IDENTIFICATION OF POTENTIAL CANCER BIOMARKERS THROUGH MULTIPLE CRITERIA OPTIMIZATION USING MICROARRAY DATA

by

Matilde Luz Sánchez-Peña

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

MASTER OF SCIENCE
in
INDUSTRIAL ENGINEERING

UNIVERSITY OF PUERTO RICO
MAYAGÜEZ CAMPUS
2010

Approved by:

________________________________ __________________
José M. Castro, PhD Date
Member, Graduate Committee

________________________________ __________________
Alexandra Medina-Borja, PhD Date
Member, Graduate Committee

________________________________ __________________
Clara E. Isaza Brando, PhD Date
Member, Graduate Committee

________________________________ __________________
Mauricio Cabrera-Ríos, PhD Date
President, Graduate Committee

________________________________ __________________
Anand D. Sharma, PhD Date
Representative of Graduate Studies

________________________________ __________________
Agustin Rullán Toro, PhD Date
Chairperson of the Department

ABSTRACT

Cancer is a worldwide relevant illness given its mortality rates and associated economic and social repercussions. Genetic profiling has become one of the most important tools for cancer characterization, its diagnosis and prognosis. Microarrays are biological experiments that have been used in recent years with this end in mind due to their capacity to measure the relative genetic expression of tens of thousands of genes simultaneously. One of the principal aims using data from microarray experiments is the selection of relevant genes that can be used as surrogate measures for the state of cancer, i.e. cancer biomarker genes. Many and varied methodologies have been developed and used for this purpose ranging from the simplest statistical approaches to sophisticated Artificial Intelligence methods. The explored literature, however, shows that setting parameters for several of these approaches is often a difficult task for final users, who mainly hail from the biological and medical sciences. As a consequence, analysis results have been reported to vary across different researchers even when using the same microarray datasets. This situation is an opportunity to develop methodologies to find potential cancer biomarkers in a consistent manner.

In this work potential biomarker identification is casted as a Multiple Criteria Optimization (MCO) Problem, aiming to remove analysis subjectivity due to parameter adjustment. MCO is a methodology used to find the best compromises between two or more conflicting criteria.

The main proposition of this work is that several measures related to microarray data analysis can be seen as criteria to be optimized. It is desirable, for example, that the p-value associated to a particular gene be low when trying to determine its statistical significance. If a gene could be characterized through two or more p-values, then an MCO problem can be formulated. Solving an MCO problem results in a set of solutions representing the best compromises among all the considered criteria. These solutions are called Pareto-efficient solutions and they conform a so-called efficient frontier of the problem. This work proposes that genes on the resulting efficient frontier of an associated MCO problem could be cancer biomarkers. Among the methodologies used to solve MCO problems, Data Envelopment Analysis (DEA) has been chosen in this work because it does not require parameter setting by the user in many of its possible formulations. Furthermore, DEA can be solved through linear programming, the most tractable of optimization problems and for which inexpensive commercial software readily available. To the best extent of our knowledge, this work constitutes the first effort on using Multiple Criteria Optimization to detect potential cancer biomarkers from microarray data.

RESUMEN

El cáncer es una enfermedad importante a nivel mundial dado su nivel de mortandad y sus repercusiones sociales y económicas. Los perfiles genéticos se han convertido en una de las herramientas más importantes para la caracterización del cáncer. Los microarreglos son experimentos biológicos que se han venido usando en los últimos años para obtener dichos perfiles, dada su capacidad de medir la expresión relativa de decenas de miles de genes de manera simultánea. Una de las principales tareas al trabajar con datos que provienen de los experimentos de microarreglos, es la selección de genes relevantes que puedan ser utilizados como detectores de la presencia de la enfermedad, en otras palabras, genes biomarcadores de cáncer. Muchas y muy variadas metodologías se han desarrollado con estos propósitos, las cuales van desde los procedimientos estadísticos más simples hasta los métodos más sofisticados de Inteligencia Artificial. Sin embargo, en la literatura explorada se muestra que, en muchos de los enfoques utilizados, la definición de ciertos parámetros resulta ser una tarea difícil para los usuarios finales, los cuales provienen principalmente de los campos de biología y medicina. A consecuencia de ésto, los análisis reportados varían entre los diferentes investigadores aun cuando se utilicen los mismos datos. Esta situación es una oportunidad para desarrollar metodologías para encontrar biomarcadores potenciales de cáncer en una forma consistente.

En este trabajo la identificación de biomarcadores potenciales es tratada como un Problema de Optimización de Múltiples Criterios (MCO por sus siglas en inglés), el cual permite extraer la subjetividad que se da por el ajuste de parámetros por los usuarios. La propuesta principal de este trabajo es que muchas medidas relacionadas con el análisis de microarreglos pueden ser vistas como criterios a ser optimizados. Es deseable, por ejemplo, que el valor-p asociado a un gen en particular sea menor cuando se trata de determinar su significancia estadística. Si un gen puede ser caracterizado por medio de dos o mas valores-p, entonces es factible formular un problema de MCO. La solución de un problema de MCO resulta en un conjunto de soluciones llamadas Pareto-eficientes que conforman la frontera eficiente de tal problema. Este trabajo propone que los genes que resulten en la frontera eficiente del problema de optimización de múltiples criterios asociado pueden ser biomarcadores de cáncer. Entre las metodologías existentes para resolver problemas MCO, el Análisis Envolvente de Datos (DEA por sus siglas en inglés) se ha elegido para ser utilizado en este trabajo dado que no requiere el ajuste de parámetros por el usuario en muchas de sus posibles formulaciones. Además, DEA puede ser resuelto por medio de programación lineal, que es el problema más tratable de optimización y para el cual existe una amplia variedad de paquetes computacionales disponibles. De acuerdo con la búsqueda de literatura, llevada a cabo en esta tesis, esta constituye el primer esfuerzo en usar Optimización de Múltiples Criterios para detectar biomarcadores potenciales de cáncer a partir de datos de microarreglos.

To my family.

ACKNOWLEDGEMENTS

I am deeply thankful to my advisor Dr. Mauricio Cabrera-Ríos. I am grateful for the opportunity to work with you during all these years, I will never regret knocking on your door as an undergraduate student looking for a project. I hope there will be a lot of new endeavors in which we can work together. Thank you, especially for inviting me to join your group in Puerto Rico, I would have never imagined being here, but it happened, and you know how much it means for my professional and personal life.

I am also grateful to Dr. Clara Isaza, my co-advisor. This work would not have been possible without your advice in the biological aspects of our research. This was a true interdisciplinary effort that showed me that well-structured ideas can reach powerful results to solve relevant problems like cancer.

I thank Dr. José Castro, for his support as a member of my graduate committee, and for being a role model on how to conduct research and manage a research group. Thank to Dr. Alexandra Medina-Borja for her valuable observations to improve this work. I also thank Marina Robles for all her help on understanding biological terms and her collaboration on the validation stage of this work. I also very much enjoyed our time together as friends. Thanks to MS Lyzett Uribe and MS Hugo Perez whose work were the first efforts in this research line in our group. Thanks to all the other members of the Applied Optimization Group, graduate and undergraduate students, it has been my pleasure to work with you.

Thanks to Dr. Jonas Almeida for the opportunity to explore other research topics in Bioinformatics during my internship at the MD Anderson Cancer Center last summer. Thanks to Dr. Leticia González for being part of the beginning of this dream and always keeping an eye on my progress. Thanks to Dr. Sheila Maestre who contributed to my overall development during these years.

My sincere thanks to the faculty and staff at the Industrial Engineering Department (ININ), especially to Mayra Colon, the best secretary ever. Thanks to all my graduate students friends at ININ. A special recognition goes to Jesus Rodríguez, Orlando Mézquita and Andy Brunot, who were always willing to help me with technical problems, as well as being great friends of mine. I also thank those who were here when starting the road: Angela Anaya, Ernesto Aponte and Cynthia Rodríguez.

I am deeply grateful to my housemate Catalina Obando, who one day she promised me to be like my sister. I am glad she really did. Special thank you notes go to Reimond Rodríguez, Christopher Soto, Oscar Herrera and Alejandro Maurosa, whom always got my back; it is my pleasure and my honor to be their friend. I also want to thank other important persons in this story: Carlos Calero and Mariana Torres, Joymariel Melecio, Sixto Fernandez, Yareni Lara, Berenice Rodríguez, Esmeralda Niño, Mauricio León, Juan C. Obando, Fernando Mérida, Omar Lazo and Dennis Rosario. All of them have a special place in my heart and have taught me that the world is small when trying to find good friends.

I am thankful to Edna Aponte and families: Soto Cruz, Davila Torres, Iglesias Sepúlveda and Rosa Urbistondo. All of them opened the doors of their homes to me and made me belong, I will never forget Them.

For the same reasons, I would like to thank to all the members of the Cabrera-Isaza family, I am really glad to consider you as my adoptive family in Puerto Rico. Mauricio and Clara are special role models to me in all the aspects. Anita and Emilio have made me keep alive my tenderness. I love them as my own nephews. I want to thank the special support from my aunt Matilde Sánchez Luna, my brother in law Juan de Dios Castro and my friends Lyzett Uribe, Victor Ramirez, Juan A. Guerrero, Gustavo Alvarez, Nicolas Neri, Eduardo González, Miguel Arias and Jorge Arevalo. Even at a distance, theywere always mindful of my well-being.

Thanks to all the members of my lovely families, Serrato Sánchez, Castro Sánchez (my Simpson), Sánchez Montiel, Sánchez de León and Sánchez Coronado, who were always willing to help me when I needed, and shared each one of my troubles and joys. Thank to all my nephews, who always inspired me with their innocence.

Thanks to my brothers and sisters: Alejandra with her professionalism, Ramiro as the most devoted worker I know, Rocío with her tireless sense of humor, José Luis with his passion to reach goals and Victor with his love to take risks. I am a mixture of all of you, and thanks to that, now I am achieving this goal.

And last but not least, thanks to the most important persons in my world, my parents Ramiro and Mary, for giving me all your support during these years. Thank you for letting me learn by myself, with all its implications. Thank you for being as you are, but specially, for loving me as I am.

I truly believe that there is a biggest force in the universe that some call God, others luck, others destiny. But, whatever it is; I am deeply thankful to it for always making me be in the perfect place at the right moment, I would not change anything, I have enjoyed the story of my life to this point, and this was a truly wonderful chapter.

TABLE OF CONTENTS

Table List

Tables **Page**

Figure List

Figures **Page**

1 INTRODUCTION

1.1 Motivation

There is no doubt that cancer is a worldwide human health challenge owing to its associated mortality rates and its economic and social repercussions in society at large (1,2). Basic understanding, diagnosis, prognosis and treatment of cancer are the focuses of extensive research endeavors across multiple disciplines everywhere (3).

One of the characterization tools that have gained relevance in recent decades is that of measuring gene expression through microarrays (4-6). Microarrays are, indeed, capable to provide gene expression readings for tens of thousands of genes in a simultaneous manner, thus resulting in databases of considerable size. These large amounts of information require subsequent analyses to derive biological insight and usable medical knowledge (7-11).

Microarrays have been used for diverse purposes in cancer research, spanning from gene identification through cancer prognosis. When it comes to gene identification, biomarkers take a particularly important place. Biomarker genes are those that characterize a particular biological state such as an illness or a specific illness stage (12,13).

Although much effort has been dedicated to the analysis of microarray data to detect biomarkers, this process still faces several challenges in at least two aspects: (i) researcher-dependency of the results, and (ii) transparency of the methods used to elicit the result for the final users. One common thread in (i) and (ii) is the use of analysis methods that require the adjustment of several parameters of computational, statistical or mathematical nature that do not necessarily have a related biological or medical meaning. This implies that, often times, the final users –from Biology and Medicine- are left with black-boxes with parameters whose adjustment will significantly affect the final results of the analysis, and whose successful tuning requires understanding of other fields.

In this work, the identification of potential cancer biomarker genes is formulated as a multiple criteria optimization problem based on microarray data. The resulting problem is solved through the application of Data Envelopment Analysis. The proposed method does not require any parameter adjustment from the user, and thereby preserves the objectivity and reproducibility of the analyses. This is, indeed, the first time that gene identification is approached through multiple criteria optimization techniques based on microarray data. Furthermore, the lists of potential cancer biomarkers elicited through the methods articulated in this thesis will constitute an important contribution from the industrial engineering field to cancer research.

1.2 Objective

The objective of this thesis is to cast the gene selection problem in microarray analysis as a multiple criteria optimization problem, aiming to identify potential cancer biomarker genes. As a first approach, the resulting multiple criteria optimization problem will be solved through the application of Data Envelopment Analysis. Validation of the potential biomarker gene sets will be carried out through search on the genes' proposed role in the existing literature.

1.3 Work Organization

In Chapter 2 the background of this work is presented first with an introduction to the biological terminology, followed by a Literature Review to finally describe the tools to be used in this work in the context of the problem at hand: Multiple Criteria Optimization and Data Envelopment Analysis.

Chapter 3 presents the three different experimental structures that were used to explore the capabilities of the proposed method. The results are presented and technically evaluated. In Chapter 4 the analyses of the previous results and their validation processes are presented. One validation scheme is based in previous works, another one is supported by data, and the last one is based on the search for specific gene roles. Conclusions and future work are presented in Chapter 5.

2 BACKGROUND

2.1 Biological Background

2.1.1 Cancer

Every human cell has a defined life cycle known as the “cellular cycle”. This cycle is composed by 3 phases: (1) the origin of the cell through the mitosis process, (2) cell growth, which is achieved to perform its intended functions; and (3) cellular death in an auto programmed manner called apoptosis. This process is continuously executed during the life of every tissue renewing its functional cells. Unfortunately some internal and external factors in the organism can result in the alteration of the apoptosis process, thus resulting in cells that do not die. The accumulation of such cell forms interferes with normal tissue functions and is capable to invade other tissues.

Cancer, then, is the term used to call the uncontrolled growth of abnormal cells. When other tissues are invaded, the phenomenon is called metastasis. Metastasis is, indeed, the principal reason for cancer related deaths. The birth zone of the cancerous cells give name to the cancer type, resulting in the existence of over 100 different cancer types, i.e. stomach cancer, prostate cancer, and so on.

Cancer can be caused by genetic mutations, hormonal abnormalities, immune conditions or metabolism mutations as well as external factors such as the exposure to radiation, chemicals, infectious organisms or tobacco (1).

Different treatments can be applied to attempt cancer eradication. Surgeries, chemotherapy, radiotherapy, hormonal treatment, are among the most common ones. Given the metastasis phenomenon, it is of paramount importance that diagnosis happens in the early stages of the illness to improve the survival chances of patients. It is estimated that one third of cancers could be cured if detected early and treated adequately (2).

World Health Organization statistics shows that cancer is a leading cause of death worldwide with 7.9 million fatalities in 2007. These represented the 13% of the total deaths that year. Worldwide, the five most common types of cancer that cause death in males are, in order of importance, lung, stomach, liver, colorectal and esophageal cancer; while breast, lung, stomach, colorectal and cervical cancer are the leading ones in females.

According to the American Society of Cancer, at least 1,529,560 new cancer cases are expected to be diagnosed in 2010 in USA. About 569,490 Americans are expected to die of cancer (more than 1,500 people a day), making it the second most common cause of death behind heart diseases (1).

Even though the genetic information is not the only factor to develop cancer, its contribution is large when combined with other internal or external factors. For this reason genetic profiles for different kinds of cancer have been extensively studied through different methodologies, including microarrays (5,14-19).

Cancer diagnosis has typically been made through morphologic characterization of a tumor sample; an increasing interest to support diagnosis with genetic profiling is evident. In addition to diagnosis, cancer prognosis, which refers to the determination of the cancer stage and its most likely course of development (3,20-22), can also be carried out using genetic information.

If the expression of a reduced number of genes is recognized to be characteristic of certain kind of cancer, pharmaceutical efforts can be focused in the expression stimulus or suppression of these genes for cancer treatment. Genetic characterization can be also useful to develop tailored treatments for each patient, avoiding unnecessary exposition to chemo or radiotherapy (13). Microarray experiments aided with a variety of data processing techniques have been used to identify these important genes, usually called biomarkers. The next section, presents a brief description of biomarkers.

2.1.2 Biomarker genes

A Biological marker, or biomarker, is defined by the Biomarker Definitions Working Group (BDWG) in (12) as “a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes or pharmacologic responses to a therapeutic intervention”.

Applications of biomarkers that the BDWG has defined for disease detection and monitoring of health status include:

- Identification of patients with a disease or an abnormal condition,
- Determination of development stage of the disease,
- Determination of the disease prognosis,
- Prediction and monitoring of the clinical response of an intervention.

Some biomarkers, known as surrogate endpoints, are intended to be proxies for characteristics that reflect how a patient feels, functions or survives (12). When this kind of biomarkers is found, their reliable validation is needed. Clinical trials are required by the Federal Drug Administration (FDA) to start the use of biomarkers proposed by the research community. Thus, the validation of a biomarker is a lengthy process that spans for years.

In the specific case of cancer, cancer biomarkers have been discovered and utilized with specific purposes such as: a) early cancer detection, b) cancer diagnosis, c) cancer prognosis, d) prediction of patient response to cancer eradication therapies, and e) prediction of cancer recurrence (23).

It is expected that groups of biomarkers be analyzed jointly rather than individually to improve sensitivity of cancer diagnosis and prognosis. It is also expected that particular groups of biomarkers exist for each type of cancer. Information provided by biomarkers can also potentially aid the development of tailored treatments, perhaps without excessive exposure to chemo or radiotherapies (23).

In some cases, a particular biomarker plays a role in more than one type of cancer. For example, the mutation of BRCA1 and BRCA2 genes has been reported as a risk enhancer of Breast cancer. The mutations of these genes however, are also biomarkers of ovarian, prostate, colon and pancreatic cancer among others (23). A short list of genes reported as hereditary biomarkers genes is shown in Table 2-1.

Table 2-1. Some known genetic mutations related to a variety of cancer types. Source (23).

Gene Mutation	Related Cancer Phenotypes
BRCA1	Breast-Female, ovarian, fallopian tube, primary peritoneal and prostate cancer
BRCA2	Breast-Female, breast-male, ovarian, fallopian tube, primary peritoneal, pancreas and peritoneal cancer
APC	Colon/rectum, hepatoblastoma, brain (medulloblastoma), pancreas, small bowel, gastric, thyroid (non-medullary)
NF1	Malignant peripheral nerve sheath rumor, astocitoma, pheochromocytoma, meuroblastoma, ependymoma, rhabdomyosarcoma, glioma
MEN1	Pancreas-islet cell, amine precursor uptake and decaroxylation tumors, adrenal cortical carcinoma, carcinoid
RET	Pheochromocytoma, thyroid (medullary)

2.1.3 Microarrays

A microarray is a biological experiment where the expression of tens of thousands of genes can be measured simultaneously. Typically, a physical platform with probes capable to detect particular biological entities is involved. For genes, each probe contains a known sequence. The expression of each known sequence is measured reading the fluorescence of a deposited sample of interest. The sample is the extracted RNA from a tissue that is prepared for these purposes.

There are different kinds of microarray experiments, each of them with different preprocessing, experimental execution, image processing and data acquisition requirements and characteristics. A description of several variants of microarray experiments is presented in (6). In general, there are two kinds of microarray experiments, those known as cDNA microarrays (from complementary DNA) and oligonucleotide microarrays (5,16,17,19,24). A description of the principal characteristics, advantages and disadvantages of each type is presented next.

2.1.3.1 cDNA Microarrays

Several libraries have been built in order to use knowledge from genetic sequences already discovered. These libraries facilitate experimentation with those sequences that are kept in state of complimentary DNA (cDNA). This is the best known preservation method due to the degradation tendencies of the genetic material. To design a cDNA microarray, a defined number of known sequences are selected and requested from the existing libraries. These sequences are amplified through a process called Polymer Chain Reaction (PCR) to increase their quantity and build the array (17). This reaction is performed in a well plate and, finally, the amplified sequences are spotted through an automated process in a small glass surface in form of a dense array ordered as a grid where each place contains a known gene sequence. Figure 2-1 illustrates this process. Nowadays, this process is technically easy to replicate (18), for this reason microarray costs have decreased in later years. Although still expensive, cDNA microarrays' cost and potential to be customized are advantages when compared to oligonucleotide microarrays (5,16,17). Some limitations of cDNA microarrays include the degradation of the original DNA material from libraries during the different PCR reactions, difficulties for physical handling, and the possibility of cross contamination between sequences.

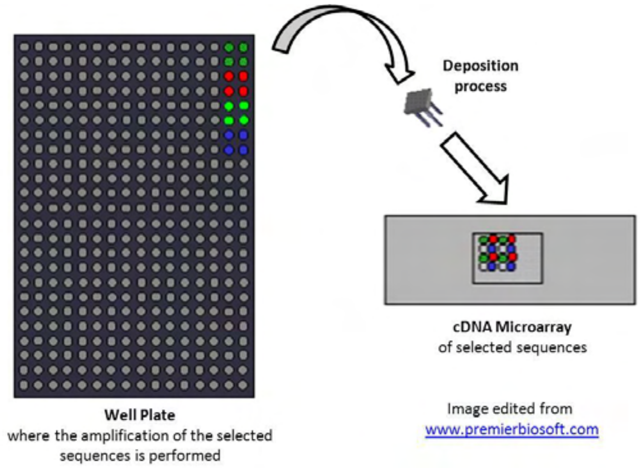


Figure 2-1. General structure of the deposition procedure in microarray manufacture, this methodology is used for cDNA and long oligonulcleotide microarrays.

2.1.3.2 Oligonucleotide Microarrays

In oligonucleotide microarrays, gene sequences are generated from the deposition of their basic components i.e. nitrogenous bases: Adenine, Cytosine, Guanine and Thymine. Readers interested in biology foundations are referred to [6]. Gene sequences can be represented by their original long sequence of nucleotide acids or by dividing them in shorter pieces. In the latter case, a single gene sequence is divided in segments for its deposition. Bases corresponding to those segments are deposited in the surface in order; when the experiment is performed, those segments are read together to interpret the expression of the gene involved.

There are at least four methods to develop oligonucleotide microarrays. The first method is photolithography, used by the producer Affymetrix (Figure 2-2). This method deposits the bases in different places. It does so depending on the area that is exposed to light in the supporting glass surface of the microarray upon a predefined order (6).

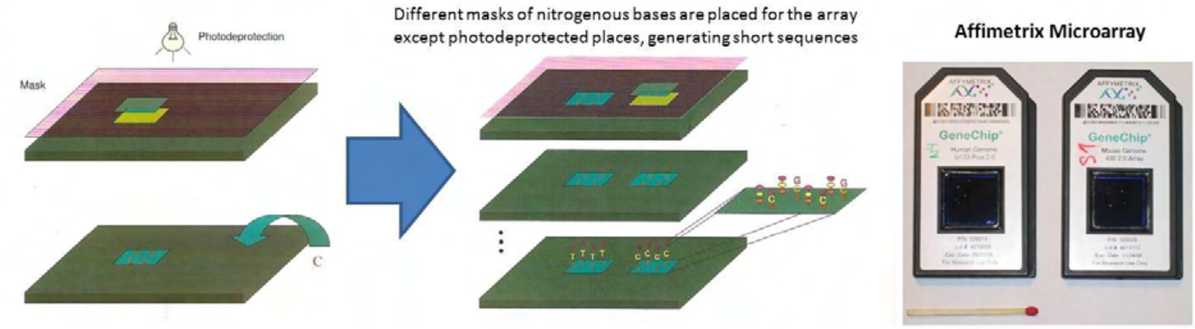
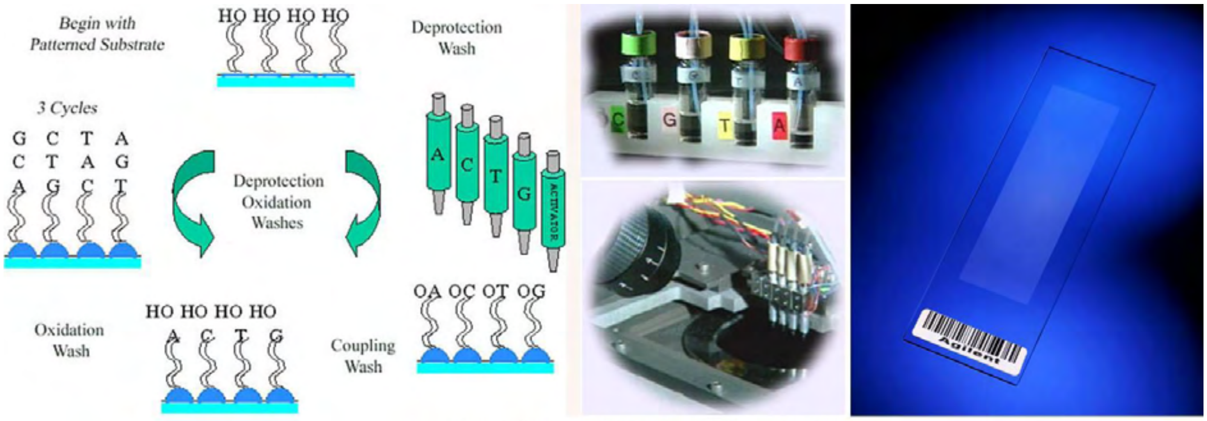

Photodeprotection
Mask
C
Different masks of nitrogenous bases are placed for the array
except photodeprotected places, generating short sequences
T T T T
C C C C
G
C
Affimetrix Microarray
GeneChip
GeneChip

Image edited from [6] and www.bioinfowordpress.com

Figure 2-2. Photolithographic process in microarray manufacturing, principal producer (Affymetrix).

The second method is the ink jet technology and is used by Agilent (Figure 2-3), Protogene, among others. It is similar to an ink jet printing deposition where the different bases are contained in cartridges and their distribution is spotted in a predefined order also (6).



Images sources: www.microfab.com, http://liebel-lab.fzk.de

Figure 2-3. Ink Jet printing process in microarray manufacturing, principal producer (Agilent).

The third method is the electrochemical synthesis (Figure 2-4) used by the producer CombiMatrix. In this method the substrate contains electrodes embedded to manage the deposition of different bases contained in solutions and washed in every step in the different individual reaction sites (6).

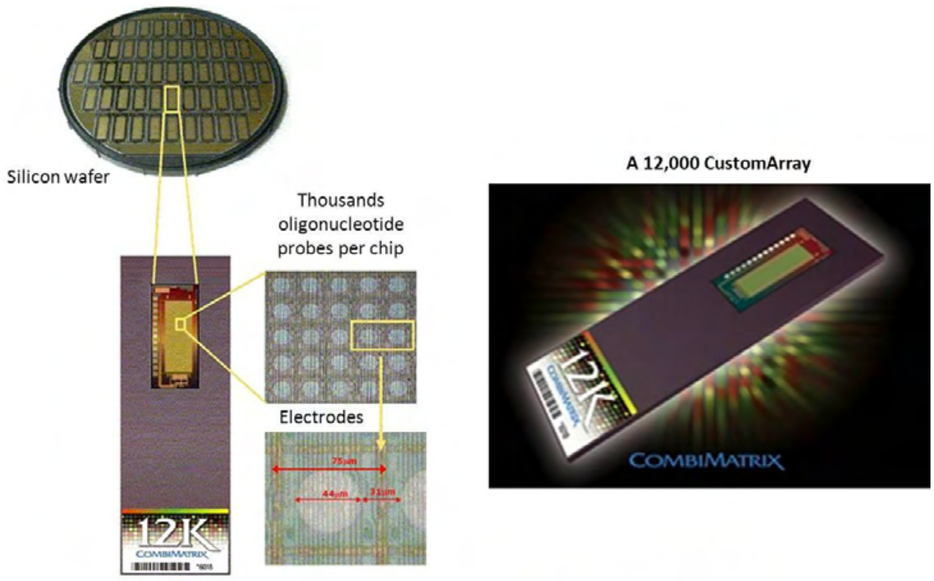


Image sources: www.combimatrix.com and www.bf-biolabs.com

Figure 2-4. Electrochemical synthesis process in microarray manufacturing, principal producer (Combimatrix).

These three methods are known as in situ synthesis methods, due to their characteristic requirement to place the known material directly in its purest state (nitrogenous bases). These methods eliminate the noise from PCR reactions and the potential cross contamination for sequence deposition that is found in cDNA microarrays. Finally, another process for oligonucleotide arrays is the previous preparation of the specific nucleotide long sequences base by base and their deposition as in the cDNA arrays. Here the deposition method and the knowledge about the generated sequence are essential (6). Figure 2-1 helps to illustrate this process as well.

2.1.3.3 Microarrays Execution

For the execution of any microarray experiment, the RNA should be extracted and amplified from a tissue of interest through a series of biological processes. The readers interested on the details of the different biological processes across the different platforms is referred to [6].

For cDNA microarrays' execution typically two kinds of tissues are prepared: control and treatment. The former refers to known samples that act as references to measure the relative changes of genetic expression on the treatment tissues. The latter could be tissues on any state of interest, including healthy, illness states or drug treatment among others. In the cDNA array case, both, control and treatment tissues should be analyzed at the same time in each platform (Figure 2-5), whereas in the Affymetrix case, one tissue is analyzed in each run without the need of using a control tissue (17).

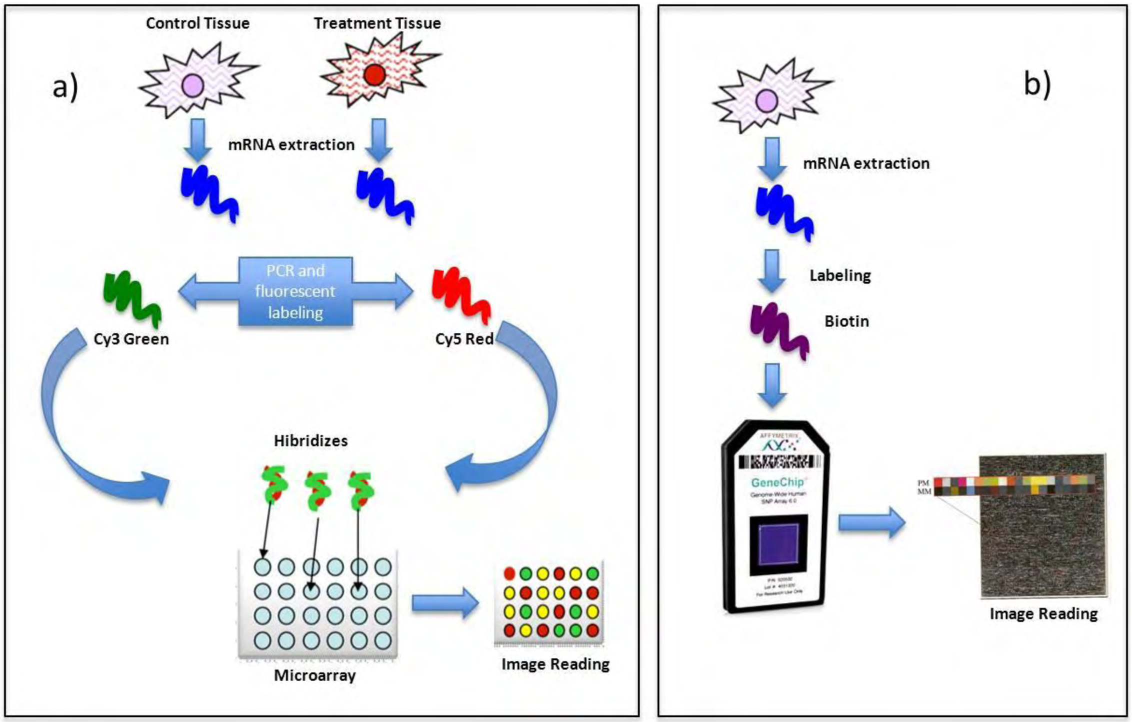
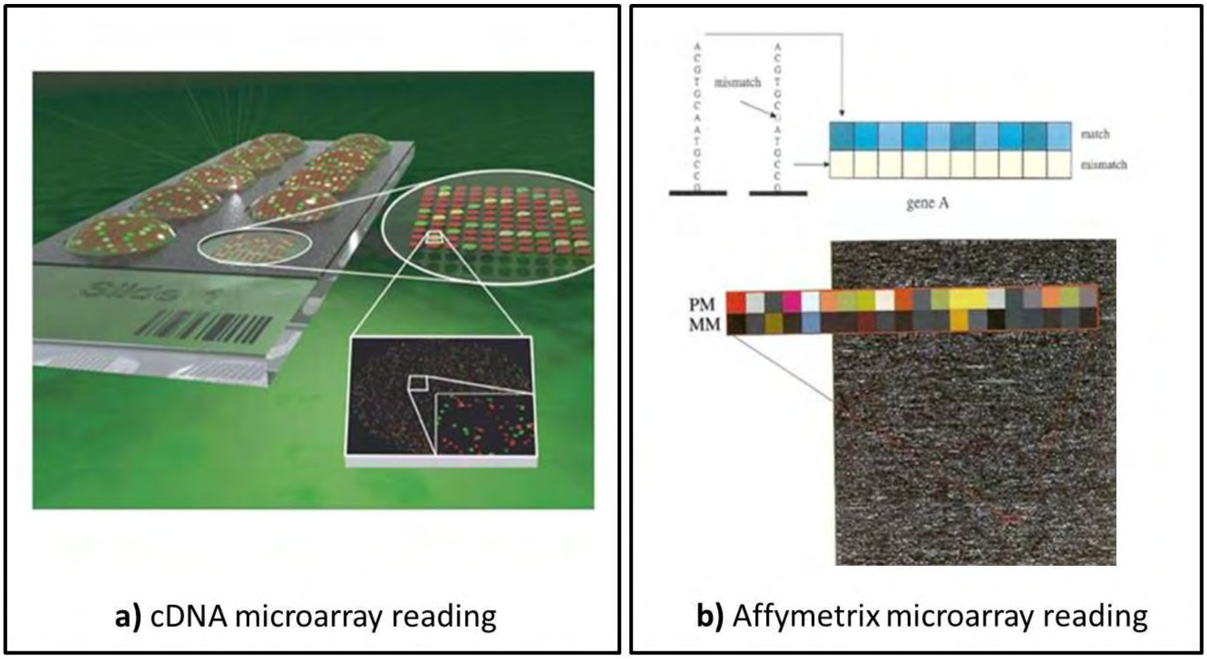


Figure 2-5. Execution diagram of a Microarray experiment, a) cDNA microarray execution, b) Oligonucleotide microarray execution.

The RNA samples should be labeled with fluorescent dyes for each state. For the cDNA arrays the labeling is typically done using Cy3 (green) for the control tissues and Cy5 (red) for the treatment tissues. In the case of the Affymetrix microarray, the samples should be labeled with Biotin dye. Once labeled, the samples are deposited in the cDNA or the Affymetrix probe (Figure 2-5).



a) cDNA microarray reading

b) Affymetrix microarray reading

Figure 2-6. General scheme for the reading of gene expression for the cDNA microarrays (a) and the Affymetrix microarrays (b)

Once the experiment is hybridized, the gene expression level is measured through the reading of the excitation of each utilized dye using a laser beam in a specialized equipment.

2.1.3.4 Interpretation of Microarrays

While each spot represents one gene in the cDNA arrays (Figure 2-6 (a)), the unit of study for each gene in the affymetrix platform is composed of two series of 20 defined sequences of oligonucleotides arranged together. One series is known as Perfect Match (PM) that contains the correct sequence of the gene, the second series, known as Mismatch (MM) contains the same sequence but with an intentionally wrong base placed in a known position (5,6). Affymetrix reading is finally obtained through the differences between PM and MM, Figure 2-6 (b).

The reading for cDNA microarrays is made through the different channels related to the control and treatment samples (green and red, as conventionally used). After processing the intensity of each channel is transformed usually through the use of $\log_2$ of the ratio of intensities ($\log_2 R/G$).

There is no consensus about what kind of platform is better for a reliable quantification of gene expression. Depending on the goals of the analysis, different researchers argue about the adequacy of one over the other. The MicroArray Quality Control consortium (MAQC) presented an effort in (25) to compare the platforms from different producers. Among other limitations, they found that, a direct comparison of expression values generated on different microarray platforms cannot be done, given their unique labeling methods. Another opportunity of improvement in the microarray technology development are the methods of mRNA extraction and the image acquisition process, since they have been recognized as sources of experimental noise (6,17).

The characteristics of each cell are given by their level of protein expressed. The level of production of these proteins are defined by the RNA. This translation however, is not perfect. Microarrays measure RNA expression, but not protein level; which could be a more informative clue in biological terms. New methodologies such as Reverse Phase Protein Arrays (RPPA) (26), have been developed in the most recent years to resolve this issue.

It is also important to note that microarray experiments are, to date, considered costly. The cost of one microarray run is estimated to be in the order of thousands of dollars (27). A large investment then has gone into creating the microarray databases available in public repositories. It is, therefore important to develop methodologies to get useful insight for cancer and other illnesses from the secondary analysis of microarray data. The work presented in this thesis moves along this line of development.

2.2 Literature Review

Since their first appearance in 1995 (4), microarray experiments have been used for many purposes due to their capability to quantify the gene expression for tens of thousands of genes in a simultaneous manner. Many approaches, as explained later, have been used to address the extraction of relevant biological and medical knowledge from the resulting large databases.

One attempt at knowledge extraction consists on determining which of the thousands of genes change their expression levels from one state to another, for example, from a state of health to a state of cancer. This problem has been referred to as gene selection or gene filtering (24,28-30). Gene filtering has been extensively explored given its potential to recognize a reduced number of genes that can provide a shortcut to diagnosis or prognosis for a particular illness. This process, with a higher focus on relevance and compactness for the group of selected genes, can also be used to detect potential biomarker genes. Identification of a smaller set of genes can offer savings in research resources to elicit useful advances in the race against cancer.

Gene filtering has been explored through a wide variety of techniques. The simplest of these is the Fold Change technique (31), which measures the number of times that the expression level of a gene in a particular state doubles the expression level for such gene in a different state of interest. Additionally, normality-based statistics approaches, like the 2 sample t-test

(7), ANOVA (8), Welch t-test (32) have also been extensively applied with gene filtering in mind. Some authors have stressed the fact that gene expression level does not follow a normal distribution (33-35), proposing the use of non-parametric statistics like the Wilcoxon Mann Whitney test (35). Some examples of these early gene filtering procedures can be found in (7,36-38).

One of the most utilized tools nowadays for gene filtering is the Significance Analysis of Microarrays (SAM) (33). SAM is a non-parametric approach that calculates a statistic, d, based on the gene expression means to standard deviation ratio for a particular gene. This approach is similar to the t-statistic, except that the distribution to be compared against is generated through a series of random permutations. This tool was first proposed by Tusher et al. in (33), nowadays it follows the methodology presented by Efron and Tibshirani in (39) in combination with the Gene Set Enrichment Analysis (GSEA) proposed by Subramanian et al (40). SAM is completely coded and available as an Excel add-in; despite this, the parameters to be set for analysis execution and final interpretation are not completely transparent for final users, usually arising from natural science fields. This fact leads to having different results that depend on the analyst. Such subjectivity denotes the opportunity to develop methodologies that are free of parameter adjustment and that converge to reproducible results across different analysts.

One of the first efforts to determine a genetic profile to characterize tumors and uncover new tumor categories was carried out by Golub et al. in 1999 (36). Previous to this study, cancer diagnosis was done just following the morphologic characteristics of biopsy cells. Golub et al. focused on supporting this first diagnostic with genetic information derived from microarrays, allowing a better distinction between cancer subtypes than the one previously achieved solely with morphological characterization. Golub et al. proposed the use of a so called neighborhood analysis, which generates an idealized expression pattern corresponding to a gene that is uniformly high in one class and uniformly low in the other. Gene expression is represented by vectors, one for each gene. Testing consists on comparing if there is an unusually high density of genes “nearby” that are similar to the defined idealized patterns i.e. if correlation exists. Finally, results are compared to randomly generated idealized expression patterns. If there is a high density of genes, this indicates that many more genes than those attributable to chance are correlated to the pattern. This approach became an important point of reference from which the interest to develop other methodologies for the same purposes grew significantly.

Applying their proposed methodology, Golub et al. selected a 50-gene set based in the correlation coefficient between Acute Myeloid Leukemia (AML) and Acute Lymphoblastic Leukemia (ALL) states. This set, called a signature, was validated with available information of the genes involved on cancer metabolism and also with their classification accuracy. This signature offers good levels of accuracy in tumor classification in their corresponding class

(AML or ALL), assigning 29 out of 34 independent samples correctly. Some subsequent methodologies for gene filtering and diagnosis trough gene characterization used Golub et al.´s database and results as benchmarks for their own efforts (9). The subsequent works conclude with different subsets of genes identified as significant, making results methodology-dependent. Convergence to a unique set of genes would accelerate its use in cancer research. There is, then, a need for methodologies that converge to a unique set of relevant genes specially if the same dataset is used.

In 2000, Alizadeh et al. (41) focused on finding the usually difficult characterization among new kinds of Diffuse Large B-Cell Lymphomas (DLBCL's), through genetic differentiation. DLBCL's are aggressive malignancies of mature B-lymphocytes. Their specific microarray "lymphochip" composed by more than four thousand genes was designed for this work. Through the use of hierarchical clustering, hundreds of differentially expressed genes are selected as related to the differentiation of two new distinguishable subgroups. This work also has been used extensively as a benchmark for other methodologies resulting on different sets of selected genes using its database, stressing again the relevance for to the identification of a unique gene set independently of the experimenters.

Dhanasekaran et al. (42) in 2001 presented a study on the identification of genetic biomarkers in prostate cancer. Using cDNA microarrays of more than 50 samples among normal and tumor tissues, the authors defined several associations between genes and prostate cancer; assessing two of them, hepsin and pim-1 genes over 700 cancer specimens, finding a significant correlation among their expression and the clinical outcome. In this case hepsin and pim-1 can be pointed as potential genetic biomarkers in prostate cancer. It would also be important to validate results of genetic signature using other databases from the same cancer type. This thesis structures the opportunity to draw conclusions from the concurrent analysis of different databases to converge to a gene signature for a specific cancer type.

In 2003 Wong et al. (43), set the hypothesis that the genetic profile of cervical cancer can be used to separate healthy samples from unhealthy ones. Their work used the Wilcoxon’s rank-sum test along with the Benjamini and Hochberg’s False Discovery Rate correction (44) to define a profile of about 40 genes. The authors were able to separate tumors at different cancer stages as well as their expected response to radiotherapy. This segregation of healthy and cancer tissues as well as the segregation among different cancer stages represent the main contribution of their work. As in the prostate cancer case, the set of genes found for cervix cancer should be also validated across different datasets to make their evidence stronger as potential biomarkers.

The last two cases were devoted to the identification of a signature for specific cancer types: prostate cancer and cervix cancer. Determining a unique set of biomarker genes for any cancer type would result in a helpful tool for diagnosis and prognosis of the illness. As one of its aims, this thesis proposes to find a set of potential biomarker genes for the general state of cancer from the analysis executed over different cancer types.

Several efforts have been carried out exploring the capabilities of simple filtering techniques combined with clustering methods for classification purposes with interesting results. Khan et al. in (45) developed an Artificial Neural Network (ANN) approach to classify four kinds of round blue-cell tumors (SRBCT's). Given that there was an opportunity to improve the accuracy of cancer subtype diagnosis for timely treatment, the authors used principal component analysis (PCA) as a filtering technique and then used the selected genes as the information for training a neural network classifier. This is one of the first documented uses of an Artificial Intelligence application for classification purposes based in a previous gene selection. A more extensive exploration of the classification methodologies can be found in (46). Even though Khan et al. case reached a good level of classification accuracy; the execution of their methodology includes the use of PCA and ANNs. The effective use of these tools requires an advanced level of knowledge in fields not necessarily dominated by natural scientists, the final users. This fact highlights the need for methodologies that are friendlier to the final users in terms of knowledge extraction.

Several advances regarding gene filtering are considered in the MammaPrint case (47-52). When a breast cancer patient is treated and the illness is eradicated, there is a chance to relapse and develop the illness either in that zone or another one altogether (metastasis).

van’t Veer et al. (47) started looking for a specific gene signature that allowed the differentiation among breast cancer patients with potential to relapse in the next five years. The authors used a three-step supervised classification method. The method starts with the calculation of the correlation coefficient between the expression for each gene and the disease outcome reducing the original number of genes from 25,000 to 231. Secondly the selected genes are ranked by magnitude. Finally, subsets of 5 genes from the ranked list are added to the predictive set of genes looking to optimize the number of genes in this set aiming to improve its classification quality. This procedure ended with a subset of 70 relevant genes. This subset was validated differentiating among patients with “good prognosis” (low probability to relapse) from those with a “poor prognosis” (high probability to relapse), with competitive results. The main advantage for this differentiation is to have the ability to prescribe tailored treatments without exposing the patient unnecessarily to radiation or chemotherapy, thereby improving patient life quality. All the efforts of this research group derived on patenting the MammaPrint chip that is actually approved by the Federal Drug Administration (FDA) to be produced and sold for its use as a prediction tool of breast cancer recurrence. The availability of cancer diagnosis tools based on biomarkers can be helpful for early detection of cancer, as well as the definition of its treatment. This thesis presents a novel approach for the selection of potential biomarker sets which, after careful validation, might help develop cancer diagnosis and prognosis tools.

In a parallel effort, Wang et al. (53) start with the same objective as the MammaPrint case of finding a breast cancer genetic signature and describe a study performed using Affymetrix microarrays. Using a completely different set of tools, composed by hierarchical clustering, univariate Cox's proportional-hazards regression and bootstrapping, Wang et al. found a 76-gene signature for prediction of distant tumor recurrence on breast cancer. On (54), differences between both sets of predictor genes are notorious. The reason for this discrepancy has been attributed to differences among microarray platforms, genes used in the arrays, experimental conditions as well as to pure chance. Only three genes are shared between the MammaPrint and the Affymetrix gene signature. This discrepancy among predictor gene sets is further explored in (54) where the MammaPrint data set is used to assess the reproducibility of the 70-gene signature previously obtained, concluding that there are many ways to construct a 70-gene predictor that offer similar levels of prediction. It is important to conciliate the existent gene sets on a stronger gene signature that can offer better diagnosis capabilities. In this thesis, the proposed method constitutes one possible venue for synthesizing conflicting results to construct a common gene set.

Lee et al. (55) explored several gene filtering and classification techniques when applied to seven databases that have been extensively cited in previous works (7,36,37,41,56-58). This work concludes with the suggestion of which method would result in the best level of classification for each database. Even though this work has given some direction on how to deal with the different available databases, it assumes that all the differences are solely tissue-related. However, the analysis presented by (54) associates the divergence in conclusions with the different gene sets even when using the same database. In this thesis the focus is on the robust identification of those potential cancer biomarker genes across multiple databases, which might allow for more stable results.

In summary, the opportunity areas identified on the reviewed literature are two. First the used statistical techniques require formal and sometimes advanced training to adequately make conclusions and extract knowledge. It must be recognized that many final users in Medicine and Biology might lack sufficient training in these disciplines. Regarding more complex methodologies as the Artificial Intelligence techniques, these have to be designed and programmed requiring previous data information, or expert decisions (45,59) offering the same limitations for final users. Second, existing methodologies often lack reproducibility because of the required adjustment of parameters by the users. These parameters tend to not have a medical, biological or sometimes statistical meaning.

There is an opportunity to develop tools that require little or none parameter adjustment that result in consistent analyses for non-statisticians and natural sciences specialists, who are the final users and who are capable to make better sense of gene selection. This is especially true in the identification of potential cancer biomarkers, their validation and their subsequent application for diagnosis and prognosis. A methodology that is reliable, objective and consistent is envisioned to this end.

The approach presented in this thesis addresses the identification of potential cancer biomarker genes through the use of Multiple Criteria Optimization.

Multiple Criteria Optimization (MCO) is used to identify the best compromising solutions when considering two or more conflicting criteria.

Some authors have referred a theoretical relation of Data Envelopment Analysis (DEA) as useful tool to solve MCO problems in [74,79]. MCO has been used to find the best compromises between two or more criteria previously in the manufacturing field (60-63). DEA, a technique in which our research group has extensive experience, is among the methods previously utilized to solve MCO problems (61-64). The MCO problem associated to biomarker finding proposed in this manuscript will be approached through the use of DEA, offering a novel strategy requiring little statistical knowledge and no parameter adjustment by the user. The details of the related methods are described next.

2.3 Methodology Background

2.3.1 Multiple Criteria Optimization (MCO)

An optimization problem involves finding the best solution from all feasible solutions. When a single criterion or performance measure is used, this is mathematically represented by a so-called objective function. Thus, the best solution is the one with the largest –in maximization- or smallest –in minimization- value of the objective function. When two or more conflicting criteria are considered, then the case is one of multiple criteria optimization (MCO). As an example of this kind of problems we will consider the selection of a car considering its retail cost altogether with its safety rating as shown in Figure 2-7. Those criteria are in conflict because the objective to the car retail cost is about minimizing and the safety rating would be intended to be maximized, but safer cars are expensive, because producers invested in design and special materials, while cheaper cars lack in safety.

Due to the conflict between the considered criteria, a unique best solution cannot be reached for this kind of problems. An MCO problem aims to locate the best compromises between the considered criteria instead.

The resulting best compromises are known as Pareto-efficient solutions, or simply efficient solutions (65), and they form an efficient frontier. Efficient solutions are those options in the feasible set of solutions for which the performance of one criterion cannot be improved without worsening at least another one (66).

The feasible space can be represented in *k* dimensions, each one associated to a criterion being considered. In Figure 2-7, the car selection example, *k*=2 conflicting criteria, placing the car retail price on the *x* axis and the safety rating on the *y* axis. All points represent the feasible space, i.e. all possible cars. Considering the minimization of the first criteria and the maximization for the second criteria in this case, the best compromises or efficient solutions are joined by straight-line segments and they conform the efficient frontier of the problem. As it can be appreciated in Figure 2-7, moving from one efficient solution to the other implies improving in one criterion but necessarily losing in the other one.

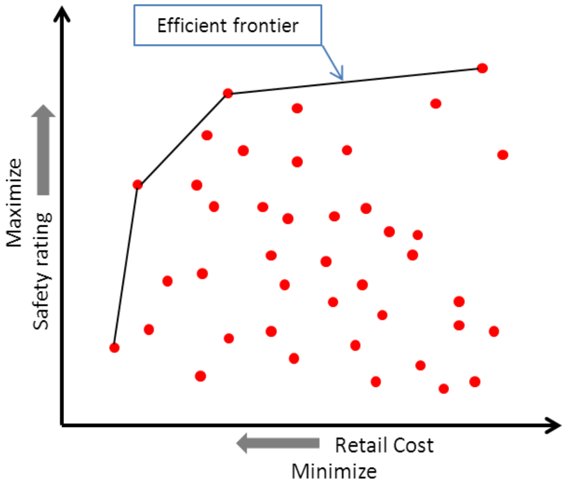
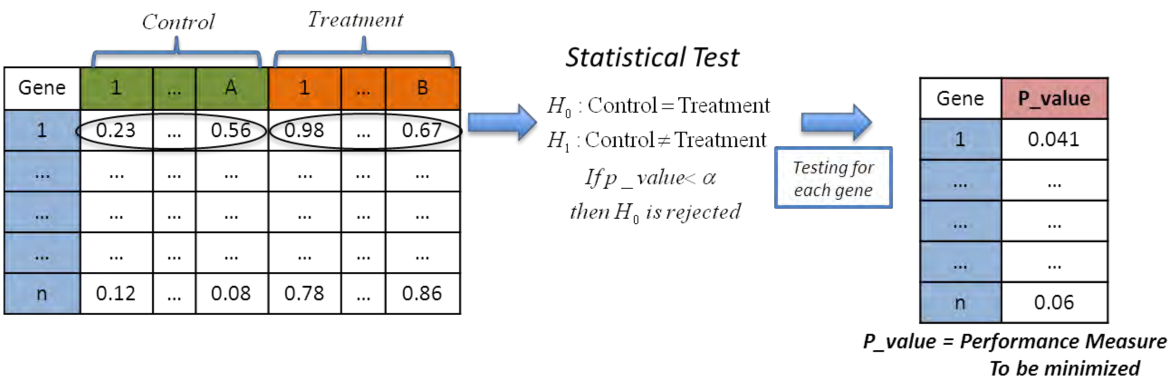


Figure 2-7. Illustration of a Multiple Criteria Optimization Problem

In MCO applications, the criteria under consideration are also called Performance Measures (PM). Different conflicting PMs can be obtained from genetic expression on microarray data to create an MCO problem. In particular for gene selection, we will be interested in PMs for each particular gene under analysis. These PMs can be obtained from statistical tests in the form of *p-values* as explained next.

Statistical testing has been used to detect significant changes for genetic expression when comparing samples of two or more different sates using microarray data e.g. treatment vs control. *P-values* obtained from statistical tests are measures used to determine gene relevance. If the computed *p-value* for a particular gene is smaller than a significance value *α*, the difference in the relative expression for that gene is considered to be large enough when contrasting both states. It follows, then, that lower *p-values* show stronger evidence of significant change between the involved states for the gene under study, and therefore, the *p-value* can be visualized as a PM to be minimized. Figure 2-8 illustrates the process of obtaining a *p-value* from a statistical test for a particular gene in a given experiment contrasting control vs. treatment tissues.

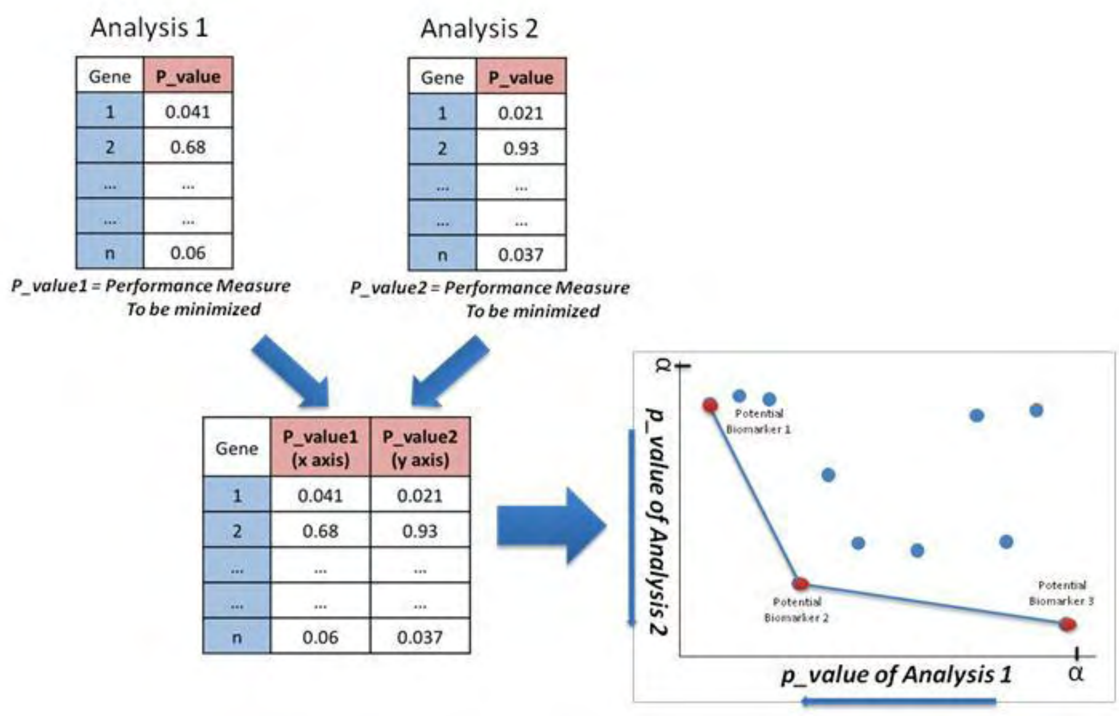


Gene	1	...	A	1	...	B
1	0.23	...	0.56	0.98	...	0.67
...	...	...	...	...	...	...
...	...	...	...	...	...	...
...	...	...	...	...	...	...
n	0.12	...	0.08	0.78	...	0.86

Gene	P_value
1	0.041
...	...
...	...
...	...
n	0.06

Figure 2-8. Scheme for a statistical analysis for genes comparing control vs. treatment. A *p-value* can be treated as Performance Measure to minimize.

If two different microarray experiments characterized through a common set of genes are analyzed through a statistic test, the results are not expected to be the same. Thus, the resulting *p-values* can be treated as conflicting PMs to be minimized on an MCO problem as illustrated in Figure 2-9.



Analysis 1

Gene	P_value
1	0.041
2	0.68
...	...
...	...
n	0.06

P_value1 = Performance Measure To be minimized

Analysis 2

Gene	P_value
1	0.021
2	0.93
...	...
...	...
n	0.037

P_value2 = Performance Measure To be minimized

Gene	P_value1 (x axis)	P_value2 (y axis)
1	0.041	0.021
2	0.68	0.93
...	...	...
...	...	...
n	0.06	0.037

Figure 2-9. Analysis of two p-values as conflicting PM to be minimized in an MCO problem. Genes located in the efficient frontier would be proposed as potential biomarkers.

This work poses that the genes associated to the best compromises resulting from the solution of the defined MCO problem would show stronger evidence to be relevant for cancer differentiation i.e. identified as potential cancer biomarkers. These genes, as shown in Figure 2-9, would be located on the efficient frontier of the associated MCO problem. Although there are many solution methods for an MCO problem (65,66), in this work, the use of DEA is proposed as a first solution approach. DEA is explained in the next section along with some of the issues that must be considered for its use to solve MCO problems.

2.3.2 Data Envelopment Analysis (DEA)

DEA is a tool that can be used to find the best compromises in an MCO problem, i.e. to locate the efficient solutions in the presence of conflict between two or more PMs.

The idea behind DEA is to use an optimization model to compute a relative efficiency score for each particular solution with respect to the rest of the candidate solutions. The resulting best compromises, identified through the maximum possible efficiency score (usually a score of 1), form the envelope of the solution set. These solutions are, indeed, efficient solutions (67-69). In typical DEA the alternatives to evaluate are called Decision Making Units (DMU´s), and these are assumed to consume different inputs (*x*´s) to produce different outputs (*y*´s). The amount of *x*´s consumed and the amount of *y*´s produced are known, thus, in the mathematical programming formulation of DEA, these are constants. Two DEA linear programming formulations proposed by Banker, Charnes and Cooper (68) as envelopment models are shown below:

$$
\begin{aligned}
&\textit{Find} \quad \theta, \lambda_j, s_i^-, s_r^+ \quad \textit{to}\\
&\textit{Minimize} \quad \theta - \varepsilon\left(\sum_{i=1}^{m} s_i^- + \sum_{r=1}^{s} s_r^+\right)\\
&\textit{Subject} \quad \textit{to}\\
&\sum_{j=1}^{n} x_{ij}\lambda_j + s_i^- = \theta x_{i0} \qquad i = 1,2,\dots,m \qquad (1)\\
&\sum_{j=1}^{n} y_{ij}\lambda_j - s_r^+ = y_{r0} \qquad r = 1,2,.\dots,s\\
&\sum_{j=1}^{n} \lambda_j = 1 \qquad \lambda_j \geq 0 \qquad j = 1,2,\dots n
\end{aligned}
$$

$$
\begin{aligned}
&\textit{Find} \quad \phi, \lambda_j, s_i^-, s_r^+ \quad \textit{to} \\
&\textit{Maximize} \quad \phi - \varepsilon\left(\sum_{i=1}^{m} s_i^- + \sum_{r=1}^{s} s_r^+\right) \\
&\textit{Subject} \quad \textit{to} \\
&\sum_{j=1}^{n} x_{ij}\lambda_j + s_i^- = x_{i0} \quad i = 1,2,\ldots, m \qquad (2) \\
&\sum_{j=1}^{n} y_{ij}\lambda_j - s_r^+ = \phi y_{r0} \quad r = 1,2,.\ldots, s \\
&\sum_{j=1}^{n} \lambda_j = 1 \quad \lambda_j \geq 0 \quad j = 1,2,\ldots\ n
\end{aligned}
$$

Where n is the number of *DMU's* to be evaluated, using m different inputs to produce s different outputs. Specifically, DMU_j consumes an amount x_{ij} of input i and produces an amount y_{rj} of output r. We assume that $x_{ij} > 0$ and $y_{rj} > 0$, s_i^- and s_r^+ are slack variables and $\varepsilon > 0$ is a so-called non-Archimedean element defined to be smaller than any positive real number, usually set to a value of $1\text{x}10^{-6}$; λ_j is the dual variable for the DMU_j. In formulation (1), DMU_0 is being compared with a hypothetical linear combination of the other *DMUs* and the value of the objective function is equal to one if there is no such linear combination for which $\sum_{j=1}^{n} \lambda_i x_{ij} < x_{i0}$ for all inputs i, while $\sum_{j=1}^{n} \lambda_i y_{ij} \geq y_{r0}$ for all outputs r (70).

Formulation (1) is called the BCC Input Oriented Envelopment Model and the Formulation (2) is called the BCC Output Oriented Envelopment Model. Given that the set of efficient solutions could differ depending upon the model orientation, because of existent alternate optima, both models are applied to each of the n candidate solutions. A particular solution with an objective function value of 1 (i.e. an efficiency score of 1) using both formulations is considered an efficient solution, and is therefore, in the envelope of the solution set (64).

In this work, it is hypothesized that when DEA is used to solve the MCO problem associated to gene selection, those genes deemed efficient are very likely to be potential biomarkers.

Even though there is a large variety of methodologies to solve MCO problems (65,66), DEA has been chosen for this work given its advantage to be based on linear programming, the most tractable optimization problem. Also, it is important to note that besides the BCC model, other DEA formulations exist. The BCC model, however, seeks for a piecewise linear frontier, that translated in terms of the graphical representation of our MCO problem is equivalent to the convex frontier that we are looking for. Then, it has the advantage of being capable to find all efficient solutions in the convex area of the efficient frontier in our problem (68).

2.3.3 Considerations using DEA to solve the proposed MCO problem

At this point, an analogy between the elements of a DEA statement and an MCO problem can be done as shown in Table 2-2. In the first column the nomenclature used in DEA is presented; the second column shows the analogue nomenclature when an MCO approach is used, and the third column shows the associated elements defined for the proposed MCO problem for potential biomarker search. The elements in the third column are explained in the following sections.

Table 2-2. Equivalency between different approaches used in this work.

DEA approach	MCO approach	Biomarker Search as MCO problem
Decision Making Units (DMU´s)	Candidate Solutions (Alternatives)	Genes
Inputs	Performance Measures to Minimize	p_value$_1$ (to be minimized)
Outputs	Performance Measures to Maximize	Transf_p_value$_2$ (to be maximized)

The work by Bouyssou (71) considers the translation of an MCO problem into a DEA problem formulation. Although the analogy proposed here goes in the opposite direction, it is useful for illustrative purposes to show the similarities between both models. In order to do that, the explanation by Bouyssou (71) is closely followed by letting $X = \{a_1, a_2, \dots, a_l\}$ be a finite set of alternatives that have been evaluated on a set of n criteria. Suppose that the preference is larger-the-better in all criteria, this can be easily inverted to accommodate the case of smaller-the-better. The evaluation of alternative a_k on criterion j is denoted by y_{jk}. Also, suppose that the evaluations of the alternatives on the criteria are strictly positive $(y_{jk} > 0)$.

