7.3407	12.7605	7.2301	5.0369	5.5304	Overexpressed

Analysis 4

ID_Ref	Identifier	Mean (HCS)	Mean (CNS)	Median (HCS)	Median (CNS)	Mean(HCS)-Mean(CNS)	Med(HCS)-Med(CNS)	Expressed in cancer
209875_s_at	SPP1	7.4312	12.0090	7.3415	12.0551	-4.5778	-4.7136	Underexpressed
				Analysis 5				
ID_Ref	**Identifier**	**Mean (HNS)**	**Mean (HCS)**	**Median (HNS)**	**Median (HCS)**	**Mean(HNS)-Mean(HCS)**	**Median(HNS-HCS)**	**Expressed in cancer**
201909_at	RPS4Y1	8.5897	10.6481	7.4258	11.5666	-2.0585	-4.1407	Underexpressed
202437_s_at	CYP1B1	6.8600	9.0555	6.8096	9.1060	-2.1955	-2.2964	Underexpressed
221728_x_at	XIST	9.2282	7.0814	10.0964	6.1053	2.1468	3.9911	Overexpressed
				Analysis 6				
ID_Ref	**Identifier**	**Mean (CNS)**	**Mean (CCS)**	**Median (CNS)**	**Median (CCS)**	**Mean(CNS)-Mean(CCS)**	**Med(CNS)-Med(CCS)**	**Expressed in cancer**
221728_x_at	XIST	9.3664	7.3730	10.1310	6.4508	1.9935	3.6801	Overexpressed

4.1.3. Analysis of Lung Cancer in Women: Never smoker vs Current Smoker in Cancer and Healthy tissues

Figure 4-5 shows the result with the same analysis described before, but using just women's tissues. For this representation the first conclusion is RAGE, which showed a large change when compared to a cancer group, regardless if the comparison was against HNS or HCS.

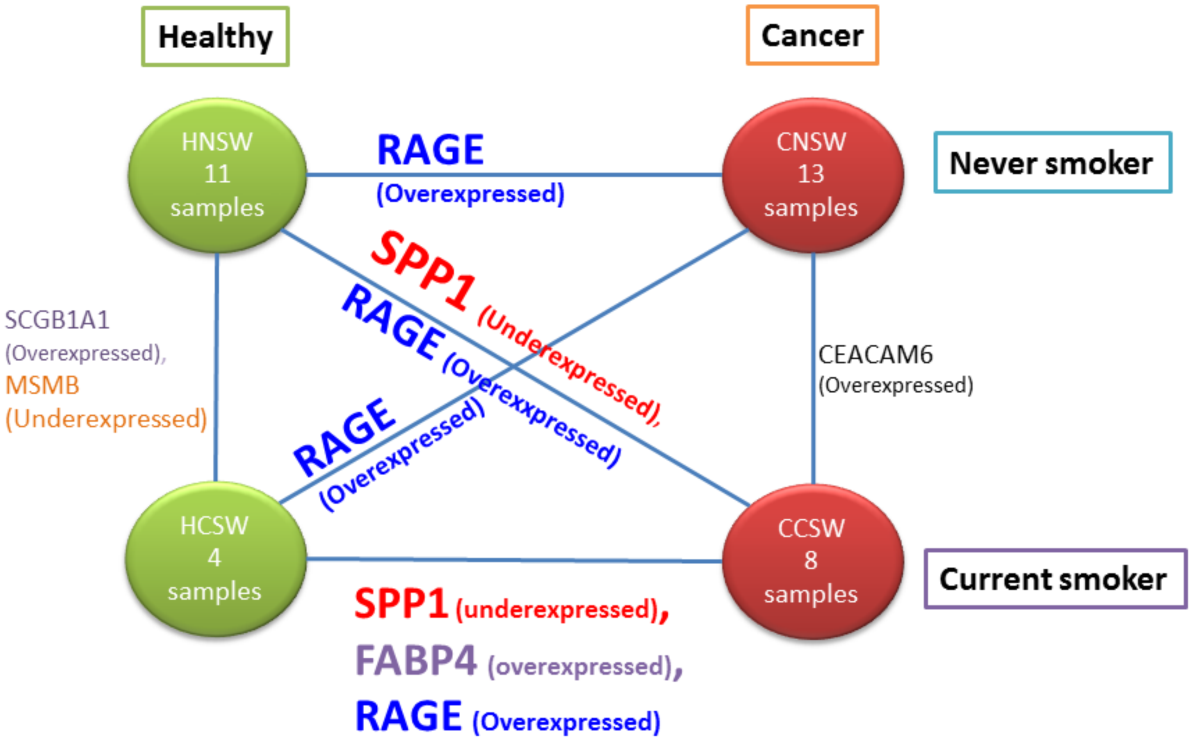


Figure 4-5: Diagram representing six analyses between four different conditions for women samples of microarray (HNS vs HCS vs CNS vs CCS). The edges of the graph represent genes of Pareto Efficient Frontier. Each case is overexpressed or underexpressed in cancer when compared to healthy tissues.

Table 4-2 shows the genes that form Pareto Efficient Frontier and the calculated values in order to apply the MCO method. If a gene is overexpressed in cancer or under-expressed in cancer it is determined if the values of the differences in mean and median are positive or negative respectively.

Table 4-2: Represent the summarizing of genes from Pareto-Efficient Frontier in the analysis for never and current smoker for women and their expressions.

Analysis 1

ID_Ref	Identifier	Mean (HNSW)	Mean (CNSW)	Median (HNSW)	Median (CNSW)	Mean(HNSW)-Mean(CNSW)	Med(HNSW)-Med(CNSW)	Expressed in cancer
210081_at	RAGE	12.236055	7.3239108	12.6324	7.12506	4.912143776	5.50734	Overexpressed

Analysis 2

ID_Ref	Identifier	Mean (HCSW)	Mean (CCSW)	Median (HCSW)	Median (CCSW)	Mean(HCSW)-Mean(CCSW)	Med(HCSW)-Med(CCSW)	Expressed in cancer
203980_at	FABP4	11.3681	7.4843175	11.3418	6.838025	3.8837825	4.503775	Overexpressed
209875_s_at	SPP1	7.822125	12.015075	7.76183	11.75325	-4.19295	-3.99142	Underexpressed
210081_at	RAGE	11.723925	7.5909588	11.856	7.630145	4.13296625	4.225855	Overexpressed

Analysis 3

ID_Ref	Identifier	Mean (HNSW)	Mean (CCSW)	Median (HNSW)	Median (CCSW)	Mean(HNSW)-Mean(CCSW)	Med(HNSW)-Med(CCSW)	Expressed in cancer
209875_s_at	SPP1	7.2148118	12.015075	7.02226	11.75325	-4.800263182	-4.73099	Underexpressed
210081_at	RAGE	12.236055	7.5909588	12.6324	7.630145	4.645095795	5.002255	Overexpressed

Analysis 4

ID_Ref	Identifier	Mean (HCSW)	Mean (CNSW)	Median (HCSW)	Median (CNSW)	Mean(HCSW)-Mean(CNSW)	Med(HCSW)-Med(CNSW)	Expressed in cancer
210081_at	RAGE	11.723925	7.3239108	11.856	7.12506	4.400014231	4.73094	Overexpressed

Analysis 5

ID_Ref	Identifier	Mean (HNSW)	Mean (HCSW)	Median (HNSW)	Median (HCSW)	Mean(HNSW)-Mean(HCSW)	Med(HNSW)-Med(HCSW)	Expressed in cancer
205725_at	SCGB1A1	12.247863	9.9364	13.2639	9.993235	2.311462727	3.270665	Overexpressed
207430_s_at	MSMB	4.5146282	7.2527475	4.24797	6.907905	-2.738119318	-2.659935	Underexpressed

Analysis 6

ID_Ref	Identifier	Mean (CNSW)	Mean (CCSW)	Median (CNSW)	Median (CCSW)	Mean(CNSW)-Mean(CCSW)	Med(CNSW)-Med(CCSW)	Expressed in cancer
203757_s_at	CEACAM6	13.174654	10.82439	13.3954	10.277045	2.350263846	3.118355	Overexpressed

4.1.4. Analysis of Lung Cancer in Men: Never smoker vs Current Smoker in Cancer and Healthy tissues

Figure 4-6 shows the results with an analysis similar to the one described before, but using just men samples. For this representation, as in previous cases SPP1 and RAGE showed large changes when compared to a cancer group, regardless if the comparison was against HNS or HCS. Also in this case SPP1 is large differentially expressed more often than RAGE.

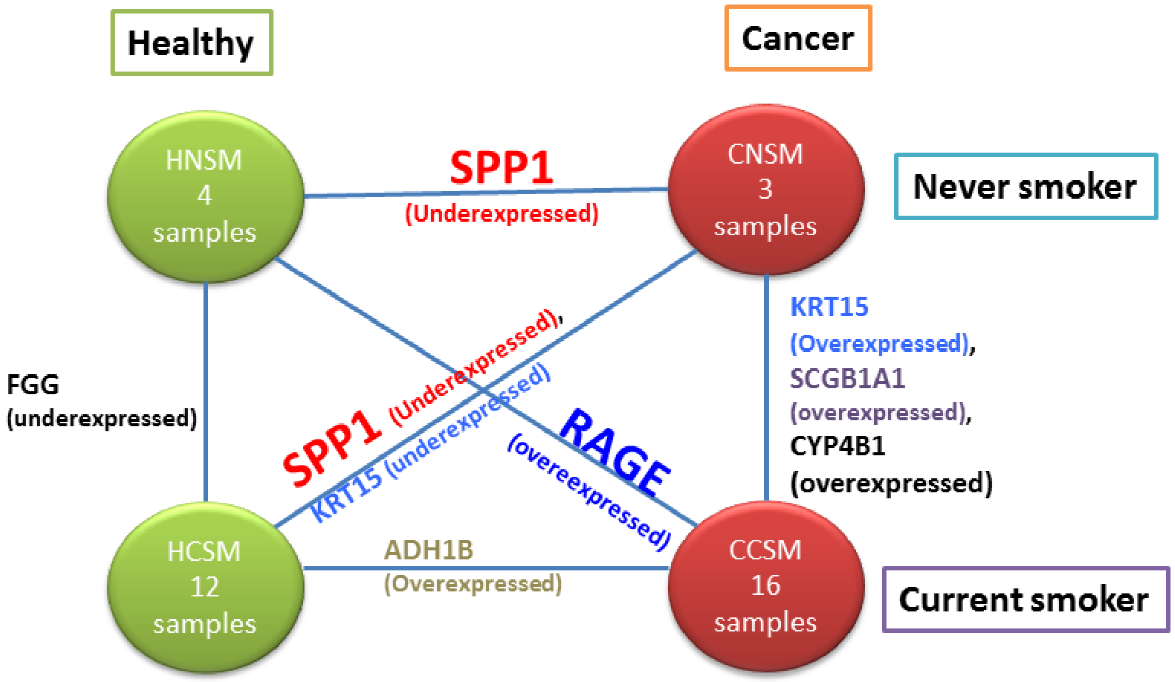


Figure 4-6: Diagram representing six analyses between four different conditions for men samples of microarray (HNS vs HCS vs CNS vs CCS). The edges of the graph represent genes of Pareto Efficient Frontier. For each case, it is overexpressed or underexpressed in cancer when compared to healthy tissue

Table 4-3 shows the genes that form Pareto Efficient Frontier and the calculated values in order to apply the MCO method.

Table 4-3: Represent the summarizing of genes from Pareto-Efficient Frontier in the analysis for never and current smoker for men and their expressions.

