IDENTIFYING THE MOLECULAR COMPONENTS THAT MATTER: A STATISTICAL MODELLING APPROACH TO LINKING FUNCTIONAL GENOMICS DATA TO CELL PHYSIOLOGY

By

VICTOR MANUEL TREVIÑO ALVARADO

A thesis submitted to

The University of Birmingham

for the degree of

DOCTOR OF PHILOSOPHY

School of Biosciences

The University of Birmingham

June 2007

ABSTRACT

Functional genomics technologies, in which thousands of mRNAs, proteins, or metabolites can be measured in single experiments, have contributed to reshape biological investigations. One of the most important issues in the analysis of the generated large datasets is the selection of relatively small sub-sets of variables that are predictive of the physiological state of a cell or tissue. In this thesis, a truly multivariate variable selection framework using diverse functional genomics data has been developed, characterized, and tested. This framework has also been used to prove that it is possible to predict the physiological state of the tumour from the molecular state of adjacent normal cells. This allows us to identify novel genes involved in cell to cell communication. Then, using a network inference technique networks representing cell-cell communication in prostate cancer have been inferred. The analysis of these networks has revealed interesting properties that suggests a crucial role of directional signals in controlling the interplay between normal and tumour cell to cell communication. Experimental verification performed in our laboratory has provided evidence that one of the identified genes could be a novel tumour suppressor gene. In conclusion, the findings and methods reported in this thesis have contributed to further understanding of cell to cell interaction and multivariate variable selection not only by applying and extending previous work, but also by proposing novel approaches that can be applied to any functional genomics data.

To my wife and children

ACKNOWLEDGEMENTS

I am infinitely grateful and indebted for the aid of genteel refined people and organizations that have made things possible I embarked and coursed this PhD research venture. To my wife, Lucia Cuéllar, for her love, support, sacrifice and positive influences, to my children Ian Leonardo, Luz Amani, and Saria Victoria for their comfortable, faithful, and familiar love and company, to my mom and dad Rita Alvarado and Miguel Cantú for their love, education and support, to Cuéllar-Chávez family for their support, to Dr. Francesco Falciani for his kind invitation, teaching, and friendship, to Darwin Trust of Edinburgh for their financial support, to the University of Birmingham for their academic services, to ITESM, Dr. José Treviño and Ing. Cuauhtémoc Durán for making the necessary arrangements to let me study without forcing me to quit my job, and to CONACyT for supplementary maintenance for my family. Other relatives and relatives-like people that were also important in my professional and intellectual development who has uninterestingly taught me and that have influenced the way I think, to Francisco Treviño and Jaime Treviño for impressing on me the fascination about Science and Biology, to José Gonzalez, Jorge Garza, and Lázaro Martínez for introducing me to the wonderful world of computer sciences.

Undoubtedly, I have enjoyed the company of other refined people I have met during these three years, Juan Carlos Cuevas-Tello and his family, Javier Gaxiola and his family, Fernando Ortega and his wife, Anabel Rodriguez, Aida Hidalgo, Anaximandro Gomez, Russell Compton, Adam Reynolds, Donatella Sarti, Nicholas Davies, Monica Falciani and Falciani's children, Dr. Karim Raza, Kekeletso Tloti, and Vibhor Gupta, and the delightful, tireless, and talkative Cristina Jimenez and Jonathan Jimenez. Thanks to all of them.

I also appreciate the kind academic and technical aid from Dr. Robert Insall, Dr. Dov Stekel, Antony Pemberton, Ann Begun, and Gill Green. I also value the collaboration and comprehension of all undergraduate Bioinformatics students who I had assisted during lectures.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS

ALL	Acute Lymphoblastic Leukaemia
ALLS	Acute Lymphoblastic Leukaemia Subclasses dataset
AML	Acute Myeloid Leukaemia
ANOVA	Analysis of Variance
BC	Breast Cancer
BP	Biological Process
BSA	Bovine serum albumin
CC	Cellular Component
CCP	Cyclic Citrullinated Peptide (antibody)
cDNA	Cloned DNA
CI	Clinical Information
COPA	Cancer Outlier Profile Analysis
CP	Capsular Penetration
CU	Chronic Uveitis
CV	Cross Validation
Cy3	Cyanine dye, Green Colour
Cy5	Cyanine dye, Red Colour
DEG	Differentially expressed genes
DLDA	Diagonal Linear Discriminant Analysis
DMARD	disease-modifying antirheumatic drug
DNA	Deoxyribonucleic Acid
ECM	Extra-Cellular Matrix
EDA	Estimation of Distribution Algorithms
ELISA	Enzyme-Linked Immunosorbent-Assay
EP	Emergent Patterns
FDR	False Discovery Rate
FG	Functional Genomics
FGD	Functional Genomics Data/Datasets
FGT	Functional Genomic Technology(ies)
FS	Forward Selection
GA	Genetic Algorithms
GEPAS	Gene Expression Pattern Analysis
GO	Gene Ontology
GS	Gleason Score
HC	Hierarchical Clustering
ICA	Independent Component Analysis
IGA	Intelligent Genetic Algorithm
IL	Interleukine
IPA	Ingenuity© Pathway Analysis (software)
IQR	Interquantile range
KNN	K-Nearest-Neighbours
LDA	Linear Discriminant Analysis
LIU	Lens-Induced Uveitis
LOOCV	Leave-one-out Cross Validation
MCMC	Markov Chain Monte Carlo
MDS	Multidimensional Scaling
MF	Molecular Function
mg	Milligrams

ml	Millilitres
MLHD	Maximum Likelihood discriminant functions
MM	Mismatched (probesets)
mRNA	Messenger RNA
MS	Mass Spectrometry
MVS	Multivariate Variable Selection
NC	Nearest Centroid
ng	Nanograms
NMR	Nuclear Magnetic Resonance
NN	Neural Networks
OOB	Out of (the) Bag
PBS	Phosphate buffered saline
PCA	Principal Component Analysis
pg	Picograms
pI	Isoelectric Point
PM	Perfect Matched (probesets)
pol	Polarization Index
QP	Quadratic programming
RA	Rheumatoid Arthritis
RF	Random Forest
RF	Rheumatoid Factor
RFE	Recursive Feature Elimination
RMA	Robust Multichip Average
RN	Relevance Networks
RNA	Ribonucleic Acid
SAM	Significance Analysis of Microarrays
SNP	Single Nucleotide Polimorphism
SOM	Self Organized Maps
SP	SwissProt
SVM	Support Vector Machines
TMSP	trimethylsilyl 2,2,3,3-tetradeuteropropionic acid
UK	United Kingdom
UVS	Univariate Variable Selection
VD	Vitreoretinal Disease
VS	Variable Selection

CHAPTER 1

System identification is a crucial step in the analysis and modelling of complex biological systems

1.1 - Making Sense of Large Scale Microarray Data: A Variable Selection Problem

Since the genomic era where large segments of DNA have been sequenced, biological sciences have been evolving from a mere descriptive or qualitative science to a quantitative one. It is now fully accepted that computational approaches are an integral part of modern biology to the extent that the development of computational models of biological systems is a top priority for most of the government funding bodies in the United Kingdom (BBSRC Cross-Committee Priorities[1] - Bioinformatics and e-Science), The United States of America (NCBI, The Bioinformatics and Computational Biology initiatives[2], NSF, DOE, and DARPA[3]), and for the European Union (European Commission, Fundamental Genomics, Bioinformatics[4]).

Biology has traditionally operated on a hypothesis driven strategy focussed on testing the involvement of a specific gene in a biological process of interest. This strategy has led to the development of a large body of knowledge on how genes

---

[1] http://www.bbsrc.ac.uk/science/areas/crosscommittee.html

[2] http://nihroadmap.nih.gov/bioinformatics/

[3]http://sciencecareers.sciencemag.org/career_development/previous_issues/articles/0630/federal_funds_and_bioinformatics_grants_a_match_made_in_heaven/(parent)/

[4] http://ec.europa.eu/research/health/genomics/index_en.htm

interact with each other in the context of complex functional pathways. Because of the qualitative nature of the approach, this knowledge has been often represented in the form of a cartoon (or more specifically a graph) showing the general topology of a functional gene network. The recent development of functional genomics technologies and consequently the progressive increase in complexity of the data that can be generated have created the necessity to develop computational techniques that allow the identification of the genes involved in a biological process and their organization in a pathway in the absence of any hypothesis.

The simplest approach for the identification of genes involved in particular biological processes (for example, response to chemical exposure) works by identifying up- and down-regulated genes using relatively standard statistical techniques. More complex bioinformatics methods have been developed to facilitate biological interpretation of gene lists (such as the one identified by differential expression) or for general data exploration and visualization purposes. Even more complex computational methods are required for developing statistical models based on molecular signatures predictive of phenotypic features of a biological system or for inferring gene regulatory networks from observational or interventional data.

In the last two cases it is important to search in the space of biological variables for a combination of informative expression profiles. Techniques designed to achieve these tasks are known as variable selection methods. So far, few researchers have applied these computational tools to their full potential in the analysis of biological data. Most of the work described in this thesis aims to develop and exploit such methodologies in biomarker identification and in understanding biological processes involved in cell to cell communication.

1.2 - Understanding cell to cell communication: The need for computational approaches

The recent development of genome-wide gene expression profiling and other functional genomics technologies has provided the scientific community with versatile tools to characterize the molecular state of cells at a genomic level. Statistical and mathematical analysis of generated datasets has identified gene signatures related to the molecular state of cells. These results have been essential to look for alternative experimental setups for biological investigations. However, most current studies performed at a genome level are based on analysis of individual cell types or tissues. Thus, important factors present in the tissue microenvironment expressed by nearby cells have not explicitly been taken into account in genome-wide studies.

As in any communication system, the process of cell-to-cell communication involves membrane proteins (transmitters), which secrete signal proteins (message) that travel through the extra-cellular matrix (media) to be detected by membrane receptors (sensors) in other cells. Receptors are coupled to signal transduction machineries delivering the message signal to final cellular effectors (interpretation). The importance of cell-to-cell communication has been studied and demonstrated in virtually every biological system of relevance [1-6].

When a system is sufficiently well understood and experimental measurements of its parameters are available (for example the binding affinity of a growth factor to its receptor) it is possible to develop computational models that represent the precise mechanism that govern the system of interest. Such models (for example based on differential equations) can be extremely useful to verify that a system is well understood. These models can in fact be used to simulate the behaviour of a system and to compare its results to observed data.

Unfortunately, very few of these parameters are known in cell to cell communication systems. For many of the gene products for which there is evidence of an involvement in this process, the precise molecular function is unknown.

It is necessary therefore to design and employ inference and data mining approaches to identify components and to infer the network structures involved in cell to cell communication from functional genomics data. Although some of these approaches are available, they have not been used in the context of cell to cell communication.

An important part of this thesis has been to address the problem of identifying molecular components involved in the interaction between normal and tumour cells. As a first approach to study cell to cell interaction, a multivariate variable selection method has been designed, implemented, and applied to show that the physiological state of a cell type can be inferred from the molecular state of adjacent cells. Afterwards, a gene selection approach that allows the identification of functional networks involved in cell to cell communication has been designed. The results, mainly reported in Chapters 5 and 6, have led in the identification of a novel putative tumour suppressor gene in cancer.

1.3 - Thesis Organisation and Summary

This thesis begins with an introductory Chapter reviewing the current techniques available for experimental and computational analysis using large scale datatsets (Chapter 2). This is followed by the description of a multivariate variable selection statistical modelling environment that has been developed in the statistical programming language R (Chapter 3). Chapter 4 describes the application of this

environment for biomarker discovery for both proteomics and NMR metabolomics data. Chapters 5 and 6 describe the application of statistical modelling to identifying genes involved in cell to cell communication between normal and tumour cells in prostate cancer. In Chapter 6, a simple metric to rank genes that are potentially involved in a particular case of cell to cell interaction has been proposed and characterised.

A large part of the work described in this thesis is based on variable selection methods. Multivariate variable selection methods seem to be more powerful than the univariate ones. From the multivariate selection methods, those using stochastic searches are the most robust, versatile, successful, and relatively fast. However, no software package was available that could be used in a variety of situations and datasets. Therefore, Chapter 3 presents GALGO, an R package that uses Genetic Algorithms (random) searches coupled with versatile fitness functions for classification and regression. This package was designed to be generic, easy to use for common large datasets, and flexible taking benefit from the free and robust R programming environment. GALGO package was inspired in a prototype tool developed in in C language during the first year of research. This C-based tool was difficult to modify, adapt, expand, and use. GALGO has been designed to surpass these difficulties.

In the microarray context, multivariate variable selection based on random search has been used successfully in the literature to solve several biological problems. However, several aspects of the multivariate search system such as redundant genes in models, generation of a gene list, model similarity, collinear genes, and specificity of gene-class have not been so far studied.

The methods and data studied in this thesis are mainly based on transcriptomics data, that is, data from the expression of genes detected by the mRNA present in cells. Although transcriptomics is by far the most used functional genomics technology, other technologies have emerged that quantify other cell aspects. Proteomics determines the amount of proteins present in a sample whereas metabolomics detects the relative amount of metabolites. These technologies produce very similar datasets to those produced by transcriptomics, with some peculiarities though. Therefore, it would be worth to know whether methods presented in this thesis would be useful to analyse this kind of data. Thus, Chapter 4 presents successful studies based on proteomics and metabolomics data using GALGO.

The first ideas to provide supportive evidence of the interaction between tumour and surrounding normal cells are shaped in Chapter 5. There, data in which surrounding normal (not tumoural) cells are predictive of tumour features is presented. One of the explanations why normal cells are, in some sense, aware of the physiological state of the tumour is that there is some interaction between normal and tumour cells. However, other logical explanation may be that surrounding normal cells are carrying the same defect than tumour cells in such a way that adjacent normal cells are lagged in the progression of malignancy transformation. Given the results obtained in which selected genes seem to be somehow related to cell to cell interactions, we believe that the former hypothesis is more realistic.

Under the assumption that there is some interaction between normal and tumour cells, in Chapter 6 it has been hypothesised how components of the assumed communication could be revealed. The proposal was based on the observation that some genes expressed in normal cells have correlations with the expression of several other genes in tumour cells but the same was not observed in the other direction (the same gene expressed in tumour were not correlated with several genes in normal

cells). To identify and rank these genes, Chapter 6 propose and characterise a descriptive metric. It is shown also that this metric displays interesting properties and analysis of the selected genes in a number of datasets is provided.

Chapter 7 concludes this manuscript with a series of general considerations on the biological findings.

1.4 - Content Acknowledgments

Some data presented in this thesis has been acquired by collaborators consisting on experimental assays or computational data processing. In Chapter 4, Dr. Karim Raza, Dr. Stephen Young, and Dr. Mike Salmon were important contributors. In Chapter 5, Dr. Mahlet G. Tadesse, Dr. Marina Vannucci, Dr. Fatima Al-Shahrour, Dr. Joaquin Dopazo, and Dr. Moray J. Campbell provided complementary data. In Chapter 6, Dr. Moray Campbell, Dr. Farida Latif, and Dr. Heiner supplied important results. All are properly referred and acknowledged as appropriate in respective Chapters.

CHAPTER 2

Linking Molecular Signatures to Cell Physiology: A Variable Selection Problem

The development of Functional Genomics Technologies (FGT) has been responsible for an important revolution in biological sciences. These techniques have contributed to characterize biological systems at an unprecedented level of detail. Such advances have boosted the development of computational methods to analyze large datasets and to generate new hypotheses on the global behaviour of biological systems. So far, studies based on a functional genomics (FG) approach have been performed to investigate the response of cells and tissues to a variety of stimuli or to associate molecular signatures to specific aspects of cell physiology. Our research interests have been focussing on the development of analysis methods to reveal a link between the molecular state of cells and the physiology of a biological system. As part of this effort, a statistical modelling techniques based on a multivariate variable selection strategy have been developed and applied. This introductory chapter summarizes the current state of the art on FGT and the statistical modelling techniques used to make sense of these complex datasets. Other background information specifically related to individual aspects of my work is contained in their respective Chapters.

2.1 - Introduction

The relatively recent introduction of FGT has contributed to change the way the experimental data is acquired and analysed in Biology (Figure 2.1). The ability to monitor the expression of thousands of genes in single experiment has increased our capacity to characterize cell identity and the dynamics of cell response to stimuli. In

many cases however this large amount of information has not immediately resulted in a better understanding of cell physiology. The interpretation of descriptive data is indeed difficult when so many variables are measured. For this reason a number of statistical and data mining approaches have been devised to facilitate the extraction of significant patterns from large scale datasets. In this context, a key strategy to formulate new meaningful hypothesis is to identify statistical properties that link physiological readouts to molecular signatures. Our interest is to develop and apply such methods to identify components of the cell machinery linking the molecular state of a cell to its physiological state. Practical examples of this are the identification of genes expressed in a bacterial cell during infection that are responsible for certain specific aspects of the host cell response and the identification of genes expressed in tumour cells that are related to tumour aggressiveness.

The link between cell physiology and molecular signatures can be achieved using appropriate statistical techniques coupled with efficient methods of variable selection (VS) which are reviewed in this chapter.

The conceptual problem of linking molecular signatures to a physiological readout has found its natural place in clinical informatics and in particular in the identification of Biomarkers of clinical significance. Most of the work being performed in the field is, therefore, referring to the problem of identifying a sufficiently small subset of genes that can explain the behaviour of disease. A clear example is the identification of markers predictive of tumour stage [7; 8] or survival after therapy [9].

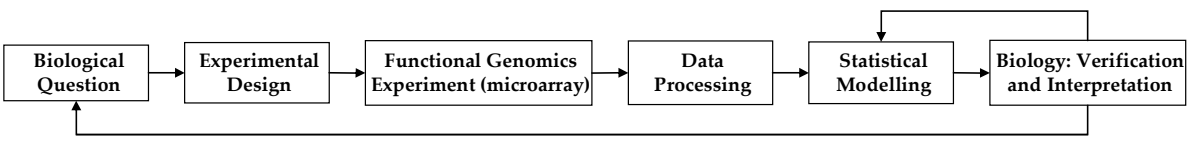


Figure 2.1 - Schematic representation of biology knowledge generation using FG.

Developing statistical models to identify molecular markers of clinical significance is not the only application of machine learning techniques in Biology. Once gene expression profiles are found to be linked to a specific aspect of cell physiology, a hypothesis can be made on the mechanisms and pathways underling the observed link. Figure 2.1 sketches the idealized learning cycle associated to statistical modelling of biological data.

This review aims to provide an overview of the computational and statistical methods used in the development of statistical models from FG data. In this context, several methods have been proposed that are difficult to classify in terms of method, search strategy, model size, pre-processing, and particularities. Therefore, a table summarizing the overall data processing of the methods mentioned in this chapter is provided (Table 2.1).

2.2 - Functional Genomics Technologies

In the last decade, to measure the activity of a molecule, a simple but tedious, time consuming and expensive experiment had to be performed. This was critical when measuring or comparing the activity of several molecules. FGT has evolved from these simple manually performed laboratory techniques to complex automated assays using robotics. In some sense then, FGT are just the miniaturization and automation of well established laboratory assays. Thus, FGT can be defined as a set of laboratory techniques to measure simultaneously the activity of a large number of genes, proteins, or metabolites. The final result of applying FGT in the laboratory is to produce a measure for each of the thousands of molecules for a given biological sample. This result can be represented, independently of the specific FGT, by a data matrix whose

rows are represented by variables (genes, proteins, or metabolites) and columns are represented by samples (with their associated prognostic information). This data matrix can be analysed by methods described in Sections 2.3. The next sections introduce the most commonly used FGT.

2.2.1 - Transcriptomics

The most common application of FGT is in monitoring the gene expression (gene expression profiling). The technique is based on a classic molecular biology procedure called reverse northern blot. A schematic representation of this procedure is shown in Figure 2.2. mRNA is extracted from a biological sample and reverse transcribed in the presence of a radioactive or fluorescent precursor. The reaction produces a pool of labelled complementary DNA copies (cDNAs) representative of the original mRNA pool which is called here a target. The expression of an individual gene is quantified by hybridizing the target to the gene specific cDNA (defined as a probe) which has been previously spotted on a solid surface. The amount of radioactive or fluorescent signal associated to the spot is proportional to the amount of the target gene and hence to the specific RNA originally present in the cell. Multiple cDNAs can be spotted in an ordered pattern (array) allowing the quantification of multiple genes in single experiments (see reference [10] for a recent review of the technology and applications others than transcriptomics). There are, mainly, two types of microarrays, cDNA which are commonly assayed for two samples labelled with different dyes, and oligonucleotide microarrays where only one dye is used. Both technologies (and others) generate a unique measure for every probe (Figure 2.2).

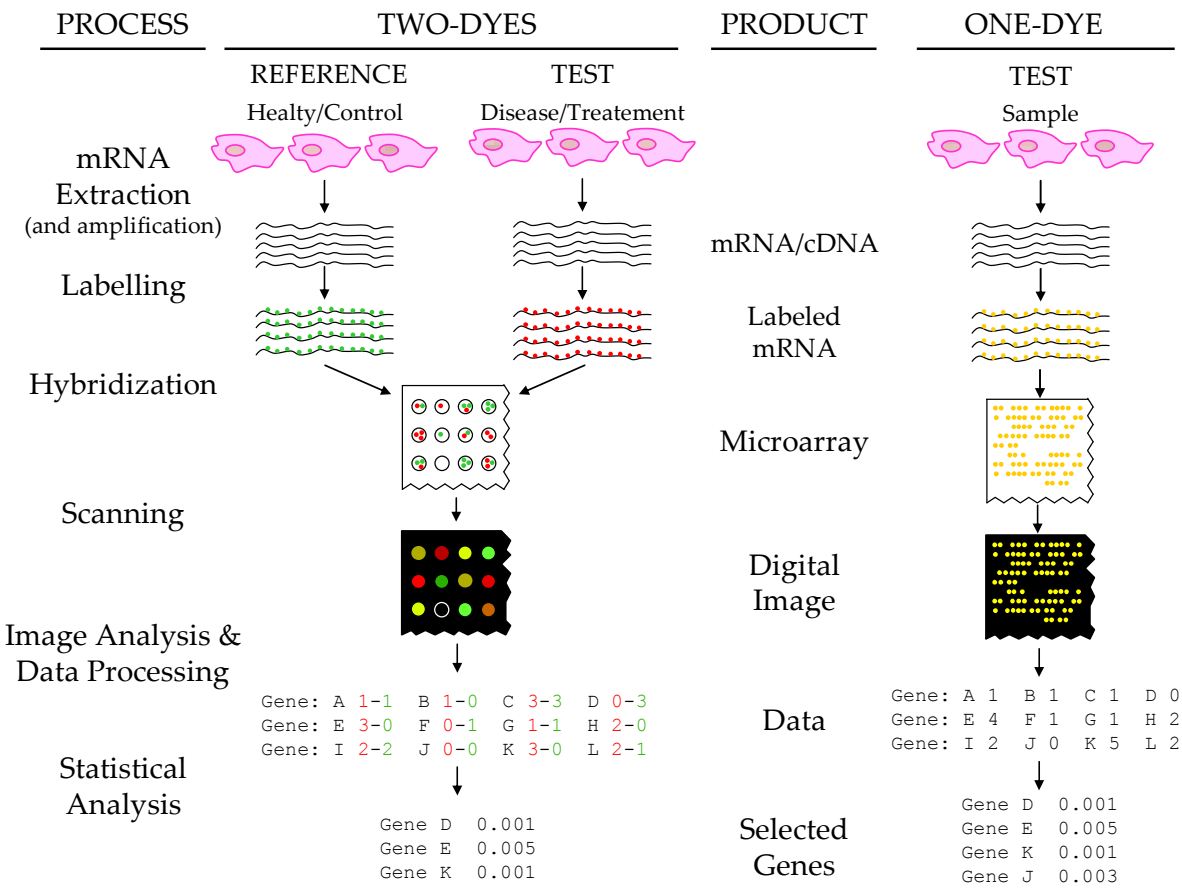


Figure 2.2 - Schematic representation of a gene expression microarray assay. Arrows represent process (left column) and pictures or text represent the product. Differences in the protocol in one- and two-dye technologies are specific to the technology rather than samples or question.

2.2.2 - Proteomics

Although at the moment expression profiling is the most commonly used FG application, there are other techniques that measure other components of the cell. Proteomics is one such technology. Proteins are the translated product of the processed mRNAs and are directly involved in translating the genetic information into function. Their precise measurements can therefore be potentially more informative of the physiological state of a cell than the mRNA levels. For practical reasons, proteomics is still not as widespread as expression profiling but is recently becoming a standard tool. Proteomics analyzes and identifies the proteins in cells, tissues, or organisms.

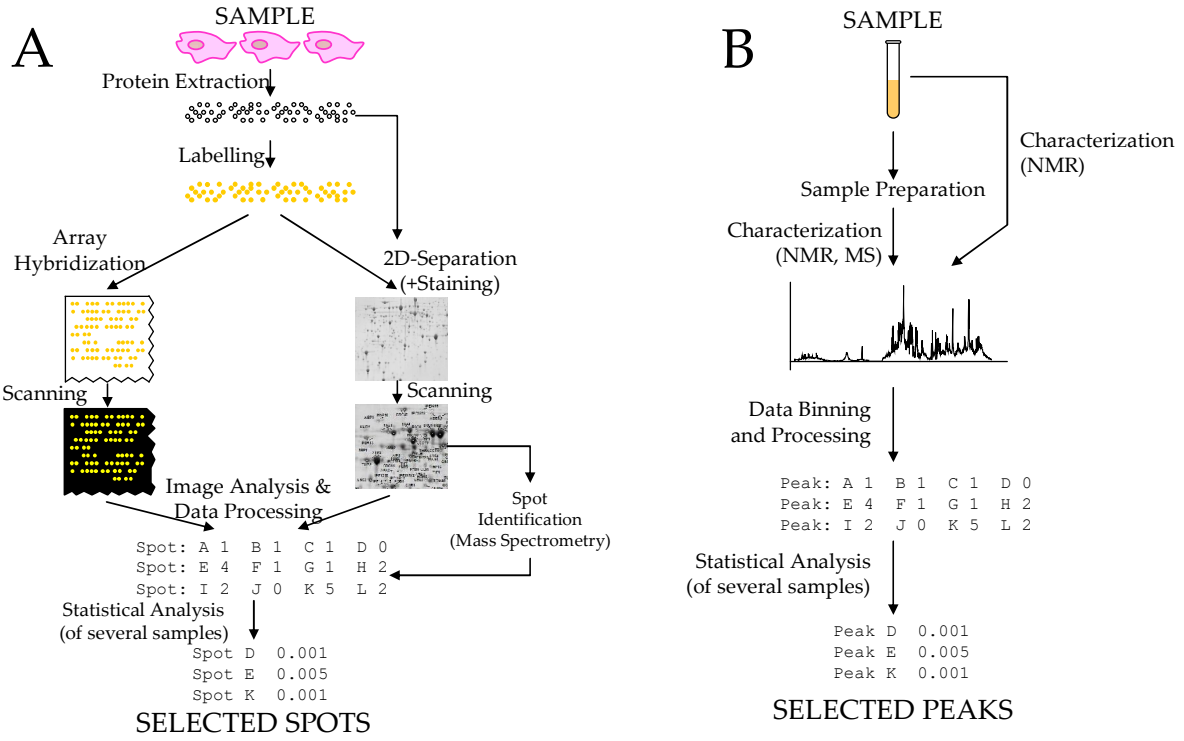


Figure 2.3 – Overview of Proteomics (A) and Metabolomics (B) assays. After data has been acquired, the data processing and statistical analysis is similar to DNA microarrays. Gel images were obtained from http://www.cogeme.abdn.ac.uk found by Google.

A simplified overview of a typical proteomic assay is shown in Figure 2.3A. The analysis may include characterization of physicochemical properties as amino acidic sequence and post-translational modifications and the description of their behaviour at function and expression level. Commonly, a 2-D gel electrophoresis is used as a separation and profiler method where the first separation is performed by isoelectric point (pI) then by weight [11]. The gel is characterised by a large number of spots representing, in many cases, a single protein. Scanned gel images are analysed to detect and quantify every spot. The resulting data is similar to a microarray dataset, hence similar dbata analysis methods can be applied. One of the problems is, however, that only a small fraction of the spots are annotated (with their corresponding gene or protein identification attached). Several methods are then applied to annotate or identify the proteins in a spot [12]. In the case of protein microarrays, the procedures needed for the data acquisition (scanning, image analysis, and normalization), and data

analysis are similar to those for DNA microarrays (although some exceptions may apply) [13]. For small-scale proteomics (a few dozens), classical manual laboratory techniques can be used, such as a series of ELISA assays, or multiplex Luminex assays [11; 14].

Independent of the proteomic technology, the dataset is a matrix whose rows represent proteins and columns represent samples. Thus, this dataset can be processed similar to a DNA microarray dataset.

2.2.3 - Metabolomics

Metabolomics is the set of techniques designed to measure the amount of metabolites present in a biological sample. Metabolites are usually small molecules or protein ligands. Metabolomics is then the study of low weight molecules (proteins and nucleic acids are therefore excluded). There are two levels of metabolite measurements, metabolic profiling whose aim is quantify a very limited number of metabolites, and metabolic fingerprinting whose goal is to provide a global screening [15]. Metabolic fingerprinting is commonly performed by nuclear magnetic resonance (NMR) or by mass spectrometry (MS). An overview of the process is depicted in Figure 2.3B. Sample preparation is critical in MS-based metabolomics [15] whereas it could be sometimes simpler in NMR-based metabolomics [16]. MS is a technique to measure ion mass-to-charge ratios. An ion source unpins and ionizes molecules in a sample. Molecules then fly under the influence of an electric and magnetic field. Ionized molecules are therefore separated by their mass-to-charge ratio. A spectrum is then collected. To identify specific molecules, their mass-to-charge ratio is compared to those produced by known molecules. NMR on the other hand uses the spin property of atoms with an odd number of protons or neutrons, such as ^{1}H, ^{13}C, ^{15}N, ^{19}F and ^{31}P. These nuclei possess an overall intrinsic magnetic moment and angular momentum. However, electrons also

alter the nuclear spin, hence molecules sharing electrons in different configurations will shift the frequency of the energy needed for resonance. This effect, known as *chemical shift*, is reported as a relative measure from a reference frequency. To measure ^{1}H or ^{13}C, the common reference used is that from tetramethylsilane. Both techniques (NMR and MS) generate a metabolic profile (spectra) representing the relative concentration (vertical axis in Figure 2.3B) of metabolites (horizontal axis in Figure 2.3B) in a sample. The profile is characterized by several peaks each of which would represent, mainly, a single metabolite. Data processing involves reducing noise and background, spectra alignments, peak annotation, removal of signals due by water or solvents, normalization, and binning [15; 16]. The final result is then a dataset of metabolites for a number of samples. Statistical analysis can proceed similarly to a DNA microarray dataset.

2.2.4 - Other "omics"

Recent generalizations of the above approaches have derived new terms such as lipidomics and glycomics (which perform large-scale studies of lipids and sugars respectively). It would not be a surprise that new terms are defined and re-defined in a few years[1].

2.3 - Data Processing

As any experimental device, FGT produce noisy data. The objective of data processing is removing the noise and systematic variability produced by devices and laboratory instruments and reveals, at a certain level, the genuine biological signal. Another

[1] See the site http://www.genomicglossaries.com/content/omes.asp

implicit goal is to generate data that is independent of the technology used. In addition, different FGT may need specific processing issues but a number of concepts described here can also be applied or adapted. This chapter describes the data processing principles using microarray technology as a reference technology making appropriate mention to the generality of the process when needed. The schematic representation of the procedure for array-based FGT is shown in Figure 2.4. The following paragraphs will introduce these processes.

2.3.1 - Image Analysis: Spot recognition and background subtraction

Image analysis is the process of converting the image generated from the microarray scanner to generate a measure of every spot. In this context, the number of pixels per

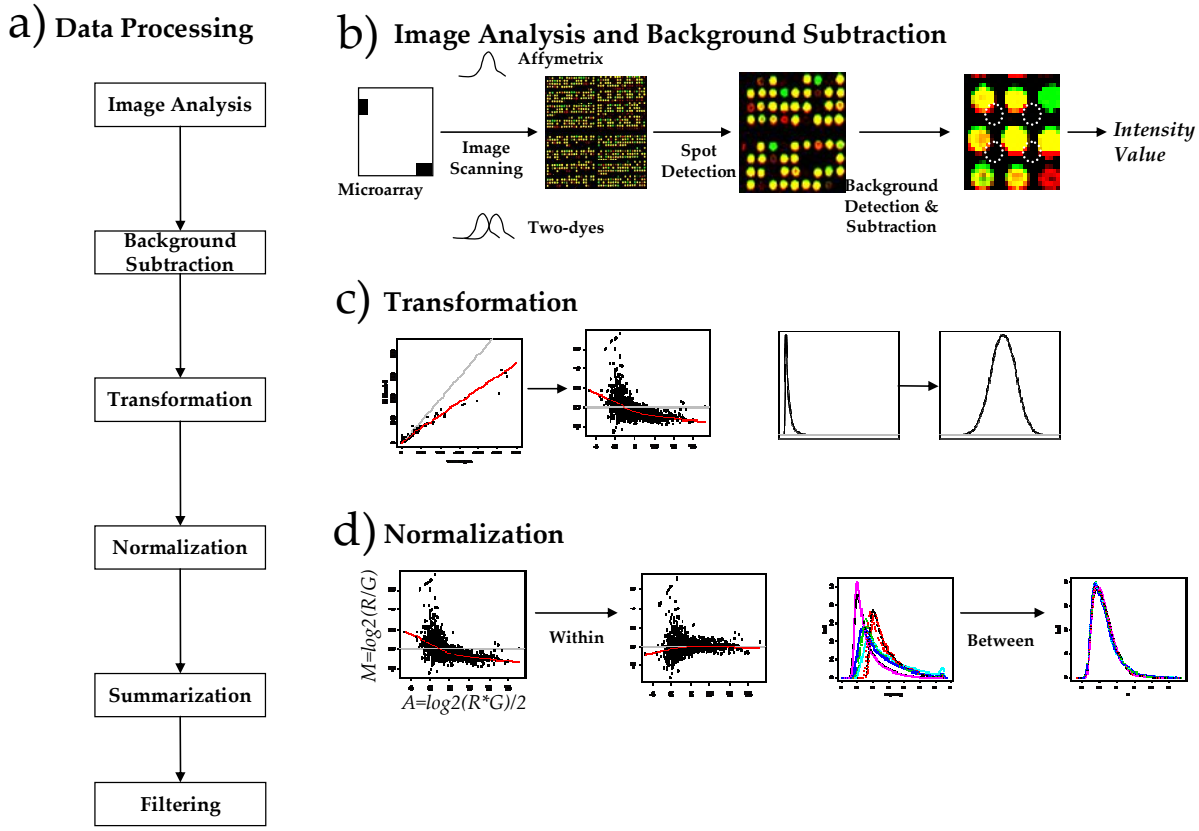


Figure 2.4 - General scheme for pre-processing procedures. (a) Pre-processing scheme. (b) Image analysis and background subtraction. (c) Log transformations. (d) Within and between Normalization.

spot, spot shape, and pixel intensity of every spot could vary significantly from experiments or slide quality (normally related to different stocks). This variation can be a systematic error or noise and is intended to be removed by the normalization process. The main sources of systematic error are the dye bias, labelling efficiency, and manual labours in the laboratory during RNA extraction and probe purification. Correction is generally algorithm-dependent [17; 18]. The process of image acquisition involves several steps, which are reviewed in Leung *et al.* [19]. Briefly, the process can be summarized in three steps: image acquisition (scanning), spot identification (segmentation), and signal quantification (see Figure 2.4b). Spot identification is an essential and a non trivial step because of the high density of spots located in the array. Specific algorithms for spot recognition are detailed elsewhere [20]. Once the spot is identified, individual pixels forming the spot can be quantified. The background can be estimated from areas outside the spot itself by different methods [21-26]. Finally, the spot can be converted to a single value. This is typically the median value of the pixel signal distribution. Slide manufacturers commonly provide their own scanner and software both tuned for optimal results. The values of the data are however different, in one-dye based experiments such as Affymetrix, values represent signal whereas two-colour based arrays represent a ratio of signals. Therefore, data processing and normalization may differ depending of the type of data available.

2.3.2 - Transformation

In one-dye microarrays, the data is heavily biased towards low signal intensity. Therefore, a transformation that distributes the values smoothly is used. For this, Logarithm base two (log_2) is preferred because it behaves reasonably well (Figure 2.4c), in addition an increment by one unit represents a double value which facilitates the interpretation into fold-changes. Other transformations such as trigonometric hyperbolic functions have been proposed whose advantage is that they deal with

negatives values [27]. For two-dye microarrays whose readout is a ratio between the two channels, a log_2 transformation is also used to represent fold changes. However, ratios are subject to noise when the denominator is close to zero.

2.3.3 - Normalization

Systematic errors are introduced in labelling, hybridization, and scanning procedures. Normalization is the process to correct for these systematic errors without removing or altering biological variation. There are two different types of normalization. These are: *within* slide normalization and *between* slide normalization (Figure 2.4d). *Within* normalization refers to normalization applied to the same slide and it is applicable, commonly, for two-dye technologies correcting for dye and spatial bias [28]. *Between* normalization is used when at least two slides are analyzed and it requires that both slides are measured on the same scale and that their values are independent of the device parameters used to generate such measurements. Several methods have been reported and compared for normalization [21; 23; 29; 30].

2.3.4 - Gene and Probeset Summarization

Summarization is the process of producing a single value for a gene that has been measured by several probes. This process is necessary since microarrays include several measures of the same transcript. For instance, Affymetrix arrays are based on oligonucleotide probesets which consist of around 20 oligonucleotides designed to represent individual transcripts. Other array technologies, which may rely on PCR generated DNA fragments spanning larger regions of the transcript, print the same fragment in duplicate or triplicate across the same slide. Therefore, methods that produce a unique measure for every gene that is the closest to a true gene expression value are used [22; 31]. However, this process may be optional.

2.3.5 - Filtering

Genome-wide microarray experiments generate large amounts of data that is influenced by noise. Several algorithms and statistical approaches are sensitive to the quality of data, thus, low-quality data could lead to wrong conclusions. On the other hand, computational resources are finite, so large amounts of data may lead to futile processing cycles and confusing results. Filtering then removes data due to bad quality, uninformative, reliability, internal assay controls, proximity to background, and manual marking [28; 32; 33]. Consequently filtering reduces the number of variables to analyse.

2.4 - Understanding Genome Wide Data is a High Dimensional Problem

Although the data generated using different FGT are acquired in a different format, their analysis has common issues. One of the most important issues is a consequence of the extremely large number of variables measured in an experiment [34; 35]. The most obvious problem linked to the number of genes being measured is that the traditional *p-value* is, in fact, not reliable in all cases when tens of thousands of variables are assessed. This problem is commonly called a *multiplicity test.* A number of methods have been recently proposed to address this issue. Some of the proposed approaches are based on the application of an a-posteriori correction [36; 37]. Other approaches are based on the application of an error model estimated from experimental data [32; 38]. In both cases there is an attempt to estimate the false discovery rate (FDR). Another important issue is that statistical tests, such as the t-test, assume that all variables are independent of each other. This assumption is clearly not true in a biological system where genes

display a strong correlation with several others (for example, because they share a common activator). Multivariate approaches, which test combinations of variables at the time, seem to be more appropriate in this context[1]. However, the specific variables involved and the optimal number of them to consider together is a parameter that needs to be determined using objective criteria. In addition, it is not feasible to compute every single combination of sets of variables for very large datasets.

2.5 - Computational Methods for the Analysis of Microarray Data

A common task is to identify genes differentially expressed between two or more experimental conditions. A number of statistical tests have been used for this purpose; some compare means (t-test) or variances (f-test). Other methods are adaptations or non-parametric versions of these two approaches. Recently approaches that are more tailored to the analysis of microarray data have been proposed (for a review of these methods see Speed [39]). More advanced statistical approaches are, however, required to predict cell physiology from its molecular signatures. In its simplest formulation this task can be considered a classification or a regression problem. Because of the large number of genes involved, an efficient VS method must be part of the analysis strategy. Machine learning approaches are particularly suitable for this task. The further sections will introduce several of the approaches that have already been applied to relate molecular profiles to cell physiology.

[1] Even though sometimes univariate methods produce models displaying similar accuracy and number of genes to those models generated by multivariate methods. See Chapter 3.

2.5.1 - Detection of Differential Expressed Genes

The most common and basic question in the analysis of FGT data is whether variables (genes, proteins, metabolites) appear to be down-regulated or up-regulated between two or more classes of samples. For this, a statistical test is used to test the hypothesis that a gene is not differentially expressed (Figure 2.5). To detect differentially expressed genes, intuitive and formal statistical approaches have been proposed. The most used intuitive approach proposed in early microarray studies is fold change that is defined as the logarithm base 2 of the ratio between the expression value of the sample divided by the reference [40; 41]. Genes whose fold change is larger than certain (arbitrary) value, are selected for further analyses. Although fold change is a very useful measure, the weaknesses of this criterion are the overestimation for genes expressed at low level in the reference, the value that determines a "significant" change is subjective, and the tendency to omit small but significant changes in gene expression levels. From the statistical approaches, the common t-test is the easiest option, though not the best [42],

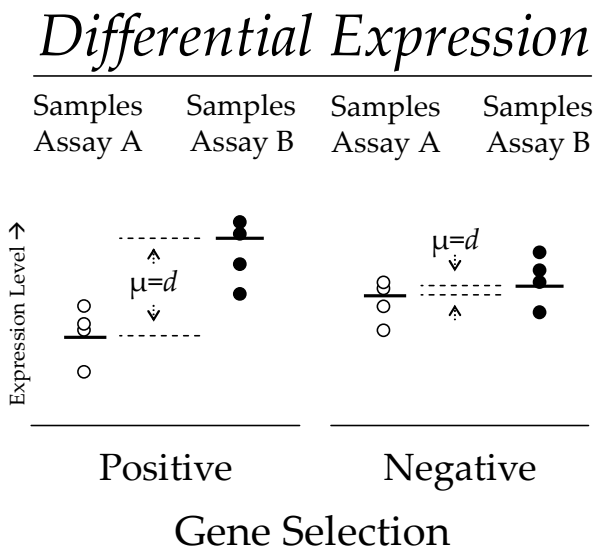


Figure 2.5 – Detection of Differential Expressed Genes. Large differences in gene expression are likely to be genuine differences between two groups of samples (A and B) whereas small differences are unlikely to be truly differences. Samples can be biological replicates or unreplicated populational samples.

for two groups of samples whose data follows a parametric distribution. The equivalent method for more than two groups of samples is the analysis of variance (ANOVA). These options apply for both one- and two-dye microarrays. If the data is non parametric, Wilcoxon or Mann-Whitney tests may be applied. Although these classical statistical tests are useful, other tests more suitable to microarray data have been proposed. For example, SAM is one of the fundamental methods in the context of relating genes to physiology [43]. Kim *et al.* [42] provides a comparison of differential expression statistical tests, including SAM [43], B-Statistic [44], samroc [45], Zhao-Pan [46], Bayes T-Statistics [47], MMM [48], and fold change.

Most of the methods to detect differential expression are based on the assumption that the majority of samples display similar expression values for a set of samples in a given class. However, the criteria of group samples in a class are commonly based on certain physiological, morphological, or molecular peculiarities. These criteria nevertheless may mask sample sub-classes that are not easy to spot. To address this issue[1], Tomlins *et al.* developed *cancer outlier profile analysis* (COPA). This method has been designed to detect "outlier genes" which show an increase in expression only in a subset of cancer samples [49]. COPA procedure is based on a non-parametric version of a standardization procedure which was further generalized [50].

2.5.2 - Unsupervised Classification: Describing the Relationship between the Molecular State of Biological Samples

One key issue in the analysis of microarray data is finding genes with a similar expression profile across a number of samples. Co-expressed genes have the potential

[1] Another way to overcome this limitation is removing the assumption of any class grouping which is seen as an unsupervised method. These methods are revised in the next section.

to be regulated by the same transcriptional factors or to have similar functions (for example, belonging to the same metabolic or signalling pathways). The detection of co-expressed genes may therefore reveal potential clinical targets, genes with similar biological functions, or expose novel biological connections between genes. On the other hand, description of the degree of similarity of biological samples at the transcriptional level may be desired [51]. It is expected that such analysis confirms that samples with similar biological properties tend to have a similar molecular profile. Although this is true it has also been demonstrated that the molecular profile of samples is also reflecting disease heterogeneity and therefore it is useful in discovering novel disease sub-classes [52]. From the methodological prospective, these questions can be addressed using unsupervised clustering methods. Common unsupervised methods applied to FGD (some revised in [53]) are Hierarchical Clustering [51; 52; 54], Principal Component Analysis [55-58], and Self Organized Maps [59]. These and other unsupervised methods are available in several software packages such as R (http://cran.r-project.org), GEPAS [60], TIGR T4 [61], [54], GeneSpring [62] and Genesis [63].