Alternative a_i is said to dominate alternative a_k if $y_{ji} \geq y_{jk}$ for $j = 1,2,\dots,n$, at least one of these inequalities being strict. An alternative $a \in X$ is said to be efficient in *X* if no alternative in *X* dominates it. If it is possible to find a set of strictly positive weights $w_1, w_2, \dots, w_n$ such that the weighted sum of the criteria for alternative a_i is larger or equal than the weighted sum for any other alternative in *X*, then a_i is efficient (in *X*). Because the considered weights can be normalized, their sum in this case is restricted to be equal to 1. Model (3) shows the resulting model when alternative a_* is being evaluated, this is the primal linear programming formulation for the MCO problem, while Model (4) shows its dual formulation. The latter is equivalent to the BCC output oriented version of the DEA problem as shown below.

$$\begin{aligned}
&\textit{Find} \quad w_j, D \quad \textit{to} \\
&\textit{Minimize} \quad D \\
&\textit{Subject} \quad \textit{to} \\
&\sum_{j=1}^{n} (y_{j*} - y_{jk}) w_j + D \geq 0, \quad k = 1,2,\dots,l \\
&\sum_{j=1}^{n} w_j y_{j*} = 1 \\
&w_j \geq \varepsilon \qquad j = 1,2,\dots n
\end{aligned} \tag{3}$$

$$\begin{aligned}
&\textit{Find} \quad \lambda_k, M, s_j, \quad \textit{to}\\
&\textit{Minimize} \quad M + \varepsilon \sum_{j=1}^{n} s_j\\
&\textit{Subject to}\\
&y_{j*}M + \sum_{k=1}^{l} (y_{j*} - y_{jk})\lambda_k + s_j = 0, \quad j = 1,2,\dots,n\\
&\sum_{k=1}^{l} \lambda_k = 1\\
&\lambda_k \geq 0, \quad s_j \geq 0, \quad M \; \textit{unrestricted}
\end{aligned} \qquad (4)$$

Our proposed approach to biomarker search can be seen as presented in Figure 2-10 with candidate solutions represented by one input and one output.

As discussed previously, the Performance Measures (*p_values*) to be evaluated in the stated MCO problem are intended to be both minimized at the same time. *P_values* are obtained from a statistical test defined to detect significant changes of expression for each gene, the smaller the *p_value* the stronger the evidence favoring a significant change of expression for a particular gene.

In order to use DEA to find the convex efficient frontier of the MCO problem, one must consider different characteristics of the chosen model. For organization purposes, a checklist of these characteristics as described by Sarkis in (72) is presented next.

2.3.3.1 *At least one input and one output must be considered*

In DEA the *DMU´s* to be evaluated use different inputs (*x´s*) in order to obtain different outputs (*y´s*). Given its usual economic implications, the DEA formulation expresses the objective of finding the best compromises when minimizing the utilized inputs and maximizing the obtained outputs. This situation can be seen in Figure 2-10, where the desired frontier is located in the north-west limit of the set. Given the nature of the DEA objectives and its difference with the proposed MCO problem, which generally aims for the minimization of the different PMs considered; a suitable transformation of some of the PMs should be used to provide the maximization instance required for DEA. Here, is important to note, that there is no cause-effect relation between the PMs used in this work, opposing to the typical relation when economical entities are evaluated through DEA.

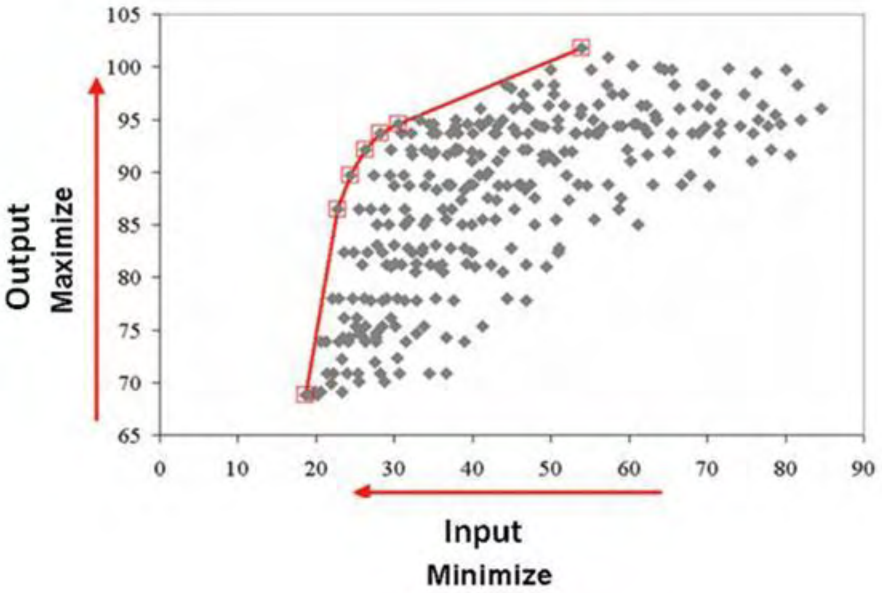


Figure 2-10. Necessary orientation of the MCO problem to be solved using a DEA model.

For the purposes of this thesis, when considering *p_values*, all of them are required to be minimized. Because DEA requires that at least one of the PMs be maximized, when considering two *p_values* one of them should be transformed as follows

$$Transf\_p\_value = (min\_p\_value + max\_p\_value) - p\_value \qquad (3)$$

In theory, *min_p_value* and *max_p_value* should be 0 and 1 respectively, but in practice those exact values are not always reached.

Figures 2-11 and 2-12 illustrate the effect of this transformation in a MCO problem considering two *p_values*. Figure 2-11 shows the original orientation of the objectives and the latter shows the proposed reorientation while Figure 2-12 shows the final orientation to apply the DEA methodology.

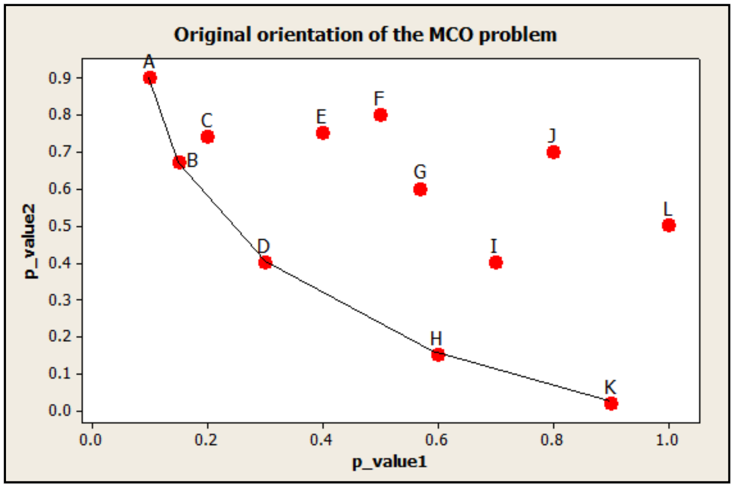


Figure 2-11. Original orientation of the proposed MCO problem using p_values

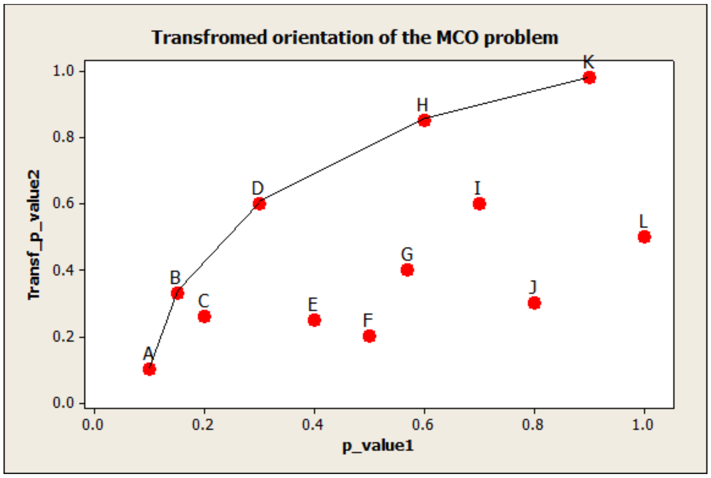


Figure 2-12. Reoriented MCO model to be addressed through DEA using p_values

When more than two *p_values* are considered in the formulation, at least one, but not all the *p_values* should be transformed.

2.3.3.2 Use of only positive numbers

DEA models require that the data meet certain characteristics, one of them being that the PMs have to be expressed in strictly positive values. From statistical properties, in theory, *min_p_val* could get 0, and *max_p_val* could get 1; then, transformation of a *p_value* equal to 1 would generate a *transf_p_value* equal to 0, generating a violation of the presented DEA rule.

In practice, *p_values* from MW test executions present values close to but not equal to zero, and *max_p_value* + *min_p_val* reaches values greater than 1, and *transf_p_value* reaches positive values. If deemed necessary, a translation as shown in Figure 2-13 could be used.

The BCC model which is used in this work, is known to be translation invariant. This feature on the model keeps the proposed methodology unchanged.

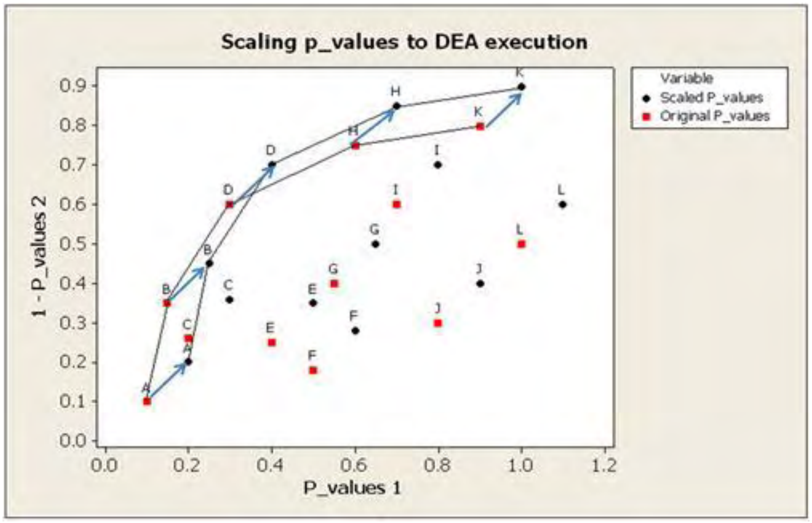


Figure 2-13. Translation of the data in order to avoid zero values for the considered alternatives

2.3.3.3 *The issue of having more alternatives than performance measures*

Some authors (73,74,72) describe the relevance of having more alternatives to be considered in the model than PMs (inputs + outputs). In this thesis, the alternatives to be analyzed are genes. The smallest database to be used in the experimentation phase contains 2,000 genes which is considered a small-sized database in microarrays. The first experimentation scheme will make use of two PMs (*p_values*). Even though other experimental schemes in this work use more than two PMs, the number of PMs will always be between three and four orders of magnitude smaller than the number of alternatives.

2.3.3.4 Returns to scale

Given that the considered PMs do not have an input-output relation, there is no need to make assumptions about returns to scale (70,75). For the practical purposes of finding alternatives that are convex efficient, it is desirable to determine the complete convex efficient frontier. This is entirely possible through the BCC model.

2.3.3.5 Preference Ranking

MCO really entails two different tasks; the search of a set of solutions and the decision making process. The decision making process of an MCO problem is known as Multiple Criteria Decision Making (MCDM) (76). Some authors (71,73,77) have denoted important issues when using DEA as a MCDM tool, the main one being that DEA results in a set of efficient entities and the decision maker has to choose among them without any discriminating criteria such as a ranking.

In the case of biomarker gene search, the objective is to find those genes conforming the efficient frontier of the defined MCO problem. However, up to this point of the research project, it is first critical to validate if the proposed biomarker genes are indeed so. Should a preference structure be required or deemed necessary to improve the results, the issue will be further investigated in the future.

2.3.3.6 Data Normalization

Given that *p_values* are always obtained in the range from 0 to 1, there is no need to balance the magnitudes among different analyses. The translation previously discussed in point number 2 would take care of not incurring into a violation of the restriction of use of positive values and, because the criteria in all cases of interest are obtained necessarily from the existent data, it is not foreseen that missing data becomes an issue.

2.3.3.7 Other Considerations

In DEA a minimal correlation between inputs and between outputs is seeked. In the cases approached in this thesis, it is expected that the *p_values* are statistically independent when using distinct databases and correlated when using a single database. In all cases, independence statistical tests will be carried out for all instances to note their behavior.

Also, because DEA is based on linear programming, it is advisable that the data subject to analysis be roughly in the same order of magnitude to avoid computational problems (78). For the cases considered in this thesis, there is no difference in magnitudes between the *p_values* and the associated transformation *Transf_p_value*, as both of them are always between 0 and 1, avoiding dimensionality issues.

Some authors (73,75,79) have denoted the importance of considering weight restrictions in DEA formulations. Because the weights in the application proposed in this thesis do not have an economic interpretation, it is deemed that weight restrictions are not necessary at this point in the context of interest.

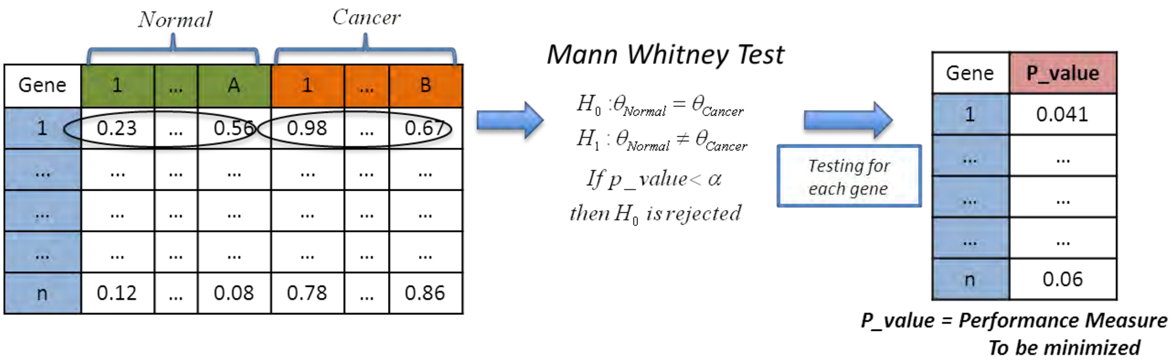
It is known in the DEA literature that there is an inherent difficulty to interpret the weights obtained through the use of DEA, and thus, much of the output cannot be readily interpreted (75). As the problem at hand requires an MCO point of view, it is expected that the efficiency scores be sufficient to provide significant results in this project.

DEA is also known to be computationally intensive; however, the preliminary results show that a database with between 10,000 and 15,000 genes can be completely analyzed in less than 30 min in one of the MacPro workstations available at our research laboratory at UPRM.

Finally, at this point it is important to stress that DEA is used in this work to identify the solutions that are convex efficient, since these are the solutions to the associated MCO problem, therefore, many properties and limitations that apply to DEA when used for benchmarking purposes do not apply to the cases here identified.

3 METHODOLOGY

In this thesis, a *p_value* is obtained for each gene through the Mann-Whitney (MW) nonparametric test for difference of medians of two populations (80) as illustrated in Figure 3-1. A low *p_value* indicates that there is a significant difference between the medians. For each gene, the first population is typically represented by a sample of relative gene expression measurements in normal tissues. Similarly the second population is represented by measurements of the same kind but in cancer tissues. The MW test has been used before in our research group for these purposes (81,82). Formally, the null hypothesis states that the medians from two different samples are equal against the alternative hypothesis of them being different (Figure 3-1). *P-values* obtained through different MW executions, as explained later, are then treated as conflicting PMs in an MCO problem.



	Normal			Cancer		
Gene	1	...	A	1	...	B
1	0.23	...	0.56	0.98	...	0.67
...	...	...	...	...	...	...
...	...	...	...	...	...	...
...	...	...	...	...	...	...
n	0.12	...	0.08	0.78	...	0.86

Mann Whitney Test
$H_0 : \theta_{Normal} = \theta_{Cancer}$
$H_1 : \theta_{Normal} \neq \theta_{Cancer}$
$If\ p\_value < \alpha$
$then\ H_0\ is\ rejected$
Testing for each gene

Gene	P_value
1	0.041
...	...
...	...
...	...
n	0.06

P_value = Performance Measure
To be minimized

Figure 3-1. Statistical evaluation through the MW non parametric statistic test

Different MW analyses can be performed for a particular gene to obtain several *p-values*, including (i) using a variety of tissue combinations from the same database, (ii) using different databases for the same type of cancer, or (iii) using databases for different cancer types, among others. These three alternatives that will be tried in this thesis.

Once with several *p-values* being considered as PMs to be minimized per gene, the representation as an MCO problem and its subsequent solution through DEA complete the proposed strategy to detect potential cancer biomarkers using microarray analysis.

It must be noted that more than two conflicting PMs can be analyzed readily (Figure 3-2). Although the case with more than 3 conflicting PMs cannot be shown graphically, the solution of the MCO problem using DEA is perfectly feasible for more than three performance measures.

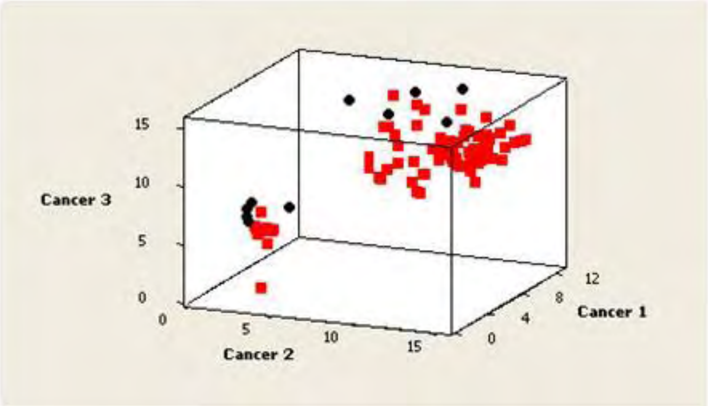


Figure 3-2. Representation of the MCO problem considering k=3 conflicting PMs, efficient frontier of the problem is represented by black points

In summary, in this thesis the search for potential cancer biomarker genes is presented for the first time as a Multiple Criteria Optimization problem. Data Envelopment Analysis will be used in this case as a first approach to solve the MCO problem and converge to a reduced set of potential cancer biomarker genes.

3.1 Data Description

In this section, the different databases to be used in the development of the proposed methodology are described. Three different databases were chosen. All the cases have been originally generated looking for particular patterns or genetic signatures of their corresponding cancer types. After being first published, many works have used these particular databases to test different methodologies and compare their results with the patterns or signatures originally found. Two databases are for Colon cancer and the third one is of Gastric cancer. A deeper description for each of them and their original publications is presented next.

The first database was presented by Alon et al. in (37). The original work is an effort to characterize expression patterns to differentiate between tumor and normal colon tissues. Oligonucleotide microarrays were used in such work. The authors applied a two-way clustering algorithm based on a deterministic annealing algorithm to genes and also to tissues. A high degree of organization in gene expressions capable to effectively separate healthy from cancer tissues was found. Also, high differentiation for clusters of genes with specific functionality was obtained using the proposed methodology. High separability between normal and cancer tissues could be obtained through this method even when the most significant genes were excluded from the data. Misfortunately, a list of genes to compare against our results when using this database is not available in the original work. This database is, however, highly regarded in the literature, thus because it has an almost

100% overlap in the original genes we decided to include it. Furthermore, with the database described next, building different analysis cases was greatly facilitated with its inclusion.

The original number of genes in the Affymetrix platform used were more than 6,500 considering human genes and Expressed Sequence Tags (ESTs). ESTs are sequences that have not been characterized in their functionality. Data available at the source just showed the 2,000 genes with the highest minimal intensity across the samples. All 2,000 genes were used in our analysis. Details about the normalization process of the original data can be found in the original reference (37). This database will be referred to as ***Colon 1*** in the following sections. Table 3-1 shows information regarding this database.

Table 3-1. General Description of Colon 1 Database.

Cancer Type	***Colon***
Original Publication	Alon et al. (37)
State 1 of tissues	Normal tissues
Number of tissues in State 1	22
State 2 of tissues	Cancer tissues
Number of tissues in State 2	40
Number of genes	2,000
Microarray Technology	Affymetrix Oligonucleotide Microarrays
Repository	http://www.molbio.princeton.edu/colondata
List of relevant genes	Not Available

The second database involves colon cancer as well. This database was first published by Notterman et al. in (57). In this work the authors compare the expression between healthy tissues, adenomas (benign tumor tissues) and adenocarcinomas (cancer tissues). Among the most important characteristics of the data is that the tumor samples are paired. This means that, from the same patient, a sample of the tumor and its paired sample from the non-tumor colon zone are taken and characterized. Data of adenoma samples were not considered in this thesis. In the original work, the authors found 19 transcripts with at least 4-10.5 fold higher mRNA expression when comparing carcinomas versus normal tissues, and 47 transcripts showed 4-38 fold lower expression in the tumor samples versus normal. Some of the identified relevant transcripts were validated through reverse transcription PCR, others were already reported as abnormally expressed in neoplastic tissue in general, or colon cancer in particular. These lists of 19 and 47 genes will be considered in Chapter 4 for validation purposes against the potential biomarker genes obtained in this thesis from that database.

The experimental platform in these experiments was the Human 6500 GeneChip Set (Affymetrix). The data were normalized by the authors across the experiments to get it centered to 50 units of intensity. The number of genes reported in the original publication is around the 6,500 but the raw data at the source show more than 7,000. To avoid elimination of relevant information no genes were excluded. Table 3-2 shows additional information on this database, from here on referred to as ***Colon 2***.

Table 3-2. General description of Colon 2 Database.

Cancer Type	***Colon***
Original Publication	Notterman et al. (57)
State 1 of tissues	Normal tissues
Number of tissues in State 1	18
State 2 of tissues	Cancer tissues
Number of tissues in State 2	18
Number of genes	7,457
Microarray Technology	Affymetrix Oligonucleotide Microarrays
Repository	http://www.molbio.princeton.edu/colondata
List of relevant genes	Available, 19 overexpressed, 47 underexpressed

The third database used is presented by Hippo et al. in (83). Their principal objective was to gain understanding at molecular level of the carcinogenesis, progression and diversity in gastric cancer. The objective was approached through a comparison of 22 samples of gastric cancer tissues against 8 healthy gastric tissues. The microarray used in this work contains 6,800 approximately, however the raw data found for this publication contained 7,129 genes. This can be attributed to the control spots that are usually placed in every microarray chips. As in the Colon 1 database, all of the original 7,129 were used in this work to avoid potentially important omissions.

Hippo et al. applied a two-way clustering algorithm to successfully distinguish cancer tissues from healthy tissues. They identified a consistent profile of 162 genes that were highly expressed in cancer tissues (overexpressed genes) and 129 highly expressed in healthy tissues (underexpressed genes). The authors report genes related to cell cycle, growth factor, cell motility, cell adhesion and matrix remodeling as highly expressed in cancer tissues. Those genes highly expressed in normal tissues are related to specific gastro intestinal functions and immune response. The important genes will be used for validation purposes in Chapter 4. As with the previous instances, characteristics of this database, from here on called Gastric, are summarized in Table 3-3.

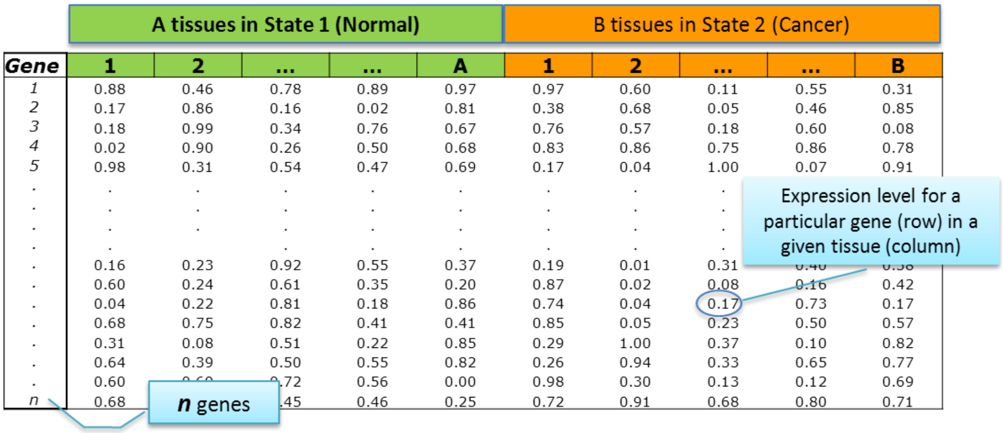
Table 3-3. General Information of the Gastric Database

Cancer Type	***Gastric***
Original Publication	Hippo et al. (83)
State 1 of tissues	Normal tissues
Number of tissues in State 1	8
State 2 of tissues	Cancer tissues
Number of tissues in State 2	22
Number of genes	7,129
Microarray Technology	Affymetrix Oligonucleotide Microarrays
Repository	http://www.ncbi.nlm.nih.gov/geo/, Serie GSE2685
List of relevant genes	Available, 162 upregulated, 129 downregulated

All selected databases follow the Minimum Information About a Microarray Experiments (MIAME) requirements, which is a standard proposed by the Microarray and Gene Expression Data Society (MGED) to facilitate microarray data sharing for interpretation and reproducibility purposes (84). The details of how the data was accessed in the work can be consulted in Appendix A.

For all the databases used it is important to notice that labels of tissues does not always correspond to their ordered position in the database, e.g. tissue Normal 29 is not necessarily in column 29 of normal tissues.

The matrix representation used throughout this work is as follows: (i) the number of genes under analysis is associated with *n* rows, (ii) the array contains the gene reference number in the first column; (iii) the next *A* columns contain the expression readings of each gene for the state 1(*A* Normal tissues); (iv) the remaining *B* columns contain the readings for state 2 (*B* Cancer tissues). Figure 3-3 shows a sketch of a sample array.



	A tissues in State 1 (Normal)					B tissues in State 2 (Cancer)				
Gene	**1**	**2**	**...**	**...**	**A**	**1**	**2**	**...**	**...**	**B**
1	0.88	0.46	0.78	0.89	0.97	0.97	0.60	0.11	0.55	0.31
2	0.17	0.86	0.16	0.02	0.81	0.38	0.68	0.05	0.46	0.85
3	0.18	0.99	0.34	0.76	0.67	0.76	0.57	0.18	0.60	0.08
4	0.02	0.90	0.26	0.50	0.68	0.83	0.86	0.75	0.86	0.78
5	0.98	0.31	0.54	0.47	0.69	0.17	0.04	1.00	0.07	0.91
.	.	.	.	.	.	.	.	.		
.	.	.	.	.	.	.	.	.		
.	.	.	.	.	.	.	.	.		
.	.		.	.	.	.	.	.		
.	0.16	0.23	0.92	0.55	0.37	0.19	0.01	0.31	0.40	0.38
.	0.60	0.24	0.61	0.35	0.20	0.87	0.02	0.08	0.16	0.42
.	0.04	0.22	0.81	0.18	0.86	0.74	0.04	0.17	0.73	0.17
.	0.68	0.75	0.82	0.41	0.41	0.85	0.05	0.23	0.50	0.57
.	0.31	0.08	0.51	0.22	0.85	0.29	1.00	0.37	0.10	0.82
.	0.64	0.39	0.50	0.55	0.82	0.26	0.94	0.33	0.65	0.77
.	0.60	0.60	0.72	0.56	0.00	0.98	0.30	0.13	0.12	0.69
n	0.68	[illegible]	.45	0.46	0.25	0.72	0.91	0.68	0.80	0.71

Expression level for a particular gene (row) in a given tissue (column)

n genes

Figure 3-3. Example of the matrix representation of microarray data

3.2 Experimentation

The software used to solve the DEA formulations of the proposed MCO problems was DEA-Solver Pro 6.0 Nd, from SAITECH Inc. The experiments were executed at the Bio-IE-Lab in the Industrial Engineering Department at UPRM. This laboratory is equipped with 4 MacPro Quad-core workstations. Windows disk partition and Microsoft Office are available for compatibility of DEA-Solver Pro. MatLab is also available in the laboratory for the execution of the MW analysis and Minitab for statistics. Routines and procedures to use these tools are described in Appendix B.

3.3 Analysis procedure

Three different cases are explored in this thesis, all of them schematized in the flowchart presented in Figure 3-4. All cases start with the selection of the database or databases to be used in the analysis. This selection has been already explained in the section 3.1 for this work. If there is just a single database available, a generation of submatrices is performed, corresponding to Case 1; otherwise, the search for genes in common between the considered databases is executed, and the matrices dimensions are reduced just to those common genes. This procedure is followed in Case 2 (multiple databases with 1 cancer type) and Case 3 (multiple databases with multiple cancer types).

The final steps for all the cases correspond to the statistical comparison through MW of either the different submatrices or matrices; the use of the obtained *p_values* to state an MCO problem and its subsequent solution. Further explanation of the specific differences between cases is presented in the following sections.

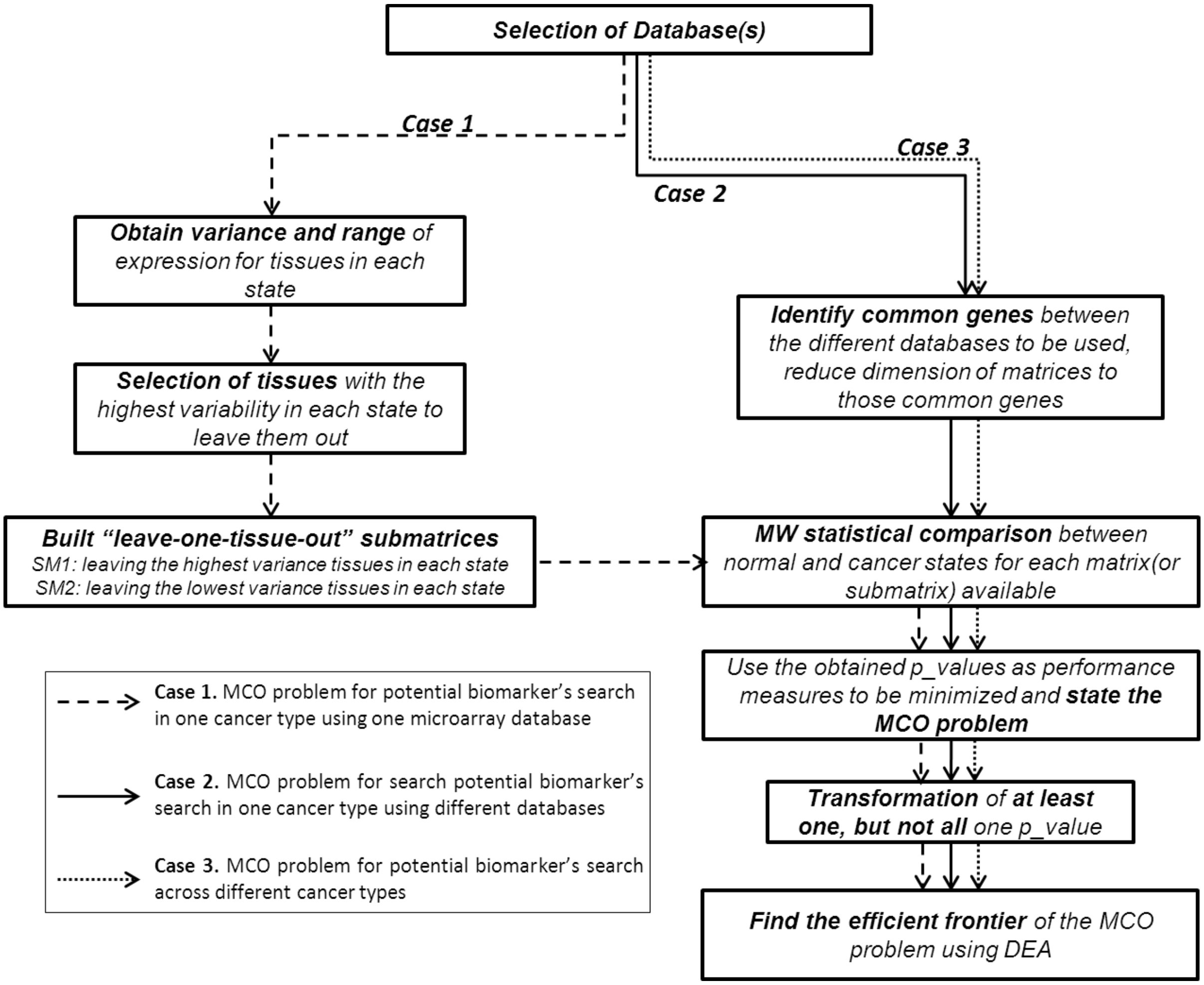

Selection of Database(s)
Case 1
Case 2
Case 3
Obtain variance and range of expression for tissues in each state
Selection of tissues with the highest variability in each state to leave them out
Built "leave-one-tissue-out" submatrices
SM1: leaving the highest variance tissues in each state
SM2: leaving the lowest variance tissues in each state
Identify common genes between the different databases to be used, reduce dimension of matrices to those common genes
MW statistical comparison between normal and cancer states for each matrix(or submatrix) available
Use the obtained p_values as performance measures to be minimized and state the MCO problem
Transformation of at least one, but not all one p_value
Find the efficient frontier of the MCO problem using DEA
Case 1. MCO problem for potential biomarker's search in one cancer type using one microarray database
Case 2. MCO problem for search potential biomarker's search in one cancer type using different databases
Case 3. MCO problem for potential biomarker's search across different cancer types

Figure 3-4. Diagram of the experimental cases proposed

3.4 Case 1 - MCO Problem for potential biomarker's search for One Cancer Type using One Microarray Database

In order to exploit the characteristics of a single microarray database, the construction of submatrices of the same dataset is explored. The submatrices are built using a "leaving-one-tissue-out" strategy for each state. To select the tissue that is excluded from the original matrix, the variance on each tissue is considered. This process is shown in Figure 3-5.

In one submatrix the excluded tissues correspond to those with the highest variance in each state (dotted loops), for the second submatrix the excluded tissues correspond to those with the lowest variance in each state (continuous loops). This strategy aims to keep different levels of variance among the tissues in each submatrix, expecting their statistical evaluation (and resulting *p_values*) to be different.

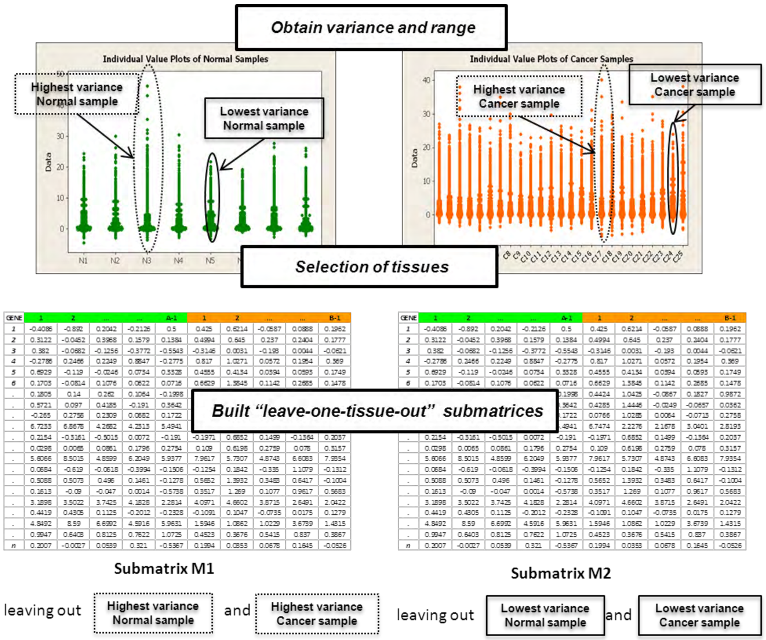

Obtain variance and range
Individual Value Plots of Normal Samples
Highest variance Normal sample
Lowest variance Normal sample
Data
N1
N2
N3
N4
N5
Individual Value Plots of Cancer Samples
Highest variance Cancer sample
Lowest variance Cancer sample
Selection of tissues
Built "leave-one-tissue-out" submatrices
GENE
Submatrix M1
leaving out
Highest variance Normal sample
and
Highest variance Cancer sample
Submatrix M2
leaving out
Lowest variance Normal sample
and
Lowest variance Cancer sample

Figure 3-5. Main processes for Case 1execution

A statistical comparison between normal and cancer states using MW is performed for each gene in both submatrices. The resulting two *p_values* will be treated as conflicting performance measures to build an MCO problem. A transformation of at least one but not all of those *p_values* has to be done using equation (3) in order to use the DEA methodology to solve it as detailed in section 2.3.3.

Once the first efficient frontier is found, their corresponding genes are removed from the original set and the models are executed again to obtain the next efficient frontier, this process is repeated until de 10th frontier is found, this process can be seen graphically in Figure 3-6. This number of efficient frontiers is explored in all the described cases.

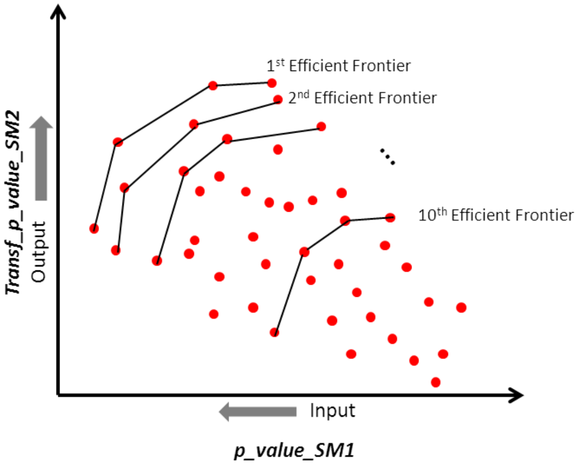


Figure 3-6. Illustration of the obtention of the first 10 frontiers

Admittedly, the number of frontiers to be searched is arbitrary at this point. A series of experiments are being carried out in our group to propose a systematic way to select this number.

Execution of Case 1 involved the three different databases described in the Data Description Section (***Colon 1***, ***Colon 2***, ***Gastric***). The results of the proposed methodology when applied to each of them are detailed below.

3.4.1 Case 1 using Colon 1 Database

Considering Colon 1 database with A = 22 and B = 40, the “leave-one-tissue-out” procedure will result in submatrices with 21 normal tissues and 39 cancer tissues.

Variance for each available tissue in both states is obtained to select which tissues are left out for each submatrix, a summary of those samples is presented in Table 3-4. From that selection, Submatrix 1 is built leaving out tissues Normal 29 and Cancer 6, and Submatrix 2 leaving out tissues Normal 12 and Cancer 10.

Table 3-4. Labels of selected tissues in Colon 1 database

	Variance	
	Lowest	***Highest***
Normal	Normal 12	Normal 29
Cancer	Cancer 10	Cancer 6

For each submatrix a *p_value* from the MW comparison between states Normal and Cancer is obtained per gene to build the MCO problem. The original orientation of the MCO problem is presented in Figure 3-7 and its corresponding representation when one *p_value* is transformed is shown in Figure 3-8. Only the transformed representation of the MCO problem will be shown for the remaining executions.

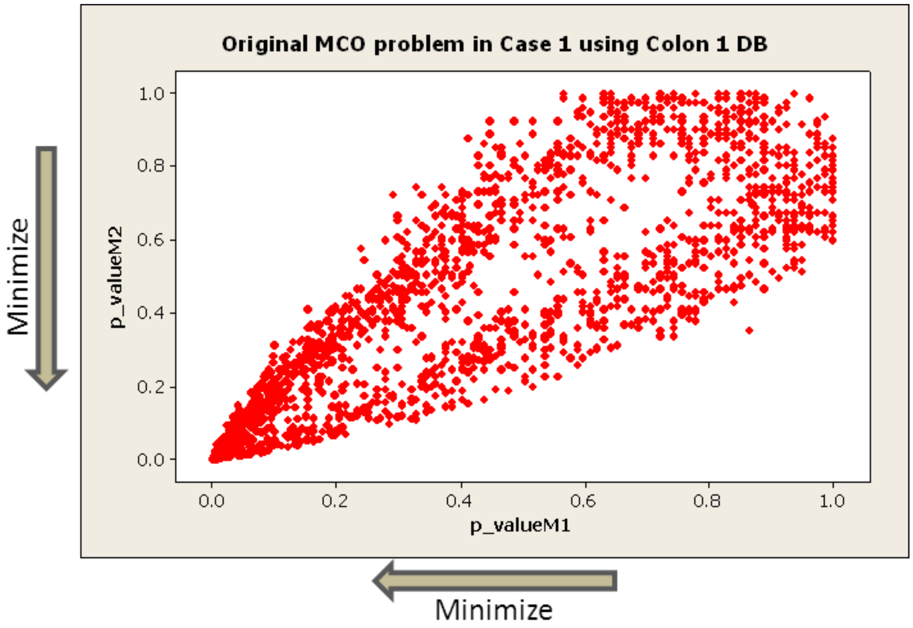


Figure 3-7. Original orientation of the MCO problem for Case 1 using Colon 1 Database (ρ = 0.852).

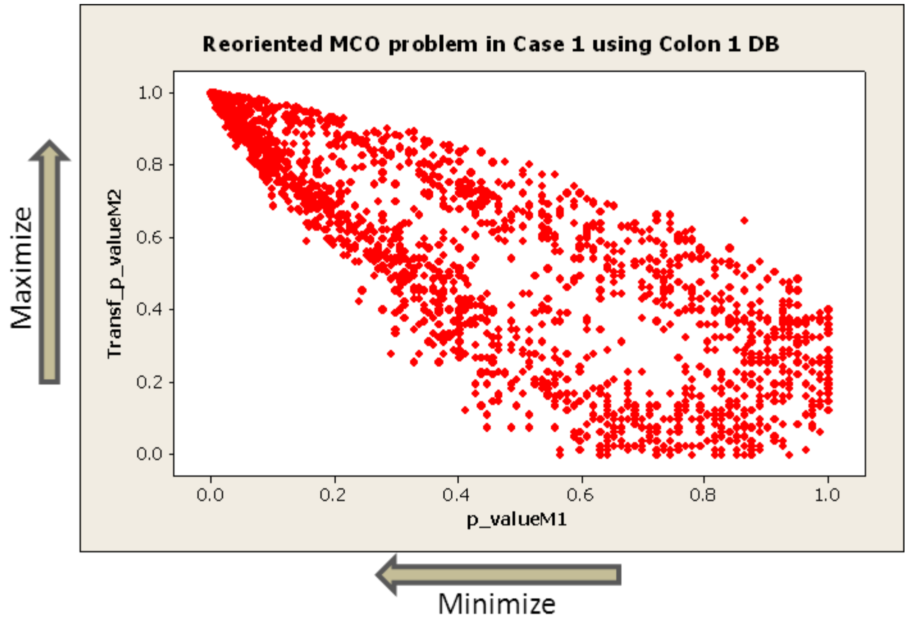


Figure 3-8. Presentation of the MCO problem after transformation of p_valueM2 in Case 1 using Colon 1 Database (ρ = -0.852).

The evolution of the different number of genes found across the first 10 evaluated efficient frontiers is detailed in Table 3-5. The first column describes the number of Genes available when the analysis for each frontier is performed, BCCI and BCCO columns describe the number of genes found through the execution of each model orientation, the Genes in common is the number of genes that were found in both analyses, constituting each convex frontier. The repeated genes column shows how many genes were found as efficient in more than one of their replicates. The last column provides the accession number of those repeated genes, when this happens, all of its replicates are removed from the original list, affecting the number of genes to analyze in the next frontier. This can be seen in the change of Genes to analyze from frontier six to seven.

Table 3-5. Evaluation of different DEA frontiers of the Case1 using Colon 1 database.

Frontier	*Genes to analyze*	*BCCI*	*BCCO*	*Common Genes*	*Repeated genes*	*Accession of repeated genes*
1	2000	2	21	2	0	
2	1998	2	19	2	0	
3	1996	2	17	2	0	
4	1994	4	15	4	0	
5	1990	2	11	2	0	
6	1988	3	11	3	1	M76378 (3)
7	1983	2	6	2	0	
8	1981	3	7	3	0	
9	1978	3	7	3	0	
10	1975	4	6	4	0	
		Total		***27***		

Results of Case 1 with Colon 1

The resulting 27 genes found in the first 10 frontiers are listed in Table 3.6. The first column describe the number of genes found, the second one details the frontier on which each of those genes was found; then, the Accession number, which corresponds to a unique identifier of a gene, and finally the gene Symbol and Name are also shown. This Table is presented here for completeness in this first case. Tables summarizing all results in this work can be found on Appendix C (Tables C1 to C7).

In this first case, 13 of the efficient genes (HSPD1, GTF3A, IL8, DES, VIP, NME1, GSN (gelsolin), HMGA1, CDH3, SRPK1, CFD (adipsin), NPM1, MT1G) have at least one reference listing them as relevant to some cancer processes (85-96).

Others, like MYL9, DARS, GUCA2B, have been proposed as potential cancer biomarkers based on the analysis with different methodologies using microarray data, however their biological validation is still pending (97-102).

HNRNPA1 has been explored as having potential relevance for cancer metabolism but the evidence is not supportive enough at this point neither to discard nor support the hypothesis (103). The rest of genes in this first list have not been investigated in their contribution in cancer processes yet as far as our literature review goes. These last are, indeed the opportunity areas to detect new biomarkers. Further analysis of these results is presented in a subsequent chapter.

Table 3-6. Resultant efficient genes for the MCO problem of Case 1 using Colon 1 database.

Gene	*Frontier*	*Accession*	*Symbol*	*Name*
1	1	M22382	HSPD1	Heat shock 60kDa protein 1 (chaperonin)
2	1	R87126		EST: yq31b10.s1
3	2	H08393	WDR77	WD repeat domain 77
4	2	R36977	GTF3A	General transcription factor IIIA
5	3	J05032	DARS	aspartyl-tRNA synthetase
6	3	M26383	IL8	interleukin 8
7	4	X63629	CDH3	cadherin 3, type 1, P-cadherin (placental)
8	4	H40095		EST: yn85b03.s1
9	4	Z50753	GUCA2B	guanylate cyclase activator 2B (uroguanylin)
10	4	M63391	DES	desmin
11	5	J02854	MYL9	myosin, light chain 9, regulatory
12	5	X12671	HNRNPA1	heterogeneous nuclear ribonucleoprotein A1
13	6	U09564	SRPK1	SFRS protein kinase 1
14	6	H43887	CFD	Complement factor D (adipsin)
15	6	M76378	CSRP1	cysteine and glycine-rich protein 1
16	7	M36634	VIP	vasoactive intestinal peptide
17	7	T86473	NME1	Non-metastatic cells 1, protein (NM23A) expressed in
18	8	H06524	GSN	Gelsolin
19	8	R84411	SNRPB	Small nuclear ribonucleoprotein polypeptides B and B1
20	8	X14958	HMGA1	high mobility group AT-hook 1
21	9	T92451	TPM2	Tropomyosin 2 (beta)
22	9	M26697	NPM1	Nucleophosmin (nucleolar phosphoprotein B23, numatrin)
23	9	T71025	MT1G	Metallothionein 1G
24	10	X86693	SPARCL1	SPARC-like 1 (hevin)
25	10	T47377	S100P	S100 calcium binding protein P
26	10	U30825	SRSF9	Serine/arginine-rich splicing factor 9
27	10	D31885	ARL6IP1	ADP-ribosylation factor-like 6 interacting protein 1

3.4.2 Case 1 using Colon 2 database

When Colon 2 database is used for the execution of Case 1, the process is analogous to the previous instance. The selected tissues to build the “leave-one-tissue-out” submatrices are presented in Table 3-7. Graphical representation of the reoriented MCO problem, after transformation of one *p_value* is shown in Figure 3-9, and the evaluations of the 10 first efficient frontiers of the MCO problem are presented in Table 3-8.

Table 3-7. Labels of selected tissues in Colon 2 database

	Variance	
	Lowest	***Highest***
Normal	Normal 34	Normal 8
Cancer	Cancer 29	Cancer 9

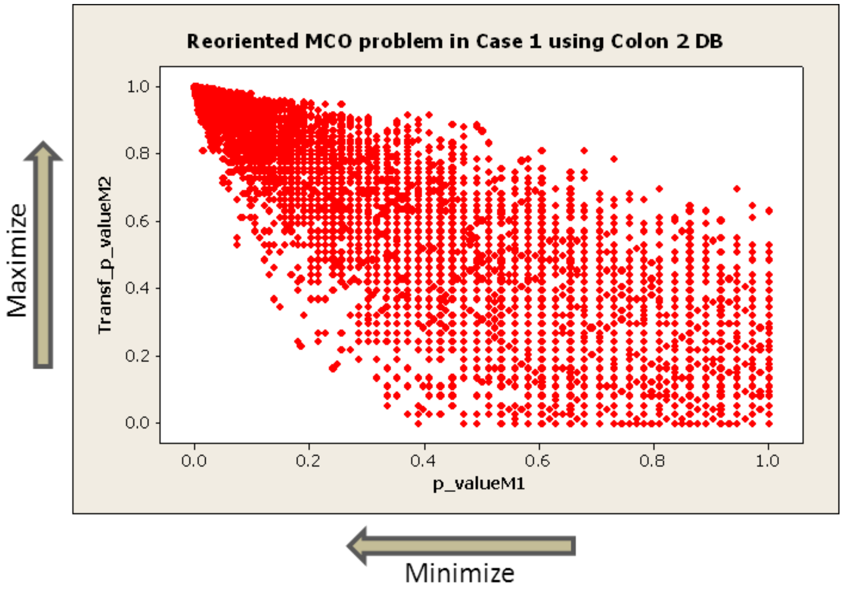


Figure 3-9. Presentation of the MCO problem after transformation ot p_valueM2 for Case 1 using Colon 2 Database (ρ = -0.888).

Table 3-8. Evaluation of different DEA frontiers of the Case 1 using Colon 2 database.

Frontier	***Genes to analyze***	***BCCI***	***BCCO***	***Genes in common***	***Repeated genes***	***Accession of repeated genes***
1	7457	11	249	11	2	Z49269 (3), H57136 (2)
2	7443	11	236	11	2	X56597 (2), M36981 (2)
3	7430	7	225	7	1	M76378 (3)
4	7421	8	216	8	0	
5	7413	9	208	9	2	X54942 (2), L11708 (2)
6	7402	4	197	4	1	H43887 (2)
7	7397	5	193	5	1	J03037 (3)
8	7390	10	186	10	1	R87126 (2)
9	7379	5	176	5	1	M84526 (2)

10	7373	9	170	9	1	Z46629 (3)
			Total	***79***		

Results of Case 1 with Colon 2 database

The first 10 efficient frontiers have a total of 79 genes. General information for these genes is presented in Table 2C of Appendix C.

Thirty seven of the 79 genes found were overexpressed, the rest 42 were underexpressed. Among the overexpressed, genes like GRTF3A, HSPD1, CKS2 identified in the previous intances were found again with this new dataset. Some interesting results of this case include the fact that two genes that code for family NME were found (NME1 and NME2), furthermore, the NME1 gene was found twice through this method due to the existence of replicates under accession synonyms in the database. The reduced NME expression has been related to an increase in the metastatic potential and a more aggressive disease in a variety of cancer types (104).

3.4.3 Case 1 using Gastric database

Applying the same process of analysis to the Gastric cancer database, the selected tissues to build the submatrices are described in Table 3-9. Representation of its corresponding MCO problem statement is shown in Figure 3-10. And the evaluation of the first ten efficient frontiers is presented in Table 3-10.

Table 3-9. Labels of selected tissues in Gastric database

	Variance	
	Lowest	***Highest***
Normal	Normal 7	Normal 5
Cancer	Cancer 8	Cancer 17

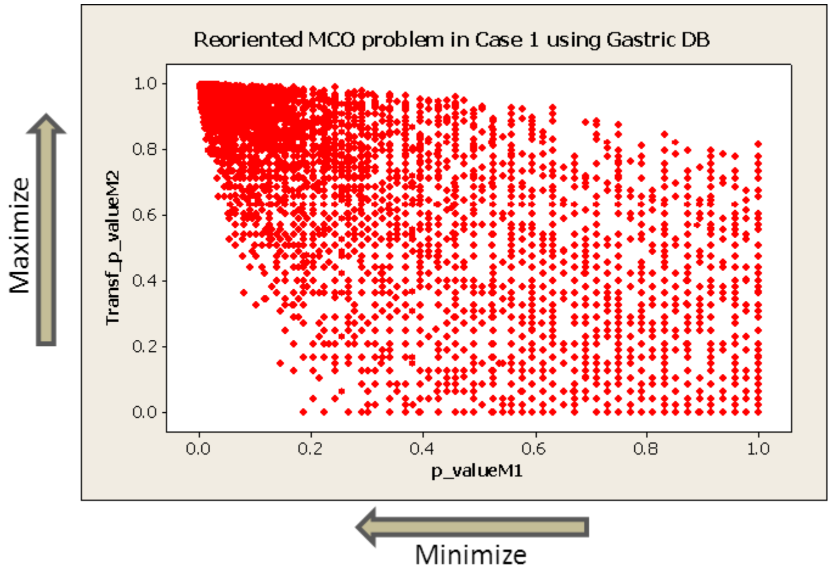


Figure 3-10. Presentation of the MCO problem after the transformation of p_valueM2 for Case 1 using the Gastric Database (ρ = -0.793).

Table 3-10. Evaluation of different DEA frontiers of the Case 1 using the Gastric database

Frontier	*Genes to analyze*	*BCCI*	*BCCO*	*Genes in common*	*Repeated genes*	*Accession of repeated genes*
1	7129	1	129	1	0	
2	7128	1	128	1	0	
3	7127	5	127	5	0	
4	7122	9	122	9	1	D29675 (2)
5	7112	33	113	32	0	
6	7079	4	82	4	0	
7	7075	3	78	2	0	
8	7073	4	77	4	0	
9	7069	3	74	3	1	
10	7066	15	72	13	0	
			Total	***74***		

Results of Case 1 with Gastric database

The first 10 efficient frontiers had a total of 74 genes. The most relevant information for these genes is presented in Table 3C in Appendix C.

Twenty five genes were overexpressed and the other 49 genes showed an underexpressed behavior. Among the overexpressed genes, CKS2 and CKS1B that belong to the cytokine family are showed. Also HMGA1 and HMGB1 that belong to the high mobility group were found. Another couple of related genes was COL1A2 and COL1A1, both belonging to collagen protein, the heat shock protein HSP90AB1 was also found in this set of overexpressed genes. In the underexpressed genes Choline Acetyltransferase (CHAT) was found, along with Aldehyde dehydrogenase (ALDH2) among many others. Many of the genes found here are explored in deep in Chapter 4, to establish their relevance in cancer.

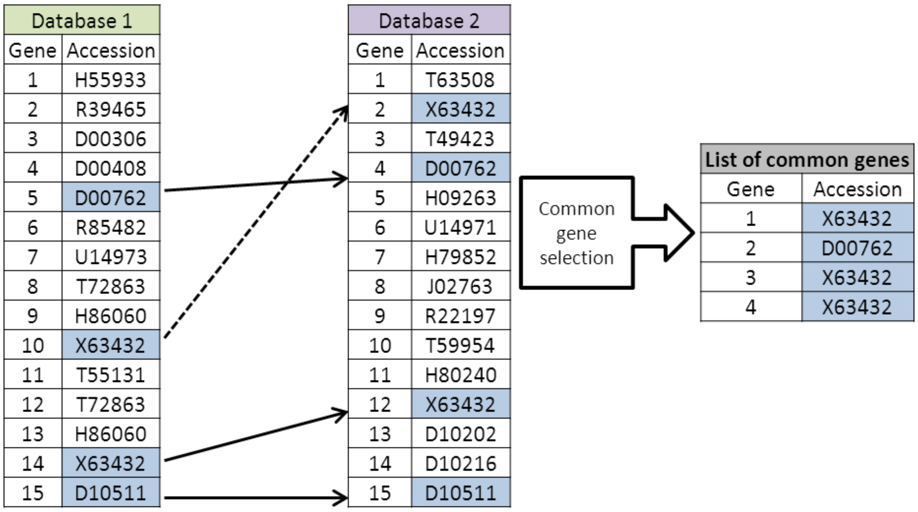
3.5 Case 2 - MCO Problem for potential biomarkers' search on One Cancer Type using Different Microarray Databases

As described previously, microarrays are custom made depending upon the objectives of a particular study. From this fact, when two microarray databases are compared, their explored genes are not necessarily the same even when studying the same cancer type. Also, different sets of microarray experiments are built under different physical conditions and executed with different resources.

However, for those genes that could be found in common between two different databases these discrepancies could be exploited. If analyses performed with different databases arrive to the same conclusion, the found genes would be robust to those differences. Results would point to potential biomarkers for a specific cancer type that are robust to differences on experimental data.

Case 2 of the proposed methodology is, then, built under this premise. Two independent databases from the same cancer type are analyzed simultaneously, and *p_values* of the genes present in both analyses are used to build the MCO problem.

The two databases related to colon cancer are used in this case. The common genes to both databases were selected. This process is shown with an illustrative example in Figure 3-11, where two lists of 15 accession numbers are used. The two lists have just four genes in common, but their position in each list is different. Tools like the vlookup function in MS Excel to find those elements present in both lists have been used in this work for this task.



Database 1	
Gene	Accession
1	H55933
2	R39465
3	D00306
4	D00408
5	D00762
6	R85482
7	U14973
8	T72863
9	H86060
10	X63432
11	T55131
12	T72863
13	H86060
14	X63432
15	D10511

Database 2	
Gene	Accession
1	T63508
2	X63432
3	T49423
4	D00762
5	H09263
6	U14971
7	H79852
8	J02763
9	R22197
10	T59954
11	H80240
12	X63432
13	D10202
14	D10216
15	D10511

List of common genes	
Gene	Accession
1	X63432
2	D00762
3	X63432
4	X63432

Figure 3-11. Illustrative example of the selection of the common genes between the considered Databases.

Once the common genes are selected, their corresponding readings in each database are gathered. To generate the statistical evaluation between the healthy and cancer states for each gene, the MW test is used as in the previous case. As a result of the application of the MW test for each database, two *p_values* are obtained, ***p_valueDB1*** and ***p_valueDB2***.

As in Case 1, in order to build the MCO problem, at least one but not all *p_values* will be transformed using equation (3). Specifically in this case, *p_values* from the second database (***p_valueDB2***) will be transformed (***transf_p_valueDB2***), while those corresponding to the first database (***p_valueDB1***) are left with no transformation. With the stated MCO problem, DEA is applied for its solution and the first 10 frontiers are found.

3.5.1 Case 2 using Colon 1 and Colon 2 Databases

Case 2 is executed using Colon 1 and Colon 2 databases. The first one contains 2,000 genes while the second one 7,457 genes. There are *1,988 genes present in both databases*. The genes in common were detected by their Accession number.

Readings for the genes in common were gathered in both databases, and the MW comparison between state normal and cancer was executed for each database. *P_values* were stored in ***p_value_Colon1*** and ***p_value_Colon2*** respectively. The required transformation was applied to ***p_value_Colon2*** using ecuation (3) to obtain ***Transf_p_value_Colon2***. The final representation of the MCO problem is shown in Figure 3-12. Solution of this transformed problem is addressed through DEA and its ten first frontiers are found. The evolution of the different frontiers found and the corresponding number of genes analyzed in each of them is detailed in Table 3-11.

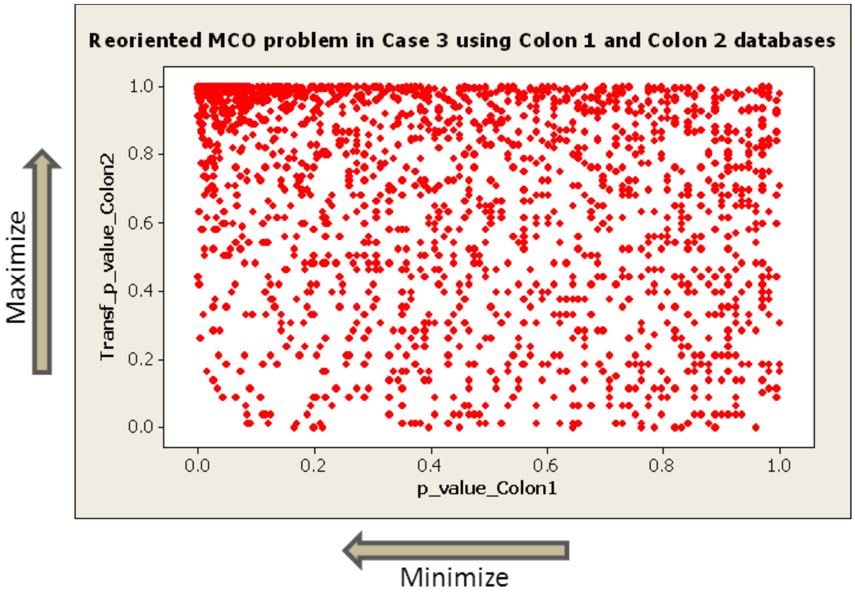


Figure 3-12. Reoriented MCO problem for Case 2 using Colon 1 and Colon 2 Databases.

Table 3-11. Amount of genes found in the first 10 efficient frontiers for the execution of Case 2 using Colon1 and Colon 2 databases.

Frontier	*Genes to analyze*	*BCCI*	*BCCO*	*Genes in common*	*Repeated genes*	*Accession of repeated genes*
1	1988	2	114	2	0	
2	1986	3	112	3	0	
3	1983	3	110	3	0	
4	1980	6	108	6	0	
5	1974	3	102	3	0	
6	1971	2	99	2	0	
7	1969	4	97	4	1	M76378 (3)
8	1963	4	92	4	1	Z49269 (2)
9	1958	9	93	9	0	
10	1949	5	85	5	0	
			Total	***41***		

Results of Case 2 with Colon 1 and Colon 2 databases

As reported in Table 3-11, there were 41 genes located in the first 10 frontiers of the MCO problem related to this execution. The complete list of those genes with their accession number, symbol and name is presented in Table 4C in Appendix C.

When comparing results of the execution of Case 1 using Colon 1 and Colon 2 databases with this instance of Case 2, many genes are consistently found. Among them, HSPD1, GTF3A and NME1. The first two are already present in patents of genetic signatures related to cancer (105,106); the latter is present in an experimental tool already available for commercialization in chips to detect tumor suppressor genes (107). GSN, DES, VIP and HMGB1 are also found through all the analyses, supporting their evidence as potential biomarkers. Further evaluation of a selected set of common genes obtained through the different executions is presented in Chapter 4 with validation purposes.

3.6 Case 3 - MCO Problem for potential biomarkers' search Across Different Cancer Types

In the same way that discrepancies are expected between databases from the same cancer type, larger differences are expected when databases of different cancer types are analyzed. Genes in common are, indeed, harder to find too.

When a particular set of genes has been explored in different databases pertaining to more than one cancer type, independent analyses using those databases can be performed. Genes resulting relevant for both independent analyses would be robust to discrepancies even between cancer types, *i.e.* these could be called potential biomarkers for the general state of cancer.

When considering databases from two different cancer types, the MCO problem obtained from their statistical analyses is similar to the two dimension MCO problem solved in Cases 1 and 2. The increase of dimensions in the analysis can be straightforwardly done. Adding a cancer type would result in adding a performance measure in the MCO problem to be finally translated in terms of DEA for its solution.

In an extreme case, if databases for all the existing cancer types were available and used with this scheme, the resulting genes would correspond to potential cancer biomarker genes to the general state of cancer regardless the original zone of the tumors. If *k* different cancer types exist, the MCO problem would have *k* performance measures to be considered and the DEA problem to solve would have *k* dimensions.

Methodologically, the structure of this Case can be seen as an extension of Case 2, but the different databases would come from different cancer types. The different stages of the analysis are described next.

To select the databases for this case, it should be considered that these do not have to be of the same cancer type, however, they should have an overlap in their original inclusion of genes. The states in each cancer type database should be ‘normal’ compared to ‘cancer’. The number of tissues in each state can be different for each cancer type database. The scale of the data in each cancer type database can be different too. Being capable to accommodate these differences is a mayor advantage of DEA.

The search process is similar to that described in Case 2 and illustrated in Figure 3-11 in section 3.2.

The MW statistical comparison for each gene in common to the databases involved is executed. Their related *p_values* (***p_valueC1*** and ***p_valueC2***) are used to state the MCO problem. As explained for Cases 1 and 2 a transformation of at least one but not all *p_values* is performed, in this case, ***p_valueC2*** was chosen to be transformed.

Once the MCO problem is stated, it is solved through the use of DEA to find the efficient frontier.As has been described in the previous cases, the frontier searching process is executed until the 10th efficient frontier is found.

For this Case all the 3 different combinations are tried. The first two analyses keep the same two dimension structure that has been explored in Cases 1 and 2. The third analysis is an exploration of the methodology potential through the use of all the available databases to build an MCO problem with three performance measures, and a consequent DEA structure with two inputs and one output.

The first combination involves Colon 1 and Gastric databases. The second one considers Colon 2 and Gastric databases. The last one make use of all three available databases. These three structures and the number of common genes used for their analyses are summarized in Table 3-12. The execution and results of these combinations are described in the following sections.

Table 3-12. Different experimental combinations for Case 3.