Analysis 1

ID_Ref	Identifier	Mean (HNSM)	Mean (CNSM)	Median (HNSM)	Median (CNSM)	Mean(HNSM)-Mean(CNSM)	Med(HNSM)-Med(CNSM)	Expression in cancer
209875_s_at	SPP1	6.914805	11.9426	6.966145	12.0506	-5.027795	-5.084455	Underexpressed

Analysis 2

ID_Ref	Identifier	Mean (HCSM)	Mean (CCSM)	Median (HCSM)	Median (CCSM)	Mean(HCSM)-Mean(CCSM)	Med(HCSM)-Med(CCSM)	Expression in cancer
209613_s_at	ADH1B	10.504545	6.040956	10.60635	5.60672	4.46358875	4.99963	Overexpressed

Analysis 3

ID_Ref	Identifier	Mean (HNSM)	Mean (CCSM)	Median (HNSM)	Median (CCSM)	Mean(HNSM)-Mean(CCSM)	Med(HNSM)-Med(CCSM)	Expression in cancer
210081_at	RAGE	12.7672	7.21564	12.7797	6.83025	5.55156	5.94945	Overexpressed

Analysis 4

ID_Ref	Identifier	Mean (HCSM)	Mean (CNSM)	Median (HCSM)	Median (CNSM)	Mean(HCSM)-Mean(CNSM)	Med(HCSM)-Med(CNSM)	Expressed in cancer
204734_at	KRT15	5.5394642	9.25312	5.518315	10.2703	-3.713655833	-4.751985	Underexpressed
209875_s_at	SPP1	7.30083	11.9426	7.30887	12.0506	-4.64177	-4.74173	Underexpressed

Analysis 5

ID_Ref	Identifier	Mean (HNSM)	Mean (HCSM)	Median (HNSM)	Median (HCSM)	Mean(HNSM)-Mean(HCSM)	Med(HNSM)-Med(HCSM)	Expressed in cancer
219612_s_at	FGG	6.363085	9.956473	6.44629	9.908025	-3.593388333	-3.461735	Underexpressed
Analysis 6								
ID_Ref	**Identifier**	**Mean (CNSM)**	**Mean (CCSM)**	**Median (CNSM)**	**Median (CCSM)**	**Mean(CNSM)-Mean(CCSM)**	**Med(CNSM)-Med(CCSM)**	**Expressed in cancer**
204734_at	KRT15	9.25312	6.909603	10.2703	6.531385	2.343516875	3.738915	Overexpressed
205725_at	SCGB1A1	11.223457	8.356614	11.3493	7.902555	2.866842917	3.446745	Overexpressed
210096_at	CYP4B1	9.4844233	6.566811	8.94934	6.112965	2.917612083	2.836375	Overexpressed

Table 4-4: Summary from Pareto efficient frontier genes and their related cancer

Gene name	Examples of cancer types related to genes	References
RAGE	**Pancreas, colon and prostate, colorectal, gastric, liver, lung**	[48, 49] [55]–[58]
SPP1	**Oral, lung, bone, bladder, prostate, cervical, breast, head and neck, liver**	[54]–[57], [58], [59]–[61]
XIST	Meninges, breast, ovarian	[67]–[69]
RPS4Y1	Meninges	[67]
CYP1B1	Lung, cervical, head and neck, prostate	[70]–[73]
WOMEN		
FABP4	Prostate and breast, ovarian	[74], [75]
CEACAM6	Head and neck, breast, colon, lung	[76]–[79]
MSMB	Prostate	[80]
SCGB1A1*	Smokers	[81]
MEN		
ADH1B	Esophageal, colorectal, head and neck	[82]–[84]
CYP4B1	Bladder	[85]
KRT15	Lung, ovarian	[86], [87]
FGG	Liver	[88]
SCGB1A1*	Smokers	[81]

Table 4-4 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 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 the gene with an asterisk (SCGB1A1), which is not related to any type of cancer. From this table, three important conclusions are obtained. First, those genes found in the literature as biomarkers, validate our method. Secondly, those genes not found in the literature as biomarkers but that are associated with other types of cancer, such as, XIST, RPS4Y1, CYPIB1, FABP4, among others, could eventually be validated and proposed as lung cancer biomarkesr with the precursor that they are important genes for other type of cancer. 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.

The MatLab tool developed in our group has a high discrimination rate. For the first case, where the genes are declared biomarkers, it is possible to remove these genes from the global data and run the tool again. This process can be repeated according to the purpose of the analysis. Such is the case of Juan Rosas, member of our research group, who searches for signaling pathways. He used MCO to find genes in different efficient frontiers and later explores the relations between them.

4.2. Pseudo Meta-Analysis with four performance measures: a prototype for meta-analysis.

In previous analyses (Section 5.1) two PMs (absolute value of differences of means and absolute value of differences of medians) were used. In this section, MCO meta-analysis is carried out using four PMs, which were the “absolute value of differences of medians” for each group [16]. The medians were used for their nonparametric characteristics, as it has been habitual in analyzes previously carried out by our group. Continuing with the case, the difference in medians between the groups of cancer and healthy tissues is calculated for each gene of the 22,283 genes in the database. These groups are: HNS (15 samples) vs CNS (16 samples), HNS (15 samples) vs CCS (24 samples), HCS (16 samples) vs CNS (16 samples), HCS (16 samples) vs CCS (24 samples) see Figure 4-7.

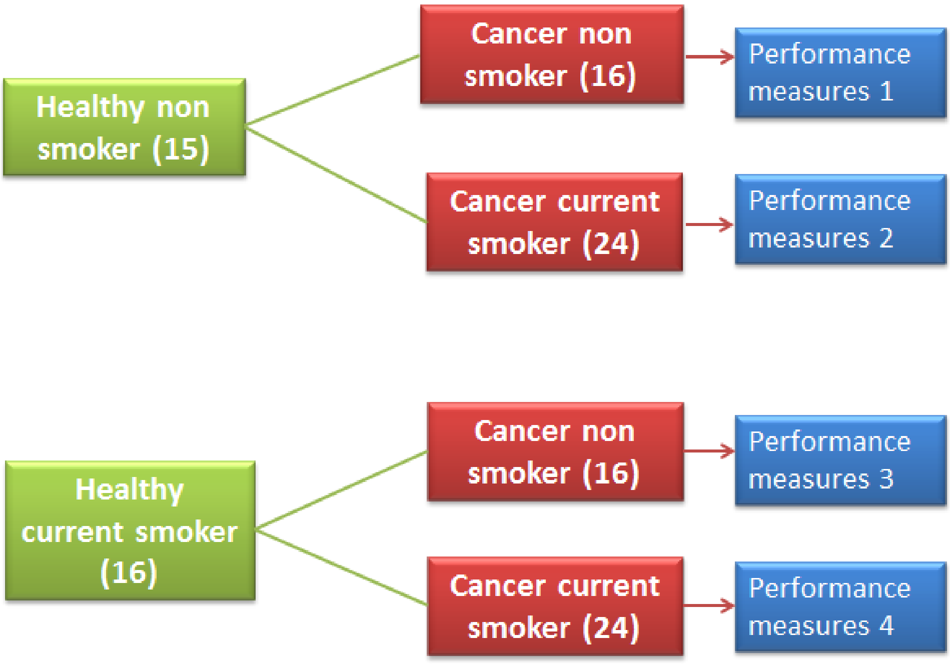


Figure 4-7: Groups for meta-analysis with four PM's

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. The data was analyzed in six groups of 4,000 genes and the tool worked on computer with 4GB of memory and 1.70 GHz CPU. For the, four groups of 6,000 genes, the tool worked on computer with 4GB of memory and 1.77 GHz CPU. Finally, the tool with four groups of 7,000 genes worked on computer with 4GH of memory RAM and 2.66 GHz CPU. That process did to make sure that the grouping scheme did not introduce any bias in the results. Owing to a low memory of computer we divided the data in groups in ascending order [49], but to prove to the consistence and trustworthiness of the program we ran the data in 3 different computers. In all these analyses the global Pareto efficient frontier included two genes as a result. 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.

CHAPTER 5: LEUKEMIA GENETIC BIOMARKERS

5.1. Second case study: Leukemia

The NCHS estimates 52,380 new cases of leukemia for 2014 in the U.S., causing 24,090 deaths [1]. Leukemia is a cancer of the blood and invades the entire body by bloodstream. There are different types of leukemia. Acute lymphoblastic leukemia (ALL) is the most common type of leukemia in young children. Acute myelogenous leukemia (AML) is the most common type of acute leukemia in adults. Chronic lymphocytic leukemia (CLL) is the most common type of chronic leukemia in adults, and chronic myelogenous leukemia (CML), which affects mainly adults [89]. This cancer is described as untreated (sick without treatment), in remission (survival), or recurrent. Leukemia has been divided by stages, going from Rai 0 to Rai IV, where stage 0 is considered low risk, stage I and II are considered intermediate risk, and stage III and IV are high risk [90].

The database used for this case is GSE2403,which was first used by S. Fält et al. [91]. This database has 12,625 genes and 21 samples: 11 in stable state and 10 progressive. This database contains both stable (healthy) and progressive (cancer) samples presenting leukemia; however, according to the authors, stable (in remission) samples do not need treatment, while progressive (new Leukemia cases or recurrent) samples need treatment. For this reason, the stable samples are here labeled “Healthy (H)” and the progressive ones as “Cancer (C)” as to not introduce further descriptors. Also, they take thirteen samples of men and eight samples of women (See Figure 5-1). The samples were taken from patients between 49 to 82 years old, and they suffer CLL.

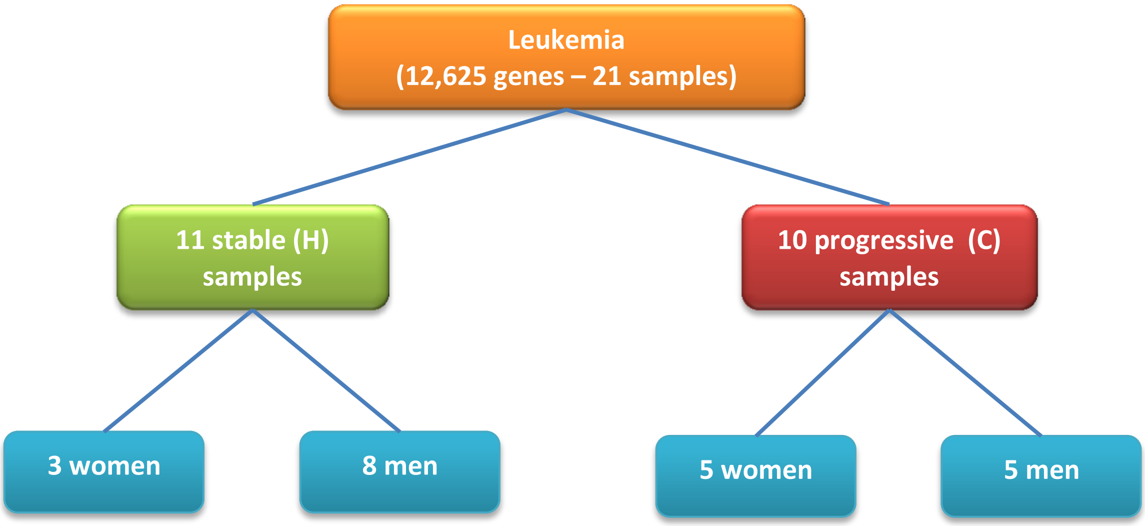


Figure 5-1: Represents the organization of the database used (GSE2403) for this case.

5.2. Analysis of Leukemia: Healthy vs cancer samples

In this case the 11 H and 10 C samples were taken from the database. With these samples the absolute value of the differences in means and medians for H and C samples were calculated for each gene. The 12,625 genes were divided in two groups: one with 7000 and another 5,625 with genes [49], as in the previous cases. For each group, the locally non dominated subset was discovered (Figure 5-2), and the total number of selected genes for the two groups were three genes: 31687_f_at, 38833_at and 41165_g_at.

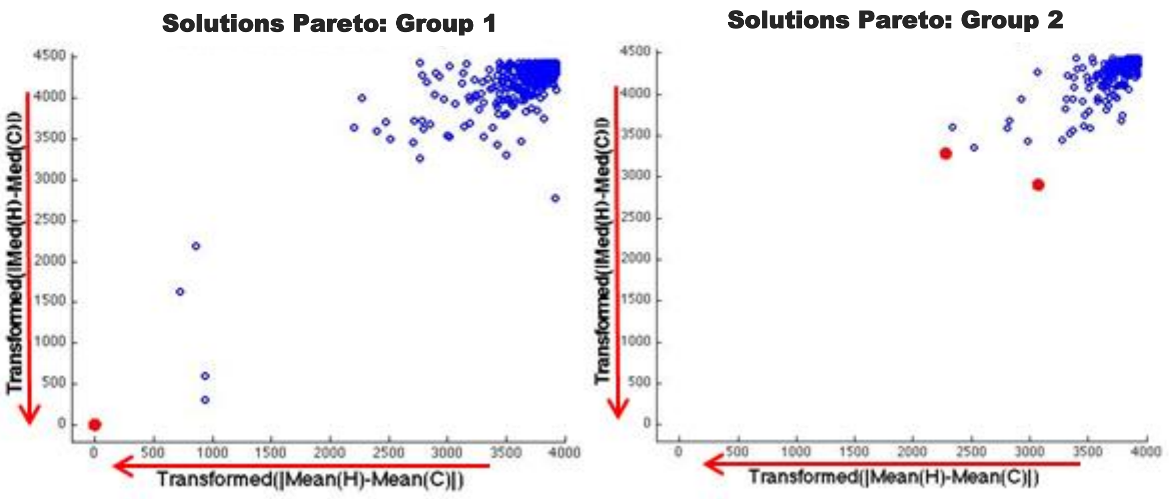


Figure 5-2: Local Pareto-Efficient frontiers of two groups. For the first group, one gene is at the local Pareto-Efficient frontier, and two genes for the second group.