2.5.3 - Supervised Classification: Predicting the Relationship between the Molecular State of Biological Samples

Supervised classification is the process of predicting sample categories once their representative patterns have been learned from data (Figure 2.6). In order to forecast the class, a rule is needed that relates data (variables) to the class. There are a number of methods that generate this rule. These methods are referred in statistics and machine learning as *classifiers*. Every classifier makes specific assumptions about the data and uses a model or algorithm to yield the rule. The number of guesses that were predicted by the classifier can be high or low depending on how good the classifier learned the data. Therefore, a robust statistical procedure referred as *error estimation procedure* is

used to evaluate the accuracy of classifiers. In addition, error should be general. That is, it should be approximately the same independently of the subset of samples used for its estimation. A concept called *overfitting* refers to the event where a model may describe well trained data but cannot describe test data that was not used for training. In this context, it is generally accepted that models with a high number of variables tend to over-fit the data. Simon argues:

> "*Complex models have so many parameters that they can fit all of the random variations in the training data well. They find predictors and nonlinear functions that account for the random variations in the training data, but these discovered relationships do not represent real effects that exist in independent data; consequently, predictive ability is poor.*" [64; 65].

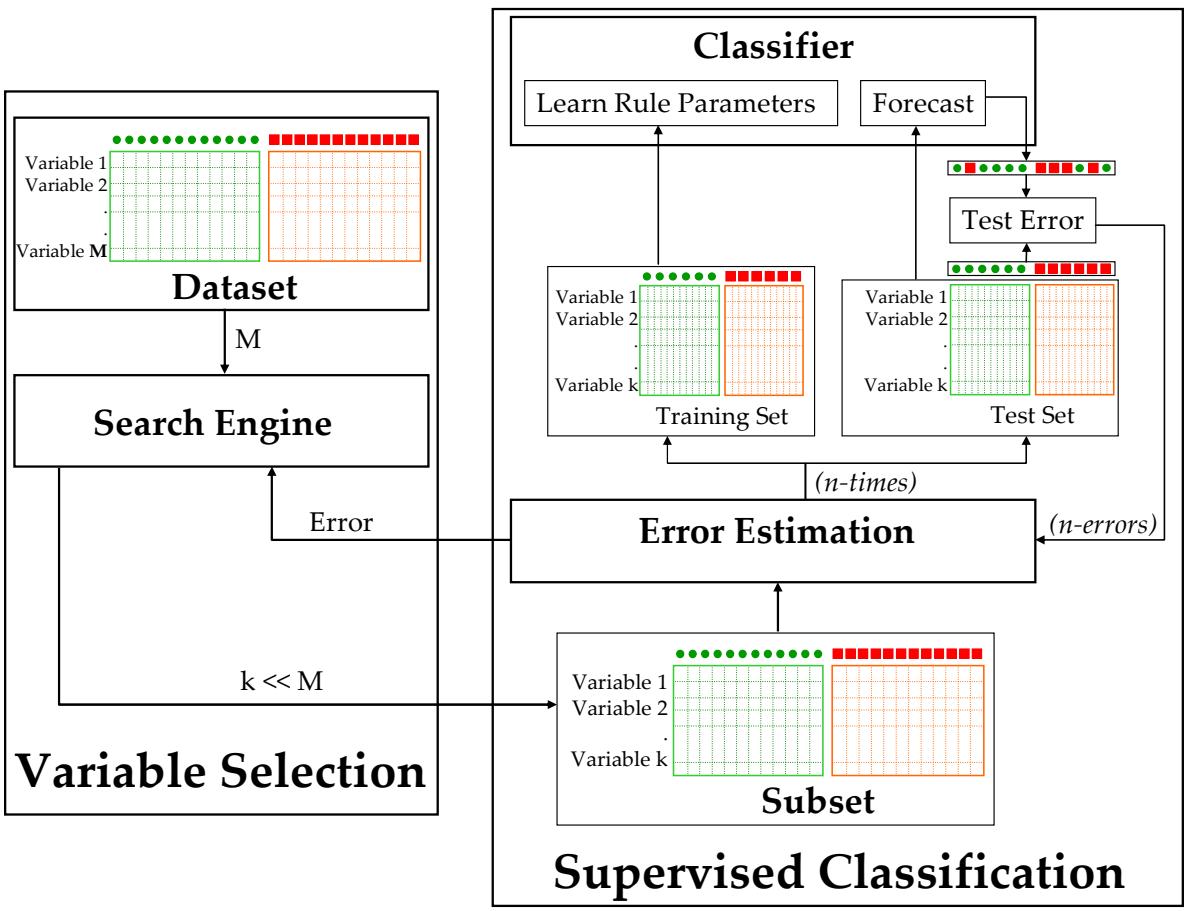


Figure 2.6 –Supervised Classification and Variable Selection.

Thus, within a rule, some variables would be more important than others, and some variables could be not needed at all. Therefore, supervised classification is commonly linked to the identification of signatures (genes, proteins, or metabolites) that are genuinely useful for the prediction.

In summary, supervised classification involves classifiers and error estimation procedures that are connected to VS strategies (Figure 2.6, for VS methods, see next section). Classifiers are algorithms or mathematical models that are able to make categorical predictions. Classifiers are generally multivariate, that is, several variables are considered in the model. A number of classifiers that have been used in the context of FGD [66-72]. Some of the most used are Nearest and Shrunken Centroids [73], K-Nearest-Neighbours [74], Classification Trees and Random Forest [75; 76], Discriminant Analysis [77-79], Neural Networks [80-82], and Support Vector Machines [83; 84].

Error Estimation is the process to determine the error in prediction made by classifiers. The true error made by a classifier in a population would be the proportion of wrong predictions from the total number of samples in the population. In practice, however, the number of samples is limited. Thus, a procedure to approximate the error has to be used. In the context of FG, the samples are scarce and the procedure to estimate error should be careful selected. A robust error estimation procedure also serves to compare the performance of different classifiers. Although there are classifiers that allow direct error estimation (at least on training data), in general, the procedure to estimate the error is performed outside the classifier. The procedure is based on defining two sets of samples. One used to train the classifier whereas the other is used to test the classifier comparing the class prediction made by the classifier against their original classes (see Figure 2.6). The differences among error estimation procedures are on how the training and testing sets are defined and how the error is mathematically determined. The most common error estimation procedures are *resubstitution* which employs the same

samples used for training as the testing cases, *cross-validation* [74; 75; 79; 85] that use a subset of samples for training and the rest for testing in such a way that all samples are part of the testing set only once, *bootstrap* [86-88] and *out of bag* [75] which use random sample subsets, and *bolstered estimation* [88] that consider the amount of the error in prediction. For a comprehensive comparison of error estimation strategies used by published work revised here, see Table 2.1.

For heuristic multivariate model construction that is designed in two steps, two error estimations may be required: (*i*) inside, in the VS procedure, and (*ii*) outside, in the final designed model. Wessels *et al.* propose averaging 100 times a 3-fold-CV for test error and 10-fold-CV for train error [89]. This research thesis has, also, followed a similar strategy (see Chapters 3, 5, and 6).

2.5.4 - Regression

In regression, likewise in supervised classification, the objective is the prediction of values of a dependent variable given the values of independent variables. The main difference between regression and supervised classification is that regression attempts to predict a continuous variable whereas supervised classification tries to predict a categorical variable such as class. Strictly, supervised classification is a special case of regression where the regressive variable is categorical. For example, the *logit* and *probit* are commonly used as classifiers however both are regression models that predict quasi-categorical variables. Therefore the issues detailed in the Supervised Classification section such as training and testing the model and error estimation procedures also applies to regression (Figure 2.6).

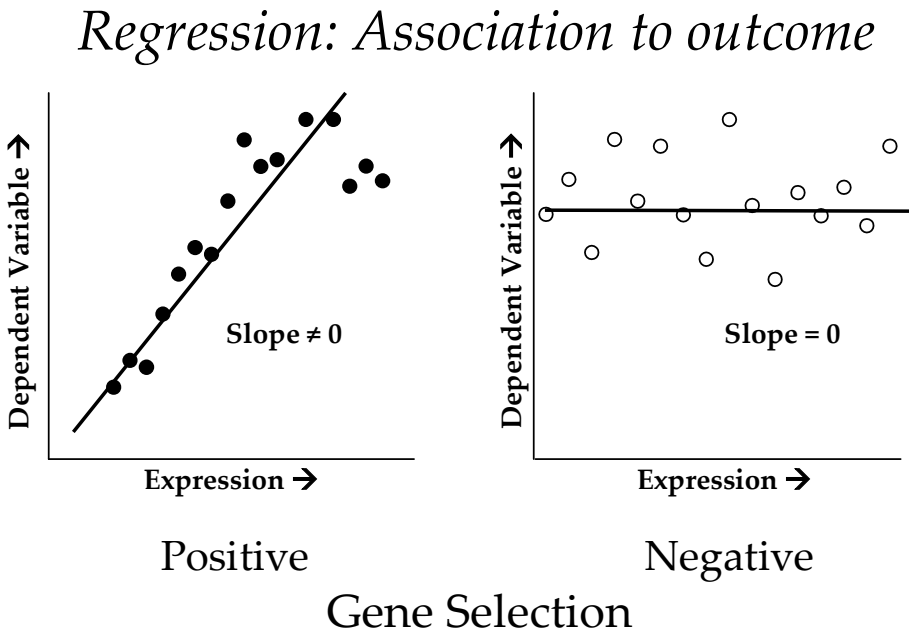


Figure 2.7 - Selection procedure for genes associated to outcome. The expression of a positive gene (horizontal axis in left plot) is highly correlated with the associated outcome (vertical axis). For a non-associated gene (right plot), the gene expression (horizontal axis) is not correlated to outcome (vertical axis).

Examples of the independent variables are the concentration of metabolites, proteins, response to treatment, growth, clinical outcome, or any other molecular, morphological, or physiological measure whose numerical representation makes sense progressively. The mathematical model that relates the independent variable to dependent variables (genes, proteins, or metabolites) is, commonly, a linear model (Figure 2.7). When the response variable is not linear, a transformation is used to make it linear if possible otherwise a non-linear model has to be employed. The work done by Antonov *et al.* [90] can be seen as a linear regression model where genes correlated to classes were successfully selected.

A special case of regression to predict survival times where the data is censored is known as *Survival Analysis*. In rigorous terms, the objective is to infer the distribution of survival times in a population of diseased people from the survival data acquired in the course of a clinical study. Since the study has a finite duration, the correct survival time

can only be measured in the case that the patient is deceased in the course of the study. If the patient is not deceased in the course of the study it is defined as *censored*. This issue complicates the estimation of a survival function. In this context, a commonly used method to estimate the survival function is the *Kaplan-Meier estimator* or *Product Limit Estimator*. Kaplan-Meier plot is a visual technique commonly used in clinics to evaluate the distinction of survival times in different populations. An example of a Kaplan-Meier plot and variable selection related to survival times is shown in Figure 2.8. To find the variables related to survival curves, a Cox Proportional Hazard regression model is commonly used [91].

Using a univariate Cox model, Lapointe *et al.* have related genes to discover novel subtypes of prostate cancer [9]. Multivariate methods have been also proposed [92].

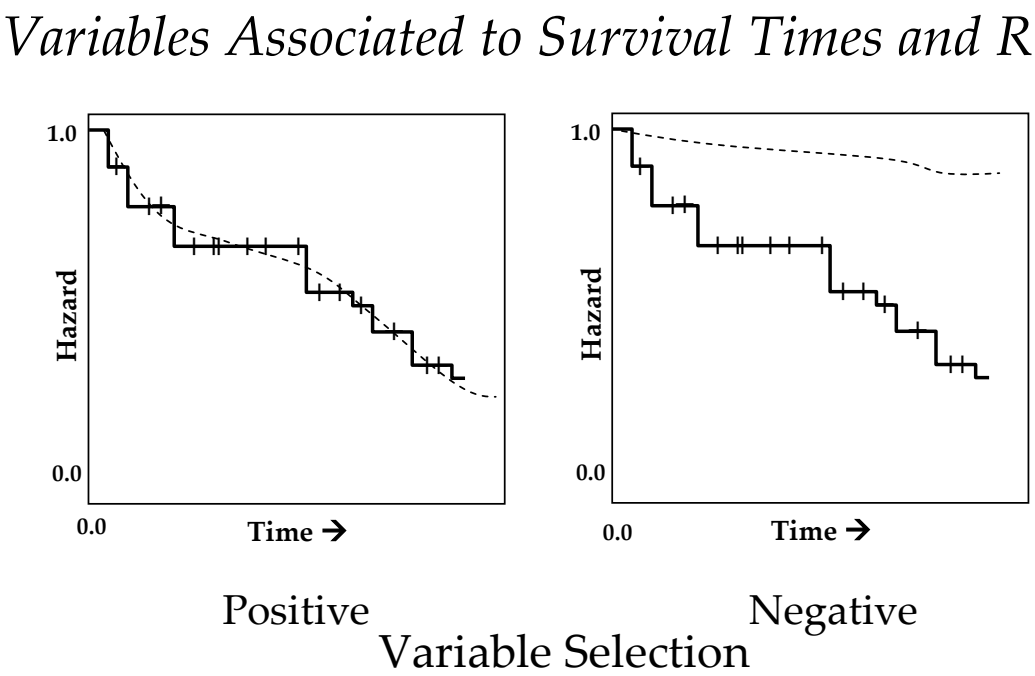


Figure 2.8 – Variable selection for Survival Times. When a gene included in a survival or hazard model (e.g. Cox model), if the resulted fitted curve fits reasonable well with the original survival or hazard curve, that gene is then considered as related to survival therefore selected (left plot) otherwise it is not selected (right plot).

2.6 - Variable Selection Methods

Variable selection methods (VSM) are procedures to select those variables more suitable for a given problem. This is important in the context of FG where the number of variables (genes, proteins, or metabolites) is huge. Although classifiers are commonly multivariate and could deal with such number of variables, their accuracy is drastically decreased by the high content of uninformative and noisy data. This happens because classifiers (or regression models) tend to use all variables and can not determine efficiently which variables are more suitable for the prediction. Therefore, VSM are coupled to predictive models to select those variables more appropriate for the prediction. VSM can be categorized as univariate and multivariate depending on how many variables are assessed at the time. Such methods are described next. A comparison of the results applying multivariate and univariate methods for FGD is detailed in Chapter 3.

2.6.1 - Univariate Variable Selection Approaches

Univariate variable selection (UVS) approaches are inspired in the simplest and commonest approach to distinguish two populations: comparing their central value (mean or median). Nevertheless, other properties such as variance are also used. In this view, only one variable is considered at a time and the test is somehow blind to the remaining set. Thus, a statistical test is performed for each variable and the selection is simply the collection of variables that are significant. For example, if the problem is a supervised classification, the VS procedure could be simply the selection of those variables that are significantly different among classes; if the problem is a regression model, the selection could be those variables that are significantly correlated with the outcome. At this stage, no predictor (classifier or regressive model) has been built. Thus, univariate approaches are based on the assumption that variables that are, by

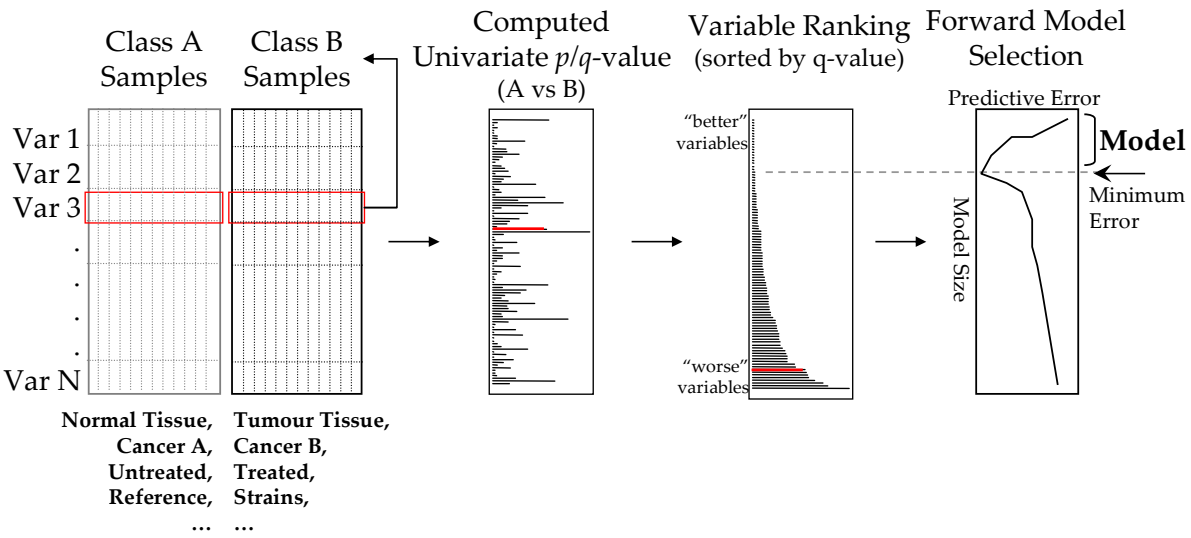


Figure 2.9 - Univariate Variable Selection. A statistical test is used to relate genes to classes. The associated *p-values* are sorted and fed into a forward selection procedure that selects the top genes depending upon the minimum error. Error is estimated by an error estimation procedure coupled with a classifier (see Figure 2.6).

themselves, related to the problem (classification or regression) could be part of good predictors. Given that the univariate selection is "decoupled" to the predictor, the major drawback is that the building of a predictive model is uncertain. Nevertheless, in practice, univariate VSM are commonly successful. In addition, they have the advantage that they are simple, fast, and have none or a few parameters to optimize. To build the predictor, a forward selection procedure is commonly used. The overall process is depicted in Figure 2.9.

Some of the most used univariate selection methods are Golub's centroids [7], Shrunken Centroids [73], t- and F-Tests [93], Wavelets [94] (see [95-97] for details), and other classical statistical tests. In general, any method that estimates differential expressed genes could be easily adapted as a UVS approach. In this context, a recent work has

compared[1] 10 parametric and non-parametric differential expressed methods [98].

2.6.2 - Multivariate Variable Selection Approaches

The UVS approaches, described in previous sections, select variables in a univariate manner to build a multivariate predictor. Multivariate Variable Selection (MVS) techniques on the other hand are also used in combination with multivariate classification or regression methods. However in this approach, variables are selected by testing several variables in combination. MVS approaches are generally better that the univariate counterpart (see Chapter 3). The main reason is that univariate approaches cannot detect differences when a variable displays similar distributions between classes whilst a set of variables with distributions alike across classes can be good discriminators using multivariate approaches (Figure 2.10). Another reason why multivariate approaches surpass the univariate ones is that only one variable could not explain completely the behaviour of the sample class. So, perhaps another variable could complement this case. Thus, multivariate methods can take advantage of combinations of variables.

[1] Methods compared were SAM, ANOVA, empirical bayes t-statistic, template matching, maxT, between group analysis (BGA), area under the receiver operating characteristic (ROC) curve, the Welch t-statistic, fold change, rank products, and randomly selected genes. A brief description of these methods is provided in the original paper.

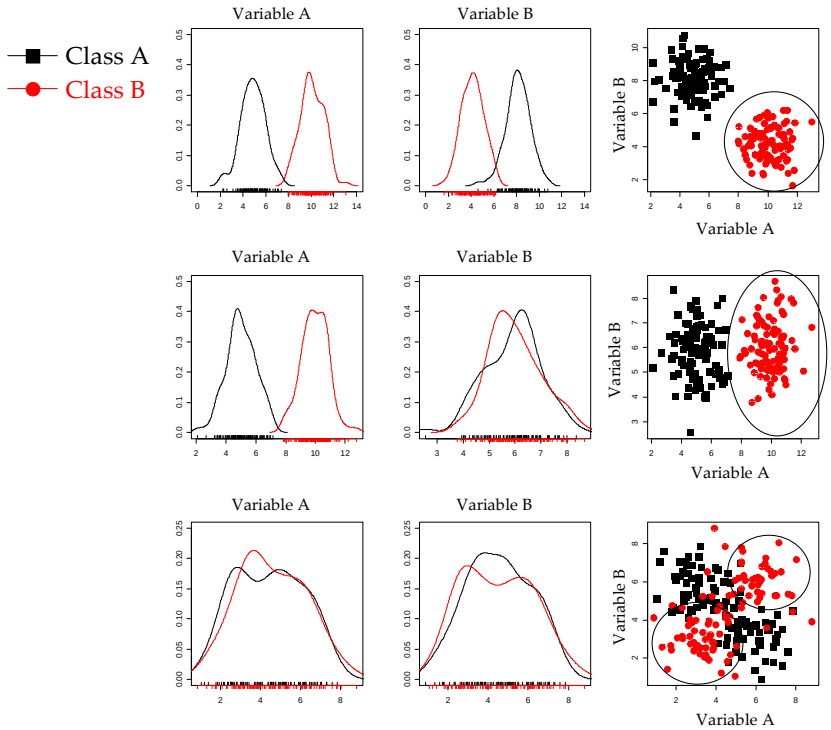


Figure 2.10 – Combinations in the distribution of two variables. The distributions of variables among classes are shown in left and central panels as seen in univariate methods, sometimes well separated and others almost identical. Variables with almost identical distributions cannot discriminate between classes using univariate methods. Alternatively, multivariate methods could distinguish classes when at least one variable is a good discriminator (right-top and right-middle panels), but in addition, under certain conditions, multivariate methods could also distinguish classes reasonable well when variables are distributed evenly (right-bottom panel). Right panels are a simplistic multivariate representation. Circles enclose samples of a target class.

The problem is, nevertheless, to search for combinations of variables that together are good predictors with a minimum of error under the realistic use of resources and time. An oversimplified generic overview of the MVS process is sketched in Figure 2.11. In this figure, the process after the MVS can be also viewed as a second MVS; nevertheless, the depicted process is closer to the trend in the literature. Unlike UVS approaches, the generation of models can be the result of the MVS or from a further model selection. In general, the process is similar to the univariate ones replacing the univariate test by the MVS.

In the analysis of FG data, several search strategies have been tested. Some of them are general search engines independent of the predictor such as Genetic Algorithms [74; 79; 99-101] (see [102; 103] for details), Random Walk [104], Forward Search [105], Exhaustive Search [106-108], Backward Elimination (see Chapter 3), Estimation of Distribution Algorithms [109], and Markov Chain Monte Carlo [110] (see [87; 111; 112] for details) while some others, are specific search strategies coupled with a specific predictor such as back-propagation in Neural Networks [59; 113; 114], random search with classification trees in Random Forest [75; 76; 115], nested coefficients in Wavelets, loadings in Principal Component Analysis [56; 58; 116] and Independent Component Analysis [117-121], Recursive Feature Elimination [122], Ensembles [123], and Ideal Features [90] for support vector machines, and sub-networks in correlation-like matrices [124].

In the following, Genetic Algorithms will be described because an implementation of

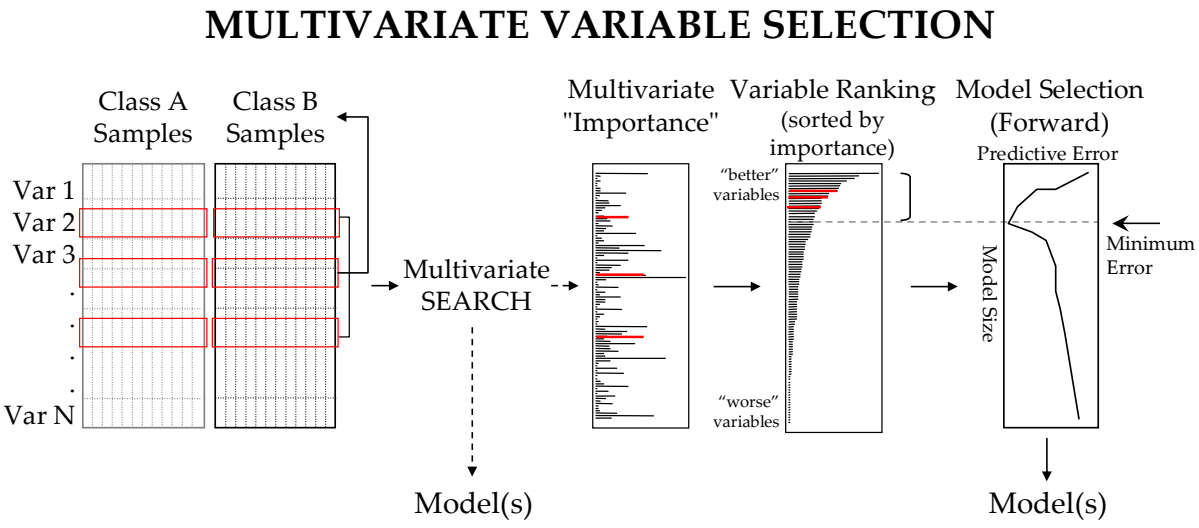


Figure 2.11 – Multivariate Variable Selection. A simplistic overview of MVS approaches. Models can be designed, at least, in two manners: directly from the multivariate search or from a further process after variable ranking or pre-filtering using perhaps a forward selection procedure or any other MVS approach.

this method is presented in Chapter 3.

2.6.2.1 - Genetic Algorithms

Genetic Algorithms (GA) is a general search strategy created by John Holland since 1970 [102] and have been popularized by Goldberg [103]. The name was inspired by how genomes evolve in a living organism. GA is popular because it can be adapted to specific problems relatively easily.

The schematic representation and flow of a GA is shown in Figure 2.12. The initial concept is the creation of an artificial string called a *chromosome*. Imitating life in organisms, a chromosome in a GA is created by artificial genes where an artificial gene represents a parameter of the model of interest. In our case, an artificial gene could be

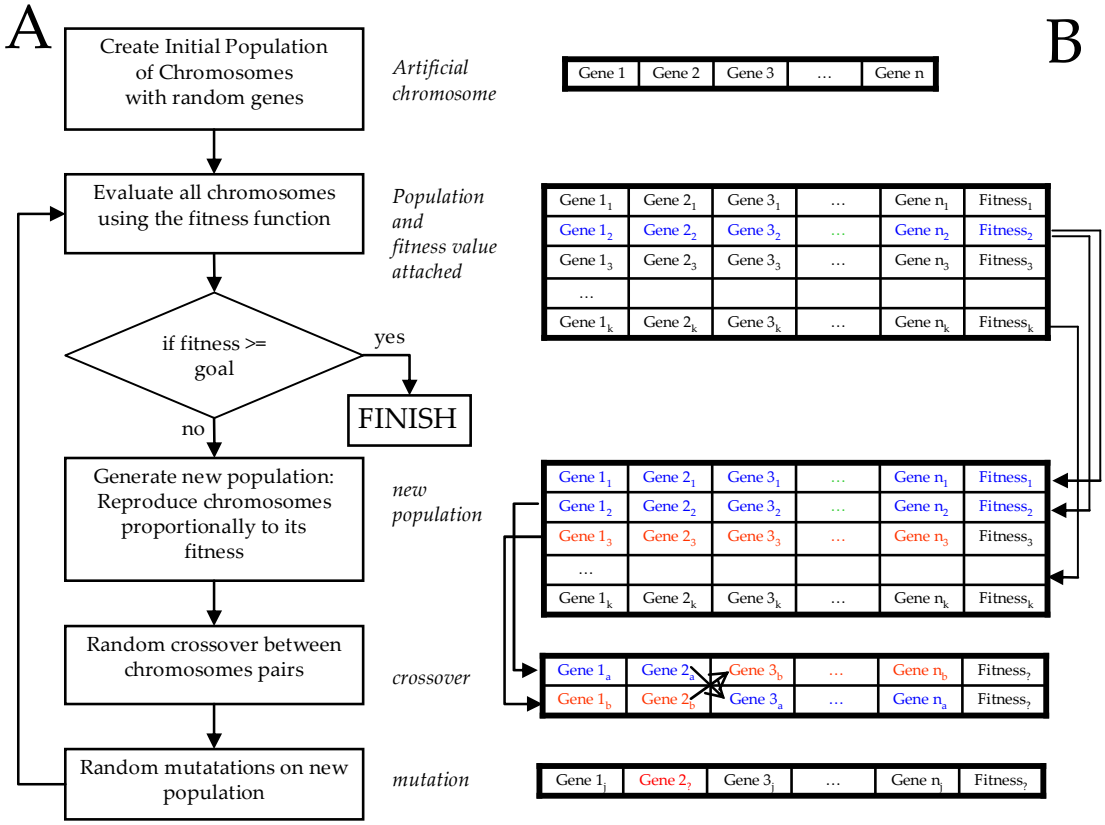


Figure 2.12 - Schematic representation of Genetic Algorithms (GA). (A) Flow chart. (B) Operation over GA chromosomes.

any gene of the set of genes measured in the microarray or any variable from any other FG data. The second concept, called *fitness*, is a function that evaluates the GA chromosome and assigns a measure on how good this chromosome is relative to a goal (mimicking nature). In classification problems, the function could be the number of correctly classified samples divided by the number of total samples. The procedure starts by creating an initial population of random chromosomes. Subsequently the fitness function is evaluated for every chromosome in the population. A new generation of chromosomes is generated by picking chromosomes randomly with a probability proportional to the fitness obtained for every chromosome (like natural selection). Therefore, the new generations will be dominated by those chromosomes that were highly evaluated in the fitness function. The new generation of chromosomes is then mated using the concept of crossover and eventually mutated at random with a specific probability. The mutation in our case is selecting a different gene or variable. Finally, the new generation is evaluated again and the process is repeated until a fixed number of generations or a chromosome reaches a goal. The process from the creation of new chromosomes until a solution is reached is called evolution.

The objective of search strategies, such as GA, is to find the combination of parameters (genes) that produce a maximum for the fitness function. This combination is known as global maximum. In GA in general, like in any stochastic search strategy, there is uncertainty that a global maximum was achieved because the procedure is dependent on a random processes (initial random population, random mutation, and random crossover). However, the solutions found are presumed to be close to this optimal point and sometimes are called near-optimal solutions. In order to be more confident that the selected parameters in an evolution of a GA search obey a near-optimal solution, sometimes independent runs are performed and different solutions are analyzed.

The work of Li *et al.* [74] was the first proposal of GA for gene selection from microarray data (because it is a seminal paper for the work presented here, a higher level of detail will be provided and will be referred as GA/KNN). They used a GA as the VSM and K-Nearest-Neighbour (KNN) as a classifier. The fitness function implemented is a measure of how many samples were classified correctly in a training set using a leave-one-out cross validation (LOOCV) method:

$$f = \sum_{i=1}^{N} C(i)$$

$$C(i) = \begin{Bmatrix} 1 & pc_i = class_i \\ 0 & otherwise \end{Bmatrix}$$

where $C(i)$ is the class prediction of the chromosome being evaluated for sample i, pc_i is the predicted class for sample i considering the distance to any other sample $j \neq i$ (leave-one-out), $class_i$ is the original class for sample i.

A common variant in GA known as niche [103] was used in GA/KNN. The use of elitism forces the best chromosome of every generation to pass, as is, to the next generation which is finally used to avoid bad movements. However, excessive elitism cycles could lead to populations trapped in attractors which may result in no evolution of their chromosomes. GA/KNN used 10 niches evolved in parallel with an elitism scheme. The probability of mutation was 1 (always mutate) and the number of mutations varied randomly from 1 to 5, except in chromosome lengths less than 11 where only 1 mutation was allowed. In Li *et al.* GA/KNN, one solution is the result of one evolution, that is, a run of a GA algorithm until a fitness function is greater than a threshold or the number or maximum generations is reached. Only solutions greater than the threshold of 31/34 were considered. After a number of solutions have been polled, the frequency for a particular gene present in every solution is assumed to be the relative importance of that gene for KNN classification. Then, the top t most important genes are used as predictors. t was decided by plotting a graph for the error

in classification versus *t* in a forward selection manner (Figure 2.11). One of their results suggested that the higher the number of solutions collected the more stable the resulting frequency. They recommended polling solutions until a plot of the ranks for independent runs are close to a straight line. The chromosome size was fixed. They tested chromosome sizes of 5, 10, 20, 30, 40, and 50 concluding that larger chromosome sizes stabilize gene ranks and therefore produce better reproducibility. Thus, they suggested running their method on several chromosome lengths from 20 to 50. They also showed that for small chromosome lengths the genes selected are dominated by a few genes (attractor genes) whereas other genes appear when size is increased. Li *et al.* successfully tested and selected genes that distinguish between classes for lymphoma [125] and colon data [52].

Although the work of Li *et al*. is a milestone for the work presented in this thesis, their implementation shows some disadvantages. First, they used all available data in the selection procedure. This leads to a misinterpretation of the comparison of independent runs. It has been seen that gene selection, for a limited number of samples, is somehow dependent on the subset of samples [126]. Therefore, in comparing independent runs using all data, they are comparing random effects within the GA procedure rather than the selected variables. Second, they suggested using large chromosome sizes because it stabilizes the gene-frequency-ranks between different sizes. Nevertheless, it will be shown that test accuracy decreases while increasing chromosome size (see Chapter 5). Thus, increasing chromosome size leads, unfortunately, to overfitting. These problems were apparently derived by the combination of using all the data and LOOCV. By using both, there are no degrees of freedom in the selection and error estimation procedures between one run and another, except by the GA randomness.

Instead of using a non-parametric KNN classifier, Ooi *et al.* have used MLHD discriminant functions [79]. In that work, the fitness function was based on subtracting

the sum of error rates in training and test sets as $f(S_i) = 2 - (E_C + E_I)$ where E_C is the LOOCV error rate in the training set and E_I is the error rate in the test set. Error rate was computed fitting the parameters for the training set and evaluating the function for every test sample to get the fraction of misclassified samples. They speculated that relatively high crossover rates ($p_c >= 0.8$) together with high mutation rates ($p_m >= 0.002$) are more likely to produce a good model. The termination criterion considers only the number of generations which was fixed to 120 and the best chromosome evaluated in all generations was retained. Instead of limiting the chromosome length to a particular value, they implemented variable length approach using the first artificial gene as chromosome length (Figure 2.13). A mutation in the chromosome length (position 0) produces a change in the effective chromosome size.

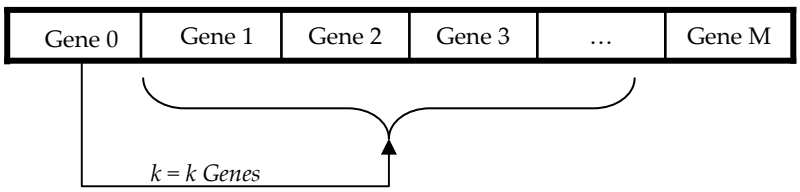


Figure 2.13 - Schematic representation of variable length chromosome. The maximum length, in genes, of the classification model is *M*. However, the active model consists of *k* which is encoded as the first gene *"0"*.

Ooi *et al.* showed that running their GA/MLHD algorithm several times produced very similar chromosomes. Therefore, unlike Li *et al.*, Ooi *et al.* do not run the algorithm several times to pool chromosomes and select by frequency. They argue for doing so different models are built in a further procedure which has an effect on the complexity present in multiple classes. Finally, they think that MLHD is a better classifier for multiple classes against KNN because the distance metric used in KNN are invariably less sensitive as data dimensionality increases [127] and because the GA/KNN does not use the resulting near-optimal chromosomes.

Likewise in the Li *et al.*, one of the disadvantages in the Ooi *et al.* implementation is the use of all samples in the selection procedure. Although they included training and test sets, they used both sets in the fitness function. Thus they biased the search for those models that perform well in the test set which may explain why they found similar models in different runs. It would be expected that by using variable sized chromosomes may lead in a large number of generations needed to optimize for this additional parameter, surprisingly they reported 120 as an upper-limit which seems too low. Their variable size implementation may display undesired properties because genes are mutated randomly. Thus, previously added genes are less likely to be removed than new profitable genes.

Other successful uses of GA include the work of Fröhlich who linked GA to SVM as classifier [99] where the GA search contains the gene subset and specific SVM parameters. A more refined but complex approach was proposed by Ho *et al.* using fuzzy rules defined as "*if gene A is up-regulated and gene B is down-regulated, then the probability of disease K is high*" [100].

2.7 - Introducing Functional Analysis and Biological Interpretation to Gene Sub-Sets

The methods mentioned here do not use biological knowledge for the selection of the genes. The selection criterion is simply to obtain the best possible model within the space that has been explored. For example, if our interest is developing models that explain the physiological state of tumour cells by means of gene expression profiles observed in nearby normal cells, from a biological prospective, the expected type of genes could be considered. Because the interaction between physically separated cells is only possible through the diffusion of secreted factors our algorithms could be forced to

explore preferentially a subset of genes fitting our expectation. Similarly, when modelling the behaviour of bacterial cells, operon structures could be used as well as knowledge available in the literature.

How can this be achieved? It is not sufficient to select a subset of genes and ignore all other possible solutions. Instead, a conditional probability of a certain gene to be included in the model dependent on its function or combine different types of data could be defined and used [128; 129]. In a Bayesian framework this can be done defining a prior distribution dependent on the function of each individual gene. Regardless of the modelling technique used, knowledge can be included in the VS strategy. In a GA this can be easily achieved setting different mutation rates that are dependent on pre-existing knowledge. The inclusion of biological knowledge in searching for models has been already done in different applications. For example, Gavaert *et al.* [130] included clinical information (CI) for Bayesian networks in three manners, (*i*) as if CI were additional variables integrating to the original matrix of gene expression, (*ii*) mixing learned models for gene expression and CI by weights, and (*iii*) learning the structure separately then mixing both structures. They concluded that only (*ii*) and (*iii*) were better than not making any mixing.

2.8 - Concluding Remarks

Microarray technology has been successfully used to relate gene transcription to different behaviours like cell cycle [116] and diseases like lymphomas [7] and prostate cancer [8; 9].

Microarray data is affected by systematic errors and noise. Pre-processing techniques are required to avoid these undesired effects keeping biological variability as intact as

possible.

One important problem using microarray data is a consequence of its high volume of data. Both classical statistical procedures and state of the art machine learning approaches have been proposed to select the information related to some desired observable outcome. Multivariate methods have been recently applied in this sense and proven to provide robust solutions even though univariate solutions are not found.

Multivariate methods coupled to stochastic search strategies such as GA or MCMC have been demonstrated to be model-independent, fast solution finding, relatively easy implementation, and easy translation from model to biology. Another important property of these stochastic methods is that the inclusion of biological knowledge is feasible. In the special case of multivariate models of two independent variables, an exhaustive search is feasible.

To build a multivariate model several univariate and multivariate methodologies use a ranking strategy that is then fed to a forward selection procedure (Figure 2.11). Recall that in these cases the forward selection explores the space starting always with the top variables and proceeds in order. So, in this case the space starting with variable two is never explored. However, in principle, instead of using a forward selection approach, any more versatile or explorative VS approach could be used (included the full version of forward selection). The motivation to use specifically forward selection in such limitation should correspond to historical reasons rather than to technical ones. So, to me, it should be expected that using better search engines such as GA or MCMC should improve the model in size and accuracy.

A very large number of supervised and non supervised methods have been applied and tested to analyse large scale functional genomics data. In this review, several methods

from classical statistics (t-test, ANOVA), data exploration (clustering and PCA), and state of the art machine learning techniques (ICA, NN, SVM, Fuzzy Logic, Wavelets) have been introduced. Table 2.1 shows a comparison on the methods used to analyse FG data for selected references cited in this chapter. Many of these methods have already become common analytical tools in bioinformatics. The importance of error estimation methods to avoid overfitting has also been stressed. Several methods can be adapted in order to study different biological systems like host pathogen interaction using microarray or other FG experiments.

Table 2.1 - Comparison of methodologies from publications reviewed in this Chapter. Model building column is specified only when it is different from the combinatorial search strategy.

Aim	Datasets	Processing-Filtering	Genes Used	Search Size	Comb. Search Strategy	Model Building Method	Classifier	Train/ Test	Train Error	Reference
Prediction	72 Leukemia		6817	1	Exhaustive	S2N Ratio	Centers+ Voting	38/34	Loocv	Golub *et al.*
Discovery							SOM			
Recurrence Genes	62C + 41N Prostate (Survival)	mean cent F/B > 1.5x >=3x var	5153/ 26260	1	Exhaustive		SAM	62/62	Perm	Lapointe *et al.*
Distinction	Gleason			1	Exhaustive		SAM	62/62	Perm	
Distinction	Capsular			1	Exhaustive		SAM	62/62	Perm	
Relapse	52T + 50N Prostate	mead cen <5 fold	12600	4+			KNN		Loocv	Singh *et al.*
Prediction	T/N	S2N (Golub)	456	1	Exhaustive		Correlation			
Prediction	Gleason	r corr perm	5265	1	Exhaustive		Correlation			
Distinction	8 Hum CL IR/nIR	gene avg	6800	1	Exhaustive		SAM	8/8	Perm	Tusher *et al.*
Prediction	88 Round Blue Cell Tumors (SRBCT)		2308	1	Exhaustive	Shrunken Centroids	Discriminant Shrunken Centroids	63/25	10cv	Tibshirani *et al.*
Outliers Reduction	Rat nervous system		112	ALL	ROBPCA		Discriminant Analysis		Loocv	Hubert *et al.*
Prediction	Mice cancer		2050							
Prediction	72 Leukemia (Golub)	Variance top	3930/ 7129	ALL	djPCA + Forward	Residuals + PCA score	F-test: disjoint PCA Residuals	38/34	Perm	Bicciato *et al.*
Prediction	SRBCT	Variance top	2308	ALL	djPCA + Forward	Residuals + PCA score	F-test: disjoint PCA Residuals	63/25	Perm	
Association Cell Cycle	Cell Cycle Scerevisae		4579	ALL			SVD			Alter *et al.*
Prediction	62 Colon (Alon)	Log	2000	5-50	GA	Forward	KNN	40/17	Loocv	Li *et al.*
Prediction	47 Lymphoma	Log2 (Alizadeh)	4026	5-50	GA	Forward	KNN	34/13	Loocv	
Prediction MultiClass	61 Cancer CL (Ross)	Ratio Variance top	1000/ 9703	<(11-15)	GA		MLHD	2/1	Loocv	Ooi *et al.*
Prediction MultiClass	198 GCM (Ramaswamy)	variance top	1000/ 16063	<(11-15)	GA		MLHD	144/54	loocv	
Prediction	22T + 21N Breast (Hedenfalk)		3226	Varied	MCMC Gibbs		Probit		Loocv	Eun Lee *et al.*
Prediction	72 Leukemia (Golub)		7129	Varied	MCMC Gibbs		Probit	38/34	Loocv	
Prediction	36 Drug Treated	Kruskal-Wallis	4700	Varied	RF		Decision trees	36/OOB	Loocv	Gunther *et al.*
Prediction	72 Leukemia (Golub)		7129	All-1 …	RFE-SVM		SVM	31/31	Loocv	Guyon *et al.*
Prediction (Proteomics)	69T+253N Prostate		15154	Opt	SVM		JOIN/ENS		Resub	Jong *et al.*
Prediction	22N+40T Colon (Alon)	M=0, SD=1	2000	20-1000	GA		SVM-Linear	50/12 x50 (10CV)	svm/ge	Fröhlich *et al.*
Prediction	Lymphoma	0~1	4026	ALL, W	NN		NN	2/3	cv	O'Neill *et al.*
Prediction	72 Leukemia (Golub)	MHT, t-Ratio	50/ 7129	ALL, W	NN	Weights	NN	38/34	Resub	Bicciato *et al.*
Discovery MultiClass	72 Leukemia (Golub)	Min~Max < 100	5855/ 7129	ALL	SOM	Adapted Golub	SOM	38/34		Hsu *et al.*
Prediction	Several	Varied	Varied	2	Exhaustive		SVM-like		Loocv	Grate *et al.*

(continuation)

Aim	Datasets	Processing-Filtering	Genes Used	Search Size	Comb. Search Strategy	Model Building Method	Classifier	Train/ Test	Train Error	Reference (et al.)
Distinction	59 Calibration		12640	1	Forward		Hotelling T2	1/0	Perm	Lu *et al.*
Distinction	82C+72N Liver (Chen)	(Chen)	11386/ 17400	1	Forward		Hotelling T2	1/0	Perm	
Distinction	34M + 44F Breast (van't Veer)			1	Forward		Hotelling T2	1/0	Perm	
Prediction	90C+22N Gastric (Chen) (Others)	Golub	Varied	Selected	Clustering Model		Model +Variance		Loocv	Qiu *et al.*
Prediction	Multiple Tumor	Forward Variance	Varied	Opt	Maximum Margins		Voting	144/54	Loocv	Antonov *et al.*
Prediction	Mice	Diff exp genes			R.Walk		Small Distance		kFold-cv	Xiao *et al.*
Prediction	Golub, Armstrong, Alon		7129, 12582, 2000	Varied	Ranks+ Average Clusters		KNN, SVM, C4.5, Bayes	72/72, 72/72, 62,62	Loocv	Wang *et al.*
Prediction	Golub			1	Exhaustive	Logist Reg	KNN		Varied	Guan *et al.*
Prediction	Several	Top30+ PCA+ Clustering	Several	All	100 NN		Voting	Varied	splits	Liu *et al.*
Prediction	Several		2000~ 12625	All	fastICA	A coefficients	Nearest Centroids	2/3 , 1/3	Loocv	Huang *et al.*
Association	SRBCT AML+ALL, Breast Cancer	none	2308 7129 ?	1	Exhaustive	Wavelets				Subramani *et al.*
Prediction	Rheumatoid Arthritis		999	Varied 1~300	MCMC		Probit	2/3, 1/3	Loocv	Sha *et al.*
Prediction	Breast Cancer	published	70	7	Exhaustive		KNN, Cart, LDA	2/3, 1/3	Bolstered	Choudhary *et. al.*

CHAPTER 3

The Development of a Statistical Modelling Environment Based on a Truly Multivariate Variable Selection Strategy

Large datasets are generated by genome-wide studies. Selecting those variables involved in certain sample phenotypes is a challenging task. Univariate and multivariate methods have been proposed to this end. Multivariate approaches seem to be more appropriate because they consider combinations and interactions of variables in a single model. However, there is no multivariate variable selection framework that can be easily configured and used in a variety of situations. Therefore, this chapter shows GALGO, a generalized multivariate variable selection framework. The usefulness of GALGO is demonstrated. An illustrative tutorial is also presented. Other uses of GALGO have been presented in Chapters 4 and 5 which includes the study of cell-to-cell communication and the study of functional genomics data other than transcriptomics.