Experimental executions for Case 3	***Cancer 1 DB***	***Cancer 2 DB***	***Cancer 3 DB***	***Number of common genes***
Combination 1	Colon 1	Gastric	N/A	674
Combination 2	Colon 2	Gastric	N/A	2,360
Combination 3	Colon 1	Colon 2	Gastric	674

3.6.1 Case 3 using Colon 1 and Gastric Databases

As detailed in Table 3-12 there were 674 genes in common between Colon1 and Gastric databases. The MW statistical test is applied to readings of those genes for each database, obtaining their related p_values (***p_value_Colon1*** and ***p_value_Gastric***). The transformation needed is performed to ***p_value_Gastric***, obtaining ***Transf_p_value_Gastric***. The graphic representation of this MCO problem is shown in Figure 3-13. Then the corresponding ten first efficient frontiers are found through DEA. The frontiers' evolution is shown in Table 3-13, with a total of 84 genes. Complete information of the identified genes is presented in Table 5 of Appendix C.

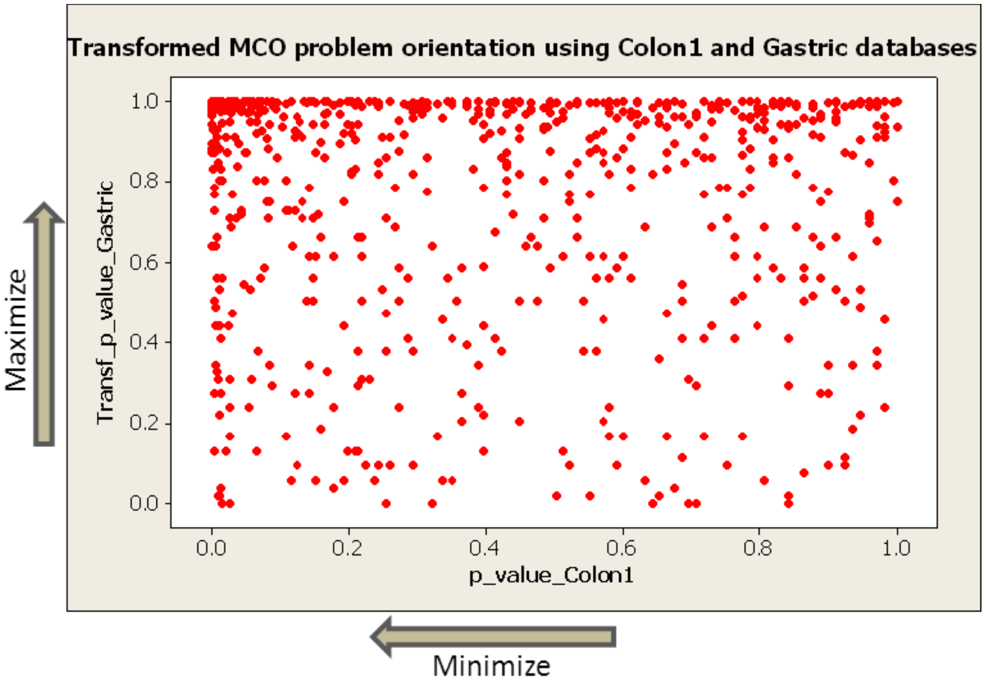


Figure 3-13. Reoriented MCO problem for Case 3 using Colon 1 and Gastric Databases.

Table 3-13. Frontiers' evolution for Case 3 using Colon 1 and Gastric Database.

Frontier	*Genes to analyze*	*BCCI*	*BCCO*	*Common Genes*	*Repeated genes*
1	674	4	25	4	0
2	670	8	27	8	0
3	662	6	24	6	0
4	656	1	27	8	0
5	647	8	22	8	0

6	638	9	20	9	0
7	630	10	18	10	0
8	620	10	18	10	0
9	610	11	17	11	0
10	599	11	13	10	0
			Total	***84***	

Results of Case 3 execution with Colon 1 and Gastric databases

Among the most relevant genes found in this list was Carbonic anhydrase IX (CA9) which has already been characterized as a cancer biomarker gene (108) and is located in the fifth frontier of this analysis. Relevance of Calpain 2(CPN2), found in the seventh frontier has been already explored in some other cancer types (109,110). The same pair of kinases described in last case (CKS2 and CKS1B) are also shown in this analysis. Both of them appeared in the first frontier of the MCO problem along with HSPD1(heat shock protein). HSP90AA1, another heat shock protein is also detected although in the third frontier. HMGB1, a gene of the high mobility group, is shown in the second frontier. Some of the genes identified in this instance were also found in the first three described schemes, strengthening the evidence of their relevance in cancer, as well as the performance of the proposed method. As will be seen in the validation section, 10 out of the 84 identified genes were identified with already reported functions in cancer processes.

3.6.2 *Case 3 using Colon 2 and Gastric Databases*

Considering Colon 2 and Gastric databases, there were 2,360 genes in common. The readings these genes were gathered from the original databases and their MW statistical comparisons between healthy and cancer states are executed. The obtained *p_values* were stored in ***p_value_Colon2*** and ***p_value_Gastric***. The required transformation of one performance measure is executed into that obtained through gastric database, ***Transf_p_value_Gastric***. Representation of the MCO problem corresponding to this case is shown in Figure 3-14. A summary of evaluation for the first ten frontiers of this case is presented in Table 3-14. There are a total of 85 resulting genes. Complete information of these genes is presented in Table 6 of Appendix C.

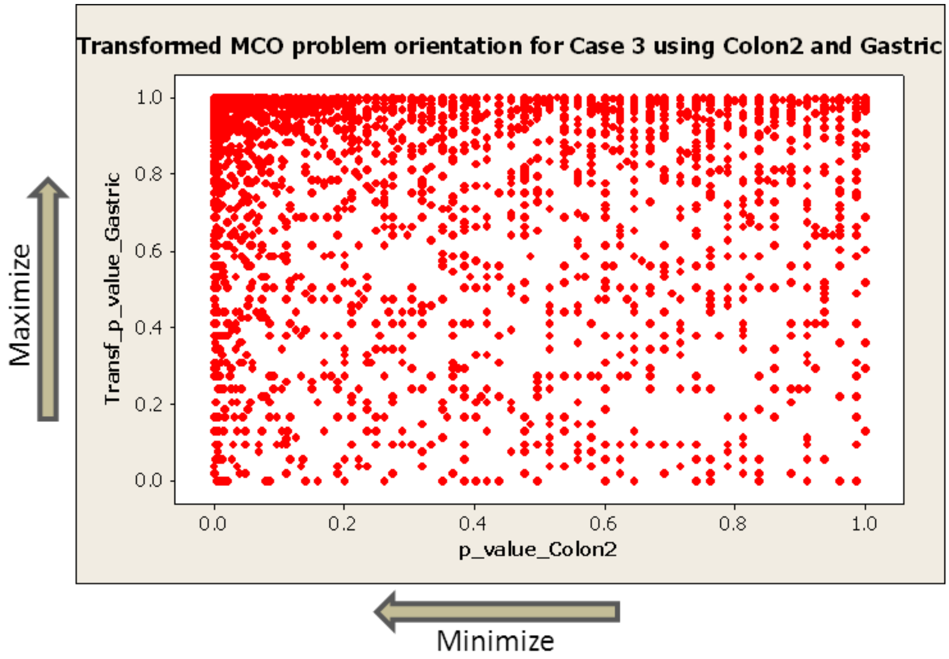


Figure 3-14. Transformed orientation of the MCO problem for Case 3 using Colon 2 and Gastric databases.

Table 3-14. Evolution of the amount of genes found in the first 10 efficient frontiers, of the MCO problem for Case 3 using Colon 2 and Gastric databases

Frontier	*Genes to analyze*	*BCCI*	*BCCO*	*Genes in common*	*Repeated genes*
1	2360	7	61	5	0
2	2355	6	59	5	0
3	2350	9	58	8	0
4	2342	8	54	8	0
5	2334	6	50	6	0
6	2328	11	51	10	0
7	2318	12	53	12	0
8	2306	13	47	13	0
9	2293	13	45	13	0
10	2280	14	41	13	0
			Total	***93***	

Results of Case 3 with Colon 2 and Gastric databases

The high mobility gene HMGB1 was found in the first frontier of this analysis, being this the fifth analysis were this gene shows relevance. Also CKMT2, CLEC3B and KRT9 appeared in this first frontier, along with the protein described in locus Z29574.

Persistent genes from along this analyses include the kinases CKS2 and CKS1B, VIP, HSPD1, SET, HMGB1. In this analysis HSP90AA1 was found, this gene belong to the same family than the reported HSPD1, the heat shock protein family. Among the genes found, AKT2 corresponds to one of the human homologues of v-akt, the transducted oncogene of the AKT8 virus which induces lymphomas in mice; its alterations have been reported to reflect poor prognosis in ovarian cancer patients (111).

The persistent genes across the different analyses will be considered for an extensive literature validation described in Chapter 4.

3.6.3 Case 3 using Colon 1, Colon 2 and Gastric Databases

This instance of Case 3 is executed using the three available databases. The overlap between the three databases was 674 genes. For this instance, MW comparisons are performed for each database to obtain their *p_values* (***p_valueColon1***, ***p_valueColon2*** and ***p_value_Gastric***). Transformation of that correspondinhg to Gastric was performed to obtain ***Transf_p_value_Gastric***. Finally the MCO problem is defined, considering the first two *p_values* as performance measures to be minimized and the third one to be maximized. Representation of this three dimensional MCO problem is shown in Figure 3-15. Similar to previous executions, the first ten frontiers found in the build MCO problem are explored. The number of genes found in each frontier is detailed in Table 3-15. Their corresponding list with complete information of the 222 genes found is given in Table 7 of Appendix C (C7).

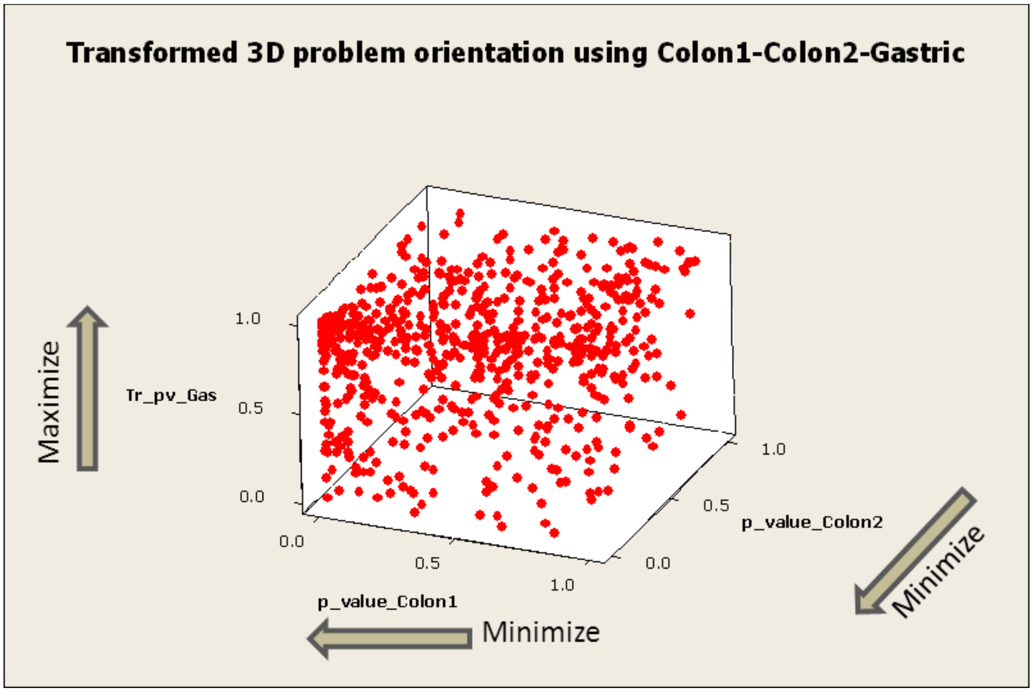


Figure 3-15. Reorientation of the MCO problem built for Case 3 using Colon1, Colon 2 and Gastric Databases.

Table 3-15. Evolution of the different 10 frontiers found for the Case 3 using Colon1-Colon2-Gastric databases.

Frontier	*Genes to analyze*	*BCCI*	*BCCO*	*Genes in common*	*Repeated genes*	*Accession of repeated genes*
1	674	6	27	6	0	
2	668	13	30	13	0	
3	655	15	27	15	0	
4	640	21	29	20	0	
5	620	23	27	22	0	
6	598	26	28	25	0	
7	573	24	28	24	0	
8	549	31	33	30	0	
9	519	36	37	35	1	M27749 (2)
10	483	32	34	32	0	
			Total	***222***		

Results of Case 3 with Colon 1, Colon 2 and Gastric databases

This list is the longest among the results presented, an extensive review of each of those genes found results impractical, but the similarities with other lists obtained can be easily analyzed. The first frontier for this last case, showed CKS CKS1B, VIP HSPD1 MYL9 and HMGB1. All of these six genes were also found for at least other two of the presented analyses, stressing their potential relevance in cancer. HSP90AA1 was also found, located in the second frontier; given its relation with HSPD1 (belonging to the same family), there could be biological insight to be explored in terms of genes relations across 3 analysis schemes. These kinds of hypothesis are left as future work.

Further analysis of the resulting genes through these seven experimental executions, is presented in Chapter 4. The different lists of genes obtained are first analyzed and grouped in order to focus the attention of the validation processes in those consistently present in different results.

4 ANALYSES AND VALIDATION

In this chapter different validation schemes are applied to the lists of genes obtained in this thesis. The first validation scheme compares the resulting potential biomarkers to those genes referenced as relevant in their original publications. The second validation scheme compares the behavior for these potential biomarkers across different databases. In this scheme if a specific gene is ***overexpressed*** (i.e. expressed higher in cancer than in healthy) or ***underexpressed*** (expressed lower in cancer than in healthy) in a particular database, it is verified that the same gene shows a consistent behavior in a different cancer database.

The third scheme corresponds to a validation based on literature review for those genes selected as having a high evidence of relevance using the proposed methodology. Gene function at the cellular level, functional group, as well as evidence of participation in the metabolism of cancer or other diseases are cited to support the selected genes as potential cancer biomarkers. It is important to note that an experimental biological validation goes beyond the efforts of this work.

Given the different validation schemes to be evaluated, the analysis of the genes consistently present in the different executions is presented in this chapter. The 7 different analyses executed and described in Chapter 3, are listed in Table 4-1.

Table 4-1. Reference of the different lists of genes obtained through the proposed method.

List	*Experimental Execution*	*Number of potential biomarker genes*
List 1	**Case 1** - Colon 1	27
List 2	**Case 1** - Colon 2	79
List 3	**Case 1** - Gastric	74
List 4	**Case 2** - Colon 1 - Colon 2	41
List 5	**Case 3** - Colon 1 - Gastric	85
List 6	**Case 3** - Colon 2 - Gastric	93
List 7	**Case 3** - Colon 1 - Colon 2 - Gastric	222

To describe the different subsets of genes analyzed in these validation procedures, a series of graphical analyses are presented in Figures 4-1 to 4-3.

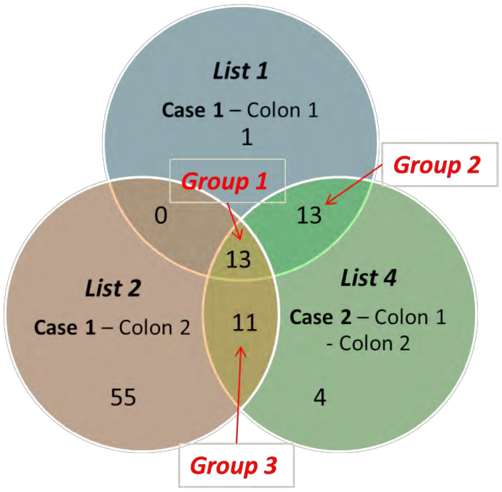


Figure 4-1. Venn diagram of the different analyses performed with Colon 1 and Colon 2 databases.

The Venn diagram in Figure 4-1, shows the interaction between the different sets of genes found with the cases involving Colon Cancer. The diagram shows an intersection between lists 1, list 2 and list 4 with 13 genes. These 13 genes are the most robust across the analyses involving colon cancer. This group of persistent relevant genes is called Group 1 as detailed in Table 4-2. The additional 13 genes in the intersection of lists 1 and 4 are the second most robust in this case are called Group 2. The intersection between lists 1 and 2 is called Group 3, and it contains 11 genes.

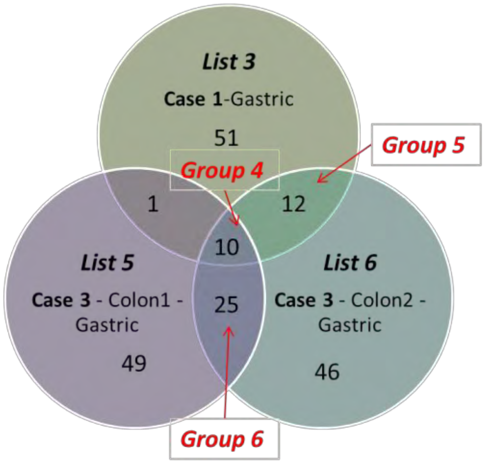


Figure 4-2. Venn diagram of the different analyses performed using Gastric database.

Figure 4-2 shows the Venn diagram of the analyses performed considering Gastric cancer. Ten genes are strongly evidenced to be potential biomarkers by being in the intersection of the three lists. These are called Group 4. The remaining intersection, then, receive the next consideration in priority and are called Groups 5 and 6 with 12 and 25 genes, respectively. The intersection between list 3 and list 5 is not considered as a group since it is just one gene, however, its validation under the second scheme is executed.

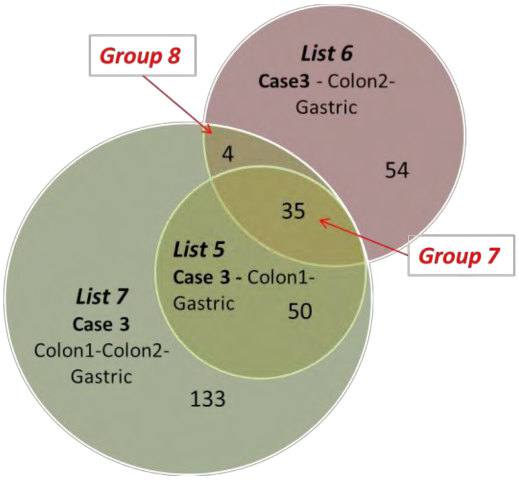


Figure 4-3. Venn diagram considering Cases type 3.

Figure 4-3 shows the set of genes found through the consideration of Lists 5, 6 and 7. Thirty five genes are located in the intersection of these lists, rendering them the most important to look up. This group is referred as Group 7, the remaining 4 genes that make up the intersection between lists 6 and 7 are called Group 8. All groups are detailed in Table 4-2.

Table 4-2. Groups built with intersections between lists of results

Group	*Intersection*	*Description*	*Number of potential biomarkers*
1	*List1-List2-List4*	Cases 1 and 2 with Colon Databases	13
2	*List1-List4*	Two Cases 1 using Colon DBs	13
3	*List2-List4*	One Case 1 and One Case 2 with Colon DBs	11
4	*List3-List5-List6*	Cases 1 and 3 with Gastric Database	10
5	*List3-List6*	One Case 1 and One Case 3 with Gastric DB	12
6	*List6-List5*	Two Cases 3	25
7	*List5-List6-List7*	Executions of Case 3	35
8	*List6-List7*	Two Cases 3	4

Due to their relevance, the chosen groups of potential biomarkers are now subjected to the validation schemes previously defined. The results are discussed next.

4.1 Validation scheme 1: Comparison against existing genetic signatures

When a specific genetic signature was reported in the original publication, our results are compared against it. In this first scheme only Lists 1, 2 and 3 are considered.

4.1.1 Validation of Case 1 using Colon 1 database

In this case, the intention of the original publication (37) was not to find a specific list of relevant genes; but differentiated patterns of genes depending upon their function. However, these data have been used in several works to generate gene signatures to characterize colon cancer. Eight of those works were selected for comparison (97-101,112,102,113). In six of them (98,99,101,112,102,113) the purpose was tissue classification while the other two (97,100) identified potential biomarker genes although without any validation through classification.

The 27 genes from Case 1 using data from Colon 1 are presented in Table 4-3, where the reported frontier localization, the accession number, symbol and gene name are presented. The last column contains those references within the eight selected where the same gene was identified as relevant for their analysis.

Table 4-3. Comparison of the selected genes using the proposed methodology against existent references using the same dataset.

	Frontier	*Accession*	*Symbol*	*Name*	*References*
Overexpressed genes	3	M26383	DARS	interleukin 8	[90][84][91][86][85][87][89]
	2	H08393		WD repeat domain 77	[90][84] [86][85][87][89]
	3	J05032	GTF3A	aspartyl-tRNA synthetase	[90][84][85][87][89]
	2	R36977	WDR77	General transcription factor IIIA	[84][86][85][87]
	4	X63629	IL8	cadherin 3, type 1, P-cadherin (placental)	[90][84][87][89]
	5	X12671		heterogeneous nuclear ribonucleoprotein A1	[90][85][87][89]
	10	T47377	CFD	S100 calcium binding protein P	[90][86][87]
	1	M22382	HSPD1	Heat shock 60kDa protein 1 (chaperonin)	[84][87][89]
	4	H40095	CDH3	EST: yn85b03.s1	[84][87][89]
	8	X14958	HNRNPA1	high mobility group AT-hook 1	[90][87]
	6	U09564	GUCA2B	SFRS protein kinase 1	[87][89]
	7	T86473	DES	Non-metastatic cells 1, protein (NM23A) expressed in	[87]
	8	R84411	MYL9	Small nuclear ribonucleoprotein polypeptides B and B1	[87]
	9	M26697	SRPK1	Nucleophosmin (nucleolar phosphoprotein B23, numatrin)	[87]
	10	U30825	CSRP1	Serine/arginine-rich splicing factor 9	[87]
	10	D31885	VIP	ADP-ribosylation factor-like 6 interacting protein 1	[87]
Underexpressed genes	1	R87126	NME1	EST: yq31b10.s1	[90][84][86][85][87][89][88]
	5	J02854	HMGA1	myosin, light chain 9, regulatory	[90][84][86][85][87][89][88]
	4	Z50753	GSN	guanylate cyclase activator 2B (uroguanylin)	[90][84][86][85][87][89]
	6	M76378	NPM1	cysteine and glycine-rich protein 1	[90][84][86][85][87][89]
	4	M63391	SNRPB	desmin	[90][84][85][87][89][88]
	9	T92451	S100P	Tropomyosin 2 (beta)	[90][84][87][89][88]
	10	X86693	ARL6IP1	SPARC-like 1 (hevin)	[90][84][87][89][88]
	7	M36634	MT1G	vasoactive intestinal peptide	[90][84][87][89]
	8	H06524	SPARCL1	Gelsolin	[90][87]
	9	T71025	SRSF9	Metallothionein 1G	[90][87]
	6	H43887	TPM2	Complement factor D (adipsin)	[87]

An interesting overlap of results is presented with Chen et al. in (100) where all the 27 genes coincide with our results. In (100), the authors noted discrepancies between different criteria considered for gene selection and propose the use of ranking algorithms. To the best of our knowledge, this publication (100), represents the closest attempt to address the opportunity areas cited in this thesis, represent a good starting point for the efforts of our work.

Many of the 27 genes under scrutiny have a scientific literature reference that evidences their role in cancer. That is the case of Interleukin-8 (IL-8) which, besides being selected by our method, it was reported in 7 of the 8 references presented. IL-8 has been recently shown to contribute to the progression of human cancer through its potential functions as mitogenic, angiogenic and motogenic factor (85). The NME1 gene is known as one of thirteen identified tumor metastasis suppressor genes (114); reduced expression of their correspondent family (NME) was associated with increased metastatic potential and more aggressive disease in human breast (115), hepatocellular, ovarian and gastric carcinoma and melanoma as identified in (116-119) through the review of (120). NME1 has also been associated with a poor prognosis in Acute Myeloid Leukemia (AML) (121).

4.1.2 Validation of Case 1 using Colon 2 database

The original reference for Colon 2 database (57) identified 66 significant genes; 19 of them overexpressed and 47 underexpressed. The single gene overlapping with our results, with accession number L11708 named hydroxysteroid (17-beta) dehydrogenase 2, has been confirmed as a potential biomarker gene for breast cancer (122,123).

4.1.3 Validation of Case 1 using Gastric database

For this case, 162 genes were reported as overexpressed and 129 as underexpressed in the original work (83). Our list of 74 selected genes present an overlap of 32 genes, 9 of them overexpressed and 23 underexpressed. Table 4-4 present the list of the genes in common between the original publication and the list obtained in this work.

Table 4-4. List of the genes obtained using the proposed methodology that are also present in the original publication.

<table>
<tr><th colspan="4"><em>9 overexpressed genes overlapping between execution of Case 1 using Colon 1 and its original reference</em></th></tr>
<tr><th><em>Frontier</em></th><th><em>Accession</em></th><th><em>Symbol</em></th><th><em>Name</em></th></tr>
<tr><td>5</td><td>X54942</td><td>CKS2</td><td>CDC28 protein kinase regulatory subunit 2</td></tr>
<tr><td>5</td><td>X54941</td><td>CKS1B</td><td>CDC28 protein kinase regulatory subunit 1B</td></tr>
<tr><td>5</td><td>X54667</td><td>CST4</td><td>cystatin S</td></tr>
<tr><td>5</td><td>D21063</td><td>MCM2</td><td>minichromosome maintenance complex component 2</td></tr>
<tr><td>9</td><td>L40379</td><td>TRIP10</td><td>thyroid hormone receptor interactor 10</td></tr>
<tr><td>10</td><td>Z74616</td><td>COL1A2</td><td>Collagen, type I, alpha 2</td></tr>
<tr><td>10</td><td>Z74615</td><td>COL1A1</td><td>Collagen, type I, alpha 1</td></tr>
<tr><td>10</td><td>X74801</td><td>CCT3</td><td>Chaperonin containing TCP1, subunit 3 (gamma)</td></tr>
<tr><td>10</td><td>M86752</td><td>STIP1</td><td>Stress-induced-phosphoprotein 1</td></tr>
<tr><th colspan="4"><em>23 underexpressed genes overlapping between execution of Case 1 using Colon 1 and its original reference</em></th></tr>
<tr><td>2</td><td>AC002077</td><td>GNAT1</td><td>guanine nucleotide binding protein (G protein), alpha transducing activity polypeptide 1</td></tr>
<tr><td>3</td><td>Z29574</td><td>TNFRSF17</td><td>tumor necrosis factor receptor superfamily, member 17</td></tr>
<tr><td>4</td><td>U57094</td><td>RAB27A</td><td>RAB27A, member RAS oncogene family</td></tr>
<tr><td>4</td><td>M75110</td><td>ATP4B</td><td>ATPase, H+/K+ exchanging, beta polypeptide</td></tr>
<tr><td>4</td><td>M63154</td><td>GIF</td><td>gastric intrinsic factor (vitamin B synthesis)</td></tr>
<tr><td>4</td><td>M62628</td><td></td><td>Human alpha-1 Ig germline C-region membrane-coding region, 3' end.</td></tr>
<tr><td>5</td><td>X76223</td><td></td><td>H.sapiens MAL gene exon 4</td></tr>
<tr><td>5</td><td>X53961</td><td>LTF</td><td>lactotransferrin</td></tr>
<tr><td>5</td><td>X05997</td><td>LIPF</td><td>lipase, gastric</td></tr>
<tr><td>5</td><td>U70663</td><td>KLF4</td><td>Kruppel-like factor 4 (gut)</td></tr>
<tr><td>5</td><td>U19948</td><td>PDIA2</td><td>protein disulfide isomerase family A, member 2</td></tr>
<tr><td>5</td><td>M63962</td><td>ATP4A</td><td>ATPase, H+/K+ exchanging, alpha polypeptide</td></tr>
<tr><td>5</td><td>M61855</td><td>CYP2C9</td><td>cytochrome P450, family 2, subfamily C, polypeptide 9</td></tr>
<tr><td>5</td><td>D63479</td><td>DGKD</td><td>diacylglycerol kinase, delta 130kDa</td></tr>
<tr><td>5</td><td>D26129</td><td>RNASE1</td><td>ribonuclease, RNase A family, 1 (pancreatic)</td></tr>
</table>

6	Z48314	MUC5AC	mucin 5AC, oligomeric mucus/gel-forming
6	X51698	TFF2	trefoil factor 2
8	J05412	REG1A	regenerating islet-derived 1 alpha
8	D14695	HERPUD1	homocysteine-inducible, endoplasmic reticulum stress-inducible, ubiquitin-like domain member 1
9	S76942	DRD4	dopamine receptor D4
10	X65614	S100P	S100 calcium binding protein P
10	X52003	TFF1	Trefoil factor 1
10	M12759		Human Ig J chain gene, exons 3 and 4

The pair of kinases, CKS2 and CKS1B selected by our methodology was also shown in the list of relevant genes of the original paper. Also two collagen proteins (COLA1 and COLA2) were detected by both methods. KLF4 belongs to the same family of KLF9 which is validated through literature in section 4.3.

Even though the overlap between the reported list and the results shown by the proposed method is just of 32 genes of 291, the original publication does not show any biological validation of their list.

4.2 Validation scheme 2: Comparison of expression profiles from different databases

This validation process is based on the evaluation of the direction of change in expression of the selected genes. A gene that is shown to be significantly underexpressed or overexpressed by any method can be verified to keep this behavior consistent in an independent database.

This validation process is performed for all the groups defined before for intersections between two or more sets of selected genes, illustrative example of these profiles can be seen in Figure 4-4, where the direction of expression change is drawn for each of the databases.

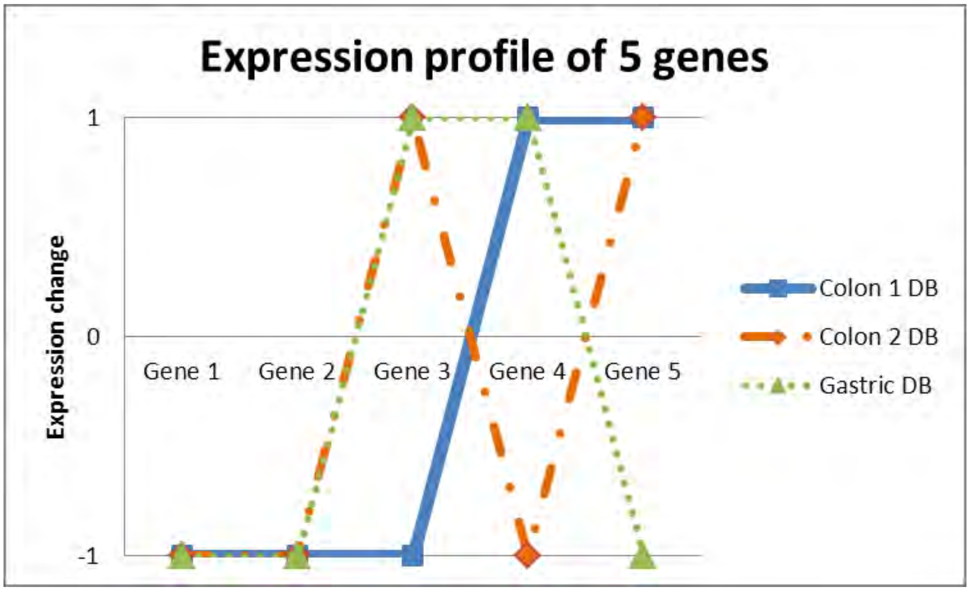


Figure 4-4. Example of the Expression profile of 5 genes with validation purposes.

These profiles can be read as follows: level +1 represents a positive change in expression from healthy state to cancer state (an overexpressed gene), level -1 represents a negative change in expression from healthy to cancer states (an underexpressed gene). This change in expression is measured between the values of the medians in the different states. Different lines represent the expression explored in different databases. If the profiles follow the same pattern, the original profile is validated, if not, further analysis is advised. In Figure 4-4 an example of 5 genes is presented, solid line represents the changes in expression for the five different genes when it is analyzed using the Colon 1 database. A segmented line describes the changes in expression of the same five genes considering data from Colon 2 and the dotted green line represents the expression profile for the 5 genes when using data from the Gastric database. These profiles would validate the behavior of two of the five genes explored, because the third gene is underexpressed in Colon 1 while overexpressed in Colon 2 and Gastric. The 4th gene also presents discrepancies, being overexpressed in Colon 1 and Gastric databases and underexpressed in Colon 2.

This analysis can be performed for all the intersections between two or three databases. These intersections can be easily visualized in Figures 4-1, 4-2 and 4-3. Their description can be consulted in Table 4-2.

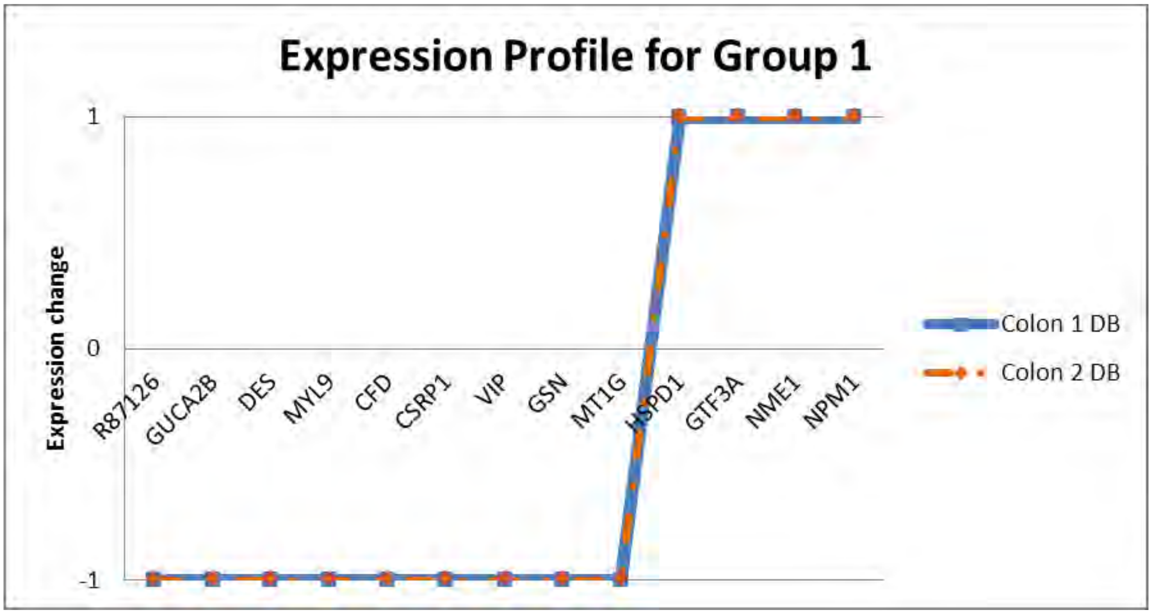


Figure 4-5. Expression profile for Group 1, correspondent to the intersection of analyses performed just with databases from Colon cancer.

Figure 4-5 shows the expression profile for Group 1, where the symbols on the x-axis corresponds to the genes. For this profile, Colon 1 was used as reference. The behavior is completely consistent across both databases, 9 of these genes were reported as underexpressed and 4 of them as overexpressed. The Desmin (DES) protein is associated with one of the underexpressed genes in this profile; previous evidence has been reported of the downregulation of this protein being relevant for cancer development (124). Among the overexpressed genes, the NME1 is found as consistently overexpressed in both databases. The reported relationship between this gene and cancer metabolism calls for an underexpression to enhance the metastasis process (125). This discrepancy with what was detected in the databases must be investigated in the future.

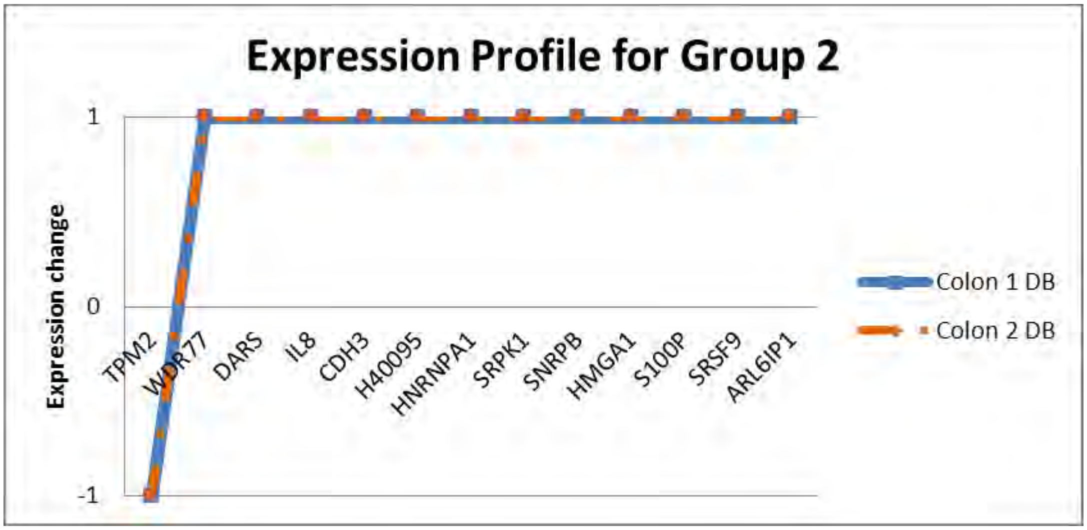


Figure 4-6. Expression Profile 2, representing the intersection of analyses 1 and 4.

Among the genes detected as common between Lists 1 and 4 (genes in Group 2), one of the 13 genes is presented as underexpressed, Tropomyosin 2-beta (TPM2), while the other 12 were shown overexpressed. Among the overexpressed genes, Interleukin-8 (IL8) has been already reported as relevant for cancer (85). The behavior of these genes is also consistent considering both colon databases.

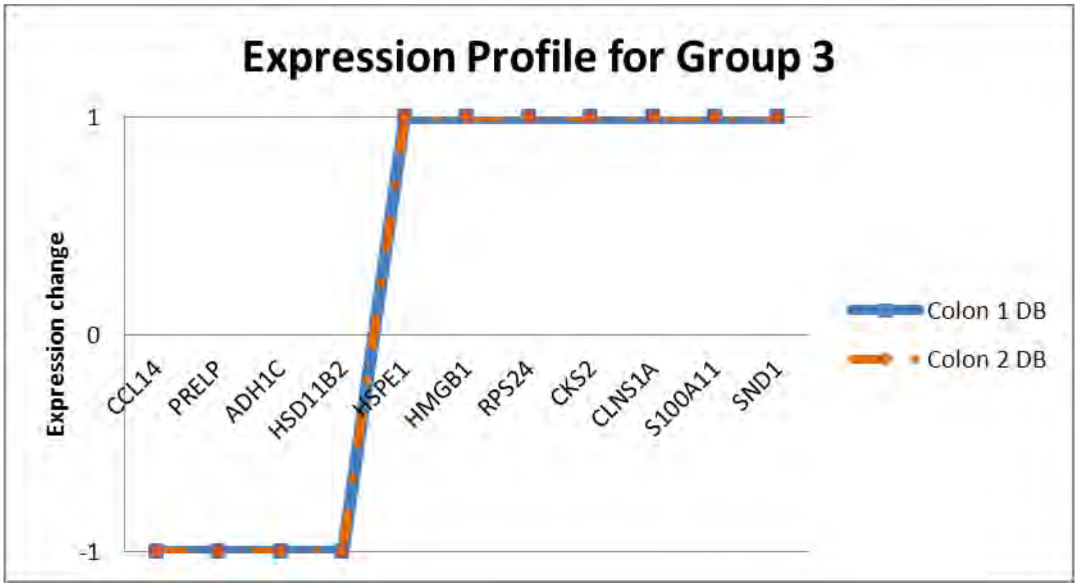


Figure 4-7. Expression profile 3, correspondent to the intersection of analyses 1 and 2.

In the expression profile for Group 3 (intersection between lists 2 and 4), 11 genes in common were identified. Four of them are underexpressed and 7 overexpressed; all of them matched across both databases.

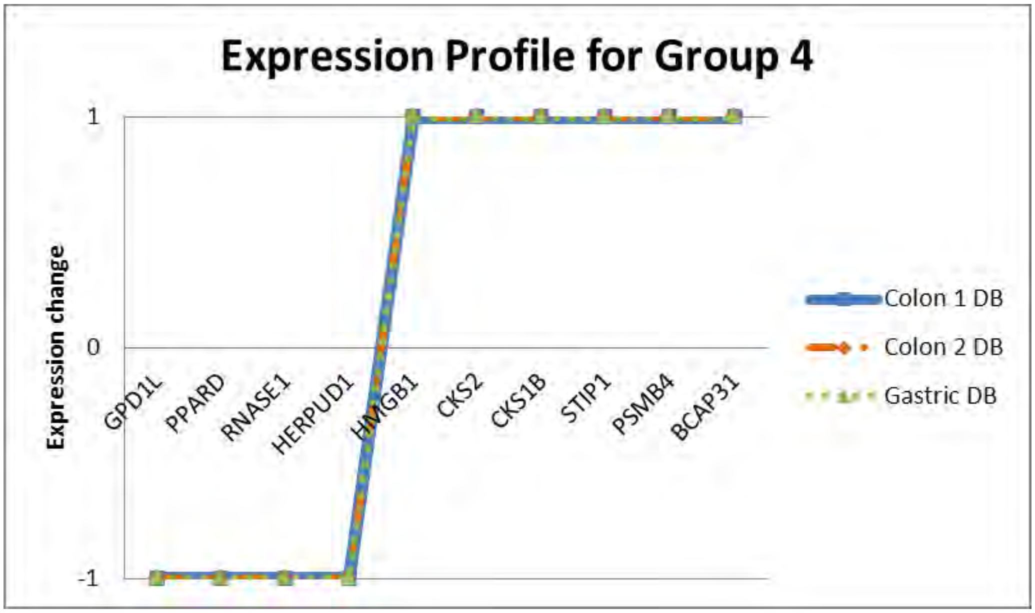


Figure 4-8. Expression profile of Group 4, representing the overlapping of lists 3, 5 and 6.

In the profile for Group 4, all the involved genes kept the same expression change behavior across the three databases. Among the overexpressed genes, there are two kinases (CKS2 and CKS1B) that have been referred as potential actors in cancer development. RNASE1 and HERPUD1, identified among the 4 underexpressed genes are already identified as potential contributors in cancer.

In the expression profile for Group 5, four genes were consistently underexpressed (TNFRSF17, ARP4B, X76223, GNB3) and three of them consistently overexpressed (PTPN12, SMS, CST4) when considering Colon 2 and Gastric databases. The other five genes (KRT9, CKMT2, ATP4A, MLLT3, FBP1), show a different change in expression when considering data from one database or the other. Although discrepancy is undesirable from the analytical point of view, there is evidence for different cancer types expressing the same gene differently. The use of these profiles facilitate detecting such discrepancies and explore the reasons behind them. When a discrepancy is detected, *p_values* from the MW test when comparing normal to cancer states using data from each database are evaluated; if the *p_value* for the discrepant database is not significant, the conflict is ignored for the next validation scheme.

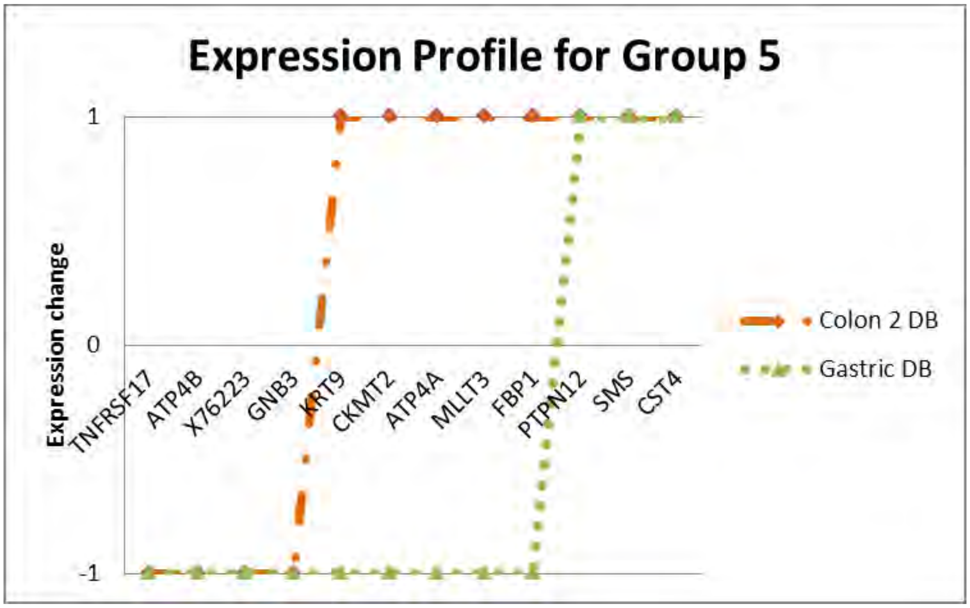


Figure 4-9. Expression profile 4, representing the overlapping between analyses 3 and 6.

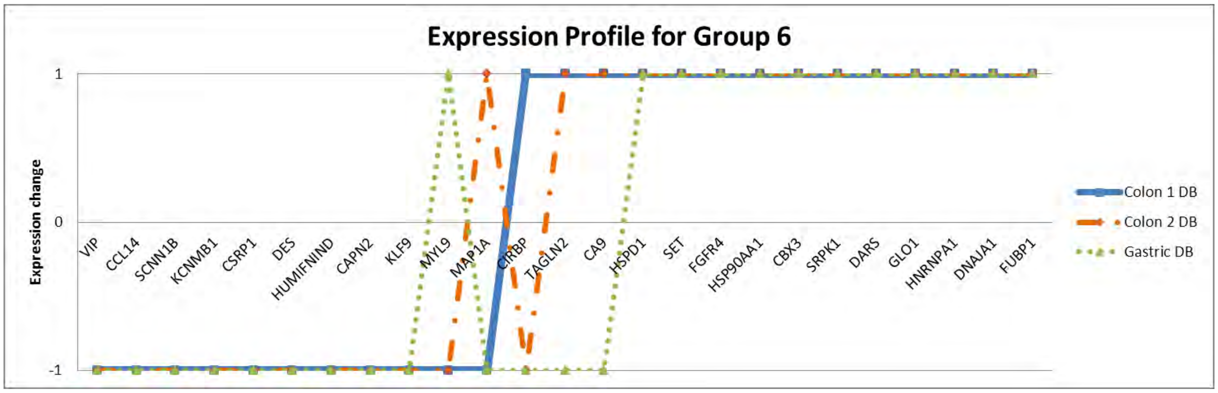


Figure 4-10. Expression profile 6, to analyze the overlapping between analyses 5 and 6.

The profile for group 6 is the longest and it shows the highest number of discrepancies. Nine of the original 25 genes are consistently underexpressed in the three databases, and 11 genes were consistently overexpressed. Gene MYL9 was underexpressed in both Colon cancer databases while overexpressed in gastric database, MAP1A is underexpressed in Colon 1 and Gastric databases but not in Colon 2. CIRBP is underexpressed in Colon2 and Gastric but not in Colon 1; and TAGLN2 is overexpressed both Colon databases and underexpressed in Gastric. These discrepancies should also be explored for biological explanation.

The single gene in common between lists 3 and 5 is named ACTN1 (Actinin, alpha 1). This gene was underexpressed in Colon 1 while overexpressed in Colon 2 and Gastric databases (see Appendix D, Figure set 8).

Even though Group 7 was composed by the 35 genes in the intersection between List 5, 6 and 7, its genes are also present in the intersections of Groups 4, 5 and 6, which was already described, thus, no further exploration was necessary. The 4 genes in intersection of List 6 and List 7 (Group 8) are not found in any other profile. Their behavior is detailed in Figure 4-11.

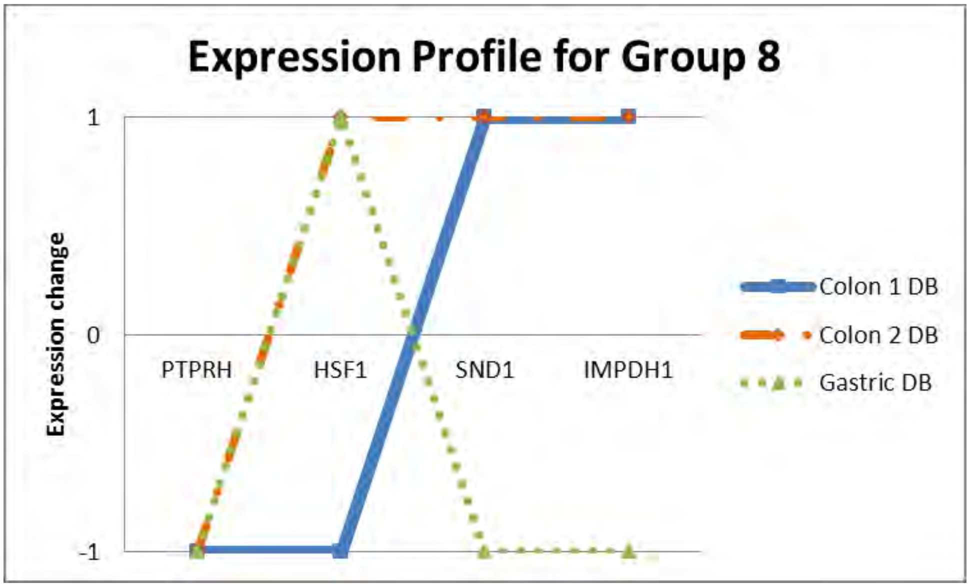


Figure 4-11. Expression profile for group 8, representing the intersection between analyses 6 and 7.

There is coincidence in the expression behavior for three genes in Figure 4-11: PTPRH (underexpressed), SND1 (overexpressed) and IMPDH1 (overexpressed). HSF1 is overexpressed in Colon 2 and Gastric databases and underexpressed in Colon 1 database.

In Appendix D, all individual value plots to evaluate the change in expression for each gene graphically are presented to support the generation of each profile.

4.3 Validation Scheme 3: Literature review

This validation procedure consists on searching for the role of the selected genes in the existent literature. The sources for this information were the Gene Expression Omnibus (GEO) repository, the Information Hyperlinked over Proteins (iHOP) site (126), the geneBank repository as well as a direct search of articles that cite the gene as a cancer biomarker. A summary with functions and interactions associated to cancer development was built with references related to the study of the involved gene,. For this evaluation in literature only those genes located in the intersection of three sets are considered. First the 13 genes conforming Group 1 (Figure 4-1), then, the 10 genes corresponding to Group 4 (Figure 4-2) are analyzed. Finally the 35 genes in the intersection of the analyses of the different executions of Case 3 (Group 7 in Figure 4-3), are analyzed. Among the 35 genes in Group 7, 15 of them were also present in Group 1 or Group 4, leaving just 20 new genes to be considered for analysis. Considering 13 genes from Group 1, 10 from Group 4 and the remaining 20 from Group 7 that were not listed previously, a total of 43 genes are subjected to this validation procedure.

The *p_value* analysis described in the last section for discrepant genes found through the expression profiles was performed for all 43 genes selected. Through this analysis, some of the genes presented as conflicting using profiles are shown as non-conflicting here. *P_values* and individual value plots using each database are presented in Apendix E for most of the 43 selected genes described in this section.

Functions of the genes under scrutiny are shown in Tables 1F, 2F and 3F (Appendix F) section. Main functions are summarized in Tables 4-6, 4-7 and 4-8. In these tables, characteristics of the genes contributing to processes of normal cell development and cancer development are cited. The main cellular processes considered were cell proliferation, metastasis and apoptosis, all other functions were clustered in the column "others". The related cancer types and other related diseases are described, for the functional comparison. The respective bibliographical references are cited directly on the table.

Table 4-5. Summary of literature review of the 13 genes from Group 1, intersection of cases mainly related to colon cancer.

	Cell processes				***Functional Comparisons***	
Gene Symbol	***Cell Prolifeation***	***Metastasis***	***Apoptosis***	***Others***	***Related Cancer types***	***Other related diseases***
			Overexpressed in all the databases			
HSPD1	(127,86)				Pancreatic (127,86,128-130)	Hereditary Spastic Paraplegia(128)
GTF3A	(126)				Down syndrome-associated Acute Myeloid Leukemia (87)	
NME1	(131)	(131)			Breast (125), ovarian (131),	
NPM1	(126)				AML(88,132)	
			Underexpressed in all the databases			
EST: yq31b10.s1						
GUCA2B						
DES		(89)			CRC (133), rhabdomyosarcoma (124,89)	Desmin-related myophaty (134)
MYL9		(126)				
CFD						
CSRP1				(135)	Hepatocellular carcinoma (136)	
VIP					Breast (135,90)	

GSN		(126)	All (Tumor activator) (137)	Hepatitis B-associated liver cirrosis (138)
MT1G	(126)		Prostate (91), hepatoblastoma (139) renal (140)	

Referring to the genes in Table 4-5, 10 of 13 genes were found in the literature relating to the cell processes described. CFD and GUCA2B do not have any cancer-related reference and EST: yq31b10.s1 is an expressed sequence tag without any reported function yet. A description of the relevant genes is presented in the following paragraphs, where references found for each specific gene are briefly described.

Gene HSPD1, also referenced as HSP60, is known as coding for the heat shock 60kDa protein, and it is one of the four principal chaperonin proteins reported in Table 4-6. This family is important for cell signaling and protein traffic in the presence of stress(129). After a cellular assault, the need for this kind of proteins increases markedly as a defense mechanism to allow cells to survive otherwise lethal conditions (127). This gene has been reported as consistently high in different stages of prostate cancer (86), and thus been proposed as potential biomarker for this cancer type (130). In our results, this gene is overexpressed, supporting the literature evidence of its potential as biomarker.

The overexpression of the General transcription factor IIIA (GTF3A) was successfully validated as related to Down Syndrome Acute Myeloid Leukemia (87). GTF3A is reported as overexpressed in our results, supporting the AML evidence.

NME1 is already known as a metastasis suppressor gene, Single nucleotide polymorphisms (SNP) in the promoter region of the NME1 gene has been found to be associated with breast cancer prognosis (125). An inverse relationship between its expression and metastasis potential has been observed for some solid tumors (131). In our results, it was detected as overexpressed in cancer tissues. This behavior could be interpreted as the non-metastasic stage of the involved samples in the study, however further investigation is needed to clarify it.

Three of the genes (HSPD1, GTF3A and NME1) are already used in patented genetic signatures for the diagnosis or prognosis of different cancer states (105-107).

Mutations in nucleophosmin NPM1 represent the most frequent molecular aberrations in adult patients of acute myeloid leukemia (AML). Its mutation has been proved as a useful marker for minimal residual disease (MRD) in this cancer type (132). It has been demonstrated that patients with mutations in NPM1 show favorable prognostic (88). There is no detail of the direction of expression change that is generated by such mutation. This information would be useful to evaluate if the overexpression change found in our analysis is consistent with other findings.

The gene Desmin (DES) has been proved as a specific marker for rhabdomyosarcomas (89), however, the direction of the expression change in this cancer type follow a different direction than that found in this work (underexpressed). This gene has also been proposed as a potential oncofetal serum tumor marker for Colorectal Cancer (CRC) where its overexpression has been proved correlated to the presence of the disease (133). However, underexpression of Desmin in other cancer types, besides rhabdomyosarcomas and CRC, correspond to underexpression, coinciding with the evidence shown by the data used here.

For myosin, light chain 9 regulatory (MYL9) a gene involved in cell locomotion, there is no biological validation for its relevance in cancer, however it has been found as significantly expressed in different data-based analyses (141,142). The situation is similar for CFD, the Complement factor D (adipsin) and GUCA2B (guanylate cyclase activator 2B (uroguanylin)) is a similar case (145) are reported in data-based analysis with no biological validation (98,143,144).. Both genes, CFD and GUCA2B, were underexpressed in the used data.

The expression of CSRP1 (cysteine and glycine-rich protein 1) has been explored in hepatocellular carcinoma, concluding that it was inactivated (underexpressed) in cancerous cells by aberrant methylation. This gene and CAV1 may serve as important biomarkers of this malignancy (136). This underexpression coincides with the behavior shown in our results.

A hybrid of the vasoactive intestinal peptide (VIP), which in its basal form is involved in vasodilation, has been shown to inhibit breast cancer development (135,90). Even though the basal form of VIP is not reported as relevant for cancer, it is here proposed for further investigation according to our results.

Gelsolin (GSN) is an actin binding protein that modulates a variety of physiological process by interacting differently with the actin cytoskeleton. It has been shown that underexpression of gelsolin is present in several types of cancer cells. It significantly reduces the invasive and motile properties of cells, and it has been reported as underexpressed in some cancer types like lung and bladder (146,147). Its behavior is confirmed as a tumor activator in vitro, although further investigations concerning the role of these gene in tumor progression in vivo are pending (137).This information supports our findings regarding Gelsolin being underexpressed in colon cancer.

MT1G (Metallothionein 1G) has been validated with associated with tumor aggressiveness in prostate cancer and might be a marker of locally advanced disease (91). This means that it is underexpressed in tumors. Its hypermethylation has been also proposed as a potential prognostic marker for hepatoblastoma (139). Its reported behavior perfectly matched our findings.

Additional literature review was performed to validate the 10 gene set reported in Group 4. Table 4-6 summarizes this information.

Table 4-6. Summary of literature review of the 10 genes for Group 4 cases mainly related to gastric cancer.

	Cell processes				***Functional comparisons***	
Gene Symbol	***Cell Proliferation***	***Metastasis***	***Apoptosis***	***Others***	***Related Cancer types***	***Other related diseases***
			Overexpressed in all the databases			
HMGB1	(148)				Gastric (149), Colon (150), breast (151), Melanoma(152),	

			AML (153)	
CKS2	(126)		Bladder (154) Cervical (155) Breast (156)	
CKS1B	(126)		Multiple Myeloma (157), breast (158)	
STIP1	(159)		Ovarian (159)	
PSMB4		(126)		
BCAP31		(160)		
Underexpressed in all the databases				
GPD1L			Bladder (161)	Brugada syndrome (162) some cardiac diseases (163)
PPARD	(164)		Colon (165)	
RNASE1			Pancreas(166)	
HERPUD1			Prostate(167,168)	

Referring to Table 4-6, all genes had functionality related to cell development reported in the literature. A description for each gene is presented below.

HMGB1 has been implicated in a variety of biological in important processes. It has been reported to contribute to cellular signaling, cellular migration and tumor invasion. Increased expression of HMGB1 has been reported for several differential tumor types, including breast carcinoma, melanoma, gastrointestinal stromal tumors and acute myeloblastic leukemia as reported by (150) by review of (153,151,149,152). This gene has been reported as overexpressed, which matches our results.

The mus musculus (mouse) homologue of Cyclin-dependent kinase subunit 2 (Cks2) has been identified as a transcriptional target downregulated by the tumor suppressor p53. P53 is a tumor suppressor protein that is a principal factor in regulation of growth arrest as well as apoptosis. It is known to be mutated in the majority of human tumors and acts by engaging in complexes with other proteins or functions (148). CKS2 expression has been reported as strongly correlated to bladder (154), breast (156) and cervical (155) cancers. CKS2 is proposed to be downregulated by p53. When p53 is not working, CKS2 would have an increase in expression. Our analysis of colon cancer detects, indeed, an overexpression for CKS2.

CKS1B belongs to the family of the cyclin kinase subunit (CKS1), which interacts with cyclin-dependent kinases and plays an important role in cell cycle progression. Some authors have referred to this protein as an adverse prognostic factor in multiple myeloma (157). Its overexpression has been also related to poor overall survival in human breast cancer (158). Supported by our analyses, the overexpression of this gene can be proposed as a biomarker for colon cancer too.

A referred biomarker for ovarian cancer that promotes cancer cell proliferation is the Stress-induced-phosphoprotein 1 (STIP1) (159). Its overexpression is found in our analyses, posing the hypothesis of STIP1 being a biomarker for colon cancer.

The proteasome is responsible for the degradation of all short-lived proteins and 70/90% of all long-lived proteins, regulating processes such as cell cycle progression, DNA transcription, angiogenesis, DNA repair/misrepair, apoptosis/survival, among others (169). PSMB4 is a subunit of the proteasome and has been detected by our method, as overexpressed. Given that any direction of the change in this protein would result in the described cancer-related processes, the overexpression in our analyses results suggests its relevance for Colon and Gastric cancers.

Gene BCAP31 (B-Cell receptor associated protein 31) is a tumor suppressor gene. It is an integral protein of the endoplasmic reticulum membrane and substrate of caspase-8, a known regulator of apoptosis (160). It is one of 17 genes located in chromosome X related to apoptosis and has been studied to explain the excesses of cancer risk in males (170). The overexpression for BCAP31 in our results contradicts the literature evidence. This is left for further exploration in the literature.

Mutations of Glycerol-3-phosphate dehydrogenase 1-like (GPD1L) have been related to some cardiac diseases (163) and Brugada syndrome (162). Its behavior, like that in other lipogenic enzimes has been explored in Bladder cancer (161). In our results, this gene is underexpressed, however, the reported change in the literature was of overexpression. This discrepancy is also suggested to further analysis.

Peroxisome proliferators-activated receptors (PPARs) are members of the nuclear receptor superfamily and have three different isoforms: PPARα, PPARδ, and PPARγ. PPARs. These are ligand-activated transcription factors implicated in tumor progression, differentiation, and apoptosis. Activation of PPAR isoforms lead to both anticarcinogenesis and anti-inflammatory effect (164). Thus, an underexpressed behavior is expected for this kind of proteins in cancer state as was reported in our findings for the PPARδ (PPARD) gene. Its contribution for a specific cancer type is hypothesized in (165). The authors' hypothesis is supported by results presented in (164). Underexpression is then, a match for our results.

For the case of RNASE1, there exists some evidence to consider this human ribonuclease as a possible tumor marker for pancreatic cancer. In (166) authors report that the elevated serum RNASE levels in patients with pancreatic cancer are due to the tumor cells, raising the possibility to use human serum RNASE1 as tumor marker for this cancer type. This gene was detected as underexpressed in all databases used. However, this discrepancy could be explained by other unexplored biological relations that can be the subject of further analyses.

HERPUD1 is involved in the endoplasmic reticulum (ER) stress response pathway, its underexpression has been analyzed as correlated to prostate cancer, suggesting the involvement of the ER stress pathway in prostate tumorigenesis (167). Henriksen et al. in (168) conclude that its underexpression in prostate cancer predicts the occurrence of metastases almost perfectly. This evidence coincides with the behavior of this gene when using colon and gastric databases in our analysis.

Table 4-7 describes the characteristics considered for genes in Group 7. As it has been explained, just 20 of them are new for evaluation, because 15 were reported in the analyses of Groups 1 and 4.

Table 4-7. Summary of literature review of the 20 genes of Group 7; intersection of instances for Case 3.

	Cell processes				***Functional comparisons***	
Gene Symbol	***Cell Proliferation***	***Metastasis***	***Apoptosis***	***Others***	***Related Cancer types***	***Other related diseases***
			Overexpressed in all the databases			
HSP90AA1	(171)				Lung (171),	
CBX3	(126)					
SRPK1	(172)				Breast, Colon, Pancreas (95)	
FGFR4	(173)	(174)			Breast (174) melanoma (173) prostate (175) lung (176) soft tissue sarcoma (177)	
DARS				(178)		
HNRNPA1	(179)				CRC(180), lung (181)	
FUBP1	(182)				NSCLC (182)	
SET	(183)				AML (184)	
DNAJA1	(185)				Glioblastomas (186)	

GLO1				(187)	Leukemia (188), prostate (187)
Overexpressed in Colon DBs and underexpressed in Gastric DB					
TAGLN2		(189)			
CA9	(108)				Kidney (190) cervix (191,192)
Underexpressed in all the DBs					
CIRBP	(193)				
CCL14					
SCNN1B					Renal Clear cell carcinoma (140)
KCNMB1					
KLF9	(194)		(195)		Endometrial and breast (196), Colon (194)
HUMIFNIND					
CAPN2			(109)		Prostate (109), breast (110)
Underexpressed in Colon 1 and Gastric and overexpressed in Colon 2					
MAP1A			(197)		

Among the genes summarized in Table 4-7, 16 of the original 20 have biographical references of their relevance in cancer related processes. Only four of them (CBX3, DARS, SET, HUMIFNIND) were not referred as having a relevant role in those kinds of processes. A brief description found for them using iHOP (126) or Entrez descriptions from GEO are presented here. Details for each of the 16 genes found are described next.