Then, the locally non-dominated subsets (i.e., the three genes obtained before), were used to obtain the globally-optimal Pareto Efficient Frontier, as seen in Figure 5-3.

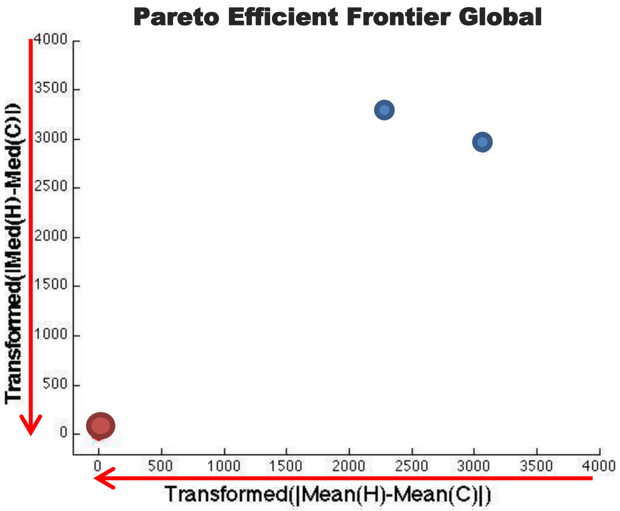


Figure 5-3: Represents the final globally-optimal Pareto-Efficient Frontier, which consists of 31687_f_at gene.

In this analysis we could observe that only one gene out of 12,625 showed significant changes in its relative expression. The gene 31687_f_at or HBB, hemoglobin, beta [92] is a protein also known as beta-globin. In the literature this gene is proposed to as biomarker in PK Chong et al, "*Hemoglobin* *subunit beta (HBB) is a potential biomarker for predicting response to Gefitinib in NSCLC patients*" In this case it is important to analyze the relation between non-small cell lung cancer (NSCLC) and Leukemia. (See Table 5-1)

Table 5-1: Represent the summarizing of gene from Pareto-Efficient Frontier in the Leukemia analysis and their expressions.

ID_REF	Mean(H)	Mean(C)	Median(H)	Median(C)	Mean(H)-Mean(C)	Med(H)-Med(C)	Expression in cancer
31687_f_at	7076.109091	11006.57	5494.3	9932.45	-3930.46091	-4438.15	Underexpressed

CHAPTER 6: BREAST CANCER microRNA BIOMARKERS

6.1. Third case study: microRNAs in breast Cancer

In 2012, breast cancer was one of the most common causes of death by cancer in the world. According to the International Agency for Research on Cancer, 1.7 million cases or 11.9% of total cases [46] around the world. Specifically, in 2014 in United States breast cancer is expected to be associated to 235,030 new cases and 40,430 deaths [1].

This chapter describes the application of the proposed method to microRNA experiments, a newer member of the high-throughput biological experiments (omics). microRNAs (miRNAs) are considered as master regulators of gene expression. miRNAs could be able to regulate up to 30% of protein-coding genes in the human genome. Furthermore, miRNAs have been associated with the development of several diseases [93]. The existence of circulating miRNAs in the blood of cancer patients has raised the possibility that miRNAs may serve as a novel diagnostic marker [94].

For these type of miRNA microarray, the traditionally normalization methods are not applicable. Due to this normalization, methods are based on two assumptions. First, the total number of genes to evaluate is large (>10,000), and the expression levels of the majority of genes is preserved constantly. On the contrary, miRNA microarrays are generally spotted in low number. Due to that, the total number of miRNA is less than 2000 [22]. Also, the expression of miRNAs is different to genes. In this chapter, it is demonstrated how the MCO method was able to handle this novel type of –omics experiments.

The database used in this chapter is divided by 27 case samples (breast cancer) and 29 control samples (healthy) and was provided by Dr. Isaza, from Ponce School of Medicine (unpublished data).

6.2. First data analysis

In this analysis, 27 C samples and 29 H samples were used. A total of 384 miRNAs are included in the database. After revision, 125 miRNAs were removed: 94 of them did not have enough readings and 31 had less than three replicates in healthy or cancer samples. Finally, 259 miRNAs were analyzed using two PMs, absolute value of differences of means and absolute value of differences of medians. The results showed five miRNAs (See Figure 6-1): hsa-miR-9-000583 (FAM, NFQ), hsa-miR-219-000522 (FAM, NFQ), hsa-miR-365-001020 (FAM, NFQ), hsa-miR-625-002431 (FAM, NFQ) and hsa-miR-652-002352 (FAM, NFQ) (See Table 6-1).

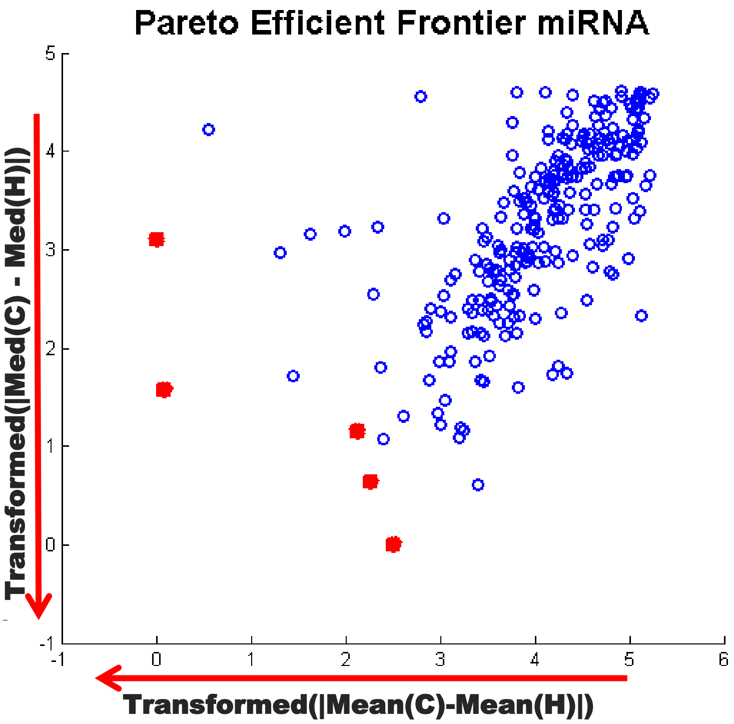


Figure 6-1: Represents the final globally-optimal Pareto-Efficient Frontier, which consists of five miRNAs

In this first analysis, five miRNAs showed a significant changed in their relative expression from a total of 259 miRNAs. This analysis evidenced the transferability of the method proposed to other types of data. In other words, the method could be used in any type of biological database in which is possible to obtain the PMs for healthy vs cancer samples. This type of biological data is part of the 'omics' groups. At this time, in the literature there is not enough information about these five microRNAs. These are undergoing more analyses related to breast cancer at Ponce School Medicine and Health Sciences by Dr. Isaza. Table 6-1 shows the genes that form the Pareto Efficient Frontier.

Table 6-1: Represent the summarizing of miRNA from Pareto-Efficient Frontier in the breast cancer analysis and their expressions.

Detector	Mean (H)	Mean (C)	Median (H)	Median (C)	Mean(H-Mean(C)	Median(H)-Median(C)	Expression in cancer
hsa-miR-9-000583 (FAM,NFQ)	30.3902	33.5226	30.9568	34.4201	-3.1324	-3.4633	Overexpressed
hsa-miR-219-000522 (FAM,NFQ)	35.2830	30.0280	34.7969	33.2758	5.2550	1.5211	underexpressed
hsa-miR-365-001020 (FAM,NFQ)	31.5740	34.3179	30.1818	34.7958	-2.7439	-4.6140	Overexpressed
hsa-miR-625-002431 (FAM,NFQ)	36.1277	30.9577	36.6335	33.5928	5.1699	3.0406	underexpressed
hsa-miR-652-002352 (FAM,NFQ)	27.9800	30.9634	27.9513	31.9310	-2.9834	-3.9797	Overexpressed

6.3. Second data analysis

In a second analysis, the data was divided by DNA Repair Capacity (DCR) in four groups: low DCR-H, high DCR-H, low DCR-C and high DCR-C. A total of 384 of miRNAs were taken into account and 136 miRNAs were removed due to lack of enough replicates (less than three replicates). Finally, 256 miRNAs were analyzed using the following four cases.

6.3.1. Low DRC healthy vs. Low DRC cancer

Ten low DRC-H samples and 17 low DRC-C samples were used from the database. From the initial 256 miRNAs, 22 miRNAs were removed from the database due to lack of readings or replicates in any group. Finally, with the 226 miRNAs, the MCO results showed four miRNAs: hsa-let-7a-000377 (FAM, NFQ) is underexpressed in low DCR-H. RNU48-001006 (FAM, NFQ) is underexpressed in low DCR-H. Hsa-miR-365-001020 (FAM, NFQ) is underexpressed in low DCR-H. Hsa-miR-627-001560 (FAM, NFQ) is overexpressed in Low DCR-H (See Figure 6-2).

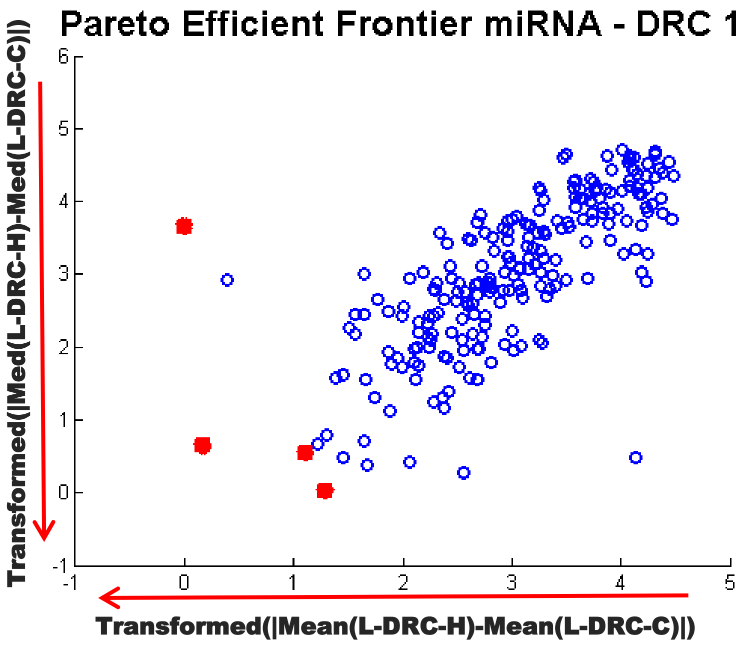


Figure 6-2: Represents the final globally-optimal Pareto-Efficient Frontier, which consists of four miRNAs

Table 6-2 shows the summarizing of miRNAs that form the Pareto Efficient Frontier.

Table 6-2: Represent the summarizing of miRNAs from Pareto-Efficient Frontier in the breast cancer analysis and their expressions.

Detector	Mean(L-DRC-H)	Mean (L-DRC-C)	Median (L-DRC-H)	Median (L-DRC-C)	Mean(L-DRC-H)-Mean(L-DRC-C)	Med(L-DRC-H)-Med(L-DRC-C)	Expression in cancer
hsa-let-7a-000377 (FAM,NFQ)	26.7975	30.1677	27.0273	31.2117	-3.3702	-4.1844	overexpressed
RNU48-001006 (FAM,NFQ)	27.1073	31.4208	28.0333	32.1213	-4.3134	-4.0879	overexpressed
hsa-miR-365-001020 (FAM,NFQ)	31.4041	34.5915	30.0910	34.7958	-3.1874	-4.7048	overexpressed
hsa-miR-627-001560 (FAM,NFQ)	26.1536	21.6778	23.9459	22.8787	4.4758	1.0672	underexpressed

6.3.2. High DRC healthy vs. High DRC cancer

Fourteen high DRC-H samples and 4 high DRC-C samples were used. From the initial 256 miRNAs, 77 miRNAs were removed as in the previous cases. Finally, a total of 171 miRNAs were analyzed. The MCO results showed only one miRNA, hsa-miR-133b-002247 (FAM, NFQ). See Figure 6-3.