3.1 - Introduction

Functional Genomics Technologies (reviewed in Chapter 2) allow us to measure thousands of transcripts, proteins, and metabolites in single experiments. In order to make sense of these complex datasets, methods with select relatively small subsets of biological measurements that are associated to cell function are essential. In this context, many statistical modelling techniques have been applied to microarray data [see Chapter 2 and reference 131]. These approaches can be subdivided into univariate and multivariate analysis. Univariate approaches test one feature at a time for their

ability to discriminate a dependent variable (Figure 2.15). The top-most significant features are then used to develop a statistical model [for an extensive comparison of classification methods in the context of a UVS strategy see reference 132]. Multivariate approaches take into consideration that variables are influencing a biological outcome in the context of networks of interacting genes rather than in isolation and can take into consideration synergy between genes, proteins, or metabolites. Although these approaches have been very successful, there are still issues in the development of multivariate models from large datasets. These issues are related to the extremely large number of possible models that would need to be evaluated to identify the most predictive. In order to address these issues, stochastic search strategies have been developed and tested on Functional Genomics datasets (FGD). Among these, Markov Chain Monte Carlo (MCMC) and Genetic Algorithms (GA) procedures have been used successfully in the analysis of microarray data [74; 79; 110]. Although a comparative study of these methods, analyzing functional genomics data, has not been published, the GA procedure seems to be computationally more efficient. Chapter 2 described the GA procedure in section 2.6.2.7. Briefly, GA starts from a random population of models and evolve good local solutions by mimicking the process of natural selection using mechanisms such as higher rate of replication of the more effective variable subsets (chromosomes), mutation to generate variants, and crossover to improve combinations [103]. The fact that sets of variables (arranged in chromosomes) are tested in combination during the selection process ensures a truly MVS. Past implementations of GA in FGD [74; 79] are limited to a specific GA procedure, classification methods, and error estimation strategy. Moreover, their output is a plain text file that is often difficult to interpret. In order to address these issues, the GALGO package has been developed. To provide flexibility, GALGO was implemented in an object-oriented fashion and has been coupled with a general-purpose fitness function to guide the variable selection. In the initial release, several fitness functions for solving classification problems were included, documentation for implementing ad-hoc fitness functions to solve general

optimization problems were also made available (the package, manual, and supplementary information can be downloaded from http://www.bip.bham.ac.uk/bioinf/galgo.html). This chapter will show an overview of the implementation, a quick tutorial of GALGO package, and a comparison of the accuracy and size of the models designed using GALGO for classification problems from microarray data with a number of other methods. The results suggest that GALGO is a useful tool in the analysis of large scale FGD.

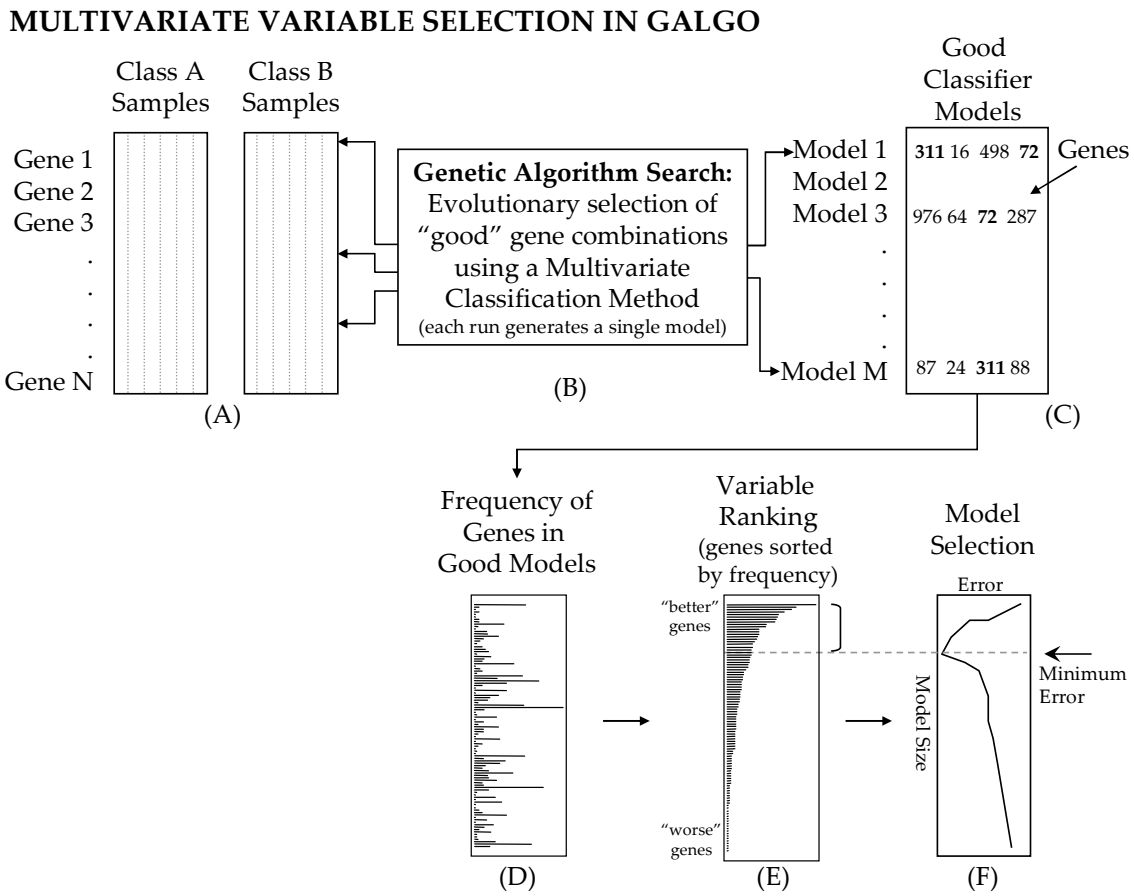


Figure 3.1 - Schematic representation of MVS in GALGO. From a dataset of two classes of samples (A), a genetic algorithm (B) searches and evolves combination of genes (chromosomes representing a multivariate model) that distinguish between classes using a classification method. A number of models are generated performing this procedure several times (C). These models may differ in gene content but with similar high classification accuracy. Genes appearing multiple times in different models suggest these genes are important for the classification problem in a multivariate context. Therefore, the number of times (frequency) a gene appears in a model is computed (D). These frequencies are used to rank genes (E). Then, a forward selection strategy is used to select a representative model that generates the lowest error (F).

3.2 - Implementation

The GALGO package has been conceived as an implementation of MVS using a GA in an object-oriented paradigm under the R language (Figure 3.1). GALGO uses a GA procedure for selecting models with a high fitness value and implements functions for the analysis of the populations of selected models as well as functions to reconstruct and characterize representative summary models [74].

3.2.1 - The GA procedure and GALGO Object-Oriented Design

GALGO uses Genetic Algorithms for selecting variable subsets (detailed in section 2.6.2.7 in Chapter 2). In short, the procedure starts from a random population of variable subsets of given size (defined as chromosomes). Each chromosome is assessed for its ability to predict a dependent variable and has a certain level of accuracy. The general principle is to replace the initial population with a new population that includes variants of chromosomes with higher classification accuracy and to repeat the process enough times to achieve a desired level of accuracy. The progressive improvement of chromosome population is driven by a number of operators that mimic the process of natural selection (selection, mutation, and crossover). In order to increase the proportion of the solution space that is explored, independent chromosomes populations can be evolved in partially isolated environments, named *niches* [103]. Chromosomes can occasionally migrate from one niche to another ensuring that particularly good solutions can recombine. In GALGO, a collection of niches is called *world*. A detailed description of the procedure is available in GALGO manual and tutorial (http://www.bip.bham.ac.uk/bioinf/ galgo.html).

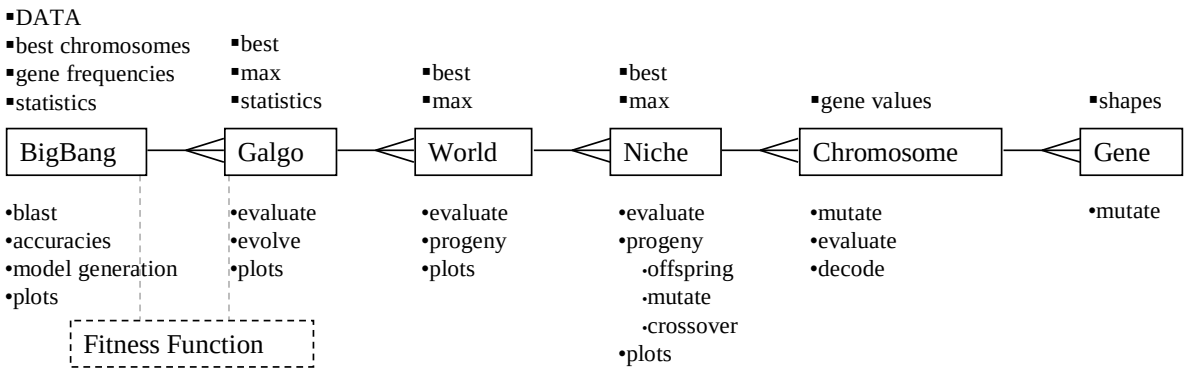


Figure 3.2 - Simplified object-oriented structure of the GALGO package. Boxes represent objects, which are connected by one-to-many relationships hierarchically. Major object properties are marked with solid squares above boxes whereas core methods are marked with solid circles below boxes. Dashed box represents the fitness function, which are included in GALGO for several classification methods. Dashed lines represent logical connections.

The object design of the GALGO package reflects the structure just described (Figure 3.2). In GALGO, the *Gene* object represents a variable (representing mRNA, proteins, metabolites, or any other measure) whereas the *Chromosome* object stands for a set of *n* variables that will be included in the multivariate model that will be evaluated using a fitness function. A *Niche* object organizes chromosomes in populations whereas the *World* object may include several niches. The *Galgo* object arranges all these objects, implements the GA evolutionary process, and saves the best chromosome. Finally, a *BigBang* object collects the result of several GA searches for further analysis. All these objects have properties that allow users to control each part of the process. To provide further flexibility, common GA operators such as *Reproduction*, *Mutation*, *Crossover*, *Migration*, and *Elitism* have been included as rewritable methods [see reference 103 for details about this operators]. Another important characteristic of GALGO is that the user can add custom defined properties to add new functionality.

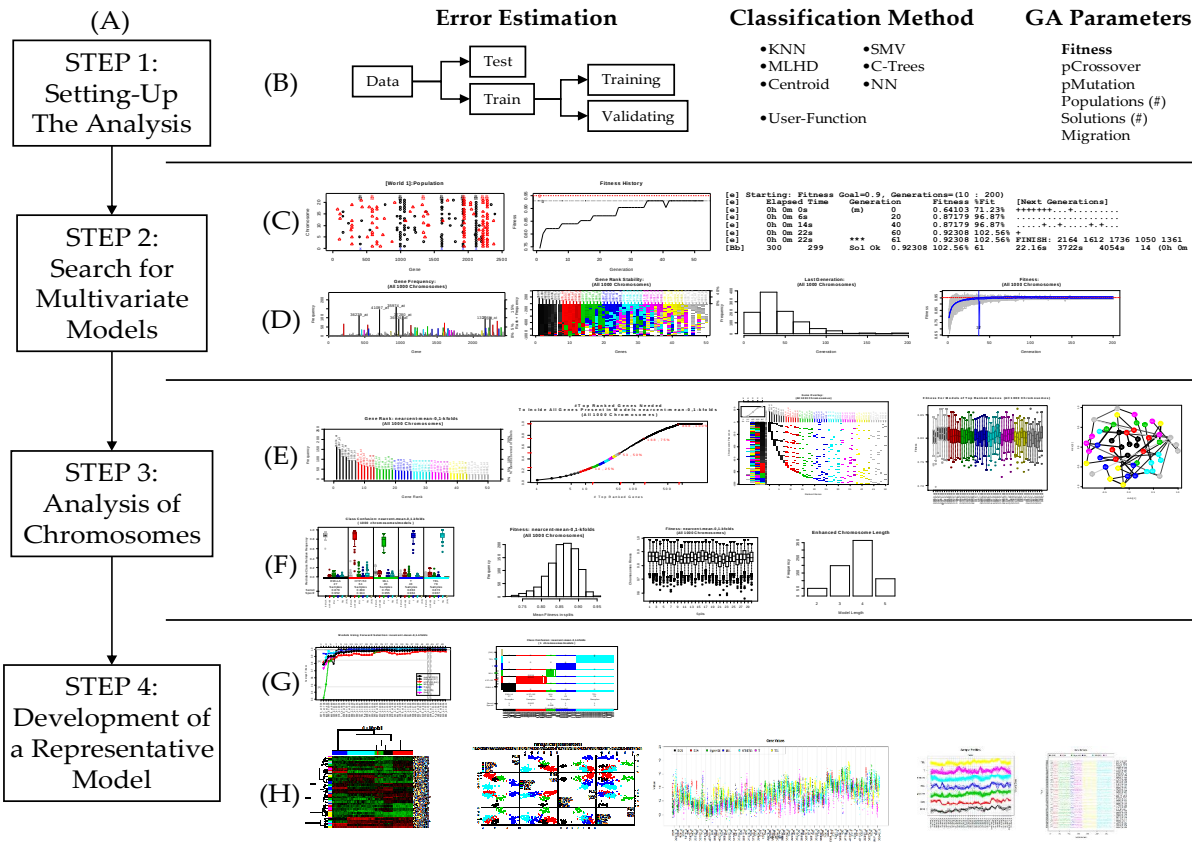


Figure 3.3 - Implementation and application of GALGO package. (A) General strategy (see text). (B) Specification of data input, error estimation, classification (fitness function) and GA parameters. (C) Monitoring evolution of GA process. The values of chromosomes inside the GA across generations and the evolution of the maximum fitness can be traced. (D) Monitoring models acquired. Every time the GA generates a final chromosome, this chromosome is collected and can be used to view the frequency for each gene, the stability of their ranks, the generation distribution, and the evolutionary fitness distribution. These plots serve as diagnostics during and after the whole process. (E) Analysis of genes in models. Collected chromosomes are analyzed in most frequent genes, number of genes present in chromosomes, top-genes overlapping in chromosomes and gene dependency in chromosomes. (F) Analysis of accuracy and specificity of models. The models or chromosomes can be assessed for: classification accuracy in test and/or train cases, confusion matrix, overall accuracy distribution, accuracy per data split, and determination of important genes within chromosome. (G) Development of representative model. (H) Model and chromosome visualization. Heatmaps and PCA decomposition plots can be used from the representative model or any other chromosome.

3.2.2 - Classification Methods in GALGO

In the current implementation, parametric and non-parametric classification methods such as KNN, MLHD discriminant functions, nearest centroid, SVM, classification trees, NN, and RF have been included (see Chapter 2 for an overview of these methods). The

first three were implemented in C whereas the others were adapted from original R packages.

3.3 - Application

The aim in this section is to describe a typical application of GALGO in a real dataset. The analysis protocol has been subdivided into four steps (for an overview of the whole process, see Figure 3.3, for details, see GALGO manual, tutorial, and supplementary material in http://www.bip.bham.ac.uk/bioinf/galgo.html).

3.3.1 - Step 1: Setting-Up the Analysis

In this initial stage of the analysis the user specifies the input data, the dependent variable (e.g. class labels), the statistical model (classifier), the desired accuracy (fitness), the error estimation scheme, and the parameters that define the GA search environment. Gene expression values can be provided in a common text file or as a matrix object, which may be the result of pre-processing using other R tools such as Bioconductor [133]. The classification method may be one of the six already implemented supervised classification methods, or a user-defined function. The error estimation can be defined at two levels: the classic training and test validation strategy using a single or multiple random splits, and inside the training process using k-fold cross-validation, random splits, or re-substitution error [see section 2.5.3.2 in Chapter 2 and reference 131]. The common GA parameters are automatically configured but can also be specified.

3.3.2 - Step 2: Searching for Relevant Multivariate Models

Every evolutionary cycle in the GA procedure starts from a random population of chromosomes and may lead to a diverse collection of good local solutions. For this reason a sufficiently large number of chromosomes should be selected in order to have a good representation of the solution space. Ideally such numbers should be sufficiently large to ensure that all solutions that can be found with the GA procedure are represented in the population of selected chromosomes. In order to make this possible two real time monitors that provide information on the chromosome composition have been designed. These are the level of convergence of the solutions and the evolution of the fitness values. These diagnostic plots are useful tools to assess when the searches converge to a stable population that can then be analyzed further.

3.3.3 - Step 3: Refinement and analysis of the Population of Selected Chromosomes

The chromosomes selected from the GA procedure have a fixed length defined at the first step of the analysis. Although the models have the desired classification accuracy, there is the possibility that not all genes included in the model contribute significantly to the fitness value. Therefore, a backward selection strategy (revised in section 2.6.2.6 in Chapter 2) has been implemented to derive a chromosome population where only genes effectively contributing to the classification accuracy of the model are included (refinement). This function can also be used within the selection process in step 2.

GALGO implements a number of functions for the analysis of the chromosome populations. These produce a text or graphical output and describe: 1) occurrence of genes in the model population, 2) gene composition of models, 3) model accuracy, 4) relative importance of genes within models, 5) evolution of the fitness function during chromosome selection, and 6) prediction of new samples. In addition, gene signatures

associated to specific chromosomes can be visualized using two dimensional clustering heatmaps, PCA plots, gene profiles, or class profiles.

3.3.4 - Step 4: Development of a Representative Statistical Model

The aim of this part of the analysis is to develop a single representative model from the population of selected chromosomes. In order to do so, a forward selection strategy has been implemented (see section 2.6.2.5 in Chapter 2) based on the step-wise inclusion of the most frequent genes represented in the chromosome population [74]. Every model developed with GALGO can be stored and used to predict the identity of novel samples.

3.4 - Quick GALGO Tutorial

This section describes a typical application of GALGO in biomarker discovery using large scale expression profiling data. The aim of this illustrative analysis is to identify gene sets that are predictive of disease type in a panel of leukaemia patients. This tutorial will describe the main and basic functionality implemented in GALGO to introduce the reader to the entire process. More advanced features in each step are detailed in the GALGO manual (http://www.bip.bham.ac.uk/bioinf/galgo.html). This tutorial assumes certain knowledge about the R programming environment (http://cran.r-project.org). The analysis pipeline implemented below is summarised in a schematic form in Figure 3.3. This tutorial uses the dataset described in section 3.5.2.1 -.

3.4.1 - Step 1 – Setting-Up the Analysis

The GALGO package includes a data-frame object (named ALL) that contains the normalized expression values of the datasets used in this tutorial. The object is a matrix

in which rows are genes and columns are samples. The identity of the samples is defined in another object (ALL.classes). Both objects are loaded making available for the user using the function *data*() as follows. In R type:

```
> library(galgo)
> data(ALL)
> data(ALL.classes)
```

Of course, user data from an external text file can be loaded specifying the parameter *file* instead of *data* and perhaps *classes* (refer to the GALGO manual for details) in the wrapper function. The wrapper function *configBB.VarSel* is used to specify the data, the parameters for the GA search, the classification method, the error estimation method, and any user-defined parameter. This function builds a BigBang object that contains the data and the values of all parameters and it eventually stores the results of the analysis.

To set up the GA search type in R:

```
> bb.nc <- configBB.VarSel(
data=ALL,
classes=ALL.classes,
classification.method="nearcent",
chromosomeSize=5,
maxSolutions=300,
goalFitness = 0.90,
main="ALL-Tutorial",
saveVariable="bb.nc",
saveFrequency=30,
saveFile="bb.nc.Rdata")
```

The code above configures a BigBang object that will store 300 chromosomes (*maxSolutions=300*) which will contain 5 genes (*chromosomeSize=5*) that correspond to models developed using a nearest centroid classifier (*classification.method="nearcent"*) with a classification accuracy of at least 90% (*goalFitness=0.9*). The other parameters define the name of the saved object that is created (*saveVariable="bb.nc"*), the frequency

of saving the results in a file (*saveFrequency=30*) and the name of the file where the results are saved (*saveFile="bb.nc.Rdata"*).

The wrapper function *configBB.VarSel* can also be used to configure additional functions. Please refer to the GALGO manual for an extensive description of the *configBB.VarSel* parameter specification. To show the available parameters and their descriptions type:

```
> ?configBB.VarSel
```

3.4.2 - Step 2 - Evolving Models/Chromosomes

Once the BigBang and Galgo objects are configured properly, the user is ready to start the procedure and to collect chromosomes that are good predictive models of the tumour class. This is achieved by calling the method *blast*().

In R type:

```
> blast(bb.nc)
```

This procedure can last a long time, from minutes to hours, depending on the degree of difficulty in the classification problem, in the classification method, and in the GA search parameters. The default configuration displays the course of BigBang and Galgo objects to the console (controlled by the verbose parameter) including the approximated remaining time.

This is an example of the text output for one GA cycle (61 generations):

```
[e] Starting: Fitness Goal=0.9, Generations=(10 : 200)
[e]     Elapsed Time    Generation      Fitness %Fit    [Next Generations]
```

```
[e]     0h 0m 0s        (m)     0       0.64103 71.23%  +++++++...+.........
[e]     0h 0m 6s                20      0.87179 96.87%  ....................
[e]     0h 0m 14s               40      0.87179 96.87%  .....+..+.....+.+...
[e]     0h 0m 22s               60      0.92308 102.56% +
[e]     0h 0m 22s       ***     61      0.92308 102.56% FINISH: 2164 1612...
[Bb]    300     299     Sol Ok  0.92308 102.56% 61      22.16s  3722s   4054s   14 (0h
0m 14s )
```

Lines starting with "[Bb]" correspond to the current collection of the BigBang object. This line shows respectively the number of evolutions (300 in this case), the number of evolutions that have reached the goal fitness (299), the status of the last evolution (Sol Ok – the goal fitness was reached), the fitness value of the best chromosome from the last evolution (0.92408) along with its percentage relative to the goal fitness (102.56%), the number of generations it needed (61), the process time spent in last evolution (22.16 seconds), the accumulated process time spent in all evolutions (3,722 seconds), the accumulated real time (4,054 seconds, which considers the time spent by saving the object and other operative system delays), and the remaining time needed to collect the previously specified number of chromosomes (14 seconds).

Lines starting with "[e]" represent the output of the evolutionary process (the genetic algorithm search). The first line of each evolution shows the goal fitness and the constraints in generations. Successive lines show, in columns, the elapsed time, the current number of generation (by default refreshed every 20 generations) and the current best fitness along with the percentage relative to the goal fitness. The last column summarizes the behaviour of the next generations, "+" means that maximum fitness of the current population has increased, "-" means that it has decreased, and "." means that it has not changed. "G" appears occasionally when the fitness goal has been reach but the algorithm can not end because of a constraint in the number of generations.

The default configuration would show three plots summarizing the characteristics of the population of selected chromosomes. This plot is shown in Figure 3.4. The top-most

plot shows the number of times each gene has been present in a stored chromosome, by default the top 50 genes are coloured and the top 7 are named. The middle plot shows the stability of the rank of the top 50 genes, which is designed to aid in the decision to stop or continue the process once the top ranked genes are stabilized. When genes have many changes in ranks, the plot shows different colours; hence the rank of these genes is unstable. Commonly the top 7 "black" genes are stabilized quickly, in 100 to 300 solutions, whereas low ranked "grey" genes would require many thousands of solutions to be stabilized. The plot at the bottom is the distribution of the last generation of the GA process to have produced a solution. It is intended to show how difficult is the search problem for the current configuration of GA. If peaks are observed at either end, a configuration change is advisable (see GALGO manual).

Once the blast method ends, you can continue with the analysis step. Nevertheless, the blast process can be interrupted (by typing the ctrl-c keys in Linux or esc in windows) and the results analyzed straight away. It is recommended to break the process in the evolution stage, not in the BigBang update stage as that may disrupt the object. The process can be resumed by typing the blast command again. The result of the last evolution might be lost but the accumulated results should remain intact.

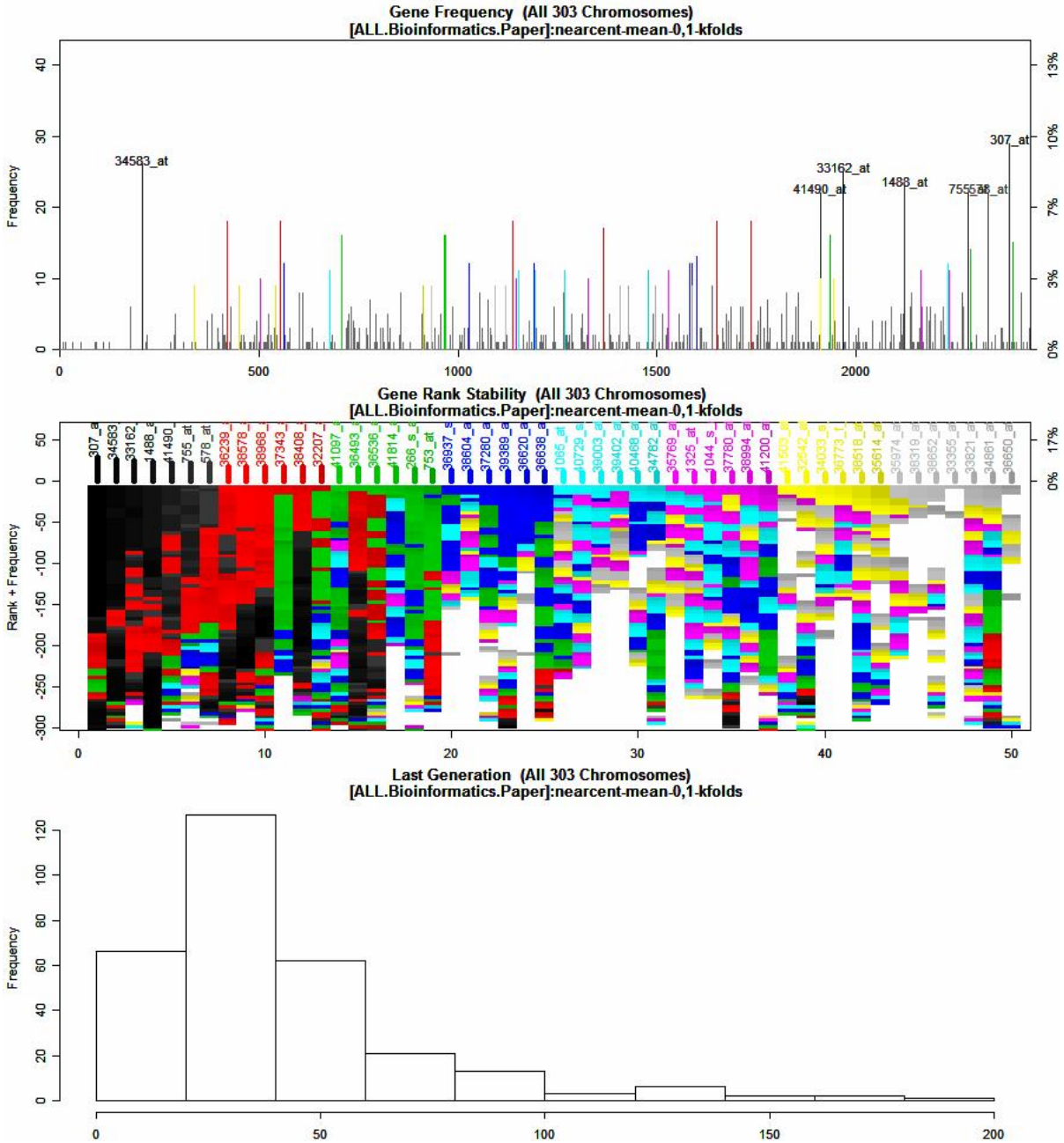


Figure 3.4 - Default monitoring of accumulated chromosomes in the BigBang object.

Resuming the process will have the effect of restarting the Galgo object as in any cycle. The possibility to interrupt the process is very useful for initial exploratory analysis since the most updated results can be analysed and can be saved anyway using the saveObject method. Instead of interrupting the process, one can open a new R console and benefit from the use of progressive saving strategy that updates the current object called "bb.nc" into a file named "bb.nc.Rdata" once at least 30 solutions have been

reached or the saveObject method has been called (controlled by saveVariable, saveFile, and saveFrequency parameters respectively). To do this, a previously saved object can be loaded in GALGO using the loadObject method in a new R console window:

```
> library(galgo)
#change directory to yours
> loadObject("bb.nc.Rdata")
```

Once the file is loaded, the loadObject method displays a summary of the loaded variables and their classes and you can proceed to the analysis step.

GALGO also has the functionality to summarise the population of chromosomes within each generation. The code below shows the modifications to the definition of the BigBang Object that are required to activate this function (marked in red).

```
> x11()
> x11()
> bb.nc <- configBB.VarSel(
data=ALL,
classes=ALL.classes,
classification.method="nearcent",
chromosomeSize=5,
maxSolutions=300,
goalFitness = 0.90,
saveVariable="bb.nc",
saveFrequency=30,
saveFile="bb.nc.Rdata",
main="ALL-Tutorial",
callBackFuncGALGO=plot,
callBackFuncBB=function(...) {dev.set(2);plot(...);dev.set(3);}
)
```

The topmost plot in Figure 3.5 shows the current values of the genes in chromosomes in order to show the explorative process. The middle plot shows the evolution of the fitness relative to the goal in the course of generations. The plot at the bottom shows the history of the maximum chromosome.

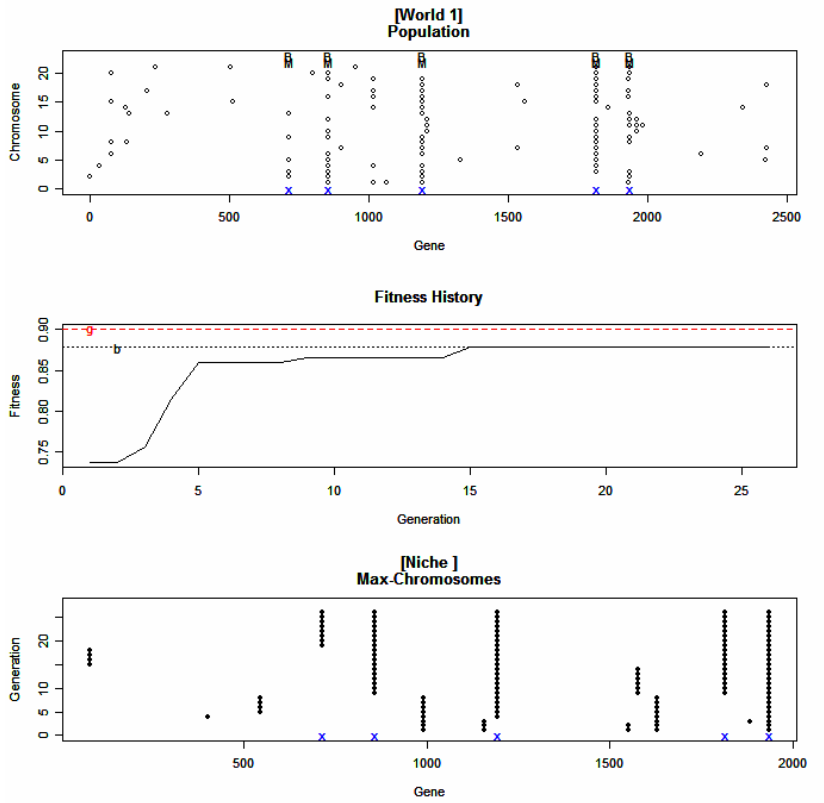


Figure 3.5 - Real-time monitoring of the Genetic Algorithm search. The horizontal axis of the top and bottom plots display unranked gene indexes. The vertical axis of the top panel is displaying the chromosome index whereas the vertical axis of the bottom panel is displaying the generation number. In the middle plot the horizontal axis is displaying the generation whereas the vertical axis is displaying the fitness value.

3.4.3 - Step 3 - Analysis and Refinement of Populations Chromosome

3.4.3.1 - Are we getting solutions?

The first question to answer is whether one is actually getting acceptable solutions. By default, configBB.VarSel configures the BigBang object to save all chromosomes even if they did not reach the goalFitness value. The reason is that one may need to assess the success of the configured GA search in all searches, not only in those that reach solutions. To analyze the success of the configured GA search, one can look at the evolution of the fitness value across generations, using the code below.

```
> plot(bb.nc, type="fitness")
```

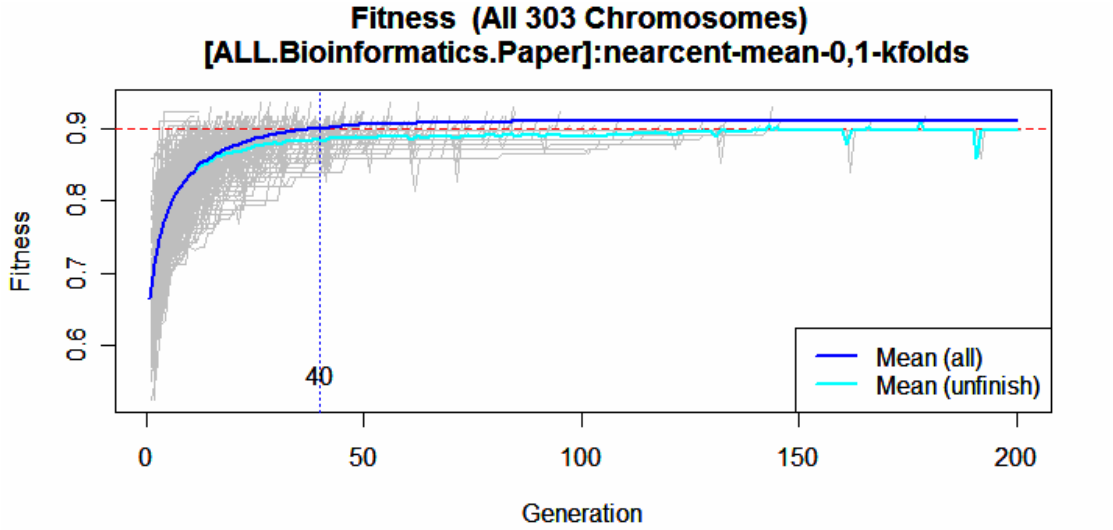


Figure 3.6 - Evolution of the maximum fitness across generations in 303 independent searches.

Figure 3.6 shows that on average, the search is reaching a solution in generation 40, which is very sensible. The blue and cyan lines show the average fitness for all chromosomes and for those that have not reached a goal respectively. These lines intend to delimit an empirical "confidence interval" for the fitness across generations. Grey lines describe the course of each evolution. The characteristic plateau effect could be useful to decide if the search is not working to reach our goal, which is marked with a dotted line. See the GALGO manual if you cannot reach solutions.

It is possible to separate the evolutions that have reached the goal using the following code.

```
> par(mfrow=c(2,1))
> plot(bb.nc, type="fitness", filter="solutions")
> plot(bb.nc, type="fitness", filter="nosolutions")
```

The "filter" parameter can be used almost in any function and in any plot type.

3.4.3.2 - What is the overall accuracy of the population of selected models?

Once the chromosomes have been selected it is needed to assess the classification accuracy of the corresponding models using one of the three error estimation strategies described in **BOX 1** (see Chapter 2 for details) that were specified in the first step. The default configuration will estimate the accuracy of the models using strategy 3 as described in **BOX 1**.

Use the following command to plot the overall accuracy.

```
> plot(bb.nc, type="confusion")
```

The output of this function is shown in Figure 3.8. The horizontal axis represents the individual samples grouped according to class whereas the vertical axis represents the predicted classes. The barcharts represent the percentage of models that classify each sample in a given class. For example, samples in the second column (marked in red) belong to the HYP+50 class. These are, on average, correctly classified 85.6% of the times. However, on average, they are "wrongly" classified 2.5% of the times as EMLLA, 5.4% of the times as MLL, 1.5% as T, and 5% as TEL. The plot also reports the value of sensitivity and specificity of the prediction. These are measures of the overall prediction per class. The sensitivity of the prediction for a given class *k* is defined as the proportion of samples in *k* that are correctly classified. The specificity for a given class *k* is defined as the number of true negatives divided by the sum of true negatives and false positives.

BOX 1: Error estimation strategies in GALGO

There are several methods to estimate classification accuracy. These are all based on the fundamental principle that a correct estimate of accuracy must be performed on a set of samples that has not been used to develop the model itself. Classical approaches involve splitting data into training and test sets. The training set is used to estimate the parameters of the model whereas the test set is left aside and it is used to assess the accuracy of the model itself. This approach is considered the most appropriate when a large number of samples is available. However, when the number of samples is relatively small, as it is the case of a typical functional genomics experiment, the test set could be too small to estimate the classification accuracy with acceptable precision. In order to estimate the accuracy with small datasets it is possible to use a different statistical technique called *cross-validation* (see Chapter 2). The dataset is split in *k* different training and test sets. The classification accuracy is then defined as the average of the classification accuracies calculated, by default, on the test sets for each of the *k* splits. GALGO uses a technique called bootstrapping [86] to generate the splits.

Within GALGO one can use three main strategies for estimating classification accuracy. In the first strategy a simple cross-validation or resubstitution error strategy is used to compute the value of the fitness function that guide chromosome selection in the GA procedure. The classification accuracy of the selected chromosome is defined as the fitness value (Figure 3.7A). The second strategy (Figure 3.7B) is a classic training and test procedure where the accuracy is estimated on the test data. In the GA process, the value of the fitness function is estimated by cross validation on the training data. Other approaches, such as .632 bootstrap [86], combine training and test accuracies, which can be specified as error weights through the parameter *classification.test.error = c(.368, .632)* for training and test respectively. The third strategy is to select the chromosomes as in the second strategy and to compute the classification accuracy of the selected chromosomes as the average of the classification accuracy estimated on *k* data splits as exemplified in Figure 3.7C. GALGO defines the initial split (common to both strategies) as Split 1.

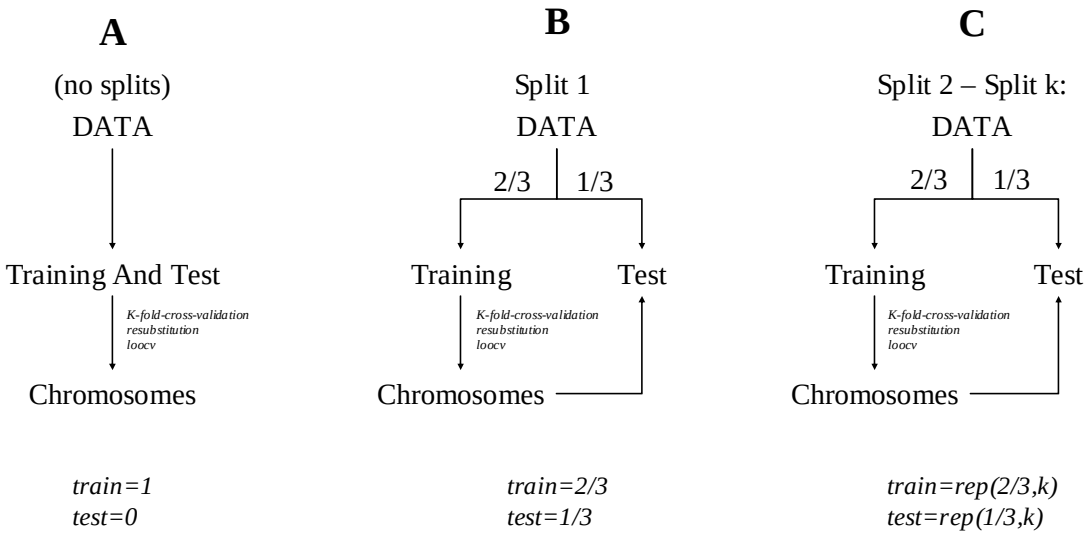


Figure 3.7 - Schematic Representation of the Estimation of Classification Accuracy. (A) Strategy 1, using all data as training and test. (B) Strategy 2, classical training and test. (C) Strategy 3, *k* repetitions of the strategy 2. The respective values of the parameters, *train* and *test*, needed to perform each strategy is shown at the bottom of the schema.

To obtain the confusion matrix, specificity, and sensitivity measures in a numeric format use the following code.

```
> cpm <- classPredictionMatrix(bb.nc)
> cm <- confusionMatrix(bb.nc, cpm)
> sec <- sensitivityClass(bb.nc, cm)
> spc <- specificityClass(bb.nc, cm)
```

cpm is a matrix with the number of times every sample as been predicted as any other class. For instance, let's analyze the first rows of cpm.

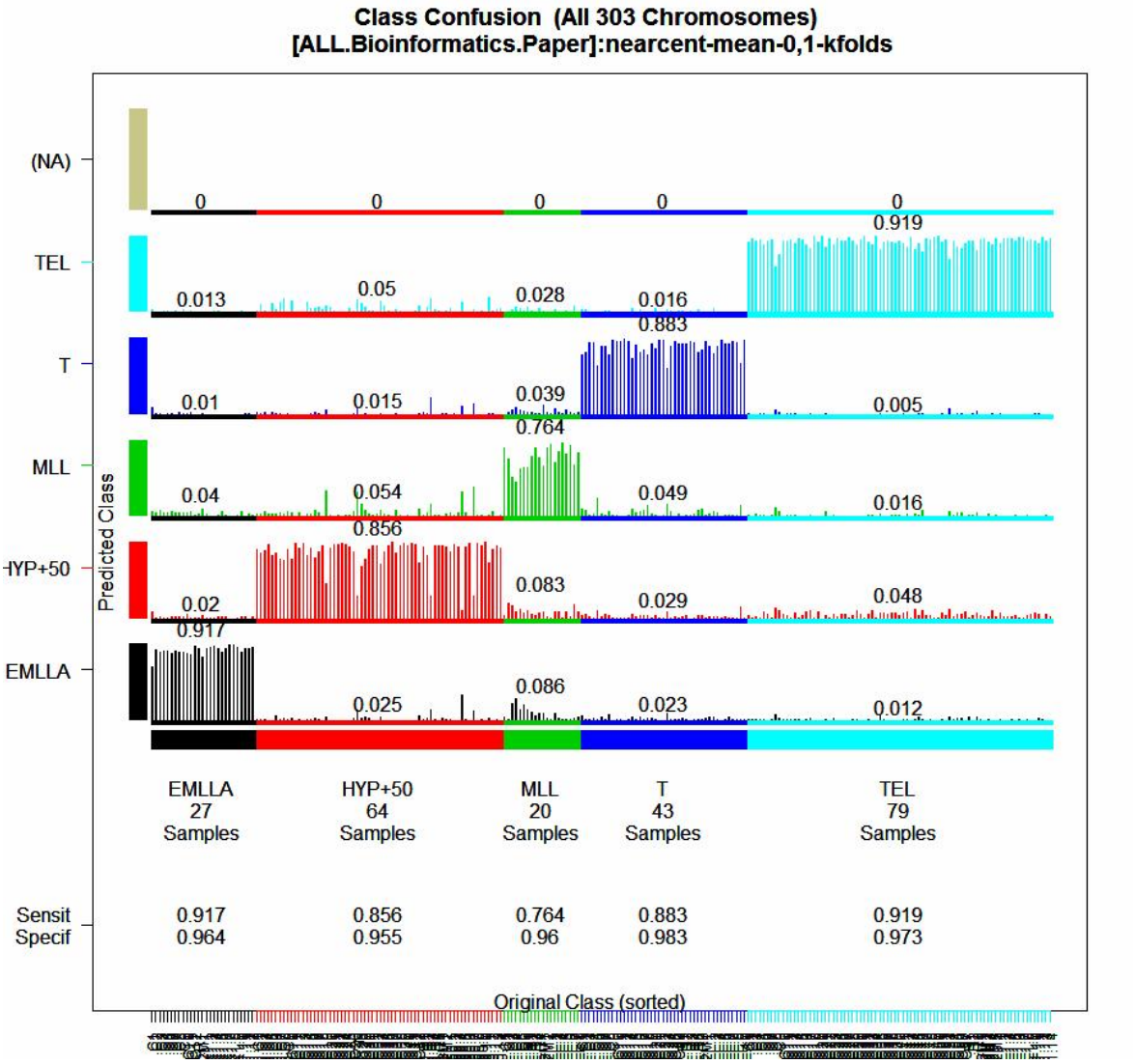


Figure 3.8 - Overall classification accuracy.

```
> cpm[1:3,]
             EMLLA HYP+50  MLL    T TEL (NA)
E2A.PBX1.C1 10837   1418 1062 1473 663    0
E2A.PBX1.C2 13654    101  767  113 212    0
E2A.PBX1.C3 13729    262 1225  218  19    0
```

The output above shows that E2A.PBX1.C2 sample has been predicted 14847 times (the row sum), in which 13654 times (92%) as been predicted as EMLLA, 101 times (0.6%) as HYP+50 and so on. The number of predictions depends on how the error is estimated in terms of training and test sets, and the number of chromosomes (type ?classPredictionMatrix.BigBang). By default, the prediction is made on test sets only for each chromosome (303 in the plot shown here). Initially, configBB.VarSel function generated 150 random test sets and each test set was made using 1/3 of the samples. Thus, on average, a sample would be predicted 303 * 150/3 times, that is 15150 times, which is comparable to 14847 for the second sample. In certain circumstances, some classification methods cannot make a prediction based on the data, "(NA)" column summarise those cases (for nearest centroid method, it will be always 0).