HSP90AA1, also known as HSP90, belongs to the heat shock protein family (such as HSP60 contained in Table 4-6). Functionality of this family has already been described in (127,86,129). Low expressions of HSP90 have been related to better survival rates in Non-small cell lung cancer (NSCLC) (171). In Gallegos-Ruiz et al. (171), the authors suggest that targeting of HSP90 will have a clinical impact for NSCLC patients. This fact could relate the overexpression of this gene in our results to colon cancer diagnosis.

SRPK1 is a protein serine kinase that regulates the activity of RS-proteins (arginineserine-rich proteins), a group of nuclear factors controlling a variety of physiological processes including RNA processing and spliceosome assembly (198 by review of 199). Underegulation by siRNA of the expression for SRPK1 in cancer cell lines is known to reduce cell proliferation (198). Targeting SRPK1 is a promising tool that might prove therapeutically effective for tumors that overexpress this protein (95). This behavior coincides with our results, where overexpression was detected.

Allele Arg388 for gene FGFR4 (Fibroblast growth factor receptor 4) has been related with cancer progression and metastasis in breast carcinoma (174), and has been proposed as a potential marker for progression of melanoma (173). A gene allele is one of the two or more forms of the DNA sequence of a particular gene, sometimes referred to as single nucleotide polymorphisms (SNPs)(173). Evidence of its relationship with prostate cancer (175), lung cancer (176) and soft tissue sarcoma (177) has been reported. This evidence corroborates the overexpression behavior of this gene in our results, posing it as potential marker for colon cancer.

Evidence of relevance of gene HNRNPA1 (Heterogeneous nuclear ribonucleoprotein A1) in colon cancer has been reported by Ma et al. in (180) where the overexpression of the gene is correlated to the tumor severity. Its alteration has also has been deemed relevant for lung cancer (181). Behavior of its protein family has been reported to contribute in tumor development and progression (179). This behavior coincides with the expression of the gene reported in our findings.

There is some evidence of the Far upstream element (FUSE) binding protein 1 (FUBP1) to be overexpressed in tumor cell lines including Non-Small Cell Lung Cancer (NSCLC). Its coordinated expression of microtubule-destabilizing factors is also a critical step to facilitate microtubule dynamics and subsequently increase proliferation and motility of tumor cells (182). This overexpression is also shown in this gene in our results regarding colon and gastric cancer.

SET is an oncoprotein that participates in a diversity of cellular functions including cell proliferation (183). Its interaction with PP2A, which has been suggested be named I_2^{PP2A}, has been studied to contribute positively to acute myeloid leukemogenesis (184). This gene was identified as overexpressed in our findings, matching with the literature evidence.

Gene DnaJ homolog, subfamily A-member 1, corresponding to symbol DNAJA1 (also known as HDJ2) is a co-chaperone of Hsp70 that has been reported as contributing to the resistance to radiotheraphy of glioblastomas, the most aggressive and common of brain tumors (186). In our results DNAJA1 was overexpressed in colon and gastric data, possibly invoking the same effect.

Glyoxalase I (GLO1) has been found overexpressed in prostate tumor cells (187) suggesting that it may play a role in cancer homeostasis and survival, i.e. a potential biomarker gene. This protein has also shown evidence as a resistant factor to antitumor agent-induced apoptosis in human leukemia cells (188). These two evidences match with the behavior of GLO1 in our results, where GLO1 was overexpressed in all the databases.

Protein Transgelin-2 (TAGLN2) has been reported as overexpressed in Colorectal Cancer (CRC) with supportive biological validation. Overexpression of TAGLN2 was associated with lymph node and distant metastasis, advanced clinical stage of CRC and shorter overall survival in CRC. It has been proposed as a biomarker to predict CRC progression and prognosis (189). In our results, TAGLN2 was found overexpressed in both colon cancer databases, matching with the reviewed evidence. In gastric cancer, however, it showed the opposite direction. To the best of our knowledge there is no evidence of a proposed direction of change in expression for this cancer type, thus, it is suggested that its expression be investigated for biomarking characteristics.

Carbonic Anhydrase IX (CA9) has been extensively reported for its contribution in cancer processes as it is highly overexpressed in many types of cancer. CA9 has been used as a target for anticancer drug development (200). It has also been reported as an endogenous marker for hypoxic cells in cervical cancer (201). In our results this gene is overexpressed for colon databases but underexpressed for gastric. This discrepancy should be further explored to determine if it is due to difference in the cancer type or to data quality.

The Cold inducible RNA binding protein (CIRP) is one of the major cold-inducible RNA binding proteins known in human cells. It has been demonstrated that CIRP has a stimulatory effect on proliferation. CIRP has been reported as overexpressed in human tumors, considering it as a potential proto-oncogenic protein given its spectrum of characteristics like: ability to increase general protein synthesis, association with proteins that are known to be involved in tumorigenesis, involvement in immortalization of primary cells and overexpression in human malignancies (193). However, our results show underexpression of this gene using colon and gastric cancer data. This discrepancy suggests further exploration.

In (140) SCNN1B is cited as participating in the control of reabsorption of sodium in kidney, colon, lung and sweat glands, their results show this gene as importantly underexpressed in renal cell carcinoma samples. The authors postulate this gene as potential biomarker for this cancer type. Their results coincide with the behavior of this gene in our analyses.

A role for gene KCNMB1 (Potassium large conductance calcium-activated channel, subfamily M, beta member 1) in cancer metabolism has not been reported yet. However, the role for its similar KCNMA1, a protein belonging to the same family, has been reported as relevant for breast cancer invasion and metastasis to brain (202). Further investigation might be directed towards linking these genes in the context of cancer.

Many members of the Kruppel-like factors (KLF) family have been shown to be relevant to human cancers by their identified abilities to mediate in signaling processes related to the control of cell proliferation, apoptosis, migration and differentiation. In this family, the downregulation of Kruppel-like factor 9 (KLF9) has been hypothesized as involved in the carcinogenesis of human colorectal cancer with its supporting evidence presented in (194). Our results also detected an underexpression, adding to the existing evidence. Also, evidence of correlation between the activity of KLF members and the pathogenesis of endometrial and breast cancers is presented in (196) where authors suggest that a better understanding of the mode of actions of KLFs and their functional networks may lead to the development new therapeutics.

The family of microtubule-associated proteins is a growing family that includes products of oncogenes, tumor suppressors and apoptosis regulators. Existent evidence suggests the alteration of microtubule dynamics may be one of the critical events in tumorigenesis and tumor progression (197). Microtubule-associated protein 1A (MAP1A) belongs to this family and by the similarity of their elements, it could be hypothesized that this protein also contributes to the described processes.

Among the activities of CAPN2 (calpain-2), some works have shown experimental evidence suggesting that the epigenetic activation of calpain plays an important role in the invasion of human prostate cancer and that it can be targeted to reduce tumor progression (109). Coincidence of its behavior varies within the reported databases, and further biological analysis is suggested.

An important opportunity is found in those genes appearing in our meta-analysis that have not been reported to have a role in cancer processes. CBX3 gene (Chromobox homolog 3) is known as contributor in the cell cycle but there is no evidence of its relevance in cancer. CCL14 (chemokine (C-C motif) ligand 14) have been reported as potentially relevant in embryo implantation (203), but no reference has been found about this gene in terms of cancer. HUMIFNIND, the “human interferon gamma treatment inducible mRNA” is reported as a nucleotide sequence without any proved role in cell development neither in normal or cancerous progression.

As it can be seen through this validation scheme, 37 of 43 evaluated genes already have reported evidence for their biological relevance for cellular processes related to cancer. Most of them also have documented evidence for in vivo or in vitro validation for their role in specific cancer types. The remaining 6 genes would be proposed as candidates for experimental validation to verify their contribution in cellular processes related to cancer. If these last genes can be confirmed as biomarkers, a major contribution of this research will be established. Therefore, these genes are here proposed as potential biomarkers for the related cancer types.

5 CONCLUSION AND FUTURE WORK

In this thesis, a method to identify potential cancer genetic biomarkers from microarray data was introduced. The main contribution of the work comes from representing the gene identification problem as a multiple criteria mathematical optimization problem for the first time. Two important benefits of such representation are (i) the possibility of eliminating parameter adjustment by final users, and thus (ii) promoting results consistent convergence and repeatability across different analysts. This representation was also shown to permit to elicit solutions in an effective and efficient manner.

Besides the devise and validation of a competitive analysis method, the most important result from this thesis is the actual identification of potential biomarkers for colon and gastric cancer. These genes are not only backed by our mathematical and statistical analysis, but by a focused investigation of their role in the cell functions, aiming to identify those that have been biologically related to cancer. In doing this, a true interdisciplinary approach has been followed that attempts to link analysis capability from Industrial Engineering with the generation of useful knowledge for Cancer Biology and Research.

The list of potential biomarkers for colon cancer involve EST: yq31b10.s1, GUCA2B, MYL9, CFD, CBX3, CCL14 and HUMIFNIND. The last three are also proposed as potential biomarkers for gastric cancer. A major contribution to cancer research will be provided by this work if these genes can be confirmed as biomarkers in the future.

As a novel and first approach, it was important to establish that it is not only feasible but attractive to use Data Envelopment Analysis (DEA) to solve the multiple criteria optimization problem built within this research line. Two characteristics that make DEA critical to consider as a strong candidate for future analysis endeavors are (1) convergence consistency and (2) lack of the requirements of parameter adjustment by the final user. Both of them are indeed strengths of the particular formulation used in this research, called the BCC model in the literature in recognition of the original authors Banker, Charnes and Cooper, and both of them are owed in great part to having a convenient linear mathematical programming structure.

Finally, this thesis helped to pave the way for collaboration between the Bio IE Lab at UPRM and the Integrative Bioinformatics Group at the UT MD Anderson Cancer Center. The proposed method is now being considered to form the base for a collaborative effort to extract knowledge from proteomics data.

For future work at the Bio IE lab at UPRM, it is recommended that different DEA formulations as well as other multiple criteria optimization techniques be assessed in their qualities to contribute analysis convenience, precision, and –to a lesser extent- speed. In these explorations, it must be kept in mind that the final users stand in areas that are very different from the traditional data-analysis disciplines, so a simple and effective transfer across these lines must be favored. Indeed, starting with a graphical analysis to solve the multiple criteria optimization problem might be a worthwhile effort in the group.

Another challenge is that of transferring the method to proteomics, where again a cross disciplinary approach is recommended as to keep the results relevant and the methods simple yet never simplistic.

REFERENCES

1. American Cancer Society. Cancer Fact & Figures 2010. Atlanta: American Cancer Society;

2. Danaei G, Vander Hoorn S, Lopez AD, Murray CJ, Ezzati M. Causes of cancer in the world: comparative risk assessment of nine behavioural and environmental risk factors. The Lancet. 2005 Nov 19;366(9499):1784-1793.

3. Berns A. Cancer: Gene expression in diagnosis. Nature. 2000 Feb 3;403(6769):491-492.

4. Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science. 1995 Oct 20;270:467.

5. Barrett JC, Kawasaki ES. Microarrays: the use of oligonucleotides and cDNA for the analysis of gene expression. Drug Discovery Today. 2003 Jan 21;8(3):134-141.

6. Draghici S. Data Analysis Tools for DNA Microarrays. Chapman & Hall.

7. Dudoit S, Yang YH, Callow MJ, Speed TP. Statistical Methods for identigying differentially expressed genes in replicated cDNA microarray expreiments. Statistica Sinica. 2002;12:111-139.

8. Kerr MK, Afshari CA, Bennett L, Bushel P, Martinez J, Walker NJ, et al. Statistical Analysis of a Gene Expression Microarray Experiment with Replication. Statistica Sinica. 2002;12:203-217.

9. Pan W. A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. Bioinformatics. 2002 Apr 1;18(4):546-554.

10. Saeys Y, Inza I, Larranaga P. A review of feature selection techniques in bioinformatics. Bioinformatics. 2007 Oct 1;23(19):2507-2517.

11. Townsend J, Hartl D. Bayesian analysis of gene expression levels: statistical quantification of relative mRNA level across multiple strains or treatments. Genome Biology. 2002;3(12):research0071.1 - research0071.16.

12. Atkinson AJ, Colburn WA, DeGruttola VG, DeMets DL, Downing GJ, Hoth DF, et al. Biomarkers and surrogate endpoints: Preferred definitions and conceptual framework*.

Clin Pharmacol Ther. 2001 Mar;69(3):89-95.

13. Dalton WS, Friend SH. Cancer Biomarkers--An Invitation to the Table. Science. 2006 May 26;312(5777):1165-1168.

14. Burgun A, Bodenreider O. Accessing and Integrating Data and Knowledge for Biomedical Research. IMIA Yearbook 2008.

15. Yuen T, Wurmbach E, Pfeffer RL, Ebersole BJ, Sealfon SC. Accuracy and calibration of commercial oligonucleotide and custom cDNA microarrays. Nucl. Acids Res. 2002 May 15;30(10):e48.

16. Duggan DJ, Bittner M, Chen Y, Meltzer P, Trent JM. Expression profiling using cDNA microarrays. Nat Genet.

17. Leung YF, Cavalieri D. Fundamentals of cDNA microarray data analysis. Trends in Genetics. 2003 Nov;19(11):649-659.

18. Cheung VG, Morley M, Aguilar F, Massimi A, Kucherlapati R, Childs G. Making and reading microarrays. Nat Genet.

19. Watson A, Mazumder A, Stewart M, Balasubramanian S. Technology for microarray analysis of gene expression. Current Opinion in Biotechnology. 1998 Dec;9(6):609-614.

20. Fan J, Ren Y. Statistical Analysis of DNA Microarray Data in Cancer Research. Clinical Cancer Research. 2006;12(15):4469-4473.

21. Yeatman TJ. The Future of Clinical Cancer Management: One Tumor, One Chip. American Surgeon. 2003 Jan;69(1):41.

22. Lisa M. McShane, Douglas G. Altman, Willi Sauerbrei. Identification of Clinically Useful Cancer Prognostic Factors: What Are We Missing? Journal of the National Cancer Institute [Internet]. 2005 Jul 20;Available from: http://proquest.umi.com/pqdweb?did=876678081&Fmt=7&clientId=45091&RQT=309&VName=PQD

23. Tainsky MA. Genomic and proteomic biomarkers for cancer: A multitude of opportunities. Biochimica et Biophysica Acta (BBA) - Reviews on Cancer. 2009 Dec;1796(2):176-193.

24. Macoska JA. The Progressing Clinical Utility of DNA Microarrays. CA Cancer J Clin. 2002 Jan 1;52(1):50-59.

25. MAQC Consortium. The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. Nat Biotech. 2006 print;24(9):1151-1161.

26. Reverse phase protein array: validation of a novel proteomic technology and utility for analysis of primary leukemia specimens and hematopoietic stem cells.

27. Peng X, Wood C, Blalock E, Chen K, Landfield P, Stromberg A. Statistical implications of pooling RNA samples for microarray experiments. BMC Bioinformatics. 2003;4(1):26.

28. Olson NE. The Microarray Data Analysis Process: From Raw Data to Biological Significance. NeuroRX. 2006 Jul;3(3):373-383.

29. Allison DB, Cui X, Page GP, Sabripour M. Microarray data analysis: from disarray to consolidation and consensus. Nat Rev Genet. 2006 Jan;7(1):55-65.

30. Schmidt U, Begley CG. Cancer diagnosis and microarrays. The International Journal of Biochemistry & Cell Biology. 2003 Feb;35(2):119-124.

31. Chen Y, Dougherty ER, Bittner ML. Ratio-based decisions and the quantitative analysis of cDNA microarray images. Journal of Biomedical Optics. 1997;2(4):364-374.

32. Pan W. A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. Bioinformatics. 2002 Apr 1;18(4):546-554.

33. Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. Proceedings of the National Academy of Sciences of the United States of America. 2001;98(9):5116-5121.

34. Zhang S. A comprehensive evaluation of SAM, the SAM R-package and a simple modification to improve its performance. BMC Bioinformatics. 2007;8(1):230.

35. Troyanskaya OG, Garber ME, Brown PO, Botstein D, Altman RB. Nonparametric methods for identifying differentially expressed genes in microarray data. Bioinformatics. 2002 Nov 1;18(11):1454-1461.

36. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, et al. Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. Science. 1999;286(5439):531-537.

37. Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, Mack D, et al. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proceedings of the National Academy of Sciences of the United States of America. 1999;96(12):6745-6750.

38. Ross DT, Scherf U, Eisen MB, Perou CM, Rees C, Spellman P, et al. Systematic variation in gene expression patterns in human cancer cell lines. Nat Genet. 2000 Mar;24(3):227-235.

39. Efron B, Tibshirani R. On testing the significance of sets of genes. Annals of Applied Statistics. 1(1):107-129.

40. Subramanian A, Kuehn H, Gould J, Tamayo P, Mesirov J. GSEA-P: a desktop application for Gene Set Enrichment Analysis. Bioinformatics [Internet]. 2007 Dec 1;Available from: http://proquest.umi.com/pqdweb?did=1390613461&Fmt=7&clientId=45091&RQT=309&VName=PQD

41. Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature. 2000 Feb 3;403(6769):503-511.

42. Dhanasekaran SM, Barrette TR, Ghosh D, Shah R, Varambally S, Kurachi K, et al. Delineation of prognostic biomarkers in prostate cancer. Nature. 2001 print;412(6849):822-826.

43. Wong YF, Selvanayagam ZE, Wei N, Porter J, Vittal R, Hu R, et al. Expression Genomics of Cervical Cancer. Clinical Cancer Research. 2003;9(15):5486-5492.

44. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society. 1995;57(1):289-300.

45. Khan J, Wei JS, Ringner M, Saal LH, Ladanyi M, Westermann F, et al. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nat Med. 2001 Jun;7(6):673-679.

46. Lu Y, Han J. Cancer classification using gene expression data. Inf. Syst. 2003;28(4):243-268.

47. van 't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AAM, Mao M, et al. Gene

expression profiling predicts clinical outcome of breast cancer. Nature. 2002 Jan 31;415(6871):530-536.

48. van de Vijver MJ, He YD, van 't Veer LJ, Dai H, Hart AAM, Voskuil DW, et al. A Gene-Expression Signature as a Predictor of Survival in Breast Cancer. N Engl J Med. 2002;347(25):1999-2009.

49. Fan C, Oh DS, Wessels L, Weigelt B, Nuyten DSA, Nobel AB, et al. Concordance among Gene-Expression-Based Predictors for Breast Cancer. N Engl J Med. 2006;355(6):560-569.

50. Glas A, Floore A, Delahaye L, Witteveen A, Pover R, Bakx N, et al. Converting a breast cancer microarray signature into a high-throughput diagnostic test. BMC Genomics. 2006;7(1):278.

51. Buyse M, Loi S, van't Veer L, Viale G, Delorenzi M, Glas AM, et al. Validation and Clinical Utility of a 70-Gene Prognostic Signature for Women With Node-Negative Breast Cancer. J. Natl. Cancer Inst. 2006;98(17):1183-1192.

52. Bueno-de-Mesquita J, Linn S, Keijzer R, Wesseling J, Nuyten D, van Krimpen C, et al. Validation of 70-gene prognosis signature in node-negative breast cancer. Breast Cancer Research and Treatment. 2009 Oct 1;117(3):483-495.

53. Wang Y, Klijn JG, Zhang Y, Sieuwerts AM, Look MP, Yang F, et al. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. The Lancet. 2005 Feb 19;365(9460):671-679.

54. Ein-Dor L, Kela I, Getz G, Givol D, Domany E. Outcome signature genes in breast cancer: is there a unique set? Bioinformatics. 2005 Jan 15;21(2):171-178.

55. Lee JW, Lee JB, Park M, Song SH. An extensive comparison of recent classification tools applied to microarray data. Computational Statistics & Data Analysis. 2005 Apr 1;48(4):869-885.

56. Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. Proceedings of the National Academy of Sciences of the United States of America. 1998;95(25):14863-14868.

57. Notterman DA, Alon U, Sierk AJ, Levine AJ. Transcriptional Gene Expression Profiles of Colorectal Adenoma, Adenocarcinoma, and Normal Tissue Examined by Oligonucleotide Arrays. Cancer Res. 2001;61(7):3124-3130.

58. Garber ME, Troyanskaya OG, Schluens K, Petersen S, Thaesler Z, Pacyna-Gengelbach M, et al. Diversity of gene expression in adenocarcinoma of the lung. Proceedings of the National Academy of Sciences of the United States of America. 2001;98(24):13784-13789.

59. Herrero J, Valencia A, Dopazo J. A hierarchical unsupervised growing neural network for clustering gene expression patterns. Bioinformatics. 2001 Feb 1;17(2):126-136.

60. Castro C, Cabrera-Ríos M, Lilly B, Castro JM, Mount-Campbell CA. Identifying The Best Compromises Between Multiple Performance Measures In Injection Molding (IM) Using Data Envelopment Analysis (DEA). J. Integr. Des. Process Sci. 2003;7(1):77-86.

61. Castro CE, Rios MC, Castro JM, Lilly B. Multiple criteria optimization with variability considerations in injection molding. Polymer Engineering and Science [Internet]. 2007 Apr 1;Available from: http://find.galegroup.com/gtx/infomark.do?&contentSet=IAC-Documents&type=retrieve&tabID=T002&prodId=AONE&docId=A161846332&source=gale&srcprod=AONE&userGroupName=uprmayaguez&version=1.0

62. Loera V, Castro J, Diaz J, Mondragon O, Cabrera-Rios M. Setting the Processing Parameters in Injection Molding Through Multiple-Criteria Optimization: A Case Study. Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on. 2008;38(5):710-715.

63. Marroquín MGV, Peña MLS, Castro CE, Castro JM, Cabrera-Ríos M. Use of data envelopment analysis and clustering in multiple criteria optimization. Intell. Data Anal. 2008;12(1):89-101.

64. Castro C, Cabrera-Rí M, Lilly B, Castro JM, Mount-Campbell CA. Identifying the best compromises between multiple performance measures in injection molding (IM) using Data Envelopment Analysis (DEA). Journal of Integrated Design & Process Science. 2003 Mar;7(1):77.

65. Ehrgott M. Multicriteria Optimization. 2nd ed. Heidelberg New York: Springer; 2005.

66. Deb K. Multi-Objective Optimizacion using Evolutionary Algorithms. England: John Wiley & Sons Ltd.; 2004.

67. Charnes A, Cooper WW, Rhodes E. Measuring the efficiency of decision making units. European Journal of Operational Research. 1978;2:429-444.

68. Cooper WW, Seiford LM, Shu J, editors. Handbook on Data Envelopment Analysis, 2nd ed. 2004.

69. Charnes A, Cooper W, Rhodes E. Evaluating Program and Managerial Efficiency: An Application of Data Envelopment Analysis to rogram Follow Through. Management Science. 1981 Jun;27(6).

70. Stewart TJ. Data Envelopment Analysis and Multiple Criteria Decision Making: A Response. Omega. 1994 Mar;22(2):205-206.

71. Bouyssou D. Using DEA as a tool for MCDM: some remarks. Journal of the Operation Reserch Society. 1999;50(1999):974-978.

72. Sarkis J. Preparing Your Data for DEA [Internet]. In: Modeling Data Irregularities and Structural Complexities in Data Envelopment Analysis. 2007. Available from: http://dx.doi.org/10.1007/978-0-387-71607-7_17

73. Dyson RG, Allen R, Camanho AS, Podinovski VV, Sarrico CS, Shale EA. Pitfalls and protocols in DEA. European Journal of Operational Research. 2001 Jul 16;132(2):245-259.

74. Li X.-B., Reeves G.R.[1]. A multiple criteria approach to data envelopment analysis. European Journal of Operational Research. 1999 Jun 16;115:507-517.

75. Belton V, Vickers SP. Demystifying DEA-A Visual Interactive Approach Based on Multiple Criteria Analysis. J Oper Res Soc. 1993 print;44(9):883-896.

76. Mastinu G, Gobbi M, Miano C. Optimal Design of Complex Mechanical Systems. Springer; 2006.

77. Seiford LM, Thrall RM. Recent developments in DEA : The mathematical programming approach to frontier analysis. Journal of Econometrics. Oct;46(1-2):7-38.

78. Taha HA. Operations Research an Introduction. 7th ed. USA: Prentice Hall; 2003.

79. Stewart TJ. Relationships between Data Envelopment Analysis and Multicriteria Decision Analysis. The Journal of the Operational Research Society. 1996 May;47(5):654-665.

80. Hollander M, Wolfe DA. Nonparametric Statistical Methods. Second. Wiley-Interscience, John Wiley & Sons, Inc.; 1999.

81. Uribe Mastache L, Perez Vicente H, Cabrera-Ríos M, Isaza-Brando C. Análisis estadístico no paramétrico para la detección de cáncer a partir de datos de microarreglos:

Resultados Preliminares. Cuernavaca, Morelos, México:

82. Perez Vicente H, Uribe Mastache L, Cabrera-Ríos M, Isaza.Brando C. Diagnóstico de cáncer a partir de datos de microarreglos. In: Congreso Internacional en Innovación y Desarrollo Tecnológico. Cuernavaca, Morelos, México: 2008.

83. Hippo Y, Taniguchi H, Tsutsumi S, Machida N, Chong J, Fukayama M, et al. Global Gene Expression Analysis of Gastric Cancer by Oligonucleotide Microarrays. Cancer Research. 2002 Jan 1;62(1):233 -240.

84. Alvis Brazma, Pascal Hingamp, John Quackenbush, Gavin Sherlock, Paul Spellman, Chris Stoeckert, et al. Minimum information about a microarray experiment (MIAME)--toward standards for microarray data. Nature Genetics. 2001 Dec;29(4):365.

85. Xie K. Interleukin-8 and human cancer biology. Cytokine & Growth Factor Reviews. 2001 Dec;12(4):375-391.

86. Cornford PA, Dodson AR, Parsons KF, Desmond AD, Woolfenden A, Fordham M, et al. Heat Shock Protein Expression Independently Predicts Clinical Outcome in Prostate Cancer. Cancer Research. 2000 Dec 12;60(24):7099 -7105.

87. Malagó Jr. W, Sommer CA, Del Cistia Andrade C, Soares-Costa A, Possik PA, Cassago A, et al. Gene Expression Profile of Human Down Syndrome Leukocytes. Croatian Medical Journal. 2005;46(4):647-656.

88. Verhaak RGW, Goudswaard CS, van Putten W, Bijl MA, Sanders MA, Hugens W, et al. Mutations in nucleophosmin (NPM1) in acute myeloid leukemia (AML): association with other gene abnormalities and previously established gene expression signatures and their favorable prognostic significance. Blood. 2005 Dec 1;106(12):3747-3754.

89. Altmannsberger M, Weber K, Droste R, Osborn M. Desmin is a specific marker for rhabdomyosarcomas of human and rat origin. Am J Pathol. 1985 Jan 1;118(1):85-95.

90. Zia H, Hida T, Jakowlew S, Birrer M, Gozes Y, Reubi JC, et al. Breast Cancer Growth Is Inhibited by Vasoactive Intestinal Peptide (VIP) Hybrid, a Synthetic VIP Receptor Antagonist. Cancer Research. 1996;56(15):3486 -3489.

91. Henrique R, Jerónimo C, Hoque MO, Nomoto S, Carvalho AL, Costa VL, et al. MT1G Hypermethylation Is Associated with Higher Tumor Stage in Prostate Cancer. Cancer Epidemiology Biomarkers & Prevention. 2005 May 1;14(5):1274 -1278.

92. Hristov AC, Cope L, Di Cello F, Reyes MD, Singh M, Hillion JA, et al. HMGA1

correlates with advanced tumor grade and decreased survival in pancreatic ductal adenocarcinoma. Mod Pathol. 2009 Oct 9;23(1):98-104.

93. HIBI K, GOTO T, MIZUKAMI H, KITAMURA Y, SAKURABA K, SAKATA M, et al. Demethylation of the CDH3 Gene Is Frequently Detected in Advanced Colorectal Cancer. Anticancer Research. 2009;29(6):2215-2217.

94. Yeong Min H, Spiegelman BM. Adipsin, the adipocyte serine protease: gene structure and control of expression by tumor necrosis factor. Nucleic Acids Research. 1986 Nov 11;14(22):8879 -8892.

95. Hayes GM, Carrigan PE, Miller LJ. Serine-Arginine Protein Kinase 1 Overexpression Is Associated with Tumorigenic Imbalance in Mitogen-Activated Protein Kinase Pathways in Breast, Colonic, and Pancreatic Carcinomas. Cancer Research. 2007 Mar 1;67(5):2072 -2080.

96. Kwiatkowski DJ. Functions of gelsolin: motility, signaling, apoptosis, cancer. Current Opinion in Cell Biology. 1999 Feb 1;11(1):103-108.

97. Xiong M, Fang X, Zhao J. Biomarker Identification by Feature Wrappers. Genome Research. 2001 Nov 1;11(11):1878 -1887.

98. Yap Y, Zhang X, Ling MT, Wang X, Wong YC, Danchin A. Classification between normal and tumor tissues based on the pair-wise gene expression ratio. BMC Cancer. 2004;4(1):72.

99. Furlanello C, Serafini M, Merler S, Jurman G. Entropy-based gene ranking without selection bias for the predictive classification of microarray data. BMC Bioinformatics. 2003;4(1):54.

100. Chen J, Tsai C, Tzeng S, Chen C. Gene selection with multiple ordering criteria. BMC Bioinformatics. 2007;8(1):74.

101. Zhang H, Song X, Wang H, Zhang X. MIClique: An Algorithm to Identify Differentially Coexpressed Disease Gene Subset from Microarray Data. Journal of Biomedicine and Biotechnology. 2009;2009(Article ID 642524):9.

102. Bo T, Jonassen I. New feature subset selection procedures for classification of expression profiles. Genome Biology. 2002;3(4):research0017.1 - research0017.11.

103. Christian K, Lang M, Maurel P, Raffalli-Mathieu F. Interaction of Heterogeneous Nuclear Ribonucleoprotein A1 with Cytochrome P450 2A6 mRNA: Implications for

Post-Transcriptional Regulation of the CYP2A6 Gene. Molecular Pharmacology. 2004;65(6):1405-1414.

104. MacDonald NJ, la Rosa D, Steeg PS. The potential roles of nm23 in cancer metastasis and cellular differentiation. European Journal of Cancer. Jul;31(7-8):1096-1100.

105. Beechem J, Wang L, Love B, Rogers J. METHODS AND KITS FOR DETECTING PROSTATE CANCER BIOMARKERS.

106. Knudsen S. METHODS AND DEVICES FOR IDENTIFYING BIOMARKERS OF TREATMENT RESPONSE AND USE THEREOF TO PREDICT TREATMENT EFFICACY.

107. Tumor Suppressor Genes DNA Methylation PCR Array [Internet]. [cited 2010 Oct 17];Available from: file:///C:/Documents%20and%20Settings/Matilde/Desktop/Validacion%20Colon/NME1/NME1_v1.htm

108. Saarnio J, Parkkila S, Parkkila A, Haukipuro K, Pastorekova S, Pastorek J, et al. Immunohistochemical Study of Colorectal Tumors for Expression of a Novel Transmembrane Carbonic Anhydrase, MN/CA IX, with Potential Value as a Marker of Cell Proliferation. Am J Pathol. 1998 Jul 1;153(1):279-285.

109. Mamoune A, Luo J, Lauffenburger DA, Wells A. Calpain-2 as a Target for Limiting Prostate Cancer Invasion. Cancer Research. 2003;63(15):4632 -4640.

110. Libertini SJ, Robinson BS, Dhillon NK, Glick D, George M, Dandekar S, et al. Cyclin E Both Regulates and Is Regulated by Calpain 2, a Protease Associated with Metastatic Breast Cancer Phenotype. Cancer Research. 2005 Dec 1;65(23):10700 -10708.

111. Bellacosa A, de Feo D, Godwin A, Bell D, Cheng J, Altomare D, et al. Molecular alterations of the AKT2 oncogene in ovarian and breast carcinomas. International Journal of Cancer. 1995 Aug;64(4):280-5.

112. Wang X, Gotoh O. Microarray-Based Cancer Prediction Using Soft Computing Approach. Cancer Informatics. 2009 May 26;2009(CIN-7-Wang-et-al):123.

113. Zhang H, Yu C, Singer B, Xiong M. Recursive partitioning for tumor classification with gene expression microarray data. Proceedings of the National Academy of Sciences of the United States of America. 2001 Jun 5;98(12):6730 -6735.

114. Steeg PS, Bevilacqua G, Kopper L, Thorgeirsson UP, Talmadge JE, Liotta LA, et al.

Evidence for a Novel Gene Associated With Low Tumor Metastatic Potential. Journal of the National Cancer Institute. 1988 Apr 6;80(3):200 -204.

115. Qu S, Long J, Cai Q, Shu X, Cai H, Gao Y, et al. Genetic Polymorphisms of Metastasis Suppressor Gene NME1 and Breast Cancer Survival. Clinical Cancer Research. 2008;14(15):4787 -4793.

116. Bevilacqua G, Sobel ME, Liotta LA, Steeg PS. Association of Low nm23 RNA Levels in Human Primary Infiltrating Ductal Breast Carcinomas with Lymph Node Involvement and Other Histopathological Indicators of High Metastatic Potential. Cancer Research. 1989;49(18):5185 -5190.

117. Nakayama T, Ohtsuru A, Nakao K, Shima M, Nakata K, Watanabe K, et al. Expression in Human Hepatocellular Carcinoma of Nucleoside Diphosphate Kinase, a Homologue of the nm23 Gene Product. Journal of the National Cancer Institute. 1992;84(17):1349 -1354.

118. Mandai M, Konishi I, Koshiyama M, Mori T, Arao S, Tashiro H, et al. Expression of Metastasis-related nm23-H1 and nm23-H2 Genes in Ovarian Carcinomas: Correlation with Clinicopathology, EGFR, c-erbB-2, and c-erbB-3 Genes, and Sex Steroid Receptor Expression. Cancer Research. 1994 Apr 1;54(7):1825 -1830.

119. Flørenes VA, Aamdal S, Myklebost O, Maelandsmo GM, Bruland ØS, Fodstad Ø. Levels of nm23 Messenger RNA in Metastatic Malignant Melanomas: Inverse Correlation to Disease Progression. Cancer Research. 1992 Nov 1;52(21):6088 -6091.

120. Zaza G, Yang W, Kager L, Cheok M, Downing J, Pui C, et al. Acute lymphoblastic leukemia with TEL-AML1 fusion has lower expression of genes involved in purine metabolism and lower de novo purine synthesis. Blood. 2004 Sep 1;104(5):1435-1441.

121. Yokoyama A, Okabe-Kado J, Wakimoto N, Kobayashi H, Sakashita A, Maseki N, et al. Evaluation by Multivariate Analysis of the Differentiation Inhibitory Factor nm23 as a Prognostic Factor in Acute Myelogenous Leukemia and Application to Other Hematologic Malignancies. Blood. 1998 Mar 15;91(6):1845-1851.

122. Plourde M, Manhes C, Leblanc G, Durocher F, Dumont M, Sinilnikova O, et al. Mutation analysis and characterization of HSD17B2 sequence variants in breast cancer cases from French Canadian families with high risk of breast and ovarian cancer. J Mol Endocrinol. 2008 Apr 1;40(4):161-172.

123. Bhavani V, Srinivasulu M, Ahuja Y, Hasan Q. Role of BRCA1, HSD17B1 and HSD17B2 methylation in breast cancer tissue. Cancer Biomarkers. 2009;5(4-5):207-213.

124. Dias P, Kumar P, Marsden H, Morris-Jones P, Birch J, Swindell R, et al. Evaluation of desmin as a diagnostic and prognostic marker of childhood rhabdomyosarcomas and embryonal sarcomas. Br J Cancer. 1987 print;56(3):361-365.

125. Qu S, Long J, Cai Q, Shu X, Cai H, Gao Y, et al. Genetic Polymorphisms of Metastasis Suppressor Gene NME1 and Breast Cancer Survival. Clinical Cancer Research. 2008;14(15):4787 -4793.

126. Hoffmann R, Valencia A. A gene network for navigating the literature. Nature Genetics [Internet]. 2004;36(664). Available from: iHOP - http://www.ihop-net.org/

127. So A, Hadaschik B, Sowery R, Gleave M. The Role of Stress Proteins in Prostate Cancer [Internet]. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2430682

128. Christensen J, Nielsen M, Hansen J, Füchtbauer A, Füchtbauer E, West M, et al. Inactivation of the hereditary spastic paraplegia-associated <i>Hspd1</i> gene encoding the Hsp60 chaperone results in early embryonic lethality in mice. Cell Stress and Chaperones. 2010 Nov 1;15(6):851-863.

129. Mosser DD, Morimoto RI. Molecular chaperones and the stress of oncogenesis. Oncogene. 0000 print;23(16):2907-2918.

130. Johansson B, Pourian MR, Chuan Y, Byman I, Bergh A, Pang S, et al. Proteomic comparison of prostate cancer cell lines LNCaP-FGC and LNCaP-r reveals heatshock protein 60 as a marker for prostate malignancy. Prostate. 2006;66(12):1235-1244.

131. Leary JA, Kerr J, Chenevix-Trench G, Doris CP, Hurst T, Houghton CRS, et al. Increased expression of the nme1 gene is associated with metastasis in epithelial ovarian cancer. Int. J. Cancer. 1995;64(3):189-195.

132. Papadaki C, Dufour A, Seibl M, Schneider S, Bohlander SK, Zellmeier E, et al. Monitoring minimal residual disease in acute myeloid leukaemia with NPM1 mutations by quantitative PCR: clonal evolution is a limiting factor. British Journal of Haematology. 2009;144(4):517-523.

133. Ma Y, Peng J, Liu W, Zhang P, Huang L, Gao B, et al. Proteomics Identification of Desmin as a Potential Oncofetal Diagnostic and Prognostic Biomarker in Colorectal Cancer. Molecular & Cellular Proteomics. 2009;8(8):1878 -1890.

134. Van Spaendonck-Zwarts K, Van Hessem L, Jongbloed J, De Walle H, Capetanaki Y,

Van Der Kooi A, et al. Desmin-related myopathy: a review and meta-analysis. Clinical Genetics. 2010;:no.

135. Moody TW, Mantey SA, Fuselier JA, Coy DH, Jensen RT. Vasoactive intestinal peptide-camptothecin conjugates inhibit the proliferation of breast cancer cells. Peptides. 2007 Sep;28(9):1883-1890.

136. Yuichi H, Arai M, Imazeki F, Tada M, Mikata R, Fukai K, et al. Methylation Status of Genes Upregulated by Demethylating Agent 5-aza-2′-Deoxycytidine in Hepatocellular Carcinoma. Oncology International Journal for Cancer Research and Treatment. 2006;71(1-2):9.

137. Van den Abbeele A, De Corte V, Van Impe K, Bruyneel E, Boucherie C, Bracke M, et al. Downregulation of gelsolin family proteins counteracts cancer cell invasion in vitro. Cancer Letters. 2007 Sep 18;255(1):57-70.

138. Marrocco C, Rinalducci S, Mohamadkhani A, D'Amici G, Zolla L. Plasma gelsolin protein: a candidate biomarker for hepatitis B-associated liver cirrhosis identified by proteomic approach. Blood Transfus. 2010 Jun;8 Suppl 3:s105-12.

139. SAKAMOTO LHT, DE CAMARGO B, CAJAIBA M, SOARES FA, VETTORE AL. MT1G Hypermethylation: A Potential Prognostic Marker for Hepatoblastoma. Pediatric Research. 2010;67(4):387-393 10.1203/PDR.0b013e3181d01863.

140. Dalgin GS, Drever M, Williams T, King T, DeLisi C, Liou LS. Identification of Novel Epigenetic Markers for Clear Cell Renal Cell Carcinoma. The Journal of Urology. 2008 Sep;180(3):1126-1130.

141. Wang X, Gotoh O. Inference of Cancer-specific Gene Regulatory Networks Using Soft Computing Rules. Gene Regulation and Systems Biology. 2010 Mar 31;2010(1964-GRSB-Inference-of-Cancer-specific-Gene-Regulatory-Networks-Using-Soft-Compu.pdf):19.

142. Gorlov I, Byun J, Gorlova O, Aparicio A, Efstathiou E, Logothetis C. Candidate pathways and genes for prostate cancer: a meta-analysis of gene expression data. BMC Medical Genomics. 2009;2(1):48.

143. Adib TR, Henderson S, Perrett C, Hewitt D, Bourmpoulia D, Ledermann J, et al. Predicting biomarkers for ovarian cancer using gene-expression microarrays. Br J Cancer. 0000 print;90(3):686-692.

144. Zhang X, Yap Y, Wei D, Chen F, Danchin A. Molecular diagnosis of human cancer type by gene expression profiles and independent component analysis. Eur J Hum Genet. 2005 Dec;13(12):1303-11.

145. AHMED FE, VOS P, IJAMES S, LYSLE DT, ALLISON RR, FLAKE G, et al. Transcriptomic Molecular Markers for Screening Human Colon Cancer in Stool and Tissue. Cancer Genomics - Proteomics. 2007;4(1):1-20.

146. Sagawa N, Fujita H, Banno Y, Nozawa Y, Katoh H, Kuzumaki N. Gelsolin suppresses tumorigenicity through inhibiting PKC activation in a human lung cancer cell line, PC10. Br J Cancer. 0000 print;88(4):606-612.

147. Tanaka M, Müllauer L, Ogiso Y, Fujita H, Moriya S, Furuuchi K, et al. Gelsolin: A Candidate for Suppressor of Human Bladder Cancer. Cancer Research. 1995;55(15):3228 -3232.

148. Rother K, Dengl M, Lorenz J, Tschöp K, Kirschner R, Mössner J, et al. Gene expression of cyclin-dependent kinase subunit Cks2 is repressed by the tumor suppressor p53 but not by the related proteins p63 or p73. FEBS Letters. 2007 Mar 20;581(6):1166-1172.

149. Choi YR, Kim H, Kang HJ, Kim N, Kim JJ, Park K, et al. Overexpression of High Mobility Group Box 1 in Gastrointestinal Stromal Tumors with KIT Mutation. Cancer Research. 2003 May 1;63(9):2188 -2193.

150. Völp K, Brezniceanu M, Bösser S, Brabletz T, Kirchner T, Göttel D, et al. Increased expression of high mobility group box 1 (HMGB1) is associated with an elevated level of the antiapoptotic c-IAP2 protein in human colon carcinomas. Gut. 2006 Feb 1;55(2):234 -242.

151. Brezniceanu M, Volp K, Bosser S, Solbach C, Lichter P, Joos S, et al. HMGB1 inhibits cell death in yeast and mammalian cells and is abundantly expressed in human breast carcinoma. FASEB J. 2003;17(10):1295-1297.

152. Poser I, Golob M, Buettner R, Bosserhoff AK. Upregulation of HMG1 Leads to Melanoma Inhibitory Activity Expression in Malignant Melanoma Cells and Contributes to Their Malignancy Phenotype. Mol. Cell. Biol. 2003 Apr 15;23(8):2991-2998.

153. Court EL, Ann Smith M, Avent ND, Hancock JT, Morgan LM, Gray AG, et al. DNA microarray screening of differential gene expression in bone marrow samples from AML, non-AML patients and AML cell lines. Leukemia Research. 2004 Jul;28(7):743-

753.

154. Kawakami K, Enokida H, Tachiwada T, Gotanda T, Tsuneyoshi K, Kubo H, et al. Identification of differentially expressed genes in human bladder cancer through genome-wide gene expression profiling. Oncol Rep. 2006;16(3):521-31.

155. Lyng H, Brovig R, Svendsrud D, Holm R, Kaalhus O, Knutstad K, et al. Gene expressions and copy numbers associated with metastatic phenotypes of uterine cervical cancer. BMC Genomics. 2006;7(1):268.

156. Lin C, Strom A, Li Kong S, Kietz S, Thomsen J, Tee J, et al. Inhibitory effects of estrogen receptor beta on specific hormone-responsive gene expression and association with disease outcome in primary breast cancer. Breast Cancer Research. 2007;9(2):R25.

157. Chang H, Jiang N, Jiang H, Saha MN, Qi C, Xu W, et al. CKS1B nuclear expression is inversely correlated with p27Kip1 expression and is predictive of an adverse survival in patients with multiple myeloma. Haematologica. 2010 Sep 1;95(9):1542-1547.

158. Slotky M, Shapira M, Ben-Izhak O, Linn S, Futerman B, Tsalic M, et al. The expression of the ubiquitin ligase subunit Cks1 in human breast cancer. Breast Cancer Research. 2005;7(5):R737 - R744.

159. Wang T, Chao A, Tsai C, Chang C, Chen S, Lee Y, et al. Stress-induced Phosphoprotein 1 as a Secreted Biomarker for Human Ovarian Cancer Promotes Cancer Cell Proliferation. Molecular & Cellular Proteomics. 2010;9(9):1873 -1884.

160. Breckenridge DG, Nguyen M, Kuppig S, Reth M, Shore GC. The procaspase-8 isoform, procaspase-8L, recruited to the BAP31 complex at the endoplasmic reticulum. Proceedings of the National Academy of Sciences of the United States of America. 2002 Apr 2;99(7):4331 -4336.

161. Turyn J, Schlichtholz B, Dettlaff-Pokora A, Presler M, Goyke E, Matuszewski M, et al. Increased Activity of Glycerol 3-phosphate Dehydrogenase and Other Lipogenic Enzymes in Human Bladder Cancer. Horm Metab Res. 2003;35(10):565,569.

162. Makiyama T, Akao M, Haruna Y, Tsuji K, Doi T, Ohno S, et al. Mutation Analysis of the Glycerol-3 Phosphate Dehydrogenase-1 Like (GPD1L) Gene in Japanese Patients With Brugada Syndrome. Circulation Journal. 2008;72(10):1705-1706.

163. London B, Michalec M, Mehdi H, Zhu X, Kerchner L, Sanyal S, et al. Mutation in Glycerol-3-Phosphate Dehydrogenase 1-Like Gene (GPD1-L) Decreases Cardiac Na+ Current and Causes Inherited Arrhythmias. Circulation. 2007 Nov 13;116(20):2260-

2268.

164. Yasui Y, Kim M, Tanaka T. PPAR Ligands for Cancer Chemoprevention. PPAR Research. 2008;2008:10.

165. Mackenzie GG, Rasheed S, Wertheim W, Rigas B. NO-Donating NSAIDs, PPARδ, and Cancer: Does PPARδ Contribute to Colon Carcinogenesis? PPAR Research. 2008;2008:11.

166. Fernández-Salas E, Peracaula R, Frazier ML, De Llorens R. Ribonucleases expressed by human pancreatic adenocarcinoma cell lines. European Journal of Biochemistry. 2000;267(5):1484-1494.

167. Segawa T, Nau ME, Xu LL, Chilukuri RN, Makarem M, Zhang W, et al. Androgen-induced expression of endoplasmic reticulum (ER) stress response genes in prostate cancer cells. Oncogene. 2002 Dicember 12;21(57):8749-8758.

168. Hendriksen PJ, Dits NF, Kokame K, Veldhoven A, van Weerden WM, Bangma CH, et al. Evolution of the Androgen Receptor Pathway during Progression of Prostate Cancer. Cancer Research. 2006 May 15;66(10):5012 -5020.

169. Pajonk F, McBride WH. The Proteasome in Cancer Biology and Treatment. Radiation Research. 2001 Nov 1;156(5):447-459.

170. Biggar RJ, Bergen AW, Poulsen GN. Impact of X Chromosome Genes in Explaining the Excess Risk of Cancer in Males. American Journal of Epidemiology. 2009 Jul 1;170(1):65 -71.

171. Gallegos Ruiz MI, Floor K, Roepman P, Rodriguez JA, Meijer GA, Mooi WJ, et al. Integration of Gene Dosage and Gene Expression in Non-Small Cell Lung Cancer, Identification of HSP90 as Potential Target. PLoS ONE. 2008 Mar 5;3(3):e0001722.

172. Grosso AR, Martins S, Carmo-Fonseca M. The emerging role of splicing factors in cancer. EMBO Rep. 2008 Nov;9(11):1087-1093.

173. Streit S, Mestel DS, Schmidt M, Ullrich A, Berking C. FGFR4 Arg388 allele correlates with tumour thickness and FGFR4 protein expression with survival of melanoma patients. Br J Cancer. 2006;94:1879 - 1886.

174. Bange J, Prechtl D, Cheburkin Y, Specht K, Harbeck N, Schmitt M, et al. Cancer progression and tumor cell motility are associated with the FGFR4 Arg(388) allele. Cancer Res. 2002;62:840 - 847.

175. Wang J, Stockton DW, Ittmann M. The fibroblast growth factor receptor-4 Arg388 allele is associated with prostate cancer initiation and progression. Clin Cancer Res. 2004;10:6169 - 6178.

176. Spinola M, Leoni V, Pignatiello C, Conti B, Ravagnani F, Pastorino U, et al. Functional FGFR4 Gly388Arg polymorphism predicts prognosis in lung adenocarcinoma patients. J Clin Oncol. 2005;23:7307 - 7311.

177. Morimoto Y, Ozaki T, Umehara N, Ohata N, Yoshida A, Shimizu K, et al. Single nucleotide polymorphism in fibroblast growth factor receptor 4 at codon 388 is associated with prognosis in high-grade soft tissue sarcoma. Cancer. 2003;98(10):2245-2250.

178. Park SG, Schimmel P, Kim S. Aminoacyl tRNA synthetases and their connections to disease. Proceedings of the National Academy of Sciences. 2008;105(32):11043 -11049.

179. Carpenter B, MacKay C, Alnabulsi A, MacKay M, Telfer C, Melvin WT, et al. The roles of heterogeneous nuclear ribonucleoproteins in tumour development and progression. Biochimica et Biophysica Acta (BBA) - Reviews on Cancer. 2006 Apr;1765(2):85-100.

180. Ma Y, Peng J, Zhang P, Huang L, Liu W, Shen T, et al. Heterogeneous Nuclear Ribonucleoprotein A1 Is Identified as a Potential Biomarker for Colorectal Cancer Based on Differential Proteomics Technology. Journal of Proteome Research. 2009 Oct 2;8(10):4525-4535.

181. Pino I, Pío R, Toledo G, Zabalegui N, Vicent S, Rey N, et al. Altered patterns of expression of members of the heterogeneous nuclear ribonucleoprotein (hnRNP) family in lung cancer. Lung Cancer. 2003 Aug;41(2):131-143.

182. Singer S, Malz M, Herpel E, Warth A, Bissinger M, Keith M, et al. Coordinated Expression of Stathmin Family Members by Far Upstream Sequence Element-Binding Protein-1 Increases Motility in Non–Small Cell Lung Cancer. Cancer Research. 2009 Mar 15;69(6):2234 -2243.

183. Vera J, Jaumot M, Estanyol JM, Brun S, Agell N, Bachs O. Heterogeneous nuclear ribonucleoprotein A2 is a SET-binding protein and a PP2A inhibitor. Oncogene. 2005 online;25(2):260-270.

184. Li M, Makkinje A, Damuni Z. The Myeloid Leukemia-associated Protein SET Is a Potent Inhibitor of Protein Phosphatase 2A. Journal of Biological Chemistry. 1996 May

10;271(19):11059 -11062.

185. Mitra A, Shevde L, Samant R. Multi-faceted role of HSP40 in cancer. Clinical and Experimental Metastasis. 2009 Aug 1;26(6):559-567.

186. Wang C, Liao Y, Mischel PS, Iwamoto KS, Cacalano NA, McBride WH. HDJ-2 as a Target for Radiosensitization of Glioblastoma Multiforme Cells by the Farnesyltransferase Inhibitor R115777 and the Role of the p53/p21 Pathway. Cancer Research. 2006 Jul 1;66(13):6756 -6762.

187. Davidson Sd, Cherry Jp, Choudhury Ms, Tazaki H, Mallouh C, Konno S. Glyoxalase i activity in human prostate cancer: a potential marker and importance in chemotherapy. The Journal of Urology. 1999 Feb;161(2):690-691.

188. Sakamoto H, Mashima T, Kizaki A, Dan S, Hashimoto Y, Naito M, et al. Glyoxalase I is involved in resistance of human leukemia cells to antitumor agent-induced apoptosis. Blood. 2000 May 15;95(10):3214-3218.

189. Zhang Y, Ye Y, Shen D, Jiang K, Zhang H, Sun W, et al. Identification of transgelin-2 as a biomarker of colorectal cancer by laser capture microdissection and quantitative proteome analysis. Cancer Science. 2010;101(2):523-529.

190. Bui MHT, Seligson D, Han K, Pantuck AJ, Dorey FJ, Huang Y, et al. Carbonic Anhydrase IX Is an Independent Predictor of Survival in Advanced Renal Clear Cell Carcinoma. Clinical Cancer Research. 2003 Feb 1;9(2):802 -811.

191. Loncaster JA, Harris AL, Davidson SE, Logue JP, Hunter RD, Wycoff CC, et al. Carbonic Anhydrase (CA IX) Expression, a Potential New Intrinsic Marker of Hypoxia: Correlations with Tumor Oxygen Measurements and Prognosis in Locally Advanced Carcinoma of the Cervix. Cancer Research. 2001;61(17):6394 -6399.

192. Lim HY, Ahn M, Chung HC, Gardner TA, Kao C, Lee S, et al. Tumor-specific gene therapy for uterine cervical cancer using MN//CA9-directed replication-competent adenovirus. Cancer Gene Ther. 2004 May 28;11(8):532-538.

193. LLeonart M. A new generation of proto-oncogenes: Cold-inducible RNA binding proteins. Biochimica et Biophysica Acta (BBA) - Reviews on Cancer. 2010 Jan;1805(1):43-52.

194. Kang L, Lü B, Xu J, Hu H, Lai M. Downregulation of Krüppel-like factor 9 in human colorectal cancer. Pathology International. 2008;58(6):334-338.

195. Black AR, Black JD, Azizkhan-Clifford J. Sp1 and krüppel-like factor family of transcription factors in cell growth regulation and cancer. J. Cell. Physiol. 2001;188(2):143-160.

196. Simmen RCM, Pabona JMP, Velarde MC, Simmons C, Rahal O, Simmen FA. The emerging role of Kruppel-like factors in endocrine-responsive cancers of female reproductive tissues. J Endocrinol. 2010 Mar 1;204(3):223-231.

197. Bhat KM, Setaluri V. Microtubule-Associated Proteins as Targets in Cancer Chemotherapy. Clinical Cancer Research. 2007 May 15;13(10):2849 -2854.

198. Sanz G, Mir L, Jacquemin-Sablon A. Bleomycin Resistance in Mammalian Cells Expressing a Genetic Suppressor Element Derived from the SRPK1 Gene. Cancer Research. 2002;62(15):4453 -4458.

199. Fu XD. The superfamily of arginine/serine-rich splicing factors. RNA. 1995;1(7):663-680.

200. Thiry A, Dogné J, Masereel B, Supuran CT. Targeting tumor-associated carbonic anhydrase IX in cancer therapy. Trends in Pharmacological Sciences. 2006 Nov;27(11):566-573.

201. Olive PL, Aquino-Parsons C, MacPhail SH, Liao S, Raleigh JA, Lerman MI, et al. Carbonic Anhydrase 9 as an Endogenous Marker for Hypoxic Cells in Cervical Cancer. Cancer Research. 2001 Dec 15;61(24):8924 -8929.

202. Khaitan D, Sankpal U, Weksler B, Meister E, Romero I, Couraud P, et al. Role of KCNMA1 gene in breast cancer invasion and metastasis to brain. BMC Cancer. 2009;9(1):258.

203. Hannan NJ, Salamonsen LA. CX3CL1 and CCL14 Regulate Extracellular Matrix and Adhesion Molecules in the Trophoblast: Potential Roles in Human Embryo Implantation. Biology of Reproduction. 2008 Jul 1;79(1):58 -65.

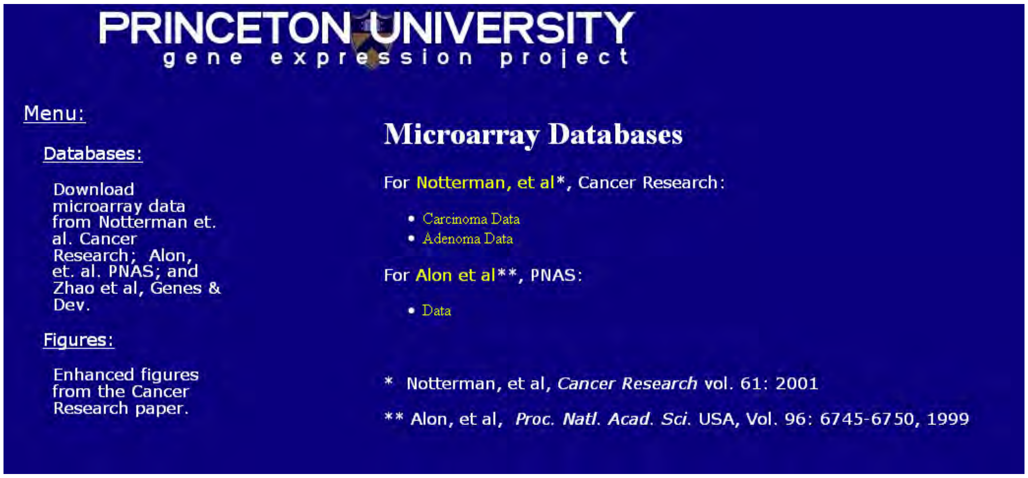
204. Henrique R, Jerónimo C, Hoque MO, Nomoto S, Carvalho AL, Costa VL, et al. MT1G Hypermethylation Is Associated with Higher Tumor Stage in Prostate Cancer. Cancer Epidemiology Biomarkers & Prevention. 2005 May 1;14(5):1274 -1278.

205. Futami J, Tsushima Y, Murato Y, Tada H, Sasaki J, Seno M, et al. Tissue-Specific Expression of Pancreatic-Type RNases and RNase Inhibitor in Humans. DNA and Cell Biology. 1997;16(4):413-419.

APPENDIX A – HOW TO ACCESS MICROARRAY DATABASES

It is common that scientific papers using microarray data provide an internet address –usually the laboratory´s or research group´s- where the raw data can be accessed. Sometimes, these data are also available in official public repositories sponsored by government agencies or private repositories sponsored usually by pharmaceutical industries. Once the data source is identified in the reference paper, the format and structure of the data should be understood for its good use.

In this section the different procedures to access the data involved in this work are presented. The data for the two cases of colon can be found in http://microarray.princeton.edu/oncology. The main page when this repository is accessed is presented below in Figure A1.



PRINCETON UNIVERSITY
gene expression project

Menu:

Databases:

Download microarray data from Notterman et. al. Cancer Research; Alon, et. al. PNAS; and Zhao et al, Genes & Dev.

Figures:

Enhanced figures from the Cancer Research paper.

Microarray Databases

For Notterman, et al*, Cancer Research:

- Carcinoma Data
- Adenoma Data

For Alon et al**, PNAS:

- Data

* Notterman, et al, *Cancer Research* vol. 61: 2001

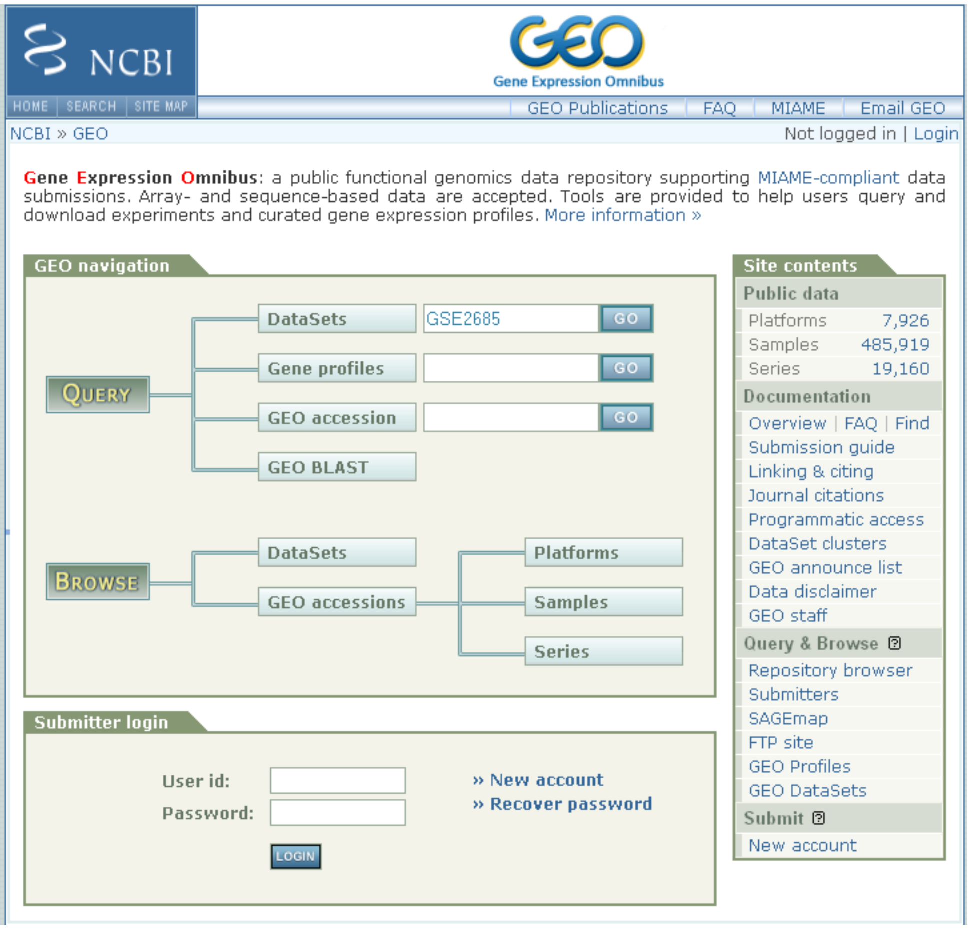
** Alon, et al, *Proc. Natl. Acad. Sci.* USA, Vol. 96: 6745-6750, 1999

Figure A1. Web repository of the data from the Princeton University gene expression project.

Data from Alon et al. (37) (Colon 1 database) were accessed originally in .txt format, then exported and edited using MS Excel. Information of the samples was also accessed in .txt format and paired with their correspondent samples following the description in http://genomics-pubs.princeton.edu/oncology/affydata/index.html. The result of this process was the matrix with readings of 62 tissues in 2,000 genes to be analyzed.

Data from Notterman et al. (57) (Colon 2 database) was available in the same repository that belongs to the Princeton University gene expression project. Data in the different states (healthy and cancer) was available in MS Excel format. All the information about samples and the accession numbers of the genes involved in the experiments were available in the same documents. Manipulation of this database was easier than that of Colon 1 database given the format of the original files.

The data of Hippo et al. (83) (Gastric database) was accessed through the Gene Expression Omnibus (GEO) repository. This repository is sponsored by the National Center of Biotechnology Information (NCBI) and can be located in http://www.ncbi.nlm.nih.gov/geo/. GEO is a powerful tool to access genetic data in many different ways. Figure 2 shows the main screen of the tool.