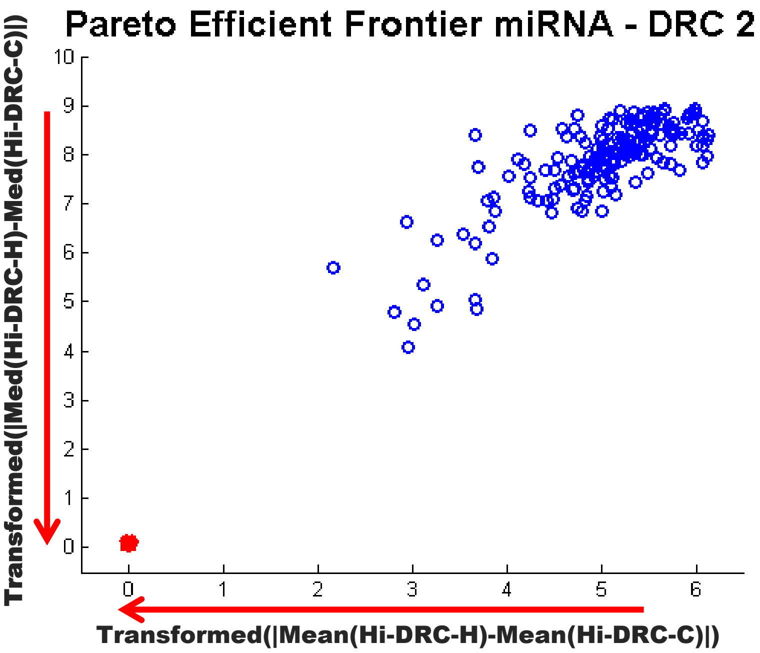


Figure 6-3: Represents the final globally-optimal Pareto-Efficient Frontier, which consists of one miRNA.

Table 6-3 shows the miRNA that form the Pareto Efficient Frontier for this case.

Table 6-3: Represent the summarizing of miRNA from Pareto-Efficient Frontier in the breast cancer analysis and expressions.

Detector	Mean(Hi-DRC-H)	Mean (Hi-DRC-C)	Median (Hi-DRC-H)	Median (Hi-DRC-C)	Mean(Hi-DRC-H)-Mean(Hi-DRC-C)	Med(Hi-DRC-H)-Med(Hi-DRC-C)	Expression in cancer
hsa-miR-133b-002247 (FAM,NFQ)	30.0527	36.1958	28.0686	36.9911	-6.1431	-8.9225	overexpressed

6.3.3. Low DRC healthy vs. High DRC healthy

Ten low DRC-H samples and 14 high DRC-H samples were used. From the initial 256 data, 19 miRNAs were removed from the database due to lack of readings or replicates in any group. A total of 229 miRNAs were analyzed. The MCO results showed four miRNAs: hsa-miR-134-001186 (FAM, NFQ), RNU48-001006 (FAM, NFQ), hsa-miR-379-001138 (FAM, NFQ), and hsa-miR-889-002202 (FAM, NFQ). See Figure 6-4.

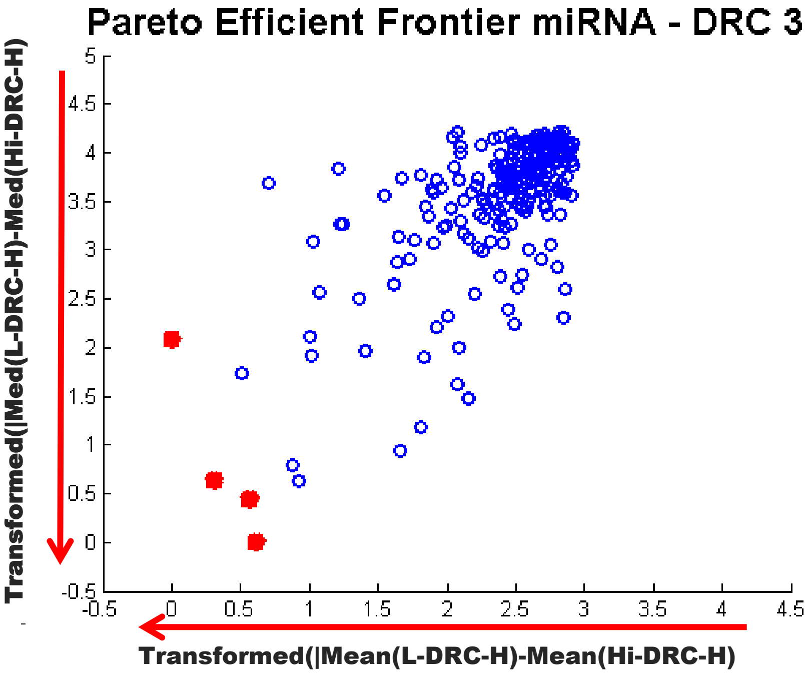


Figure 6-4: Represents the final globally-optimal Pareto-Efficient Frontier, which consists of four miRNAs

Table 6-4 shows the miRNA that form Pareto Efficient Frontier and the calculated values in order to apply the MCO method.

Table 6-4: Represent the summarizing of miRNAs from Pareto-Efficient Frontier in the breast cancer analysis and their expressions.

Detector	Mean(L-DRC-H)	Mean (Hi-DRC-H)	Median (L-DRC-H)	Median (Hi-DRC-H)	Mean(L-DRC-H)-Mean(Hi-DRC-H)	Med(L-DRC-H)-Med(Hi-DRC-H)	Expression in cancer
hsa-miR-134-001186 (FAM,NFQ)	30.5928	33.2027	31.2458	34.8367	-2.6099	-3.5909	overexpressed
RNU48-001006 (FAM,NFQ)	27.1073	30.0215	28.0333	30.1769	-2.9142	-2.1436	overexpressed
hsa-miR-379-001138 (FAM,NFQ)	29.3594	31.6635	29.5276	33.7454	-2.3041	-4.2179	overexpressed
hsa-miR-889-002202 (FAM,NFQ)	30.8396	33.1913	30.2792	34.0602	-2.3517	-3.7810	overexpressed

6.3.4. Low DRC cancer vs. High DRC cancer

Seventeen low DRC-C samples and 14 high DRC-C samples were used. From the initial data, 77 miRNAs were removed due to the number of lectures or replicates in any group. Finally, a total of 171 miRNAs were analyzed. The MCO results showed three miRNAs: hsa-miR-486-3p-002093 (FAM, NFQ), hsa-miR-502-3p-002083 (FAM, NFQ), hsa-miR-889-002202 (FAM, NFQ). See Figure 6-5.

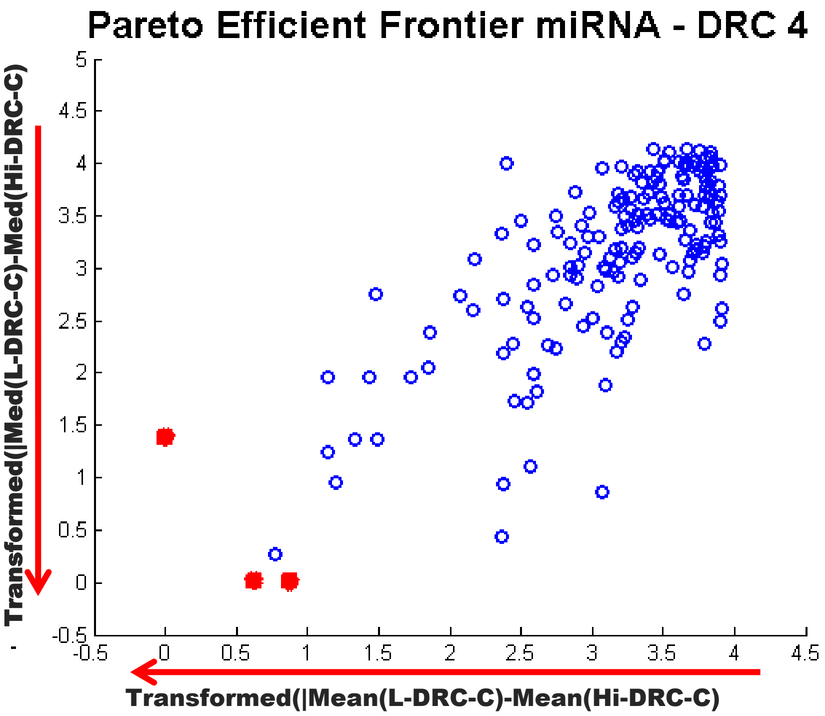


Figure 6-5: Represents the final globally-optimal Pareto-Efficient Frontier, which consists of three miRNAs.

Table 6-5 shows the miRNA that form Pareto Efficient Frontier and the calculated values in order to apply the MCO method.

Table 6-5: Represent the summarizing of miRNA from Pareto-Efficient Frontier in the breast cancer analysis and expressions.

Detector	Mean (L-DRC-C)	Mean (Hi-DRC-C)	Median (L-DRC-C)	Median (Hi-DRC-C)	Mean(L-DRC-C)-Mean(Hi-DRC-C)	Med(L-DRC-C)-Med(Hi-DRC-C)	Expression in cancer
hsa-miR-486-3p-002093 (FAM,NFQ)	32.7547	35.7990	32.9069	37.0522	-3.0443	-4.1453	overexpressed
hsa-miR-502-3p-002083 (FAM,NFQ)	33.0337	36.3284	32.9711	37.1065	-3.2947	-4.1354	overexpressed
hsa-miR-889-002202 (FAM,NFQ)	33.2527	37.1692	34.5913	37.3564	-3.9165	-2.7650	overexpressed

Figure 6-6 shows the results summary of the four analyses about DRC. In this graphic, the edges contain miRNAs that showed significant changes in their relative expression. For this representation, hsa-miR-889-002202 (FAM, NFQ) showed significant changes when comparing high DRC to low DRC groups (both in healthy and cancer tissues). Due to the fact that this research is ongoing at Ponce School of Medicine further analysis and conclusions are left for the future. This case study, however support the capability and transferability of the methods in this thesis.

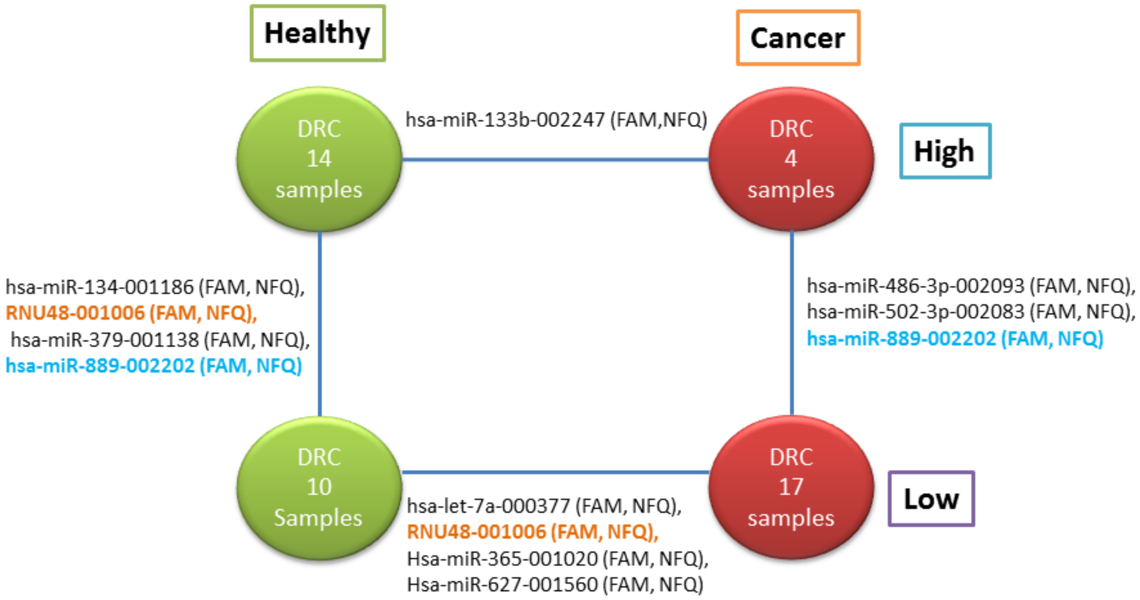

Healthy
Cancer
DRC
14
samples
hsa-miR-133b-002247 (FAM,NFQ)
DRC
4
samples
High
hsa-miR-134-001186 (FAM, NFQ),
RNU48-001006 (FAM, NFQ),
hsa-miR-379-001138 (FAM, NFQ),
hsa-miR-889-002202 (FAM, NFQ)
hsa-miR-486-3p-002093 (FAM, NFQ),
hsa-miR-502-3p-002083 (FAM, NFQ),
hsa-miR-889-002202 (FAM, NFQ)
DRC
10
Samples
DRC
17
samples
Low
hsa-let-7a-000377 (FAM, NFQ),
RNU48-001006 (FAM, NFQ),
Hsa-miR-365-001020 (FAM, NFQ),
Hsa-miR-627-001560 (FAM, NFQ)

Figure 6-6: Diagram representing four analyses between four different conditions for DRC samples of miRNA. The edges of the graph represent genes of Pareto Efficient Frontier.

Currently Dr. Isaza is working with the result of these analyses in the Ponce School of Medicine. In her research, she is studying the potential relation between DNA repair capacity and the miRNAs expression of miR-365 and miR-889.

CHAPTER 7: MCO COMPARED TO THE USE OF A VOLCANO PLOT

7.1. 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 [95]. This method is implemented in different software packages. The MCO method proposed in this thesis 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 $\log_{10-}$ transformed p-values from the gene specific t-test against the $\log_2$ 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 [96].