To evaluate the error in the first training set (as they were evolved), one can use the following changes in parameters.

```
> plot(bb.nc, type="confusion", set=c(1,0), splits=1,
filter="solutions")
```

set parameter specify that the error estimation should be computed in the training set only. *splits* parameter limit the estimation to one partition, the original used to evolve the chromosomes. *filter* specify that only chromosomes that reach the goal fitness will be evaluated. In this plot (not shown) some samples do not show their respective "bar", which indicates that those samples were never predicted. This is because a limit for the evaluation was used to the train set in the split #1, which should contain 155 samples approximately (2/3=66%).

To evaluate a single chromosome or any other model in the same circumstances use the following code.

```
> plot(bb.nc, type="confusion", set=c(0,1), splits=1,
chromosomes=list(bb.nc$bestChromosomes[[1]]))
```

In this case, the bars do not represent an average prediction because each test sample were predicted once (1 model in 1 split only).

3.4.3.3 - Is the rank of the genes stable?

Stochastic searches (such as GA) are very efficient methods to identify solutions to an optimization problem (e.g. classification). However they are exploring only a small portion of the total model space. The starting point of any GA search is a random population. Different searches therefore are likely to provide different solutions. In order to extensively cover the space of models that can be explored, it is necessary to collect a large number of chromosomes. GALGO offers a diagnostic tool to determine when the GA searches reach some degree of convergence. Our approach is based on the analysis of the frequency that each gene appears in the chromosome population. As chromosomes are selected, the frequency of each gene in the population will change until no new solutions are found. Therefore, monitoring the stability of gene ranks (based on their frequency) offers the possibility to visualize that the GA has detected a representative population of models.

To produce the rank stability plot type:

```
> plot(bb.nc, type="generankstability")
```

By default, the most frequent 50 genes are shown in 8 different colours with about 6 or 7 genes per colour. Figure 3.9 and Figure 3.10 shows two extreme examples. Horizontal axis in these figures shows the genes ordered by rank. Vertical axis shows the gene frequency (in the top part of the y axis) and the colour coded rank of each gene in previous evolutions. Consequently, for a given gene, changes in ranks are marked by different colours (below the frequency). Figure 3.9 shows that the first 7 black genes have been stable at least during the last 50 solutions whereas some red genes have recently swapped from green. Thus, red and green genes are not yet stable; this is because 303 chromosomes are not enough to stabilize these genes. Probably, 1000 chromosomes would generate more stable results, nevertheless, the more chromosomes the better. For a comparison, Figure 3.10 shows the result for the same run used here but using 1,000 chromosomes, which shows more stability in ranks, at least, for black, red, and perhaps green and blue marked genes. Thus, top genes are being stabilized in order; first black genes, then red, green and so on.

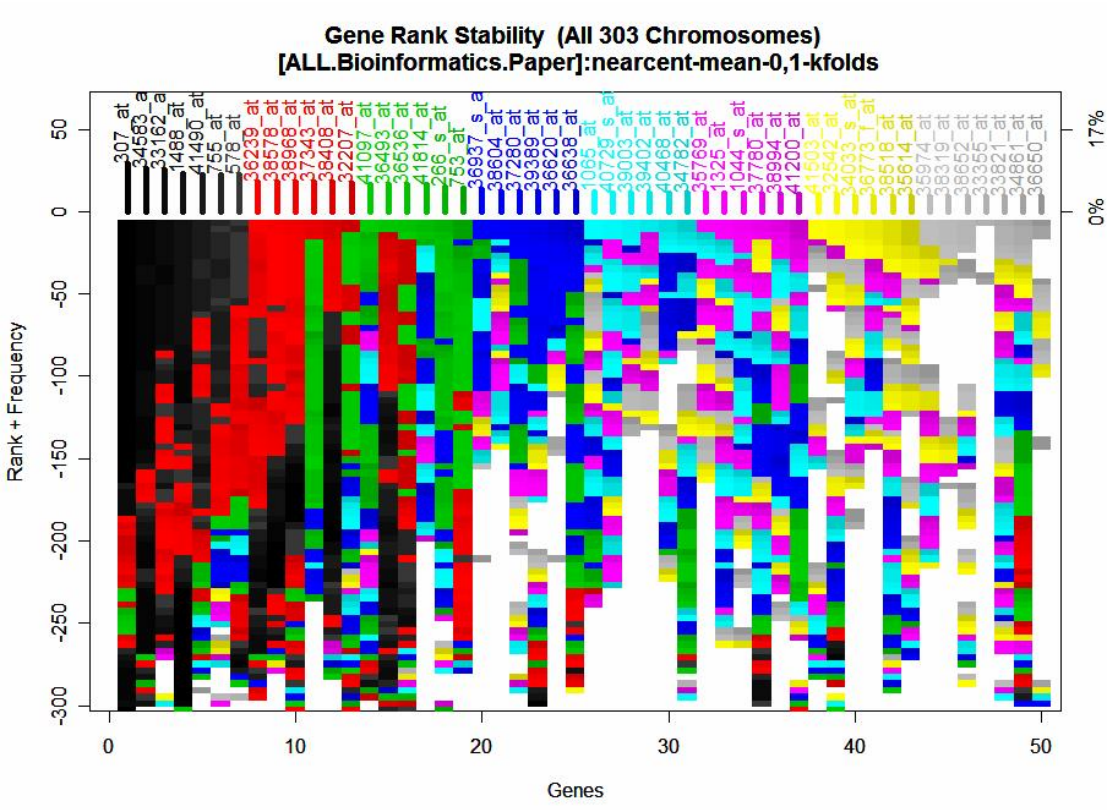


Figure 3.9 - Gene Ranks across past evolutions.

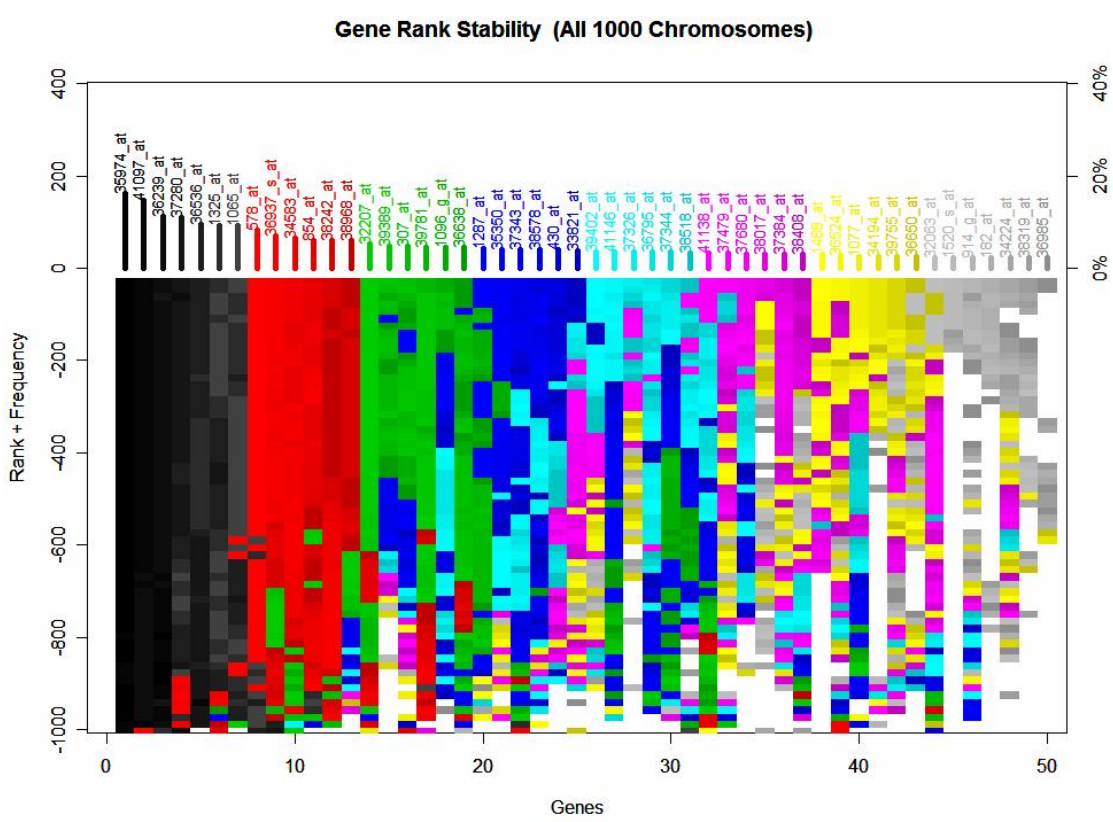


Figure 3.10 - Rank Stability in 1000 chromosomes.

3.4.3.4 - Are all genes included in a chromosome contributing to the model accuracy?

The chromosome size is fixed by an initial parameter in GALGO. This implies that some of the genes selected in the chromosome could not be contributing to the classification accuracy of the correspondent model. GALGO offers the possibility to identify these genes and remove them from the chromosomes. This can be done after the selection is completed or within the selection process itself. In order to perform this task a backward selection procedure has been implemented (see Chapter 2). Briefly, the methodology works as follows. A given gene is removed from the chromosome. The classification accuracy of the resulting shorter chromosome is then computed. If this is not reduced, another elimination cycle is performed. If the classification accuracy is

reduced the gene is left in the chromosome and another elimination cycle is performed. The process continues until all genes have been tested.

In order to perform this procedure outside the blast method type:

```
> rchr <- lapply(bb.nc$bestChromosomes[1:300],
robustGeneBackwardElimination, bb.nc, result="shortest")
```

The distribution of the size of the refined chromosome population can be plotted using the following function.

```
> barplot(table(unlist(lapply(rchr,length))), main="Length of
Shortened Chromosomes")
```

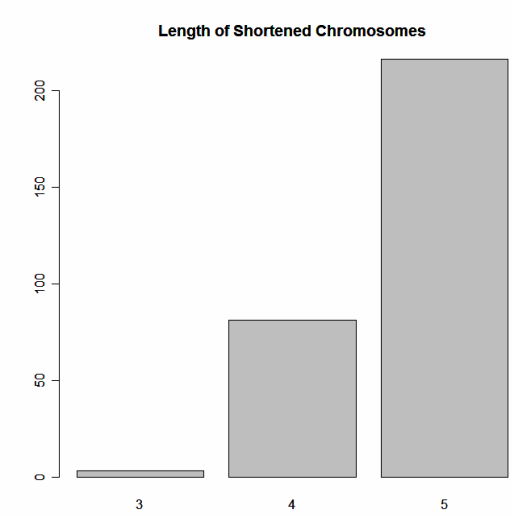


Figure 3.11 - Refinement of chromosomes.

Figure 3.11 shows that a large proportion of the chromosomes require all five genes to accurately classify the samples. Considering that the problem trying to be solved here is a five-class problem (multi-class), the fact that in this example the majority of the models actually need five genes is not particularly surprising. However, it is common to build models with more genes than classes; indeed the majority of the datasets

actually contain only two classes (e.g. treated-untreated, cancer-normal, wild-mutant, etc). Therefore, it is advisable to perform this analysis regularly.

3.4.4 - Step 4 - Developing Representative Models

The GA procedure provides a large collection of chromosomes. Although these are all good solutions of the problem, it is not clear which one should be chosen for developing a classifier, for example, of clinical importance or for biological interpretation. For this reason there is a need to develop a single model that is, to some extent, representative of the population. The simpler strategy to follow is to use the frequency of genes in the population of chromosomes as criteria for inclusion in a forward selection strategy [74]. The model of choice will be the one with the highest classification accuracy and the lower number of genes. However GALGO also stores alternative models with similar accuracy and larger number of genes. This strategy ensures that the most represented genes in the population of chromosomes are included in a single summary model.

This procedure should be applied to the population of chromosomes generated by initial GA search. However, it can also be applied to the population of chromosomes that is the result of backward selection procedure explained in the previous section.

The forward selection model can be generated by typing:

```
> fsm <- forwardSelectionModels(bb.nc)
> fsm$models
> ?forwardSelectionModels.BigBang # Help System
```

Figure 3.12 shows the results from forward selection procedure. The selection is done evaluating the test error using the fitness function in all test sets. The output is a list of values including the models whose fitness is higher than 99% of the maximum (or above a specified value using "minFitness" parameter). *fsm* object contains the ***best*** models (29 in this case). The model labelled as 12, containing the 33 most frequent genes, was the best model in terms of accuracy. The other 28 models included in *fsm* are 99% as close to the best model. Models included in the result can be viewed in heat maps, PCA space, or profiles. To visualize the best model in a heatmap plot use the following code.

```
> heatmapModels(bb.nc, fsm, subset=12)
```

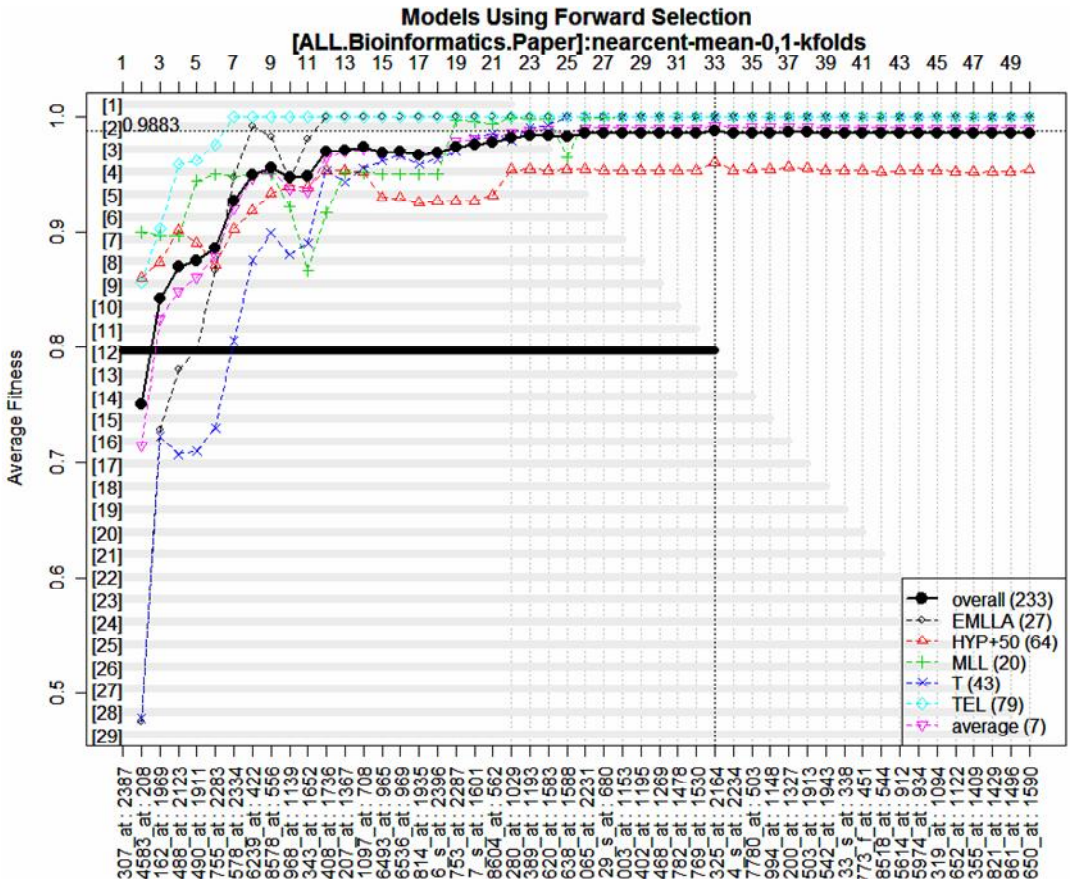


Figure 3.12 - Forward selection using the most frequent genes. Horizontal axis represents the genes ordered by their rank. Vertical axis shows the classification accuracy. Solid line represents the overall accuracy (misclassified samples divided by the total number of samples). Coloured dashed lines represent the accuracy per class. 1 model resulted from the selection whose fitness value is maximum (black thick line), but 29 models were finally reported because they were very similar in absolute value.

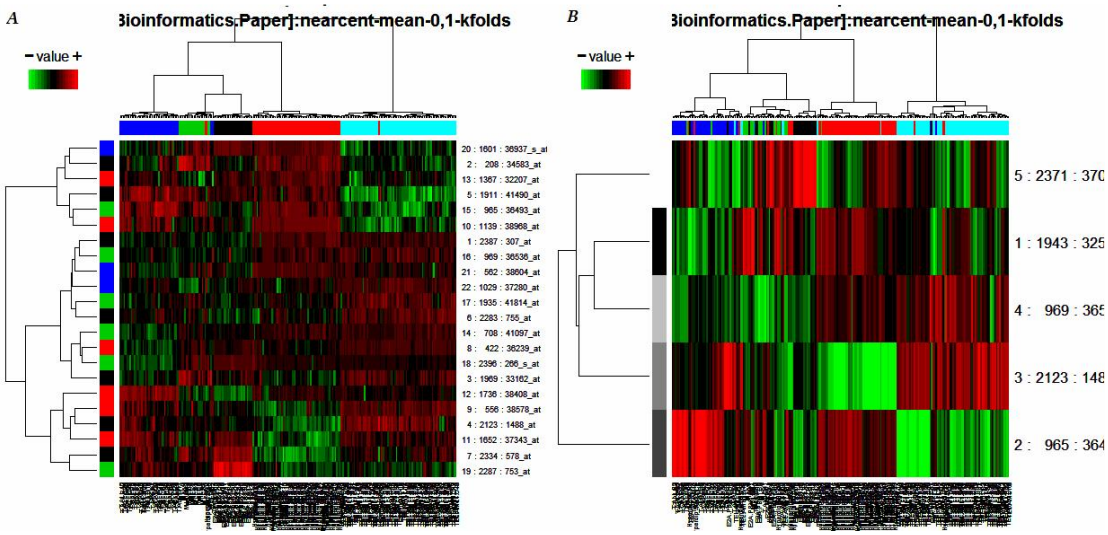


Figure 3.13 - Heatmaps. From a model resulted from forward selection (A) and an original evolved chromosome (B).

Details for visualization of models (or chromosomes) are given in the GALGO manual. The classification accuracy can be plotted extracting the information for any of these specific models, as in the example below (plot not shown).

```
> plot(bb.nc, type="confusion",
chromosomes=list(fsm$models[[1]]))
> cpm.1 <- classPredictionMatrix(bb.nc,
chromosomes=list(fsm$models[[1]]))
> cm.1 <- confusionMatrix(bb.nc, cpm.1)
> mean(sensitivityClass(bb.nc, cm.1))
[1] 0.9863334
> mean(specificityClass(bb.nc, cm.1))
[1] 0.9965833
```

From the mean values of sensitivity and specificity one can conclude that the selected model is, by far, more accurate than any original evolved chromosome.

3.4.5 - Visualizing Models and Chromosomes

Gene signatures associated within individual chromosomes or in a representative model (derived by forward selection) can be visualised in GALGO using a number of

graphical functions. This section will illustrate the use of heat maps, PCA, and other methods. For, the typical heat map format, use the following commands.

```
> heatmapModels(bb.nc, fsm, subset=1) # forward
> heatmapModels(bb.nc, bb.nc$bestChromosomes[1])
```

The results are shown in Figure 3.13[1].

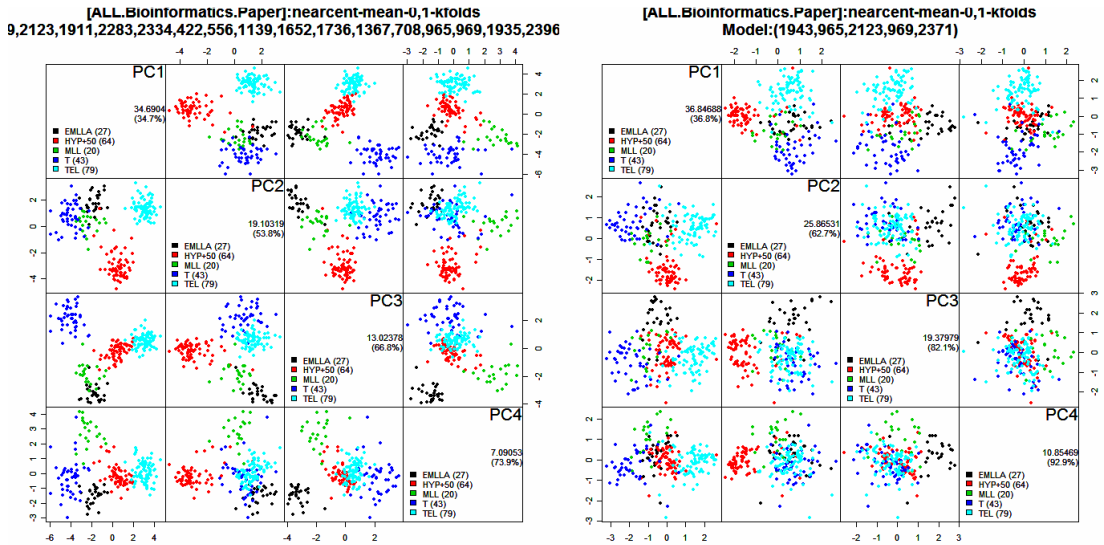


Figure 3.14 - Depiction of a model (left) and a chromosome (right) in PCA space.

In order to visualise the relation of samples using the genes selected in a chromosome or in a representative model one can also use principal component analysis representation. In order to do this, type the following command (see Figure 3.14).

```
> pcaModels(bb.nc, fsm, subset=1)
> pcaModels(bb.nc, bb.nc$bestChromosomes[1])
```

[1] Remember that the hierarchical clustering of samples given in the heatmap is the product of an unsupervised algorithm, which may differ from the classification method of our choice. Therefore, the relative sample order in the heatmap, the original class, and the predicted class by the model may all be different. Nevertheless, many of the times, the hierarchical clustering gives a good overview.

By default, only the first four components are shown, which can be changed specifying the *npc* parameter.

Another useful way to show a model is using the profiles of samples within a class as shown in Figure 3.15, which is the result from the code.

```
> plot(bb.nc, fsm$models[[1]], type="sampleprofiles")
```

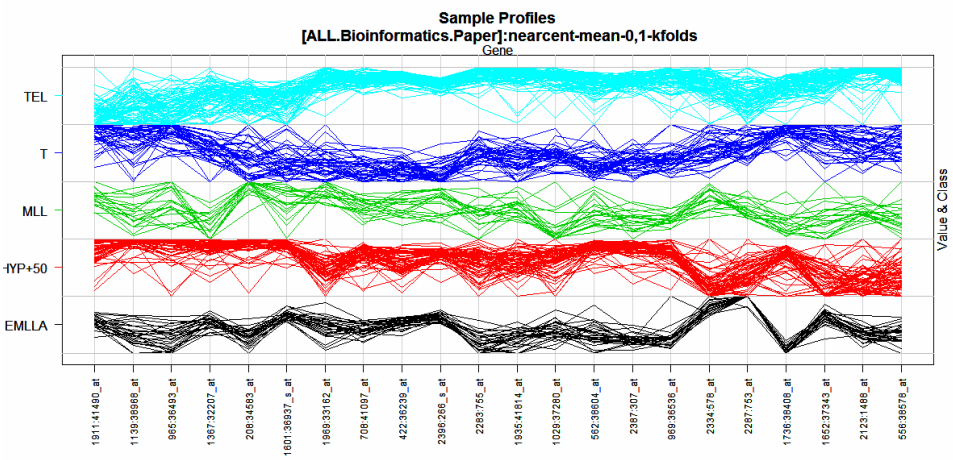


Figure 3.15 - Sample profiles per class.

3.4.6 - Gene content in Models

Another interesting analysis is how top-ranked genes are included in models, that is, what is the gene-composition of models. Figure 3.16 shows the composition of the models in terms of top-ranked genes. By default, the chromosomes are sorted by their most top-ranked genes; hence, chromosomes with similar top-ranked-genes are stacked together. Chromosomes are shown in vertical and genes in horizontal. For example, one can see easily which genes have been combined with the first top-gene.

```
> plot(bb.nc, type="geneoverlap",cex=.75)
```

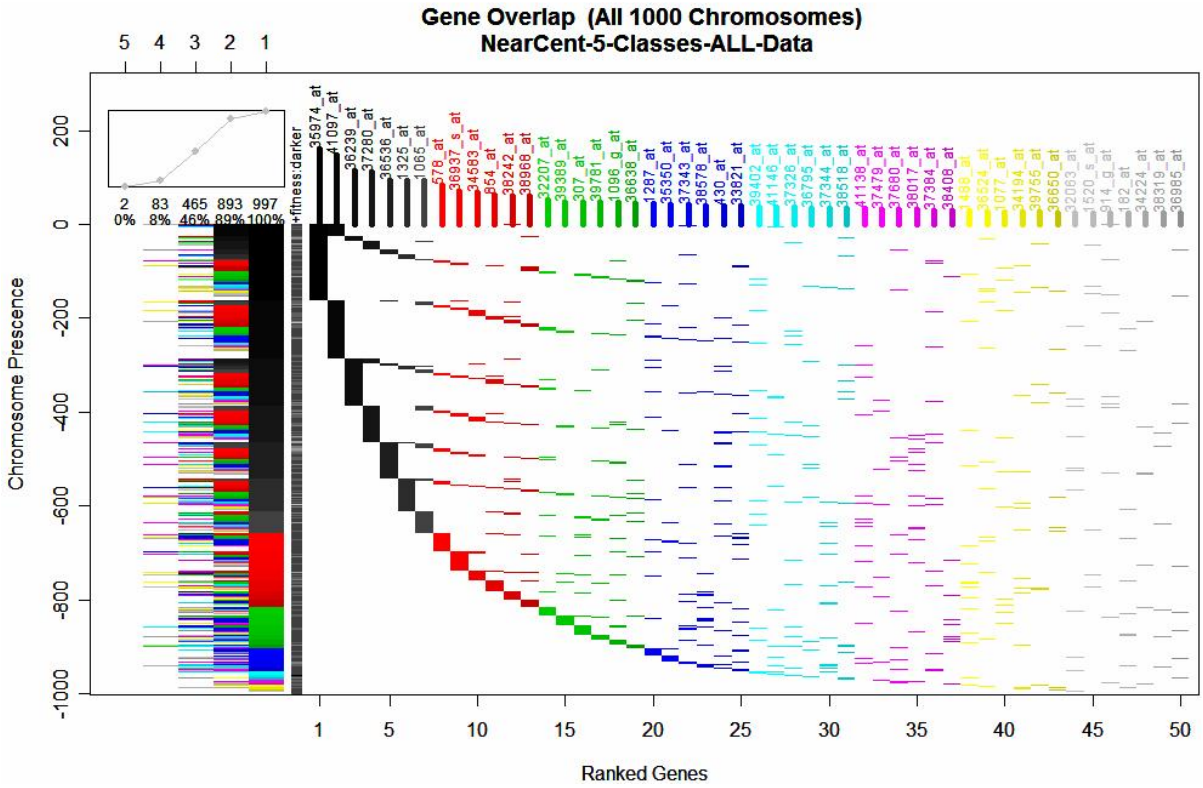


Figure 3.16 - Overlapped genes in models.

3.4.7 - Predicting Class Membership of Unknown Samples

An important characteristic of any model is their ability to make predictions. Models designed using GALGO can be evaluated with a complete unknown or blind dataset. The following code exemplifies how to make predictions in a new "dummy" dataset for all chromosomes collected in the BigBang object.

```
> data(ALL)
# dummy data: the first 15 samples from original ALL data
# which all must be from EMLLA class
> dummy <- ALL[,1:15]
> ?predict.BigBang
> cpm <- predict(bb.nc, newdata=dummy,
func=classPredictionMatrix, splits=1:10)
> cpm
> plot(bb.nc, cpm, type="confusion")
```

In the above code, *dummy* was temporarily appended to the original data. Then *classPredictionMatrix* was run for all chromosomes. *splits* is a parameter used in *classPredictionMatrix* (which was used to illustrate the use of user-parameters for any function specified in *func*). The result of the plot is shown in Figure 3.17 where the new data were labelled as "UNKNOWN". The black bars in these samples indicate that they were predicted as EMLLA (as expected).

To predict new data using an individual model, one may use the *classPredictionMatrix* method using the *chromosomes* parameter (see ?classPredictionMatrix.BigBang), such as in the following code.

```
> cpm <- predict(bb.nc, newdata=dummy,
func=classPredictionMatrix, chromosomes=fsm$models[1])
> cpm
> plot(bb.nc, cpm, type="confusion")
```

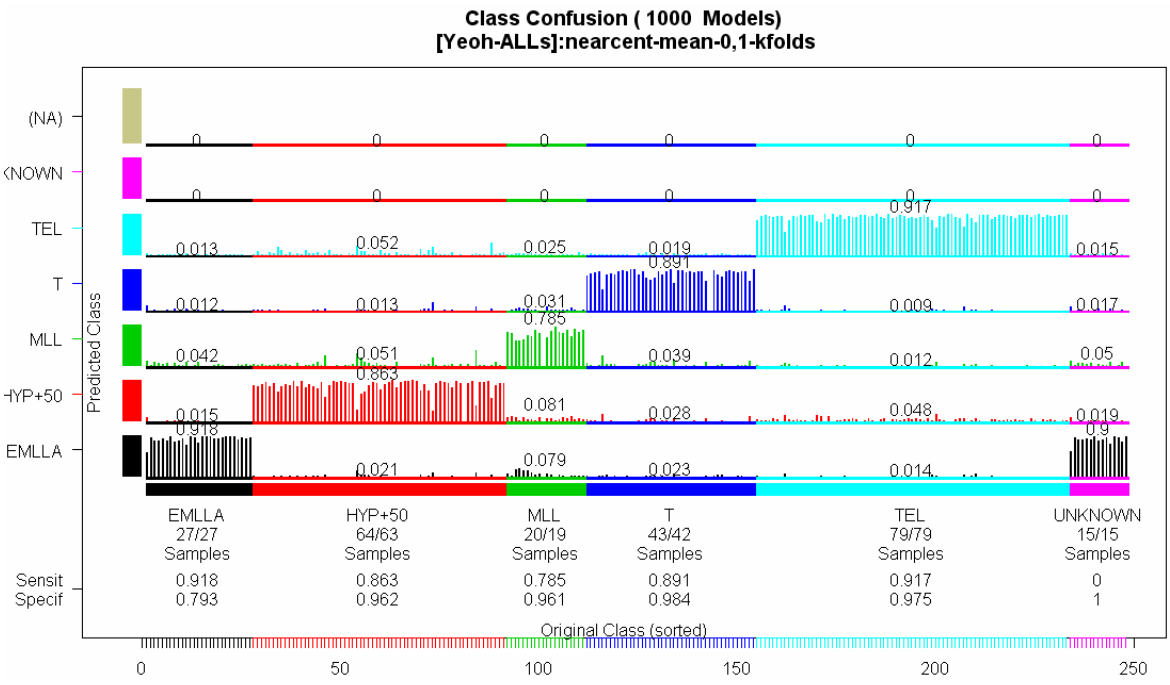


Figure 3.17 - Prediction for unknown samples (the last 15 samples in the right).

3.4.8 - Tutorial Summary

The *configBB.VarSel* configures the necessary objects and specifies the right parameters to make the entire process to work in different contexts and testing strategies with the classification method of your choice. In addition, the implementation of new classification methods is simplified providing your specific fitness function in the *classification.userFitnessFunc* parameter (type ?configBB.VarSel).

It has been shown in this section how to build multivariate statistical models for a classification problem using GALGO. So far, a basic analysis with the dataset included has been performed. The GALGO manual contains more advanced analysis explaining many of the available options in each step that can be customized for particular data, classification methods, GA searches, user defined fitness functions, error estimation, process parallelization, GA parameters, and troubleshooting.

3.5 - A Comparison between GALGO and Univariate Variable Selection Methods

To show the utility of GALGO, a comparison of a UVS strategy (Figure 2.15) using F-statistic and d-statistic versus the MVS implemented in GALGO has been performed (Figure 3.1). This comparison includes a number of classification methods in UVS and MVS. The models developed have been analysed in respect to classification accuracy, number of genes required to achieve the highest classification accuracy and the identity of the genes selected in the models. In order to make sure that the comparison is of general validity three different datasets have been used.

The results support the use of MVS in developing statistical models and in particular support the use of GALGO as a general software environment for model selection.

3.5.1 - Methods

3.5.1.1 - Variable selection

In this comparison three variable selection strategies have been used. These are: The F-test, d-statistics and GA.

3.5.1.2 - Classification methods

F-statistics and GA have been compared with the following methods: 1) Diagonal Linear Discriminant Analysis (DLDA)[1], 2) Support Vector Machines (SVM), 3) Random Forest (RF) and 4) K-Nearest-Neighbours (KNN). Another well established tool PAM [see reference 73 and section 2.6.1.2 in Chapter 2] that uses a *d*-statistics based in centroid distances in combination with shrunken nearest centroids was also compared versus GA in combination with nearest centroid (NC).

3.5.1.3 - Construction of a representative model

To generate a representative model, a Forward Selection (FS) strategy in both multivariate and univariate model selection was used (Figures 2.15, 2.17, and Figure 3.1). Briefly, an initial model is created using the first two genes from an ordered list of genes (using p-values or d-statistic for UVS or gene frequency using GALGO). Then, the model is assessed using a classification method to estimate the classification error. Subsequently, the model is lengthened with the next gene in the ordered list and the resulting model is re-assessed. This cycle continues until all genes in a list have been

[1] The MLHD method implemented in GALGO is equivalent linear discriminant analysis (LDA).

included. The model whose classification error is the lowest is chosen. In the case of a draw, the smallest model is selected.

3.5.1.4 - Method-specific gene signatures

To determine the degree of overlap between the different models at a gene identity level, a pool gene set containing the genes from the models generated with all five methods have been built. For each gene in a given model, the number of times it appears in the pool set was counted. Genes appearing only once were defined as model-specific.

3.5.1.5 - Implementation

To develop classifiers with a UVS strategy (F- or d-statistics), the web-based tool TNASAS [60] (http://tnasas.bioinfo.cipf.es), which is part of the Gene Expression Pattern Analysis Suite (GEPAS, http://gepas.bioinfo.cipf.es) have been used. All classification methods tested in combination with GA have been used in the GALGO implementation with default settings with the addition of a backward elimination step for model enhancement. The representative models were developed from 1,000 chromosomes.

3.5.2 - Datasets

3.5.2.1 - ALL-Subclasses Dataset (ALLS)

This dataset, developed by Yeoh *et al.* [134], describes the expression profile of 327 acute lymphoblastic leukaemia (ALL) patients representing 7 different disease sub-classes. The authors have used Affymetrix GeneChips. In this comparison the five largest classes were selected (EMLLA, Hyp+50, MLL, T, and TEL including respectively 27, 64,

20, 43, and 79 samples). The original dataset, downloaded from http://www.stjuderesearch.org/data/ALL1/ comprising 12,600 genes, have been filtered to eliminate the most invariant genes. Briefly, the standard deviation and difference between maximum and minimum expression values were calculated for each gene. The genes were ranked by these values, and if they were within the top 15% for either, were selected for further analysis. The dataset after filtering contained the expression values for 2,435 genes.

3.5.2.2 - ALL-AML Dataset (ALL/AML)

This dataset developed by Golub *et. al.* [7] describes the transcriptional state of 47 acute lymphoblastic leukaemia (ALL) and 25 acute myeloid leukaemia (AML) patients. Data were processed as in the original publication. Briefly, intensity values were re-scaled such that overall intensities for each chip are equivalent. This was done by fitting a linear regression model using the intensities of all genes with "P" (present) calls in both the first sample (baseline) and each of the other samples. The inverse of the "slope" of the linear regression line becomes the (multiplicative) re-scaling factor for the current sample. This was performed for every chip (sample) in the dataset except the baseline which gets a re-scaling factor of one. A further processing step was performed to eliminate genes that were not detected in the majority of the samples. For this reason every gene that was not expressed (Flagged as M or A) in more then 80% of the samples was filtered out.

3.5.2.3 - Breast Cancer Dataset (BC)

This dataset was developed by van't Veer *et al.* [135] representing 78 patients subdivided in two groups with different clinical outcomes in which 44 patients with no metastases developed within the first five years versus 34 patients positive for

metastases within the first five years. Data were normalized as described in the original publication. Genes were filtered out when the confidence level that a gene's mean ratio is significantly different from 1 were larger than 0.001 using all samples.

3.5.3 - Results

Table 3.1, Table 3.2, and Table 3.3 show the results of the analysis performed using GALGO with five different classification methods (NC, KNN, SVM, MLHD, RF) on three datasets (BC, ALL/AML, ALLS). The tables report the classification accuracy and model size for the best representative models developed using forward selection and for the top five individual chromosomes selected by the GA search. These tables show that in all cases GALGO can identify accurate models of a relatively small size. Figure 3.18, Figure 3.19, and Figure 3.20 summarize the results of the comparison between GALGO and UVS strategies.

In the *Breast Cancer* dataset (Table 3.1 and Figure 3.18) GALGO produced models with higher classification accuracy regardless of the classification method used. The error was diminished by about the half. Except for one case (RF), the size of the models developed with GALGO was comparable or smaller than the models developed with the UVS strategy (Figure 3.18). The largest difference was observed in models developed with the KNN method (these require 2920 genes with univariate model selection and 31 genes with GALGO).

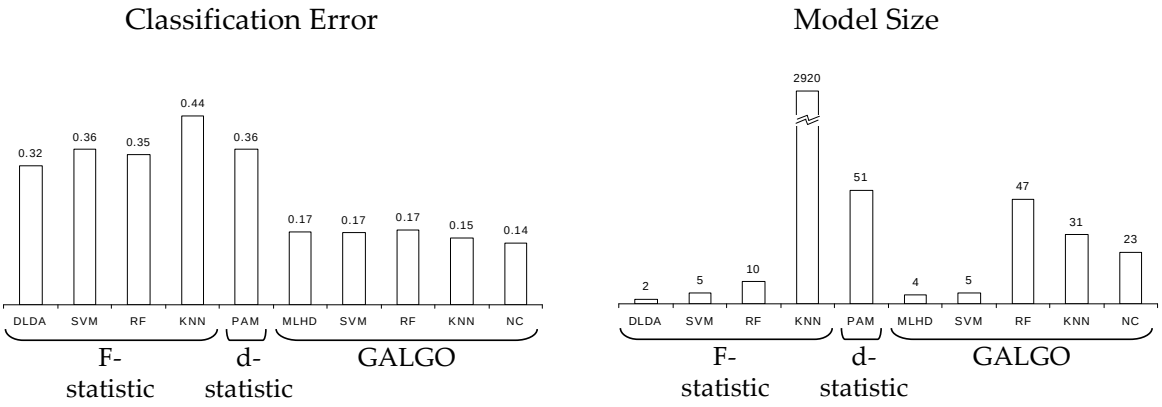


Figure 3.18 – Results from Breast Cancer dataset.

Table 3.1 – GALGO results for Breast Cancer dataset. GA – Genetic Algorithms, FS – Forward Selection, BE – Backward Elimination. DLDA – Diagonal Linear Discriminant Analysis, PAM – Shrunken Centroids, PAMR – Shrunken Centroids R package, KNN – K-Nearest-Neighbours, SVM – Support Vector Machines, NC – Nearest Centroid, MLHD – Maximum Likelihood Discriminant Functions, RF – Random Forest.

BREAST CANCER (2 Classes)										
Method	*KNN*		*SVM*		*NC*		*MLHD*		*RF*	
Model	*Size*	*Error*	*Size*	*Error*	*Size*	*Error*	*Size*	*Error*	*Size*	*Error*
GA+FS 1st	32	0.16	12	0.17	35	0.15	4	0.18	47	0.17
GA+FS 2nd	33	0.16	9	0.18	11	0.15	-	-	37	0.18
GA 1st	5	0.20	5	0.17	5	0.18	5	0.17	5	0.18
GA 2nd	5	0.21	5	0.18	5	0.19	5	0.18	5	0.24
GA 3rd	5	0.22	5	0.19	5	0.19	5	0.18	5	0.24
GA 4th	5	0.22	5	0.19	5	0.19	5	0.19	5	0.25
GA 5th	5	0.22	5	0.20	5	0.19	5	0.19	5	0.25
GA+BE+FS 1st	31	0.15	12	0.17	23	0.14	4	0.18	40	0.18
GA+BE+FS 2nd	32	0.15	-	-	9	0.15	-	-	14	0.19
GA+BE 1st	5	0.21	5	0.17	3	0.17	4	0.17	4	0.18
GA+BE 2nd	3	0.21	4	0.17	4	0.17	3	0.17	3	0.23
GA+BE 3rd	4	0.21	2	0.18	3	0.18	4	0.17	2	0.24
GA+BE 4th	4	0.21	2	0.18	5	0.18	4	0.18	4	0.24
GA+BE 5th	3	0.22	2	0.18	4	0.18	3	0.19	3	0.24

In the *ALL-AML* dataset (Table 3.2 and Figure 3.19) the classification accuracy of models developed with univariate and multivariate models was comparable (GALGO gave models in the range between 3% and 10% of error whereas the univariate methods gave models with error in the range between 3% and 7%). However, the models developed

using GALGO were markedly smaller in size (a range of 4 to 49 genes respect to 79 to 1697 in the UVS).

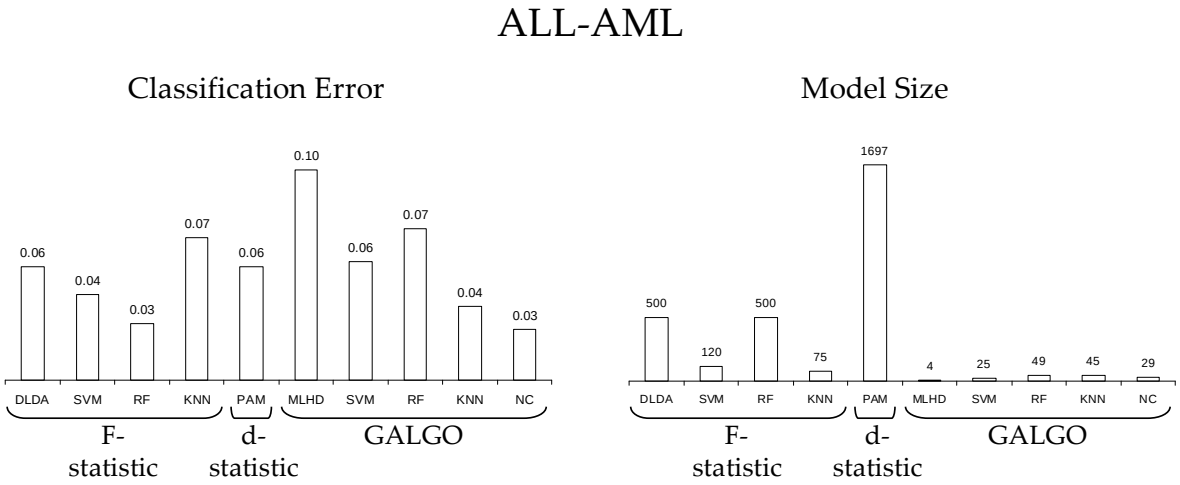


Figure 3.19 – Results from ALL-AML dataset.

Table 3.2 – GALGO results for ALL-AML dataset. Abbreviations as in table 1.