Gene Expression Omnibus: a public functional genomics data repository supporting MIAME-compliant data submissions. Array- and sequence-based data are accepted. Tools are provided to help users query and download experiments and curated gene expression profiles. More information »

DataSets: GSE2685

User id:

Password:

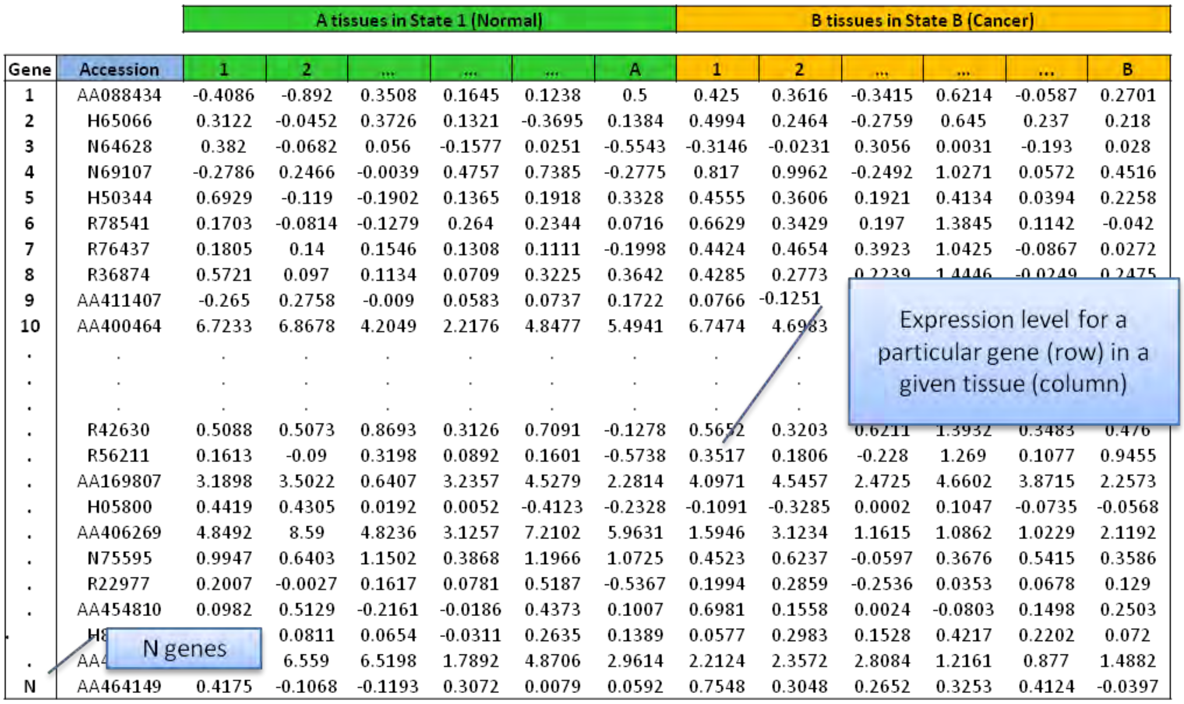
Public data	
Platforms	7,926
Samples	485,919
Series	19,160

Figure A2. Initial page for the GEO repository queries.

Different queries can be submitted to the repository through the use of their interface (Figure A2). A data set can be queried by their GSE (GEO Series) key, which is commonly provided in the associated publications. This kind of queries returns complete sets of experiments summarized and reported. Complete information regarding the different kinds of queries that can be performed through this tool can be consulted in http://www.ncbi.nlm.nih.gov/geo/info/faq.html#retrievals.

For the case of gastric cancer, the location of the raw dataset was not specified in the original publication, but it was found directly in the searching options of the repository. The number of the GEO Data Series was GSE2685. The reference of the platform used for the experiment was GPL80 corresponding to Hu6800-Affymetrix Human Full Length HuGeneFL Array. The 30 samples were identified with the labels GSM51763 to GSM51792.

When the data query is solved, GEO shows the arrays in different electronic formats like HTML, .xls, RTF, among others. Then, the user can choose the most convenient format to manage the data depending on the objectives of the analyses. In this case, all the accessed data was managed in MS Excel. The data was edited to resemble a matrix format containing information of the number of genes (N) involved in the experiments (Gene column), the accession number of each gene (Accession column), and all experimental readings of each gene on each tissue. The tissues were organized columns and the genes in rows. Samples in normal state were placed first (Normal1, Normal2,…, NormalA) and the cancer tissues were placed next (Cancer1, Cancer2,… CancerB). An illustrative example of a matrix of this sorts is presented in Figure 3, where the matrix has N = 7,129 genes explored, A = 5 normal samples, and B = 7 Cancer tissues. In order to avoid omissions, the labels used by the authors was kept throughout the analyses in this thesis.



		A tissues in State 1 (Normal)						B tissues in State B (Cancer)					
Gene	Accession	1	2	...	...	...	A	1	2	...	...	...	B
1	AA088434	-0.4086	-0.892	0.3508	0.1645	0.1238	0.5	0.425	0.3616	-0.3415	0.6214	-0.0587	0.2701
2	H65066	0.3122	-0.0452	0.3726	0.1321	-0.3695	0.1384	0.4994	0.2464	-0.2759	0.645	0.237	0.218
3	N64628	0.382	-0.0682	0.056	-0.1577	0.0251	-0.5543	-0.3146	-0.0231	0.3056	0.0031	-0.193	0.028
4	N69107	-0.2786	0.2466	-0.0039	0.4757	0.7385	-0.2775	0.817	0.9962	-0.2492	1.0271	0.0572	0.4516
5	H50344	0.6929	-0.119	-0.1902	0.1365	0.1918	0.3328	0.4555	0.3606	0.1921	0.4134	0.0394	0.2258
6	R78541	0.1703	-0.0814	-0.1279	0.264	0.2344	0.0716	0.6629	0.3429	0.197	1.3845	0.1142	-0.042
7	R76437	0.1805	0.14	0.1546	0.1308	0.1111	-0.1998	0.4424	0.4654	0.3923	1.0425	-0.0867	0.0272
8	R36874	0.5721	0.097	0.1134	0.0709	0.3225	0.3642	0.4285	0.2773	0.2239	1.4446	-0.0249	0.2475
9	AA411407	-0.265	0.2758	-0.009	0.0583	0.0737	0.1722	0.0766	-0.1251				
10	AA400464	6.7233	6.8678	4.2049	2.2176	4.8477	5.4941	6.7474	4.6983				
.	.	.	.	.	.	.	.	.	.				
.	.	.	.	.	.	.	.	.	.				
.	.	.	.	.	.	.	.	.	.				
.	R42630	0.5088	0.5073	0.8693	0.3126	0.7091	-0.1278	0.5652	0.3203	0.6211	1.3932	0.3483	0.476
.	R56211	0.1613	-0.09	0.3198	0.0892	0.1601	-0.5738	0.3517	0.1806	-0.228	1.269	0.1077	0.9455
.	AA169807	3.1898	3.5022	0.6407	3.2357	4.5279	2.2814	4.0971	4.5457	2.4725	4.6602	3.8715	2.2573
.	H05800	0.4419	0.4305	0.0192	0.0052	-0.4123	-0.2328	-0.1091	-0.3285	0.0002	0.1047	-0.0735	-0.0568
.	AA406269	4.8492	8.59	4.8236	3.1257	7.2102	5.9631	1.5946	3.1234	1.1615	1.0862	1.0229	2.1192
.	N75595	0.9947	0.6403	1.1502	0.3868	1.1966	1.0725	0.4523	0.6237	-0.0597	0.3676	0.5415	0.3586
.	R22977	0.2007	-0.0027	0.1617	0.0781	0.5187	-0.5367	0.1994	0.2859	-0.2536	0.0353	0.0678	0.129
.	AA454810	0.0982	0.5129	-0.2161	-0.0186	0.4373	0.1007	0.6981	0.1558	0.0024	-0.0803	0.1498	0.2503
.	H8		0.0811	0.0654	-0.0311	0.2635	0.1389	0.0577	0.2983	0.1528	0.4217	0.2202	0.072
.	AA4		6.559	6.5198	1.7892	4.8706	2.9614	2.2124	2.3572	2.8084	1.2161	0.877	1.4882
N	AA464149	0.4175	-0.1068	-0.1193	0.3072	0.0079	0.0592	0.7548	0.3048	0.2652	0.3253	0.4124	-0.0397

Expression level for a particular gene (row) in a given tissue (column)

N genes

Figure A3. Representation of the microarray matrix to be used in the execution of the methodology.

Appendix B – Computer programs

Appendix B1 Program For MW

The MatLab code to execute the Mann-Whitney statistical analysis to obtain *n* p_values corresponding to *n* genes, organized in two different states (normal and cancer), each state with more than 1 tissue (replicate).

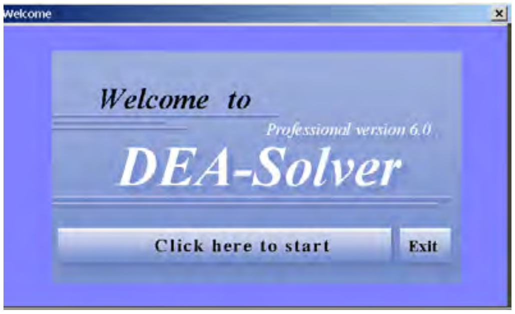
```
Function

[Filtrados
p_Value]=filtrado(N_genes,Matriz,Index_fil_sanos,Index_fil_cancer,alpha)

% Data selection, if needed
Datos=Matriz;
Datos(:,1)=[];
Fil_sanos=Datos(:,Index_fil_sanos);
Fil_cancer=Datos(:,Index_fil_cancer);

% Applying ranksum test overa all the genes
p_Value=zeros(N_genes,1);
H0=p_Value;
for i=1:N_genes
    [p_Value(i)
H0(i)]=ranksum(Fil_sanos(i,:),Fil_cancer(i,:),'alpha',alpha);
end

% Relevant genes are filtered (H0 = 1), if needed
Filtrados=Matriz;
%p_Value(H0==0,:)=[];
%Filtrados(H0==0,:)=[];
p_Value(H0==0&H0==1,:)=[];
Filtrados(H0==0&H0==1,:)=[];
```

This code is embedded in a suite with different processes of microarray data analysis that includes Gene Filtering, Separability and Classification of samples. The suite was developed in the Bio IE lab at UPRM.

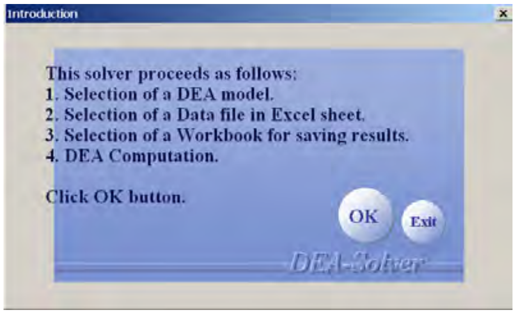
APPENDIX B2 – DATA ENVELOPMENT ANALYSIS (DEA) SOFTWARE

DEA-Solver Pro, a commercially available software, was used to apply DEA to the multiple criteria optimization problems posed in this thesis. DEA-Solver pro, is an MS Excel Add-In, that requires Windows XP. A brief explanation on how to use this tool is presented here.

DEA Solver Pro must be installed in the computer´s hard drive and used with a license dongle. The location of the installation folder is typically: C://SAITECH/DEA-Solver/DEA-SOLVERPRO-60Nd .xls. When accessing this address, the steps to run an analysis are shown below.



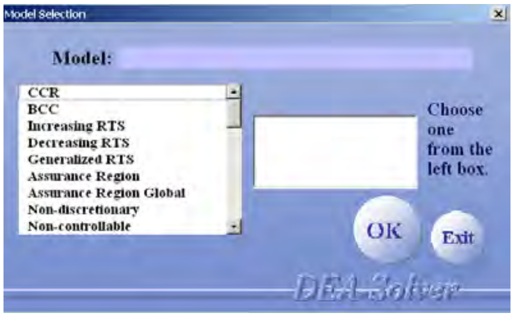
1.- Welcome screen for the DEA-Solver Pro tool in excel. When clicking start, the instructions are displayed



This solver proceeds as follows:
1. Selection of a DEA model.
2. Selection of a Data file in Excel sheet.
3. Selection of a Workbook for saving results.
4. DEA Computation.

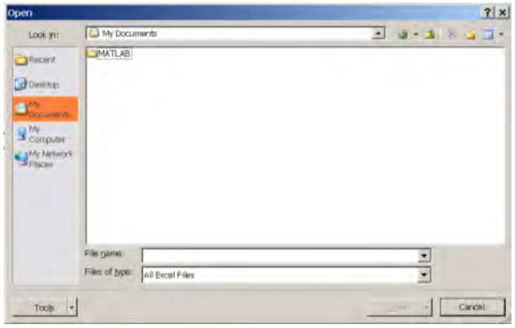
Click OK button.

2.- Instructions screen. Here, the different steps for the analysis through DEA-Solver Pro are described. Clicking OK is followed by the start of the process



Choose one from the left box.

3.- Model selection screen. Here the model selected to be applied to our data is selected. There are 157 available models, considering Input and Output orientations. In this work just the *BCC-I and BCC-O* have been chosen.



4.- Data Selection. In this step the program asks for the data to be used in the model execution. Data should be stored in .xls format, version 2003. Inputs and outputs to be considered should be identified with a previous (I) or (O) in their labels respectively.

If everything is correct with the data to be used, a dialog box appears where it is only necessary to click RUN.

When the size of the problem is greater than 1,000 DMUs to explore, a dialog box permits the user to request either a complete report or a summarized report. Depending upon the objectives of the study, one or the other can be chosen. In this thesis, all the reports were in the summarized format.

APPENDIX C – TABLES OF COMPLETE GENE SELECTION RESULTS

Table 1C. Resultant efficient genes for the MCO problem of Case 1 using Colon 1 database.

Gene	*Frontier*	*Accession*	*Symbol*	*Name*
1	1	M22382	HSPD1	Heat shock 60kDa protein 1 (chaperonin)
2	1	R87126		EST: yq31b10.s1
3	2	H08393	WDR77	WD repeat domain 77
4	2	R36977	GTF3A	General transcription factor IIIA
5	3	J05032	DARS	aspartyl-tRNA synthetase
6	3	M26383	IL8	interleukin 8
7	4	X63629	CDH3	cadherin 3, type 1, P-cadherin (placental)
8	4	H40095		EST: yn85b03.s1
9	4	Z50753	GUCA2B	guanylate cyclase activator 2B (uroguanylin)
10	4	M63391	DES	desmin
11	5	J02854	MYL9	myosin, light chain 9, regulatory
12	5	X12671	HNRNPA1	heterogeneous nuclear ribonucleoprotein A1
13	6	U09564	SRPK1	SFRS protein kinase 1
14	6	H43887	CFD	Complement factor D (adipsin)
15	6	M76378	CSRP1	cysteine and glycine-rich protein 1
16	7	M36634	VIP	vasoactive intestinal peptide
17	7	T86473	NME1	Non-metastatic cells 1, protein (NM23A) expressed in
18	8	H06524	GSN	Gelsolin
19	8	R84411	SNRPB	Small nuclear ribonucleoprotein polypeptides B and B1
20	8	X14958	HMGA1	high mobility group AT-hook 1
21	9	T92451	TPM2	Tropomyosin 2 (beta)
22	9	M26697	NPM1	Nucleophosmin (nucleolar phosphoprotein B23, numatrin)
23	9	T71025	MT1G	Metallothionein 1G
24	10	X86693	SPARCL1	SPARC-like 1 (hevin)
25	10	T47377	S100P	S100 calcium binding protein P
26	10	U30825	SRSF9	Serine/arginine-rich splicing factor 9
27	10	D31885	ARL6IP1	ADP-ribosylation factor-like 6 interacting protein 1

Table 2C. Resultant efficient genes for the MCO problem of Case 1 using Colon 2 database.

Amount	*Frontier*	*Accession*	*Symbol*	*Name*
1	1	Z50753	GUCA2B	guanylate cyclase activator 2B (uroguanylin)
2	1	Z49269	CCL14	chemokine (C-C motif) ligand 14
3	1	X64559	CLEC3B	C-type lectin domain family 3, member B
4	1	T96548	ACTG2	Actin, gamma 2, smooth muscle, enteric
5	1	T64297	FABP1	Fatty acid binding protein 1, liver
6	1	T55741	MYLK	Myosin light chain kinase
7	1	R36977	GTF3A	General transcription factor IIIA
8	1	M97496	GUCA2A	guanylate cyclase activator 2A (guanylin)
9	1	M77836	PYCR1	pyrroline-5-carboxylate reductase 1
10	1	J02854	MYL9	myosin, light chain 9, regulatory
11	1	H57136	FXYD1	FXYD domain containing ion transport regulator 1
12	2	X73502	KRT20	keratin 20
13	2	X56597	FBL	Fibrillarin
14	2	U37019	CNN1	calponin 1, basic, smooth muscle
15	2	U17077	MALL	mal, T-cell differentiation protein-like
16	2	R61502	TRAP1	TNF receptor-associated protein 1
17	2	R08183	HSPE1	Heat shock 10kDa protein 1 (chaperonin 10)
18	2	M83670	CA4	carbonic anhydrase IV
19	2	M36981	NME2	non-metastatic cells 2, protein (NM23B) expressed in
20	2	L29254	SORD	sorbitol dehydrogenase
21	2	H20426	NME1	Non-metastatic cells 1, protein (NM23A) expressed in
22	2	D63874	HMGB1	high-mobility group box 1
23	3	T76971	MT1F	Metallothionein 1F
24	3	T51961	PCNA	Proliferating cell nuclear antigen
25	3	T48804	RPS24	Ribosomal protein S24
26	3	M80244	SLC7A5	solute carrier family 7 (cationic amino acid transporter, y+ system), member 5
27	3	M76378	CSRP1	cysteine and glycine-rich protein 1
28	3	M63603	PLN	phospholamban
29	3	M26697	NPM1	nucleophosmin (nucleolar phosphoprotein B23, numatrin)
30	4	T86473	NME1	Non-metastatic cells 1, protein (NM23A)
31	4	T78104	PRELP	Proline/arginine-rich end leucine-rich repeat protein
32	4	T71025	MT1G	Metallothionein 1G
33	4	M36634	VIP	vasoactive intestinal peptide
34	4	M18079	FABP2	fatty acid binding protein 2, intestinal
35	4	H09351	MCM7	Minichromosome maintenance complex component 7
36	4	D00137	ADH1B	alcohol dehydrogenase 1B (class I), beta polypeptide
37	4	H06524	GSN	Gelsolin
38	5	X54942	CKS2	CDC28 protein kinase regulatory subunit 2
39	5	U34994	PRKDC	protein kinase, DNA-activated, catalytic polypeptide
40	5	U17899	CLNS1A	chloride channel, nucleotide-sensitive, 1A
41	5	T52362	CLNS1A	Chloride channel, nucleotide-sensitive, 1A
42	5	T51571	S100A11	S100 calcium binding protein A11

Table 2C (continued)

Amount	*Frontier*	*Accession*	*Symbol*	*Name*
43	5	M63391	DES	Desmin
44	5	M30448	CSNK2B	casein kinase 2, beta polypeptide
45	5	M12272	ADH1C	alcohol dehydrogenase 1C (class I), gamma polypeptide
46	5	L11708	HSD17B2	hydroxysteroid (17-beta) dehydrogenase 2
47	6	Z18951	CAV1	caveolin 1, caveolae protein, 22kDa
48	6	T49732	SNRPD2	Small nuclear ribonucleoprotein D2 polypeptide 16.5kDa
49	6	R71676	PA2G4	Proliferation-associated 2G4, 38kDa
50	6	H43887	CFD	Complement factor D (adipsin)
51	7	X12369	TPM1	tropomyosin 1 (alpha)
52	7	U22055	SND1	staphylococcal nuclease and tudor domain containing 1
53	7	U14631	HSD11B2	Hydroxysteroid (11-beta) dehydrogenase 2
54	7	T51913	CRYAB	Crystallin, alpha B
55	7	J03037	CA2	Carbonic anhydrase II
56	8	X52679	SMPD1	sphingomyelin phosphodiesterase 1, acid lysosomal
57	8	X16396	MTHFD2	methylenetetrahydrofolate dehydrogenase (NADP+ dependent) 2, methenyltetrahydrofolate cyclohydrolase
58	8	U33286	CSE1L	CSE1 chromosome segregation 1-like (yeast)
59	8	R88575	TKT	Transketolase
60	8	R87126		EST: yq31b10.s1
61	8	R11676	CDC20	Cell division cycle 20 homolog (S. cerevisiae)
62	8	M95936	AKT2	V-akt murine thymoma viral oncogene homolog 2
63	8	M37583	H2AFZ	H2A histone family, member Z
64	8	M22382	HSPD1	Heat shock 60kDa protein 1 (chaperonin)
65	8	H65842	ACP1	Acid phosphatase 1, soluble
66	9	X05231	MMP1	Matrix metallopeptidase 1 (interstitial collagenase)
67	9	R75843		EST: yi59f12.s1
68	9	M84526	CFD	complement factor D (adipsin)
69	9	L02785	SLC26A3	Solute carrier family 26, member 3
70	9	H20709	MYL6	Myosin, light chain 6, alkali, smooth muscle and non-muscle
71	10	Z46629	SOX9	SRY (sex determining region Y)-box 9
72	10	T98555	POLR1D	Polymerase (RNA) I polypeptide D, 16kDa
73	10	T95048	RPS15A	Ribosomal protein S15a
74	10	T46924	ABP1	Amiloride binding protein 1 (amine oxidase (copper-containing))
75	10	R50129	AT5G39220	Hydrolase, alpha/beta fold family protein
76	10	H54425		EST: yq91a08.s1
77	10	H29320	GNL3	Guanine nucleotide binding protein-like 3 (nucleolar)
78	10	D21262	NOLC1	Nucleolar and coiled-body phosphoprotein 1
79	10	D15049	PTPRH	Protein tyrosine phosphatase, receptor type, H

Table 3C. Resultant efficient genes for the MCO problem of Case 1 using the Gastric database.

Amount	*Frontier*	*Accession*	*Symbol*	*Name*
1	1	Z29074	KRT9	keratin 9
2	2	AC002077	GNAT1	guanine nucleotide binding protein (G protein), alpha transducing activity polypeptide 1
3	3	Z29574	TNFRSF17	tumor necrosis factor receptor superfamily, member 17
4	3	X76342	ADH7	alcohol dehydrogenase 7 (class IV), mu or sigma polypeptide
5	3	U21689	GSTP1	glutathione S-transferase pi 1
6	3	L13744	MLLT3	myeloid/lymphoid or mixed-lineage leukemia (trithorax homolog, Drosophila); translocated to, 3
7	3	J05401	CKMT2	creatine kinase, mitochondrial 2 (sarcomeric)
8	4	U57094	RAB27A	RAB27A, member RAS oncogene family
9	4	U50136	LTC4S	leukotriene C4 synthase
10	4	M75110	ATP4B	ATPase, H+/K+ exchanging, beta polypeptide
11	4	M63154	GIF	gastric intrinsic factor (vitamin B synthesis)
12	4	M62628		Human alpha-1 Ig germline C-region membrane-coding region, 3' end.
13	4	M31328	GNB3	guanine nucleotide binding protein (G protein), beta polypeptide 3
14	4	HG4051-HT4321_at	CHAT	Choline Acetyltransferase
15	4	HG2604-HT2700_at	PAN2	Pan-2
16	4	D29675		Homo sapiens inducible nitric oxide synthase gene, promoter and exon 1
17	5	Z49099	SMS	spermine synthase
18	5	X99101	ESR2	estrogen receptor 2 (ER beta)
19	5	X89750	TGIF1	TGFB-induced factor homeobox 1
20	5	X81817	BCAP31	B-cell receptor-associated protein 31
21	5	X76223		H.sapiens MAL gene exon 4
22	5	X54942	CKS2	CDC28 protein kinase regulatory subunit 2
23	5	X54941	CKS1B	CDC28 protein kinase regulatory subunit 1B
24	5	X54667	CST4	cystatin S
25	5	X53961	LTF	lactotransferrin
26	5	X05997	LIPF	lipase, gastric
27	5	U70663	KLF4	Kruppel-like factor 4 (gut)
28	5	U36759	PTCRA	pre T-cell antigen receptor alpha
29	5	U27325	TBXA2R	thromboxane A2 receptor
30	5	U21931	FBP1	fructose-1,6-bisphosphatase 1
31	5	U19948	PDIA2	protein disulfide isomerase family A, member 2
32	5	S68616	SLC9A1	solute carrier family 9 (sodium/hydrogen exchanger), member 1
33	5	S54005	TMSB10	thymosin beta 10
34	5	M63962	ATP4A	ATPase, H+/K+ exchanging, alpha polypeptide
35	5	M61855	CYP2C9	cytochrome P450, family 2, subfamily C, polypeptide 9
36	5	L38025	CNTFR	ciliary neurotrophic factor receptor
37	5	L17131	HMGA1	high mobility group AT-hook 1
38	5	L07592	PPARD	peroxisome proliferator-activated receptor delta
39	5	J04988	HSP90AB1	heat shock protein 90kDa alpha (cytosolic), class B member 1

Table 3C (continued)

Amount	***Frontier***	***Accession***	***Symbol***	***Name***
40	5	HG4272-HT4542_at	MET	Hepatocyte Growth Factor Receptor
41	5	HG3432-HT3618_at		Fibroblast Growth Factor Receptor K-Sam, Alt. Splice 1
42	5	D63874	HMGB1	high-mobility group box 1
43	5	D63479	DGKD	diacylglycerol kinase, delta 130kDa
44	5	D50914	BOP1	block of proliferation 1
45	5	D42047	GPD1L	glycerol-3-phosphate dehydrogenase 1-like
46	5	D26600	PSMB4	proteasome (prosome, macropain) subunit, beta type, 4
47	5	D26129	RNASE1	ribonuclease, RNase A family, 1 (pancreatic)
48	5	D21063	MCM2	minichromosome maintenance complex component 2
49	6	Z48314	MUC5AC	mucin 5AC, oligomeric mucus/gel-forming
50	6	X51698	TFF2	trefoil factor 2
51	6	M21259		Human small nuclear ribonucleoprotein (snRNP) E gene, 3' intergenic region.
52	6	L34587	TCEB1	transcription elongation factor B (SIII), polypeptide 1 (15kDa, elongin C)
53	7	U66052		Clone W2-6 mRNA from chromosome X
54	7	M95178	ACTN1	actinin, alpha 1
55	8	U18235	ABCA2, LACS2	ATP-binding cassette, sub-family A (ABC1), member 2. LACS2 (LONG-CHAIN ACYL-COA SYNTHETASE 2); long-chain-fatty-acid-CoA ligase
56	8	J05412	REG1A	regenerating islet-derived 1 alpha
57	8	HG2417-HT2513_at		Dynein, Heavy Chain, Cytoplasmic
58	8	D14695	HERPUD1	homocysteine-inducible, endoplasmic reticulum stress-inducible, ubiquitin-like domain member 1
59	9	S76942	DRD4	dopamine receptor D4
60	9	M93425	PTPN12	protein tyrosine phosphatase, non-receptor type 12
61	9	L40379	TRIP10	thyroid hormone receptor interactor 10
62	10	Z74616	COL1A2	Collagen, type I, alpha 2
63	10	Z74615	COL1A1	Collagen, type I, alpha 1
64	10	X74801	CCT3	Chaperonin containing TCP1, subunit 3 (gamma)
65	10	X65614	S100P	S100 calcium binding protein P
66	10	X52003	TFF1	Trefoil factor 1
67	10	X05409	ALDH2	Aldehyde dehydrogenase 2 family (mitochondrial)
68	10	U61397	SUMO1	SMT3 suppressor of mif two 3 homolog 1 (S. cerevisiae)
69	10	M86752	STIP1	Stress-induced-phosphoprotein 1
70	10	M12759		Human Ig J chain gene, exons 3 and 4
71	10	HG2702-HT2798_r_at		Serine/Threonine Kinase (Gb:Z25424)
72	10	HG2279-HT2375_at		Triosephosphate Isomerase
73	10	HG2148-HT2218_f_at		Mucin 3, Intestinal (Gb:M55406)
74	10	D50582		Homo sapiens gene for inward rectifier K channel, complete cds

Table 4C. Description of the 41 different genes found in the execution of Case 2 using Colon1 and Colon 2 databases.

Amount	*Frontier*	*Accession*	*Symbol*	*Name*
1	1	R36977	GTF3A	General transcription factor IIIA
2	1	R87126		EST: yq31b10.s1
3	2	H08393	WDR77	WD repeat domain 77
4	2	M22382	HSPD1	Heat shock 60kDa protein 1 (chaperonin)
5	2	Z50753	GUCA2B	Guanylate cyclase activator 2B (uroguanylin)
6	3	M26383	IL8	Interleukin 8
7	3	J02854	MYL9	Myosin, light chain 9, regulatory
8	3	H40095		EST: yn85b03.s1
9	4	J05032	DARS	Aspartyl-tRNA synthetase
10	4	M36634	VIP	Vasoactive intestinal peptide
11	4	X63629	CDH3	Cadherin 3, type 1, P-cadherin (placental)
12	4	H43887	CFD	Complement factor D (adipsin)
13	4	D63874	HMGB1	High-mobility group box 1
14	4	M26697	NPM1	Nucleophosmin (nucleolar phosphoprotein B23, numatrin)
15	5	R08183	HSPE1	Heat shock 10kDa protein 1 (chaperonin 10)
16	5	T86473	NME1	Non-metastatic cells 1, protein (NM23A) expressed in
17	5	X12671	locus	Human gene for heterogeneous nuclear ribonucleoprotein (hnRNP) core protein A1.
18	6	U09564	locus HSU09564	Human serine kinase mRNA, complete cds.
19	6	T71025	MT1G	Metallothionein 1G
20	7	U17899	CLNS1A	Chloride channel, nucleotide-sensitive, 1A
21	7	X54942	CKS2	CDC28 protein kinase regulatory subunit 2
22	7	M76378	locus HUMCRP04	Homo sapiens cysteine-rich protein (CRP) gene, exons 5, 6 and complete cds.
23	7	M63391	locus HUMDES	Homo sapiens desmin gene, complete cds.
24	8	H06524	GSN	Gelsolin
25	8	Z49269	locus Z49269	H.sapiens gene for chemokine HCC-1
26	8	T51571	S100A11	S100 calcium binding protein A11
27	8	X14958	HMGA1	High mobility group AT-hook 1
28	9	U14631	HSD11B2	Hydroxysteroid (11-beta) dehydrogenase 2
29	9	T47377	S100P	S100 calcium binding protein P
30	9	U22055	SND1	Staphylococcal nuclease and tudor domain containing 1
31	9	R84411	SNRPB	Small nuclear ribonucleoprotein polypeptides B and B1
32	9	X12466	SNRPE	Small nuclear ribonucleoprotein polypeptide E
33	9	M12272	ADH1C	Alcohol dehydrogenase 1C (class I), gamma polypeptide
34	9	U30825	SRSF9	Serine/arginine-rich splicing factor 9
35	9	D31885	ARL6IP1	ADP-ribosylation factor-like 6 interacting protein 1
36	9	T51023	HSP90AB1	Heat shock protein 90kDa alpha (cytosolic), class B member 1
37	10	T92451	TPM2	Tropomyosin 2 (beta)
38	10	H24030	CCT3	Chaperonin containing TCP1, subunit 3 (gamma)

Table 4C (continued)

Amount	*Frontier*	*Accession*	*Symbol*	*Name*
39	10	T78104	PRELP	Proline/arginine-rich end leucine-rich repeat protein
40	10	X55715	RPS3	Ribosomal protein S3
41	10	T48804	RPS24	Ribosomal protein S24

Table 5C. Complete list of the 85 genes found in the Case 3 execution with Colon 1 and Gastric Databases.

Amount	*Frontier*	*Accession*	*Symbol*	*Name*
1	1	X54942	CKS2	CDC28 protein kinase regulatory subunit 2
2	1	X54941	CKS1B	CDC28 protein kinase regulatory subunit 1B
3	1	M22382	HSPD1	Heat shock 60kDa protein 1 (chaperonin)
4	1	D63874	HMGB1	High-mobility group box 1
5	2	X63629	CDH3	Cadherin 3, type 1, P-cadherin (placental)
6	2	X55715	RPS3	Ribosomal protein S3
7	2	U26312	CBX3	Chromobox homolog 3
8	2	M36634	VIP	Vasoactive intestinal peptide
9	2	J05032	DARS	Aspartyl-tRNA synthetase
10	2	D42047	GPD1L	Glycerol-3-phosphate dehydrogenase 1-like
11	2	D38551	RAD21	RAD21 homolog (S. pombe)
12	2	D21261	TAGLN2	Transgelin 2
13	3	X15183	HSP90AA1	Heat shock protein 90kDa alpha (cytosolic), class A member 1
14	3	X12671	locus: X12671	Human gene for heterogeneous nuclear ribonucleoprotein (hnRNP) core protein A1
15	3	U25138	KCNMB1	Potassium large conductance calcium-activated channel, subfamily M, beta member 1
16	3	M63391	locus: HUMDES	Homo sapiens desmin gene, complete cds.
17	3	M16937	HOXB7	Homeobox B7
18	3	D26129	locus: HUMRNASA	Homo sapiens mRNA for ribonuclease A, complete cds
19	4	X87159	SCNN1B	Sodium channel, nonvoltage-gated 1, beta
20	4	X81817	locus: X81817	H.sapiens BAP31 mRNA.
21	4	U30825	SRSF9	Serine/arginine-rich splicing factor 9
22	4	U21090	POLD2	Polymerase (DNA directed), delta 2, regulatory subunit 50kDa
23	4	U09564	locus: HSU09564	Human serine kinase mRNA, complete cds
24	4	U05040	FUBP1	Far upstream element (FUSE) binding protein 1
25	4	U02493	NONO	Non-POU domain containing, octamer-binding
26	4	M15841	SNRPB2	Small nuclear ribonucleoprotein polypeptide B
27	4	L08069	DNAJA1	DnaJ (Hsp40) homolog, subfamily A, member 1
28	4	J02854	MYL9	Myosin, light chain 9, regulatory
29	5	Z25521	CD47	CD47 molecule
30	5	D00596	locus: HUMTS1	Homo sapiens gene for thymidylate synthase, complete cds.
31	5	X66839	CA9	Carbonic anhydrase IX
32	5	X53586	ITGA6	Integrin, alpha 6
33	5	M86752	STIP1	Stress-induced-phosphoprotein 1
34	5	M76378	locus: HUMCRP04	Homo sapiens cysteine-rich protein (CRP) gene, exons 5, 6 and complete cds
35	5	D43950	CCT5	Chaperonin containing TCP1, subunit 5 (epsilon)

Table 5C (continued)

Amount	*Frontier*	*Accession*	*Symbol*	*Name*
36	5	D26600	PSMB4	Proteasome (prosome, macropain) subunit, beta type, 4
37	6	Z49269	locus: Z49269	H.sapiens gene for chemokine HCC-1.
38	6	X07979	locus:X07979	Human mRNA for integrin beta 1 subunit.
39	6	U10324	ILF3	Interleukin enhancer binding factor 3, 90kDa
40	6	M93651	SET	SET nuclear oncogene
41	6	M34344	locus: HUMGPIIB3	Human platelet Glycoprotein IIb (GPIIb) gene, exon 30.
42	6	L07592	locus: HUMPPARA	Human peroxisome proliferator activated receptor mRNA, complete cds.
43	6	D31885	ARL6IP1	ADP-ribosylation factor-like 6 interacting protein 1
44	6	D31716	KLF9	Kruppel-like factor 9
45	6	D13315	GLO1	Glyoxalase I
46	7	X86693	SPARCL1	SPARC-like 1 (hevin)
47	7	X74295	ITGA7	Integrin, alpha 7
48	7	X72727	HNRNPK	Heterogeneous nuclear ribonucleoprotein K
49	7	D00761	PSMB1	Proteasome (prosome, macropain) subunit, beta type, 1
50	7	U32519	G3BP1	GTPase activating protein (SH3 domain) binding protein 1
51	7	U24166	MAPRE1	Microtubule-associated protein, RP/EB family, member 1
52	7	M82919	GABRB3	Gamma-aminobutyric acid (GABA) A receptor, beta 3
53	7	M23254	CAPN2	Calpain 2, (m/II) large subunit
54	7	L03840	FGFR4	Fibroblast growth factor receptor 4
55	8	X82103	COPB1	Coatomer protein complex, subunit beta 1
56	8	X75208	EPHB3	EPH receptor B3
57	8	X07290	ZNF3	Zinc finger protein 3
58	8	X05610	COL4A2	Collagen, type IV, alpha 2
59	8	U20998	SRP9	Signal recognition particle 9kDa
60	8	M95178	ACTN1	Actinin, alpha 1
61	8	M94556	SSBP1	Single-stranded DNA binding protein 1
62	8	M37583	H2AFZ	H2A histone family, member Z
63	8	M31303	locus:HUMOP18A	Human oncoprotein 18 (Op18) gene, complete cds.
64	8	J03040	SPARC	Secreted protein, acidic, cysteine-rich (osteonectin)
65	9	X74330	PRIM1	Primase, DNA, polypeptide 1 (49kDa)
66	9	X67155	KIF23	Kinesin family member 23
67	9	X13482	SNRPA1	Small nuclear ribonucleoprotein polypeptide A'
68	9	M31994	locus: HUMALDC13	Homo sapiens aldehyde dehydrogenase (ALDH1) gene, exon 13 and complete cds.
69	9	M26683	HUMIFNIND	Human interferon gamma treatment inducible mRNA.
70	9	M25809	ATP6V1B1	ATPase, H+ transporting, lysosomal 56/58kDa, V1 subunit B1
71	9	M23114	ATP2A2	ATPase, Ca++ transporting, cardiac muscle, slow twitch 2
72	9	L38951	KPNB1	Karyopherin (importin) beta 1
73	9	L12350	THBS2	Thrombospondin 2
74	9	L05144	locus: HUMPHOCAR	Homo sapiens (clone lamda-hPEC-3) phosphoenolpyruvate carboxykinase (PCK1) mRNA, complete cds.
75	9	D78134	CIRBP	Cold inducible RNA binding protein
76	10	U14577	MAP1A	Microtubule-associated protein 1A
77	10	D00760	PSMA2	Proteasome (prosome, macropain) subunit, alpha type, 2
78	10	U09587	GARS	Glycyl-tRNA synthetase
79	10	L38929	PTPRD	Protein tyrosine phosphatase, receptor type, D

Table 5C (continued)

Amount	*Frontier*	*Accession*	*Symbol*	*Name*
80	10	D00860	PRPS1	Phosphoribosyl pyrophosphate synthetase 1
81	10	L37112	AVPR1B	Arginine vasopressin receptor 1B
82	10	L07648	locus:HUMMXI1A	Human MXI1 mRNA, complete cds.
83	10	K03460	locus:HUMTUBA2H	Human alpha-tubulin isotype H2-alpha gene, last exon.
84	10	D59253	NCBP2	Nuclear cap binding protein subunit 2, 20kDa
85	10	D14695	HERPUD1	Homocysteine-inducible, endoplasmic reticulum stress-inducible, ubiquitin-like domain member 1

Table 6C. List of information for the 93 genes found through the execution of Case 3 using Colon 2 and Gastric Databases.

Amount	*Frontier*	*Accession*	*Symbol*	*Name*
1	1	Z29574	locus: Z29574	Homo sapiens gene for BCMA peptide.
2	1	Z29074	KRT9	Keratin 9
3	1	X64559	CLEC3B	C-type lectin domain family 3, member B
4	1	J05401	CKMT2	Creatine kinase, mitochondrial 2 (sarcomeric)
5	1	D63874	HMGB1	High-mobility group box 1
6	2	X54942	CKS2	CDC28 protein kinase regulatory subunit 2
7	2	U17077	MALL	Mal, T-cell differentiation protein-like
8	2	M97496	GUCA2A	Guanylate cyclase activator 2A (guanylin)
9	2	M63603	PLN	Phospholamban
10	2	D42047	GPD1L	Glycerol-3-phosphate dehydrogenase 1-like
11	3	X54941	CKS1B	CDC28 protein kinase regulatory subunit 1B
12	3	U33286	CSE1L	CSE1 chromosome segregation 1-like (yeast)
13	3	M86752	STIP1	Stress-induced-phosphoprotein 1
14	3	M75110	ATP4B	ATPase, H+/K+ exchanging, beta polypeptide
15	3	M36634	VIP	Vasoactive intestinal peptide
16	3	M22382	HSPD1	Heat shock 60kDa protein 1 (chaperonin)
17	3	L07592	locus: HUMPPARA	Human peroxisome proliferator activated receptor mRNA, complete cds.
18	3	J02854	MYL9	Myosin, light chain 9, regulatory
19	4	Z49269	locus: Z49269	H.sapiens gene for chemokine HCC-1.
20	4	M93651	SET	SET nuclear oncogene
21	4	M93425	PTPN12	Protein tyrosine phosphatase, non-receptor type 12
22	4	M84526	CFD	Complement factor D (adipsin)
23	4	M80244	SLC7A5	Solute carrier family 7 (cationic amino acid transporter, y+ system), member 5
24	4	M77836	PYCR1	Pyrroline-5-carboxylate reductase 1
25	4	L03840	FGFR4	Fibroblast growth factor receptor 4
26	4	D26600	PSMB4	Proteasome (prosome, macropain) subunit, beta type, 4
27	5	X87159	SCNN1B	Sodium channel, nonvoltage-gated 1, beta
28	5	M30448	locus: HUMCSK2B	Human casein kinase II beta subunit mRNA, complete cds.
29	5	M14745	BCL2	B-cell CLL/lymphoma 2
30	5	L11708	HSD17B2	Hydroxysteroid (17-beta) dehydrogenase 2
31	5	D26129	locus: HUMRNASA	Homo sapiens mRNA for ribonuclease A, complete cds

Table 6C (continued)

Amount	*Frontier*	*Accession*	*Symbol*	*Name*
32	5	D21262	NOLC1	Nucleolar and coiled-body phosphoprotein 1
33	6	Z49099	locus: Z49099	H.sapiens mRNA for spermine synthase.
34	6	X76223	locus: X76223	H.sapiens MAL gene exon 4
35	6	X57766	MMP1	Matrix metallopeptidase 11 (stromelysin 3)
36	6	X16396	MTHFD2	Methylenetetrahydrofolate dehydrogenase (NADP+ dependent) 2, methenyltetrahydrofolate cyclohydrolase
37	6	X15183	HSP90AA1	Heat shock protein 90kDa alpha (cytosolic), class A member 1
38	6	M95936	AKT2	V-akt murine thymoma viral oncogene homolog 2
39	6	L22524	locus: HUMMATRY06	Human matrilysin gene, exon 6 and complete cds.
40	6	D31766	GNPDA1	Glucosamine-6-phosphate deaminase 1
41	6	D25218	RRS1	RRS1 ribosome biogenesis regulator homolog (S. cerevisiae)
42	6	D21261	TAGLN2	Transgelin 2
43	7	X81817	locus: X81817	H.sapiens BAP31 mRNA.
44	7	X66079	SPIB	Spi-B transcription factor (Spi-1/PU.1 related)
45	7	X54489	locus: X54489	Human gene for melanoma growth stimulatory activity (MGSA).
46	7	U26312	CBX3	Chromobox homolog 3
47	7	U25138	KCNMB1	Potassium large conductance calcium-activated channel, subfamily M, beta member 1
48	7	U09564	locus: HSU09564	Human serine kinase mRNA, complete cds
49	7	M80899	AHNAK	AHNAK nucleoprotein
50	7	M18079	HUMFABP	Human, intestinal fatty acid binding protein gene, complete cds, and an Alu repetitive element.
51	7	L02785	SLC26A3	Solute carrier family 26, member 3
52	7	J05032	DARS	Aspartyl-tRNA synthetase
53	7	J03507	C7	Complement component 7
54	7	D13315	GLO1	Glyoxalase I
55	8	X85740	locus: X85740	H.sapiens mRNA for C-C chemokine receptor-4.
56	8	X54162	LMOD1	Leiomodin 1 (smooth muscle)
57	8	X12671	locus: X12671	Human gene for heterogeneous nuclear ribonucleoprotein (hnRNP) core protein A1
58	8	U22055	SND1	Staphylococcal nuclease and tudor domain containing 1
59	8	U05259	locus: HSU05259	Human MB-1 gene, complete cds.
60	8	M77349	TGFBI	Transforming growth factor, beta-induced, 68kDa
61	8	M76378	locus: HUMCRP04	Homo sapiens cysteine-rich protein (CRP) gene, exons 5, 6 and complete cds
62	8	M63962	locus: HUMHKATPC	Human gastric H,K-ATPase catalytic subunit gene, complete cds.
63	8	M63391	locus: HUMDES	Homo sapiens desmin gene, complete cds.
64	8	M31328	GNB3	Guanine nucleotide binding protein (G protein), beta polypeptide 3
65	8	M26683	HUMIFNIND	Human interferon gamma treatment inducible mRNA.
66	8	M23254	CAPN2	Calpain 2, (m/II) large subunit
67	8	D78134	CIRBP	Cold inducible RNA binding protein
68	9	X66839	CA9	Carbonic anhydrase IX
69	9	X54667	CST4	Cystatin S
70	9	M61832	AHCY	Adenosylhomocysteinase
71	9	M24486	P4HA1	Prolyl 4-hydroxylase, alpha polypeptide I
72	9	M16801	NR3C2	Nuclear receptor subfamily 3, group C, member 2

Table 6C (Continued)

Amount	*Frontier*	*Accession*	*Symbol*	*Name*
73	9	L13744	MLLT3	Myeloid/lymphoid or mixed-lineage leukemia (trithorax homolog, Drosophila); translocated to, 3
74	9	L08069	DNAJA1	DnaJ (Hsp40) homolog, subfamily A, member 1
75	9	L07590	PPP2R3A	Protein phosphatase 2, regulatory subunit B'', alpha
76	9	J05272	IMPDH1	IMP (inosine 5'-monophosphate) dehydrogenase 1
77	9	J00123	locus: AH005276S2	Homo sapiens preproenkephalin precursor (PEN) gene, exon 3 and complete cds.
78	9	D43772	GRB7	Growth factor receptor-bound protein 7
79	9	D31716	KLF9	Kruppel-like factor 9
80	9	D15049	PTPRH	Protein tyrosine phosphatase, receptor type, H
81	10	Z31695	INPP5A	Inositol polyphosphate-5-phosphatase, 40kDa
82	10	Y00285	IGF2R	Insulin-like growth factor 2 receptor
83	10	X06323	MRPL3	Mitochondrial ribosomal protein L3
84	10	U21931	locus: HSLFBPS7	Human fructose-1,6-biphosphatase (FBP1) gene, exon 7, and complete cds.
85	10	U14577	MAP1A	Microtubule-associated protein 1A
86	10	U05040	FUBP1	Far upstream element (FUSE) binding protein 1
87	10	M96956	TDGF3	Teratocarcinoma-derived growth factor 3, pseudogene
88	10	M64673	HSF1	Heat shock transcription factor 1
89	10	M63928	CD27	CD27 molecule
90	10	M34181	PRKACB	Protein kinase, cAMP-dependent, catalytic, beta
91	10	L23808	MMP12	Matrix metallopeptidase 12 (macrophage elastase)
92	10	L20298	CBFB	Core-binding factor, beta subunit
93	10	D14695	HERPUD1	Homocysteine-inducible, endoplasmic reticulum stress-inducible, ubiquitin-like domain member 1

Table 7C. List of 222 genes obtained from the Case 3 executed with Colon 1, Colon 2 and Gastric databases.

Gene	*Frontier*	*Accession*	*Symbol*	*Name*
1	1	X54942	CKS2	CDC28 protein kinase regulatory subunit 2
2	1	X54941	CKS1B	CDC28 protein kinase regulatory subunit 1B
3	1	M36634	VIP	Vasoactive intestinal peptide
4	1	M22382	HSPD1	Heat shock 60kDa protein 1 (chaperonin)
5	1	J02854	MYL9	Myosin, light chain 9, regulatory
6	1	D63874	HMGB1	High-mobility group box 1
7	2	Z49269	locus: Z49269	H.sapiens gene for chemokine HCC-1.
8	2	X87159	SCNN1B	Sodium channel, nonvoltage-gated 1, beta
9	2	X63629	CDH3	Cadherin 3, type 1, P-cadherin (placental)
10	2	X55715	RPS3	Ribosomal protein S3
11	2	X15183	HSP90AA1	Heat shock protein 90kDa alpha (cytosolic), class A member 1
12	2	U26312	CBX3	Chromobox homolog 3
13	2	U25138	KCNMB1	Potassium large conductance calcium-activated channel, subfamily M, beta member 1

Table 7C (Continued)

Gene	*Frontier*	*Accession*	*Symbol*	*Name*
14	2	U09564	locus: HSU09564	Human serine kinase mRNA, complete cds.
15	2	L03840	FGFR4	Fibroblast growth factor receptor 4
16	2	J05032	DARS	Aspartyl-tRNA synthetase
17	2	D42047	GPD1L	Glycerol-3-phosphate dehydrogenase 1-like
18	2	D38551	RAD21	RAD21 homolog (S. pombe)
19	2	D21261	TAGLN2	Transgelin 2
20	3	X12671	locus: X12671	Human gene for heterogeneous nuclear ribonucleoprotein (hnRNP) core protein A1.
21	3	U22055	SND1	Staphylococcal nuclease and tudor domain containing 1
22	3	U05040	FUBP1	Far upstream element (FUSE) binding protein 1
23	3	M93651	SET	SET nuclear oncogene
24	3	M86752	STIP1	Stress-induced-phosphoprotein 1
25	3	M76378	locus: HUMCRP04	Homo sapiens cysteine-rich protein (CRP) gene, exons 5, 6 and complete cds.
26	3	M63391	locus: HUMDES	Homo sapiens desmin gene, complete cds.
27	3	M16937	HOXB7	Homeobox B7
28	3	M15841	SNRPB2	Small nuclear ribonucleoprotein polypeptide B
29	3	L08069	DNAJA1	DnaJ (Hsp40) homolog, subfamily A, member 1
30	3	L07592	locus: HUMPPARA	Human peroxisome proliferator activated receptor mRNA, complete cds.
31	3	D31716	KLF9	Kruppel-like factor 9
32	3	D13315	GLO1	Glyoxalase I
33	3	D26600	PSMB4	Proteasome (prosome, macropain) subunit, beta type, 4
34	3	D26129	locus: HUMRNASA	Homo sapiens mRNA for ribonuclease A, complete cds.
35	4	X81817	locus: X81817	H.sapiens BAP31 mRNA.
36	4	D00596	locus: HUMTS1	Homo sapiens gene for thymidylate synthase, complete cds.
37	4	X74295	ITGA7	Integrin, alpha 7
38	4	X66839	CA9	Carbonic anhydrase IX
39	4	D00761	PSMB1	Proteasome (prosome, macropain) subunit, beta type, 1
40	4	X07767	PRKACA	Protein kinase, cAMP-dependent, catalytic, alpha
41	4	X05610	COL4A2	Collagen, type IV, alpha 2
42	4	U30825	SRSF9	Serine/arginine-rich splicing factor 9
43	4	U21090	POLD2	Polymerase (DNA directed), delta 2, regulatory subunit 50kDa
44	4	U10324	ILF3	Interleukin enhancer binding factor 3, 90kDa
45	4	U02493	NONO	Non-POU domain containing, octamer-binding
46	4	M94556	SSBP1	Single-stranded DNA binding protein 1
47	4	M64673	HSF1	Heat shock transcription factor 1
48	4	M26683	locus: HUMIFNIND	Human interferon gamma treatment inducible mRNA.
49	4	L07648	locus: HUMMXI1A	Human MXI1 mRNA, complete cds.
50	4	J03040	SPARC	Secreted protein, acidic, cysteine-rich (osteonectin)
51	4	D78134	CIRBP	Cold inducible RNA binding protein
52	4	D43950	CCT5	Chaperonin containing TCP1, subunit 5 (epsilon)
53	4	D31885	ARL6IP1	ADP-ribosylation factor-like 6 interacting protein 1
54	4	D15049	PTPRH	Protein tyrosine phosphatase, receptor type, H
55	5	Z25521	CD47	CD47 molecule
56	5	Z23064	RBMX	RNA binding motif protein, X-linked

Table 7C (continued)

Gene	*Frontier*	*Accession*	*Symbol*	*Name*
57	5	X86693	SPARCL1	SPARC-like 1 (hevin)
58	5	X76057	MPI	Mannose phosphate isomerase
59	5	X53586	ITGA6	Integrin, alpha 6
60	5	X07979	locus: X07979	Human mRNA for integrin beta 1 subunit.
61	5	U24166	MAPRE1	Microtubule-associated protein, RP/EB family, member 1
62	5	U20998	SRP9	Signal recognition particle 9kDa
63	5	U14577	MAP1A	Microtubule-associated protein 1A
64	5	M91463	locus: HUMGLUT4B	Human glucose transporter (GLUT4) gene, complete cds.
65	5	M83751	MANF	Mesencephalic astrocyte-derived neurotrophic factor
66	5	M37583	H2AFZ	H2A histone family, member Z
67	5	M34344	locus: HUMGPIIB3	Human platelet Glycoprotein IIb (GPIIb) gene, exon 30.
68	5	M31994	locus: HUMALDC13	Homo sapiens aldehyde dehydrogenase (ALDH1) gene, exon 13 and complete cds.
69	5	M31303	locus: HUMOP18A	Human oncoprotein 18 (Op18) gene, complete cds.
70	5	M23254	CAPN2	Calpain 2, (m/II) large subunit
71	5	M23114	ATP2A2	ATPase, Ca++ transporting, cardiac muscle, slow twitch 2
72	5	L38951	KPNB1	Karyopherin (importin) beta 1
73	5	L28010	HNRNPF	Heterogeneous nuclear ribonucleoprotein F
74	5	L12350	THBS2	Thrombospondin 2
75	5	K03460	locus: HUMTUBA2H	Human alpha-tubulin isotype H2-alpha gene, last exon.
76	5	J05272	IMPDH1	IMP (inosine 5'-monophosphate) dehydrogenase 1
77	6	Y00815	locus:Y00815	Human mRNA for LCA-homolog. LAR protein (leukocyte antigen related).
78	6	X82103	COPB1	Coatomer protein complex, subunit beta 1
79	6	D00760	PSMA2	Proteasome (prosome, macropain) subunit, alpha type, 2
80	6	X77548	NCOA4	Nuclear receptor coactivator 4
81	6	X74330	PRIM1	Primase, DNA, polypeptide 1 (49kDa)
82	6	X72727	HNRNPK	Heterogeneous nuclear ribonucleoprotein K
83	6	X67155	KIF23	Kinesin family member 23
84	6	X15882	locus: AY029208	Homo sapiens type VI collagen alpha 2 chain precursor (COL6A2) mRNA, complete cds, alternatively spliced.
85	6	X13482	SNRPA1	Small nuclear ribonucleoprotein polypeptide A'
86	6	D12686	EIF4G1	Eukaryotic translation initiation factor 4 gamma, 1
87	6	X07290	ZNF3	Zinc finger protein 3
88	6	U32519	G3BP1	GTPase activating protein (SH3 domain) binding protein 1
89	6	M96233	locus:HUMGSTM4A	Human glutathione transferase class mu number 4 (GSTM4) gene, complete cds.
90	6	M95178	ACTN1	Actinin, alpha 1
91	6	M82919	GABRB3	Gamma-aminobutyric acid (GABA) A receptor, beta 3
92	6	M25809	ATP6V1B1	ATPase, H+ transporting, lysosomal 56/58kDa, V1 subunit B1
93	6	M24069	CSDA	Cold shock domain protein A
94	6	L38929	PTPRD	Protein tyrosine phosphatase, receptor type, D
95	6	L37936	TSFM	Ts translation elongation factor, mitochondrial
96	6	L37112	AVPR1B	Arginine vasopressin receptor 1B
97	6	L05144	locus: HUMPHOCAR	Homo sapiens (clone lamda-hPEC-3) phosphoenolpyruvate carboxykinase (PCK1) mRNA, complete cds.

Table 7C (continued)

Gene	*Frontier*	*Accession*	*Symbol*	*Name*
98	6	D43947	KIAA0100	KIAA0100
99	6	D16294	ACAA2	Acetyl-CoA acyltransferase 2
100	6	D14695	HERPUD1	Homocysteine-inducible, endoplasmic reticulum stress-inducible, ubiquitin-like domain member 1
101	6	D14812	MORF4L2	Mortality factor 4 like 2
102	7	Z30644	CLCNKB	Chloride channel Kb
103	7	X80507	locus: X80507	H.sapiens YAP65 mRNA.
104	7	D00762	PSMA3	Proteasome (prosome, macropain) subunit, alpha type, 3
105	7	D00763	PSMA4	Proteasome (prosome, macropain) subunit, alpha type, 4
106	7	X75208	EPHB3	EPH receptor B3
107	7	X71490	locus: X71490	H.sapiens mRNA for vacuolar proton ATPase, subunit D.
108	7	X01060	TFRC	Transferrin receptor (p90, CD71)
109	7	U12255	FCGRT	Fc fragment of IgG, receptor, transporter, alpha
110	7	U09587	GARS	Glycyl-tRNA synthetase
111	7	M88468	MVK	Mevalonate kinase
112	7	M88108	KHDRBS1	KH domain containing, RNA binding, signal transduction associated 1
113	7	M85289	HSPG2	Heparan sulfate proteoglycan 2
114	7	M68520	CDK2	Cyclin-dependent kinase 2
115	7	M31627	XBP1	X-box binding protein 1
116	7	M24470	GMPR	Guanosine monophosphate reductase
117	7	M22490	BMP4	Bone morphogenetic protein 4
118	7	M21984	TNNT3	Troponin T type 3 (skeletal, fast)
119	7	M14539	F13A1	Coagulation factor XIII, A1 polypeptide
120	7	L41559	PCBD1	Pterin-4 alpha-carbinolamine dehydratase/dimerization cofactor of hepatocyte nuclear factor 1 alpha
121	7	L21993	ADCY2	Adenylate cyclase 2 (brain)
122	7	L20859	SLC20A1	Solute carrier family 20 (phosphate transporter), member 1
123	7	L12723	HSPA4	Heat shock 70kDa protein 4
124	7	J04794	AKR1A1	Aldo-keto reductase family 1, member A1 (aldehyde reductase)
125	7	D59253	NCBP2	Nuclear cap binding protein subunit 2, 20kDa
126	8	Z19002	ZBTB16	Zinc finger and BTB domain containing 16
127	8	D00860	PRPS1	Phosphoribosyl pyrophosphate synthetase 1
128	8	X89986	locus: X89986	H.sapiens mRNA for NBK apoptotic inducer protein.
129	8	X85750	MMD	Monocyte to macrophage differentiation-associated
130	8	X83301	SMA5	Glucuronidase, beta pseudogene
131	8	X70944	SFPQ	Splicing factor proline/glutamine-rich
132	8	X70040	MST1R	Macrophage stimulating 1 receptor (c-met-related tyrosine kinase)
133	8	X64364	BSG	Basigin (Ok blood group)
134	8	X57206	locus: X57206	H.sapiens mRNA for 1D-myo-inositol-trisphosphate 3-kinase B isoenzyme.
135	8	X15880	COL6A1	Collagen, type VI, alpha 1
136	8	D13641	TOMM20	Translocase of outer mitochondrial membrane 20 homolog (yeast)
137	8	X06700	COL3A1	Collagen, type III, alpha 1
138	8	U29175	SMARCA4	SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily a, member 4

Table 7C (continued)

Gene	*Frontier*	*Accession*	*Symbol*	*Name*
139	8	U28686	RBM3	RNA binding motif (RNP1, RRM) protein 3
140	8	U05572	MAN2B1	Mannosidase, alpha, class 2B, member 1
141	8	M96326	locus: HUMAZCDI	Human azurocidin gene, complete cds.
142	8	D14662	PRDX6	Peroxiredoxin 6
143	8	M94630	HNRNPD	Heterogeneous nuclear ribonucleoprotein D (AU-rich element RNA binding protein 1, 37kDa)
144	8	M91670	UBE2S	Ubiquitin-conjugating enzyme E2S
145	8	M86868	GABRR2	Gamma-aminobutyric acid (GABA) receptor, rho 2
146	8	M86737	SSRP1	Structure specific recognition protein 1
147	8	M85085	CSTF2	Cleavage stimulation factor, 3' pre-RNA, subunit 2, 64kDa
148	8	M35531	FUT1	Fucosyltransferase 1 (galactoside 2-alpha-L-fucosyltransferase, H blood group)
149	8	L24203	TRIM29	Tripartite motif-containing 29
150	8	L09604	PLP2	Proteolipid protein 2 (colonic epithelium-enriched)
151	8	L06132	VDAC1	Voltage-dependent anion channel 1
152	8	J03827	YBX1	Y box binding protein 1
153	8	D37931	ANG	Angiogenin, ribonuclease, RNase A family, 5
154	8	D31883	ABLIM1	Actin binding LIM protein 1
155	8	D25217	MLC1	Megalencephalic leukoencephalopathy with subcortical cysts 1
156	9	Z24727	TPM1	Tropomyosin 1 (alpha)
157	9	X90858	UPP1	Uridine phosphorylase 1
158	9	X87838	CTNNB1	Catenin (cadherin-associated protein), beta 1, 88kDa
159	9	X82895	locus: X82895	H.sapiens mRNA for DLG2.DLG2 gene; tumor supressor gene
160	9	X80026	BCAM	Basal cell adhesion molecule (Lutheran blood group)
161	9	D13627	CCT8	Chaperonin containing TCP1, subunit 8 (theta)
162	9	X78549	PTK6	PTK6 protein tyrosine kinase 6
163	9	X74262	RBBP4	Retinoblastoma binding protein 4
164	9	X73358	AES	Amino-terminal enhancer of split
165	9	X68688	locus: X68688	H.sapiens ZNF33B gene. Kruppel-related protein; zinc finger protein; ZNF33B gene.
166	9	X65488	HNRNPU	Heterogeneous nuclear ribonucleoprotein U (scaffold attachment factor A)
167	9	X57346	YWHAB	Tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein, beta polypeptide
168	9	D14520	KLF5	Kruppel-like factor 5 (intestinal)
169	9	X54871	RAB5B	RAB5B, member RAS oncogene family
170	9	X16663	HCLS1	Hematopoietic cell-specific Lyn substrate 1
171	9	D14663	PSMD6	Proteasome (prosome, macropain) 26S subunit, non-ATPase, 6
172	9	X16354	CEACAM1	Carcinoembryonic antigen-related cell adhesion molecule 1 (biliary glycoprotein)
173	9	X02875	OAS1	2',5'-oligoadenylate synthetase 1, 40/46kDa
174	9	X02761	FN1	Fibronectin 1
175	9	U30872	CENPF	Centromere protein F, 350/400kDa (mitosin)
176	9	U28963	GPS2	G protein pathway suppressor 2
177	9	U14971	RPS9	Ribosomal protein S9
178	9	U01038	PLK1	Polo-like kinase 1
179	9	M95627	AAMP	Angio-associated, migratory cell protein

Table 7C (continued)

Gene	*Frontier*	*Accession*	*Symbol*	*Name*
180	9	M88279	FKBP4	FK506 binding protein 4, 59kDa
181	9	M84739	CALR	Calreticulin
182	9	M75126	HK1	Hexokinase 1
183	9	M32313	SRD5A1	Steroid-5-alpha-reductase, alpha polypeptide 1 (3-oxo-5 alpha-steroid delta 4-dehydrogenase alpha 1)
184	9	M31516	CD55	CD55 molecule, decay accelerating factor for complement (Cromer blood group)
185	9	M28882	MCAM	Melanoma cell adhesion molecule
186	9	M27749	IGLL1	Immunoglobulin lambda-like polypeptide 1
187	9	M22632	GOT2	Glutamic-oxaloacetic transaminase 2, mitochondrial (aspartate aminotransferase 2)
188	9	M19045	LYZ	Lysozyme
189	9	L19437	TALDO1	Transaldolase 1
190	9	J02645	EIF2S1	Eukaryotic translation initiation factor 2, subunit 1 alpha, 35kDa
191	10	Z48541	PTPRO	Protein tyrosine phosphatase, receptor type, O
192	10	Z35093	SURF1	Surfeit 1
193	10	Z17227	IL10RB	Interleukin 10 receptor, beta
194	10	Y00264	APP	Amyloid beta (A4) precursor protein
195	10	Y00097	ANXA6	Annexin A6
196	10	X81372	BPHL	Biphenyl hydrolase-like (serine hydrolase)
197	10	X73882	MAP7	Microtubule-associated protein 7
198	10	X68314	GPX2	Glutathione peroxidase 2 (gastrointestinal)
199	10	X68277	DUSP1	Dual specificity phosphatase 1
200	10	X64037	GTF2F1	General transcription factor IIF, polypeptide 1, 74kDa
201	10	X62153	MCM3	Minichromosome maintenance complex component 3
202	10	X62048	WEE1	WEE1 homolog (S. pombe)
203	10	X02152	LDHA	Lactate dehydrogenase A
204	10	U07664	locus: HSHB9HB2	Homo sapiens HB9 homeobox gene, exons 2 and 3 and complete cds.
205	10	U04241	AES	Amino-terminal enhancer of split
206	10	U01877	EP300	E1A binding protein p300
207	10	M95678	PLCB2	Phospholipase C, beta 2
208	10	M77698	YY1	YY1 transcription factor
209	10	M73481	GRPR	Gastrin-releasing peptide receptor
210	10	M59807	IL32	Interleukin 32
211	10	M22538	NDUFV2	NADH dehydrogenase (ubiquinone) flavoprotein 2, 24kDa
212	10	L40992	RUNX2	Runt-related transcription factor 2
213	10	L38696	RALY	RNA binding protein, autoantigenic (hnRNP-associated with lethal yellow homolog (mouse))
214	10	L35545	PROCR	Protein C receptor, endothelial
215	10	L32163	ZNF197	Zinc finger protein 197
216	10	L10284	CANX	Calnexin
217	10	L06895	MXD1	MAX dimerization protein 1
218	10	L02426	PSMC1	Proteasome (prosome, macropain) 26S subunit, ATPase, 1
219	10	K03192	CYP2A7	Cytochrome P450, family 2, subfamily A, polypeptide 7
220	10	J03069	locus: HUMMYCL2A	Human MYCL2 gene, complete cds. c-myc proto-oncogene; proto-oncogene; repeat region.