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

*Fold Chang*e 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):

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

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.

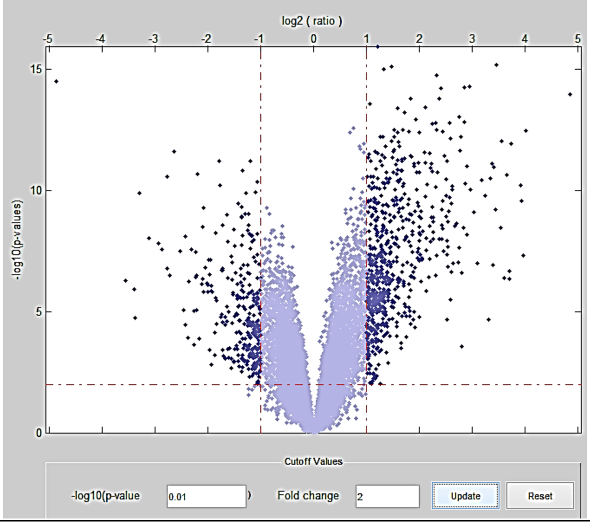
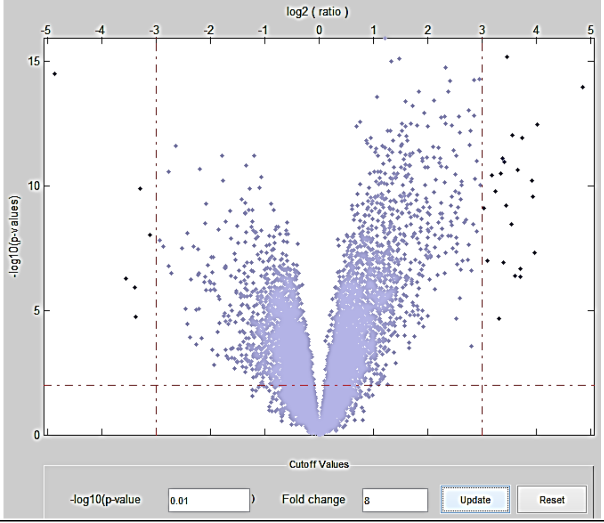
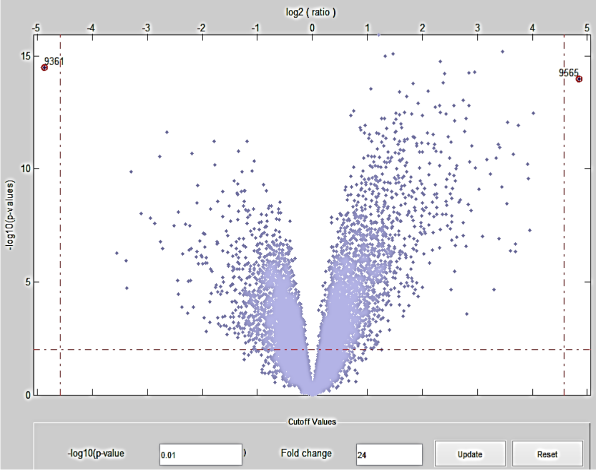
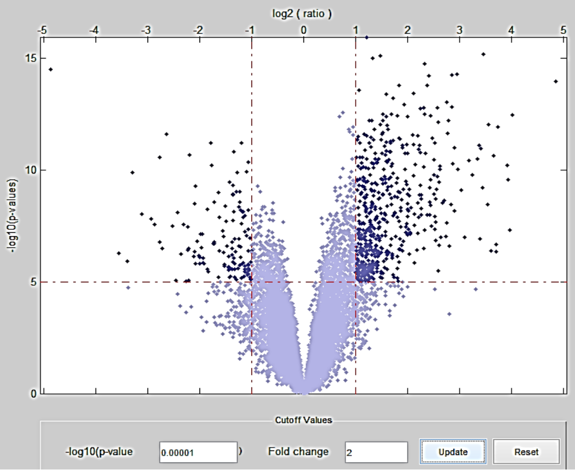
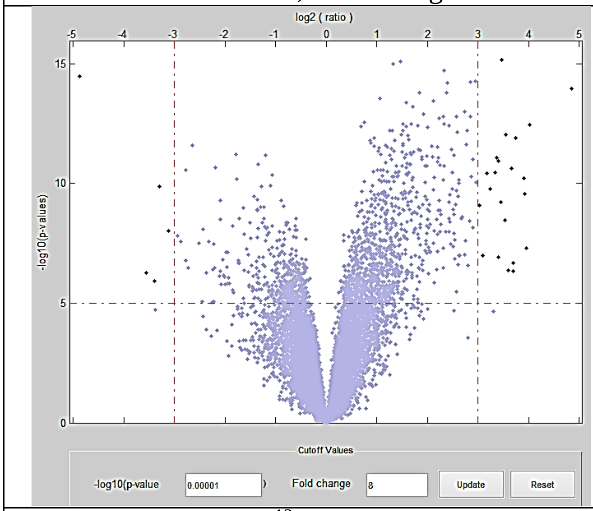
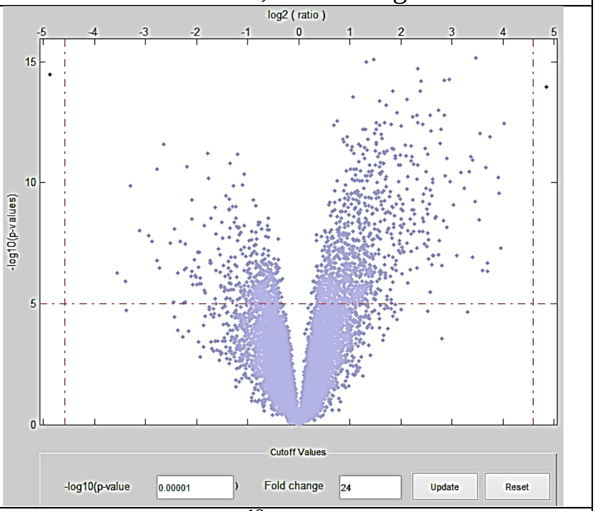
The analysis is required to choose threshold values for both measures to select important genes. The volcano plot is available in the bioinformatics toolbox for MatLab.

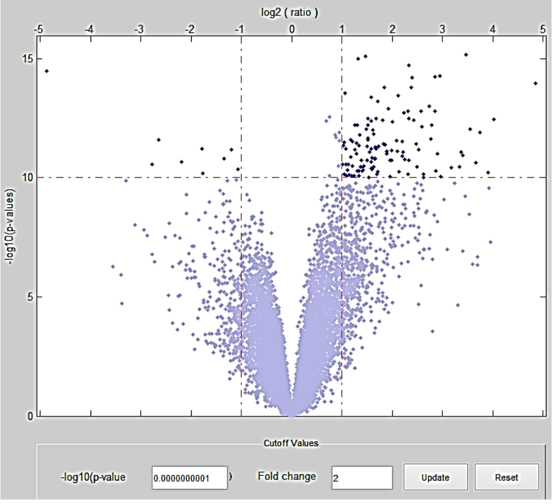
Given a particular microarray set with genetic expression levels measured. In two distinct states, the tool in MatLab obtains a p-value per gene using a t-test, and measures the FC in a logarithmic scale with base 2.

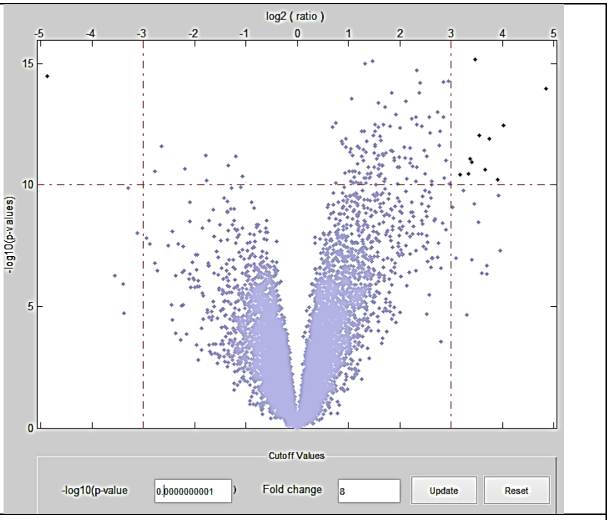
The cases of lung cancer microarray and leukemia presented in chapters 5 and 6 respectively will be revisited here for comparison purposes.

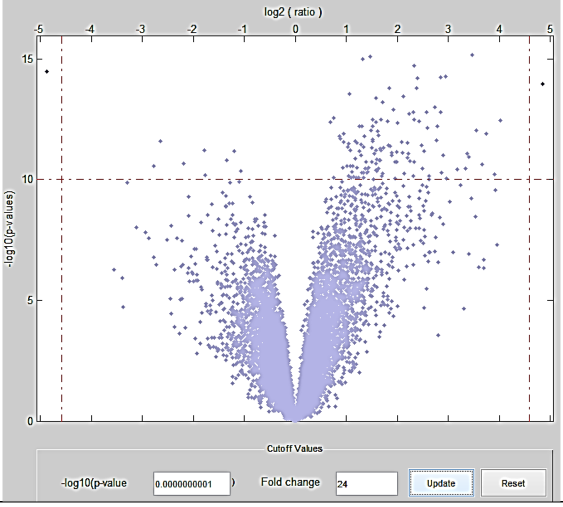
7.2. Comparison of Volcano with Lung Cancer Case

The original database GDS3257 of lung cancer was used for this analysis. Fifteen samples of HNS and sixteen samples of CNS were used to build the Volcano plot. As mentioned previously, Volcano plot requires the user to define thresholds for two parameters: p-value and FC to select genes. This analysis has many different combinations and selection of p-values and FC. However choice p-values and FCs for this analysis was according to the response of the MCO analysis. A 3^2 factorial experiment was used to explore these parameters as shown in Figure 7-1.The results are shown in Table 7-1.

P-value =10^{-2} ; Fold change = 2	P-value = 10^{-2} ; Fold change = 8
 Cutoff Values -log10(p-value 0.01) Fold change 2 Update Reset	 Cutoff Values -log10(p-value 0.01) Fold change 8 Update Reset
P-value = 10^{-2} ; Fold change = 24	P-value = 10^{-7} ; Fold change = 2
 Cutoff Values -log10(p-value 0.01) Fold change 24 Update Reset	 Cutoff Values -log10(p-value 0.00001) Fold change 2 Update Reset
P-value = 10^{-7} ; Fold change = 8	P-value = 10^{-7} ; Fold change = 24
 Cutoff Values -log10(p-value 0.00001) Fold change 8 Update Reset	 Cutoff Values -log10(p-value 0.00001) Fold change 24 Update Reset
P-value = 10^{-12} ; Fold change = 2	P-value = 10^{-12} ; Fold change = 8

<table>
<tr>
<td>

-log10(p-value 0.0000000001) Fold change 2</td>
<td>

-log10(p-value 0.0000000001) Fold change 8</td>
</tr>
<tr>
<td colspan="2">P-value = 10^{-12} ; Fold change = 24</td>
</tr>
<tr>
<td colspan="2">

-log10(p-value 0.0000000001) Fold change 24</td>
</tr>
</table>

Figure 7-1: Figures represent the results of genes with height DE using Volcano plot and varying the p-values and FC.

From Table 7-1, it can be seen how the results depend highly in the user's selection of thresholds. The combinations that fully matches the output of MCO are the ones with FC = 24 at any value of p-value.

In the case of changing FC=24 and for any p-value the genes with the most DE are two genes: one overexpressed and one underexpressed. These genes are SPP1 and RAGE, which are the same genes obtained with MCO method.

Table 7-1: 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

7.3. Comparison of Volcano with Leukemia Case

In this section, eleven healthy tissues and ten cancerous tissues were used from Leukemia database GSE2403 to build the Volcano plot. For this case, the p-values and FC were small, because the volcano tool was not permitted other values like lung cancer analysis. In this case, the relative expression for each gene required a previous transformation using $\log_2$. Due to, the data of leukemia microarray was in natural scale. After that, a 3^2 factorial design was used to explore the variation of results to changes in the thresholds for p-values and FC. The results are shown in Table 7-2.

Table 7-2: Summary of genes using volcano plot

P-value	Fold change	Differential expression	Overexpressed	Underexpressed
0.03	2	140	17	123
0.03	3	13	1	12
0.03	4	2	0	2
0.02	2	96	11	85
0.02	3	11	10	1
0.02	4	2	0	2
0.01	2	57	5	52
0.01	3	9	1	8
0.01	4	2	0	2

In this case, there is no match between the solutions found with the volcano plot and those found with MCO. Table 7-3 explores this issue. The reason behind the mismatch has done with, potentially.

Table 7-3: Table for comparison the different values between important genes from MCO and Volcano methods.