ALL-AML Dataset (2 Classes)										
Method	*KNN*		*SVM*		*NC*		*MLHD*		*RF*	
Model	*Size*	*Error*	*Size*	*Error*	*Size*	*Error*	*Size*	*Error*	*Size*	*Error*
GA+FS 1st	42	0.06	50	0.07	37	0.05	9	0.14	47	0.08
GA+FS 2nd	37	0.06	23	0.07	24	0.06	17	0.14	45	0.08
GA 1st	5	0.11	5	0.07	5	0.13	5	0.10	5	0.12
GA 2nd	5	0.12	5	0.11	5	0.15	5	0.10	5	0.15
GA 3rd	5	0.13	5	0.11	5	0.15	5	0.11	5	0.15
GA 4th	5	0.13	5	0.12	5	0.15	5	0.12	5	0.16
GA 5th	5	0.13	5	0.13	5	0.15	5	0.12	5	0.16
GA+BE+FS 1st	45	0.04	25	0.06	29	0.03	13	0.12	49	0.07
GA+BE+FS 2nd	40	0.05	24	0.06	27	0.03	34	0.13	32	0.08
GA+BE 1st	3	0.08	4	0.07	2	0.12	4	0.10	2	0.12
GA+BE 2nd	3	0.09	3	0.11	2	0.12	5	0.10	4	0.14
GA+BE 3rd	3	0.11	5	0.11	5	0.13	4	0.11	4	0.15
GA+BE 4th	3	0.11	3	0.12	4	0.14	4	0.11	5	0.15
GA+BE 5th	4	0.12	3	0.12	3	0.15	2	0.11	4	0.16

In the *ALLS* dataset (Table 3.3 and Figure 3.20) GALGO generated either model with comparable accuracy (the maximum difference in classification accuracy was 2%) or higher accuracy with respect to univariate models (1% against 17% using RF and 1%

against 13% with NC). As in the other datasets the size of the models developed using GALGO was markedly smaller than models developed with univariate methods (the range of model size in this dataset was between 4 and 49 whereas the range of model size in the univariate selected models was between 75 and 1697).

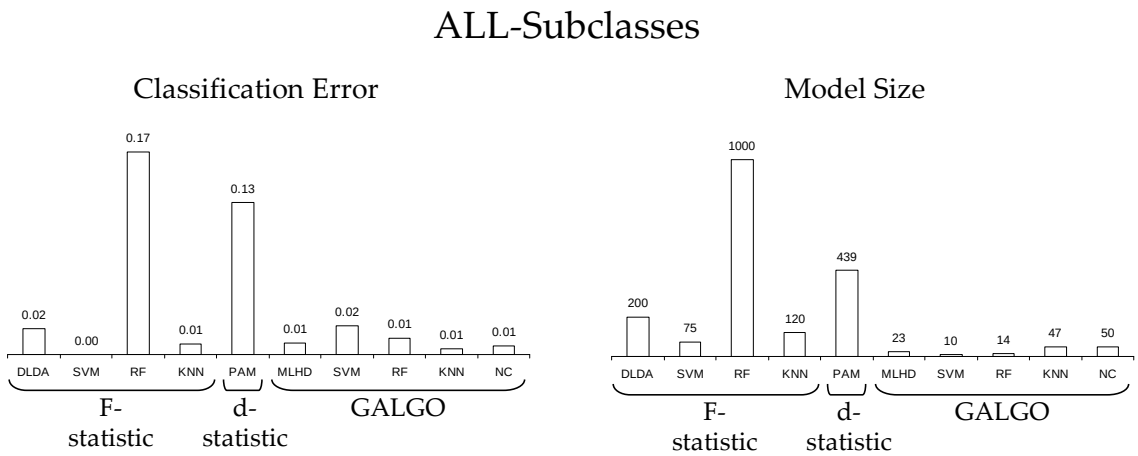


Figure 3.20 – Results from ALL dataset.

Table 3.3 – GALGO results from ALL-Subclasses dataset. Abbreviations as in table 1.

ALL-Subclasses Dataset (5 Classes)

Method	*KNN*		*SVM*		*NC*		*MLHD*		*RF*	
Model	*Size*	*Error*	*Size*	*Error*	*Size*	*Error*	*Size*	*Error*	*Size*	*Error*
GA+FS 1st	47	0.00	10	0.02	50	0.01	23	0.01	14	0.01
GA+FS 2nd	13	0.01	9	0.03	16	0.02	15	0.02	10	0.02
GA 1st	5	0.06	5	0.05	5	0.06	5	0.06	5	0.08
GA 2nd	5	0.06	5	0.05	5	0.07	5	0.06	5	0.08
GA 3rd	5	0.06	5	0.05	5	0.07	5	0.06	5	0.08
GA 4th	5	0.06	5	0.06	5	0.07	5	0.06	5	0.08
GA 5th	5	0.06	5	0.06	5	0.07	5	0.06	5	0.09
GA+BE+FS 1st	47	0.00	10	0.02	50	0.01	20	0.01	19	0.01
GA+BE+FS 2nd	13	0.01	9	0.03	16	0.02	15	0.02	10	0.02
GA+BE 1st	4	0.06	5	0.05	5	0.06	5	0.06	4	0.08
GA+BE 2nd	5	0.06	5	0.05	4	0.07	5	0.06	4	0.08
GA+BE 3rd	4	0.06	5	0.06	5	0.07	5	0.06	4	0.08
GA+BE 4th	5	0.06	5	0.06	5	0.07	4	0.06	5	0.09
GA+BE 5th	4	0.06	4	0.06	5	0.07	5	0.06	5	0.09

In two datasets the model size is dramatically different making obvious that multivariate models are very effective in identifying different gene subsets. In the Breast Cancer dataset, gene sets are of a more comparable size. However, in Breast Cancer the error was remarkably smaller.

Table 3.4 summarizes the overlap in gene composition of the models developed with the different methods in the Breast Cancer dataset. These results suggest that multivariate model selection tend to give different gene subsets respect to the UVS strategy. In interpreting these results however it should be taken into account that the classification error of models developed from UVS strategies was very high.

Table 3.4 – Method-Specific genes. The table shows the genes that have been selected by only one method, hence these genes could not be obtained by the other methods.

Method	*Model Size*	*Method-Specific Genes*	*Percent of Method-Specific Genes*
F+DLDA	2	0	0%
F+SVM	5	0	0%
F+RF	10	0	0%
F+KNN	2920	-	-
d+PAM	51	38	75%
GA+BE+MLHD	4	2	50%
GA+SVM	5	3	60%
GA+FS+RF	47	32	68%
GA+BE+FS+KNN	31	11	35%
GA+BE+FS+NC	23	11	48%

3.6 - Conclusion and Discussions

GALGO is a user-friendly R package designed for developing multivariate statistical models using large-scale "omics" data. In the context of MVS in large-scale datasets GALGO performs well and does not require any coding. For a more general use, its

object-oriented structure allows the definition of new methods by simply recoding the fitness function. GALGO allows the development and analysis of statistical models using a unique wrapping function. These characteristics make GALGO an ideal environment for both Bioinformaticians and computer minded biologists. The availability of a very broad spectrum of R libraries with general statistical (CRAN) or with specific machine learning functionality (such as MLinterfaces and ipred) makes GALGO an ideal prototyping environment for any analysis method that utilizes GA as a search strategy

The models developed have been analysed in respect to classification accuracy, number of genes required to achieve the highest classification accuracy, and the identity of the genes selected in the models. All these factors are important in determining the usefulness of variable selection methodologies. High classification accuracy is obviously a very desirable property but in order for the models to be biologically interpretable and of practical use, it is also important that the gene set is a manageable size. The identity of the genes is also a very important factor. One of the reasons why multivariate methods may be a good option is that they allow the identification of genes that contribute to a biological effect in association. These could not be discovered by UVS methods where every gene is tested in isolation. If univariate and multivariate approaches provide models with comparable classification accuracy but with different genes then the two approaches have to be considered complementary as they are likely to represent different underlying biological processes.

It will be shown in Chapter 5 that GALGO also produces similar multivariate models displaying comparable accuracies than those generated by a Markov Chain Monte Carlo using a probit classifier. In addition, GALGO provides more flexibility in classifiers, GA search, and user interface than other tools published so far.

The results shown here suggest that the methodology implemented in the R package GALGO tends to produce models with comparable or better classification accuracies than UVS strategies. The multivariate selected models generally use a smaller number of genes than univariate models in all datasets and methods tested. These results support the use of a multivariate model selection strategy in the analysis of FGD and in particularly support GALGO as a general tool.

In conclusion, GALGO is a valuable, robust, and easy to use tool for developing multivariate statistical models using multivariate variable selection.

CHAPTER 4

The Application of GALGO to Biomarker Discovery in Proteomics and Metabolomics

Genomic technologies generate large datasets for a few samples. Besides Transcriptomics, other "omics" such as Proteomics and Metabolomics are increasingly being used as research tools for functional genomics which require computational tools for data analysis. The selection of features such as transcripts, proteins, or metabolites, related to sample classes is an important task to design novel clinical tests and to guide further research revealing some of the biological components. This Chapter will show two successful multivariate variable selection analyses based on the proteomics of Rheumatoid Arthritis and metabolomics of Vitreoretinal Disease.

4.1 - Introduction

Methods previously used in transcriptomics will now be applied to proteomics and metabolomics data. In general, the work will focus on multivariate variable selection using GALGO (see Chapter 3) to analyse two case studies, which are described next.

4.1.1 - Case 1: Early Rheumatoid Arthritis

Arthritis is a disease in which joints are chronically inflamed. The inflammation is commonly accompanied by pain, swelling, and stiffness. There are a several types of Arthritis. One of these is Rheumatoid Arthritis (RA) which is considered a chronic,

inflammatory autoimmune disorder. In RA, joints are attacked by the immune system causing painful joint inflammations. To diagnose RA, several clinical and molecular factors have been used. Two of the most common molecular factors are the presence of rheumatoid factor (RF) and Cyclic Citrullinated Peptide antibody (CCP) in blood. It is known that individuals who are positive for both RF and CCP will suffer or are actually suffering RA. However, RF and CCP are both positive only in around 50% of individuals that suffer RA. Due to the inflammation caused by the immune response, several other related molecular factors are also detected in RA such as tumour necrosis factor alpha (TNF-α), interleukins IL-1, IL-6, IL-8 and IL-15, transforming growth factor beta, fibroblast growth factor and platelet-derived growth factor. Whether any of these molecules or combinations of them is better predictors of RA needs to be demonstrated. In this context, Dr. Karim Raza (IBR, Medical School, University of Birmingham, UK) [136] has designed an experimental setup that includes performing a number of clinical tests along with a proteomic approach measuring levels of more than 30 cytokines from individuals at the very early stages of RA. More precisely, samples are taken within a few weeks from the first referral to the general practitioner when it is impossible to predict the disease associated with the symptoms. Patients are followed in the subsequent months until their test is clear (typically after 6 months from referral). The aim of our modelling approach has been to develop models based on a combination of blood cytokines that are predictive of disease outcome. Our results are encouraging and show that markers other than CCP and RF can be developed that are more sensitive and specific than the currently employed CCP and RF autoantibody test. Furthermore, a model could be designed using an identified marker together with CCP and RF that are also more sensitive.

4.1.2 - Case 2: Vitreoretinal Disease

Uveitis is the inflammation of the uvea sited in the middle layer of the eye. The uvea includes the iris, ciliary body and choroid. Vitreoretinal Disease (VD) like other forms of uveitis or inflammatory disease lacks specific biomarkers that define either disease type or response to treatment. A range of possible endpoints is used to define outcome, but it is not clear how these relate to each other in different patients or studies [137]. This is particularly relevant for clinical trials when comparing a novel treatment to established therapy. Several studies have identified inflammatory mediators, such as cytokines, chemokines and growth factors or single nucleotide polymorphisms with disease type, activity, and response to treatment; however, there is no clear result that can relate to clinical response [138-142]. In this study, Dr. Stephen Young has used functional genomics analysis of metabolite fingerprints (metabolomics) in vitreous humour of patients with VD using high-resolution ^{1}H-nuclear magnetic resonance spectroscopy. The goal is the identification of metabolites that may help in the prognosis of VD. The results show that a few putative metabolites could be good predictors even though a small number of samples were used. The identified segments of the spectra (bins) were method-specific. Even though consecutive bins were treated independently by the multivariate search, a number of selected bins were agglomerated in the neighbourhood suggesting that these bins represent the same putative metabolite.

4.2 - Results

4.2.1 - Case 1: Early Rheumatoid Arthritis

It is known that the presence of CCP and RF autoantibodies are related to RA prognosis. When used in combination, their specificity is close to 100%; however the sensitivity is around 50% only [143; 144]. Our aims are therefore to identify groups of cytokines from the 30 cytokines measured that could be better predictors of RA and propose prognostic tests including these identified cytokines. Preliminary inspection of clinical data available for this study confirmed that, as expected, 50% of RA patients recruited in this study are RA and CPP double positives (Figure 4.1A). Also, as expected, only 11 samples out of 25 that are diagnosed as RA are positive for RF and CCP simultaneously (Figure 4.1B). To identify cytokines, univariate and multivariate variable selection procedures were performed. Univariate results suggest that several cytokines may also be related to RA outcome (Figure 4.2). However, the multivariate variable selection used chooses only a few variables among a number of classifier machines (Table 4.1). In general, the most recurrent cytokine was IL-1ra which was accompanied by TNFα in a couple of classifiers. Not all classifiers were able to produce

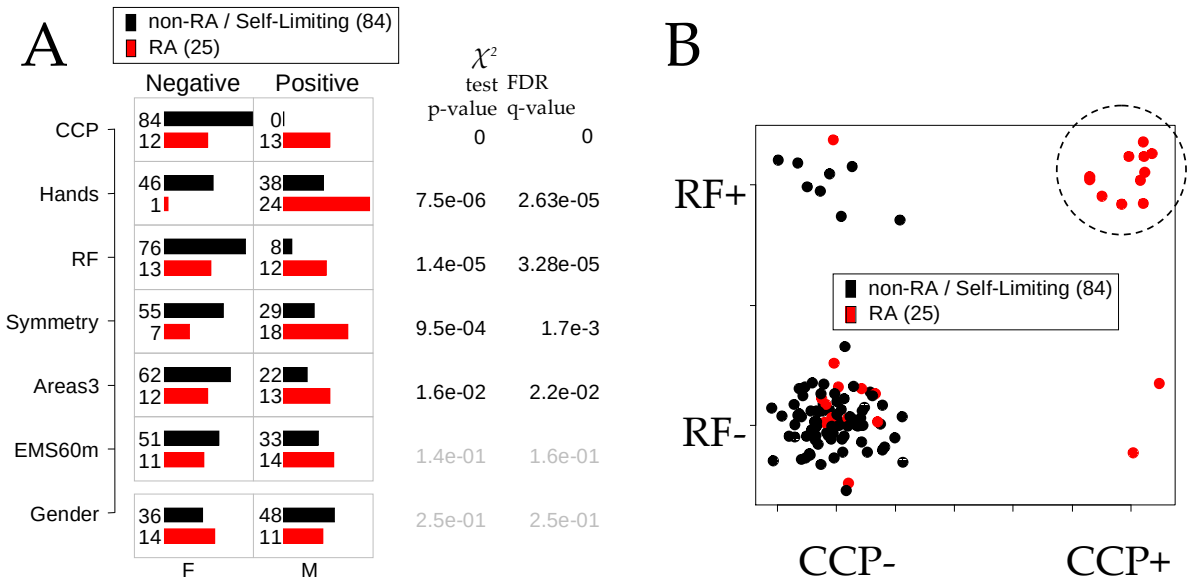


Figure 4.1 – Overview of clinical information related to RA outcome. (A) Clinical variables associated to RA along with their significant test p-value. Some of these variables are the basis for RA diagnosis. P-values correspond to contingency Chi-square test and its subsequent FDR correction. (B) RF and CCP clinical tests for each individual. Labels indicate RA outcome (in red). Individuals with both RF and CCP positive are indicated inside dotted circle. Axes in arbitrary units, high values are clinically considered positive otherwise negative.

acceptable accuracy though. RF, CT and KNN produced the most accurate models. In the following however, the study has been focused on CT and KNN because these methods would be easier to interpret and implement in clinical practice.

Table 4.1 – Summary of multivariate models designed using cytokines levels in the RA dataset. Sensitivity and Specificity are estimated relative to RA class. ROC area is the product of sensitivity and specificity. In general, higher ROC areas indicate better predictive models. Models were developed using GALGO. KNN – K-nearest-neighbours, MLHD – Maximum likelihood discriminant functions, RF – Random Forest, SVM – Support Vector Machines, CT – Classification Trees.

Classifier	*Variables used in model*	*Sensitivity*	*Specificity*	*ROC area*
KNN	**IL-1ra**, IL-2R, TNFα	0.83	0.88	0.73
MLHD	**IL-1ra**, MCP-1, MMP-3	0.70	0.88	0.62
RF	TNFα, **IL-1ra**	0.85	0.89	0.76
SVM	MMP-13, Eotaxin	0.45	0.80	0.36
CT	**IL-1ra**, IL-2	0.81	0.92	0.75

From the 109 samples, 87 samples are both CCP and RF negative in which 11 samples are actually diagnosed as RA (Figure 4.1B). In these 11 individuals, CCP and RF were not able to provide any information about the outcome. However, cytokine levels are still differentially detected between RA and non-RA individuals whose CCP and RF were negative (Figure 4.3). Therefore, once it has been shown that acceptable multivariate models could be designed using cytokine levels in the overall RA population (Table 4.1), the next question to answer was whether predictive models can be built using cytokine levels in a reduced dataset where CCP and RF are both negative. Results are shown in Table 4.2. KNN found good models whereas CT could not design predictive models. Repeated KNN runs yielded models where IL-1ra was always present but the inclusion of other variables was unstable which suggest that variables other than IL-1ra were dependent on the specific data split set for training. Although this last result indicates that perhaps more RA samples would be needed, in general, results indicate that IL-1ra may be an important prognostic factor even when CCP and RF are negative.

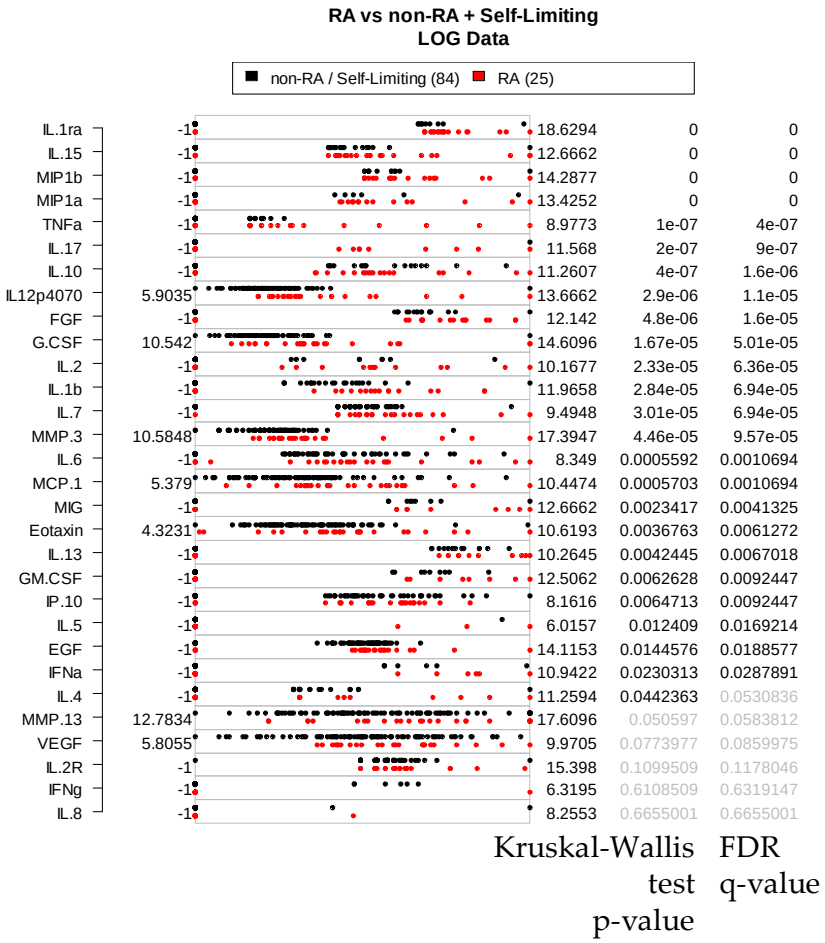


Figure 4.2 – Univariate test for association of cytokines levels to RA outcome. The figure shows that several cytokines are associated to outcome. Minimum and maximum values are indicated. Values in log-scale, zero values before log where set to -1. A non-parametric test p-value and its corresponding FDR correction are shown.

Table 4.2 – GALGO results for reduced RA dataset where CCP and RF are negative.

Classifier-Run	*Variables used in model*	*Sensitivity*	*Specificity*	*ROC area*
KNN-1	IL-1ra, IL-6			
KNN-2	IL-7, IL-1ra (best)	0.87	0.85	0.74
KNN-3	MMP-3, IL-1ra			
KNN-4	IL-10, IL-5			
CT-1…10	IL-7, IL-8, IL-5	1	0	0.00

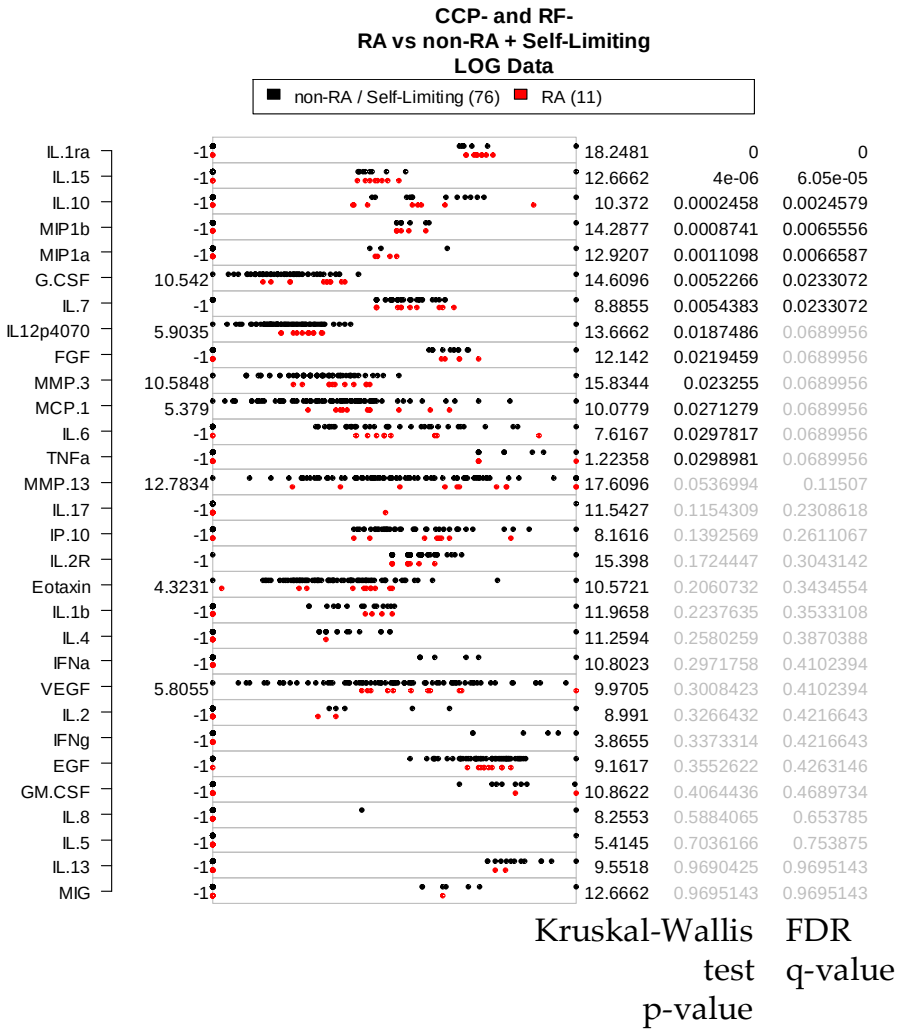


Figure 4.3 – Univariate test for association of cytokines levels in samples whose CCP and RF are negative. The figure shows that only a subset of the cytokines is related to RA outcome for samples where CCP and RF are negative. Minimum and maximum values are indicated. Values in log-scale, zero values before log where set to -1. A non-parametric test p-value and its corresponding FDR correction are shown.

The results shown above suggested that cytokine levels, especially IL-1ra, are able to be predictive of RA regardless of the RF and CCP outcome.

Then, the study was focused on whether a model could be build considering both clinical and cytokine variables. GALGO results using this combined dataset are shown in Table 4.3. In these models, IL-1ra, RF, and CCP were preferable selected. In models designed by GALGO, IL-1ra was surprisingly more frequent than the usual RF and

CCP markers (Figure 4.4). Data inspection supports the combined predictive power of IL-1ra, RF, and CCP (Figure 4.5).

Table 4.3 – Results for the RA dataset where clinical information and cytokine levels are both considered. Unmarked models were designed using GALGO whereas. For comparison purposes, some models were built manually which are marked with a star (*).

Classifier	*Variables used in model*	*Sensitivity*	*Specificity*	*ROC area*
KNN	**IL-1ra**, RF, CCP	0.87	0.91	0.79
KNN*	RF, CCP	0.52	1.00	0.52
CT	**IL-1ra**	0.82	0.91	0.75
CT*	IL-1ra, RF, CCP	0.82	0.92	0.75
CT*	RF, CCP	0.50	0.99	0.50

Altogether, these results suggest that IL-1ra is an important RA prognostic factor which can also be used in combination with RF and CCP to increase RA prediction. Consequently, a rule using these three variables was finally designed. This rule increases sensitivity from 44% to 88% at only 6% decrease in specificity (Figure 4.6). Since protein expression data were obtained with a relatively large scale technique these results will need to be validated by more robust methodologies (for example ELISA assays) that could also be more appropriate for applications in the clinic.

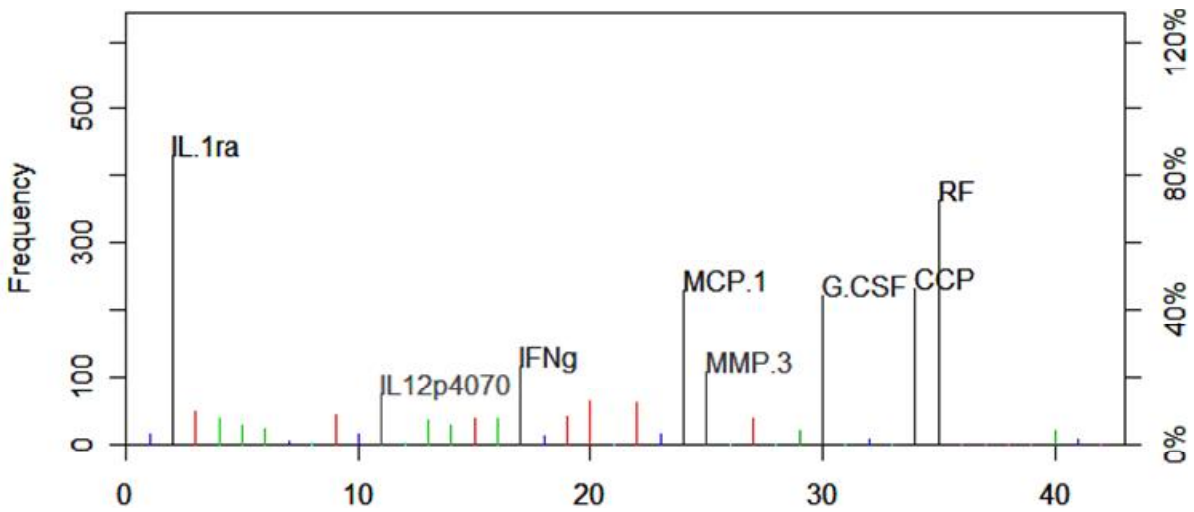


Figure 4.4 – Frequency for each variable when both clinical information and cytokines levels are considered. Representative results estimated from 500 KNN models.

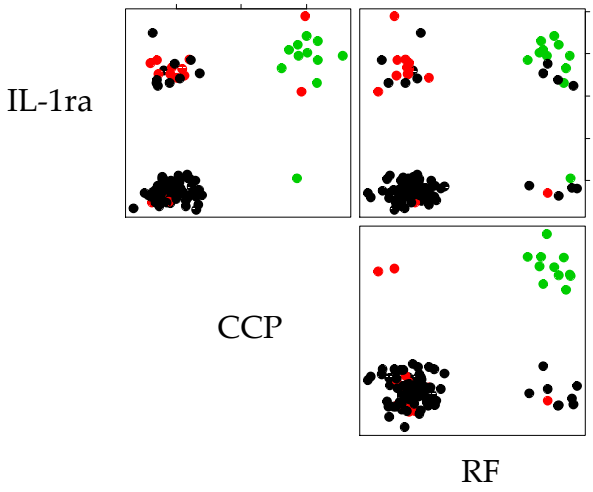
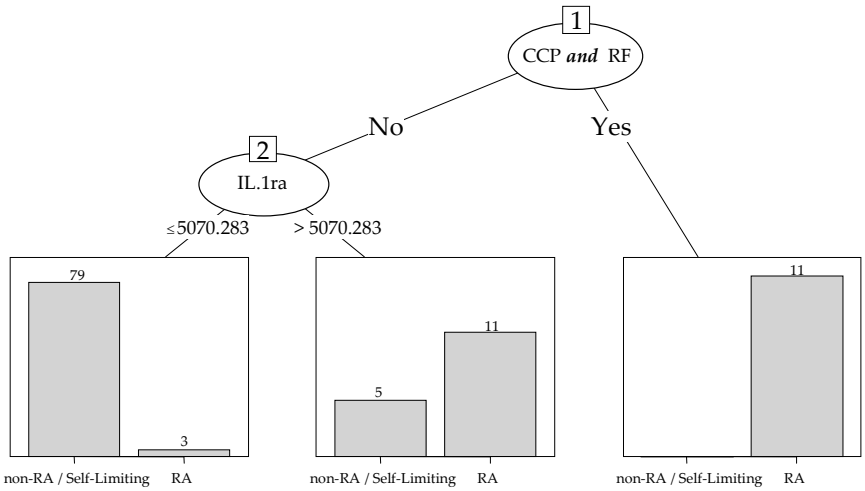


Figure 4.5 – Values of IL-1ra, CCP, and RF. Each dot represents a sample. Green dots are samples diagnosed as RA and whose CCP and RF test were both positive. Red dots stand for the remaining samples also diagnosed as RA. Black dots symbolize samples not diagnosed as RA. The vertical axis depends on IL-1ra in top panels and on CCP in bottom panel. The horizontal axis depends on RF in the right panels and on CCP in the left panel.



Prediction:	*non-RA or SL*	*RA*	*RA*
Sensitivity:	*94%*	*79%*	*44%*
Specificity:	*96%*	*69%*	*100%*

Calling RA Rule: (CCP and RF) or (IL-1ra > 5070) { **Sensitivity: 88% (22/25)**, **Specificity: 94% (1-5/84)**, **ROC Area: 0.83**

Figure 4.6 – Designed rule adding IL-1ra as RA prognostic factor. For RA prognosis, the first and common condition is whether CCP and RF are both positive. If this condition is false, the second condition decides the outcome, if IL.1ra is "positive" (greater than 5070 before log transformation), the prediction is RA otherwise it is designated as non-RA.

4.2.2 - Case 2: Vitreoretinal Disease

There are no definitive biomarkers to diagnose VD. Dr. Young *et al.* [145] used nuclear magnetic resonance spectroscopy to obtain metabolomic profiles from patients with chronic uveitis (CU) and lens-induced uveitis (LIU). The aim is therefore to select "bins" from the metabolic profile that could discriminate between CU and LIU patients. A "bin" is the result of the integration of a continuous segment of the original NMR spectrum (see Methods sections). The result is a smoothed spectrum (top profiles in Figure 4.7) formed by bins. The VD dataset consists of 1,960 bins per sample which will be considered here as variables. 18 samples were used, 10 for CU and 8 for LIU. Two

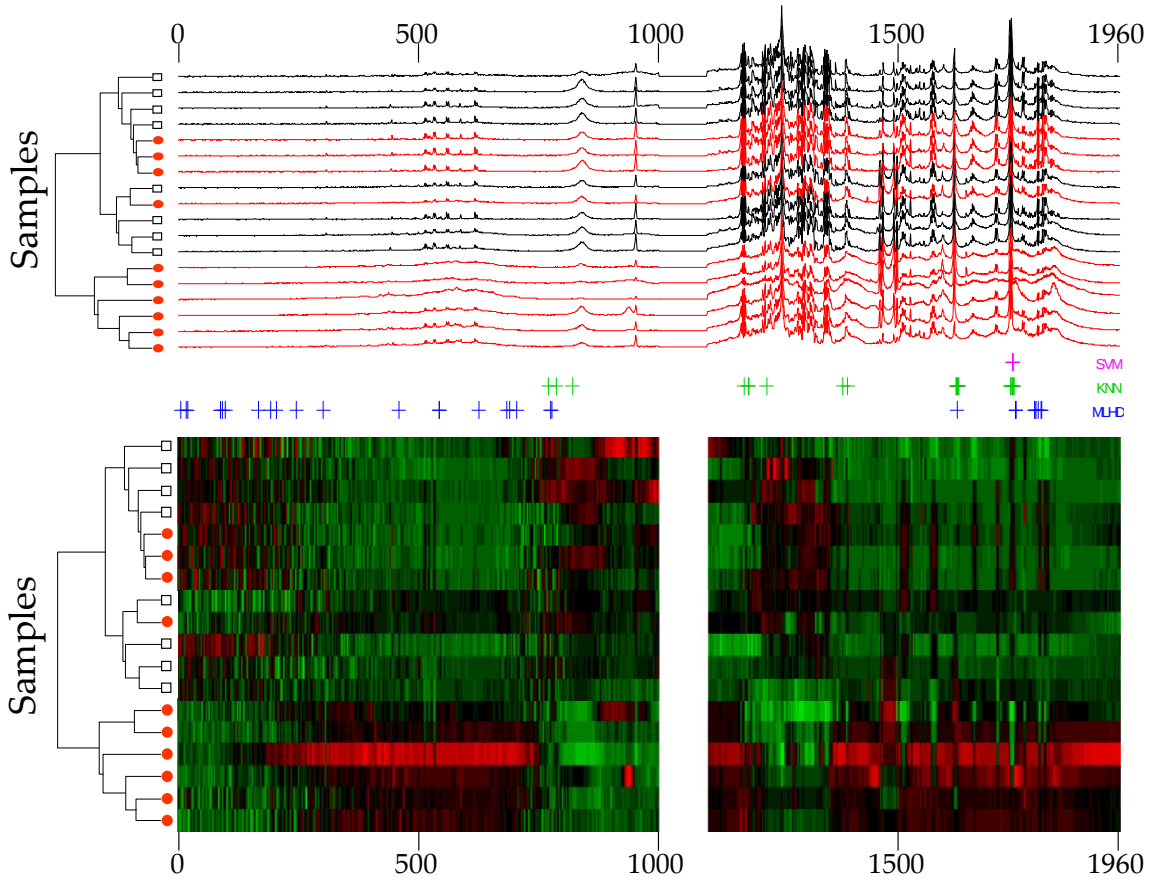


Figure 4.7 – Metabolomic profiles overview (two representations). Top panel show the raw metabolic profile (1960 bins in horizontal) whereas the bottom panel shows the classical heatmap representation. Coloured marks (+) in the middle of the figure represent selected bins from a multivariate selected representative model using the labelled classifier (SVM - purple, KNN - green, MLHD - blue). Heatmap values were mean centred, red (darker) represent higher values whereas green colours (lighter) represent lower values. The white band around 1000 in the heatmap corresponds to water region that is commonly removed before the analysis.

Table 4.4 – Representative models for Vitreoretinal Disease obtained using three classifier machines. Bins are shown ordered by their frequency in pooled models (the most frequent is shown first).

Classifier	*Pooled models*	*Bins used in representative model (number of bins)*	*Sensitivity-Specificity*	*ROC area*
KNN	2199	1733,1738,1740,1734,1736,1737,1739,1623, 1622,821,1187,1225,1384,1620,787,1178, 1624,1621,1188, 771,1625,1394 (22)	0.82-0.89	0.73
MLHD	4181	5,690,544,19,167,98,92,87,1744,705,543,459, 192,16,626,1787,302,1783,204,246,775,1798, 1784,1622,1796, 1745,684,778,1791 (29)	0.85-0.83	0.71
SVM	3638	1738, 1739, 1737 (3)	0.98-0.77	0.75

representations of the profiles of this dataset are shown in Figure 4.7 in which hierarchical clustering shows that, overall, samples of the same class are grouped together though not all of them in the same cluster. Although the goal is to find specific bins that are related to classes, hierarchical clustering indicates that sample classes can be distinguished using the whole profile. The goal is to make this distinction using specific bins of the spectrum.

In order to select bins potentially related to VD classes, GALGO was used to search for multivariate models. These models were then used to design a representative model.

Table 4.5 – Contiguous and non-contiguous bins selected in representative models for Vitroretinal disease. Contiguous bins may represent the same metabolite. In this table, "contiguous" bins were defined as those bins in Table 4.4 that are no farther than 10 bins.

Classifier	*Non-Contiguous bins*	*"Contiguous" bins*
KNN	771, 787, 821, 1178, 1225, 1384, 1394, 1620	1187-1188, 1621-1625, 1733-1740
MLHD	5, 167, 192, 204, 246, 302, 459,626, 705, 1622	16-19, 87-98, 543-544, 684-690, 775-778, 1744-1745, 1783-1793
SVM		1737-1739

Representative models using the most frequent bins from a pool of more than 2,000 multivariate selected models are shown in Table 4.4, their relative position in reference to the metabolic profiles is shown in Figure 4.7 (marked with coloured "+" in the middle of the figure). Only 3 bins were sufficient to obtain a good classifier distinguishing between CU and LIU using SVM. Interestingly, these 3 signatures are nearby each other (1737 to 1739). Considering the property of metabolic profiles in which the signal is continuous and very highly correlated, these three bins could represent the same metabolite. Other classifiers selected larger numbers of bins (22 and 29 for KNN and MLHD respectively). However, the same trend of selecting contiguous or almost contiguous bins was conserved (Table 4.5). Bins selected by SVM were also selected by KNN but not by MLHD. Nevertheless, MLHD selected bins nearby (1744 and 1745). In general, KNN and MLHD did not select similar bins (Figure 4.7).

A PCA representation of the representative models show sample separation using only those bins selected by the classifiers and the first two components (Figure 4.8) whereas the whole profile show only partial distinction for the first component (Figure 4.9). Some of the selected bins are clearly related to large groups of samples in the same class

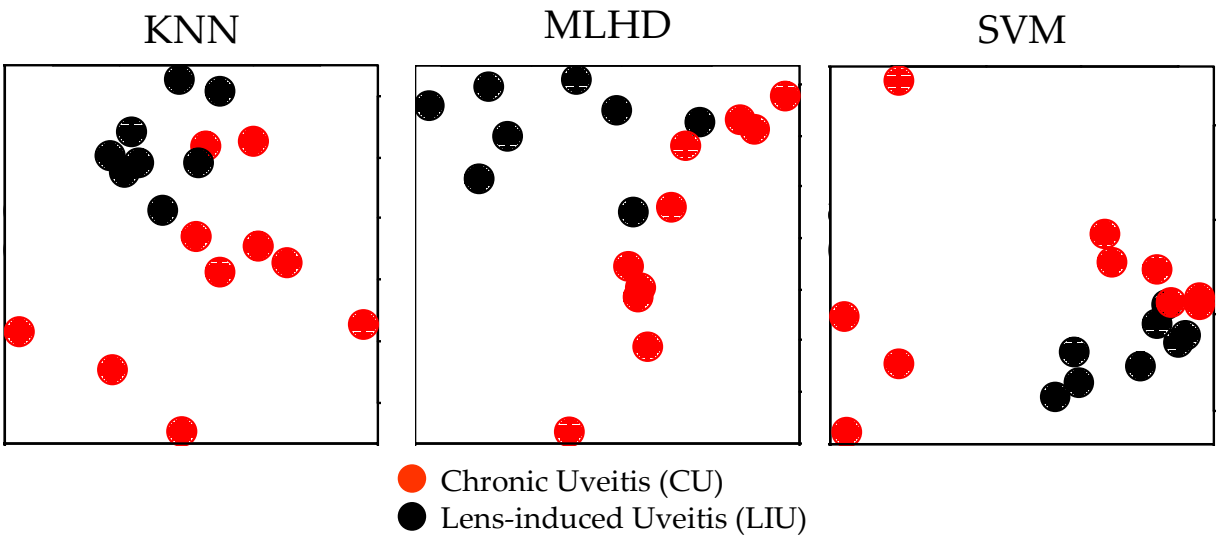


Figure 4.8 – PCA representation of representative models. Axes show the first and second principal components (horizontal and vertical respectively).

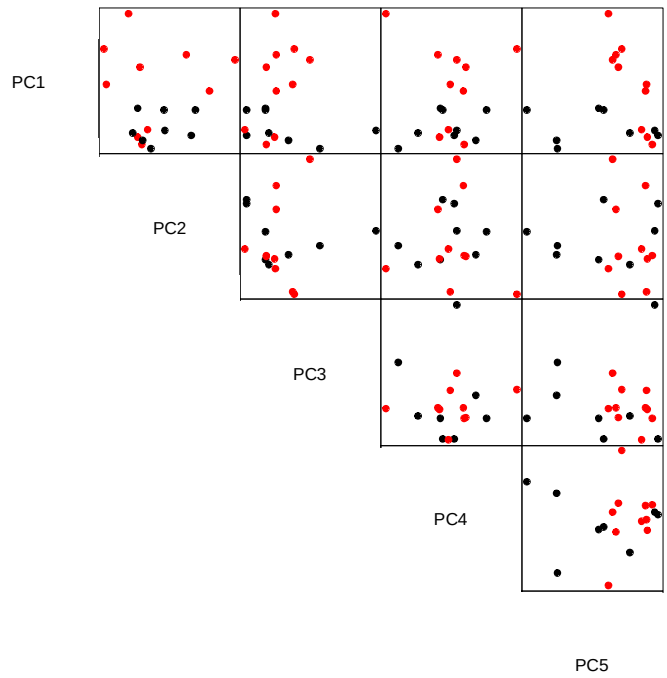
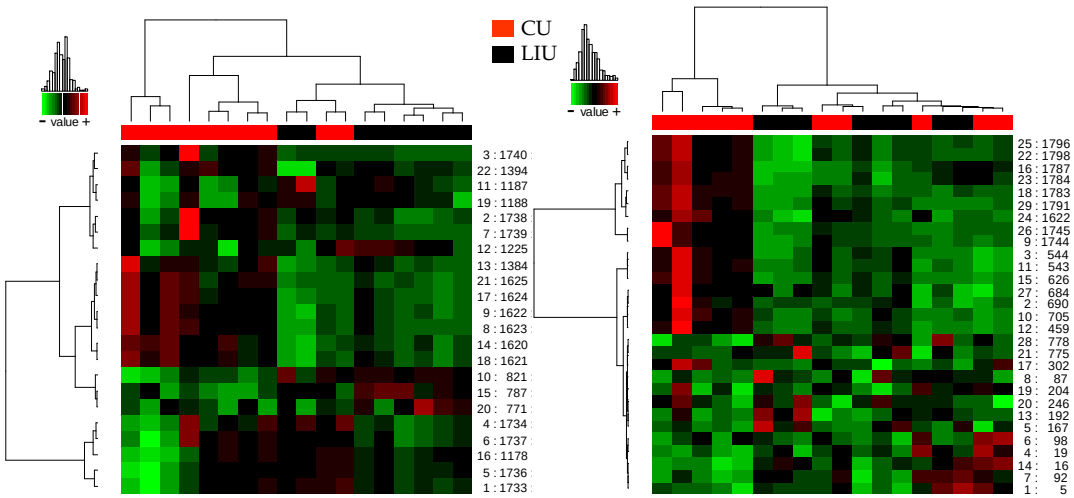


Figure 4.9 - PCA representation of the whole metabolic profile from the VD dataset. Red dots represent CU samples whereas black dots represent LIU samples.

(Figure 4.10). In summary, it was possible to select some bins that could be related to distinguish CU from LIU in VD.



KNN MLHD

Figure 4.10 – Heatmap representation of representative models for VD. Zones more dense for higher or smaller values are evident and cross related to classes and bins. Samples are shown in horizontal axis whereas bins are shown in vertical axis. Numbers at the right of each heatmap represent the relative order and their corresponding bin number.

4.3 - Discussions

4.3.1 - Case 1: Early Rheumatoid Arthritis

It was possible to select IL-1ra as a prognostic factor for RA in which the sensitivity is folded at a very low decrease in specificity. Although IL-1ra could be selected by means of the univariate tests, multivariate approaches provided several additional facts. First, it confirmed that IL-1ra is indeed important for RA prediction under a different modelling scheme. Second, multivariate models discriminate variables that were also highly ranked in the univariate approach. Third, multivariate models using only cytokines were also very accurate which can be used to guide further research. Finally, multivariate models also designed a model using IL-1ra, RF, and CCP by a blind search, that is, without constraining the inclusion of RF and CCP that would be necessary using only univariate variable selection methods.