Table 7C (continued)

Gene	*Frontier*	*Accession*	*Symbol*	*Name*
221	10	D63878	SEP2	Septin 2
222	10	D49396	PRDX3	Peroxiredoxin 3

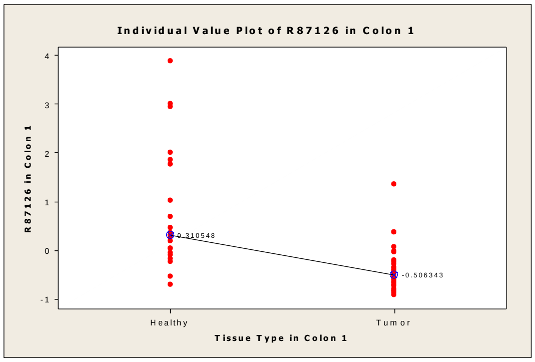
Appendix D - Graphical support of Validation Scheme 2

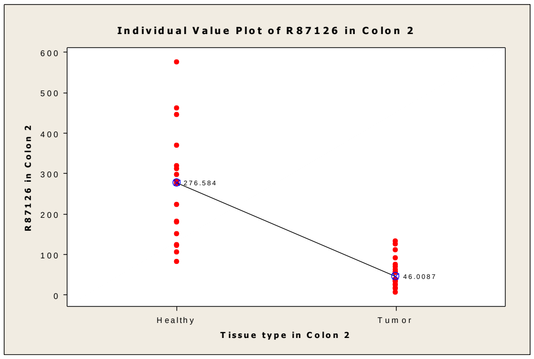
This appendix shows the Individual value plots for the gene expression values organized in different states (e.g. healthy and cancer) for a particular gene. These plots help to visualize the direction of change in the value of the medians between the states involved.

These graphs support the profiles generated in the Validation Scheme 2 in Chapter 4.

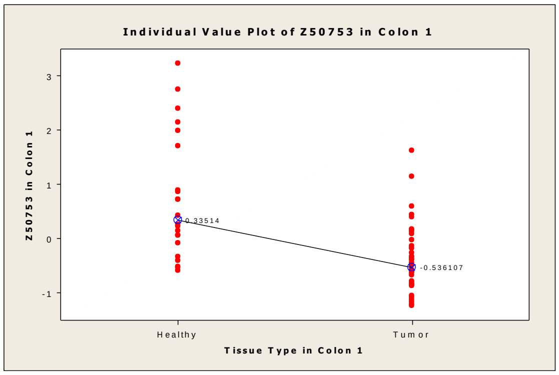
Graphs supporting Expression Profile for Group 1

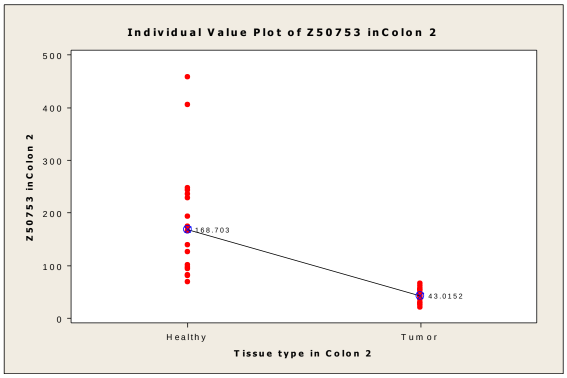
EST: yq31b10.s1 consistently underexpressed in Colon 1 and Colon 2 databases



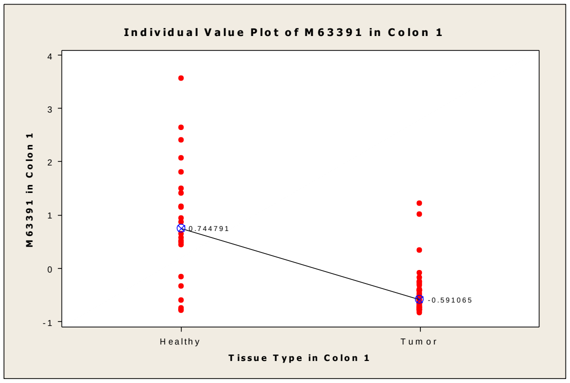


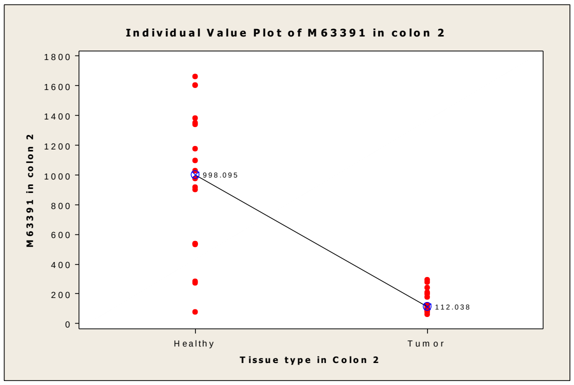
Gene GUCA2B (accession Z50753) consistently underexpressed in Colon 1 and Colon 2 databases



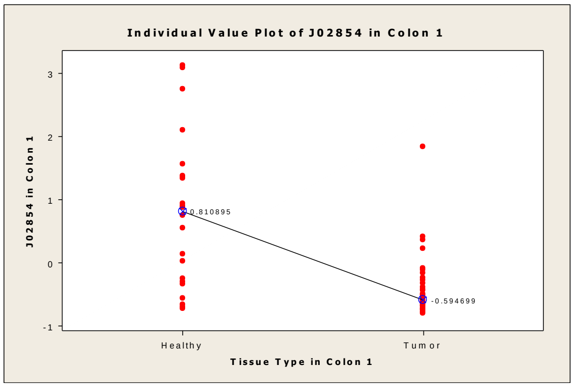


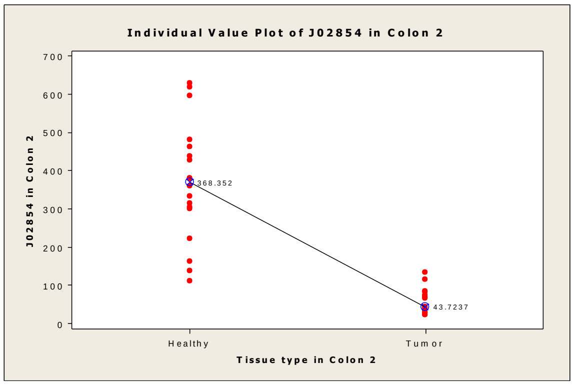
Gene Desmin (accession M63391) consistently underexpressed in Colon 1 and Colon 2 databases



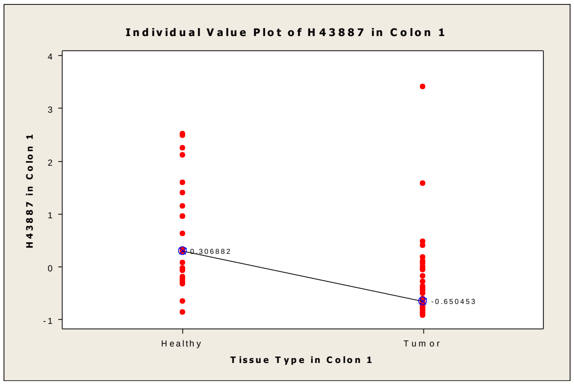


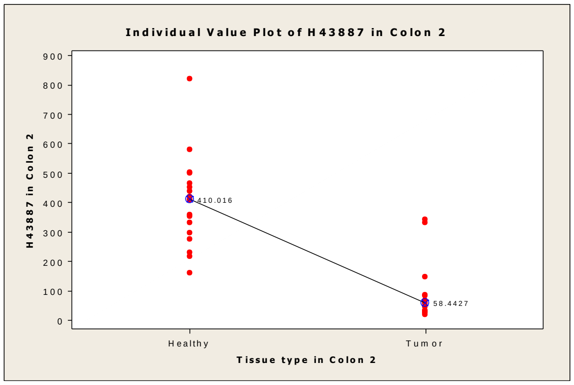
Gene MYL9 (accession J02854) consistently underexpressed in Colon 1 and Colon 2 databases



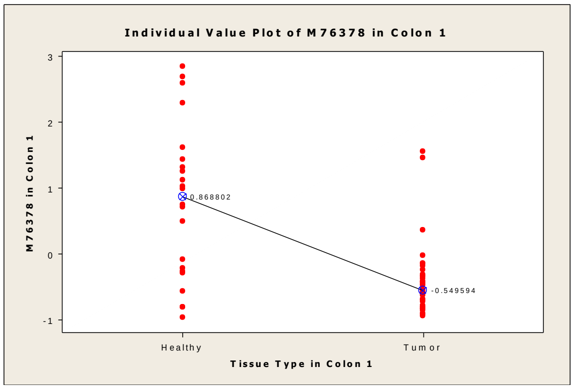


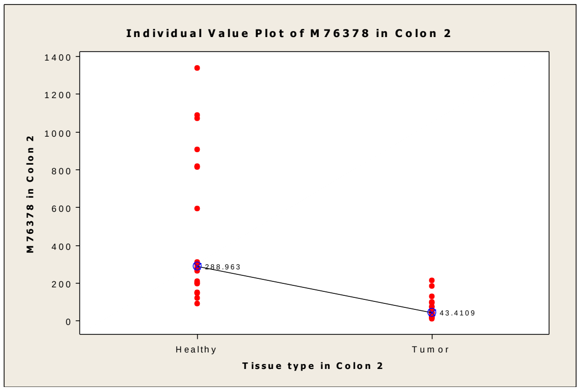
Gene CFD (adipsin), accession H43887, consistently underexpressed in Colon 1 and Colon 2 databases



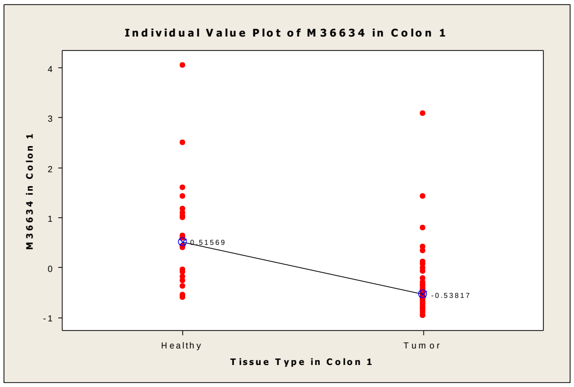


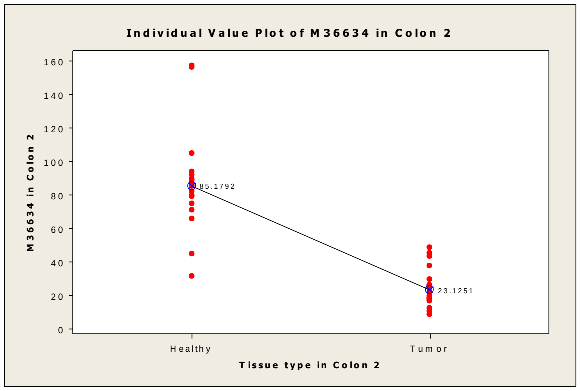
Gene CSRP1 (accession M76378) consistently underexpressed in Colon 1 and Colon 2 databases



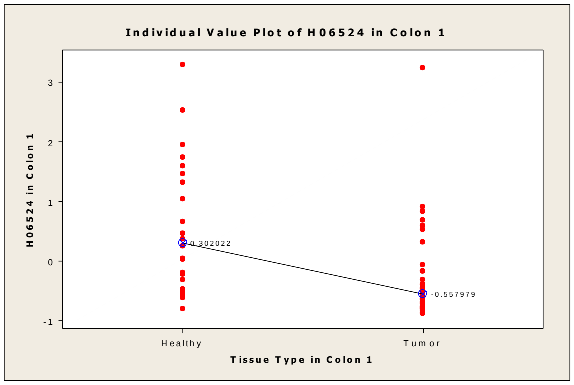


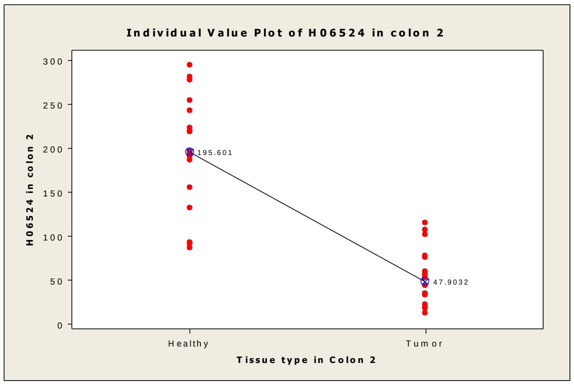
Gene VIP (accession M36634) consistently underexpressed in Colon 1 and Colon 2 databases



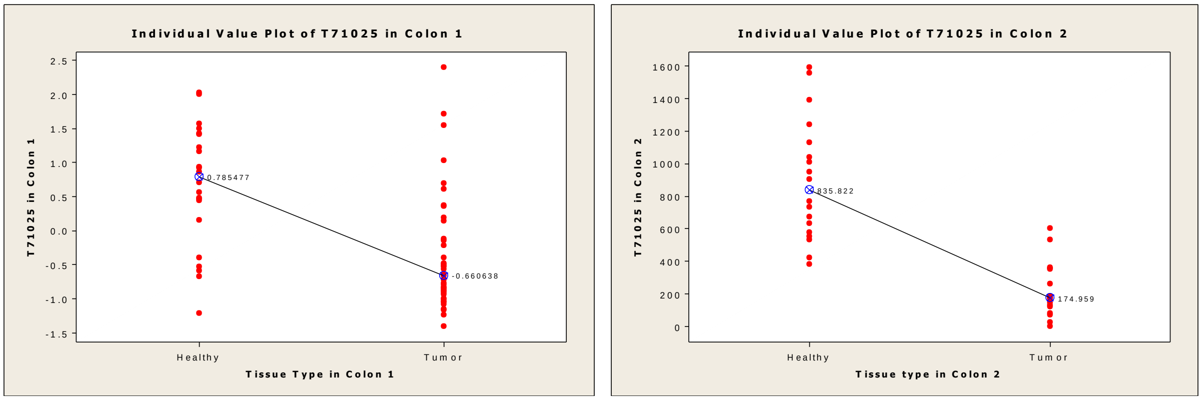


Gene GSN (Gelsolin), accession H06524, consistently underexpressed in Colon 1 and Colon 2 databases

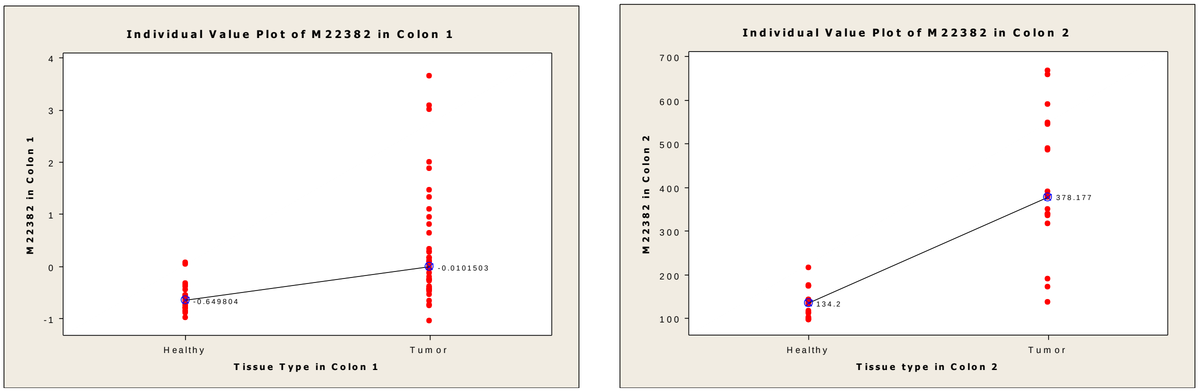




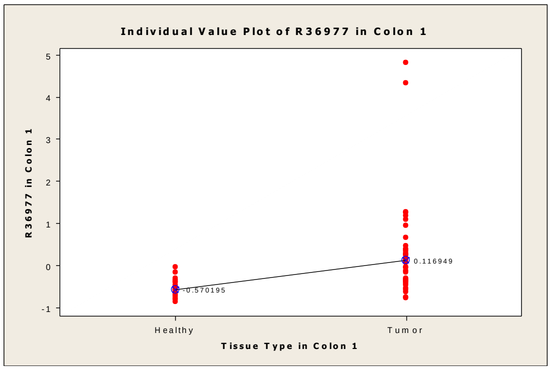
Gene MT1G (accession T71025) consistently underexpressed in Colon 1 and Colon 2 databases

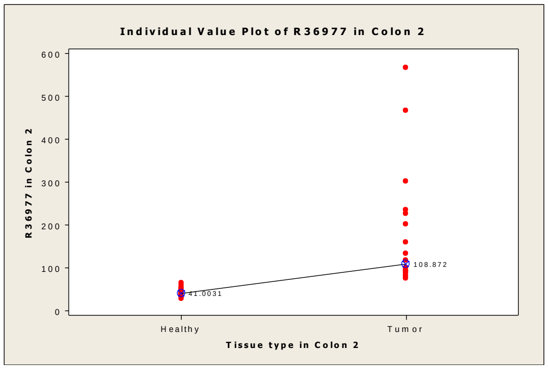


Gene HSPD1 (accession M22382) consistently overexpressed in Colon 1 and Colon 2 databases

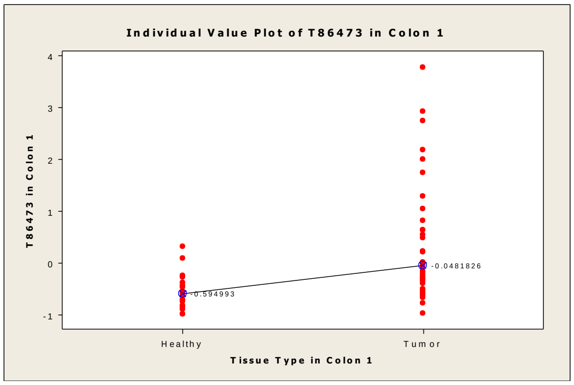


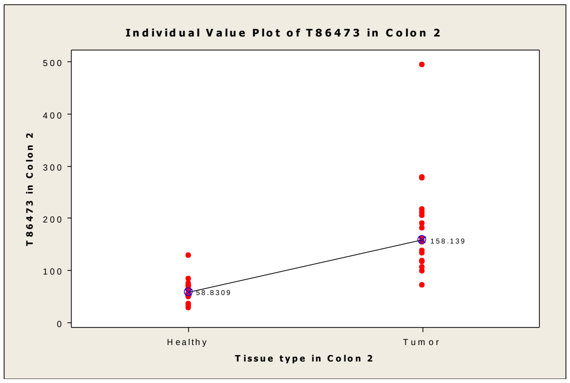
Gene GTF3A (accession R36977) consistently overexpressed in Colon 1 and Colon 2 databases



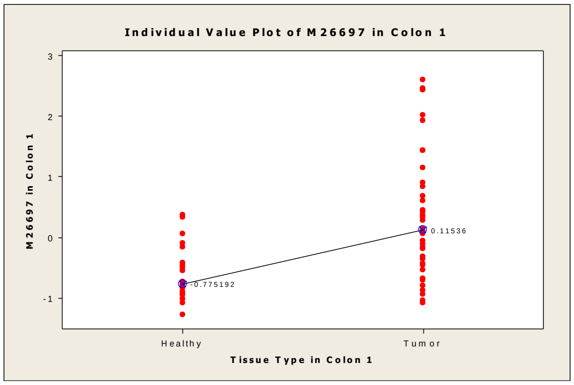


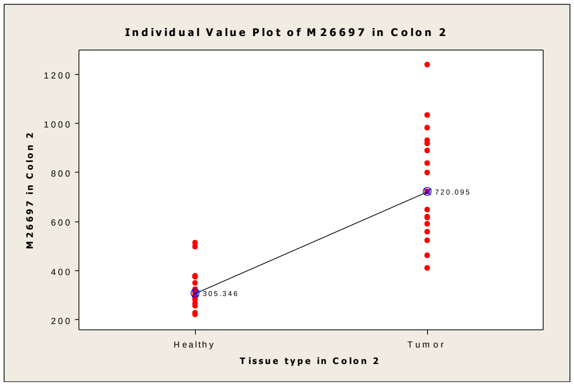
Gene NME1 (accession T86473) consistently overexpressed in Colon 1 and Colon 2 databases





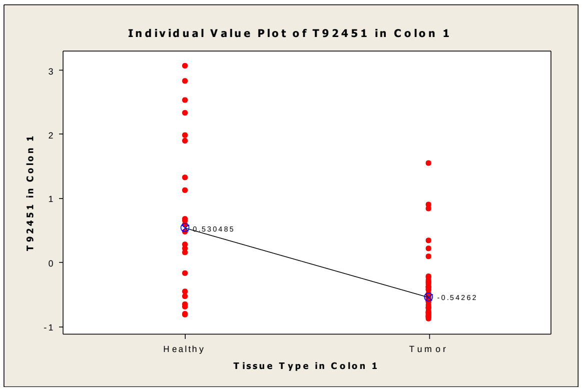
Gene NPM1(accession M26697) consistently overexpressed in Colon 1 and Colon 2 databases

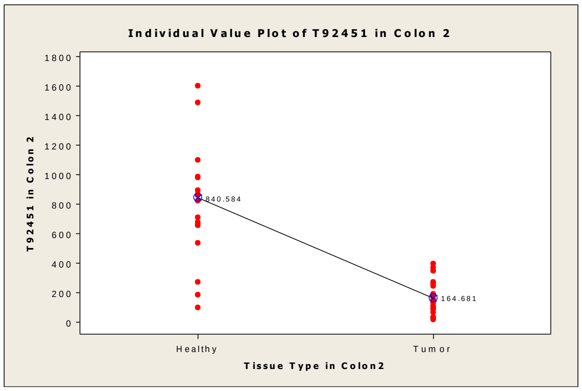




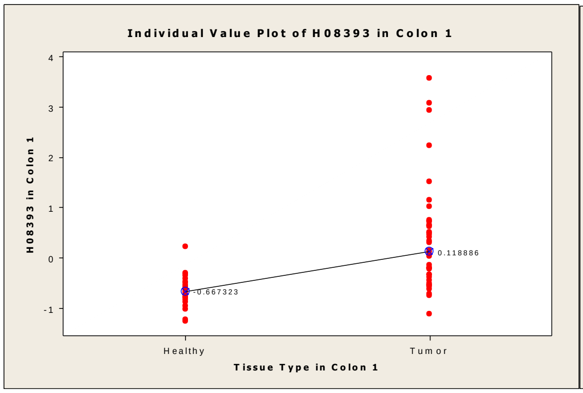
Graphs supporting Expression Profile for Group 2

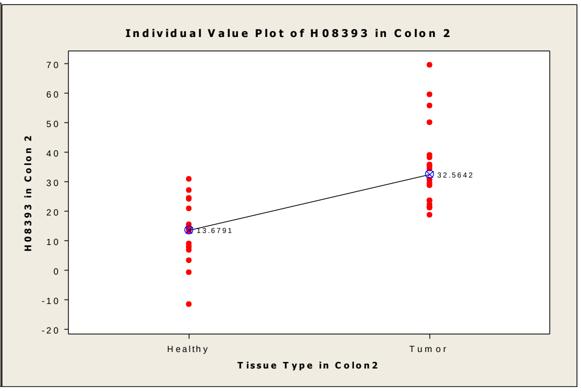
Gene TPM2 (accession T92451) consistently underexpressed in Colon 1 and Colon 2 databases



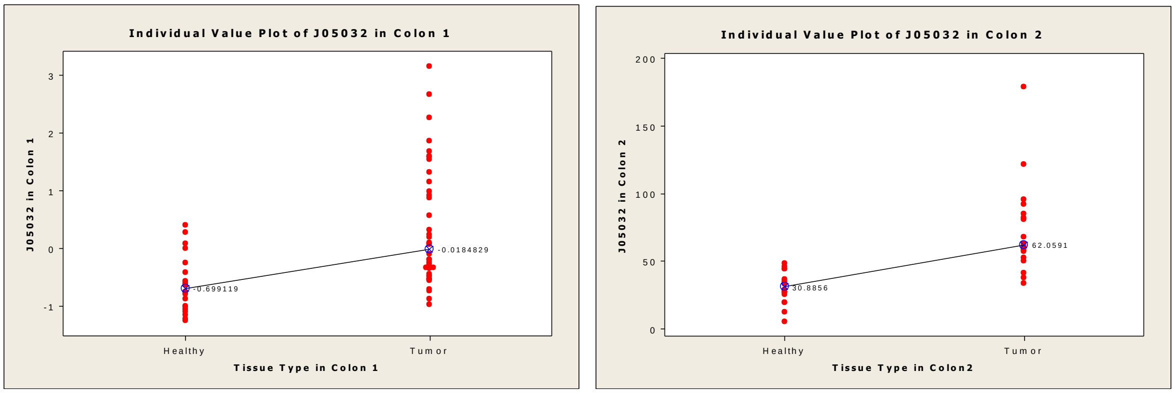


Gene WDR77 (accession H08393) consistently overexpressed in Colon 1 and Colon 2 databases

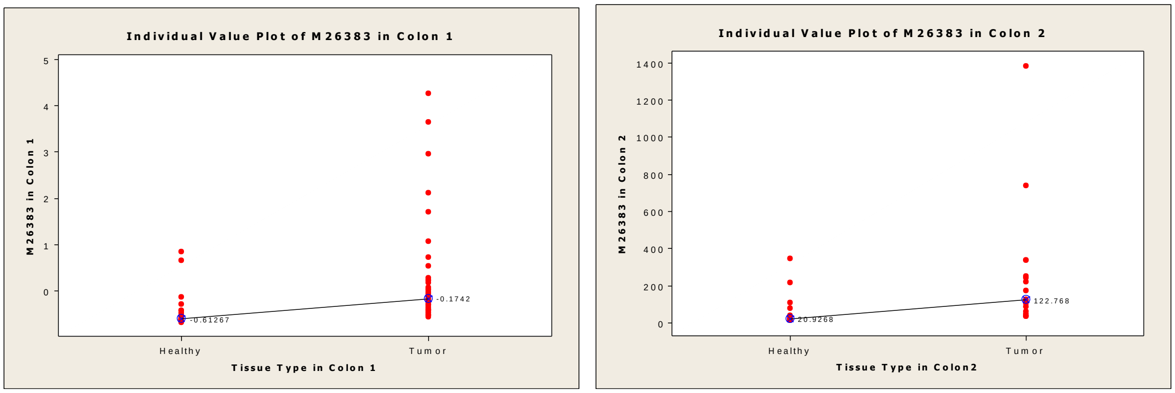




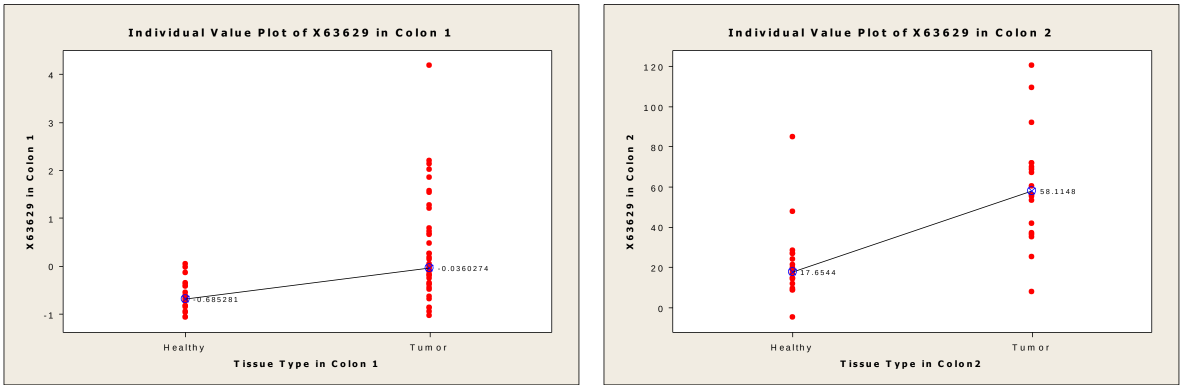
Gene DARS (accession J05032) consistently overexpressed in Colon 1 and Colon 2 databases



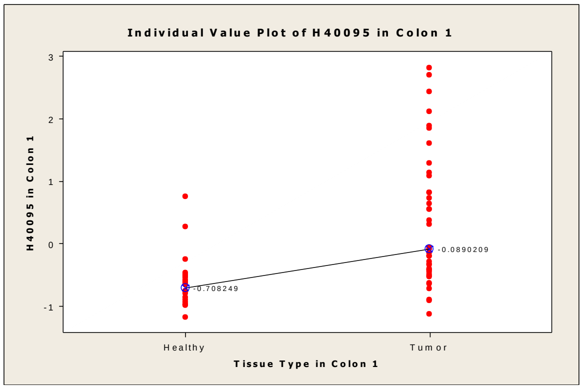
Gene IL8 (accession M26383) consistently overexpressed in Colon 1 and Colon 2 databases

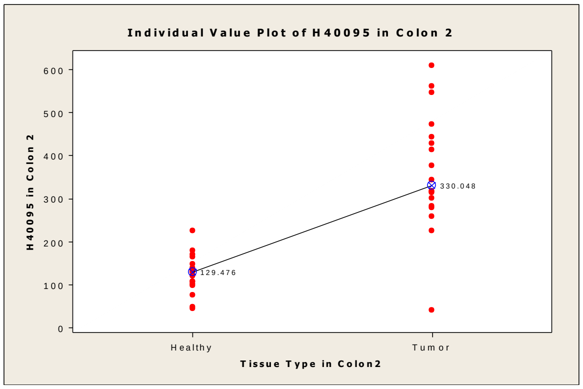


Gene CDH3 (accession X63629) consistently overexpressed in Colon 1 and Colon 2 databases

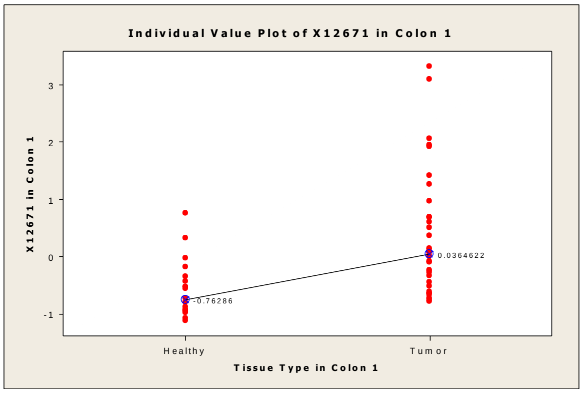


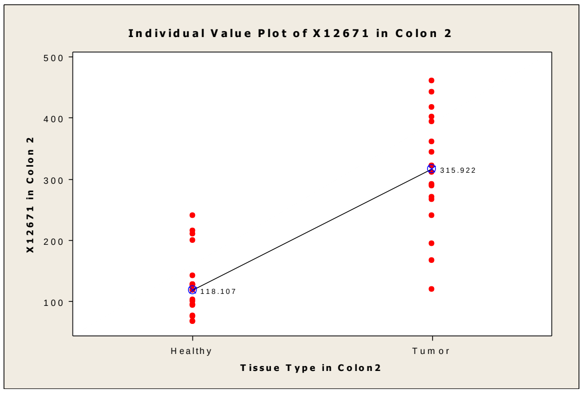
Gene EST: yn85b03.s1 (accession H40095) consistently overexpressed in Colon 1 and Colon 2 databases



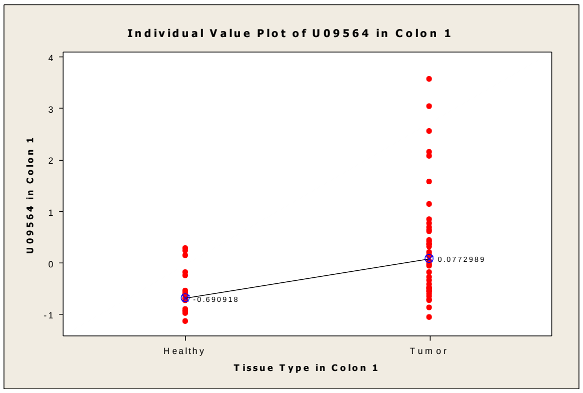


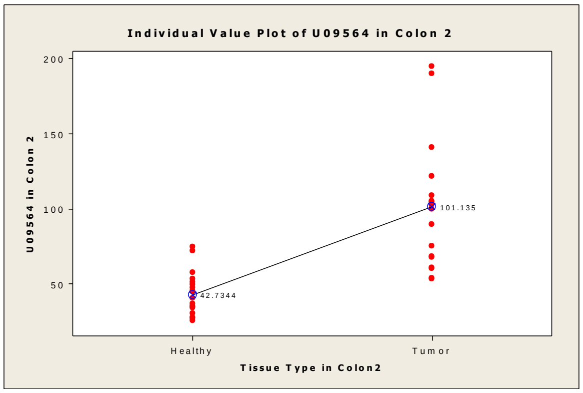
Gene HNRNPA1 (accession X12671) consistently overexpressed in Colon 1 and Colon 2 databases



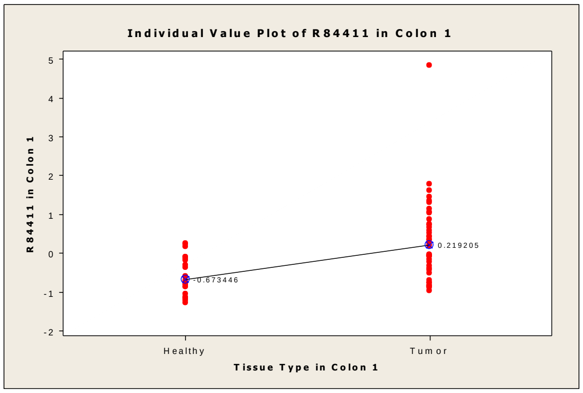


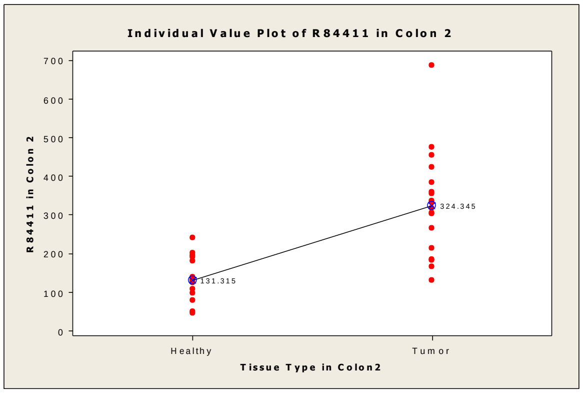
Gene SRPK1 (accession U09564) consistently overexpressed in Colon 1 and Colon 2 databases



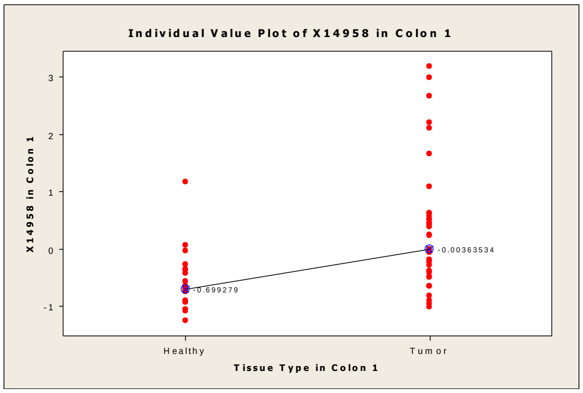


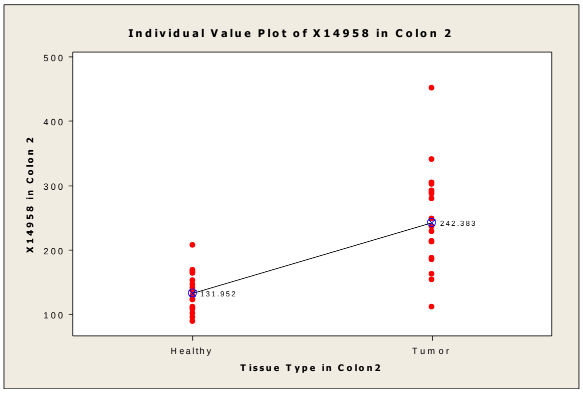
Gene SNRPB (accession R84411) consistently overexpressed in Colon 1 and Colon 2 databases



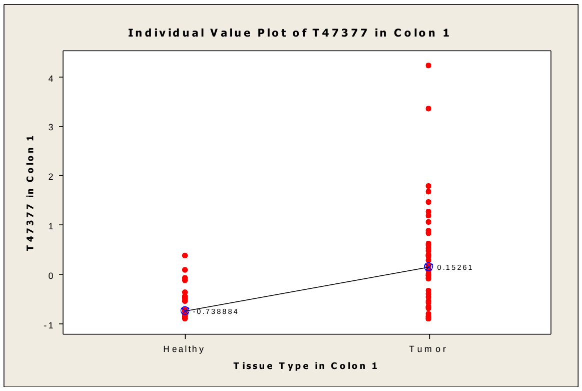


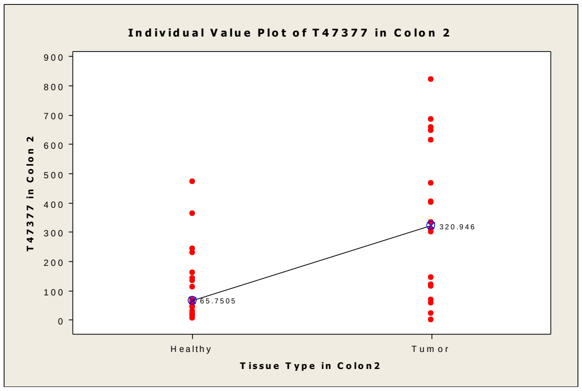
Gene HMGA1 (accession X14958) consistently overexpressed in Colon 1 and Colon 2 databases



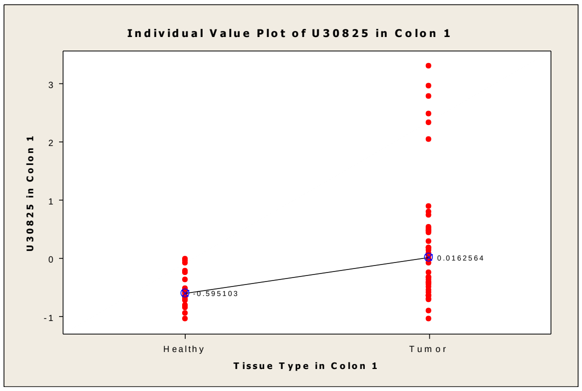


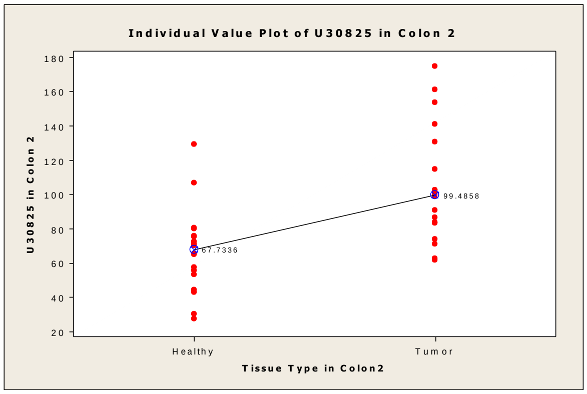
Gene S100P (accession T47377) consistently overexpressed in Colon 1 and Colon 2 databases



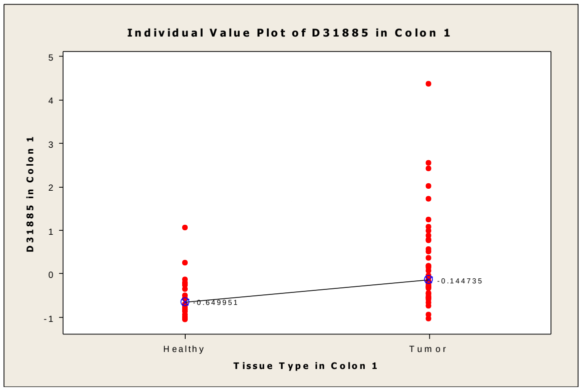


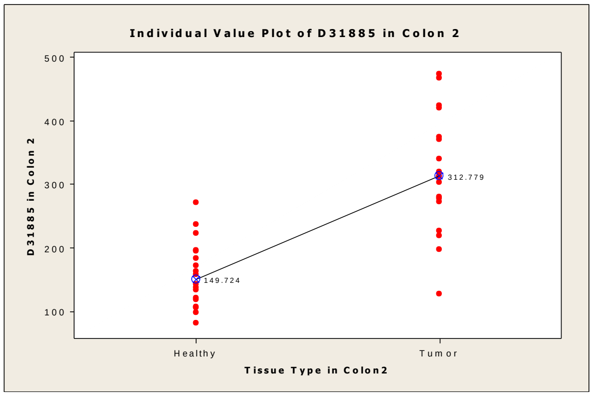
Gene SRSF9 (accession U30825) consistently overexpressed in Colon 1 and Colon 2 databases





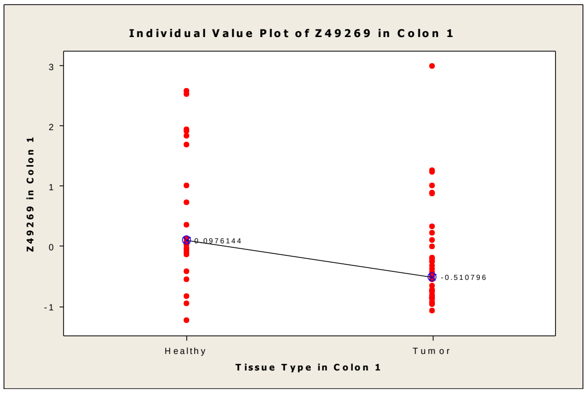
Gene ARL6IP1 (accession D31885) consistently overexpressed in Colon 1 and Colon 2 databases

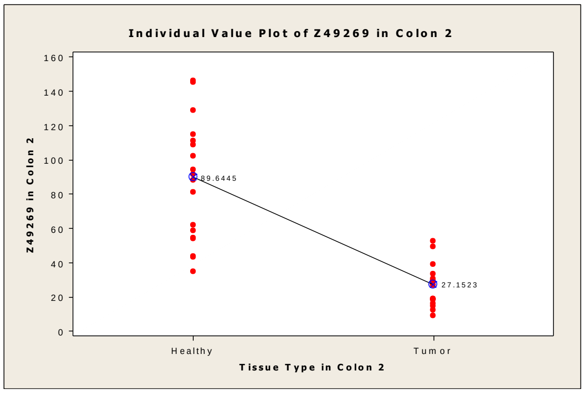




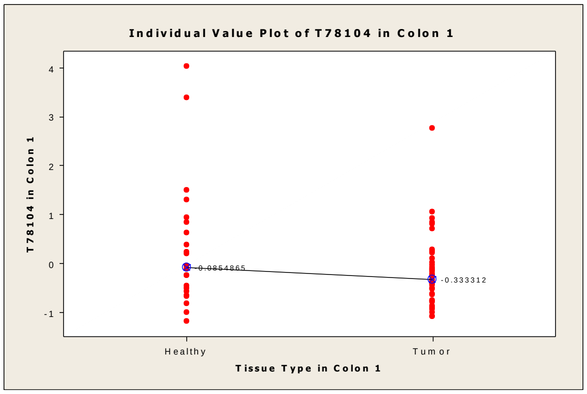
Graphs supporting Expression Profile for Group 3

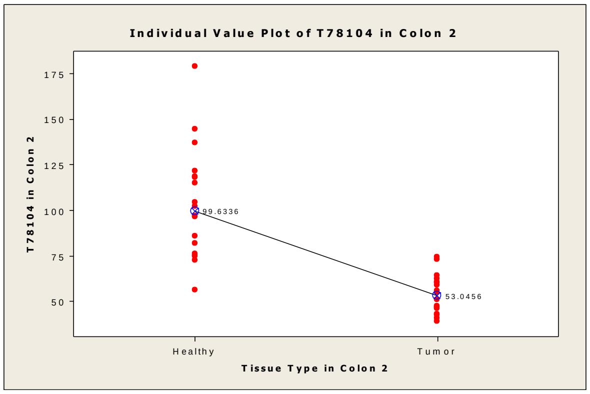
Gene CCL14 (accession Z49269), consistently underexpressed in Colon 1 and Colon 2 databases



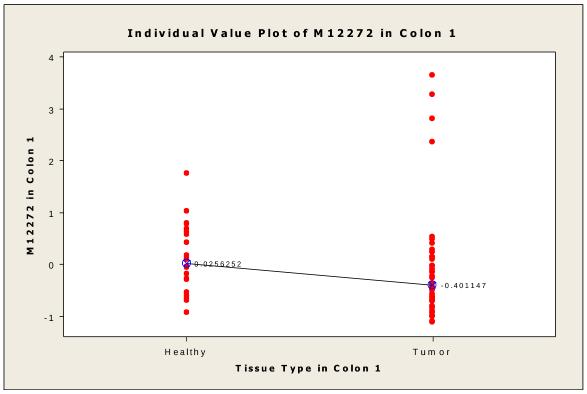


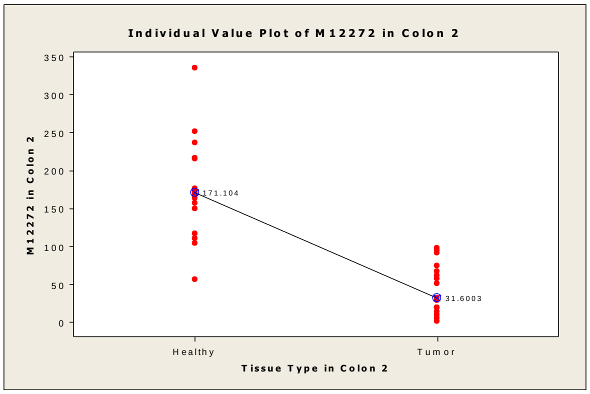
Gene PRELP (accession T78104), consistently underexpressed in Colon 1 and Colon 2 databases



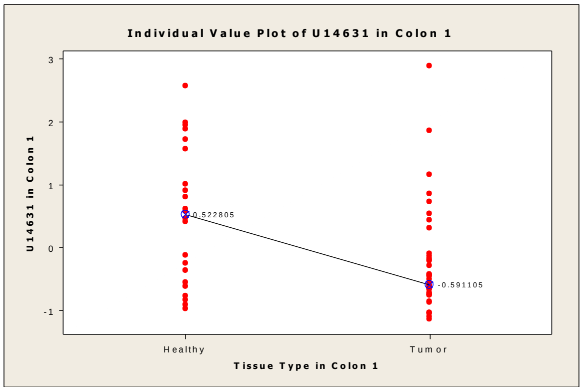


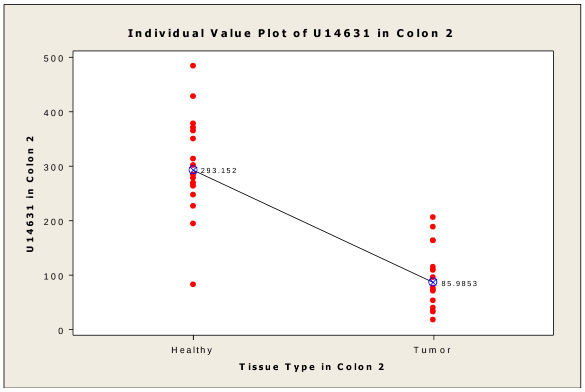
Gene ADH1C (accession M12272), consistently underexpressed in Colon 1 and Colon 2 databases



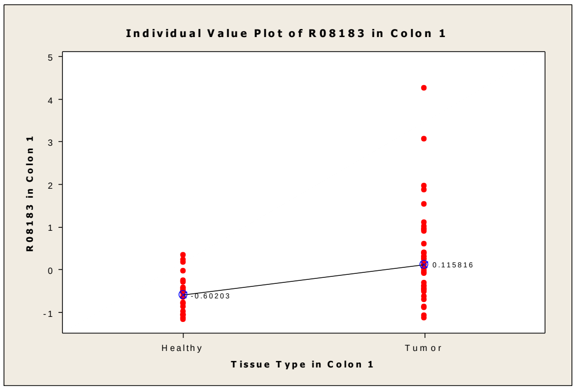


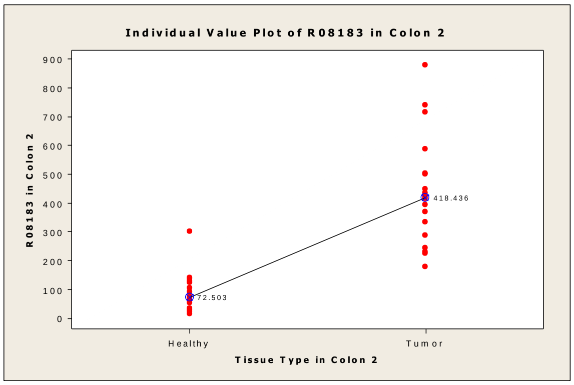
Gene HSD11B2 (accession U14631), consistently underexpressed in Colon 1 and Colon 2 databases



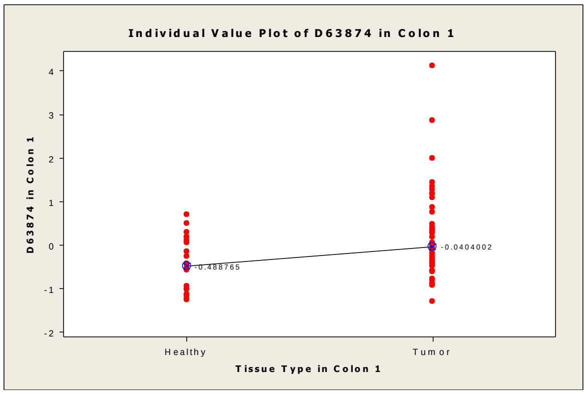


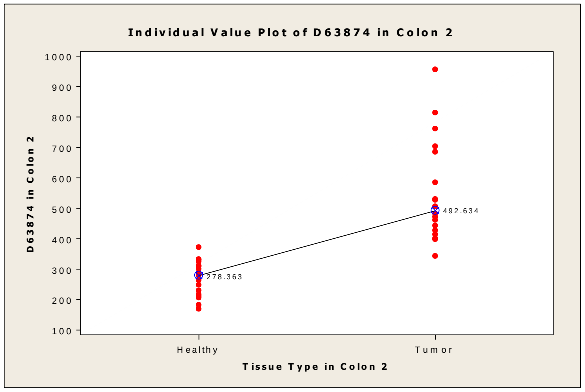
Gene HSPE1 (accession R08183), consistently overexpressed in Colon 1 and Colon 2 databases



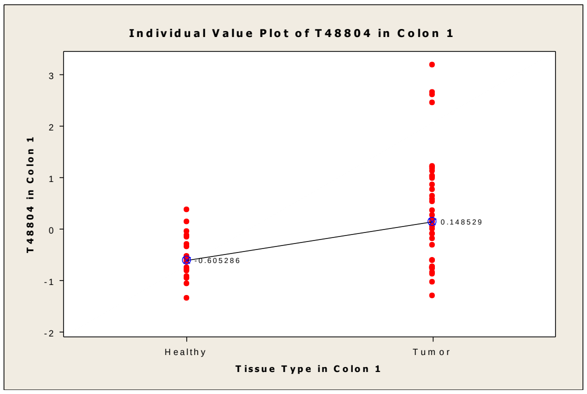


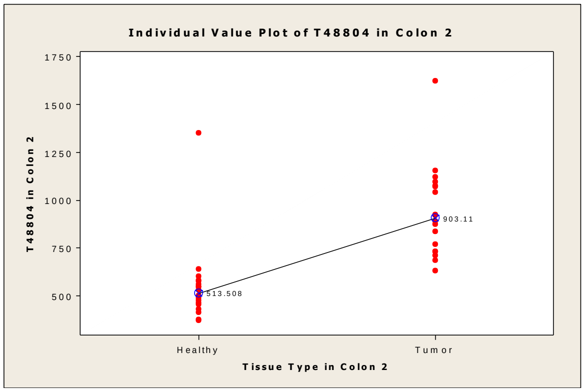
Gene HMGB1 (accession D63874), consistently overexpressed in Colon 1 and Colon 2 databases



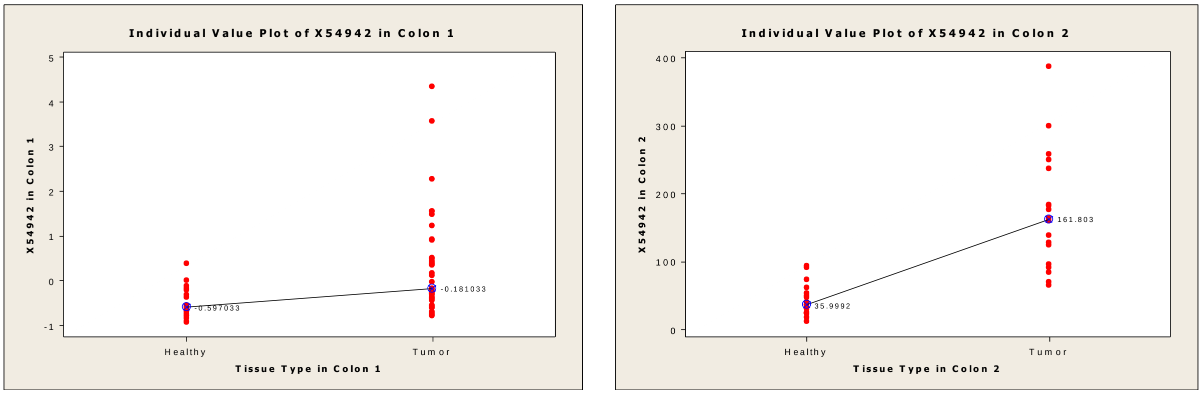


Gene RPS24 (accession T48804), consistently overexpressed in Colon 1 and Colon 2 databases

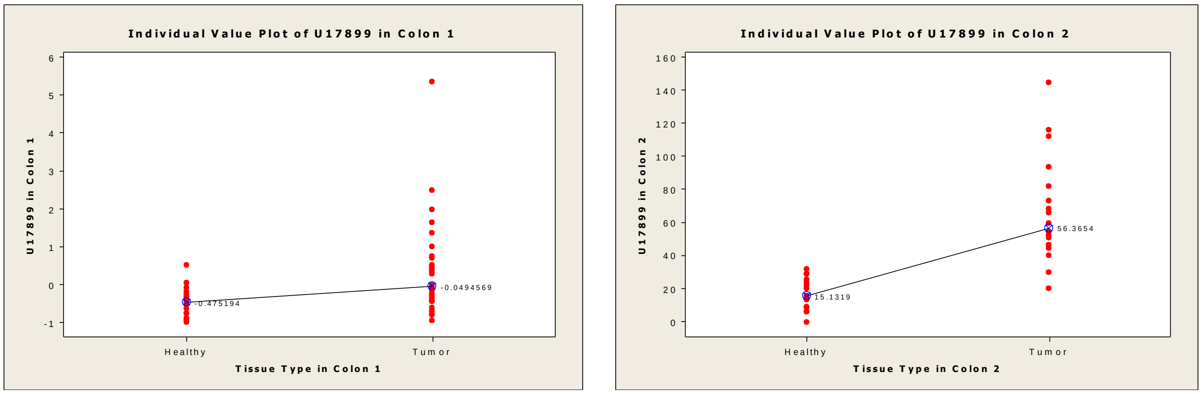




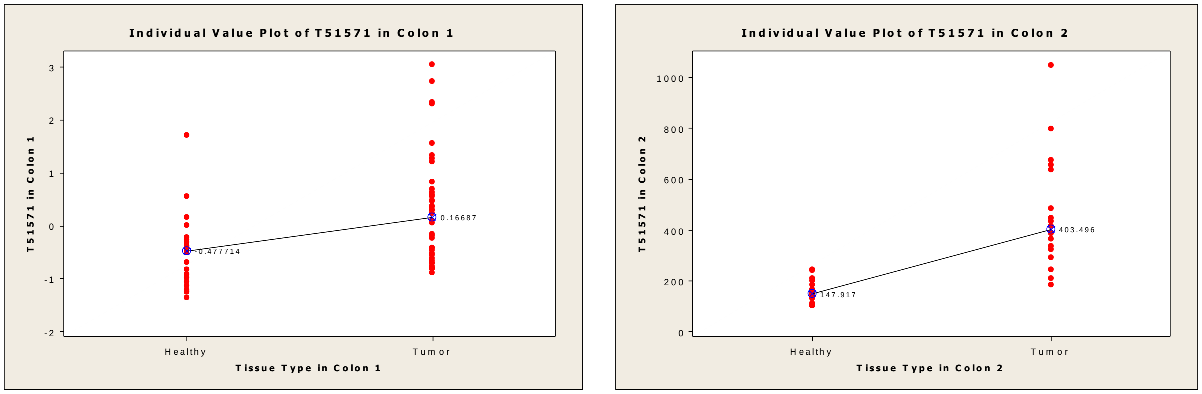
Gene CKS2 (accession X54942), consistently overexpressed in Colon 1 and Colon 2 databases



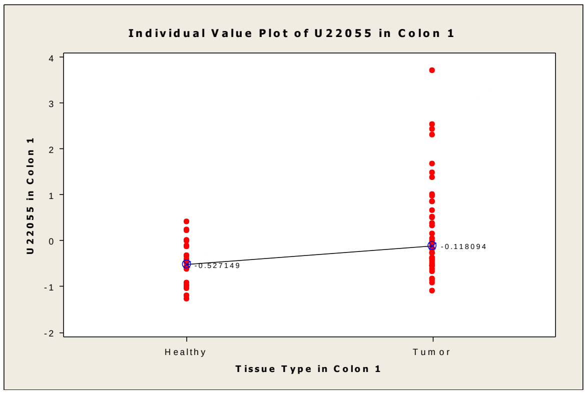
Gene CLNS1A (accession X54942), consistently overexpressed in Colon 1 and Colon 2 databases

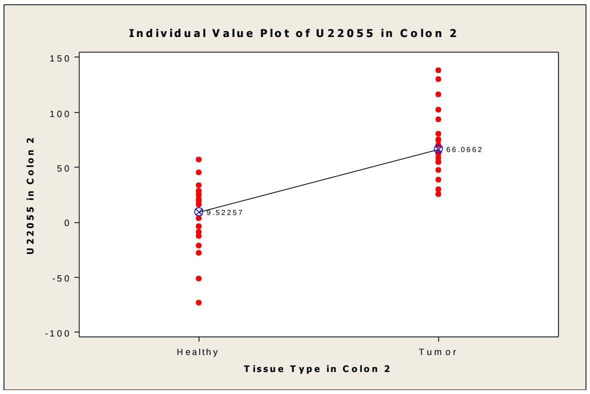


Gene S100A11 (accession T51571), consistently overexpressed in Colon 1 and Colon 2 databases



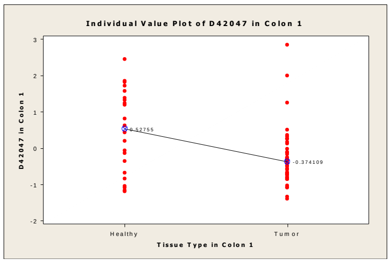
Gene SND1 (accession U22055), consistently overexpressed in Colon 1 and Colon 2 databases

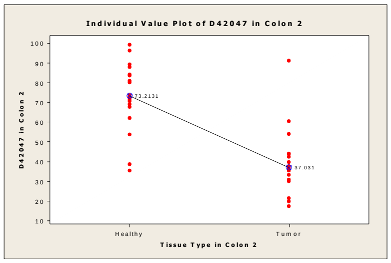


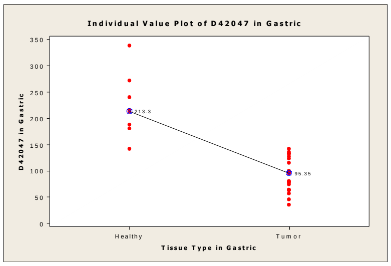


Graphs supporting Expression Profile for Group 4

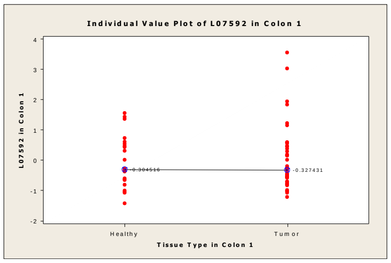
Gene GPD1L (accession D42047), consistently underexpressed in all the databases

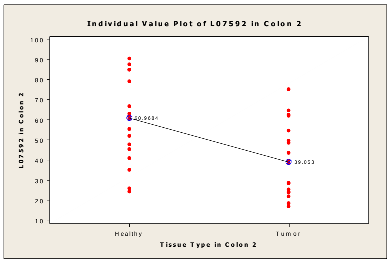


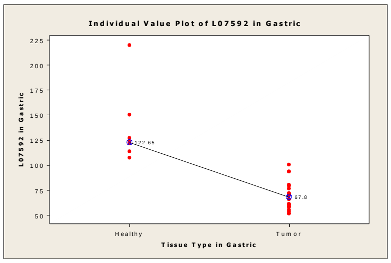




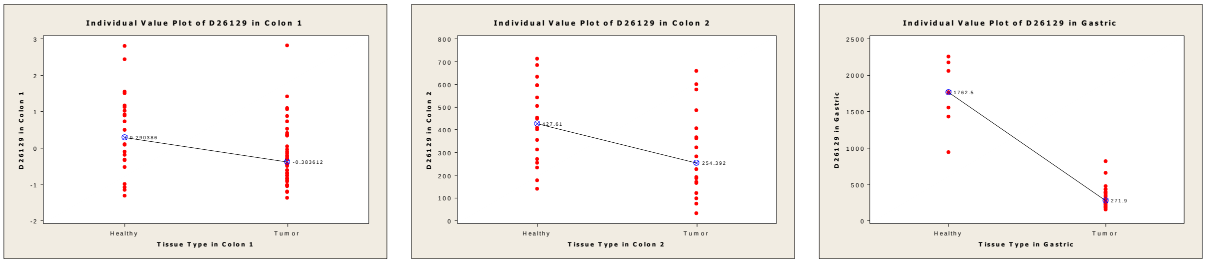
Gene PPARD (accession L07592), consistently underexpressed in all the databases