Method	Genes	|Mean(C)-Mean(H)|	|Median(C)-Median(H)|	Log2(|Mean(H)-Mean(C)|)	Log2(|Median(H)-Median(C)|)	p-value (H and C)
MCO	31687_f_at	3930.46	4438.15	11.94048279	12.11574271	0.7152
Volcano	31382_f_at	11.35	8.7	3.504851482	3.121015401	0.0001
	40354_at	10.32	7.6	3.367879325	2.925999419	0.0075

The volcano plot, as habitually coded, assumes that the data should be normally distributed. If the analyzer does not have enough expertise in the process of choosing the parameters and distinguish data, then erroneous conclusions could be obtained.

The objective to compare volcano plot method with MCO method, proposed in this thesis, was to demonstrate the difference on the perspectives of analyses. MCO tries to depart from the definition of a priori preference structures by the uses, and thus, from the use of ad-hoc threshold values among especially in the presence of multiple and potentially incommensurate performance measures.

CHAPTER 8: CONCLUSIONS AND FUTURE WORK

8.1. Conclusions

The proposed method in this thesis was evaluated through three case studies: lung cancer, leukemia, and breast cancer. The first two cases involve gene databases, while the last one considers a microRNA database.

The tool coded in MatLab in this thesis can currently analyze five criteria, implying that 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 researches, as the analysis does not require normalization, preference of objectives, parameter adjustments by user, or the definition of a threshold value.

In the case study in lung cancer, the general conclusions are: RAGE and SPP1 showed large change between a state of health and a state of cancer. Moreover, SPP1 showed large change between Healthy Current Smoker and Cancer Non Smoker, and RAGE showed large change between Healthy Never Smoker and Cancer Current Smoker. Also, XIST showed a large difference when comparing Never Smoker and Current Smoker (both in healthy and cancer tissues). This chapter developed the pseudo meta-analysis with four PMs. This last analysis supports the potential of the proposed method for meta-analysis. The case study in leukemia the gene in position 31687_f_at is known cancer biomarker also, as found in the literature. All of that give as method relevant biological evidence.

The third case of the proposed method is important because it demonstrates that the analysis strategy is not only applicable to microarrays, but also that it could be used to analyze other - omics. This means that the method could be applied to other types of data with similar experimental layout.

8.2. Future work

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, probably, experimental validation. It is certainly also recommended to effectively test the Matlab tool with different –omics.

Finally, because the perspective of MCO and its deterministic nature is different from those methods offered by the statistics field, it is necessary that further comparison be carried out in terms of habitual statistical properties and measures of the respective analyses' outputs.

REFERENCES

[1] R. Siegel, J. Ma, Z. Zou, and A. Jemal, “Cancer statistics, 2014,” *CA. Cancer J. Clin.*, vol. 64, no. 1, pp. 9–29, 2014.

[2] E. W. Sayers, T. Barrett, D. A. Benson, E. Bolton, S. H. Bryant, K. Canese, V. Chetvernin, D. M. Church, M. DiCuccio, S. Federhen, M. Feolo, I. M. Fingerman, L. Y. Geer, W. Helmberg, Y. Kapustin, S. Krasnov, D. Landsman, D. J. Lipman, Z. Lu, T. L. Madden, T. Madej, D. R. Maglott, A. Marchler-Bauer, V. Miller, I. Karsch-Mizrachi, J. Ostell, A. Panchenko, L. Phan, K. D. Pruitt, G. D. Schuler, E. Sequeira, S. T. Sherry, M. Shumway, K. Sirotkin, D. Slotta, A. Souvorov, G. Starchenko, T. A. Tatusova, L. Wagner, Y. Wang, W. J. Wilbur, E. Yaschenko, and J. Ye, “Database resources of the National Center for Biotechnology Information,” *Nucleic Acids Res.*, vol. 40, no. D1, pp. D13–D25, Dec. 2011.

[3] M. N. McCall, P. N. Murakami, M. Lukk, W. Huber, and R. A. Irizarry, “Assessing affymetrix GeneChip microarray quality,” *BMC Bioinformatics*, vol. 12, no. 1, p. 137, May 2011.

[4] “DNA Microarray 2013: A Focus on Sales Growth,” *PR Newswire*, New York, United States, 06-Feb-2013.

[5] “Research and Markets Offers Report: Microarray Markets,” *Prof. Serv. Close - Up*, Aug. 2013.

[6] A. Mohammadi, M. H. Saraee, and M. Salehi, “Identification of disease-causing genes using microarray data mining and Gene Ontology,” *BMC Med. Genomics*, vol. 4, no. 1, p. 12, Jan. 2011.

[7] S. P. Glasser and S. Duval, “Meta-Analysis,” in *Essentials of Clinical Research*, S. P. Glasser, Ed. Springer Netherlands, 2008, pp. 159–177.

[8] M. L. Sánchez-Peña, C. E. Isaza, J. Pérez-Morales, C. Rodríguez-Padilla, J. M. Castro, and M. Cabrera-Ríos, “Identification of potential biomarkers from microarray experiments using multiple criteria optimization,” *Cancer Med.*, vol. 2, no. 2, pp. 253–265, 2013.

[9] R. B. Statnikov, A. Bordetsky, and A. Statnikov, “Multicriteria analysis of real-life engineering optimization problems: statement and solution,” *Nonlinear Anal. Theory Methods Appl.*, vol. 63, no. 5–7, pp. e685–e696, Dec. 2005.

[10] K. K. Jain, *The Handbook of Biomarkers*. Portland, OR: SciTech Book News, 2010.

[11] Exigon, “What are microRNAs?” [Online]. Available: http://www.exiqon.com/what-are-microRNAs. [Accessed: 08-Aug-2014].

[12] A. L. Oom, B. A. Humphries, and C. Yang, “MicroRNAs: Novel Players in Cancer Diagnosis and Therapies,” *BioMed Res. Int.*, vol. 2014, p. 959461, 2014.

[13] “Patient’s Genetic Background,” *National Cancer Institute*. [Online]. Available: http://www.cancer.gov/cancertopics/understandingcancer/geneticbackground/page22. [Accessed: 08-Aug-2014].

[14] L. Uribe Mastache, “Metodología para el análisis de datos de Microarreglos para la detección de Cáncer,” Universidad Autonoma de Nuevo Leon, 2008.

[15] H. Perez Vicente, “Diagnóstico de cáncer a partir de datos de Microarreglos,” Universidad Autonoma de Nuevo Leon, 2008.

[16] M. L. Sanchez Pena, “Identification of Potential Cancer Biomarkers through Multiple Criteria Optimization Using Microarray Data,” University of Puerto Rico - Mayaguez campus, 2010.

[17] A. B. Rodríguez Yáñez, “Process Windows Considering two conflicting criteria: The injection molding case,” University of Puerto Rico - Mayaguez campus, 2011.

[18] National Cancer Institute, *Understanding cancer and related topics [electronic resource]: understanding gene testing*. Bethesda, Md.]: National Cancer Institute, 2011.

[19] Zu-hua Gao, “Challenges and opportunities in the era of personalized medicine,” *Oncol. Exch.*, vol. 12, no. 4, pp. 11–14, Nov. 2013.

[20] “American Cancer Society | Information and Resources for Cancer: Breast, Colon, Lung, Prostate, Skin,” 2014. [Online]. Available: http://www.cancer.org/. [Accessed: 03-May-2014].

[21] B. Alberts and D. Bray, *Essential Cell Biology*. Garland Science, 2004.

[22] B. Wang and Y. Xi, “Challenges for MicroRNA Microarray Data Analysis,” *Microarrays Basel Switz.*, vol. 2, no. 2, Jun. 2013.

[23] P. J. Hurd and C. J. Nelson, “Advantages of next-generation sequencing versus the microarray in epigenetic research,” *Brief. Funct. Genomic. Proteomic.*, vol. 8, no. 3, pp. 174–183, May 2009.

[24] A. H. Mohamed Salleh, M. S. Mohamad, S. Deris, and R. M. Illias, “A Review On Pathway Analysis Software Based On Microarray Data Interpretation,” *Int. J. Bio-Sci. Bio-Technol.*, vol. 5, no. 4, pp. 149–157, Aug. 2013.

[25] V. Tyagi and A. Mishra, “A Survey on Different Feature Selection Methods for Microarray Data Analysis,” *Int. J. Comput. Appl.*, vol. 67, no. 16, pp. 36–40, Apr. 2013.

[26] O. Lyttleton, A. Wright, D. Treanor, P. Quirke, and P. Lewis, “Extending the tissue microarray data exchange specification for inclusion of data analysis results,” *J. Pathol. Inform.*, vol. 2, p. 17, 2011.

[27] S. Draghici, *Data Analysis Tools for DNA Microarrays*. .

[28] Biomarkers Definitions Working Group., “Biomarkers and surrogate endpoints: preferred definitions and conceptual framework,” *Clin. Pharmacol. Ther.*, vol. 69, no. 3, pp. 89–95, Mar. 2001.

[29] Z. Fang, J. Martin, and Z. Wang, “Statistical methods for identifying differentially expressed genes in RNA-Seq experiments,” *Cell Biosci.*, vol. 2, no. 1, p. 26, 2012.

[30] C. E. Isaza, L. Uribe, H. A. Pérez, C. Rodríguez, and M. Cabrera-Ríos, “Cancer Diagnosis through Microarray Analysis using the Mann-Whitney statistical test,” *IIE Annu. Conf. Proc.*, pp. 736–741, 2009.

[31] A. N. Burska, K. Roget, M. Blits, L. Soto Gomez, F. van de Loo, L. D. Hazelwood, C. L. Verweij, A. Rowe, G. N. Goulielmos, L. G. M. van Baarsen, and F. Ponchel, “Gene expression analysis in RA: towards personalized medicine,” *Pharmacogenomics J.*, vol. 14, no. 2, pp. 93–106, Apr. 2014.

[32] J. Li, R. J. Coates, M. Gwinn, and M. J. Khoury, “Steroid 5-{alpha}-reductase Type 2 (SRD5a2) gene polymorphisms and risk of prostate cancer: a HuGE review,” *Am. J. Epidemiol.*, vol. 171, no. 1, pp. 1–13, Jan. 2010.

[33] P. H. G. at CDC, “Genomics|HuGENet|Reviews,” 15-Dec-2014. [Online]. Available: http://www.cdc.gov/genomics/hugenet/reviews/. [Accessed: 15-Dec-2014].

[34] X. Wang, D. D. Kang, K. Shen, C. Song, S. Lu, L.-C. Chang, S. G. Liao, Z. Huo, S. Tang, Y. Ding, N. Kaminski, E. Sibille, Y. Lin, J. Li, and G. C. Tseng, “An R package suite for microarray meta-analysis in quality control, differentially expressed gene analysis and pathway enrichment detection,” *Bioinforma. Oxf. Engl.*, vol. 28, no. 19, pp. 2534–2536, Oct. 2012.

[35] G. C. Tseng, D. Ghosh, and E. Feingold, “Comprehensive literature review and statistical considerations for microarray meta-analysis,” *Nucleic Acids Res.*, vol. 40, no. 9, pp. 3785–3799, May 2012.

[36] A. Ramasamy, A. Mondry, C. C. Holmes, and D. G. Altman, “Key issues in conducting a meta-analysis of gene expression microarray datasets,” *PLoS Med.*, vol. 5, no. 9, p. e184, Sep. 2008.

[37] S. Lu, J. Li, C. Song, K. Shen, and G. C. Tseng, “Biomarker detection in the integration of multiple multi-class genomic studies,” *Bioinformatics*, vol. 26, no. 3, pp. 333–340, Feb. 2010.

[38] Z. Gan, J. Wang, N. Salomonis, J. C. Stowe, G. G. Haddad, A. D. McCulloch, I. Altintas, and A. C. Zambon, “MAAMD: a workflow to standardize meta-analyses and comparison of affymetrix microarray data,” *BMC Bioinformatics*, vol. 15, no. 1, p. 69, Mar. 2014.

[39] M. A. Velazquez, D. Claudio, and A. R. Ravindran, “Experiments in Multiple Criteria Selection Problems with Multiple Decision Makers,” *IIE Annu. Conf. Proc.*, Jan. 2008.

[40] O. B. Augusto, F. Bennis, and S. Caro, “A new method for decision making in multi-objective optimization problems,” *Pesqui. Oper.*, vol. 32, no. 2, pp. 331–369, Aug. 2012.

[41] P.-H. Huang, J.-S. Tsai, and W.-T. Lin, “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.*, vol. 168, no. 1–4, pp. 141–158, Sep. 2010.

[42] R. Statnikov, J. Matusov, and A. Statnikov, “Multicriteria Engineering Optimization Problems: Statement, Solution and Applications,” *J. Optim. Theory Appl.*, vol. 155, no. 2, pp. 355–375, Nov. 2012.