In the reduced dataset where RF and CCP are both negative, only 11 RA samples were used, this is almost 8 times less than the 84 samples from non-RA/SL individuals. Consequently and contrary to clinical expectations, classifiers tend to give more importance in prediction to those 84 non-RA/SL samples than to the 11 RA samples. To compensate for this effect, classifiers had to be trained using similar number of samples for each class. In the case analysed here, this number should be be less than 11 to leave some samples for external testing. Thus, variable selection other than IL-1ra was presumably unstable due to this lack of data. This lack of data should be one of the reasons for the failure of CT classifier in this reduced dataset. In order to select further variables successfully and to perform validations, more RA samples would consequently be needed.

4.3.2 - Case 2: Vitreoretinal Disease

Several bins were selected that are putatively related to VD. Multivariate models designed with these bins display acceptable sensitivity and specificity. However, only 19 samples were provided in this study. Thus, results should be considered as preliminary. The selected bins have been considered for future research in which the first task would be the identification of the metabolites corresponding to those bins. The selection of some bins is supported by the inclusion of different classifier machines. This fact suggests that those bins are better candidates. It makes sense that many of the selected bins were also close to other selected bins. This is supported by the fact that metabolic profiles are a continuous signal in which the boundaries where a metabolite starts and ends need to be solved by further experiments at higher resolutions focused in that specific range of the spectra (perhaps using techniques other than NMR). In addition, continuous signal could generate auto-correlation: nearby bins could be very highly correlated. It makes sense therefore that nearby correlated signals were selected. It would be sensible to argue that the higher the number of bins selected within a nearby range of the spectra the higher the confidence that a genuine association has been found. Nearby selection nevertheless can also be affected by this continuous signal if the metabolic profiles are not aligned or binned accurately. It was assumed here that misalignment was not an issue. Putatively then, SVM was able to separate CU from LIU using only one metabolite. In addition, KNN and MLHD classifiers selected others bins which may provide additional information about other metabolites altered in VD.

4.4 - Conclusion

It has been shown that GALGO is able to design multivariate variable selected models in other functional genomics data, proteomics in the case of Rheumatoid Arthritis and metabolomics in the case of Vitreoretinal Disease. The analyses were conducted in two

biologically interesting problems and the selected variables would hopefully provide important insights about these diseases. In addition, the selected predictive models may allow the design of novel clinical tests. Besides, it has been showed that other variables may also be related. These other variables may provide additional information for further hypothesis or experiments.

GALGO is a versatile and powerful tool that can be used to aid solving several biological problems using any kind of functional genomics data or combinations of different types of data. Metabolic profiles in which the signal was continuous and highly correlated was not a problem for finding predictive models.

4.5 - Methods

4.5.1 - Case 1: Early Rheumatoid Arthritis

Cytokine data were log transformed before analysis. GALGO was used to search for multivariate variable selection [146]. A representative model was built using GALGO which uses the top most frequent variables in a large number of models. GALGO was run using 2/3 of the data as training and leaving 1/3 of the data blind for the evolutionary process. The number of random splits used to assess accuracy, sensitivity, and specificity was equal to the number of samples used in each run. Five classifier machines were used for initial screening (see Table 4.1 for classifiers' names).

Data from the experimental and clinical work was acquired and kindly provided by Dr. Karim Raza *et al.* from a previous work [136]. The experimental and clinical procedure described next was also provided by Dr. Raza. Patients were recruited through the rapid access clinic for early inflammatory arthritis at City Hospital, Birmingham, UK. Ethical permission was obtained and all patients gave written informed consent. All

patients had one or more swollen joints and a symptom duration of 3 months or less. Patients with evidence of previous inflammatory joint disease were excluded. No patient had commenced a disease-modifying antirheumatic drug (DMARD) before initial assessment. Joints were aspirated under either palpation or ultrasound guidance. Patients were included in the study if adequate synovial fluid was obtained by palpation or ultrasound-guided aspiration/lavage at initial assessment using a method described previously [147]. Patients were subsequently assessed at 1, 2, 3, 6, 12 and 18 months. If joint effusions were present at follow-up assessments, and if consent to a further arthrocentesis was obtained, then these effusions were aspirated. Patients were assigned to their final diagnostic groups at 18 months. Patients were classified as having RA in accordance with the 1987 American Rheumatism Association criteria [148], allowing criteria to be satisfied cumulatively. Although the 1987 American Rheumatism Association criteria have no exclusions, patients with alternative rheumatological diagnoses explaining their inflammatory arthritis were excluded from the RA category. Patients were diagnosed with reactive arthritis, psoriatic arthritis, and a number of miscellaneous conditions according to established criteria. 30 cytokine plus RF, CCP, and RF were measured in all patients along with other clinical information. 50 μl of serum, or of the cytokine / chemokine standard, were pre-incubated with 50 μl blocking buffer ([40% normal mouse serum (Sigma, Poole, UK), 20% goat serum (DakoCytomation Ltd., Ely, UK), 20% rabbit serum (DakoCytomation Ltd., Ely, UK)]) for 30 minutes. Eight standards were made, each containing all the chemokines / cytokines to be analysed at serial ¼ dilutions from 64 ng/ml to 3.12 pg/ml, giving a final concentration range of 32 ng/ml to 1.56 pg/ml. Each well of a 96 well plate was hydrated with PBS for 1 minute and excess PBS removed with a vacuum pump. 50 μl of the diluted serum sample, the standard, or the blocking buffer alone were added to each well (4 wells were run with blocking buffer alone). 12.5 μl of beads from the bead set was added to each well and the plate incubated at room temperature for 2 hours on a plate shaker. The wells were washed 6 times in a vacuum pump with 100 μl wash

buffer (PBS/0.05% Tween 20) and then incubated with 25 μl of the biotinylated detection antibodies diluted in 25 μl blocking buffer and 50 μl assay buffer (1% BSA (Sigma, Poole, UK) in PBS/0.05% Tween 20) at room temperature for 1 hour on a plate shaker. Each well was washed 4 times with 100 μl wash buffer. 0.5 μl of 1 mg/ml BeadlyteTM streptavidin-PE (Upstate Biotechnology, Lake Placid, NY, USA), diluted in 100 μl assay buffer, was added per well and incubated at room temperature for 30 minutes. Each well was washed once with 100 μl wash buffer. The beads from each well were resuspended in 100 μl assay buffer and transferred to eppendorf tubes for analysis using the Luminex^{100} LabMAP™ system (Luminex Corporation, Austin, TX, USA, http://www.luminexcorp.com).

4.5.2 - Case 2: Vitreoretinal Disease

GALGO [146] has been used to search for multivariate bins associated to VD. For these analyses, no further data normalization was performed. Leave-one-out cross-validation in training (LOOCV) was used for error estimation. Training data consisted of 2/3 of the samples. 19 random splits (2/3 for training and 1/3 for test) were used to assess accuracy, sensitivity, and specificity. Models were explored until 100% accuracy was found in the LOOCV training data or 100 generations were reached. Three classifiers were used: K-nearest-neighbours (KNN), maximum likelihood discriminant functions (MLHD), and support vector machines (SVM). More than 2,000 models were considered for each classifier.

The experimental work has been done in Dr. Stephen Young's laboratory at the University of Birmingham. The data and the experimental procedure described next were kindly provided by Dr. Young. Patients were recruited from the tertiary referral Vitreo-retinal unit of the Birmingham and Midland Eye Centre. Samples were obtained from patients undergoing vitrectomy (vitreous humor excised, removed, and replaced

with a clear fluid) for chronic uveitis (n=10) or lens-induced uveitis (n=8). Ethical approval and patient consent was obtained in all cases. An undiluted vitreous sample was taken at the beginning of surgery. One-dimensional ^{1}H spectra were acquired using excitation sculpting on a Bruker DRX 500MHz NMR spectrometer. Chemical shifts were calibrated with respect to the chemical shift position of the trimethylsilyl 2,2,3,3-tetradeuteropropionic acid (TMSP) resonance. Spectra were segmented into 0.005-ppm (2.5 Hz) chemical shift 'bins' between 0.2 and 9.0 ppm using ProMetab version 2. The spectral area within each bin was integrated. Bins between 4.5 and 5.0 ppm containing residual water were removed. The total spectral area of the remaining bins was normalized and the binned data describing each spectrum were then compiled into a matrix, with each row representing an individual sample. Every element was log-transformed to equalize the weightings of the smaller and larger peaks.

CHAPTER 5

Statistical Modelling for Understanding Cell-to-Cell Communication: A Supervised Classification Approach

We were interested in designing an approach to develop signatures predictive of tumour physiology from the molecular state of normal cells and ultimately to infer gene networks representing molecular interactions between the two cell types. This chapter describes the results in demonstrating that indeed the molecular state of normal cells is predictive of the tumour physiological state.

5.1 - Background

The relatively recent development of functional genomics technologies, particularly gene expression profiling, has provided the scientific community with the tools to characterize the molecular state of cells and tissues at a genome level. These technologies coupled with the ability to dissect specific cell types from a complex tissue have created an unprecedented opportunity to characterise the molecular identity of specific cell types in the context of a complex tissue. Following this approach, a number of studies have been performed using gene expression profiling technologies. Results have proven that components of the transcriptional profile of tumour cells are predictive of both tumour features and clinical outcome in a variety of human cancers [149]. These genome-wide studies however do not take into consideration components of the extra-cellular matrix (ECM) (matrix proteins, soluble growth factors and chemokines) secreted by normal cells, adjacent to the tumour site, heavily influence the biology of the tumour. Until now, stromal cells have been most often considered as the

primary candidates for playing a role in normal-tumour cell interactions [150]. These cells are known to secrete most of the enzymes involved in ECM breakdown. They produce growth factors that have a role in controlling cell proliferation, apoptosis, and migration of tumour cells and also secrete pro-inflammatory cytokines involved in chemo-attraction and activation of specific leukocytes [151]. Growth factors and cytokines are also involved in the neoplastic transformation of cells, angiogenesis, tumour clonal expansion and growth, passage through the ECM, intravasation into blood or lymphatic vessels and the non-random homing of tumour metastasis to specific sites. Many of these factors are also secreted by normal epithelial cells, immune cells and endothelial cells in the proximity of the tumour mass. The importance of the micro-environment in determining the onset and progression of cancer raises the question whether it may be possible to predict the patho-physiology and clinical outcome of the tumour from cell type specific components of the molecular state of normal cells. If possible, this would allow the identification of important components of cell to cell cross-talk involved in specifying the development of cancer. To address this question, statistical models based on a genome wide profiling of normal tissue adjacent to the tumour and that are predictive of cancer features are presented. Proof that such an approach has the potential to identify novel mechanisms involved in cancer patho-physiology is provided. In this context, two different prostate cancer microarray datasets available in the public domain have been analyzed [8; 9]. Demonstration that the molecular state of cells adjacent to the tumour is indeed predictive of its physiology is shown. Genes included in these predictive models represent components of cell to cell communication pathways with the ability to modify tumour physiology.

5.2 - Results

5.2.1 - Statistical modelling establishes a link between the molecular state of normal cells and tumour histopathological features

The initial objective of the analysis is to demonstrate that it is possible to predict relevant features of cancer from the molecular profile of normal cells. Two important aspects of prostate tumour physiology have been considered: the degree of organization of tumour cells (defined by a histopathological scoring system called Gleason Score) and the ability of tumour cells to penetrate the organ capsule (summarized by a binary histopathological score called Capsular Penetration). The level of differentiation of tumour cells measures their tendency to aggregate in glandular-like structures that are reminiscent of the organization of the normal tissue. The Gleason Score (GS) can be used to define two main classes. The first is characterized by low-grade tumours that display a highly organised structure (correspondent to a score below 7) whereas a second class is characterized by high-grade tumours cells that are dispersed in the matrix and do not show a tendency to form glandular-like structures (correspondent to a score above 6). By contrast, Capsular Penetration (CP) describes if cells have evaded the capsule that delimitate the organ itself.

The analysis aims to link then the molecular profile of normal cells to GS differentiation level (low versus high differentiation) and CP (positive versus negative). To achieve this, statistical models were developed based on the molecular profile of normal cells and predictive of the sample classes, defined on the basis of the histopathological profile of the tumour.

Firstly, classification models were developed using a combination of a univariate variable selection strategy and supervised classification techniques. In univariate

variable selection each gene is tested independently for its ability to separate the two classes of interest (for example using a simple t-test). The most differentially expressed genes are then included in the statistical model and its classification accuracy is assessed. Figure 5.1 shows that the classification accuracy of univariate models based on the molecular profile of normal cells is, in most cases, good and compare well with the classification accuracy achieved using the expression profile of tumour cells. These preliminary results suggest that the initial hypothesis may be correct. Indeed, the molecular state of the normal tissue is, at some degree, predictive of the histopathological features of the tumour. The validity of the approach is further supported by Gene Ontology (GO) analysis performed using the tool FatiScan [152]. Functional terms were defined as significantly enriched in the predictive signatures if satisfying the threshold of FDR<20% (Table 5.1). The most significant GO term enriched in genes with a higher accuracy for CP in the Singh *et al.* dataset is *membrane bound vesicles*. For GS, *cell adhesion* and *regulation of growth* terms are significantly altered. Although significant terms are rather general and do not allow a very detailed biological interpretation, they suggest the importance of components of cell communication and growth factors in contributing to the predictive power of models based on the molecular state of normal cells.

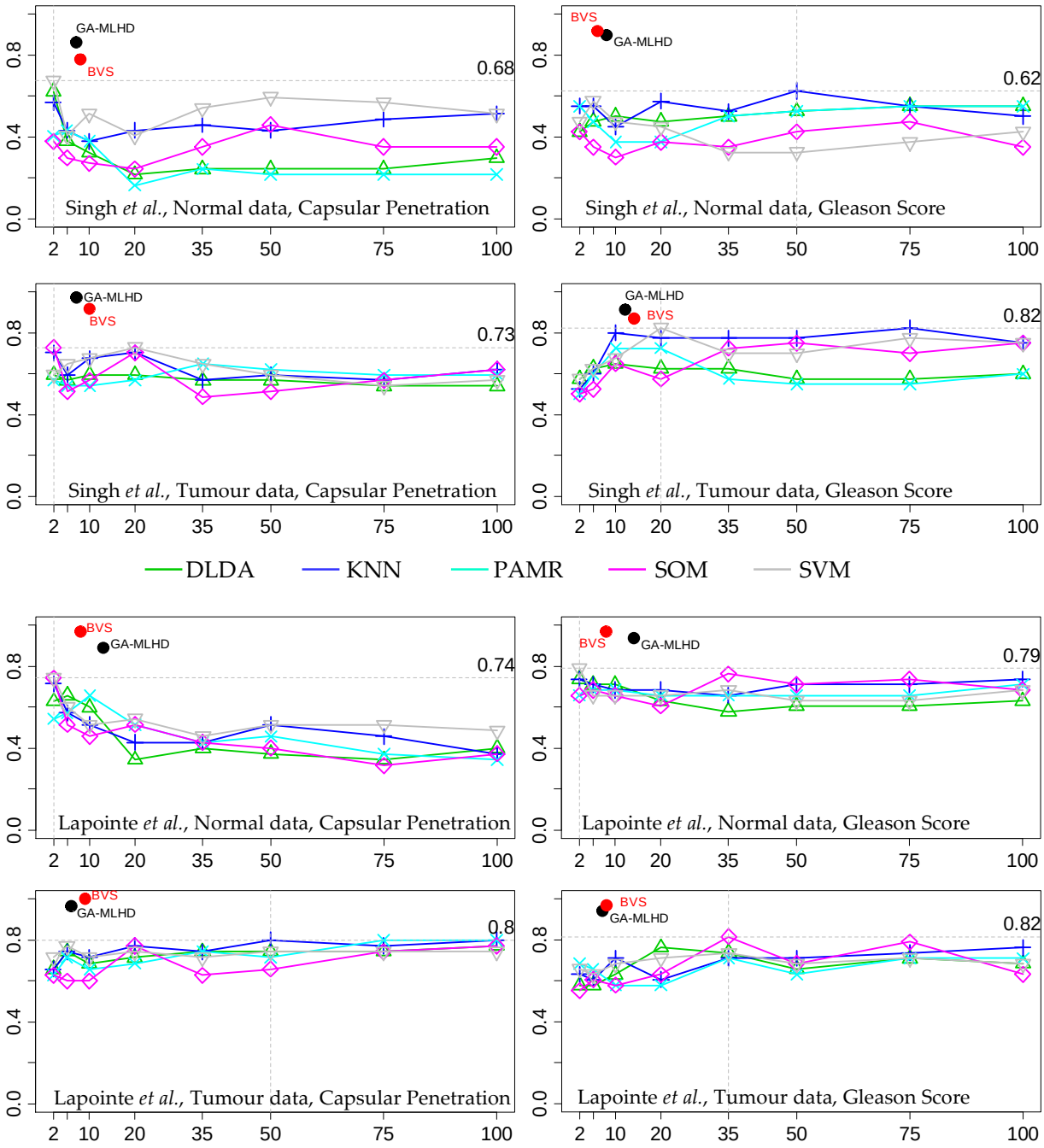


Figure 5.1 - Univariate gene selection models. Models were generated using a forward selection procedure that includes, progressively, genes ranked by a univariate statistic (F-ratio, horizontal axis). The accuracy is assessed by leave-one-out-cross-validation for a number of classification methods (vertical axis, see legends, and the Prophet tool within www.gepas.org [93]). Maximum accuracy is marked by a dotted horizontal line. Overall, this univariate gene selection generates comparable predictive models irrespective of the classification method. More accurate multivariate models generated by GA-MLHD and BVS used in this chapter are shown for comparison in red and black dots. Legends: DLDA – Diagonal Linear Discriminant Analysis, KNN – K-Nearest-Neighbours, PAMR – Shrunken Centroids, SOM – Self Organized Maps, and SVM – Support Vector Machines. See GEPAS [93] for details in F-ratio, error estimation, and classification methods. Dataset, normal or tumour data, and class is specified in each plot.

Table 5.1 - Gene Ontology analysis of genes ranked by univariate statistics. The table below summarizes the results of the Fatiscan analysis performed on the genes ranked by F-statistics (F) (the same statistics used for the development of the statistical models described in the previous section) and using the adaptive statistics (ada). The analysis has been performed using the FatiScan tool available as part of the web based data analysis toolset GEPAS. A star (*) marks terms that were also present in several further windows. The last column specifies if the term is over- (↑) or under- (↓) represented in the class of samples to which the term is associated. CP+ refers to samples whose capsular were broken otherwise referred as CP-. G7+ refer to samples whose Gleason Score is 7 or greater otherwise G6-. NA denotes analysis that did not show any significant term. CC Denotes *Cellular Component,* BP *Biological Processes,* MF *Molecular Function,* SP *SwissProt.*

Dataset	*Analysis*	*Statistic*	*Window*	*Term*	*FDR*	Significance
Singh *et al.*	Capsular-Normal	F	2	**cytoplasmic membrane-bound vesicle (GO-CC)**	0.08718700	↑ CP-
	Gleason-Normal	F	12	**regulation of growth (GO-BP)**	0.18635191	↑ G6+
		F	4	**cell adhesion (GO-BP)**	0.18635191	↓ G6-
	Capsular-Tumor	F	NA			-
	Gleason-Tumor	F	1	extracellular matrix structural constituent (GO-MF)	0.09168624	↑ G6-
		F	1	fibrillar collagen (GO(CC)	0.07888285	↑ G6-
		F	1	Hydroxylation (SP)	0.00492100	↑ G6-
		F	1	Extracellular matrix (SP)	0.09832600	↑ G6-
		F	1	Signal (SP)	0.00781821	↑ G6-
		F	1	Glycoprotein (SP)	0.00828534	↑ G6-
Lapointe *et al.*	Capsular-Normal	F	NA			-
	Gleason-Normal	F	NA			-
	Capsular-Tumor	F	1	cadmium ion binding	0.01050723	↑ CP-
		F	1	copper ion binding	0.09695348	↑ CP-
		F	1	Metal-thiolate cluster	0.01643020	↑ CP-
		F	1	Cadmium	0.01643020	↑ CP-
	Gleason-Tumor	F	1	extracellular matrix (sensu Metazoa) (GO-CC)	0.14571587	↑ G6-
		F	1	Sushi (SP)	0.09121746	↑ G6-
Singh *et al.*	Capsular-Normal	ada	13	cell communication (GO-BP)	0.03003447	↓ CP-
		ada	16	organ development (GO-BP)	0.00210180	↑ CP+
	Gleason-Normal	ada	9	Direct protein sequencing (SP)	.000001210	↓ G6-
	Capsular-Tumor	ada	16	MAPK signaling pathway (KEGG)	0.04855273	↑ CP+
	Gleason-Tumor	ada	11	Direct protein sequencing (SP)	.000009020	↓ G6-
		ada	11	transferase activity (GO-MF)	0.00951405	↑ G6-
		ada	10	cytoplasm (GO-CC)	0.02073723	↓ G6-
Lapointe *et al.*	Capsular-Normal	ada	1	response to biotic stimulus (GO-BP)	0.02075259	↑ CP-
		ada	28	cell adhesion (GO-BP)	0.02075259	↑ CP+
	Gleason-Normal	ada	27	intracellular organelle (GO-CC)	0.04916338	↑ G6-
		ada	2	Apoptosis (SP)	0.04824884	↑ G6-
		ada	2	Homeobox(SP)	0.04824884	↑ G6-
		ada	12	Direct protein sequencing(SP)*	0.04824884	↓ G6-
		ada	17	Signal(SP)*	0.04509734	↓ G7+
		ada	17	Golgi stack(SP)	0.04857050	↓ G7+
		ada	19	EGF-like domain(SP)*	0.02340546	↑ G7+
		ada	19	Signal-anchor(SP)	0.04824884	↓ G7+
		ada	22	Immune response(SP)	0.04689970	↑ G7+
		ada	26	Plasma(SP)	0.04689970	↑ G7+
		ada	26	Hyaluronic acid(SP)	0.02663752	↑ G7+
		ada	27	Lipid-binding(SP)	0.04319938	↑ G7+
		ada	27	Hyaluronic acid(SP)	0.01024267	↑ G7+
	Capsular-Tumor	ada	1	cadmium ion binding(GO-MF)	0.01050723	↑ CP-
		ada	5	membrane-bound organelle(GO-CC)	0.03310354	↓ CP-
		ada	5	intracellular organelle(GO-CC)	0.00832116	↓ CP-
		ada	22	unlocalized protein complex(GO-CC)	0.00832116	↑ CP+
		ada	1	Cadmium (SP)	0.00947798	↑ CP-
		ada	1	Metal-thiolate cluster (SP)	0.00947798	↑ CP-
	Gleason-Tumor	ada	25	Signal (SP)*	0.01810243	↑ G7+
		ada	25	Blood coagulation(SP)	0.04010447	↑ G7+
		ada	29	Hyaluronic acid(SP)	0.01810243	↑ G7+
		ada	28	Sushi	0.03150510	↑ G7+

Although these results are encouraging, this univariate methodology suffers from at least one severe limitation. Univariate variable selection procedures test each gene independently for its ability to discriminate two or more biological states. This methodology therefore ignores the fact that genes work in the context of a network of interacting gene products. Procedures that allow searching for predictive gene sets in association have been developed and tested on microarray datasets and other functional genomics platforms [79; 110; 146] and have been demonstrated to often perform better than their univariate counterparts [146]. Therefore, a second and more detailed analysis was performed based on the application of two different classification methods that use multivariate variable selection strategies for the development of predictive models. These are a Genetic Algorithm search engine coupled to a discriminant analysis as classifier, and Markov Chain Monte Carlo search coupled with a probit classifier (GAMLHD and BVS respectively, see Methods section). Using these approaches, more accurate representative models predictive of tumour features by means of the gene expression profile of normal cells were developed (Figure 5.2, Figure 5.3, Figure 5.6, Figure 5.5, and Figure 5.6). The classification accuracy and model size of these models demonstrate that comparable models were developed using the molecular state of normal and tumour cells. Representative models developed with the BVS and GA-MLHD methods are based on a very similar number of genes and have a high degree of overlap at the gene level, suggesting that these results are independent of the multivariate methodology used. Moreover, the degree of accuracy of models built with a multivariate variable selection approach is higher with respect to the univariate variable selection approach (Figure 5.1). An extensive functional analysis of multivariate models both at the level of models population and for representative models is described further in this chapter.

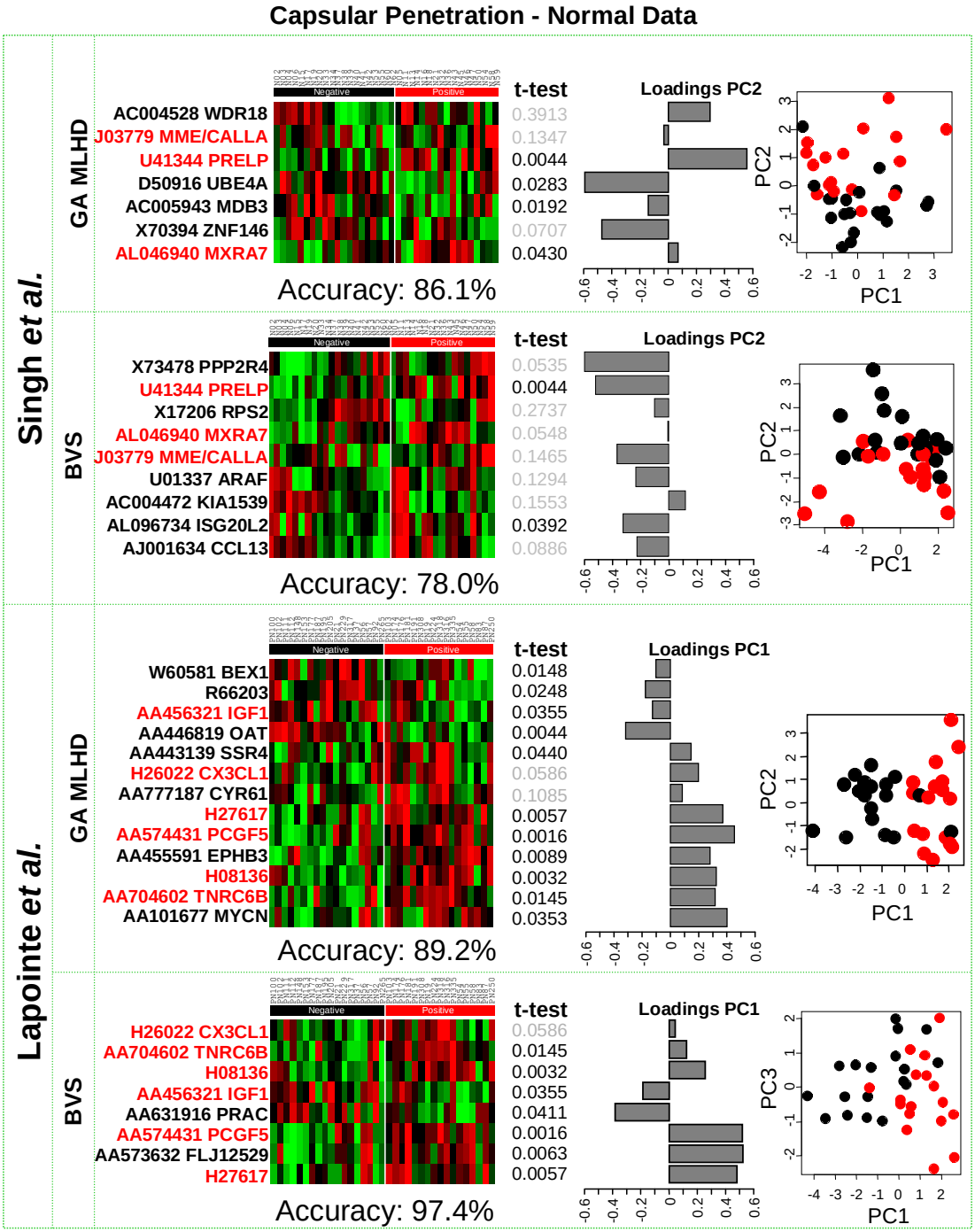


Figure 5.2 - Multivariate Models for Capsular Penetration using Normal data. Genes present in GA-MLHD and BVS for the same dataset are highlighted in red. Accuracy is estimated as described in the Material and Methods section. GeneBank accession number and gene symbol is shown. Brighter green or red colours in heatmaps represent lower or higher relative expression respectively. t-test is shown for comparison with the differential expression criteria commonly used in UVS. PCA plots and loadings are used to associate the contribution of every gene to class separation. For example, PRELP gene in top heatmap seems to contribute strongly to positive Capsular Penetration whereas MDB3 contribute weakly to negative Capsular Penetration.

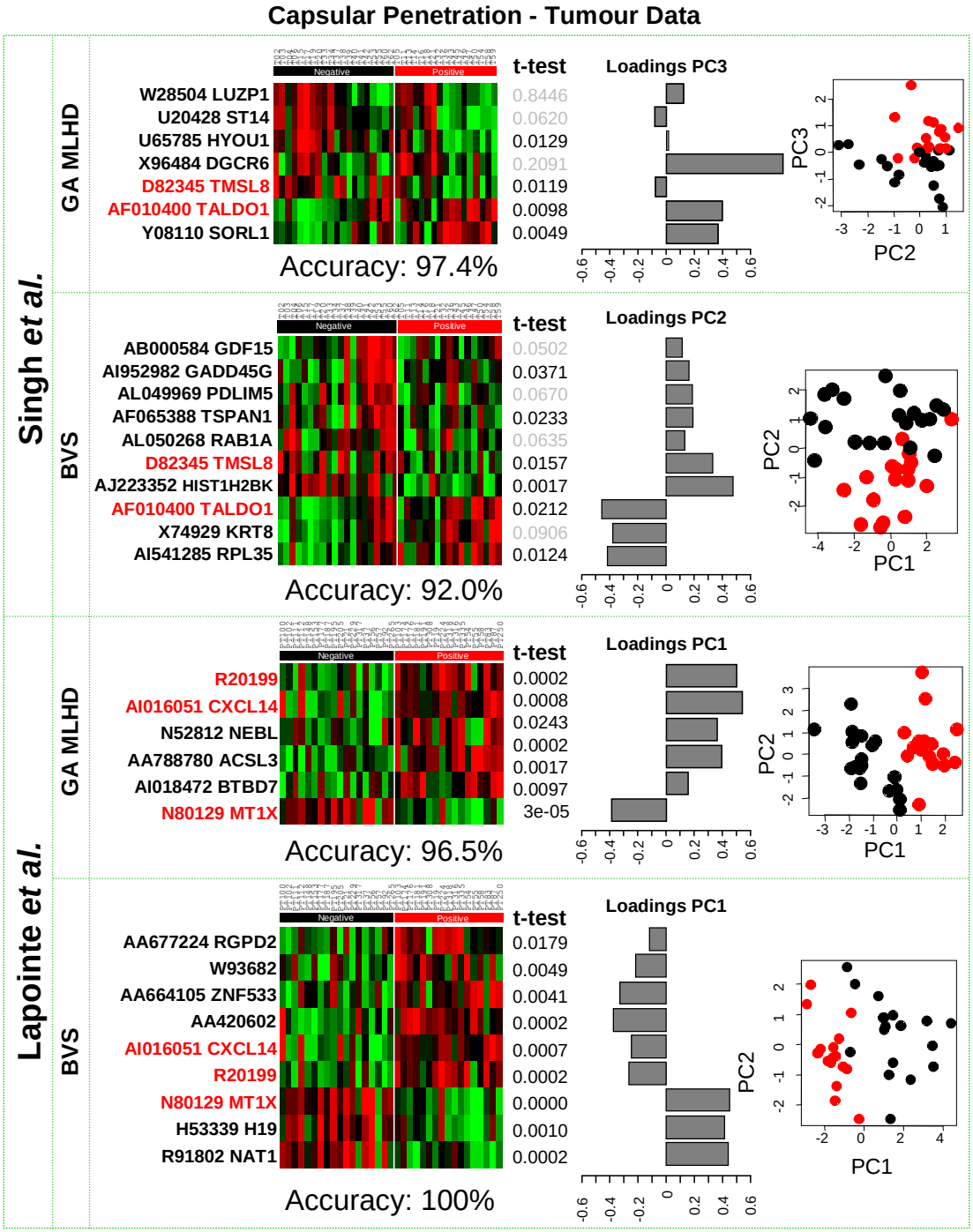


Figure 5.3 - Multivariate Models for Capsular Penetration using Tumour data. Genes present in GA-MLHD and BVS for the same dataset are highlighted in red. Accuracy is estimated as described in the Material and Methods section. GeneBank accession number and gene symbol is shown. Brighter green or red colours in heatmaps represent lower or higher relative expression respectively. t-test is shown for comparison with the differential expression criteria commonly used in UVS. PCA plots and loadings are used to associate the contribution of every gene to class separation. For example, TALDO1 gene in top heatmap seems to contribute strongly to positive Capsular Penetration whereas ST14 contribute weakly to negative Capsular Penetration.

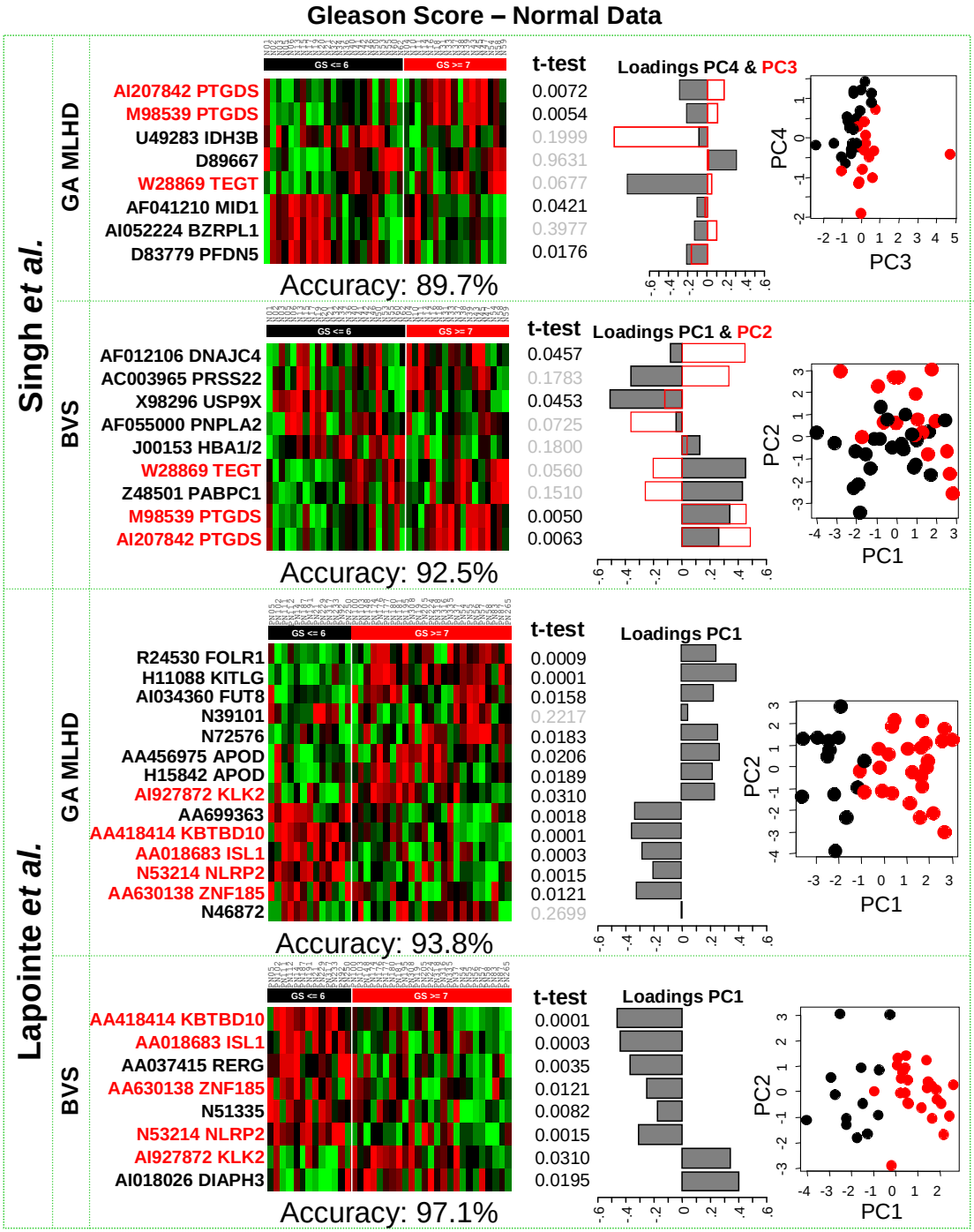


Figure 5.4 - Multivariate Models for Gleason Score using Normal data. Genes present in GA-MLHD and BVS for the same dataset are highlighted in red. Accuracy is estimated as described in the Material and Methods section. GeneBank accession number and gene symbol is shown. Brighter green or red colours in heatmaps represent lower or higher relative expression respectively. t-test is shown for comparison with the differential expression criteria commonly used in UVS. PCA plots and loadings are used to associate the contribution of every gene to class separation. For example, TEGT gene in top heatmap seems to contribute strongly to high Gleason grades whereas D89667 contribute to low Gleason Grades.

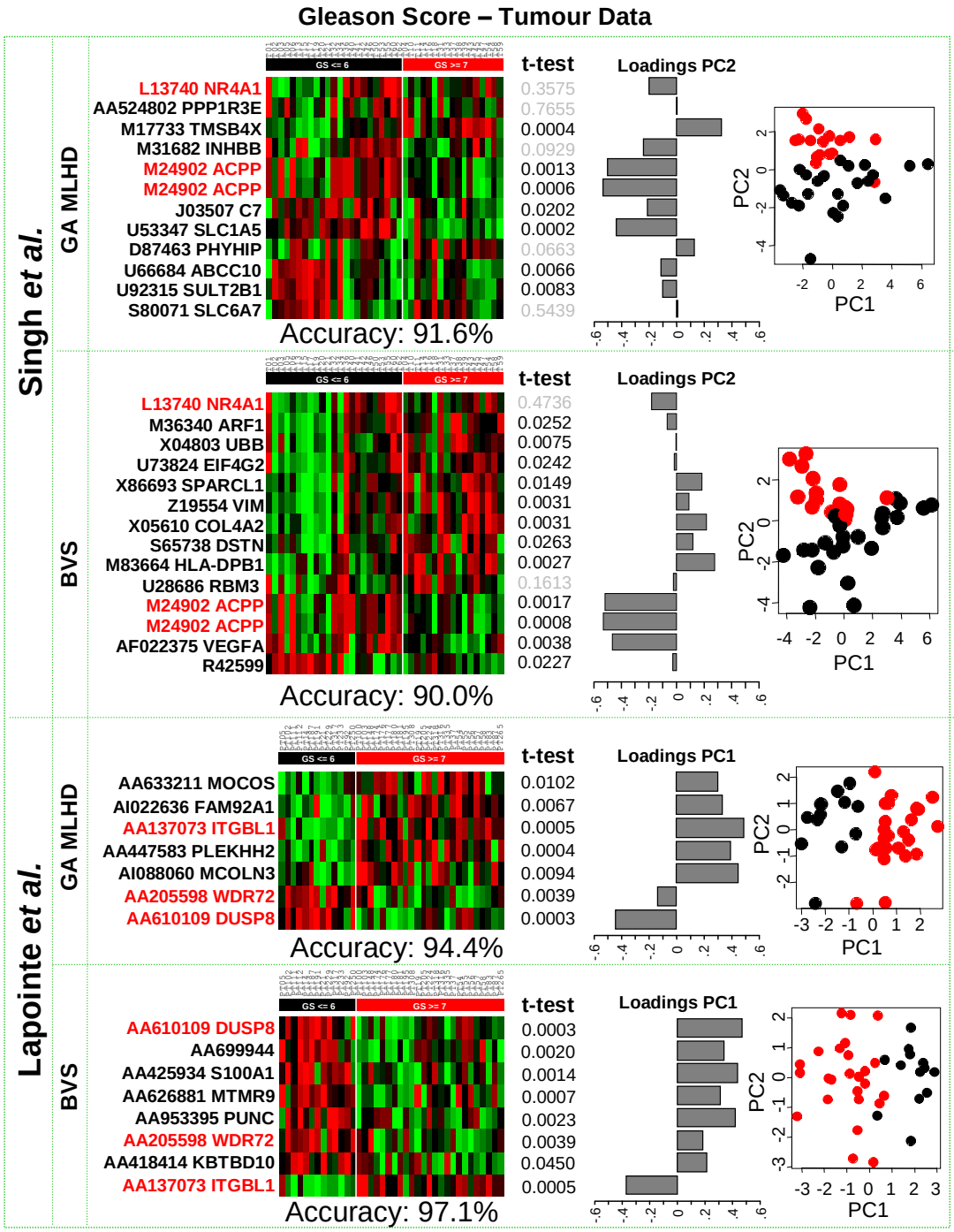


Figure 5.5 - Multivariate Models for Gleason Score using Tumour data. Genes present in GA-MLHD and BVS for the same dataset are highlighted in red. Accuracy is estimated as described in the Material and Methods section. GeneBank accession number and gene symbol is shown. Brighter green or red colours in heatmaps represent lower or higher relative expression respectively. t-test is shown for comparison with the differential expression criteria commonly used in UVS. PCA plots and loadings are used to associate the contribution of every gene to class separation. For example, ACPP gene in top heatmap seems to contribute strongly to low Gleason grades whereas TM8B4X contribute to low Gleason Grades.

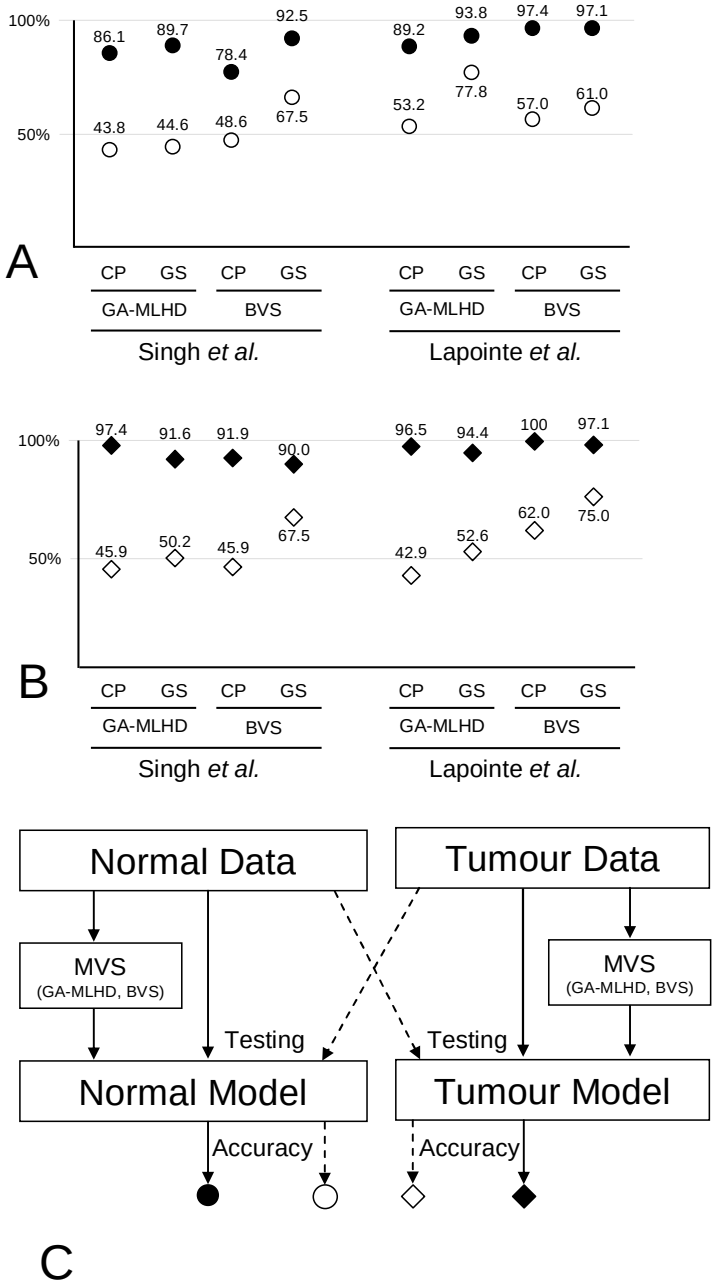


Figure 5.6 - Accuracy and tissue specificity of representative models. The predictive accuracy of the models developed using normal data (panel A, filled circles) is comparable to those models developed using tumour data (panel B, filled diamonds). When models developed using normal data are trained and tested using tumour data, the predictivity is decreased considerably (empty circles). Likewise, tumour models trained and tested with normal data are also non predictive (empty diamonds). Filled symbols follow then the step 1 whereas empty symbols follow the step 2 described in Tissue Specificity inside Material and Methods section. Panel A shows the accuracy for normal models whereas Panel B shows the accuracy for tumour models. Panel C shows the strategy described.