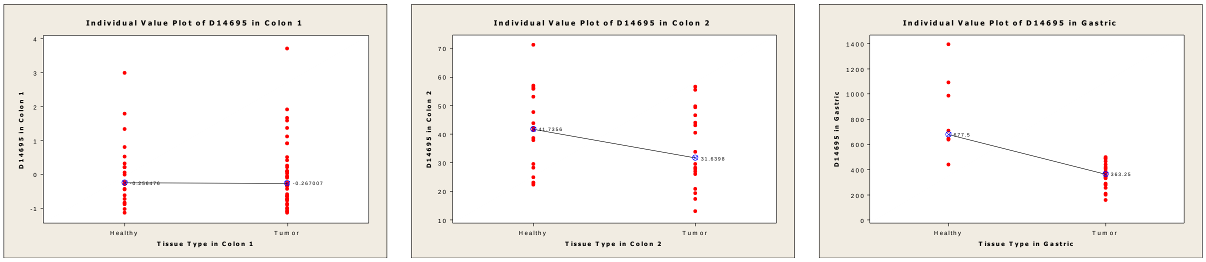




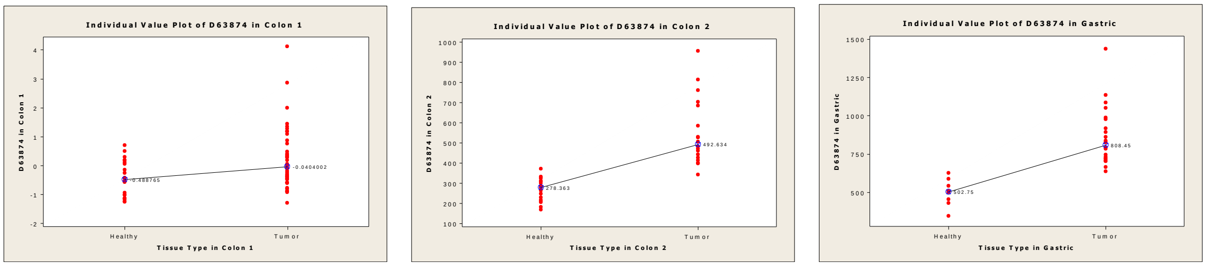
Gene RNASE1 (accession D26129), consistently underexpressed in all the databases



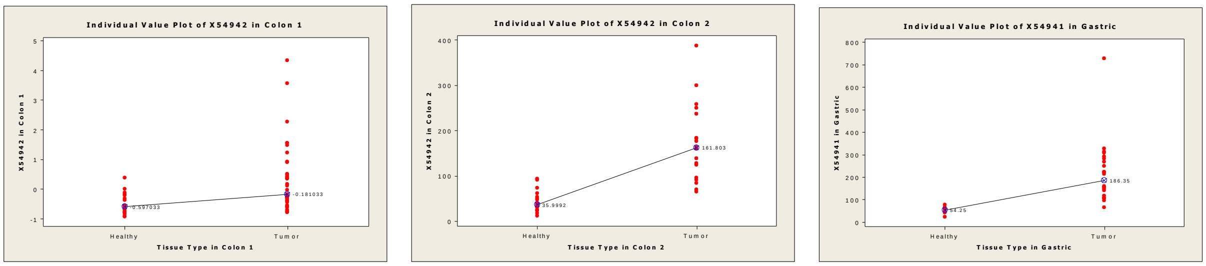
Gene HERPUD1 (accession D14695), consistently underexpressed in all the databases



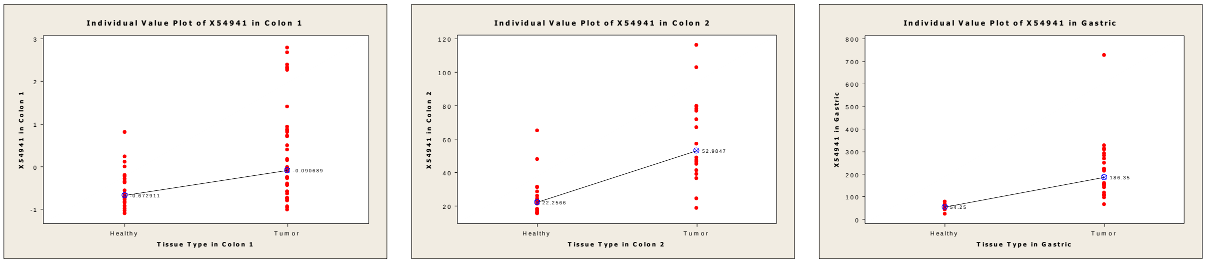
Gene HMGB1 (accession D63874), consistently overexpressed in all the databases



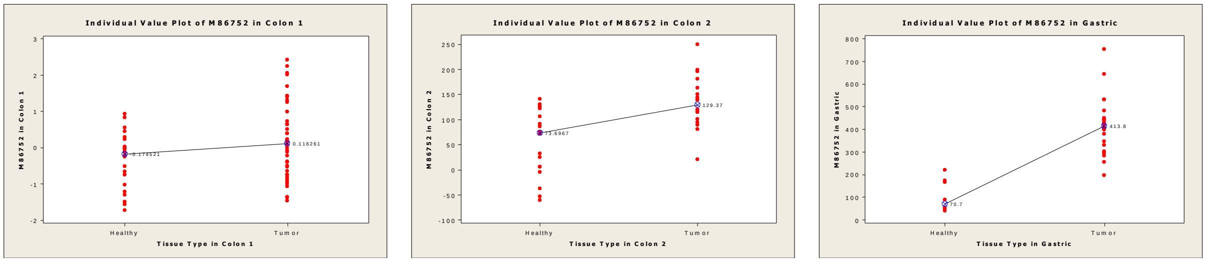
Gene CKS2 (accession X54942), consistently overexpressed in all the databases



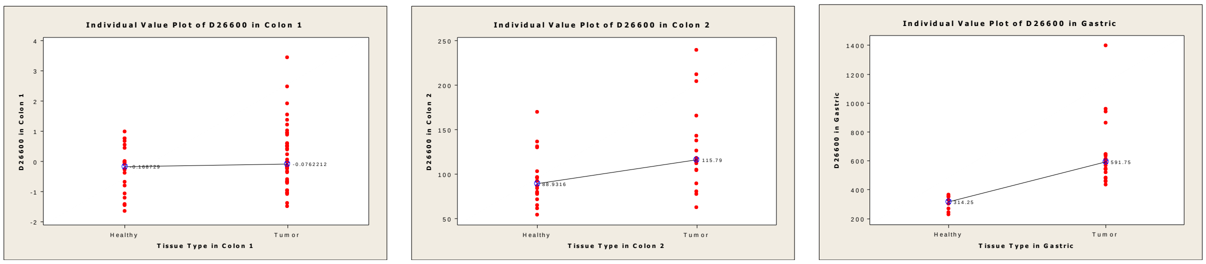
Gene CKS1B (accession X54941), consistently overexpressed in all the databases



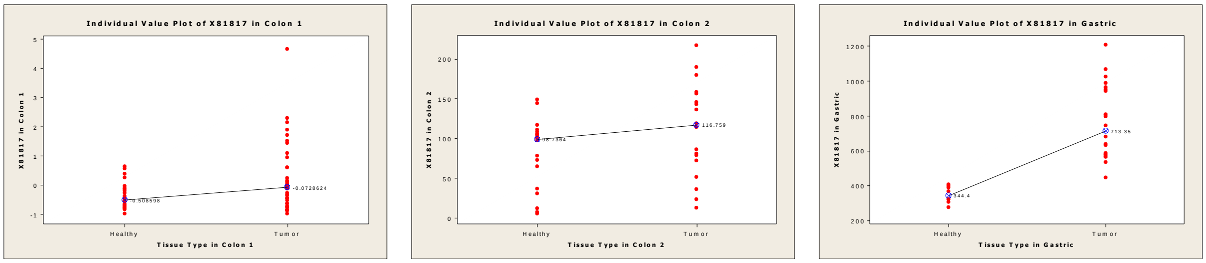
Gene STIP1 (accession M86752), consistently overexpressed in all the databases



Gene PSMB4 (accession D26600), consistently overexpressed in all the databases

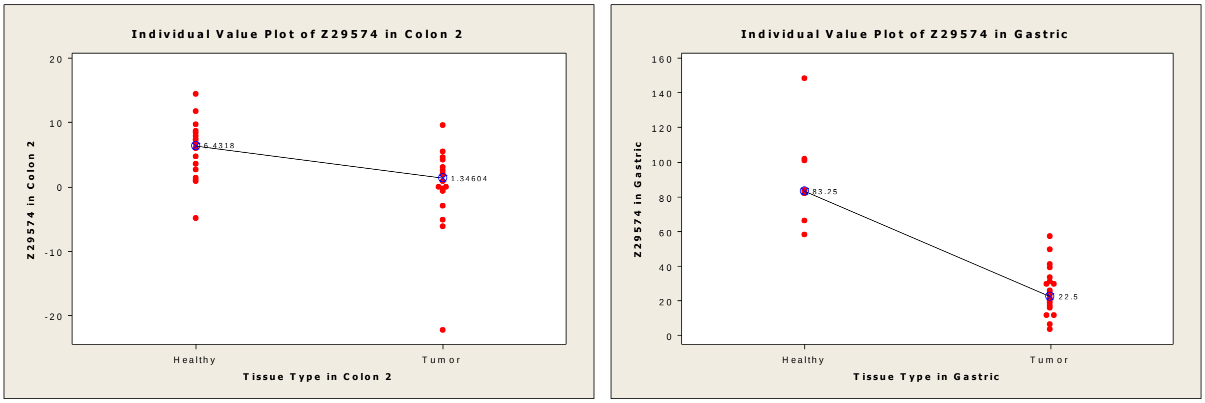


Gene BCAP31 (accession X81817), consistently overexpressed in all the databases

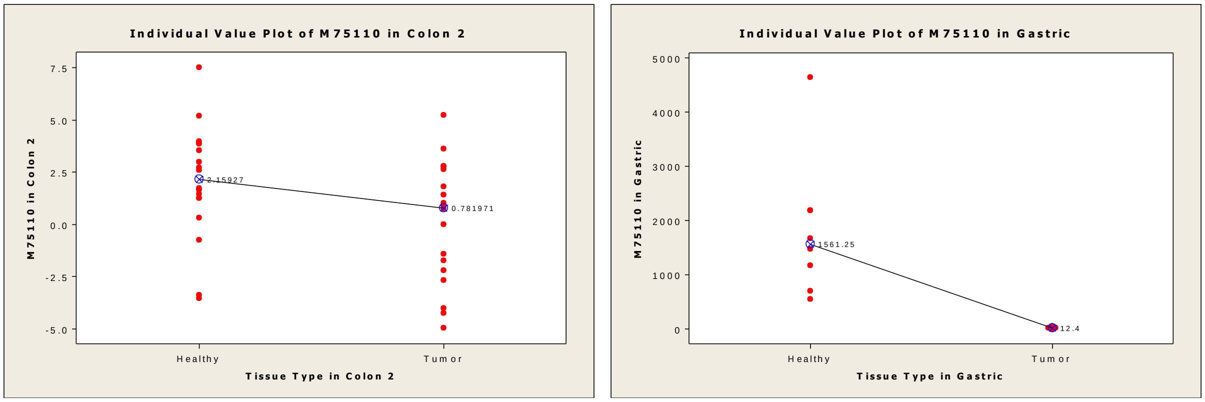


Graphs supporting Expression Profile for Group 5

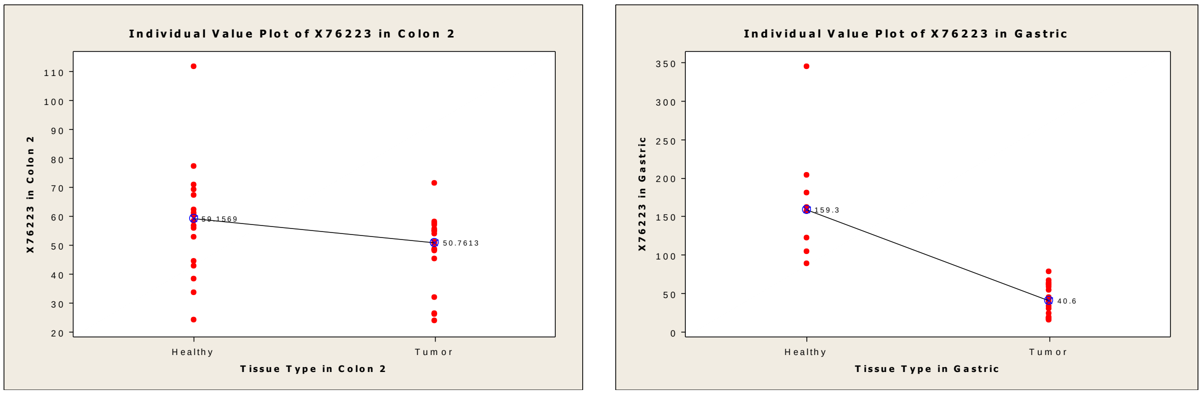
Gene TNFRSF17 (accession Z29574), consistently overexpressed in Colon 2 and Gastric databases



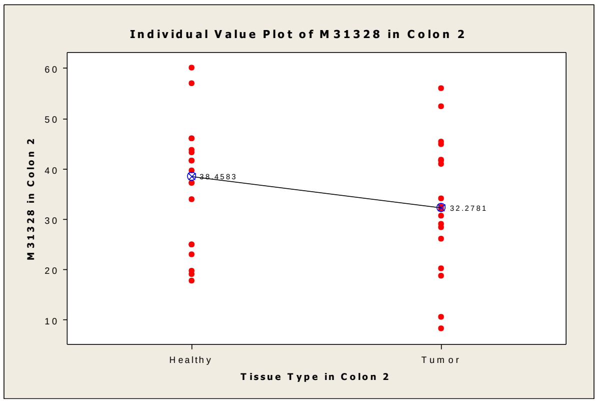
Gene ATP4B (accession M75110), consistently overexpressed in Colon 2 and Gastric databases

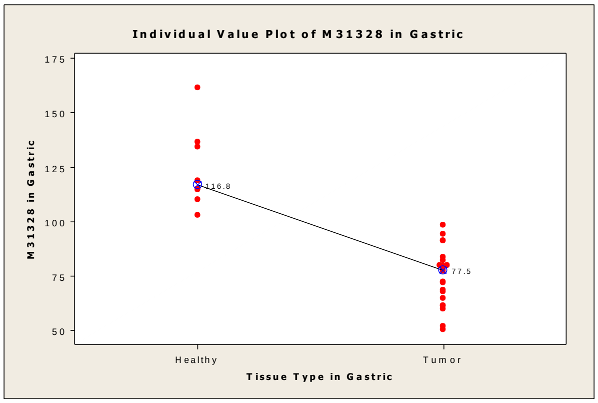


Gene with accession X76223 (H.sapiens MAL gene exon 4), consistently overexpressed in Colon 2 and Gastric databases

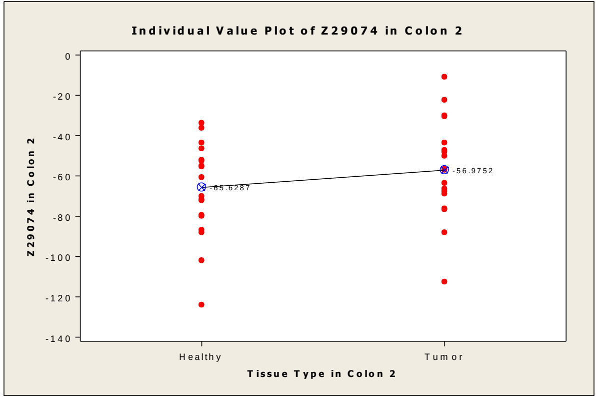


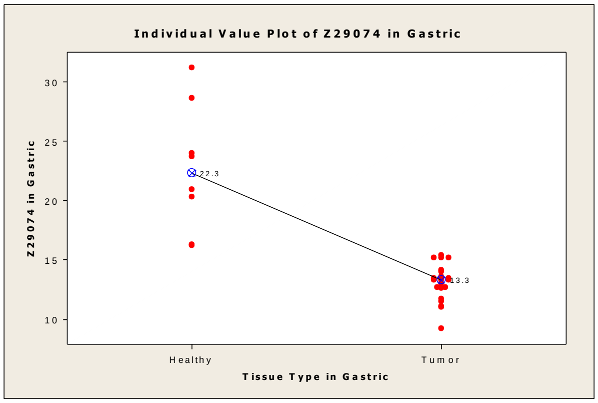
Gene GNB3(accession M31328), consistently overexpressed in Colon 2 and Gastric databases



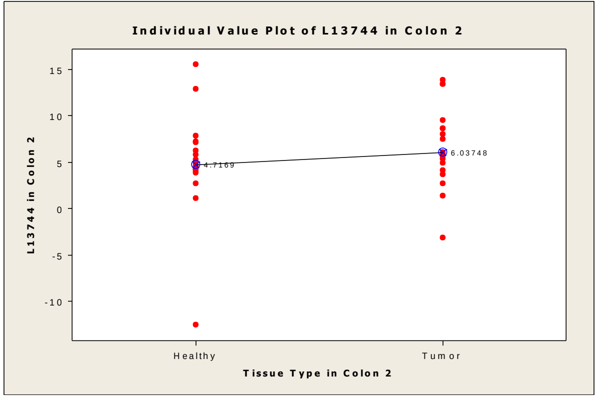


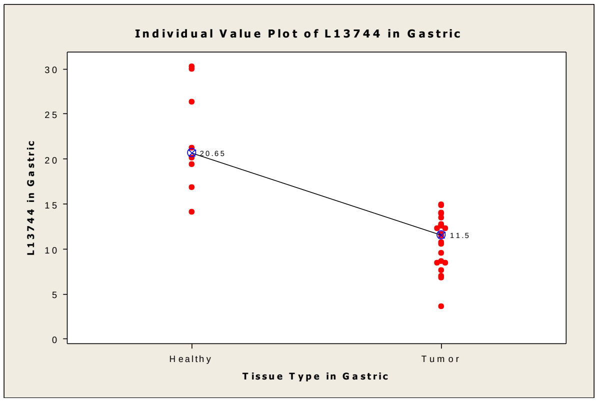
Gene KRT9 (accession Z29074), with conflicting expression, overexpressed in Colon 2 and underexpressed in Gastric



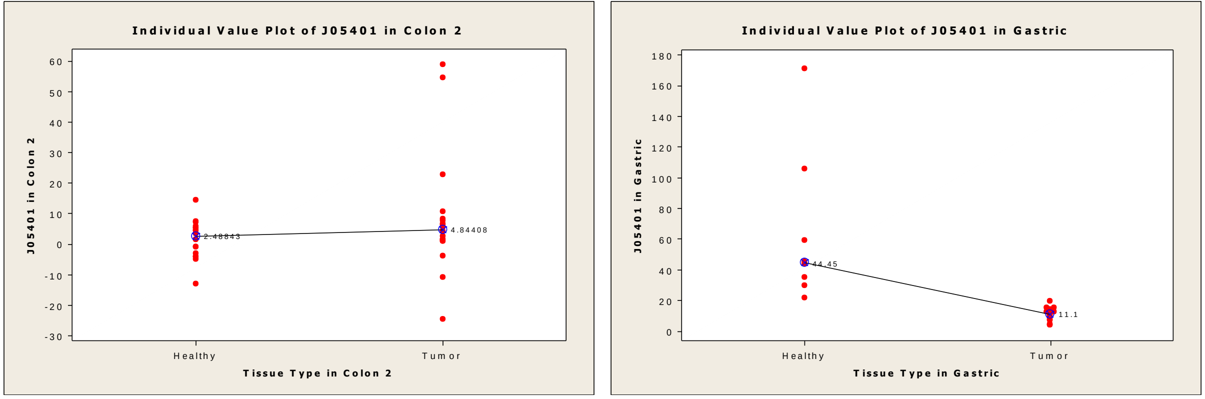


Gene CKMT2 (accession L13744), with conflicting expression, overexpressed in Colon 2 and underexpressed in Gastric.

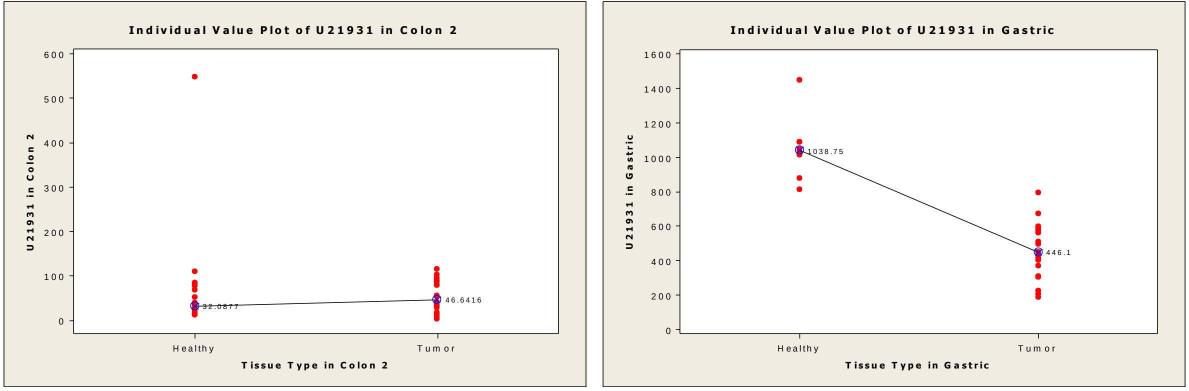




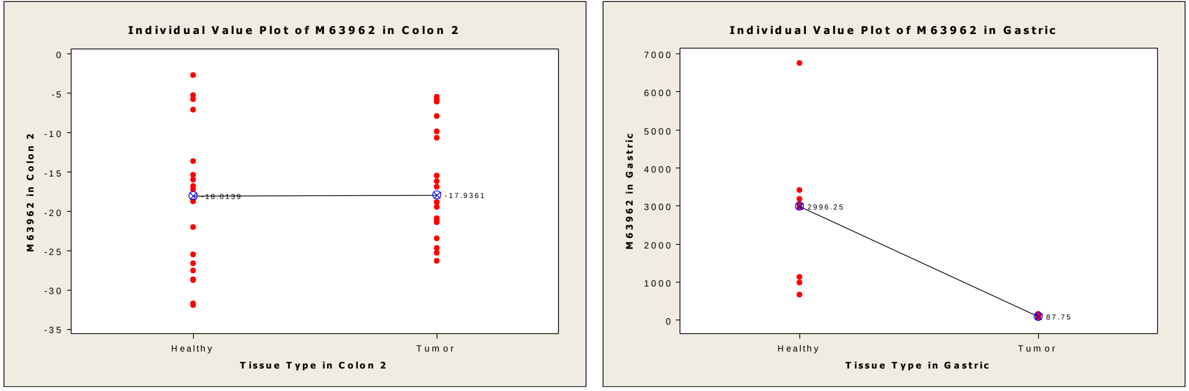
Gene ATP4A (accession J05401), with conflicting expression, overexpressed in Colon 2 and underexpressed in Gastric.



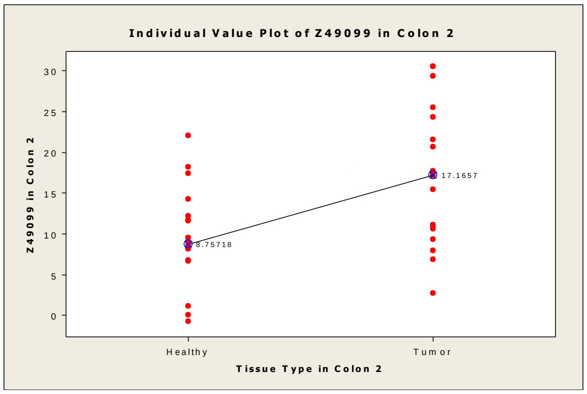
Gene MLLT3 (accession U21931), with conflicting expression, overexpressed in Colon 2 and underexpressed in Gastric

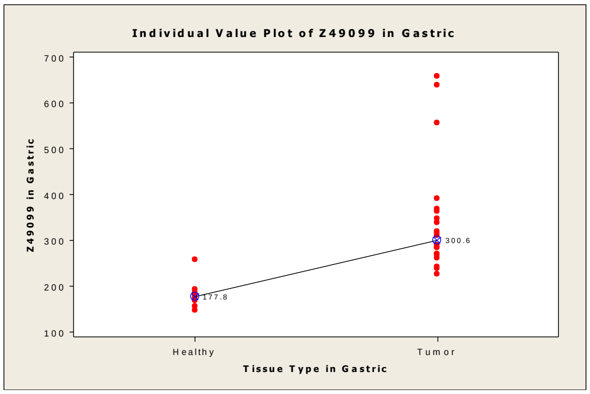


Gene FBP1 (accession M63962), with conflicting expression, overexpressed in Colon 2 and underexpressed in Gastric

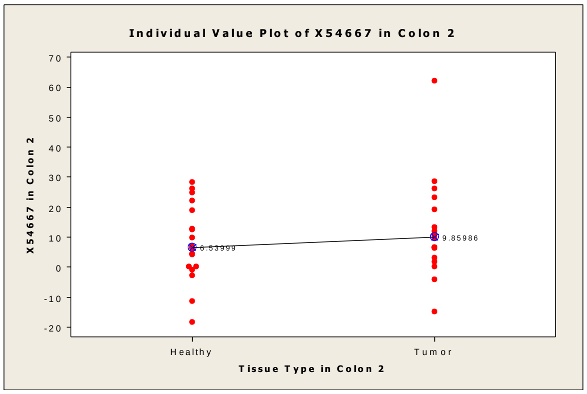


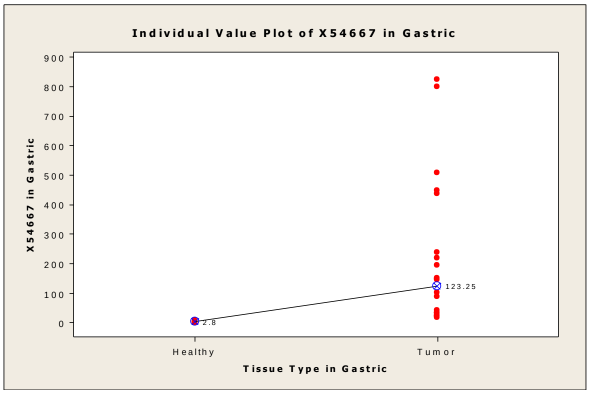
Gene PTPN12 (accession Z49099), consistently overexpressed in databases Colon 2 and Gastric



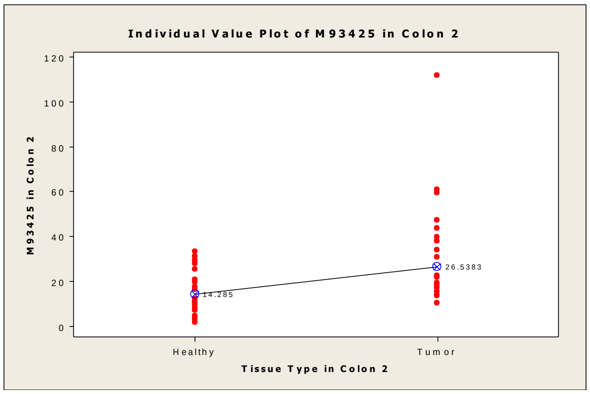


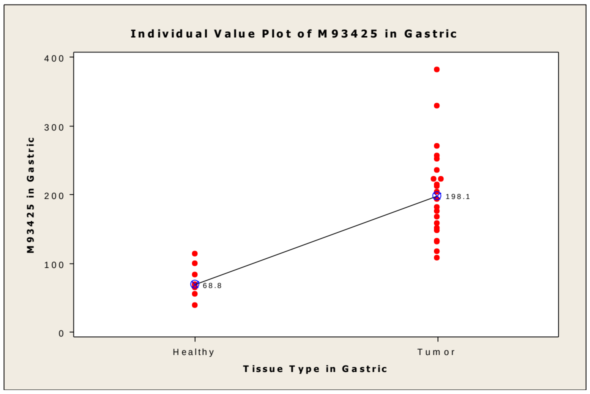
Gene SMS (accession X54667), consistently overexpressed in databases Colon 2 and Gastric



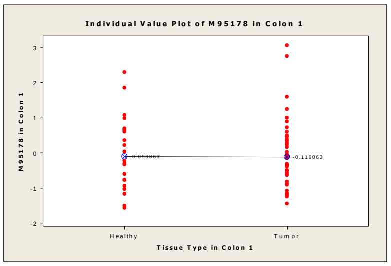


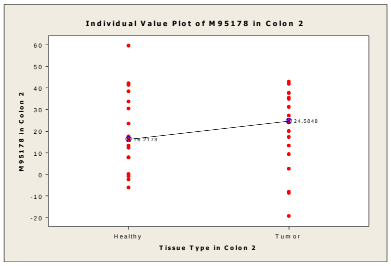
Gene CST4 (accession M93425), consistently overexpressed in databases Colon 2 and Gastric

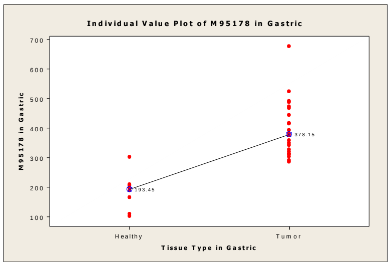




Graphs supporting one gene intersecting just List 3 and 5, ACTN1 (accession M95178), consistent across the three databases

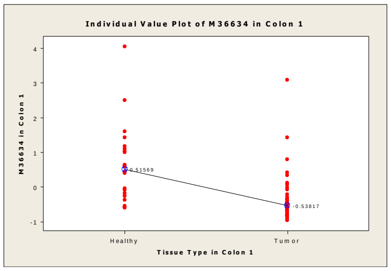


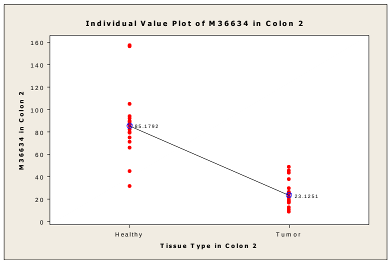


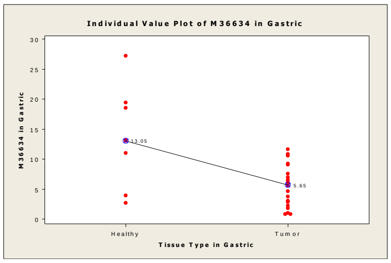


Graphs supporting Expression Profile for Group 6

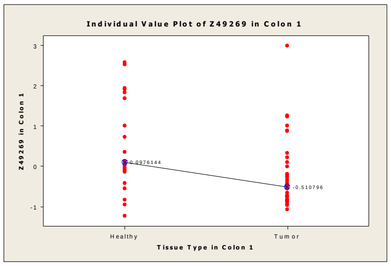
Gene VIP (accession M36634), consistently underexpressed in all the databases

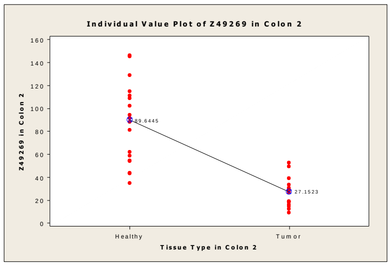


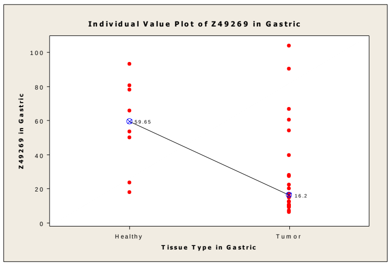




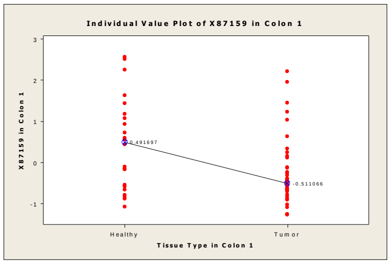
Gene CCL14 (accession Z49269), consistently underexpressed in all the databases

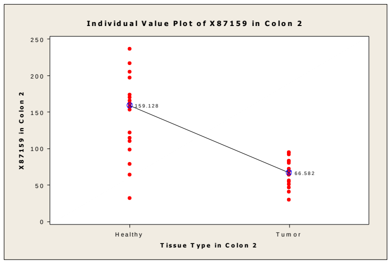


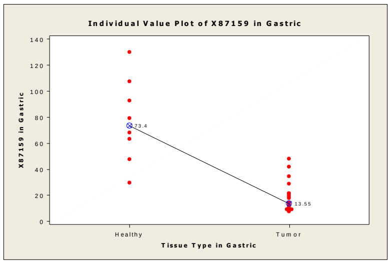




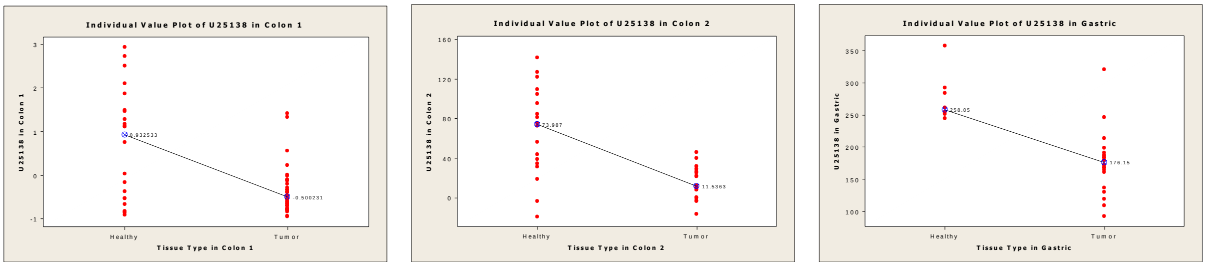
Gene SCNN1B (accession X87159), consistently underexpressed in all the databases



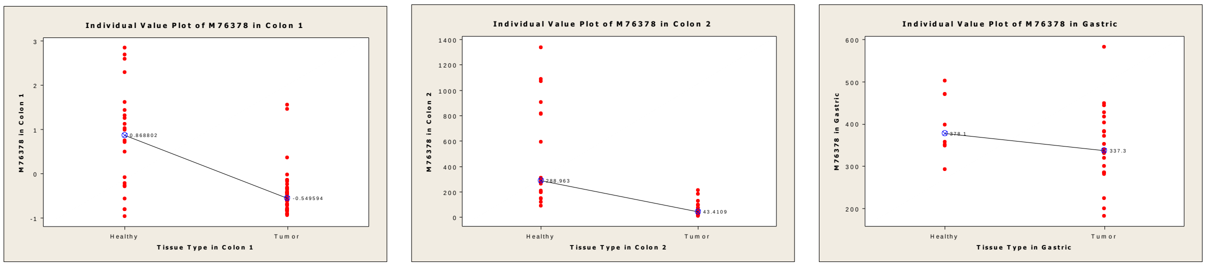




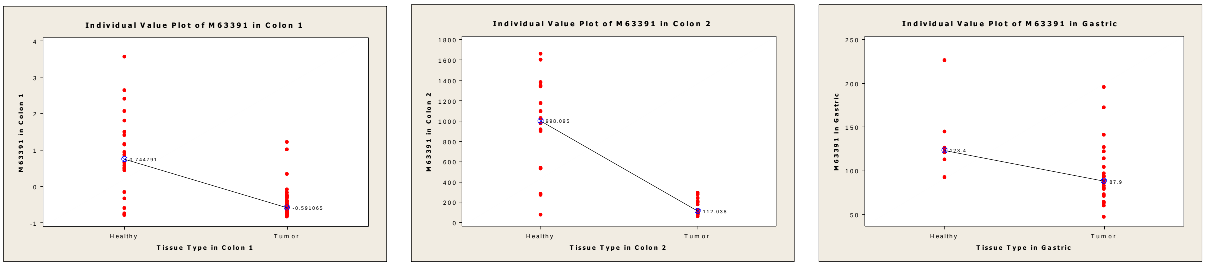
Gene KCNMB1(accession U25138), consistently underexpressed in all the databases



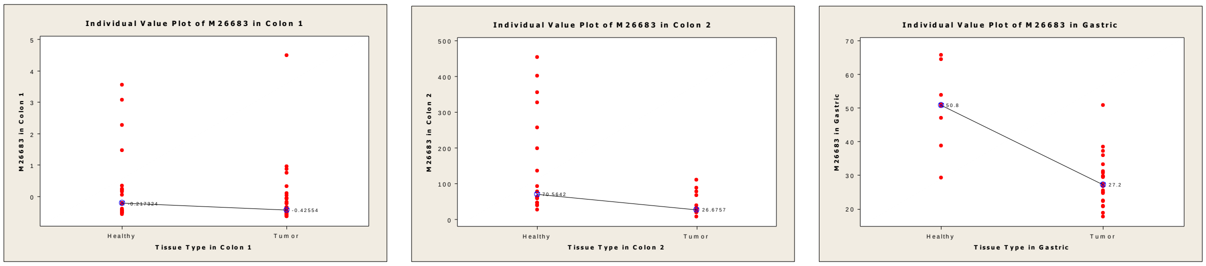
Gene CSRP1(accession M76378), consistently underexpressed in all the databases



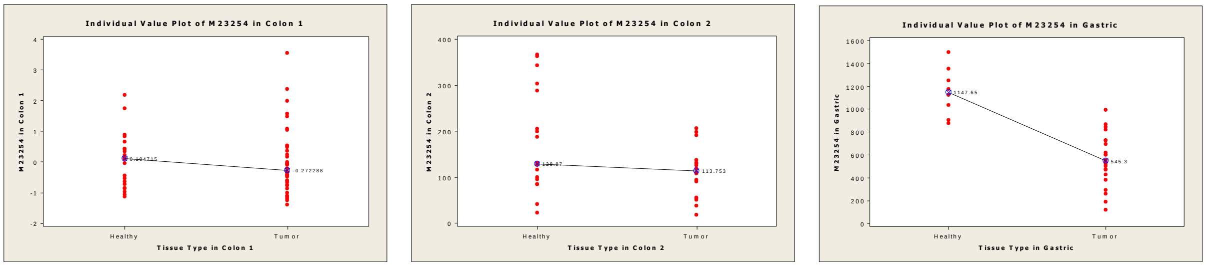
Gene DES (accession M63391), consistently underexpressed in all the databases



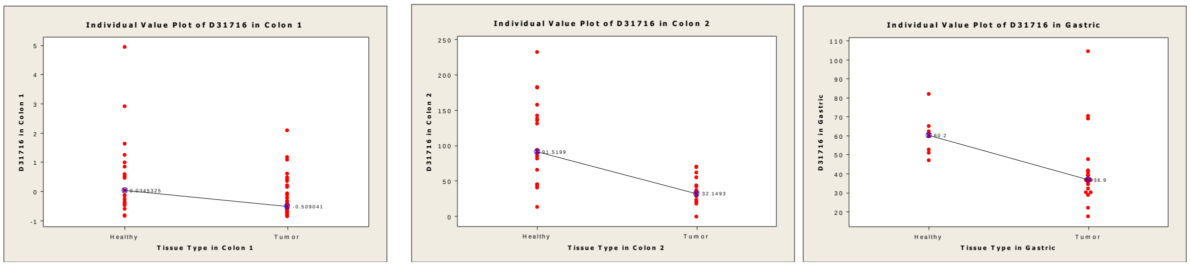
Gene HUMIFNIND (accession M26683), consistently underexpressed in all the databases



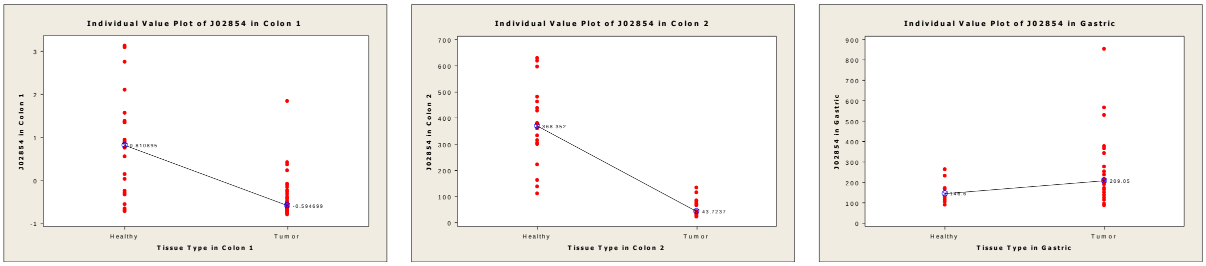
Gene CAPN2 (accession M23254), consistently underexpressed in all the databases



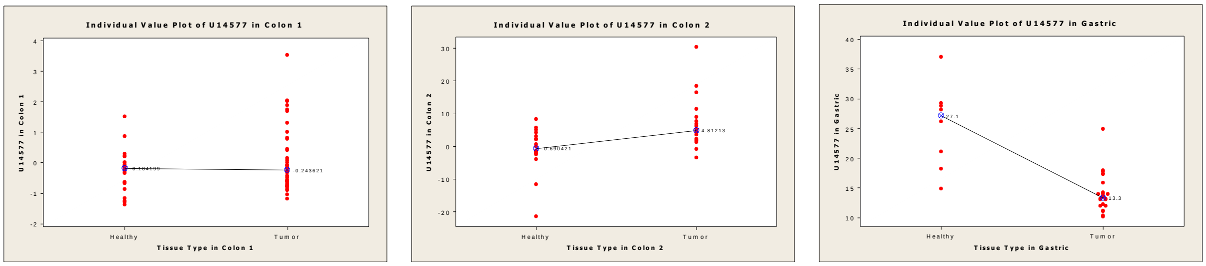
Gene KLF9 (accession M23254), consistently underexpressed in all the databases



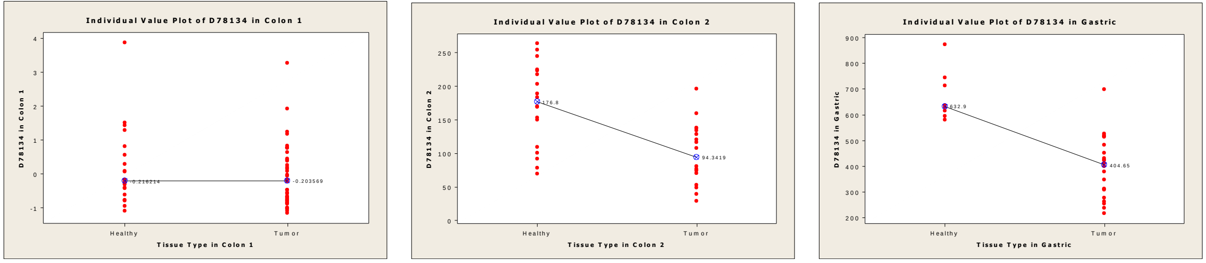
Gene MYL9 (accession J02854) with conflicting expression, underexpressed in Colon 1 and Colon 2 but overexpressed in Gastric



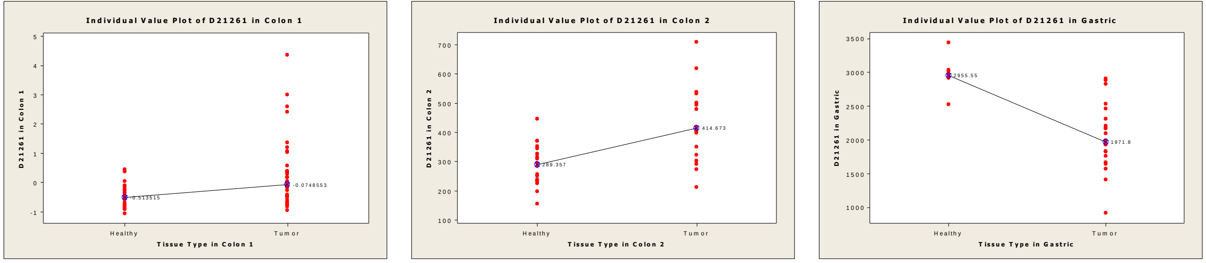
Gene MAP1A (accession U14577) with conflicting expression, underexpressed in Colon1 and Gastric and overexpressed in Colon 2



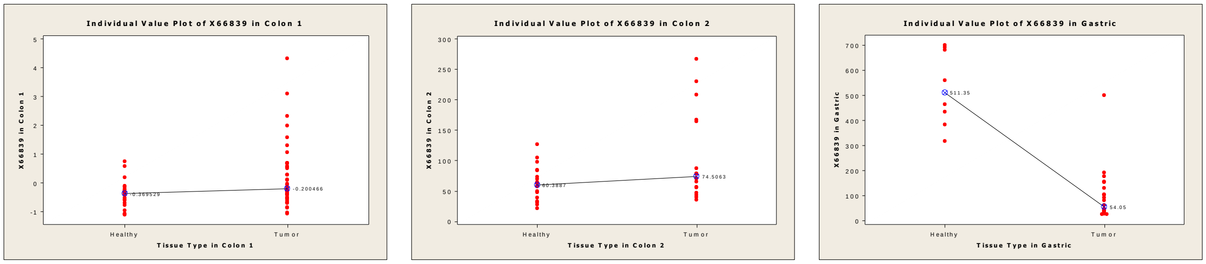
Gene CIRBP (accession D78134), conflicting behavior, overexpressed in Colon 1 and underexpresed in Colon 2 and Gastric.



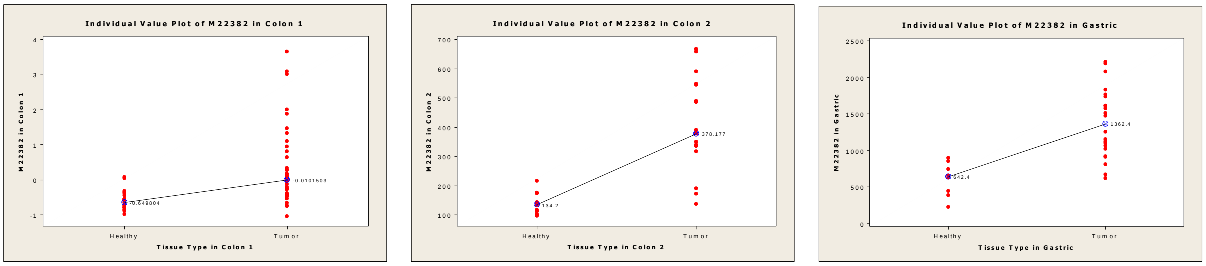
Gene TAGLN2 (accession D21261) with conflicting expression, overexpressed in Colon 1 and Colon 2 but underexpressed in Gastric



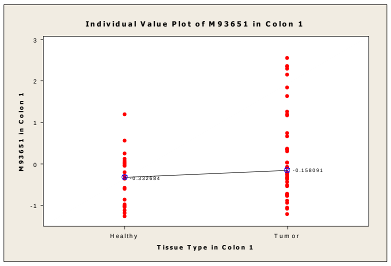
Gene CA9 (accession X66839) with conflicting expression, overexpressed in Colon 1 and Colon 2, underexpressed in Gastric

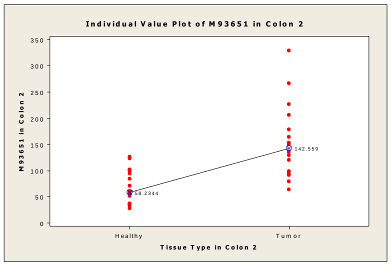


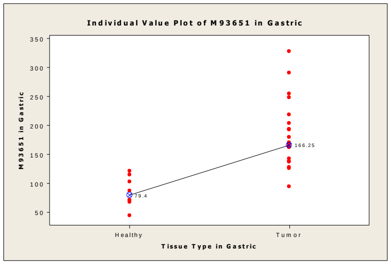
Gene HSPD1 (accession M22382), consistently overexpressed in all the databases



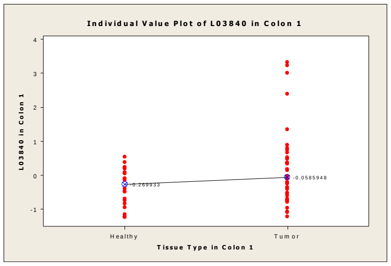
Gene SET (accession M93651), consistently overexpressed in all the databases

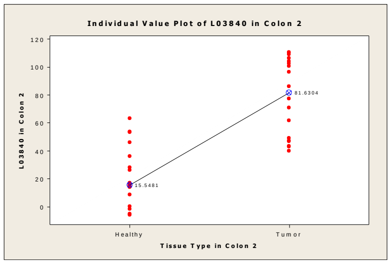


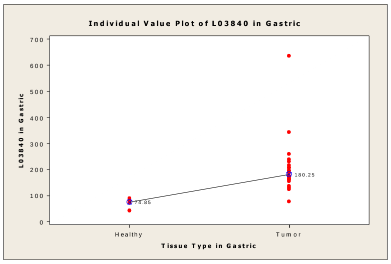




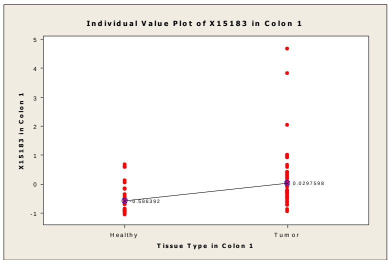
Gene FGFR4 (accession L03840), consistently overexpressed in all the databases

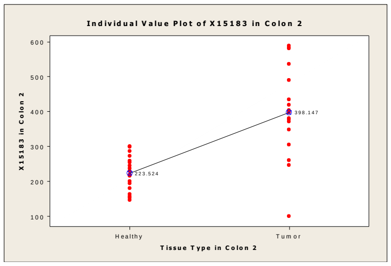


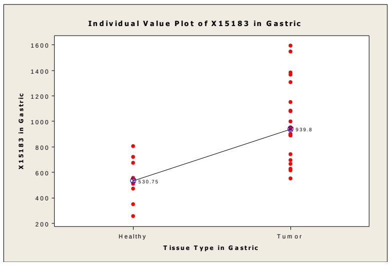




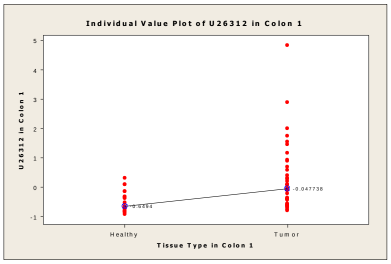
Gene HSP90AA1 (accession X15183), consistently overexpressed in all the databases

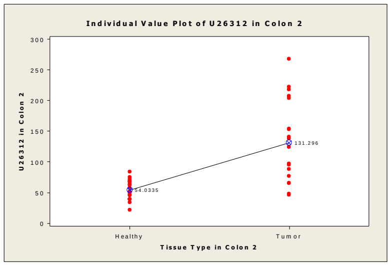


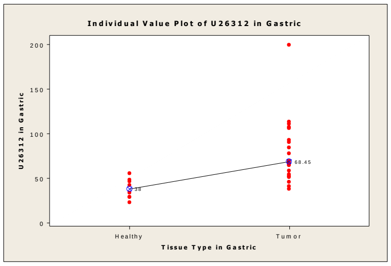




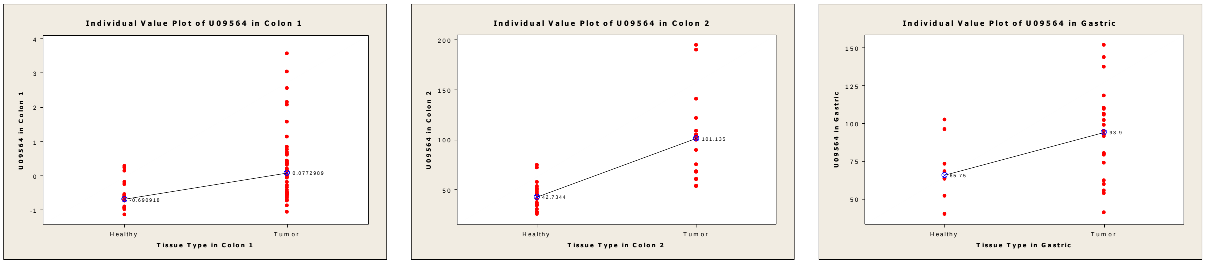
Gene CBX3 (accession U26312), consistently overexpressed in all the databases



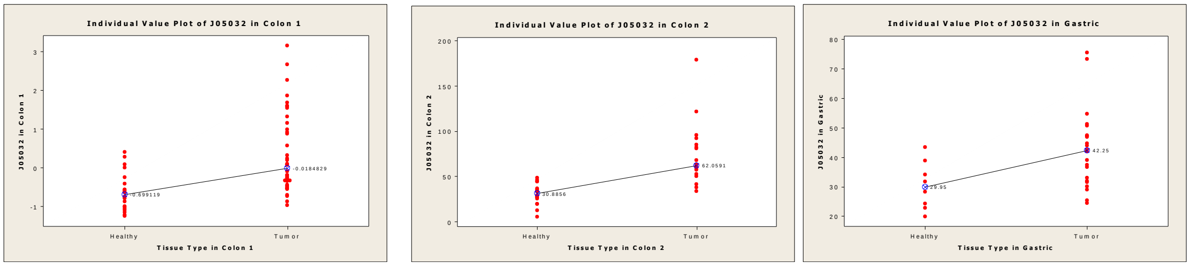




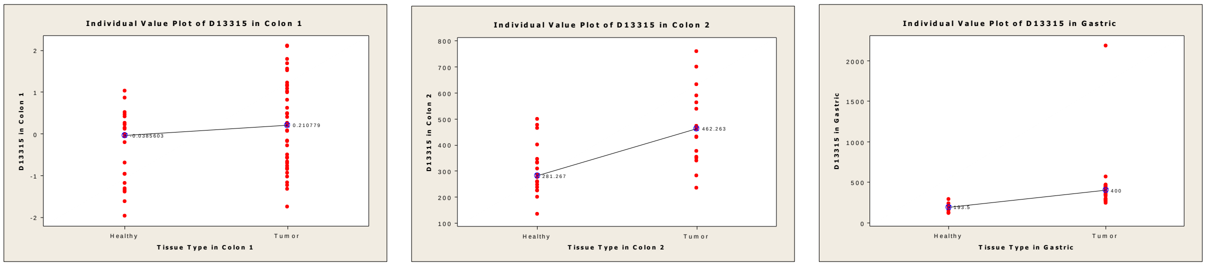
Gene SRPK1 (accession U09564), consistently overexpressed in all the databases



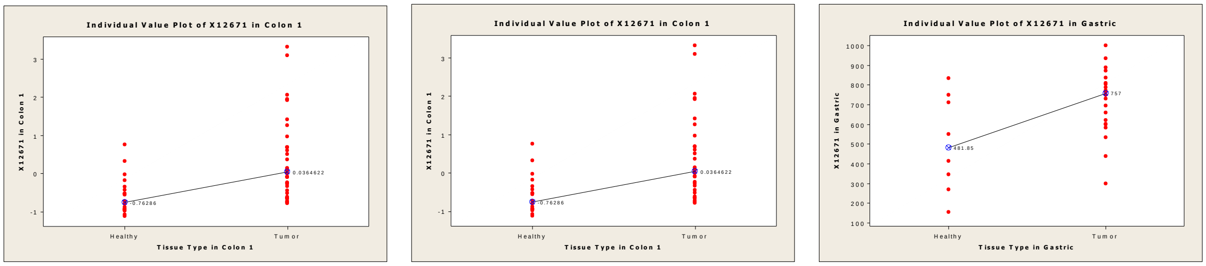
Gene DARS (accession J05032), consistently overexpressed in all the databases



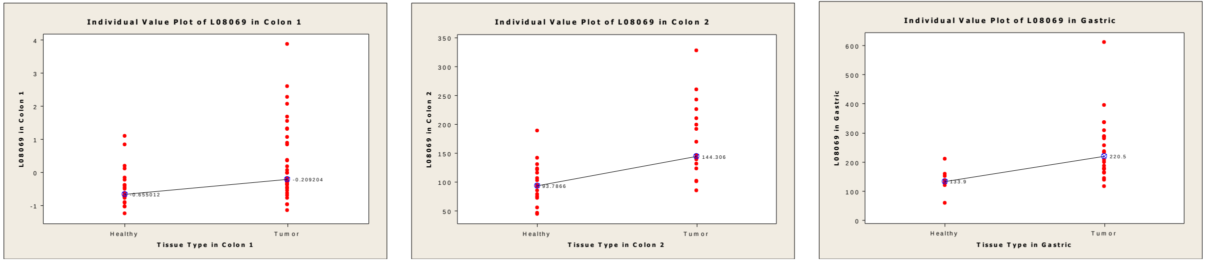
Gene GLO1 (accession D13315), consistently overexpressed in all the databases



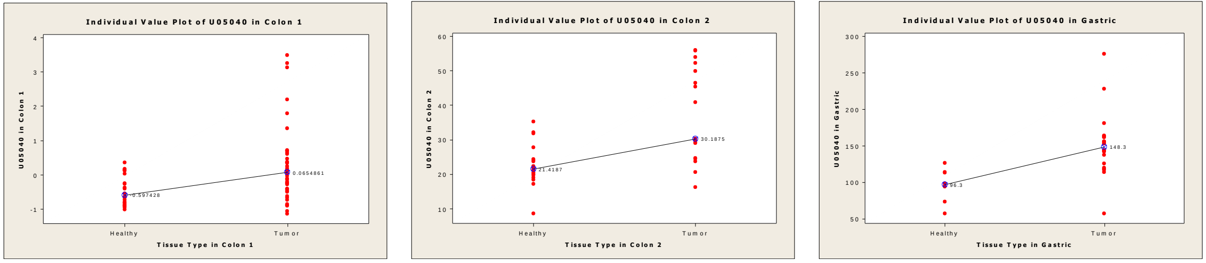
Gene HNRNPA1 (accession X12671), consistently overexpressed in all the databases



Gene DNAJA1 (accession L08069), consistently overexpressed in all the databases

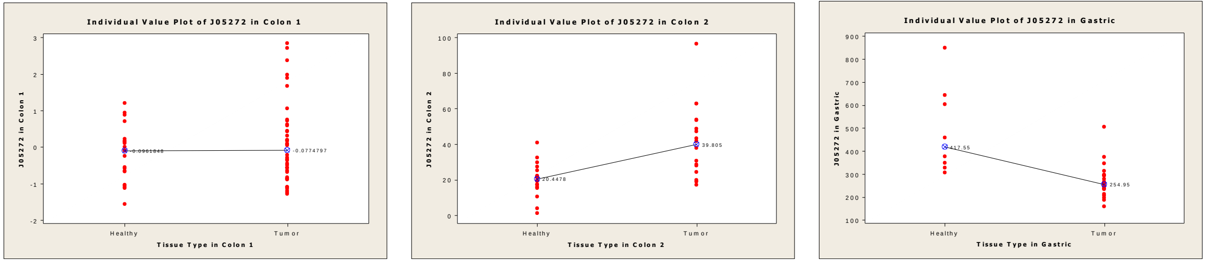


Gene FUBP1 (accession U05040), consistently overexpressed in all the databases

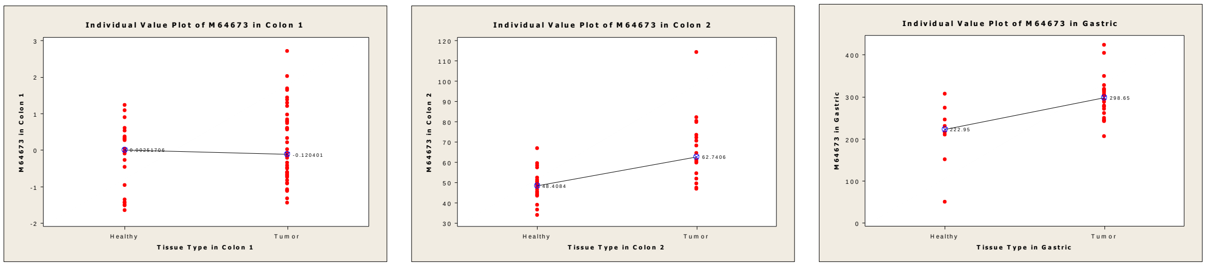


Graphs supporting Expression Profile for Group 8

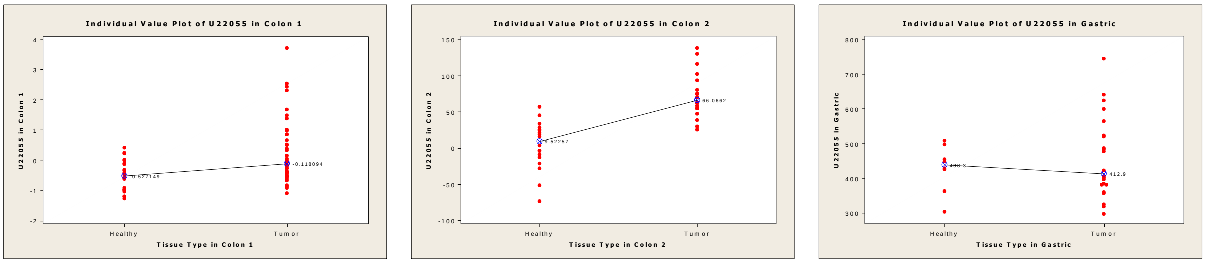
Gene HMPDH1 (accession J05272), with conflicting expression, overexpressed in Colon 2 but underexpressed in Colon1 and Gastric



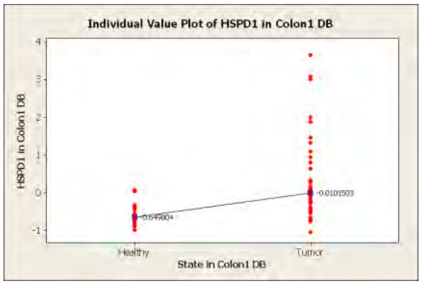
Gene HSF1 (accession M64673), with conflicting expression, overexpressed in Colon 2 and Gastric but underexpressed in Colon1



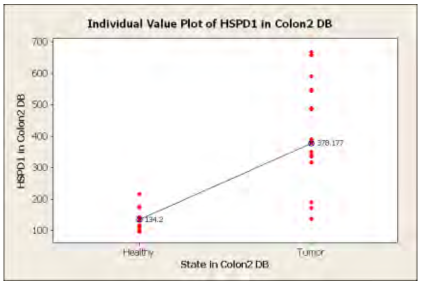
Gene SND1 (accession U22055), with conflicting expression, overexpressed in Colon 1 and Colon2 but underexpressed in Gastric



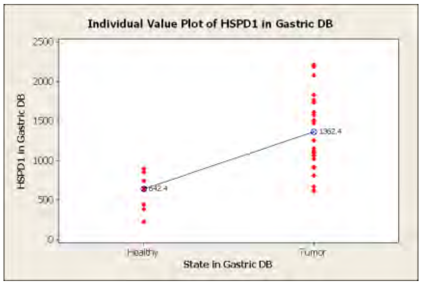
Appendix E – Statatistical validation of Expression profiles of 43 selected genes