[43] J. C. Rajapakse and P. A. Mundra, “Multiclass Gene Selection Using Pareto-Fronts,” *IEEEACM Trans Comput Biol Bioinforma.*, vol. 10, no. 1, pp. 87–97, Jan. 2013.

[44] K. Deb, *Multi-Objective Optimization using Evolutionary Algorithms*. Wiley, 2001.

[45] S. Greco and M. Ehrgott, *Multiple Criteria Decision Analysis: State of the Art Surveys*. Springer, 2005.

[46] The International Agency for Research on Cancer (IARC), “Global battle against cancer w on’t be won with treatment alone Effective prevention measures urgently needed to prevent cancer crisis,” Mar. 2014.

[47] Z. Lu, “PubMed and beyond: a survey of web tools for searching biomedical literature,” *Database J. Biol. Databases Curation*, vol. 2011, Jan. 2011.

[48] M. T. Landi, T. Dracheva, M. Rotunno, J. D. Figueroa, H. Liu, A. Dasgupta, F. E. Mann, J. Fukuoka, M. Hames, A. W. Bergen, S. E. Murphy, P. Yang, A. C. Pesatori, D. Consonni, P. A. Bertazzi, S. Wacholder, J. H. Shih, N. E. Caporaso, and J. Jen, “Gene Expression Signature of Cigarette Smoking and Its Role in Lung Adenocarcinoma Development and Survival,” *PLoS ONE*, vol. 3, no. 2, Feb. 2008.

[49] M. Cabrera-Rios, “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-Mar-2014].

[50] G. P. Sims, D. C. Rowe, S. T. Rietdijk, R. Herbst, and A. J. Coyle, “HMGB1 and RAGE in inflammation and cancer,” *Annu. Rev. Immunol.*, vol. 28, pp. 367–388, 2010.

[51] “SPP1 Gene - GeneCards | OSTP Protein | OSTP Antibody.” [Online]. Available: http://www.genecards.org/cgi-bin/carddisp.pl?gene=SPP1. [Accessed: 06-May-2014].

[52] “SPP1 (secreted phosphoprotein 1).” [Online]. Available: http://atlasgeneticsoncology.org/Genes/GC_SPP1.html. [Accessed: 06-May-2014].

[53] R. Kang, D. Tang, N. Schapiro, T. Loux, K. Livesey, T. Billiar, H. Wang, B. Van Houten, M. Lotze, and H. Zeh, “The HMGB1/RAGE inflammatory pathway promotes pancreatic tumor growth by regulating mitochondrial bioenergetics,” *Oncogene*, vol. 33, no. 5, pp. 567–577, Jan. 2014.

[54] L. J. Sparvero, D. Asafu-Adjei, R. Kang, D. Tang, N. Amin, J. Im, R. Rutledge, B. Lin, A. A. Amoscato, H. J. Zeh, and M. T. Lotze, “RAGE (Receptor for Advanced Glycation Endproducts), RAGE ligands, and their role in cancer and inflammation,” *J. Transl. Med.*, vol. 7, p. 17, 2009.

[55] M. Dahlmann, A. Okhrimenko, P. Marcinkowski, M. Osterland, P. Herrmann, J. Smith, C. W. Heizmann, P. M. Schlag, and U. Stein, “RAGE mediates S100A4-induced cell motility via MAPK/ERK and hypoxia signaling and is a prognostic biomarker for human colorectal cancer metastasis,” *Oncotarget*, vol. 5, no. 10, pp. 3220–3233, Apr. 2014.

[56] X. C. Xu, X. Abuduhadeer, W. B. Zhang, T. Li, H. Gao, and Y. H. Wang, “Knockdown of RAGE Inhibits Growth and Invasion of Gastric Cancer Cells,” *Eur. J. Histochem. EJH*, vol. 57, no. 4, Nov. 2013.

[57] A.-M. Yaser, Y. Huang, R.-R. Zhou, G.-S. Hu, M.-F. Xiao, Z.-B. Huang, C.-J. Duan, W. Tian, D.-L. Tang, and X.-G. Fan, “The Role of Receptor for Advanced Glycation End Products (RAGE) in the Proliferation of Hepatocellular Carcinoma,” *Int. J. Mol. Sci.*, vol. 13, no. 5, pp. 5982–5997, May 2012.

[58] S. T. Buckley and C. Ehrhardt, “The Receptor for Advanced Glycation End Products (RAGE) and the Lung,” *BioMed Res. Int.*, vol. 2010, Jan. 2010.

[59] M. Mardani, A. Andisheh-Tadbir, B. Khademi, M. J. Fattahi, S. Shafiee, and M. Asad-Zadeh, “Serum levels of osteopontin as a prognostic factor in patients with oral squamous cell carcinoma,” *Tumour Biol. J. Int. Soc. Oncodevelopmental Biol. Med.*, vol. 35, no. 4, pp. 3827–3829, Apr. 2014.

[60] H. Zhang, H. Liu, D. Yuan, Z. Wang, Y. Wang, and Y. Song, “Prognostic value of secreted phosphoprotein-1 in pleural effusion associated with non-small cell lung cancer,” *BMC Cancer*, vol. 14, p. 280, Apr. 2014.

[61] C. A. Dalla-Torre, M. Yoshimoto, C.-H. Lee, A. M. Joshua, S. R. de Toledo, A. S. Petrilli, J. A. Andrade, S. Chilton-MacNeill, M. Zielenska, and J. A. Squire, “Effects of THBS3, SPARC and SPP1 expression on biological behavior and survival in patients with osteosarcoma,” *BMC Cancer*, vol. 6, p. 237, Oct. 2006.

[62] A. Zaravinos, G. I. Lambrou, D. Volanis, D. Delakas, and D. A. Spandidos, “Spotlight on Differentially Expressed Genes in Urinary Bladder Cancer,” *PLoS ONE*, vol. 6, no. 4, Apr. 2011.

[63] Z. Ding, C.-J. Wu, G. C. Chu, Y. Xiao, D. Ho, J. Zhang, S. R. Perry, E. S. Labrot, X. Wu, R. Lis, Y. Hoshida, D. Hiller, B. Hu, S. Jiang, H. Zheng, A. H. Stegh, K. L. Scott, S. Signoretti, N. Bardeesy, Y. A. Wang, D. E. Hill, T. R. Golub, M. J. Stampfer, W. H. Wong, M. Loda, L. Mucci, L. Chin, and R. A. DePinho, “SMAD4-dependent barrier constrains prostate cancer growth and metastatic progression,” *Nature*, vol. 470, no. 7333, pp. 269–273, Feb. 2011.

[64] A. Thomas, U. Mahantshetty, S. Kannan, K. Deodhar, S. K. Shrivastava, C. Kumar-Sinha, and R. Mulherkar, “Expression profiling of cervical cancers in Indian women at different stages to identify gene signatures during progression of the disease,” *Cancer Med.*, vol. 2, no. 6, pp. 836–848, Dec. 2013.

[65] S. Das, R. S. Samant, and L. A. Shevde, “Hedgehog Signaling Induced by Breast Cancer Cells Promotes Osteoclastogenesis and Osteolysis,” *J. Biol. Chem.*, vol. 286, no. 11, pp. 9612–9622, Mar. 2011.

[66] G. F. Weber, G. S. Lett, and N. C. Haubein, “Osteopontin is a marker for cancer aggressiveness and patient survival,” *Br. J. Cancer*, vol. 103, no. 6, pp. 861–869, Sep. 2010.

[67] M. D. Tabernero, A. B. Espinosa, A. Maillo, O. Rebelo, J. F. Vera, J. M. Sayagues, M. Merino, P. Diaz, P. Sousa, and A. Orfao, “Patient gender is associated with distinct patterns of chromosomal abnormalities and sex chromosome linked gene-expression profiles in meningiomas,” *The Oncologist*, vol. 12, no. 10, pp. 1225–1236, Oct. 2007.

[68] S. M. Sirchia, S. Tabano, L. Monti, M. P. Recalcati, M. Gariboldi, F. R. Grati, G. Porta, P. Finelli, P. Radice, and M. Miozzo, “Misbehaviour of XIST RNA in Breast Cancer Cells,” *PLoS ONE*, vol. 4, no. 5, May 2009.

[69] K.-C. Huang, P. H. Rao, C. C. Lau, E. Heard, S.-K. Ng, C. Brown, S. C. Mok, R. S. Berkowitz, and S.-W. Ng, “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.*, vol. 1, no. 10, pp. 769–776, Aug. 2002.

[70] M. C. Morissette, M. Lamontagne, J.-C. Berube, G. Gaschler, A. Williams, C. Yauk, C. Couture, M. Laviolette, J. C. Hogg, W. Timens, S. Halappanavar, M. R. Stampfli, and Y. Bosse, “Impact of Cigarette Smoke on the Human and Mouse Lungs: A Gene-Expression Comparison Study,” *PLoS ONE*, vol. 9, no. 3, Mar. 2014.

[71] Y. Li, S.-Q. Tan, Q.-H. Ma, L. Li, Z.-Y. Huang, Y. Wang, and S.-W. Li, “CYP1B1 C4326G polymorphism and susceptibility to cervical cancer in Chinese Han women,” *Tumour Biol. J. Int. Soc. Oncodevelopmental Biol. Med.*, vol. 34, no. 6, pp. 3561–3567, Dec. 2013.

[72] E. G. Shatalova, A. J. P. Klein-Szanto, K. Devarajan, E. Cukierman, and M. L. Clapper, “Estrogen and Cytochrome P450 1B1 Contribute to Both Early- and Late-Stage Head and Neck Carcinogenesis,” *Cancer Prev. Res. Phila. Pa*, vol. 4, no. 1, pp. 107–115, Jan. 2011.

[73] J. Beuten, J. A. L. Gelfond, J. J. Byrne, I. Balic, A. C. Crandall, T. L. Johnson-Pais, I. M. Thompson, D. K. Price, and R. J. Leach, “CYP1B1 variants are associated with prostate cancer in non-Hispanic and Hispanic Caucasians,” *Carcinogenesis*, vol. 29, no. 9, pp. 1751–1757, Sep. 2008.

[74] M. K. Herroon, E. Rajagurubandara, A. L. Hardaway, K. Powell, A. Turchick, D. Feldmann, and I. Podgorski, “Bone marrow adipocytes promote tumor growth in bone via FABP4-dependent mechanisms,” *Oncotarget*, vol. 4, no. 11, pp. 2108–2123, Oct. 2013.

[75] “Adipocytes Fuel Tumor Growth at Metastatic Sites,” *Cancer Discov.*, vol. 1, no. 7, pp. 548–548, Dec. 2011.

[76] S. Cameron, L. M. de Long, M. Hazar-Rethinam, E. Topkas, L. Endo-Munoz, A. Cumming, O. Gannon, A. Guminski, and N. Saunders, “Focal overexpression of CEACAM6 contributes to enhanced tumourigenesis in head and neck cancer via suppression of apoptosis,” *Mol. Cancer*, vol. 11, p. 74, Sep. 2012.

[77] A. Mukhopadhyay, T. Khoury, L. Stein, P. Shrikant, and A. K. Sood, “Prostate derived Ets transcription factor and Carcinoembryonic antigen related cell adhesion molecule 6 constitute a highly active oncogenic axis in breast cancer,” *Oncotarget*, vol. 4, no. 4, pp. 610–621, Apr. 2013.

[78] C. Ilantzis, L. Demarte, R. A. Screaton, and C. P. Stanners, “Deregulated Expression of the Human Tumor Marker CEA and CEA Family Member CEACAM6 Disrupts Tissue Architecture and Blocks Colonocyte Differentiation,” *Neoplasia N. Y. N*, vol. 4, no. 2, pp. 151–163, Mar. 2002.

[79] B. B. Singer, I. Scheffrahn, R. Kammerer, N. Suttorp, S. Ergun, and H. Slevogt, “Deregulation of the CEACAM Expression Pattern Causes Undifferentiated Cell Growth in Human Lung Adenocarcinoma Cells,” *PLoS ONE*, vol. 5, no. 1, Jan. 2010.

[80] H. Lou, H. Li, M. Yeager, K. Im, B. Gold, T. D. Schneider, J. F. Fraumeni, S. J. Chanock, S. K. Anderson, and M. Dean, “Promoter variants in the MSMB gene associated with prostate cancer regulate MSMB/NCOA4 fusion transcripts,” *Hum. Genet.*, vol. 131, no. 9, pp. 1453–1466, Sep. 2012.

[81] R. Chari, K. M. Lonergan, R. T. Ng, C. MacAulay, W. L. Lam, and S. Lam, “Effect of active smoking on the human bronchial epithelium transcriptome,” *BMC Genomics*, vol. 8, p. 297, Aug. 2007.