5.2.2 - Signatures representative of tumour physiology are tissue-specific

Results in the previous section demonstrated that the molecular state of normal cells adjacent to the tumour is predictive of cancer histopathological features and that the predictive power of these models is comparable with those developed using the molecular profile of tumour cells. Although adjacent normal and tumour tissues are distinct, they show a high degree of molecular similarity (in the overall gene expression profile, data not shown). Therefore, the study was conducted on whether the predictive power of these models based on the molecular profile of normal cells is a reflection of either the inherent similarity between normal and tumour tissues or the relatively small degree of difference observed in the two tissues. In order to test this hypothesis, the tumour tissues expression profile of genes selected in the representative models developed from the molecular state of the normal tissue were used to predict tumour features (see section 5.5.3 - in Methods and Figure 5.6C). The specificity of signatures based on the molecular state of the tumour tissue performing the procedure in the opposite direction outlined above was also tested. With both statistical modelling approaches and in both datasets the predictive power of the non-tissue-specific signatures is closer to 50% (empty circles and empty diamonds in Figure 5.6) whereas the predictive power of the tissue-specific models is closer to 100% (filled circles and filled diamonds in Figure 5.6). This analysis suggests that the predictive power of the models shown is not a mere reflection of the overall similarity of normal and tumour tissues but instead the molecular signatures developed are highly specific for the tissues they are designed to represent.

5.2.3 - Molecular signatures of normal cells associated to tumour histopathological features represent pathways involved in cell to cell communication

There are at least two possible biological scenarios that can explain why the molecular profile of normal cells is predictive of cancer histopathological features. The first scenario involves the ability of normal cells to react and modify the micro-environment thereby affecting the physiology of tumour cells. A second scenario is based on the fact that the genetic make-up of normal cells may carry the hallmarks of the initial events of neoplastic transformation and influences further development of the disease [153]. Mutations in oncogenes, tumour suppressor genes or in genes involved in detoxification and/or repair mechanisms are known to determine the onset of the malignancy and influence clinical outcome [154]. It may be therefore possible that the expression profile of genes directly or indirectly related to mutations in tumour susceptibility genes may be predictive of disease outcome. This second possibility is realistic in respect to normal tissue only in case of germline mutations. This scenario in fact would be compatible with the presence of the mutation in both normal and tumour cells.

In order to test these hypotheses, a functional analysis of the multivariate representative models was performed. However, this would be certainly limited because of the small number of genes included which would reveal specific biological functions and would perhaps obscure other biological components. Nevertheless, the analysis with the GA-MLHD procedure identifies many alternative models that are equally predictive of tumour features. Consequently, many of the genes selected in these models may represent important biological pathways and therefore could provide important insights in the patho-physiology of the tumour. Thus, a search was performed to identify coherent functional trends amongst the genes selected not only in

representative models but also in the population of models developed using the GA-MLHD procedure. The results are detailed in the next sections.

5.2.3.1 - Functional Analysis of representative models

The biological interpretation of the representative models (representing the most frequent solutions identified by the GA-MLHD and BVS procedures, see Figure 5.2, Figure 5.3, Figure 5.4, and Figure 5.5), reinforce the importance of genes of extra-cellular or membrane localization with a demonstrated regulatory activity on tumour physiology. Models developed using the dataset by Singh *et al.* include several secreted factors and surface proteins with the ability to influence tumour physiology. These are: The neutral peptidase gene CALLA, the matrix metalloprotease encoding genes MXRA7 and ADAM22, the collagen component PRELP and the chemokine ligand CCL13. The neutral peptidase gene CALLA is a cell-surface metallopeptidase expressed by prostate epithelial cells that degrades various bioactive peptides [155] including some involved in the growth of prostate cancer. Its expression in cancer cells is drastically reduced by promoter methylation [156] and has been associated with biochemical relapse (increased levels of Prostate Specific Antigen protein in sera) after surgery [157]. CALLA has a demonstrated role in cell to cell communication by influencing prostate cancer cell invasion via a stromal-epithelial cell interaction [158]. The role of matrix metalloproteases and chemokines in influencing tumour aggressiveness is also well documented in many types of cancer. The results point in the direction of a greater importance of the expression of these factors by normal epithelial cells. Models developed from the Lapointe *et al.* dataset include the gene encoding for the CX3CL1 chemokine. This gene encodes for an important chemotactic factor that regulates the migration of human T-lymphocytes on the tumour site [159]. Of particular interest is its role in driving the migration of NK-cells, a type of immune cell involved in the host response to tumour expansion. The models developed here

also include the insulin-like growth factor 1 (IGF1), a very important component in cell to cell communication networks involved in tumour progression. IGF1 encodes a growth factor whose signal (via a regulator tyrosine kinase pathway) has been implicated in the regulation of growth and apoptosis of a number of highly prevalent tumours [160]. In prostate cancer, high levels of serum IGF1 have been correlated with higher risk of relapse [161]. The gene CYR61 encodes for an extra-cellular matrix component that promotes adhesion, migration, and proliferation of endothelial cells and fibroblasts. Its role in determining the stromal and vascular expansion correlated with to prostate hyperplasia has been well documented [162-164]. The mechanism that has been suggested involves the stimulation, via serum growth factors, of CYR61 in stromal and normal epithelial cells with consequent effect in the proliferation and migration of cells. Its biological activity would be a result of an enhanced activity of bFGF [165]. Interestingly, like the CALLA genes identified in the Singh *et al.* dataset, the CYR61 gene is inactivated in prostate cancer [165]. Its biological role, its relevance in predicting tumour CP together with the observation that its expression is diminished in the tumour cells suggest a role in the interaction of normal cells and tumour cells. Although no particular role has been suggested for SSR4 prostate cancer it is a component of the secretion machinery [166] further supporting the importance of factors secreted by the normal tissue in specifying the histo-pathological features of the tumour. Representative models also include a number of oncogenes such as MYCN and RAF and the methylation protein MBD3.

5.2.3.2 - Functional Network Analysis of the genes represented in the GA-MLHD model populations

In order to identify associations between the collections of models with known functional pathways, the selected gene signatures have been mapped on the Ingenuity database using the web-based Ingenuity Pathway Analysis (IPA) tool

(www.ingenuity.com). Ingenuity knowledge base stores curated information on the interactions between genes, maps of canonical functional pathways, and functional relationships supported by published literature and by protein-protein interaction data. Pathways of highly interconnected predictive genes are identified in this database by statistical likelihood and can be used to formulate hypotheses on the biological framework underlying the statistical models. The analysis was then performed on the 218 genes represented in the population of statistical models developed using the GA-MLHD procedure from the molecular profile of normal cells and predictive of CP. This analysis revealed that all genes included in the models and represented in the Ingenuity database were all part of a single large network representing 11 interconnected sub-networks with a significant score (Table 5.2). Figure 5.7 shows a simplified network representing the union of sub-networks 1, 3, and 4 (the untrimmed version is shown in Supplementary Figure 5.1). The figure shows that 53 of the genes represented in the model populations are associated with interconnected pathways centred on the cytokines TNF, IL4, IL13 and the growth factors TGFβ and IGF1. GO analysis of the genes represented in this network reveal a significant association with growth factor activity, response to stress, response to external stimulus and a strong association with extra-cellular region, cell to cell signalling and extra-cellular matrix terms. These GO terms indicate a direct link between the predictive power of signatures based on the molecular state of normal cells and gene products associated with the extra-cellular environment (see Supplementary Table 5.1 for the full list of significant GO terms). Figure 5.8 shows that 29 additional genes are directly linked to the oncogenes c-myc or p53 (Figure 5.8A and Figure 5.8B respectively, see also the corresponding untrimmed version in Supplementary Figure 5.2 and Supplementary Figure 5.3). This finding is consistent with the importance of p53 as a tumour suppressor gene whose mutations are associated with the development of many human malignancies. Abnormal nuclear p53 accumulation and p53 mutations have in fact been observed in prostate cancers, particularly in prostate cancers of higher tumour stage, higher tumour grade,

metastases, or androgen-independent tumours [167]. The fact that genes associated with the oncogen c-myc are part of a significant network, in combination with the recent discovery that over expression of c-myc in a transgenic mouse induces the development of prostate cancer [168] raises the interesting possibility that c-myc status may also be an important predisposition factor in the development of human prostate cancer [169]. GO analysis of the networks associated with c-myc and p53 shows the link with GO terms transcriptional regulation, nucleus, and regulation of cell proliferation (see Supplementary Figure 5.2, and Supplementary Table 5.3 for the full list of significant GO terms in c-myc and p53 sub-networks respectively). These results support the validity of the approach confirming the expectation that pathways involving cell contact and directly controlling cell migration in normal cells may be involved in sensing and/or influencing the microenvironment and so influence the aggressiveness of a tumour.

Table 5.2 - Significant Networks identified by IPA associated to CP class from the population of models of the molecular profile of normal cells developed using the GA-MLHD procedure in Singh *et al.* dataset. Column identifiers are: NTW-Network, S-Score, FG-Focus genes, FUNCTIONS-most important functional terms associated to the network as identified by the IPA software.

NTW	*GENES*	*S*	*FG*	*FUNCTIONS*
1	ACTN1, ADAM15, ADCYAP1, AFR1, ARAF, BCAM, COL6A1, CTGF, DGCR6, DNAJB2, EEF2, EPOR, ERBB3, GCLM, GRB2, IGF1, IGFBP4, IRS4, ITGA3, JUNB, KCNH2, LRPAP1, MCM2, MCM3AP, NRG2, PAK4, PAPPA, PTK6, SERPINA3, SPIN2A, SYNPO2, TUB, YWHAB, YWHAH, YWHAZ	39	26	Cellular Growth and Proliferation, Cardiovascular System Development and Function, Cellular Function and Maintenance
2	ARPC3, ARPC1B, COL18A1, COX17, COX4I1, COX5B, COX6A1, COX7C, CRIM1, DDB2, E2F1, EEF1G, EGFR, FBN1, GABPB2, GAPDH, GRP, H3F3A, IGF1R, JRK, MAP4, MDK, MYCN, NRG2, RBBP8, RPL13, RPL37, RPL13A, RPL37A, RPS2, RPS7, RUSC1, TMED9, UBB, XDH	23	18	Cell Cycle, Cancer, Cell Death
3	ADFP, ASS, ATF3, EXT1, FABP4, FOXC2, GCLC, GCLM, GPD1, GSTT1, HSD11B1, HSD17B4, IFIT1, JUNB, LBP, MGP, NFE2L2, NFRKB, NUP98, NUTF2, PMF1, PPARA, PPARD, S100A8, SCARB1, SERPINA3, SKIP (includes EG:51763), SLC16A5, SLC2A4, ST6GAL1, STMN1, TFAP2C, TNF, UBC, WWOX	21	17	Lipid Metabolism, Molecular Transport, Small Molecule Biochemistry
4	BATF, C19ORF10, CAP1, CCL11, CYC1, CYTB, DEFB103A, DHCR24, FKBP1A, FZR1, HSD11B1, HSD3B1, IGFBP4, IL4, IL9, IL11, IL13, IL4R, ITGA7, KRT1, LGALS3BP, MMP11, MVK, PLP1, S100A8, SRM, SYNGR2, TGFB1, TIMP1, UQCRB, UQCRC1, UQCRC2, UQCRFS1, UQCRFSL1, UQCRH	21	17	Carbohydrate Metabolism, Cell Signaling, Energy Production
5	ASGR2, BRD2, CDK8, CSDA, DNAJC7, DSTN, FKBP4, GADD45G, GAPDH, HOXB4, HSP90AA1, ID2, KPNA4, LY6A, MYC, NFYB, NOSIP, NR1I3, OAZ2, PPIA, PPID, PRDX2, PRDX3, PRKCD, RAD51, RB1, RCC1 (includes EG:1104), RPL32, RPL41, SNRPN, SRM, TBC1D22A, TPM1, UGT1A2, ZBTB16	17	15	Cell Cycle, Connective Tissue Development and Function, Post-Translational Modification
6	ACP1, APOD, APS, ATP1B3, BRP44, CEACAM1 (includes EG:634), CLU, CTNNB1, ERBB2, GAS6, GRIK1, GRIK2, GRIK5, IGF2R, IGHA1, IL7, IL9, ILK, KRT5, LYN, M6PRBP1, MAP2K2, MAPK1, MME, NPC2, PEBP1, PIK3CA, PTPN18, RPL17, SH2B, STAT2, STAT5B, TIMP1, UBTF, WFS1	17	15	Cellular Growth and Proliferation, Cancer, Cell Death
7	BLM, BTBD2, CCND3, CCNG1, CDKN2A, CLIC4, CRYBA4, DAZAP2, DUSP1, E4F1, ETS1, FOSL1, GSTM1, JUNB, MAF, MFAP2, MSX1, MYB, NDN, NEDD4, NFE2, NQO1, PLAGL1, PLTP, PLXNB2, PPP2R4, RNF11, RPL8, RREB1, SGK, TADA3L, TP53, TWIST1, UBE2D2, UBE3A	16	14	Gene Expression, Cell Cycle, Cell Death
8	ACACA, ACO1, AKR1A1, CCKAR, CD160, CSF2, CTSH, EFNB1, ERP29, FOS, FOXE1, GCLC, GRP, HLA-G, HNRPD, HSD3B1, IFNG, IL1RL1, INS1, KALRN, LDHB, LEP, NIPSNAP1, OPLAH, PDE3B, PMCH, POMC, PSMC5, PSMD5, RPL23A, SLC13A2, SNRPD2, SORBS1, TBX19, TIMP1	16	14	Behavior, Digestive System Development and Function, Nutritional Disease
9	ACVR1, ADORA2B, ARF5, ATF3, AVIL, CBLC, CLDN7, DDIT3, EGF, EGR2, ELL2, GABBR1, GABBR2, GGTLA1, HIRA, IGFBP4, MME, MR1, MSN, PAX7, PEA15, PLCB3, PLD2, PRKCD, PTHLH, RDX, RPS6KA1, SGK, SIM1, SLC9A3R2, SORBS3, STK11, TBC1D10A, USP19, VIL2	14	13	Cellular Assembly and Organization, Cancer, Cell Death
10	ATN1, BLM, CASP3, CCND1, CD247, CRK, CYP19A1, EPAS1, GPX4, GRP, HNRPL, HYOU1, IRS1, KCNK3, KRT5, KRT14, MYCBP2, NASP, NEU3, NFE2, PIK3CA, PINK1, PRKCD, PRKCE, PRKCZ, RASSF1, SLC25A1, SMPDL3B, SOCS6, SPINK4, TCEB2, UBTF, VEGF, XRCC2, ZAP70	13	12	Cancer, Cell Death, Hepatic System Disease
11	ACTA2, APBA2, APBB1, APOE, APP, CASP6, CD59, CLSTN1, COL18A1, CST3, CTSB, EDN1, EGR1, EWSR1, FOSB, FUS, FYN, HSPG2, MAPK10, PAX3, PIGA, PIGB, PIGH, PIGQ, PRELP, PSEN1, PSEN2, PURA, PURB, SF1, SHC3, SRF, TEAD1, TGFB2, YBX1	8	9	Cell Death, Neurological Disease, Organismal Injury and Abnormalities

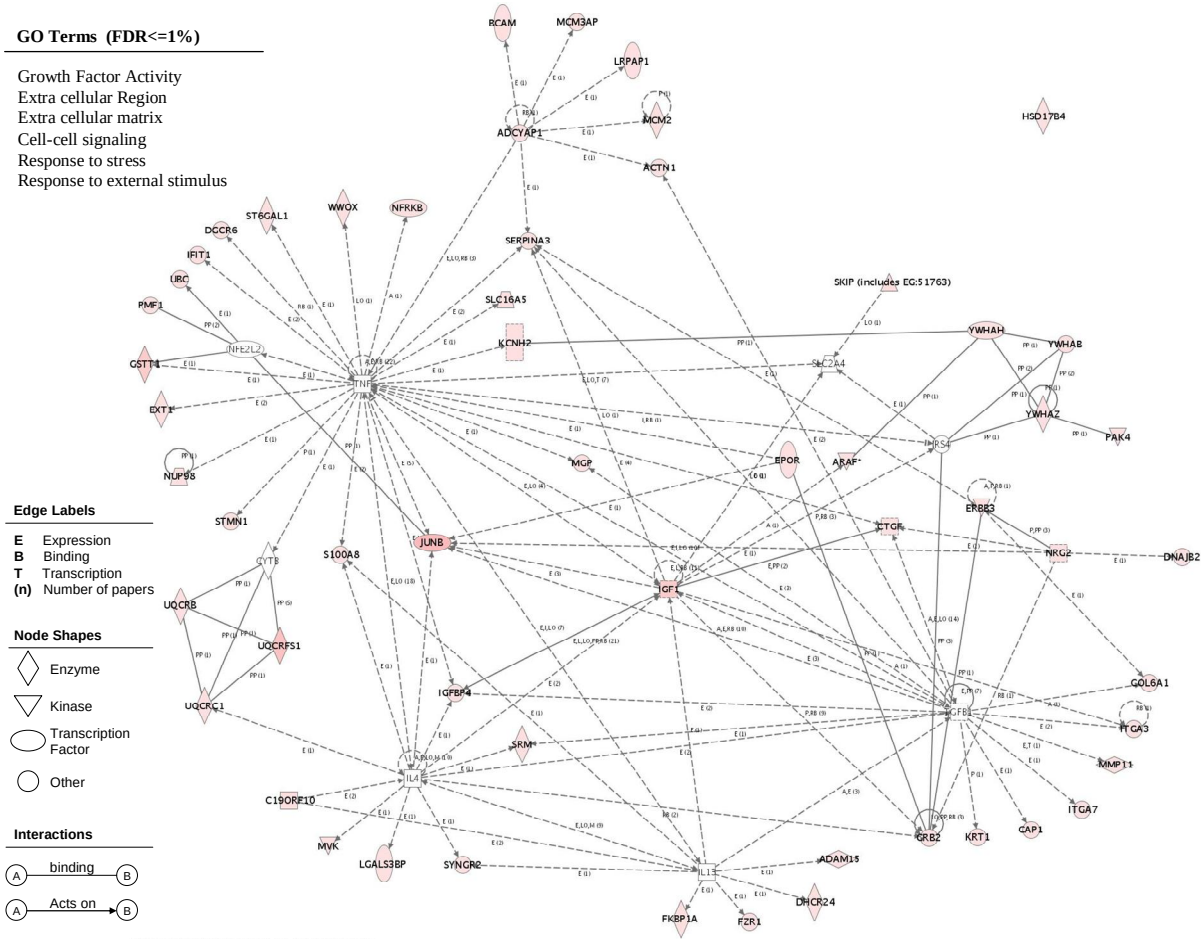


Figure 5.7 - Functional Network Analysis: genes associated to cytokine and growth factor pathways. The figure displays a network resulting from the union of three of the networks identified from the molecular signatures predictive of CP, selected using the molecular profile of the normal tissue (correspondent to networks 1, 3 and 4 of supplementary table 6). The original networks were representing 105 genes associated to five important cytokines and growth factors present in the tumour microenviroment. These are TNF, IL4, IL13, IGF1 and TGFβ. The network in the figure has been simplified in order to make easier the interpretation by pruning some of the genes not included in the statistical models. The original version of the figure is available in Supplementary Figure 5.1. The final network represents 59 genes that have been selected in the predictive models. Nodes, representing genes, with their shape representing the functional class of the gene product, and multiple edges representing the biological relationships between the nodes (see key on the left). Nodes representing genes included in predictive modes are coloured in pink. Functional analysis was performed using a Hypergeometric test and statistically significant high-level functions are reported in Supplementary Table 5.1. Some manually chosen GO terms are shown in the top-left corner.

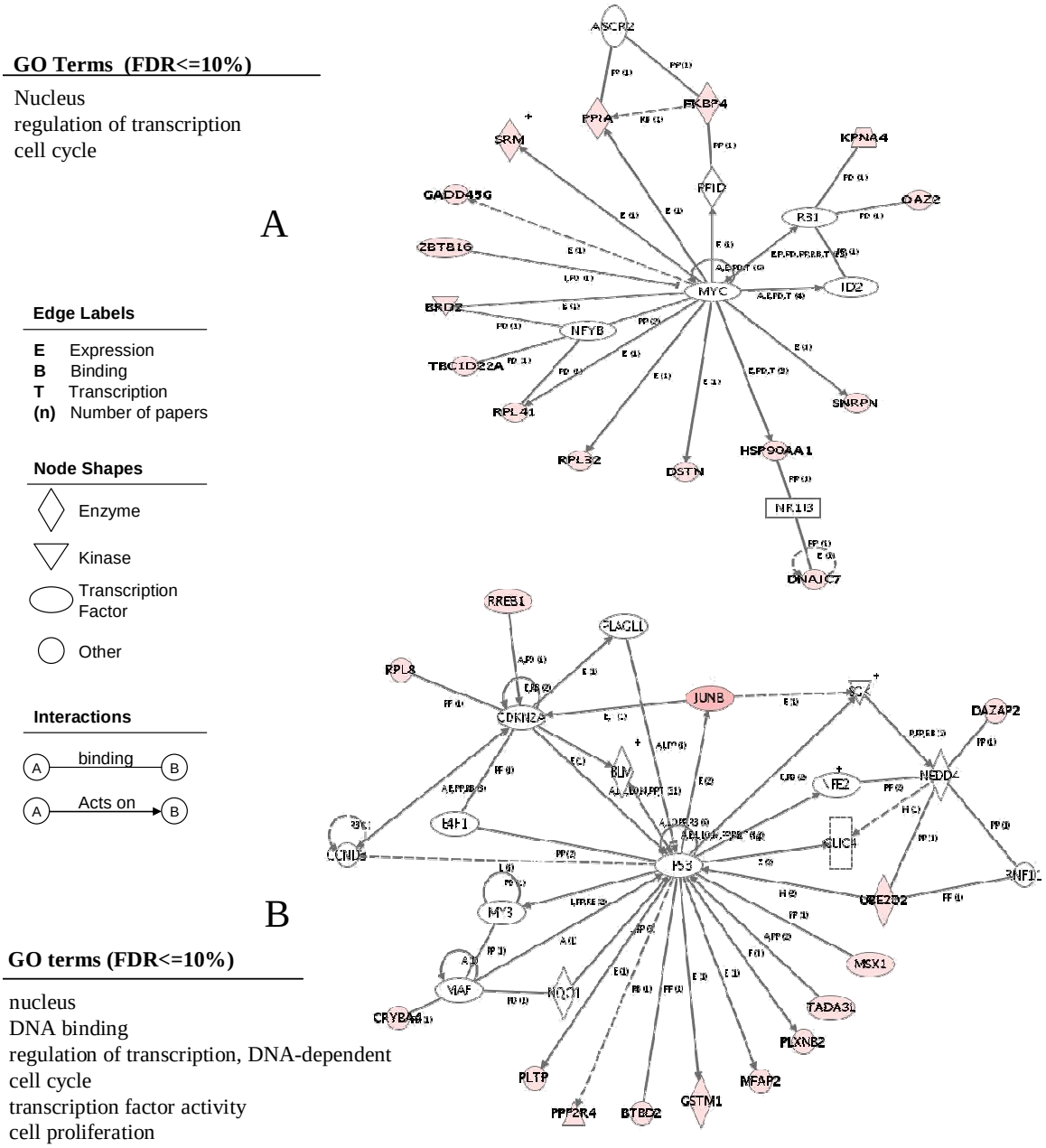


Figure 5.8 - Functional Network Analysis: genes associated to oncogenes. The figure displays two networks identified from the molecular signatures predictive of CP, selected using the molecular profile of the normal tissue. The network in panel A represents 35 genes connected to MYC. Of these, 15 were selected in the models predictive of CP. The network in panel B represented 35 connected to p53. Nodes, representing genes, with their shape representing the functional class of the gene product, and multiple edges representing the biological relationships between the nodes are represented as depicted in the key to Figure 5.7. Nodes representing genes included in predictive models are coloured in pink. Functional analysis was performed using a Hypergeometric test and statistically significant high-level functions are reported in Supplementary Figure 5.3 and Supplementary Table 5.3. Some manually chosen GO terms are shown in the top-left corner.

5.2.3.3 - *Canonical Pathway analysis of the genes represented in the GA-MLHD model populations*

In order to further characterize the function of the genes selected using the GA-MLHD procedure, a canonical pathway analysis using the IPA software was performed. This analysis has revealed additional evidence that models are representing components of the cell to cell interaction machinery and suggests that the expression of genetic predisposition factors is an important component of the models. The results of this analysis are shown in Figure 5.9 and reveal five functional pathways associated with models developed from the molecular profile of tumour cells and five pathways significantly associated to normal tissue (α=0.05). Models predictive of CP using the molecular state of the normal tissue represent *Jak/Stat signalling*, *Integrin signalling*, and *Glutathione metabolism* biological pathways. Among these, the Integrin pathway is a key component of the cell to cell interaction machinery. The Glutathione metabolism pathway plays a role in defending normal cells against carcinogens and its disruption is a predisposition factor for prostate cancer, Lin *et al.* [170] discovered a genetic defect in 88 out of 91 prostate cancer cell samples analyzed. This defect prevents the body from producing glutathione S-transferase (GST), an enzyme needed by the liver to detoxify harmful chemicals. The defect was not found in cells from healthy men. The results are

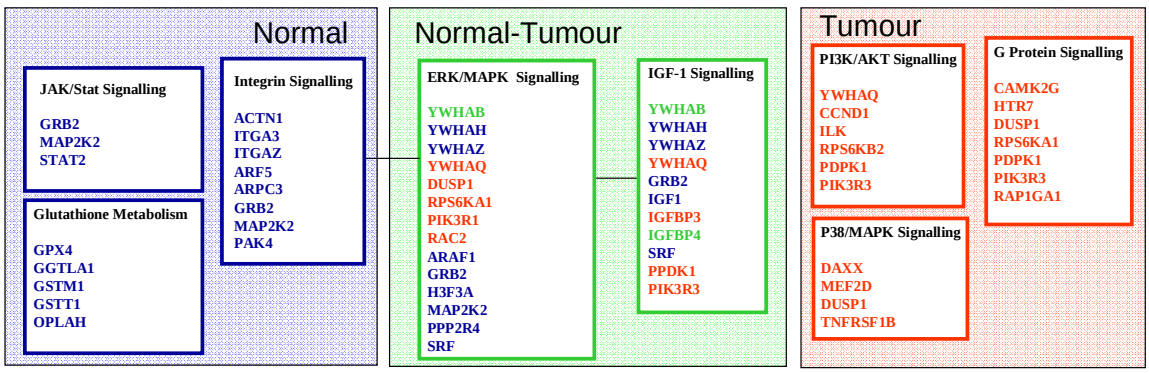


Figure 5.9 - Summary of the results of the canonical pathway analysis performed with the ingenuity software on the models developed from the analysis of the Singh *et al.* dataset to predict CP. Genes are marked in blue if specific for normal tissue and in red if specific for tumour. Genes marked in green have been found both in models developed from normal and tumour tissues.

consistent with these findings and further support the importance of the disruption of the Glutathione metabolism-dependent detoxification in the initiation or, most probably, in the progression of prostate cancer. Despite a very small degree of overlap at the gene level, two inter-related functional pathways seem to be significantly associated to models developed from the molecular profile of both normal and tumour tissues. These are the *IGF-1* and *ERK/MAPK* signalling pathways. Typical signalling pathways controlling cell proliferation and apoptosis (*PI3/AKT signalling*, *P38/MAPK signalling*, and general *G- protein signalling*) are instead specific for models built on the transcriptional profile of tumour cells. The analysis of the dataset developed by Lapointe *et al.* did provide a consistent but less comprehensive picture. Likewise the analysis performed on the Singh *et al.* dataset, the *integrin* pathway was significantly associated to models based on the transcriptional profile of the normal tissue whereas *fatty acid biosynthesis* was associated to models based on the transcriptional profile of tumour cells.

5.2.4 - Survival Analysis

Once demonstrated that it is possible to correlate the molecular state of the normal tissue to tumour features, the analysis was directed to answer whether the molecular state of normal cells is also predictive of clinical outcome. It has been possible to address this question because of the availability of survival free recurrence for a subset of patients in both the Lapointe *et al.* and Singh *et al.* datasets ("recurrence" in this context is defined by increased levels of prostate specific antigen in the sera that indicate the presence of a secondary tumour). Figure 5.10 describes the strategy used in this analysis (see Methods section). SAM was used to identify genes related to survival free of recurrence [43].

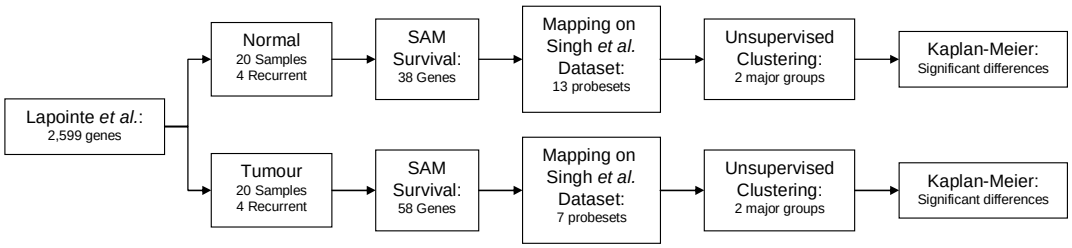


Figure 5.10 - Survival Analysis Strategy. Genes that are potentially related to survival times were detected by SAM in normal and tumour from Lapointe *et al.* dataset. These genes were mapped in Singh *et al.* dataset and subject to clustering. The significance of the difference in survival times between two groups in Singh *et al.* were assessed by the log rank test and shown by Kaplan-Meier plots.

SAM analysis identified 19 genes whose expression profile in the normal tissue is significantly associated to survival free of recurrence. The most striking characteristic of this signature is the large number of immunoglobulin genes (Figure 5.11). This result suggests that the presence of an extensive B-cell infiltrate adjacent to the tumour site at the time of surgery would be an important factor influencing recurrence. Gene signatures derived from the transcriptional profile of tumour cells that are associated to survival are radically different. Interestingly, no immunoglobulin genes or other B-cell markers have been found.

In order to gain further confidence on these results, the dataset developed by Singh *et al.* has been used as an independent validation set. Genes in common between the signatures developed from Lapointe *et al.* dataset and the dataset from Singh *et al.* were identified using Unigene. Then, a cluster analysis was performed to identify groups of patients with different survival characteristics using the data from Singh *et al.* The results are summarised in Figure 5.12. Cluster analysis shows that samples are split in two major clusters. A Kaplan-Meier analysis of these two population of patients shows significantly different survival rates (p=0.004). Other common agglomerative methods (average and complete) display similar results (data not shown). Molecular signatures

identified by SAM analysis in the tumour cells from Singh *et al.* dataset are also predictive of survival times in Lapointe *et al.* dataset (data not shown).

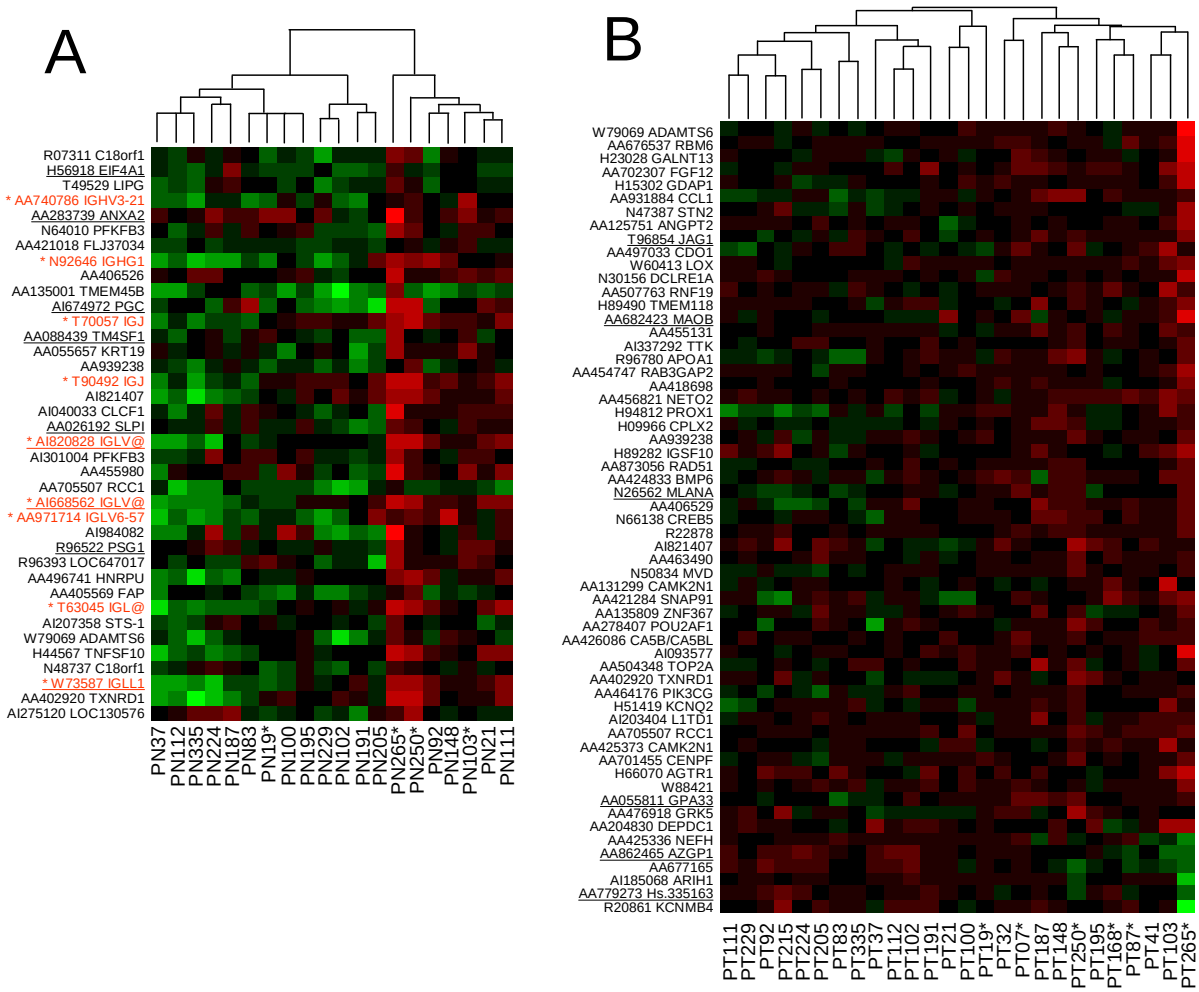


Figure 5.11 - Genes associated to survival times (Lapointe *et al.* dataset). (A) Genes obtained from normal data. (B) Genes obtained from tumour data. Immunoglobulin gene products (in rows) are highlighted in red and marked with a star (*). Genes also found in Singh *et al.* dataset are underlined. Samples (in columns) with a recurrence event are marked with a star. Brighter green or red colours represent lower or higher relative expression respectively.

Because of the limited number of patients in the dataset that have information on the clinical outcome, these results should be considered as indicative and must be confirmed with a much larger patient population.

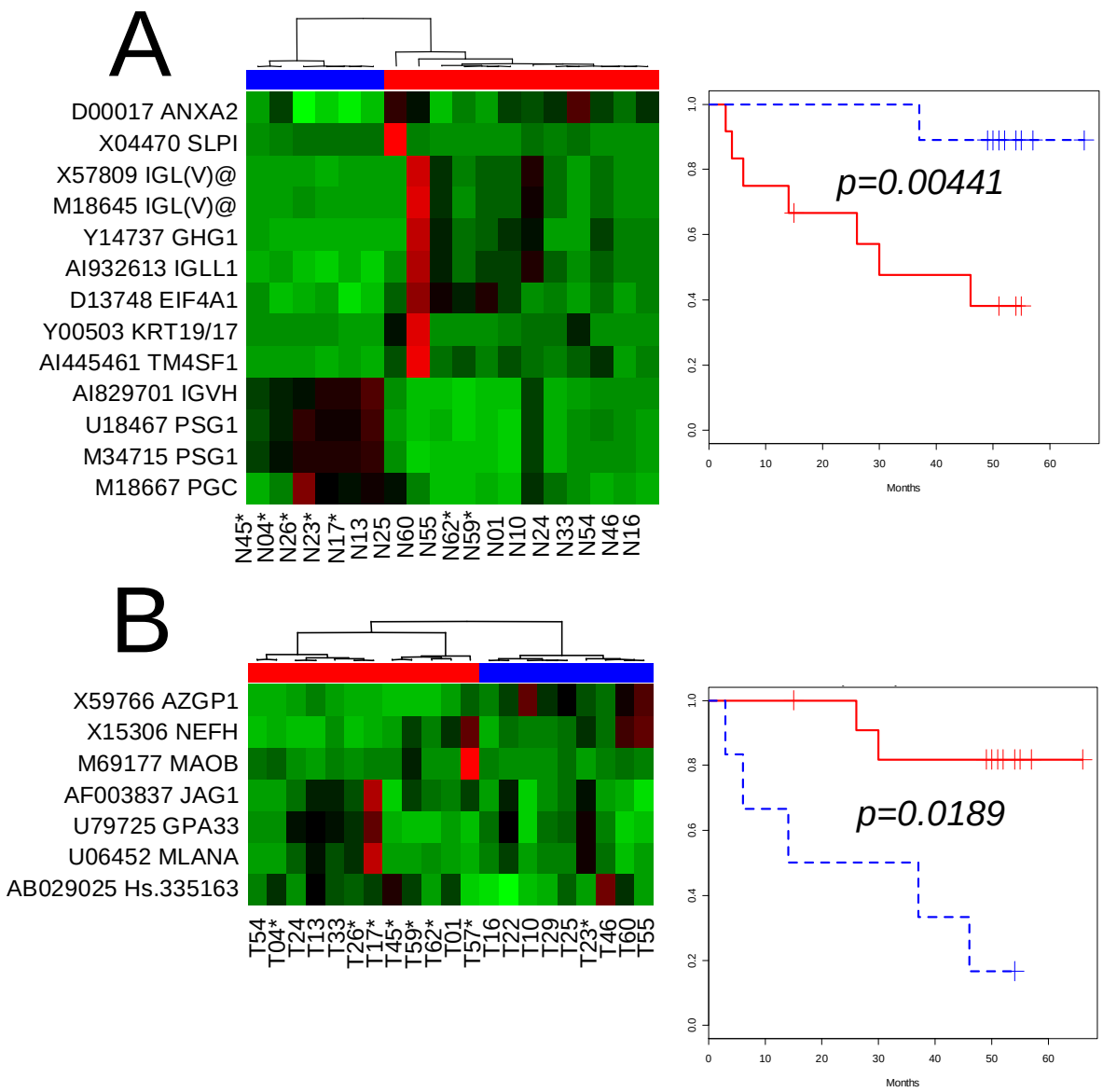


Figure 5.12 - Validation of genes associated to survival times in Singh *et al.* dataset from genes obtained using SAM analysis of Lapointe *et al.* dataset. (A) In normal data. (B) In tumour data. Samples (in columns) with a recurrence event are marked with a star (*). Brighter green or red colours represent lower or higher relative expression respectively.

5.3 - Discussion

Results have demonstrated that the expression of a set of genes in the normal tissue is predictive of important aspects of cancer physiology and probably clinical outcome. The models represent two different classes of genes. They are either important

components of the extra-cellular environment with a demonstrated biological activity on tumour cells (such as CALLA, Cytokines and chemokine ligands, matrix components and matrix remodelling proteins) or with the associated regulation of oncogenes and cancer predisposition factors (such as p53, the glutathione metabolism pathway and possibly the oncogene c-myc).

This analysis leads us to re-formulate the hypothesis that the expression of extra-cellular or membrane bound proteins with the ability to alter tumour physiology and produced by normal cells is an important factor in specifying the development of cancer. Moreover, many of these genes encode for proteins that have a demonstrated tumour suppressor activity suggesting that tumour cells have to escape the molecular restraints of their own internal signalling networks, but also must overcome the inhibitory actions of signals from adjacent cells. Furthermore, these concepts extend beyond the idea of intrinsic tumour suppression where within a given tumour cell genomic surveillance systems and other tumour suppressor pathways are activated under the cellular stress of oncogenic activation. The current findings also suggest that this concept extends to the adjacent normal tissue which also acts in a proactive manner to suppress local malignancy. The importance of the expression profile of the normal tissue in determining the development of cancer is further supported by the indicative results that the expression of genes in the normal tissue is predictive of biochemical relapse. This analysis shows the existence of molecular signatures, from the normal tissue, that specifies groups of patients with a significantly different likelihood of survival free of metastases. The biological interpretation of these signatures reveal, among various factors, that the presence of an extensive B-cell infiltrate in the normal tissue, at the time of surgery, is a component of the signatures that are associated to the development of secondary tumours and/or metastases that determine biochemical relapse.

The identification of factors expressed in normal cells and sufficient to predict tumour physiology lead to interesting hypothesis on the mechanisms involved in the development and progression of cancer and suggest a series of experiments that could be designed to confirm the validity of these hypothesis. The most obvious one is the demonstration (using for example SNP analysis) that the mutational status of some of the genes identified is associated to tumour features.

5.4 - Conclusions

In the context of prostate cancer the application of the analysis strategy described in combination with the ability to laser micro-dissect different subpopulations may allow the identification of important components in the interaction between normal, tumour epithelia, stroma and endothelial cells. Moreover, the approach described here is general and can be applied to identifying the components of the cell to cell interaction machinery in any tissue where it is possible to isolate different cell types.

5.5 - Material and Methods

5.5.1 - Datasets

The analysis is based on two independent large prostate cancer studies performed using different array technologies. In both studies, cells from tumour and adjacent normal tissues have been isolated and the extracted RNA has been hybridized on human microarrays for expression profiling. The first dataset used in the analysis is derived from a study performed by Singh *et al.* [8] where 52 samples of prostate tumours and adjacent normal tissues were collected from patients undergoing radial prostatectomy; then profiled using Affymetrix Genechip technology. The second

dataset used was collected by Lapointe *et al.* (2004) using cDNA arrays. In this study 41 paired normal and tumour specimens were removed from radical prostatectomy. Information about the histo-pathology of the tumour specimens (GS and CP) as well as survival free of recurrence was available for both datasets. For the Singh *et al.* dataset, the analysis has been performed on a subset of 40 unique patients for which matching normal and tumour samples were available. Probesets with an average expression across samples in the bottom 10% and range within top 28% were filtered. 2,752 genes matched these criteria. For the Lapointe *et al.* dataset, 39 samples were selected having tumour and matched normal data. Spots and samples with at least 75% of non-missing data were used. Missing data were median-imputed. Genes were filtered using 23% largest range. 2599 genes were then used.

5.5.2 - Statistical Modelling

The analysis aims to identify molecular signatures predictive of two binary variables representing relevant features of tumour biology. These are the degree of differentiation of the tumour (GS) and the ability of the tumour to penetrate the organ capsule (CP). For this purpose, univariate and multivariate methods have been used which are described in the following sections (general concepts are also presented in Chapter 2).

5.5.2.1 - Classification methods with univariate variable selection

A univariate variable selection strategy was tested based on an F-test in combination with several classification methods (SVM, DLDA, PAMR, KNN, SOM) as implemented in the software application Prophet available in the web-based microarray analysis suite GEPAS [93]. This application uses a step-wise variable inclusion strategy to construct increasingly large models from a list of genes ranked by the value of the F-

statistics and it also implements a cross-validation strategy for error estimation. Results of this analysis are shown in Figure 5.1.

The univariate variable selection strategy has been validated for biological significance using FatiScan [152], a functional annotation tool available in the context of the web-based analysis toolset GEPAS [93]. The aim of FatiScan is to find functional classes (defined by GO terms, InterPro, KEGG, and SwissProt keywords), namely blocks of genes that share some functional property, showing a significant asymmetric distribution towards the extremes of a list of ranked genes. This is achieved by means of a segmentation test, which consists of the sequential application of a Fisher's exact test over the contingency tables formed with the two sides of different partitions made on an ordered list of genes. The Fisher's exact test finds significantly over or under represented functional classes when comparing the upper side to the lower side of the list, as defined by any partition. The FatiScan procedure has been applied to the list of genes ranked by the value of the F-statistics as described in the previous paragraph. The application of FatiScan to such ranked lists renders blocks of functionally related genes that are over or under-represented to sample classes (CP or GS).