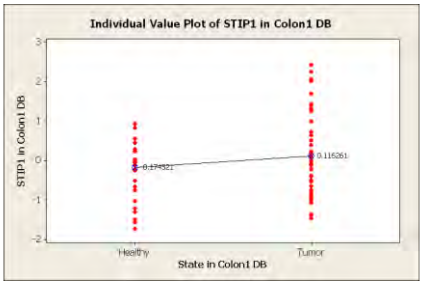
P_VALUE Colon1
2.41E-06



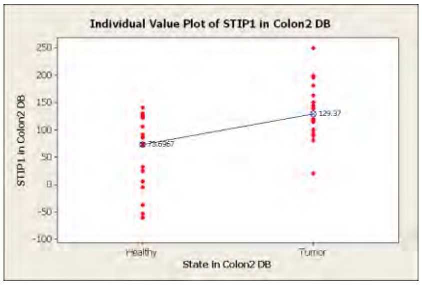
P_VALUE Colon2
2.24552E-06



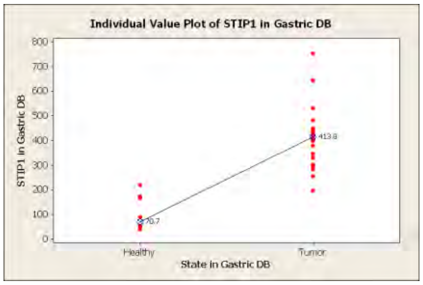
P_VALUE Gastric
0.000278426



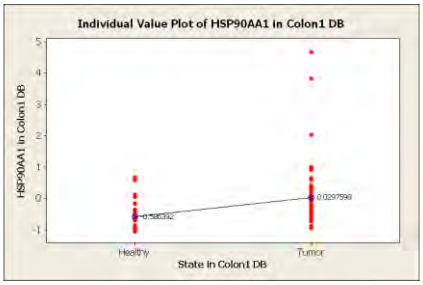
P_VALUE Colon1
0.086532988



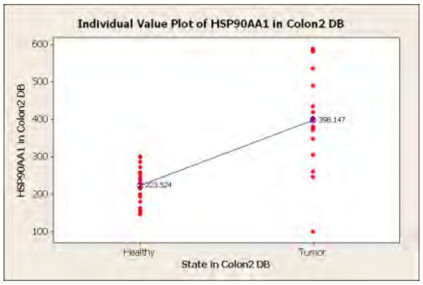
P_VALUE Colon2
0.001474381



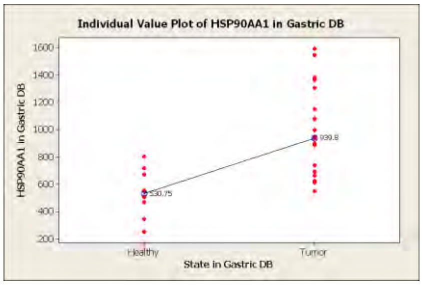
P_VALUE Gastric
4.97776E-05



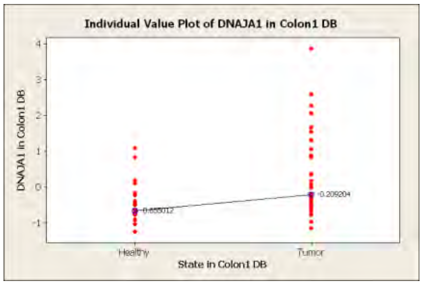
P_VALUE Colon1
0.001860511



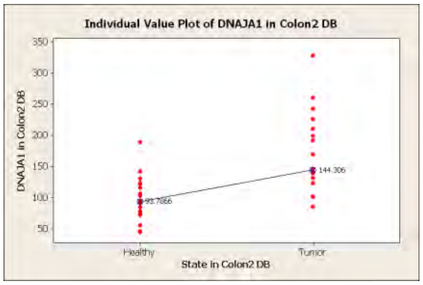
P_VALUE Colon2
2.40267E-05



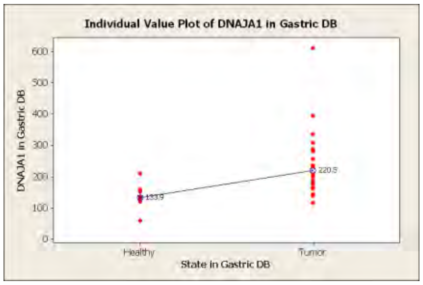
P_VALUE Gastric
0.000798829



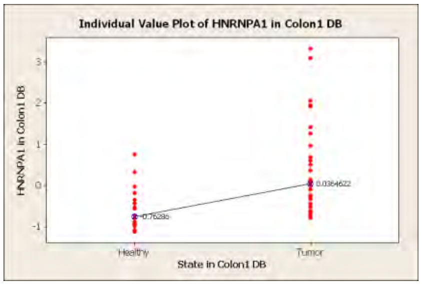
P_VALUE Colon1
0.002054658



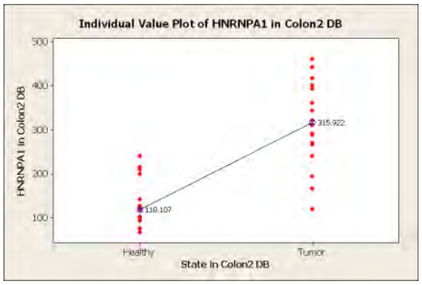
P_VALUE Colon2
0.000177418



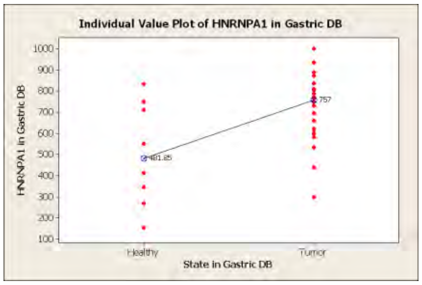
P_VALUE Gastric
0.001816442



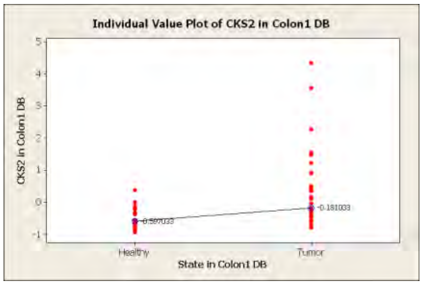
P_VALUE Colon1
2.05164E-05



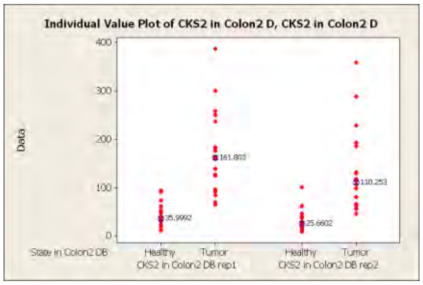
P_VALUE Colon2
4.15616E-06



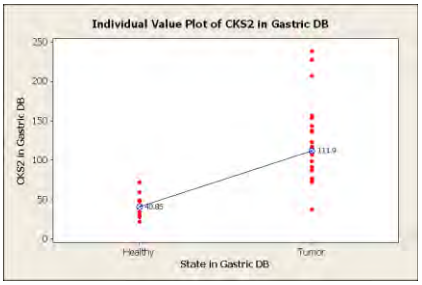
P_VALUE Gastric
0.032854679



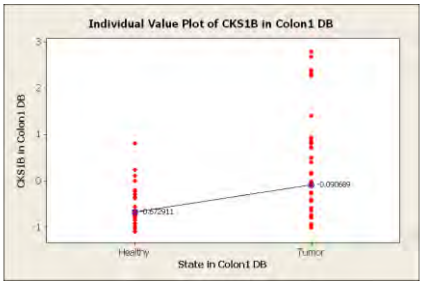
P_VALUE Colon1
0.000151618



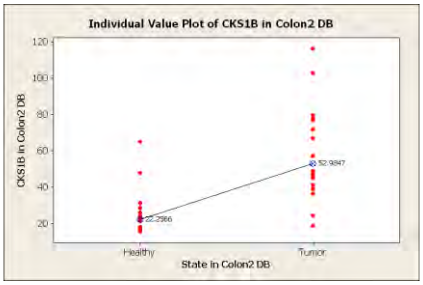
P_VALUE Colon2 ***P_VALUE Colon2, rep2***
1.40083E-06 2.24552E-06



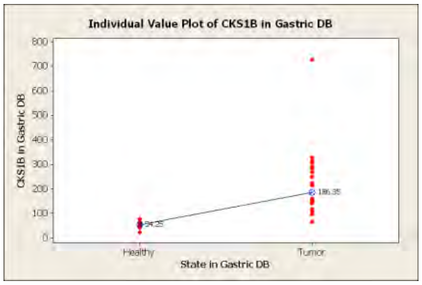
P_VALUE Gastric
9.00396E-05



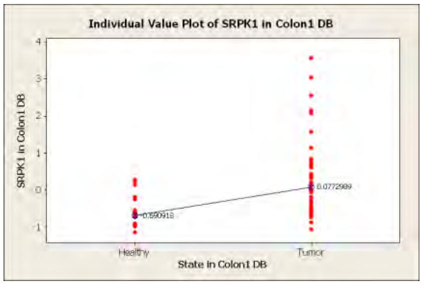
P_VALUE Colon1
0.004024381



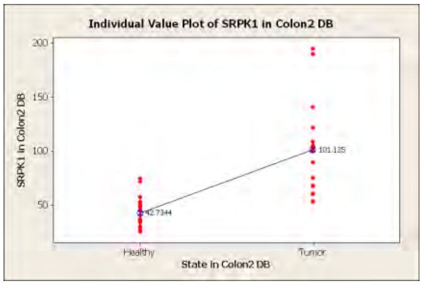
P_VALUE Colon2
3.64602E-05



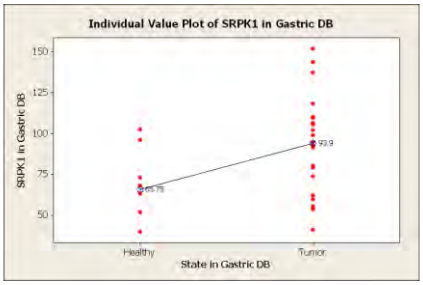
P_VALUE Gastric
6.07777E-05



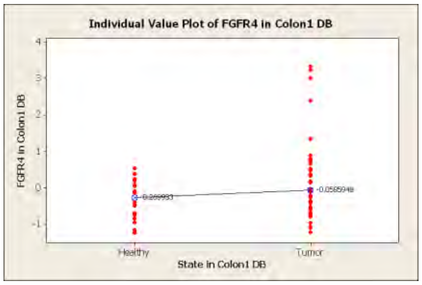
P_VALUE Colon1
2.84371E-05



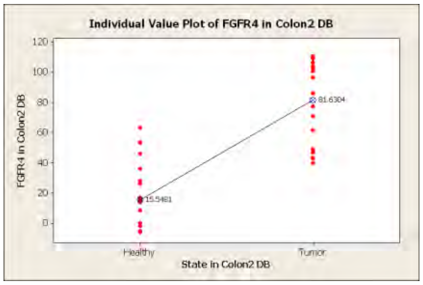
P_VALUE Colon2
3.56843E-06



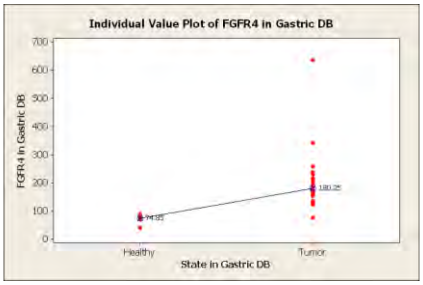
P_VALUE Gastric
0.057516482



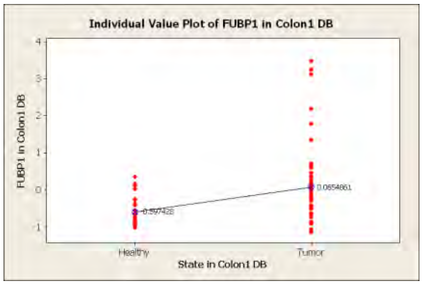
P_VALUE Colon1
0.058687946



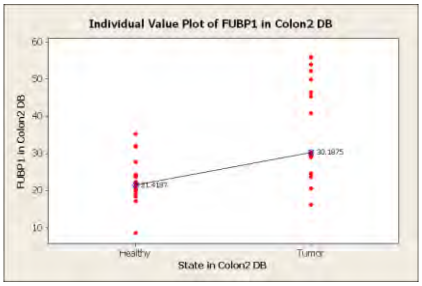
P_VALUE Colon2
1.01673E-05



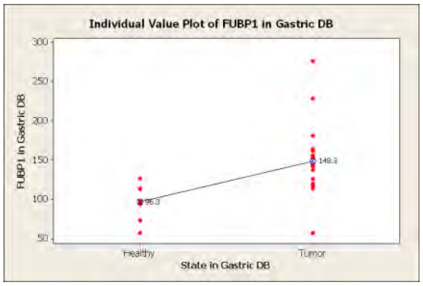
P_VALUE Gastric
9.00396E-05



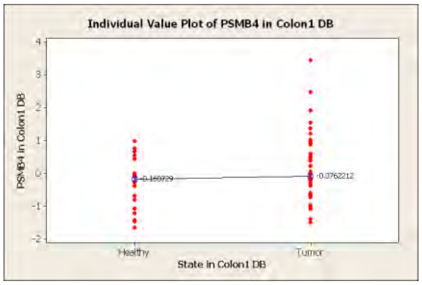
P_VALUE Colon1
0.004415835



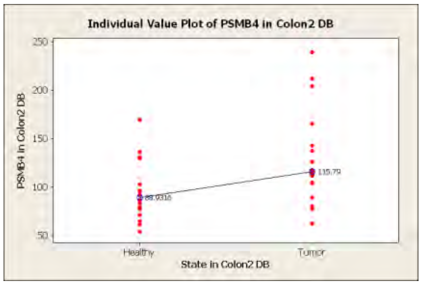
P_VALUE Colon2
0.000844225



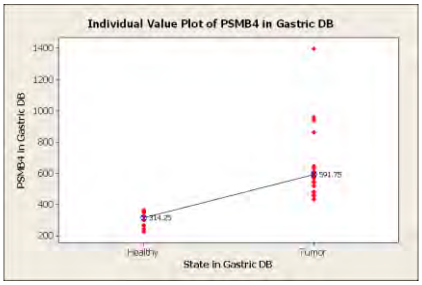
P_VALUE Gastric
0.00051885



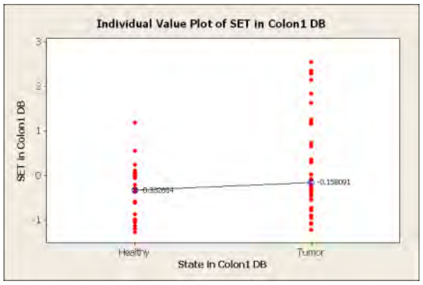
P_VALUE Colon1
0.147293926



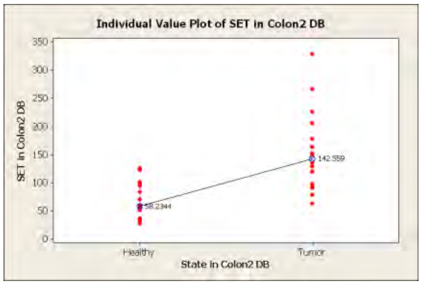
P_VALUE Colon2
0.014206802



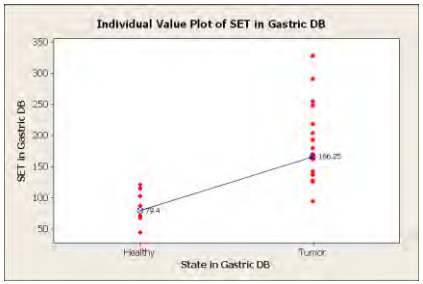
P_VALUE Gastric
4.06829E-05



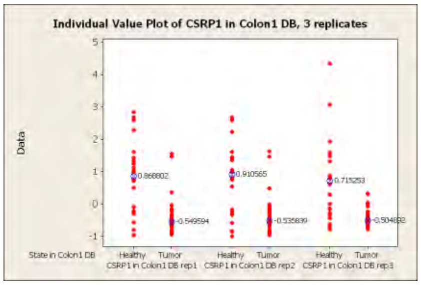
P_VALUE Colon1
0.0567494



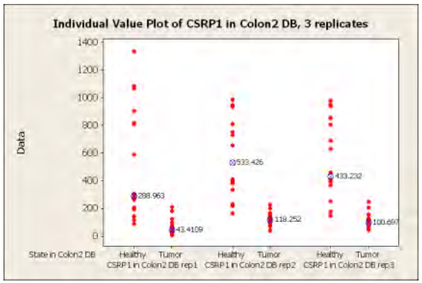
P_VALUE Colon2
8.182E-05



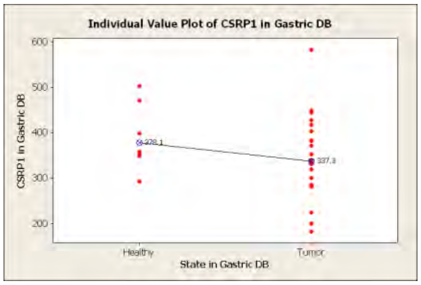
P_VALUE Gastric
7.40532E-05



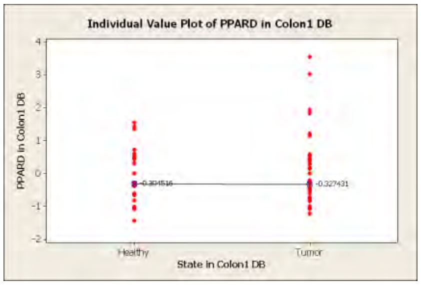
P_VALUE Colon1	***P_VALUE Colon1, rep2***	***P_VALUE colon1,rep3***
7.34111E-05	0.00016085	5.05166E-05



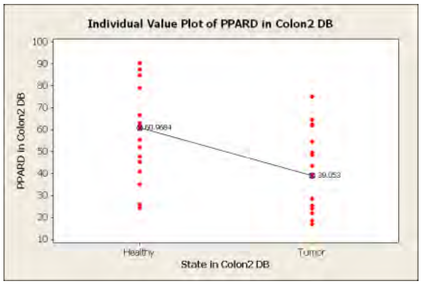
P_VALUE Colon2	***P_VALUE Colon2, rep2***	***P_VALUE Colon2, rep3***
3.56843E-06	6.25802E-07	8.66317E-07



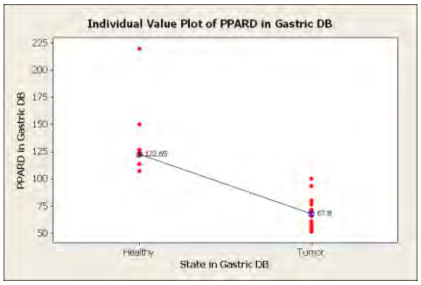
P_VALUE Gastric
0.11616411



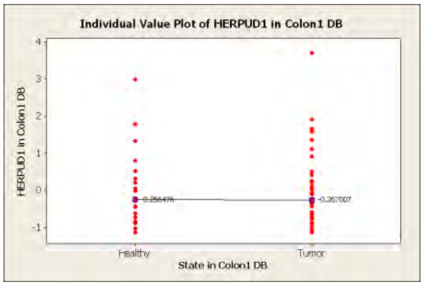
P_VALUE Colon1
0.947214527



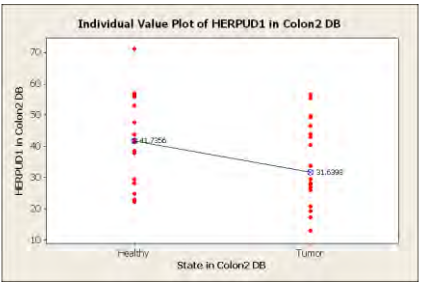
P_VALUE Colon2
0.01086823



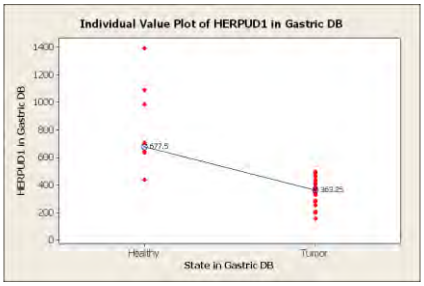
P_VALUE Gastric
4.06829E-05



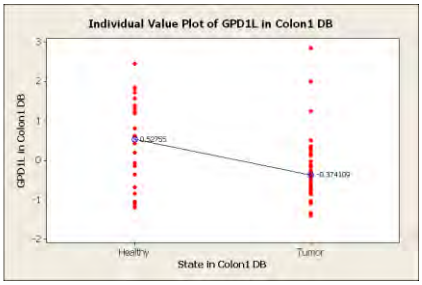
P_VALUE Colon1
0.912138265



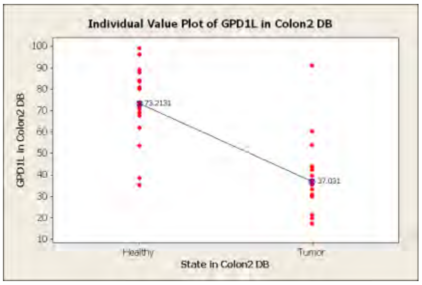
P_VALUE Colon2
0.200066091



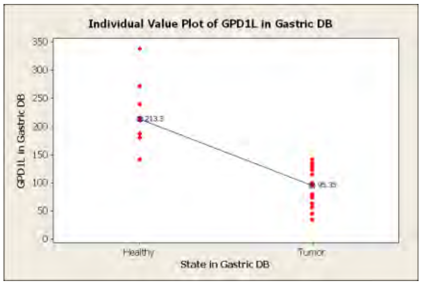
P_VALUE Gastric
9.00396E-05



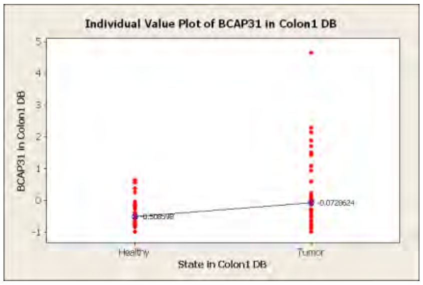
P_VALUE Colon1
0.027843742



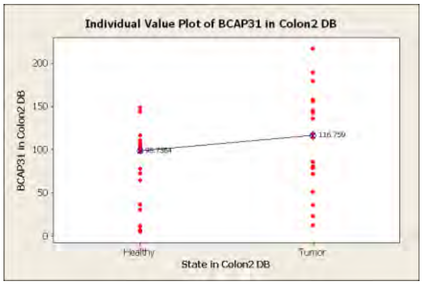
P_VALUE Colon2
5.48537E-05



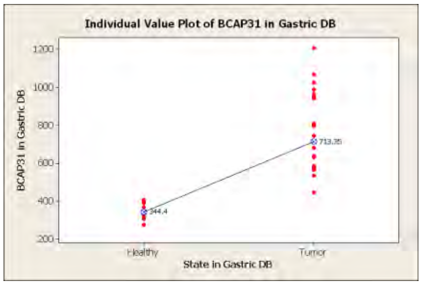
P_VALUE Gastric
4.06829E-05



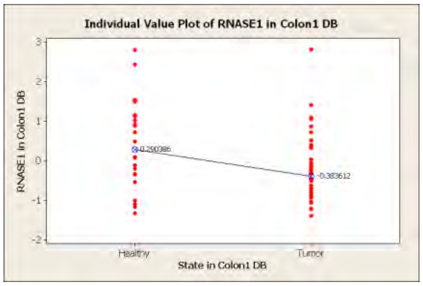
P_VALUE Colon1
0.078728193



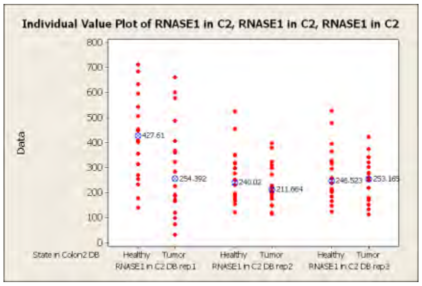
P_VALUE Colon2
0.096708126



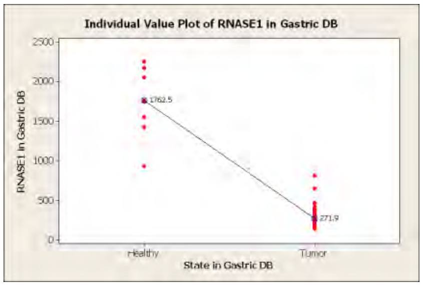
P_VALUE Gastric
4.06829E-05



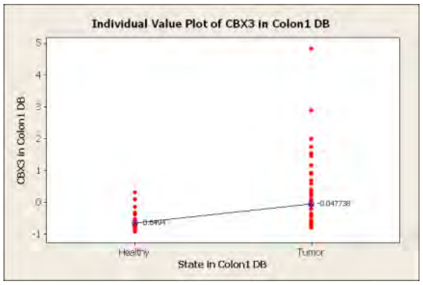
P_VALUE Colon1
0.030002569



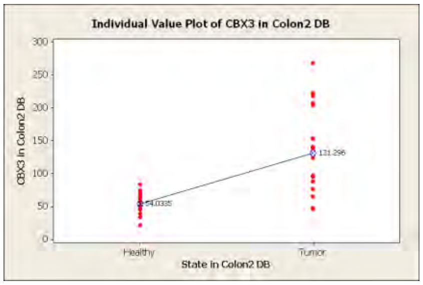
P_VALUE Colon2	***P_VALUE Colon2, rep2***	***P_VALUE Colon2, rep3***
0.038234836	0.438253572	0.516602582



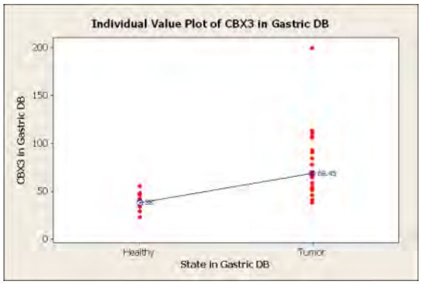
P_VALUE Gastric
4.06829E-05



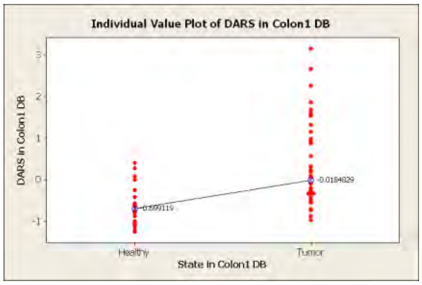
P_VALUE Colon1
0.000502451



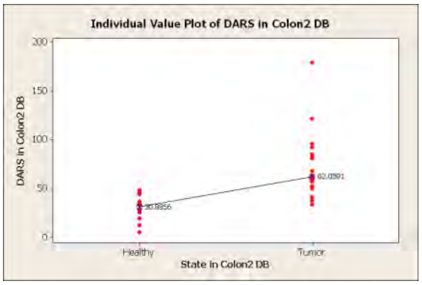
P_VALUE Colon2
7.16785E-05



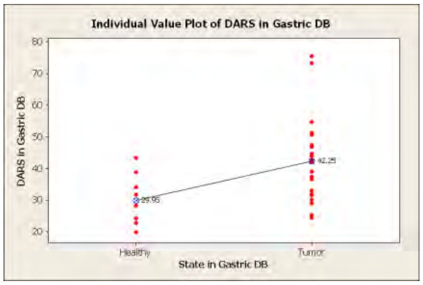
P_VALUE Gastric
0.000566857



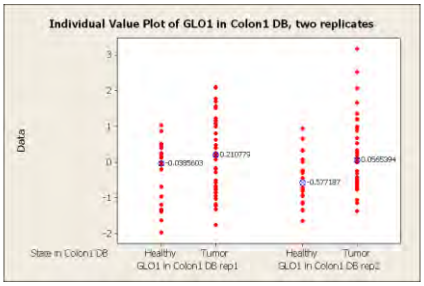
P_VALUE Colon1
1.68207E-05



P_VALUE Colon2
4.15616E-06

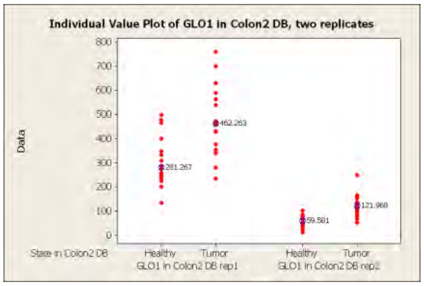


P_VALUE Gastric
0.013811074



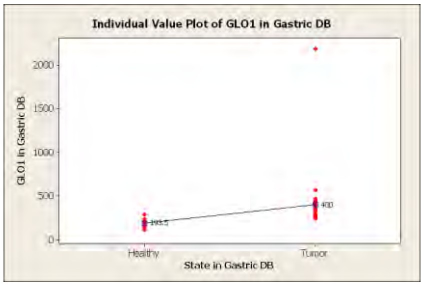
P_VALUE Colon1 ***P_VALUE Colon1, rep2***

0.051250927 0.004624229



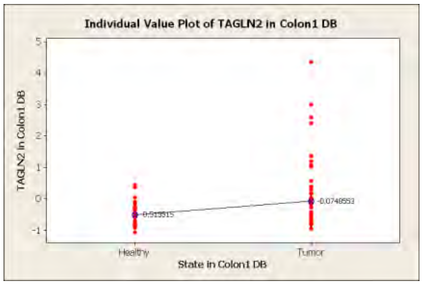
P_VALUE Colon2

0.000844225



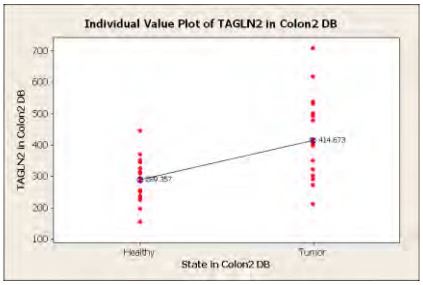
P_VALUE Gastric

0.000109248



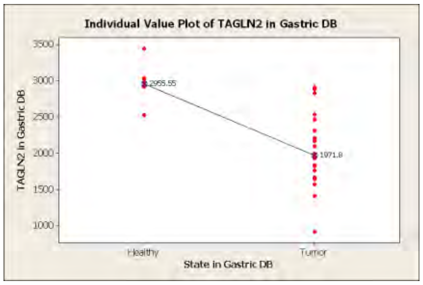
P_VALUE Colon1

0.001955373



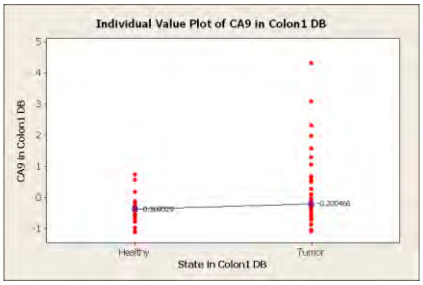
P_VALUE Colon2

0.000945592



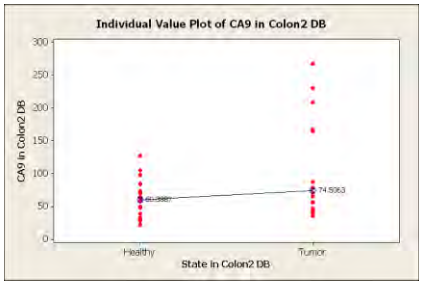
P_VALUE Gastric

9.00396E-05



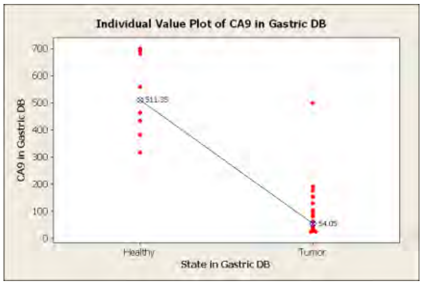
P_VALUE Colon1

0.032303729



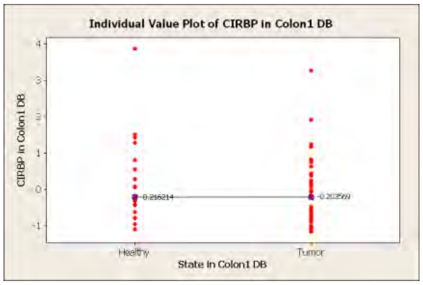
P_VALUE Colon2

0.117322705



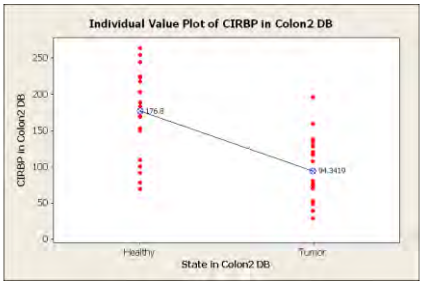
P_VALUE Gastric

8.98771E-05



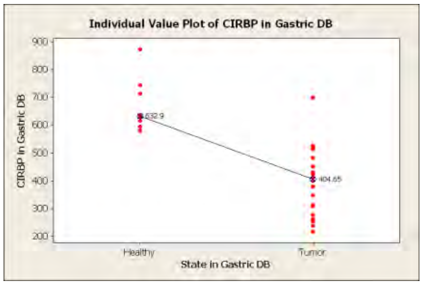
P_VALUE Colon1

0.493899714



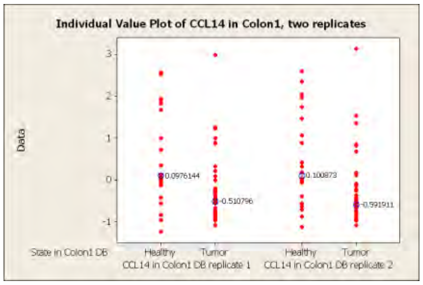
P_VALUE Colon2

0.001058135

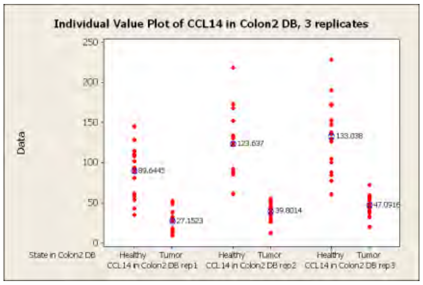


P_VALUE Gastric

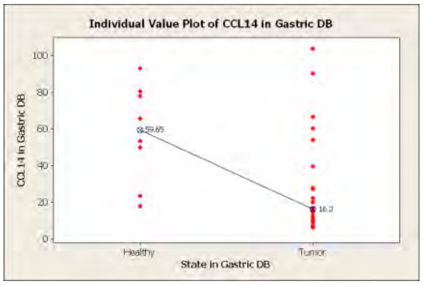
0.000109248



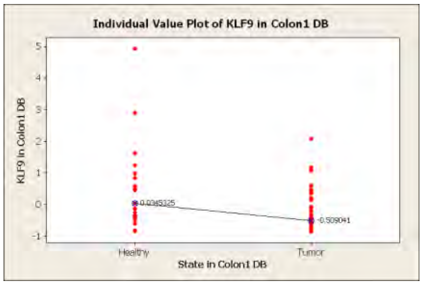
P_VALUE Colon1	***P_VALUE Colon1, rep2***
0.003496036	0.002158549



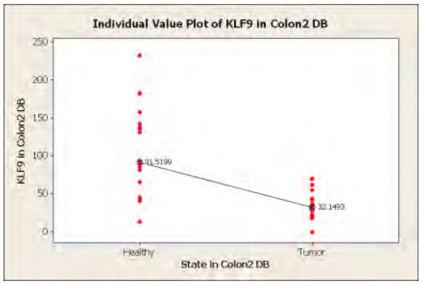
P_VALUE Colon2	***P_VALUE Colon2, rep2***	***P_VALUE Colon2, rep3***
1.01781E-06	3.22784E-07	4.50314E-07



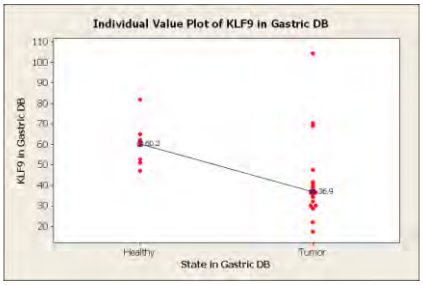
P_VALUE Gastric
0.015724754



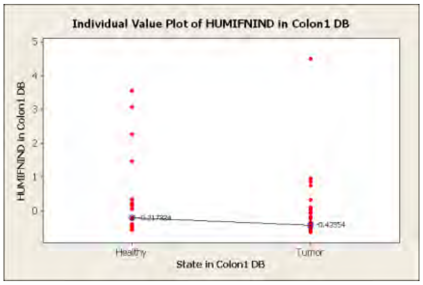
P_VALUE Colon1
0.006072622



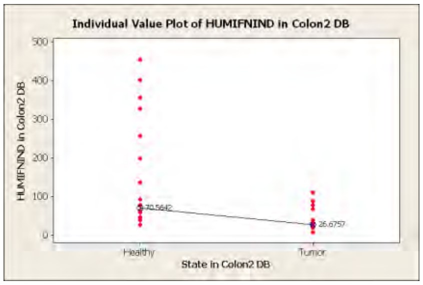
P_VALUE Colon2
0.000106306



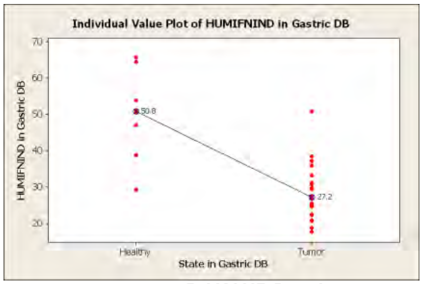
P_VALUE Gastric
0.002481501



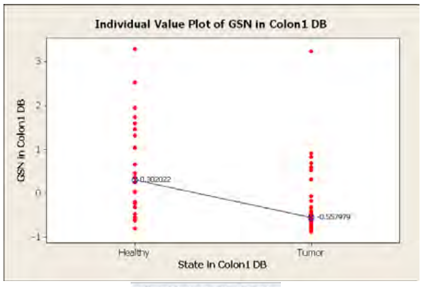
P_VALUE Colon1
0.049520236



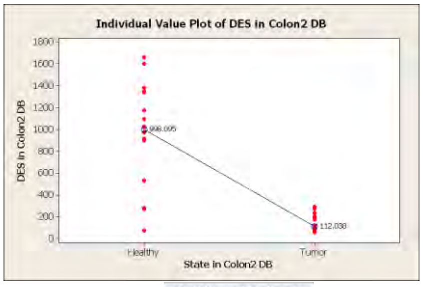
P_VALUE Colon2
0.000227899



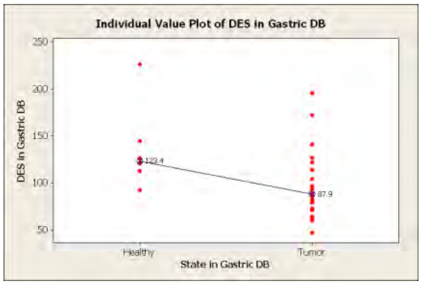
P_VALUE Gastric
0.00047603



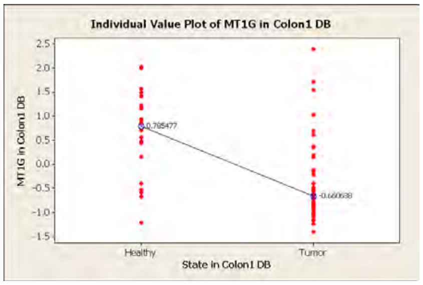
P_VALUE Colon1
3.45046E-05



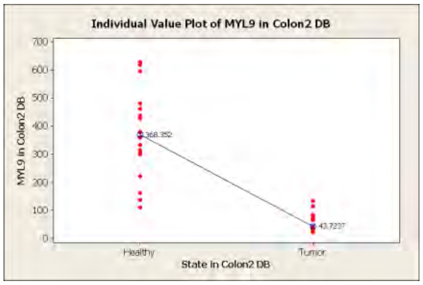
P_VALUE Colon2
6.52874E-06



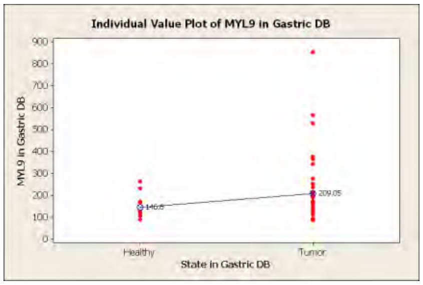
P_VALUE Gastric
0.015724754



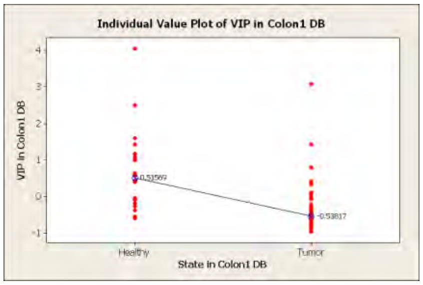
P_VALUE Colon1
2.49713E-05



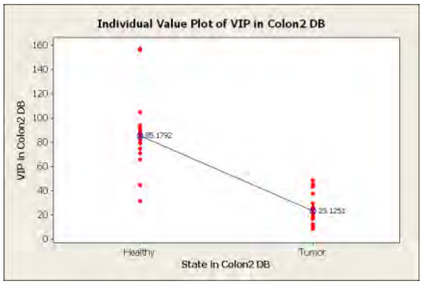
P_VALUE Colon2
4.50314E-07



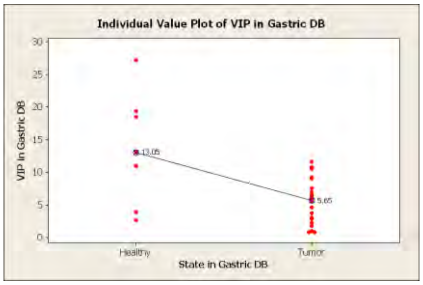
P_VALUE Gastric
0.105667399



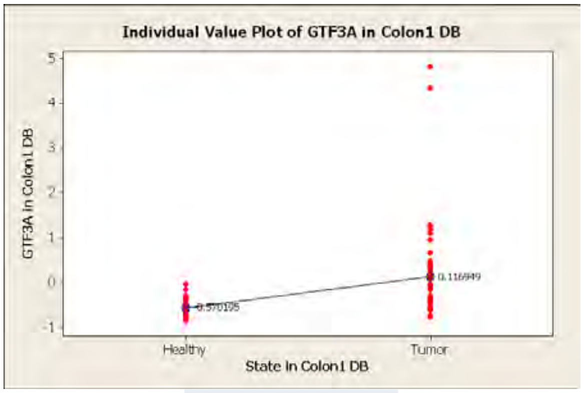
P_VALUE Colon1
5.37914E-05



P_VALUE Colon2
8.66317E-07

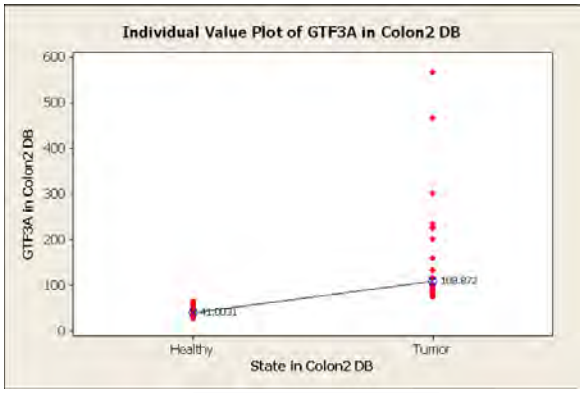


P_VALUE Gastric
0.008048548

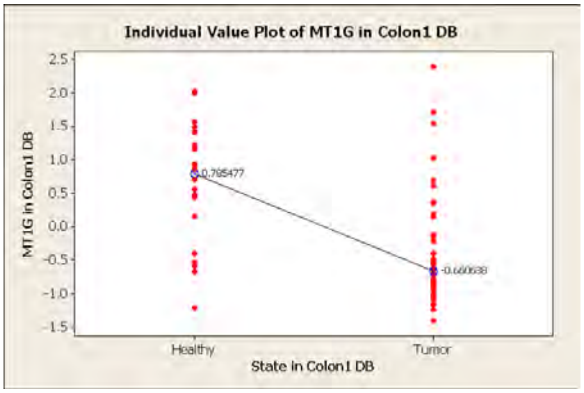


P_VALUE Colon1

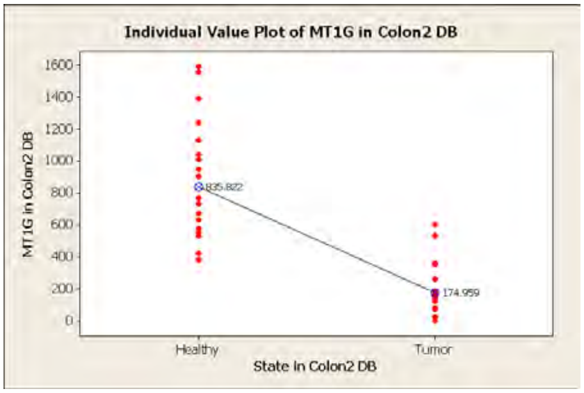
2.41384E-06



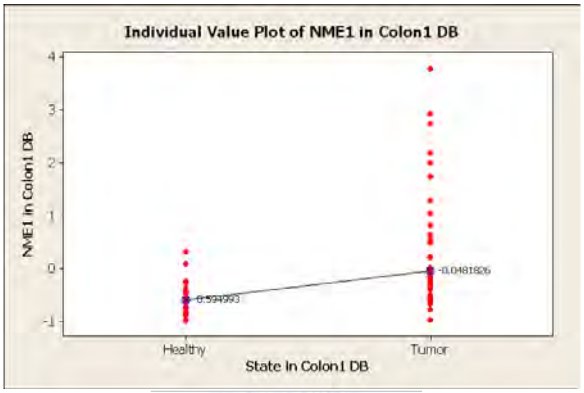
P_VALUE Colon2
3.22784E-07



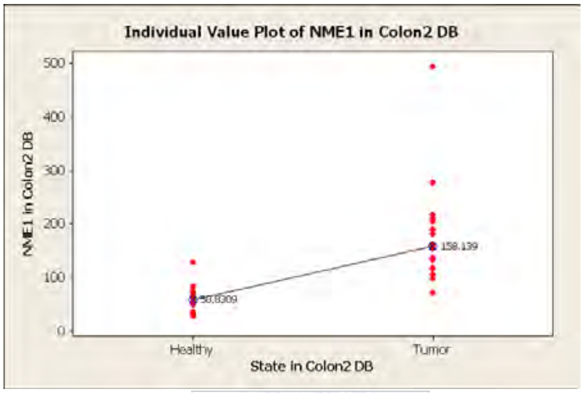
P_VALUE Colon1
0.00011945



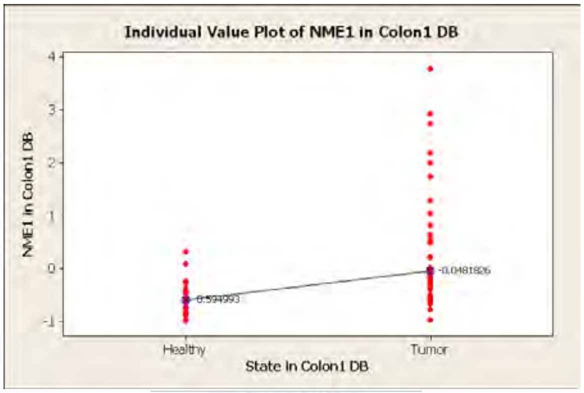
P_VALUE Colon2
1.01781E-06



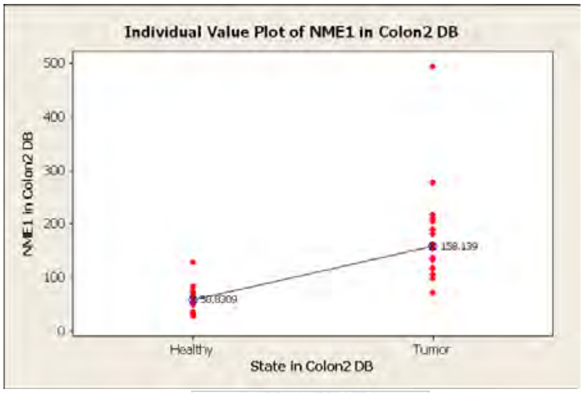
P_VALUE Colon1
5.72667E-05



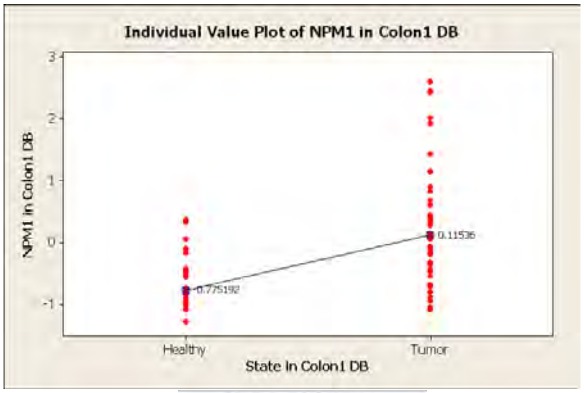
P_VALUE Colon2
1.01781E-06



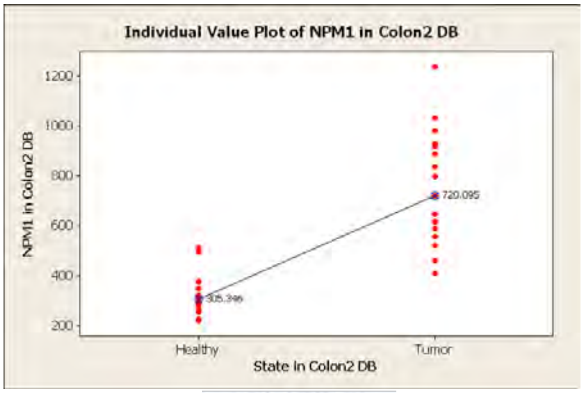
P_VALUE Colon1
5.72667E-05



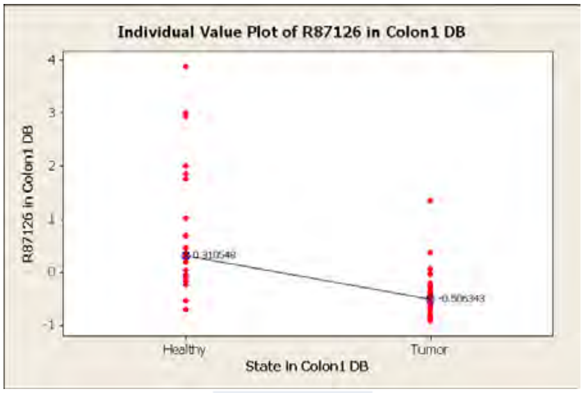
P_VALUE Colon2
1.01781E-06



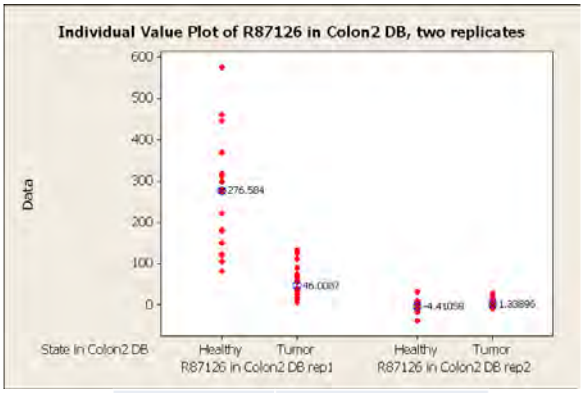
P_VALUE Colon1
9.37971E-05



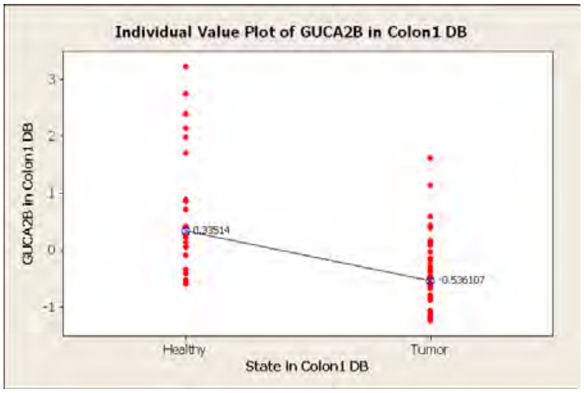
P_VALUE Colon2
6.25802E-07



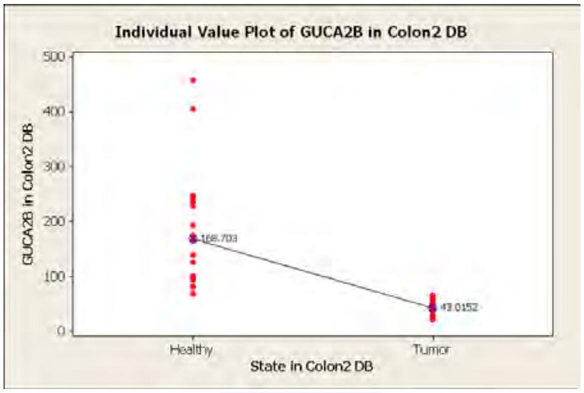
P_VALUE Colon1
6.85655E-07



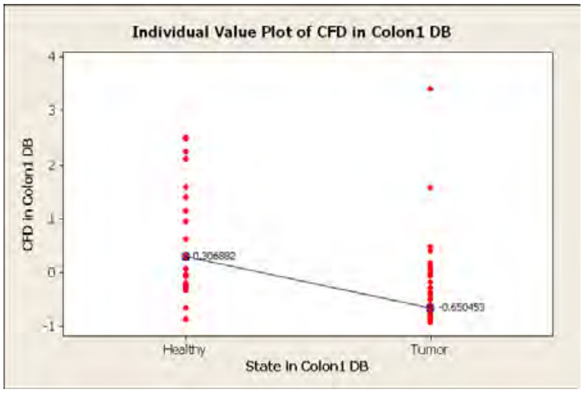
P_VALUE Colon2	P_VALUE Colon2, rep2
1.92055E-06	0.084653033



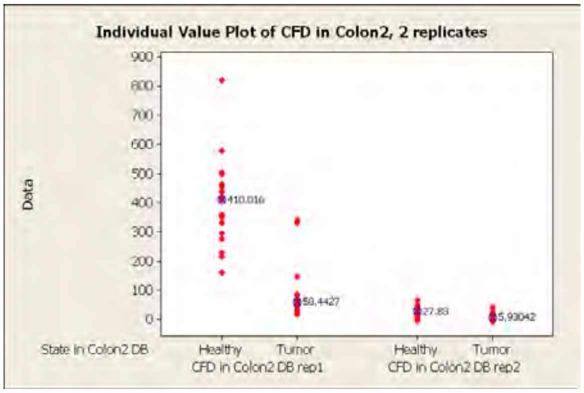
P_VALUE Colon1
2.19098E-05



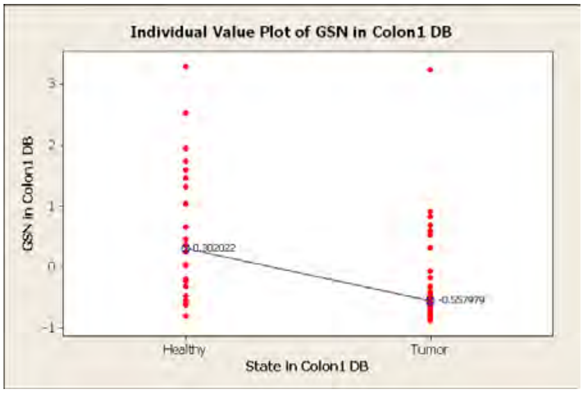
P_VALUE Colon2
3.22784E-07



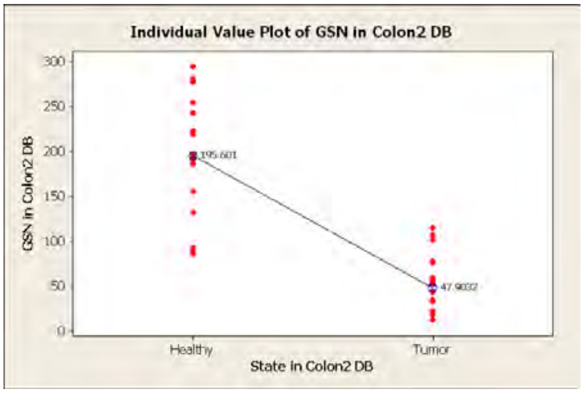
P_VALUE Colon1
3.23571E-05



P_VALUE Colon2	***P_VALUE Colon2, rep2***
1.92055E-06	0.000753017



P_VALUE Colon1
0.00020333



P_VALUE Colon2
1.40083E-06

Appendix F – Functional description of the validated genes

Table 1F. Functional description of Group 1 (13 genes) in Validation Scheme 3.

Gene	*Accession*	*Symbol*	*Name*	*Function*
1	M22382	HSPD1	Heat shock 60kDa protein 1 (chaperonin)	Prevent misfolding and promote the refolding and proper assembly of unfolded polypeptides generated under stress conditions in the mitochondrial matrix
2	R36977	GTF3A	General transcription factor IIIA	Is required for correct transcription of 5S RNA genes by RNA polymerase III
3	T86473	NME1	Non-metastatic cells 1, protein (NM23A) expressed in	Major role in the synthesis of nucleoside triphosphates other than ATP (131)
4	M26697	NPM1	Nucleophosmin (nucleolar phosphoprotein B23, numatrin)	Involved in ribosome biogenesis, centrosome duplication, protein chaperoning, histone assembly, and regulation of tumor suppressors TP53/p53 and ARF.
5	R87126		EST: yq31b10.s1	#N/A
6	Z50753	GUCA2B	guanylate cyclase activator 2B (uroguanylin)	Endogenous activator of intestinal guanylate cyclase, which synthesizes cyclic GMP (cGMP), a key component of several intracellular signal transduction pathways.[supplied by OMIM]
7	M63391	DES	desmin	Desmin are class-III intermediate filaments found in muscle cells.
8	J02854	MYL9	myosin, light chain 9, regulatory	Involved in cell locomotion
9	H43887	CFD	Complement factor D (adipsin)	Factor D cleaves factor B when the latter is complexed with factor C3b, activating the C3bbb complex, which then becomes the C3 convertase of the alternate pathway.
10	M76378	CSRP1	cysteine and glycine-rich protein 1	This gene encodes a member of the cysteine-rich protein (CSRP) family. This gene family includes a group of LIM domain proteins, which may be involved in regulatory processes important for development and cellular differentiation. The LIM/double zinc-finger motif found in this gene product occurs in proteins with critical functions in gene regulation, cell growth, and somatic differentiation.
11	M36634	VIP	vasoactive intestinal peptide	Involved in vasodilation
12	H06524	GSN	Gelsolin	Actin-modulating protein, involved in motility, signaling and apoptosis (96)
13	T71025	MT1G	Metallothionein 1G	Metallothioneins have a high content of cysteine residues that bind various heavy metals. Metallothioneins control the bioavailability of zinc, and Zinc is involved in several physiologic processes, including cell growth and proliferation (204)

Table 2E. Functional description of the Group 4 (10 genes) in Validation Scheme 3.

Gene	*Accession*	*Symbol*	*Name*	*Function*
1	D63874	HMGB1	high-mobility group box 1	DNA binding proteins that associates with chromatin and has the ability to bend DNA.
2	X54942	CKS2	CDC28 protein kinase regulatory subunit 2	Binds to the catalytic subunit of the cyclin dependent kinases and is essential for their biological function
3	X54941	CKS1B	CDC28 protein kinase regulatory subunit 1B	Binds to the catalytic subunit of the cyclin dependent kinases and is essential for their biological function
4	M86752	STIP1	Stress-induced-phosphoprotein 1	Mediates the association of the molecular chaperones HSC70 and HSP90 (HSPCA and HSPCB)
5	D26600	PSMB4	proteasome (prosome, macropain) subunit, beta type, 4	Is a multicatalytic proteinase complex that is responsible for the degradation of all short-lived proteins and 70-90% of all long-lived proteins (169)
6	X81817	BCAP31	B-cell receptor-associated protein 31	May play a role in anterograde transport of membrane proteins from the endoplasmic reticulum to the Golgi. May be involved in CASP8-mediated apoptosis
7	D42047	GPD1L	glycerol-3-phosphate dehydrogenase 1-like	Decreased enzymatic activity with resulting increased levels of glycerol 3-phosphate activating the DPD1L-dependent SCN5A phosphorylation pathway, may ultimately lead to decreased sodium current.
8	L07592	PPARD	peroxisome proliferator-activated receptor delta	Ligand-activated transcription factor. Receptor that binds peroxisome proliferators such as hypolipidemic drugs and fatty acids. Functions as transcription activator for the acyl-CoA oxidase gene.
9	D26129	RNASE1	ribonuclease, RNase A family, 1 (pancreatic)	Catalyzes the cleavage of RNA on the 3' side of pyrimidine nucleotides. Has been isolated mainly from pancreas, which is the tissue with highest expression (205)
10	D14695	HERPUD 1	homocysteine-inducible, endoplasmic reticulum stress-inducible, ubiquitin-like domain member 1	Component of the endoplasmic reticulum quality control (ERQC) system also called ER-associated degradation (ERAD) involved in ubiquitin-dependent degradation of misfolded endoplasmic reticulum proteins.

Table 3F. Functional description of the third Group 7 (originally 35 without 15 already explored, 20 genes) in Validation Scheme 3.

Gene	*Accession*	*Symbol*	*Name*	*Function*
1	X15183	HSP90AA1	Heat shock protein 90kDa alpha (cytosolic), class A member 1	Molecular chaperone. Has ATPase activity (By similarity)
2	U26312	CBX3	Chromobox homolog 3	Seems to be involved in transcriptional silencing in heterochromatin-like complexes and in the formation of functional kinetochore.
3	U09564	SRPK1	SFRS protein kinase 1	Plays a central role in the regulatory network for splicing, controlling the intranuclear distribution of splicing factors in interphase cells and the reorganization of nuclear speckles during mitosis.
4	L03840	FGFR4	Fibroblast growth factor receptor 4	Receptor for acidic fibroblast growth factor. The extracellular portion of the protein interacts with fibroblast growth factors, setting in motion a cascade of downstream signals, ultimately influencing mitogenesis and differentiation.
5	J05032	DARS	aspartyl-tRNA synthetase	Aspartyl-tRNA synthetase (DARS) is part of a multienzyme complex that charges its cognate tRNA with aspartate during protein biosynthesis. Several components of the translation apparatus show abnormal up- or down-regulation in cancer (178)
6	X12671	HNRNPA1	heterogeneous nuclear ribonucleoprotein A1	Involved in the packaging of pre-mRNA into hnRNP particles, transport of poly(A) mRNA from the nucleus to the cytoplasm and may modulate splice site selection.
7	U05040	FUBP1	Far upstream element (FUSE) binding protein 1	Regulates MYC expression by binding to a single-stranded far-upstream element (FUSE) upstream of the MYC promoter. May act both as activator and repressor of transcription
8	M93651	SET	SET nuclear oncogene	Multitasking protein, involved in apoptosis, transcription, nucleosome assembly and histone binding. Isoform 2 anti-apoptotic activity is mediated by inhibition of the GZMA-activated DNase, NME1.
9	L08069	DNAJA1	DnaJ (Hsp40) homolog, subfamily A, member 1	Co-chaperone of Hsc70. Seems to play a role in protein import into mitochondria
10	D13315	GLO1	Glyoxalase I	Catalyzes the conversion of hemimercaptal, formed from methylglyoxal and glutathione, to S-lactoylglutathione. Glyoxalase I activity is indeed higher in cancerous than in noncancerous specimens, suggesting that it may play a role in prostate cancer homeostasis and survival (187)
11	D21261	TAGLN2	Transgelin 2	The protein encoded by this gene is a homolog of the protein transgelin, which is one of the earliest markers of differentiated smooth muscle. (189)

Table 3F (continued)

Gene	*Accession*	*Symbol*	*Name*	*Function*
12	X66839	CA9	Carbonic anhydrase IX	Reversible hydration of carbon dioxide. Appears to be a novel specific biomarker for a cervical neoplasia and kidney cancer marker, associated with progression and survival (190)
13	D78134	CIRBP	Cold inducible RNA binding protein	Cold-inducible mRNA binding protein that plays a protective role in the genotoxic stress response by stabilizing transcripts of genes involved in cell survival. Seems to play an essential role in cold-induced suppression of cell proliferation.
14	Z49269	CCL14	chemokine (C-C motif) ligand 14	Has weak activities on human monocytes and acts via receptors that also recognize MIP-1 alpha. It induced intracellular Ca(2+) changes and enzyme release, but no chemotaxis, at concentrations of 100-1,000 nM, and was inactive on T-lymphocytes, neutrophils, and eosinophil leukocytes. Enhances the proliferation of CD34 myeloid progenitor cells.
15	X87159	SCNN1B	Sodium channel, nonvoltage-gated 1, beta	Sodium channel that controls the reabsorption of sodium in kidney, colon, lung and sweat glands.
16	U25138	KCNMB1	Potassium large conductance calcium-activated channel, subfamily M, beta member 1	Regulatory subunit of the calcium activated potassium KCNMA1 (maxiK) channel. Increases the apparent Ca(2+)/voltage sensitivity of the KCNMA1 channel.
17	D31716	KLF9	Kruppel-like factor 9	Transcription factor that binds to GC box promoter elements. Sp/KLF factors are involved in many growth-related signal transduction pathways and their overexpression can have positive or negative effects on proliferation. In addition to growth control, Sp/KLF factors have been implicated in apoptosis and angiogenesis (195).
18	M26683	HUMIFNIND	Human interferon gamma treatment inducible mRNA.	#N/A
19	M23254	CAPN2	Calpain 2, (m/II) large subunit	Calcium-regulated non-lysosomal thiol-protease which catalyze limited proteolysis of substrates involved in cytoskeletal remodeling and signal transduction, substrate degradation in some apoptotic pathways

Table 3F (continued)

Gene	*Accession*	*Symbol*	*Name*	*Function*
20	U14577	MAP1A	Microtubule-associated protein 1A	Structural protein involved in the filamentous cross-bridging between microtubules and other skeletal elements. MAPs are a family of proteins that bind to and stabilize microtubules, and microtubules are essential components of the cytoskeleton and play a critical role in many cellular processes, including cell division, cell motility (197)