[82] S.-J. Yang, A. Yokoyama, T. Yokoyama, Y.-C. Huang, S.-Y. Wu, Y. Shao, J. Niu, J. Wang, Y. Liu, X.-Q. Zhou, and C.-X. Yang, “Relationship between genetic polymorphisms of ALDH2 and ADH1B and esophageal cancer risk: A meta-analysis,” *World J. Gastroenterol. WJG*, vol. 16, no. 33, pp. 4210–4220, Sep. 2010.

[83] M. Crous-Bou, G. Rennert, D. Cuadras, R. Salazar, D. Cordero, H. Saltz Rennert, F. Lejbkowicz, L. Kopelovich, S. Monroe Lipkin, S. Bernard Gruber, and V. Moreno, “Polymorphisms in Alcohol Metabolism Genes ADH1B and ALDH2, Alcohol Consumption and Colorectal Cancer,” *PLoS ONE*, vol. 8, no. 11, Nov. 2013.

[84] A. M. Hakenewerth, R. C. Millikan, I. Rusyn, A. H. Herring, K. E. North, J. S. Barnholtz-Sloan, W. F. Funkhouser, M. C. Weissler, and A. F. Olshan, “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.*, vol. 20, no. 11, pp. 2438–2449, Nov. 2011.

[85] T. Sasaki, M. Horikawa, K. Orikasa, M. Sato, Y. Arai, Y. Mitachi, M. Mizugaki, M. Ishikawa, and M. Hiratsuka, “Possible Relationship Between the Risk of Japanese Bladder Cancer Cases and the CYP4B1 Genotype,” *Jpn. J. Clin. Oncol.*, vol. 38, no. 9, pp. 634–640, Sep. 2008.

[86] L. Boyero, A. Sanchez-Palencia, M. T. Miranda-Leon, F. Hernandez-Escobar, J. A. Gomez-Capilla, and M. E. Farez-Vidal, “Survival, Classifications, and Desmosomal Plaque Genes in Non-Small Cell Lung Cancer,” *Int. J. Med. Sci.*, vol. 10, no. 9, pp. 1166–1173, Jul. 2013.

[87] H. Jiang, X. Lin, Y. Liu, W. Gong, X. Ma, Y. Yu, Y. Xie, X. Sun, Y. Feng, V. Janzen, and T. Chen, “Transformation of Epithelial Ovarian Cancer Stemlike Cells into Mesenchymal Lineage via EMT Results in Cellular Heterogeneity and Supports Tumor Engraftment,” *Mol. Med.*, vol. 18, no. 1, pp. 1197–1208, Jul. 2012.

[88] W.-L. Zhu, B.-L. Fan, D.-L. Liu, and W.-X. Zhu, “Abnormal Expression of Fibrinogen Gamma (FGG) and Plasma Level of Fibrinogen in Patients with Hepatocellular Carcinoma,” *Anticancer Res.*, vol. 29, no. 7, pp. 2531–2534, Jul. 2009.

[89] National Cancer Institute, “What You Need To Know About^{TM} Leukemia - National Cancer Institute.” [Online]. Available: http://www.cancer.gov/cancertopics/wyntk/leukemia/page4. [Accessed: 17-Jun-2014].

[90] American Cancer Society, “How is chronic lymphocytic leukemia staged?” [Online]. Available: http://www.cancer.org/cancer/leukemia-chroniclymphocyticcll/detailedguide/leukemia-chronic-lymphocytic-staging. [Accessed: 17-Jun-2014].

[91] S. Fält, M. Merup, G. Gahrton, B. Lambert, and A. Wennborg, “Identification of progression markers in B-CLL by gene expression profiling,” *Exp. Hematol.*, vol. 33, no. 8, pp. 883–893, Aug. 2005.

[92] “HBB hemoglobin, beta [Homo sapiens (human)] - Gene - NCBI.” [Online]. Available: http://www.ncbi.nlm.nih.gov/gene?cmd=Retrieve&dopt=Graphics&list_uids=3043. [Accessed: 19-Jun-2014].

[93] K. Felekkis, E. Touvana, C. Stefanou, and C. Deltas, “microRNAs: a newly described class of encoded molecules that play a role in health and disease,” *Hippokratia*, vol. 14, no. 4, pp. 236–240, 2010.

[94] N. Kosaka, H. Iguchi, and T. Ochiya, “Circulating microRNA in body fluid: a new potential biomarker for cancer diagnosis and prognosis,” *Cancer Sci.*, vol. 101, no. 10, pp. 2087–2092, Oct. 2010.

[95] W. Li, “Volcano plots in analyzing differential expressions with mRNA microarrays,” *J. Bioinform. Comput. Biol.*, vol. 10, no. 6, p. 1231003, Dec. 2012.

[96] X. Cui and G. A. Churchill, “Statistical tests for differential expression in cDNA microarray experiments,” *Genome Biol.*, vol. 4, no. 4, p. 210, 2003.

Appendix 1: MatLab Code

```
%Análisis de frontera Pareto de cinco criterios
%Autor: Katia I Camacho Cáceres

dataT = load('data5Criteria.txt'); %Cargar la data
[x,y] = size(dataT); % data completa x=num filas, y=num columnas
data = dataT(:,2:end); %se toma solo las columnas de los criterios
[n,m]=size(data); %n=num filas (k=PM), m=num columnas (j = criterios)
c1 = 1000*ones(n,n,m);   % matriz primera condición con j criterios
for j=1:m
    for a=1:n
        for b=1:n
            if data(a,j) == data(b,j) %condición 1.1
                c1(a,b,j)=0;
            elseif data(a,j)<data(b,j)
                c1(a,b,j)=-1;
            end
        end
    end
end

% Procedimiento para sumar c1 para cinco criterios
c2=zeros(n,n);  %matriz segunda condición
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

% Procedimiento para encontrar conjunto dominado cd y no dominado cnd
cnd = zeros(x,y); %matriz del conjunto no dominado
cd = zeros(x,y);  % matriz del conjunto dominado
i=0; %contador para cd
j=0; %contador para cnd
for a=1:x
    sumfila=sum(c2(a,:));
     if sumfila>=5000; % conjunto dominado
        i=i+1;
        cd(i,:)=dataT(a,:);
    else  % conjunto no dominado
        j=j+1;
        cnd(j,:)=dataT(a,:);
    end
```

```
end

index = 1:x;
disp([round(index') cd]);
disp([round(index') cnd]);
%Mostrar el Conjunto no dominado en un notepad, con los datos de
%Posicion,f1,f2, f3, f4, f5c
disp('   Conjunto no dominado    ');
cnd=cnd(1:j,:);
filecnd = fopen('cnd5CriteriaBio.txt','w');
fprintf(filecnd,'%6s   %12s   %12s   %12s   %12s
%12s\r\n','Posicion','F1','F2','F3', 'F4', 'F5');
fprintf(filecnd,'%6.4f   %12.4f   %12.4f   %12.4f   %12.4f
%12.4f\r\n',cnd');
fclose(filecnd);
```

Appendix 2: Design of Experiments for Volcano Plot and Lung Cancer case

Table 1 represents the original data used in the analysis of the volcano plot for lung cancer microarray:

Table A2-1: The Upper (P-value) and Lower (Fold change) limits for Volcano plot

P-value	Fold change	Differential expression
10^{-2}	2	934
10^{-2}	8	29
10^{-2}	24	2
10^{-7}	2	649
10^{-7}	8	27
10^{-7}	24	2
10^{-12}	2	130
10^{-12}	8	12
10^{-12}	24	2

Using Minitab and the Box-Cox procedure to automatically choose the optimal the transformation $y^* = y^{-0.0983497}$ was used. The results are shown in Table 2.

Table A2-2: Transforming data, where pv = p-values, fc= fold change and Y^* is the response

pv	fc	y*
-1	-1	0.510349
-1	0	0.718081
-1	1	0.934101
0	-1	0.528952
0	0	0.723146
0	1	0.934101
1	-1	0.619576
1	0	0.783182
1	1	0.934101

Figure A2-1 shows the results for the Analysis of Variance (ANOVA) using Minitab 16.

General Regression Analysis: y* versus pv, fc

```
Regression Equation

y*  =  0.742843 + 0.0290547 pv + 0.190571 fc - 0.0273068 pv*fc


Coefficients

Term             Coef    SE Coef        T      P
Constant     0.742843  0.0056041  132.552  0.000
pv           0.029055  0.0068637    4.233  0.008
fc           0.190571  0.0068637   27.765  0.000
pv*fc       -0.027307  0.0084062   -3.248  0.023


Summary of Model

S = 0.0168124       R-Sq = 99.38%         R-Sq(adj) = 99.01%
PRESS = 0.00540478  R-Sq(pred) = 97.62%


Analysis of Variance

Source       DF    Seq SS    Adj SS    Adj MS        F          P
Regression    3  0.225952  0.225952  0.075317  266.460  0.0000062
  pv          1  0.005065  0.005065  0.005065   17.919  0.0082236
  fc          1  0.217904  0.217904  0.217904  770.909  0.0000011
  pv*fc       1  0.002983  0.002983  0.002983   10.552  0.0227363
Error         5  0.001413  0.001413  0.000283
Total         8  0.227365
```

Figure A2-1. General Regression Analysis on the number of genes deemed important through a volcano plot.

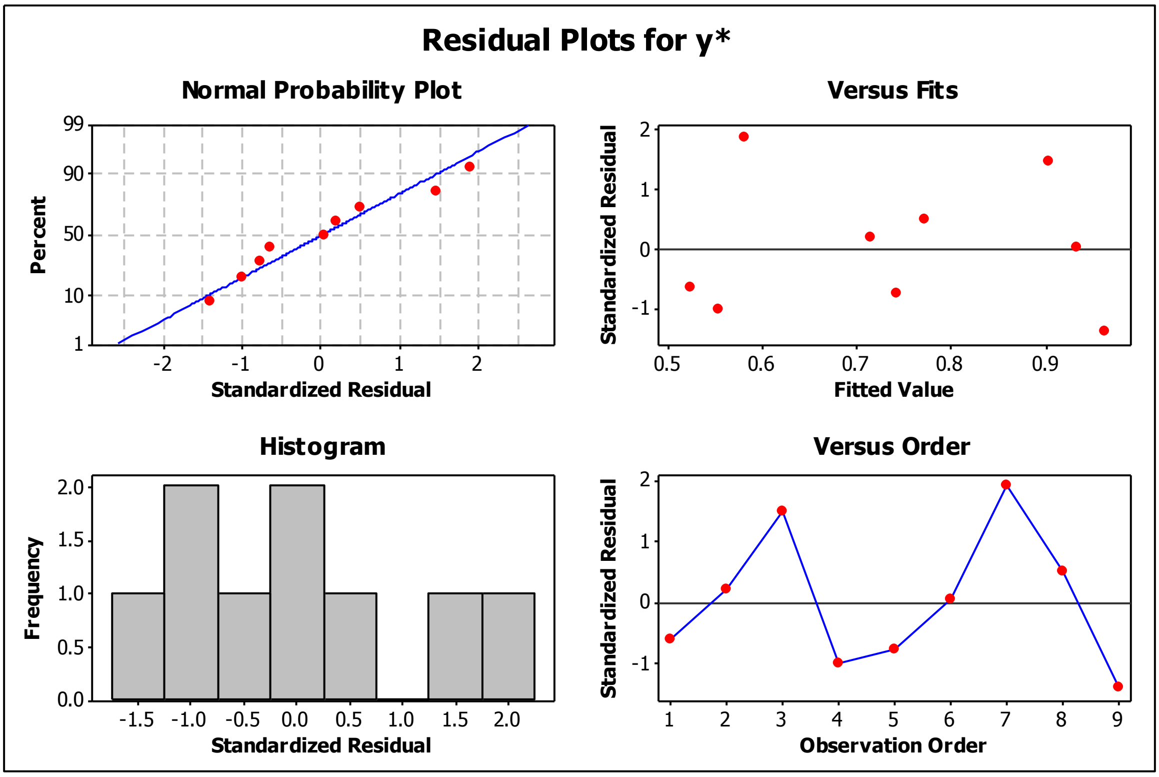


Figure A2-2: Residual plots.

In order to verify the model adequacy the following assumptions were evaluated based on the model residuals: normality, independence and equal of variances. These are graphically shown in Figure A2-2.

The Kolmogorov Smirnov normality test was performed using Minitab. Yielding a p-value of 0.150, so the assumption of normality in the residuals cannot be rejected.

For independence, the “runs test” was utilized, obtaining a p-value of 0.748. Thus, the assumption of independence in the residuals seems to be in check.

The significance value for this analysis was chosen as 0.05. All terms used in this model are statistically significant. In addition, the model explains 99.38% (R-Sq) of the variability. These results evidence how choosing different thresholds for p-value and fold change affect the selection of important genes in microarrays, independently and when set jointly.