5.5.2.2 - Classification methods with multivariate variable selection

In order to consider the effect of combinations of genes in the prediction of the histopathological variables, statistical modelling approaches combined with multivariate variable selection procedures have been used. Besides, in order to demonstrate that the results are independent of a particular methodology, multivariate classification models obtained using two independent procedures were compared. These methods differ for both the variable selection strategy and for the classification algorithms used. The first approach is a modification of the Genetic Algorithm maximum likelihood discriminant analysis (GA-MLHD) method originally proposed

by Ooi and Tan [79]. This method uses a genetic algorithm approach for variable selection coupled to an MLHD functions classifier. The GA-MLHD methodology uses an initial random population of models (called chromosomes) and evolves from them highly accurate classifiers using a process that mimics natural selection. In the implementation used (see Chapter 3), the error estimation strategy has been improved by using two-levels of cross-validations. The first level is used in the evolutionary step of the GA to evaluate the error in a subset of the dataset using a k-fold-cross-cross-validation procedure (k=5). The second level is used at the end of the evolutionary process, when all chromosomes are selected, to estimate the classification error as an average of the test error in 40 random splits (2/3 for training and 1/3 for testing) using the entire dataset. For details in error estimation strategies see Chapter 2. The resulting models obtained using a Bayesian variable selection (BVS) approach that was recently developed [110] were compared with those obtained with GA/MLHD (these BVS models were kindly provided by collaboration with Dr. Mahlet G. Tadesse and Dr. Marina Vannucci from Texas A&M University). The BVS method uses a multinomial probit model as classifier and Markov Chain Monte Carlo (MCMC) methods to search multivariate space for informative subsets of the variables. Error estimation and parameters settings have been described in Sha *et al.* [110].

5.5.2.3 - Selection of model size

The variable selection strategies employed in the implementation of the GA-MLHD and the BVS methods require the definition of the a-priori model size. In the GA-MLHD method the model size is specified as the dimension of the chromosomes whereas in the BVS method this is the parameter of the prior model, w, that determines the expected model size. Since training accuracy is not a good indicator of the optimal model-size, to choose this parameter the classification test accuracy of models of different sizes was compared (Figure 5.13). For the GA-MLHD method the optimal model size has been

defined as the value associated to the highest second-level classification accuracy averaged across a population of 10,000 chromosomes. The initial value of w for the BVS method was set to 10, or 20. Two runs for each choice of w were performed. The optimal model size was chosen according to the lowest miss-classification error (Table 5.3).

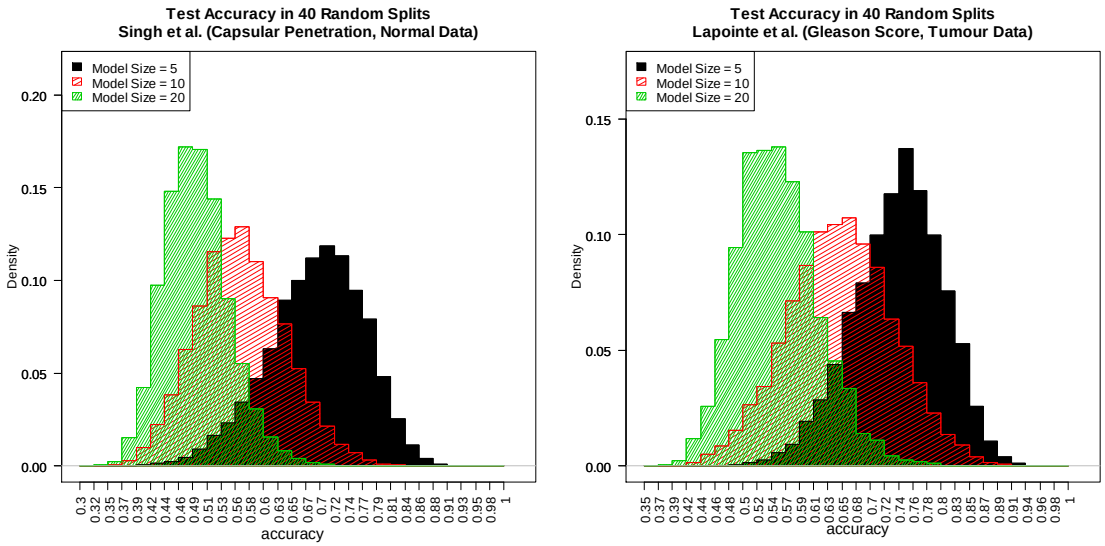


Figure 5.13 - Model size selection for GA-MLHD method. Panels show representative test accuracy distribution from 10,000 models of three sizes (5, 10, 20) for Singh *et al.* dataset (left panel) and for Lapointe *et al.* dataset (right panel). The overall accuracy tends to decrease for larger models. Accuracy distribution for other class-data combinations are similar to those shown in this figure (data not shown). Therefore model size 5 was choosen to build the representative models for GA-MLHD procedure.

5.5.2.4 - Selecting representative models

Both GA-MLHD and BVS modelling approaches provide a number of alternative models with comparable predictive values. These models tend to have a degree of overlap in their gene composition. It is therefore meaningful to select a single summary model that represents the most frequent solutions. In order to do so, for the GA-MLHD approach, a forward selection procedure was applied to the top 1% most predictive models selected using the GA procedure (as in Chapter 3 and [146]). In the case of BVS, models were developed with the genes that were included in the subsets of variables most frequently visited by the MCMC search. The final list of models was generated by

Table 5.3 - Model size selection for BVS method. The final accuracy of two BVS runs starting with model size 10 or 20 is shown. The size chosen for each combination of tissue-class-dataset is bolded. CP+N – Capsular Penetration class from Normal data, CP+T – Capsular Penetration Tumour, GS+N – Gleason Score Normal, GS+T – Gleason Score Tumour.

Class+Tissue Dataset	*Run*	*Expected Model –Size=10*	*Expected Model-Size=20*	*Class+Tissue Dataset*	*Run*	*Expected Model – Size=10*	**Expected Model-Size=20**
CP+N	Run 1	72.9%	**78.4%**	CP+N	Run 1	97.1%	**97.1%**
Singh *et al.*	Run 2	72.9%	**78.4%**	Lapointe *et al.*	Run 2	94.3%	**100%**
CP+T	Run 1	**91.9%**	89.2%	CP+T	Run 1	**100%**	100%
Singh *et al.*	Run 2	**91.9%**	94.6%	Lapointe *et al.*	Run 2	**100%**	100%
GS+N	Run 1	75%	**95%**	GS+N	Run 1	**100%**	97.4%
Singh *et al.*	Run 2	62.5%	**85%**	Lapointe *et al.*	Run 2	**97.4%**	97.4%
GS+T	Run 1	**90%**	87.5%	GS+T	Run 1	**97.4%**	89.5%
Singh *et al.*	Run 2	**90%**	90%	Lapointe *et al.*	Run 2	**97.4%**	92.1%

the union of the two chains with minimum average mis-classification error [110]. Interestingly, representative models developed with the GA-MLHD procedure largely overlaps with the pooled models from the BVS approach.

5.5.3 - Tissue specificity of representative models

An important component of the strategy is to demonstrate that molecular signatures are tissue specific hence they are not representing a mere reflection of the overall similarity between normal and tumour tissues. The strategy to demonstrate the specificity of the gene signatures obtained with the multivariate variable selection strategy implemented in the GA-MLHD procedure is described below in two steps.

5.5.3.1 - Step 1: Development of representative models

Expression data from the normal tissue samples are split between a training and test sets (respectively 2/3 and 1/3 of the original dataset). The training set is used to develop a classification model to predict cancer features with a cross-validation strategy. Once the representative models have been developed, their classification accuracy is estimated on the test set. This procedure is performed in 40 random train-test splits.

5.5.3.2 - Step 2: Specificity test

Expression data from the tumour tissue samples are split between a training and test sets (respectively 2/3 and 1/3 of the original dataset). The expression profile of genes selected in Step 1 (in the samples selected in the training set) is used to train a classification model to predict cancer features. The classification accuracy of the trained model is then estimated on the test set. This procedure is performed in 40 random train-test splits. The classification accuracy estimated in this step 2 is then compared to the classification accuracy estimated in step 1 to establish the tissue specificity of the gene signatures (Figure 5.6). In order to demonstrate the tissue specificity of models based on the molecular profile of tumour tissues, the reverse test between tumour models using normal data was also performed.

The assessment of the tissue specificity of the molecular signatures obtained with the BVS procedure has been performed similarly using a cross-validation procedure for the error estimation as described in [110].

5.5.4 - Network and Canonical Analysis of Genes Selected in the GA-MLHD Models Using the Ingenuity Software

Because of the relatively small number of genes selected in the summary models the biological interpretation is relatively simple. This analysis has been presented in section 5.2.3.1 -. However, the analysis with the GA-MLHD procedure identifies many alternative models that are equally predictive of tumour features. Consequently, many of the genes selected in these models may represent important biological pathways and therefore could provide important insights in the patho-physiology of the tumour. In order to identify associations between the collections of models with known functional

pathways, the selected gene signatures were mapped on the Ingenuity database using the web-based Ingenuity Pathway Analysis (IPA) tool (Palo Alto, www.ingenuity.com) which enables discovery, visualization, and exploration of biological interaction networks. This database store maps of canonical functional pathways and functional relationships supported by published literature and by protein-protein interaction data. This chapter focuses the discussion on the analysis of the models that are predictive of CP and are based on the molecular profile of normal cells. Nevertheless, supplementary tables report the results of network analysis in a tabular format for all datasets and all models developed.

The gene sets represented in the populations of models selected using the GA-MLHD procedure have been analyzed using IPA application. Gene lists represented in the model populations developed with normal or tumour expression data predicting CP or GS were uploaded into the application. Each gene identifier was mapped to its corresponding gene object in the Ingenuity Pathways Knowledge Base. Only genes found in this database were considered. These genes, called focus genes, were overlaid onto a global molecular network developed from information contained in the Ingenuity Pathways Knowledge Base. Networks of these focus genes were then algorithmically generated based on their connectivity according to the following procedure implemented in the IPA software application. The specificity of connection for each focus gene was calculated by the percentage of its connection to other focus genes. The initiation and the growth of pathways proceed from the gene with the highest specificity of connections. Each network had a maximum of 35 genes for easier interpretation and visual inspection. Pathways of highly interconnected genes were identified by statistical likelihood using the following equation:

$$Score = -\log_{10}\left(1-\sum_{i=0}^{f-1}\frac{C(G,i)C(N-G,s-i)}{C(N,s)}\right)$$

Where *N* is the number of genes in the genomic network, of which *G* is the total number of focus genes, for a pathway of *s* genes, *f* of which are included in focus genes. *C(n,k)* is the binomial coefficient. Pathways with a *Score* greater than 5 ($p<0.00001$) were selected for biological interpretation. GO analysis for identifying functional terms significantly enriched in the selected networks has been performed outside the IPA application using a Hyper-geometric test with FDR correction using the Benjamini and Hochberg procedure [171] as implemented in the Cytoscape plug-in BINGO [172].

Canonical pathway analysis was performed using the IPA tools and significance for the enrichment of the genes with a particular Canonical Pathway was determined by right-tailed Fisher's exact test with α=0.01 and the whole database as a reference set.

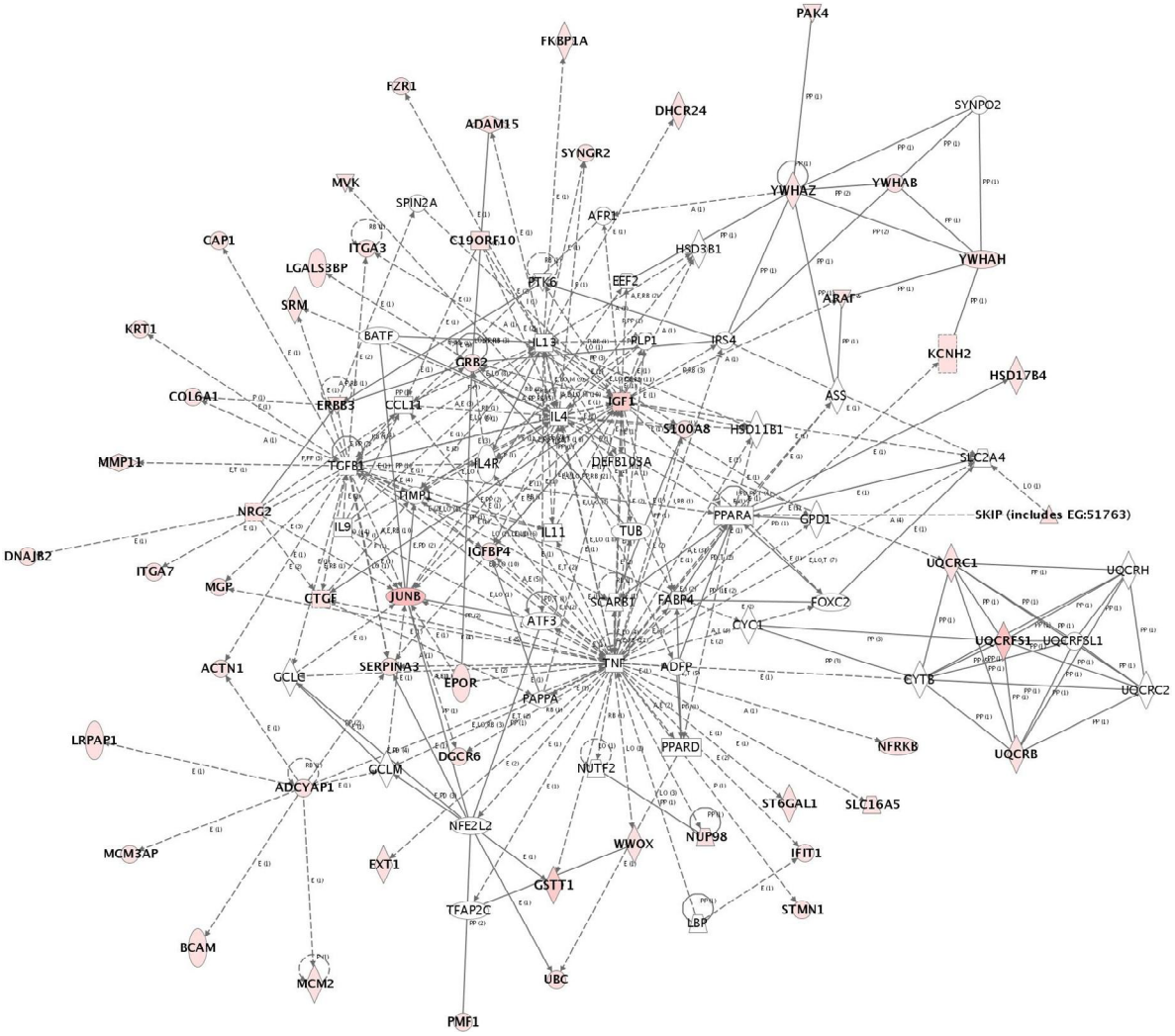
5.5.5 - Survival Analysis

Figure 5.10 describes the strategy used in this analysis. SAM was used to identify genes related to survival free of recurrence [43]. This method uses a proportional-hazard Cox model (see section 2.5.4.1 in Chapter 2) to relate genes whose expression values are associated to survival times considering censored data. SAM uses permutations of survival times to estimate a significance score (FDR) for each gene. The genes identified in initial analysis of the larger training dataset (Lapointe *et al.*) were selected using a threshold of FDR < 30%. The genes associated to survival in the molecular profile of normal and tumour tissues were mapped on Unigene to identify the Affymetrix probes from the Singh *et al.* dataset. Normal and tumour data were then used to cluster patients in the Singh *et al.* dataset. Clusters delimited by two clearly separated groups of patients were analysed by survival curves and their difference assessed by a log-rank test. The analysis was performed using different clustering algorithms to determine that results are independent of the clustering method (data not shown).

5.6 - Supplementary Material

Supplementary Table 5.1 - Gene Ontology analysis of the genes represented in the network shown in Supplementary Figure 5.1. Significant terms have been identified using a Hypergeometric test followed by FDR correction using the Benjamini and Hochberg False discovery rate (FDR) procedure. Terms with an FDR<=1% and with more than 4 genes in the term were listed in the table. Column identifiers are: N= number of genes in the GO term, FDR= False Discovery Rate.

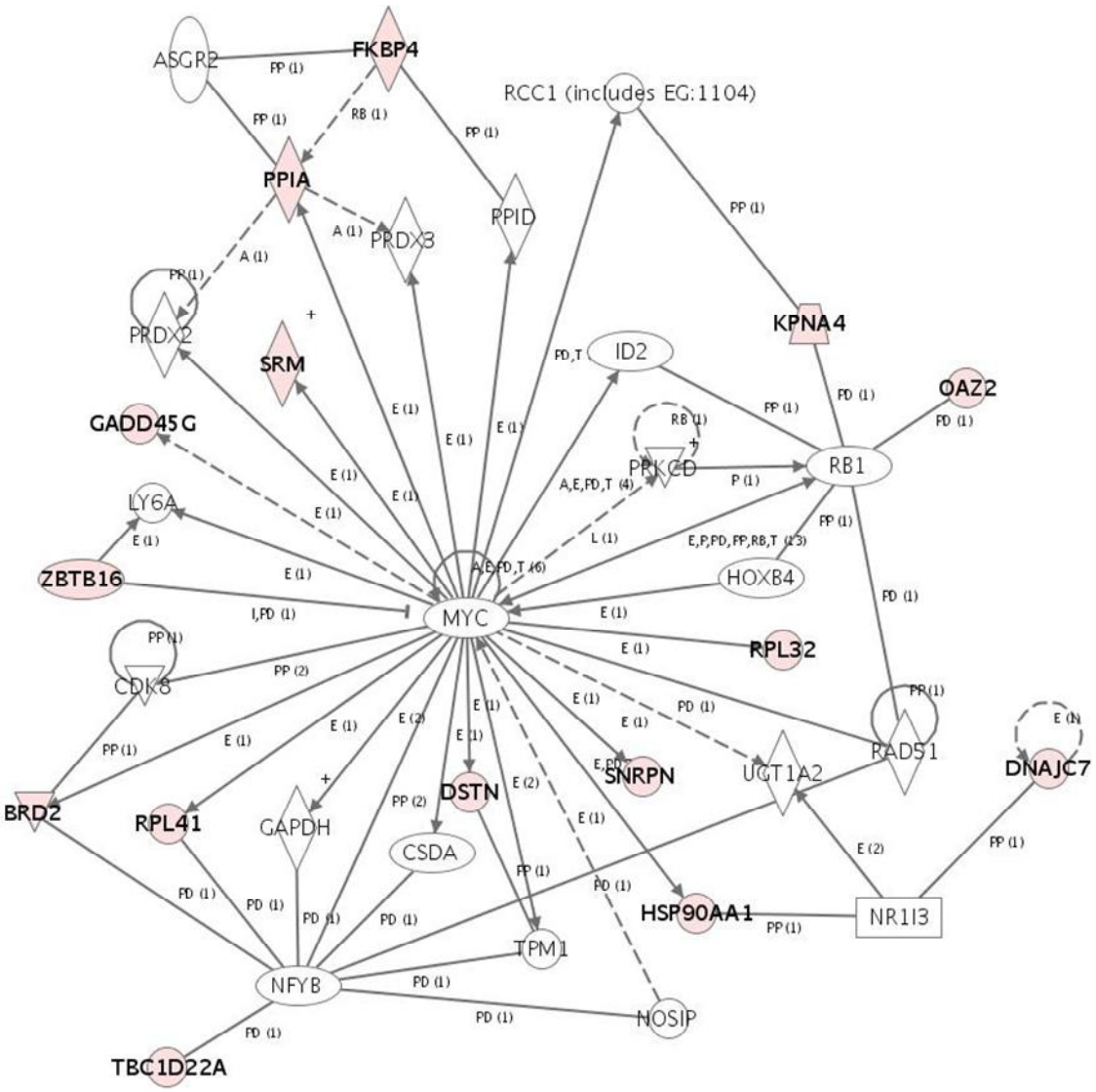
GO terms	*N*	*FDR*	*GO term*	*N*	*FDR*
cellular metabolism	57	6.60E-03	steroid metabolism	9	2.28E-05
protein binding	52	4.04E-06	electron transport	9	5.15E-03
signal transducer activity	29	6.17E-03	organelle envelope	8	3.19E-03
organismal physiological process	25	5.15E-03	envelope	8	3.35E-03
extracellular region	23	5.63E-06	extracellular matrix (sensu Metazoa)	8	3.35E-03
response to stress	19	4.82E-04	mitochondrial inner membrane	7	2.94E-04
extracellular region part	17	2.80E-05	organelle inner membrane	7	4.27E-04
immune response	17	2.82E-04	mitochondrial membrane	7	9.79E-04
defense response	17	9.79E-04	mitochondrial envelope	7	1.63E-03
response to biotic stimulus	17	1.32E-03	protein dimerization activity	7	3.19E-03
receptor binding	15	2.45E-04	mitochondrial part	7	7.63E-03
oxidoreductase activity	15	9.79E-04	oxidoreductase activity, acting on diphenols and related substances as donors, cytochrome as acceptor	6	1.55E-08
response to pest, pathogen or parasite	14	2.82E-04	oxidoreductase activity, acting on diphenols and related substances as donors	6	1.55E-08
response to other organism	14	2.82E-04	ubiquinol-cytochrome-c reductase activity	6	1.55E-08
response to wounding	13	2.80E-05	mitochondrial electron transport chain	6	5.48E-06
organelle membrane	13	2.82E-04	mitochondrial membrane part	6	1.93E-04
response to external stimulus	13	2.94E-04	hydrogen ion transporter activity	6	1.49E-03
extracellular space	12	4.46E-04	monovalent inorganic cation transporter activity	6	2.11E-03
lipid metabolism	12	2.64E-03	positive regulation of cell proliferation	6	2.94E-03
positive regulation of biological process	12	5.48E-03	regulation of organismal physiological process	6	5.15E-03
phosphorylation	12	6.60E-03	humoral immune response	6	5.26E-03
cellular lipid metabolism	11	1.84E-03	growth factor activity	6	6.17E-03
cell-cell signaling	11	4.09E-03	cholesterol metabolism	5	1.32E-03
organ development	11	5.02E-03	sterol metabolism	5	1.84E-03



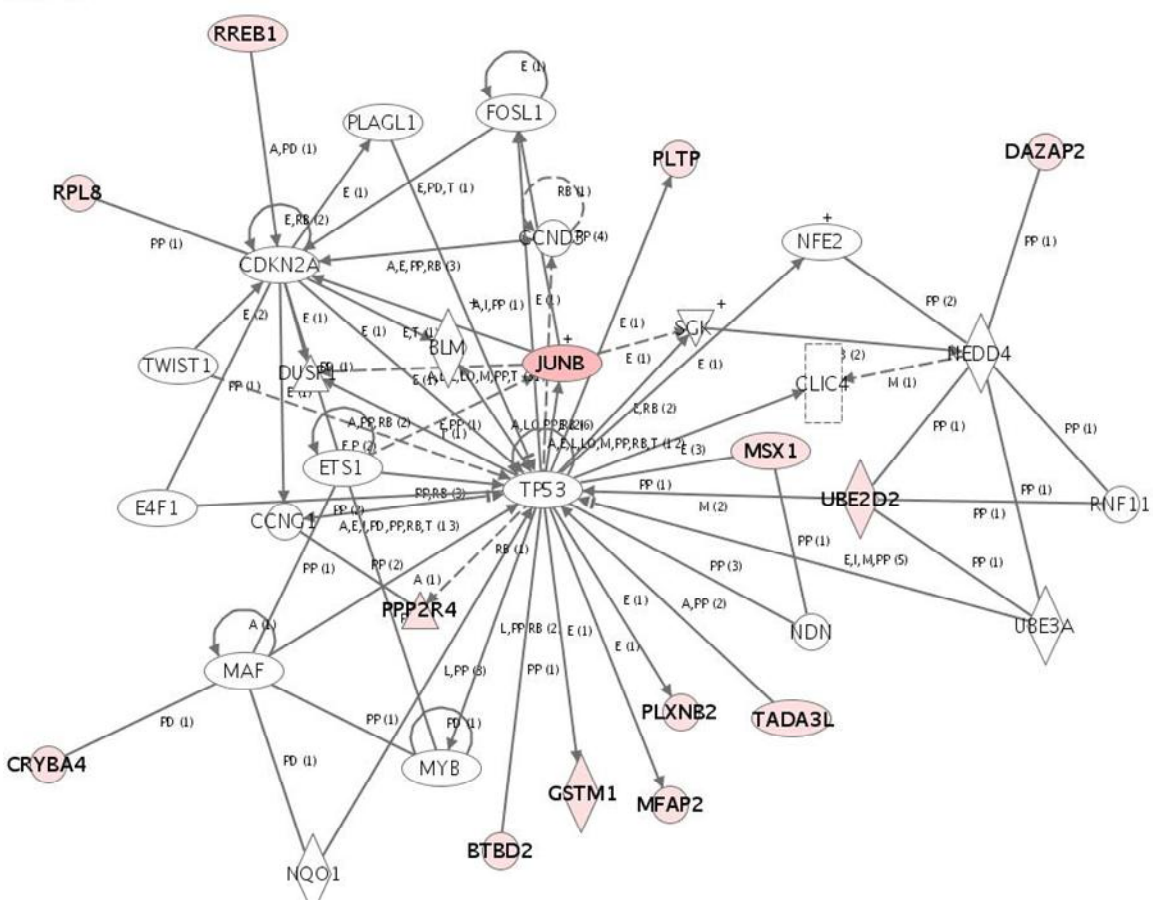
Supplementary Figure 5.1 - Network representing gene interactions between 53 of the genes selected in the statistical models (in pink) predictive of Capsular penetration based on the molecular state of normal cells. The network shows that many of the genes selected in the network are associated to cytokines and growth factor pathways. The network shown in the figure is the union of Networks 1, 3, and 4 (as annotated in Table 5.2). It complements its trimmed version shown in Figure 5.7. Keys to the symbols are as in Figure 5.7.

Supplementary Table 5.2 - Gene Ontology analysis of the genes represented in the network shown in Supplementary Figure 5.2. See Supplementary Table 5.1 for methods and annotations.

GO terms	*N*	*FDR*
cellular physiological process	29	9.88E-03
cellular metabolism	26	4.54E-03
primary metabolism	25	4.54E-03
protein binding	19	9.88E-03
intracellular organelle	19	6.77E-02
regulation of biological process	16	1.70E-02
regulation of cellular physiological process	13	4.02E-02
regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolism	11	2.76E-02
negative regulation of cellular process	6	3.09E-02



Supplementary Figure 5.2 - Network representing gene interactions between 15 of the genes selected in the statistical models predictive of CP based on the molecular state of normal cells. The network shows that many of the genes selected in the network are associated to the gene *c-myc*. The network shown in the figure annotated as network 5 in Table 5.2. It complements its trimmed version shown in Figure 5.8A. Keys to the symbols are as in Figure 5.7.



Supplementary Figure 5.3 - Network representing gene interactions between 13 of the genes selected in the statistical models predictive of Capsular penetration based on the molecular state of normal cells. The network shows that many of the genes selected in the network are linked to the oncogene p53. The network shown in the figure is annotated as network 7 in Table 5.2. It complements its trimmed version shown in Figure 5.8B. Keys to the symbols are as in Figure 5.8.

Supplementary Table 5.3 - Gene Ontology analysis of the genes represented in the network shown in Supplementary Figure 5.3. See Supplementary Table 5.1 for methods and annotations.

GO terms	*N*	*FDR*	*GO term*	*N*	*FDR*
physiological process	31	2.77E-02	regulation of transcription	14	1.74E-03
cellular physiological process	29	2.09E-02	regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolism	14	1.96E-03
Intracellular	27	3.23E-03	Transcription	14	2.03E-03
intracellular part	26	3.11E-03	regulation of cellular metabolism	14	2.68E-03
Metabolism	26	6.98E-03	regulation of metabolism	14	3.23E-03
primary metabolism	24	1.23E-02	transcription regulator activity	12	7.53E-04
cellular metabolism	24	1.85E-02	biopolymer metabolism	11	8.11E-02
intracellular organelle	22	2.09E-02	Development	10	2.77E-02
Organelle	22	2.09E-02	cell cycle	9	8.38E-04
Nucleus	20	7.53E-04	transcription factor activity	9	2.03E-03
intracellular membrane-bound organelle	20	1.85E-02	transcription from RNA polymerase II promoter	8	6.27E-04
membrane-bound organelle	20	1.85E-02	regulation of progression through cell cycle	7	1.41E-03
nucleic acid binding	18	9.79E-04	regulation of cell cycle	7	1.41E-03
protein binding	18	2.15E-02	negative regulation of cellular physiological process	7	6.14E-03
DNA binding	17	9.55E-05	negative regulation of physiological process	7	6.98E-03
regulation of cellular physiological process	17	2.03E-03	negative regulation of cellular process	7	9.33E-03
regulation of physiological process	17	2.56E-03	negative regulation of biological process	7	1.32E-02
regulation of cellular process	17	3.31E-03	protein dimerization activity	6	6.27E-04
regulation of biological process	17	6.65E-03	regulation of transcription from RNA polymerase II promoter	6	9.79E-04
nucleobase, nucleoside, nucleotide and nucleic acid metabolism	15	1.88E-02	sequence-specific DNA binding	6	7.23E-03
regulation of transcription, DNA-dependent	14	1.24E-03	cell proliferation	6	9.46E-03
transcription, DNA-dependent	14	1.41E-03	transcriptional activator activity	5	2.56E-03

Supplementary Table 5.4 - Significant Networks identified by IPA associated to GS tumour class from the population of models of the molecular profile of normal cells developed using the GA-MLHD procedure in Singh *et al.* dataset. Column identifiers as in Table 5.2.

NTW	*GENES*	*S*	*FG*	*FUNCTIONS*
1	CALR, CANX, CSNK1D, CSTA, CTNNA1, CTNNB1, DCT, DDX17, DUSP26, FCGRT, FGF9, FGF19 (includes EG:9965), GAS6, HSF4, HSPA1A, ILF3, JUND, MAPK3, MARCKSL1, NOTCH1, ORM1, P2RX7, PDIA3, PLAU, PMP22, QPCT (includes EG:25797), SERPINA5, SIM2, SORBS3, SREBF1, TAX1BP3, TFF3, TRIM28, USP9X, VCL	33	21	Cell-To-Cell Signaling and Interaction, Cell Death, Cellular Growth and Proliferation
2	AGRP, ANPEP, APBB1, APP, BCL2L1, BRD2, CCL2, CLSTN1, CNKSR1, DOK1, E2F4, IGF1, IGFBP4, KITLG (includes EG:4254), MAP2K4, MC2R, MGA (includes EG:23269), PHB, PRKCA, PRKCG, PRKCSH, PSMA1, PSMA2, PSMA4, PSMB1, PSMB5, RAF1, SRRM2, STMN1, SUPT4H1, TNF, UBC, UGCG, WDR1, ZNF267	22	16	Cell Death, Cell Cycle, Skeletal and Muscular Disorders
3	ADCYAP1, AP1B1, ASAH1, BASP1, CASP2, CCL5, CYP17A1, ECH1, EPHA2, EWSR1, FUS, GNL1, HLA-J, IFIT1, KRT8, LBP, LTBR, MAP2K6, NRP1, NRP2, PLXNA1, PSMD2, RBPMS, REL, RPSA, RXRA, SEMA3C, SEMA3F, SERPINA3, SF1, SMPD1, TNF, TRAF4, TUBB2B, YBX1 (includes EG:4904)	19	14	Cell Death, Connective Tissue Disorders, Cellular Movement
4	B2M, CAT, CD24, CD160, CTTN, FANCC, FCGR2A, FGR, GSTP1 (includes EG:2950), GYPC (includes EG:2995), HLA-A, HLA-C, HLA-E, HLA-G, IFNG, IGFBP4, JRK, KIR2DL4, MYCN, NALP12, PRDX2, PRKCA, PRKCB1, PRPF8, RAB5B, RGS10, RPS17 (includes EG:6218), SERPINA5, SF3A1, SIM1, SMARCB1, TAP1, TAPBP, TMEM109, VLDLR	19	14	Cell-To-Cell Signaling and Interaction, Hematological System Development and Function, Immune Response
5	BCL2L1, CAD, CAPN1, CCL2, CDKN1A, CEBPA, CKAP4, CSDE1, DDX17, DNMT1, EIF4A1, EIF4G1, EIF4G3, GGT1, H1FX, H2AFZ, HES1, IRF7, LGMN, LUM, MYC, NOTCH1, PABPC1, PCBP2, PCBP1 (includes EG:5093), PIAS2, REL, RPL6, SFRS4, SFRS12, STRAP, TEGT, TNFSF11, TTK, TUBB2A	19	14	Cellular Development, Hematological System Development and Function, Immune and Lymphatic System Development and Function
6	ADRB2, ANXA2, ARHGAP1, ARHGAP8, ARHGEF1, ATF1, BAIAP2, BCL2, BCL2L1, BNIP2, C1QL1, CCL2, CDC42, DIAPH1, DNMT1, ELAVL3, FOS, FUS, HRAS, IL3, IL1R1, MID1, NOTCH1, P4HB, PPP2CA, PRKCA, RHOA, RPL15, RPL27A, SLC30A3, SLC9A3R1, SOX4, SPTBN1, WASF2, YBX1 (includes EG:4904)	17	13	Cell Morphology, Cell Cycle, Cancer
7	AGT, BLM, CCL2, CCL13, COX4I1, DDOST (includes EG:1650), FAM38A, GEMIN6, GEMIN7, GPI, GPX4, GRN, HES1, HK2, HK3, IGFBP4, IL13, LSM2, LSM4, MAOB, PIAS2, PKM2, PML, POU2F1, PTGDS, RAD51L3, REL, S100A2, SMN1, SNRPE, SNRPF, SNRPG, SP1, SUMO3, TP53	17	13	Hematological System Development and Function, Immune and Lymphatic System Development and Function, Tissue Morphology
8	CD40, CD74, CSF2, DDT, H2-PB, HLA-DPA1, HLA-DPB1, IFNG, IGH-IA, IL8, IL15, IL1B, MAPK8, NFKB1, NME4, TNF, TNFRSF25, TXNRD1	6	5	Cell-To-Cell Signaling and Interaction, Hematological System Development and Function, Gene Expression

Supplementary Table 5.5 - Significant Networks identified by IPA associated to CP tumour class from the population of models of the molecular profile of normal cells developed using the GA-MLHD procedure in Lapointe *et al.* dataset. Column identifiers as in Table 5.2.

NTW	*GENES*	*S*	*FG*	*FUNCTIONS*
1	AGT, CBLB, CCDC99, CCNB2, CEACAM5, CEACAM1 (includes EG:634), COL4A1, COL8A1, CSPG2, CYP19A1, CYR61, DCLRE1A, FN1, GATA2, GHR, IGF1, IGFBP7, IL1RN, ITGA8 (includes EG:8516), LOX, LPL, MAP3K5, MAPK8, MYCN, NFE2, PDE3B, PDLIM2, PIK3CG, RFC3, SLC5A5, SORBS1, TITF1, TP53, UBE2C, WAC	67	33	Cellular Growth and Proliferation, Cell Morphology, Cell Death
2	AKT2, ANG, ARD1A, ASPH, CCL19, DACH1, EDIL3, FOS, FSCN1, GFAP, GNAI1, GSTM2, GSTP1 (includes EG:24426), IFNGR2, IGFBP6, IL11, IL16, IL4R, IRS4, ITPR3, LTBP2, NAT2, OGN, PANK1, PEX5L, POMC, PTEN, QKI, RPLP0 (includes EG:6175), RPS7, SSR4, TNF, WEE1, YWHAZ, ZAP70	19	14	Cellular Development, Connective Tissue Development and Function, Skeletal and Muscular System Development and Function
3	ALDH1A1, ARPC1A, CDH11, CNN1, COL5A2, CTNNB1, DDX5, EPHA7, EPHB3, FER (includes EG:2241), FRAS1, GRIP1, HES2, HOXC8, HRAS, IGFALS, IL1B, INS1, ISL1, JAK1, MEIS2, NALP2, PBX1, PKP1, PPM2C, RALA, RYK, SLC2A2, SOCS2, TF, THRSP, TJP1, TSPAN8, VHL, VIPR1	19	14	Tissue Development, Cellular Growth and Proliferation, Carbohydrate Metabolism
4	ACTN4, AFR1, AKT2, BEX1, BGLAP, CDCA7, CDH1, CDO1, DDX5, EP300, FABP1, FABP4, FOXP1, GTF2E2, IL11, ITGAM, LY6A, MAPK1, MDM2 (includes EG:246362), MST1R, MYC, MYCT1, MYH11, MYLPF, MYOD1, NCAM1, PLA1A, PPARG, PTMA, PTPRM, ST8SIA1, TGFB1, TJP1, TPM1, ZBTB17	16	12	Cellular Development, Cell Morphology, Cellular Growth and Proliferation
5	ADA, ATP2A1, BCL11A, CACNA1C, CACNA1D, CCL20, CEBPA, DEFB103A, FANCC, GAL, HSD11B1, IDE, IFNGR2, IGF1, IL2, IL4, IL6, IL13, IL3RA, IL4R, LCN2, MALT1, MRVI1, MS4A1, PAG1, PCLO, PDLIM2, PFKP, PLP1, PSME1, PTPN9, PTPRC, TK1, UCP1, WEE1	16	12	Cellular Development, Hematological System Development and Function, Immune and Lymphatic System Development and Function
6	CEBPA, CIITA, CNTN4, COL1A2, GRIA4, HLA-DQB2, HLA-DRA, IFNG, MTPN, MYC, NFYA, NFYB, NFYC, NSF, RFX1, RFX2, RFX3, RFX4, RFX5, SLMAP, STX16, TAF6, TAF9, TAF10, TAF12, TNF	5	5	Gene Expression, Endocrine System Disorders, Metabolic Disease

Supplementary Table 5.6 - Significant Networks identified by IPA associated to GS tumour class from the population of models of the molecular profile of normal cells developed using the GA-MLHD procedure in Lapointe *et al.* dataset. Column identifiers as in Table 5.2.

NTW	*GENES*	*S*	*FG*	*FUNCTIONS*
1	ADAM17, ARD1A, CD53, CDH16, CMA1, CSPG2, CYP1B1, FANCC, FAP, FBLN1, FGF10, FN1, GLB1, GNAQ, GNB5, GSTA2, GSTM2, GSTP1 (includes EG:2950), ITGA8 (includes EG:8516), ITGB1, KITLG (includes EG:4254), NAT2, NPNT, OGG1, POMC, PRKCB1, R9AP, RGS11, RGS13, RND3, RPS6, TACSTD2, TNFSF4, TPSB2, TYR	21	15	Cellular Movement, Embryonic Development, Cell-To-Cell Signaling and Interaction
2	CASP1, CD40LG, CTDSPL, CTNNB1, EGFR, F2, F5, FSCN1, FUT8, HLA-DQA1, IFITM1, KLK2, KLK3, MEF2C, NEDD4L, NRP1, NRP2, PLAT, PLXNA1, PODXL, PPARA, RB1, RBP4, RPLP0 (includes EG:6175), SEMA3C, SEMA6D, SERPINA3, SERPINA5, SERPINB6, SERPINE1, STX7, TAGLN3, TFAP2A, UGT2B4, YWHAZ	19	14	Cellular Movement, Tissue Development, Hematological System Development and Function
3	ACSL3, APOC2 (includes EG:344), CASP1, CCNA2, CCND3, CD14, CDH8, CHUK, CPA3, CX3CL1, EGF, ETS1, GABRE, GSK3B, HDAC3, IGFBP3, IKBKG, IL1RL1, MAPK8, MDFIC, MYB (includes EG:4602), NFKBIA, NOX4, NUAK1, OGG1, PDPN, PPP3CA, PRKCQ, RAB6IP2, RUNX1, SCG2, STK11, THAP7, TP53, UBE2C	18	13	Cell Death, Cellular Growth and Proliferation, Gene Expression
4	ABCG2, ADCYAP1, AKAP9, AR, ARF1, CCNA2, CFTR, CGA, CHUK, CLIC5, COL9A1, COL9A2, COL9A3, DQX1, FOS, GAS1, GBF1, GCG, GDF2, GSK3B, IL11, KBTBD10, KCNMA1, KRT8, KRT18, MGAT4A, PRKACA, PTPRN, PTPRN2, RASL11B, RIMS1, SNAP25, STX1A, STXBP6 (includes EG:29091), VIL2	16	12	Drug Metabolism, Molecular Transport, Small Molecule Biochemistry
5	CASP1, CCL22, CMA1, CTSL2, CX3CL1, DIAPH3, DNAJA4, FABP5, GAS6, IL2, IL13, IL16, IL1B, IL6ST, ISL1, JAK1, JAK2, KCNJ10, LHX3, MGP, MME, NALP2, NR5A2, PAG1, PFKP, PLSCR1, POU5F1, PRL, PTPN11, PYCARD, RXRA, SH2D1A (includes EG:4068), SRC, TNFSF4, TPT1	16	12	Cellular Movement, Inflammatory Disease, Cellular Growth and Proliferation
6	ACPP, AR, ARC, ASPH, ATAD4, BDNF, CD14, CENTG1, CHUK, CXCL14, DOK1, EIF4EBP1, ERBB2, FOLR1, HBD, IKBKB, IKBKG, IL8, KLK3, KRT19, LZTR1, NFKB2, NTRK2, PDE8A, PLCG1, PTEN, PXN, RTKN, SLC12A5, TAX1BP3, TGFBR1, TNFRSF10D, TNFRSF11B, TNFSF10, TSC22D3	14	11	Tissue Morphology, Organismal Survival, Cancer
7	ADAM17, AK3, CACNB1, CCL4, CCL6, CCNA2, CCND3, CCNE1, CHEK1, CHGN, CSN2, FTH1, GPM6B, HMGA1, IL4, IL9, IL1RL1, IL1RN, IL6ST, JUN, KLK3, MAOB, MAP2K4, MYC, NOG, PIGR, PIM1, PRKCD, PRNP, SHANK2, SST, SSTR2, SUPT3H, TEAD4, TNFSF11	12	10	Cell Death, Cellular Growth and Proliferation, Cellular Development

Supplementary Table 5.7 - Significant Networks identified by IPA associated to CP tumour class from the population of models of the molecular profile of tumour cells developed using the GA-MLHD procedure in Singh *et al.* dataset. Column identifiers as in Table 5.2.

NTW	*GENES*	*S*	*FG*	*FUNCTIONS*
1	AHNAK, BTG2, CCND1, CD14, CEBPA, DGCR6, DUSP1, ELAVL2, EPOR, FCGRT, FOSL1, GADD45B, GADD45G, GDF15, H1FX, HNRPD, IGFBP3, IGFBP4, IL6, ILK, JUNB, KCNMB1, KRT18, LITAF, MEF2D, MYH11, NPY, NR4A1, PDPK1, RPS6KB2, SCARB1, SPINT2, TFF3, TNFRSF1B, YWHAQ	64	35	Cellular Growth and Proliferation, Cancer, Cell Death
2	ACP1, ANG, APOE, ASAH1, BTG2, CDKN2A, CFTR, CTSB, CXCL5, CYBA, CYP7A1, DHRS3, FBP1, FCGR3A, GNL1, HLA-C, HNF4A, ITGAL, LTBR, MTTP, PIGR, PTX3, RASA1, RPL8, RREB1, SLC4A4, SMPD1, SOD3, SSR4, TALDO1, TM4SF1, TNF, TRAF4, ZAP70, ZNF202	18	15	Lipid Metabolism, Molecular Transport, Small Molecule Biochemistry
3	ANXA2, CCNB1, CCNE1, CDC42, ECHS1, EEF2, GAPDH, HNRPK, ID2, JRK, MYCN, NCL, NME1, PARD6A, PDLIM5, RPL28, RPL30, RPL31, RPL34, RPL35, RPL27A, RPS7, RPS20, RPS24, SDC1, SNCA, SORD, STAU1, TIAM1, TIMP2, TMED9, TMED10, TUBB2A, WT1, YWHAZ	18	15	Cell Cycle, Cancer, Cellular Movement
4	ADM, ARG1, BSG, CD2, COL1A2, CSF2, CTSB, DDX5, DOK2, EIF3S6, IFNG, ING1, IREB2, ITK, KIR2DS2, LCK, LILRB1, MYC, PNN, PRPF8, PRPH, PURB, RFX1, RFX2, RFXAP, RPL35, RPS15A, SFRS4, SFRS12, SLAMF1, TREM3, TYROBP, YBX1, YWHAB, ZFP36	18	15	Cellular Growth and Proliferation, Hematological System Development and Function, Immune Response
5	APOE, BLM, COMT, DAXX, DDB2, DDT, ERCC1, ERCC5, EXO1, GDF15, GH1, HIPK1, HIPK2, KLK2, KLK3, MLH1, MSH2, MSH6, MUTYH, PA2G4, PLTP, PMAIP1, PMS2, POU4F1, RAD51L3, RFC1, S100A2, SPN, TERF2, TMPRSS2, TMSL8, TP53, UBB, XPC, XRCC2	14	13	DNA Replication, Recombination, and Repair, Cancer, Gastrointestinal Disease
6	ALAS2, BTG1, CALM2, CAMK2G, CD2, CLNS1A, CRKL, DIAPH1, DOK1, EDG2, EFNB1, FCGR2B, FHL1, FHL2, FHL3, INPP5D, ITGA7, ITGB5, ITGB6, ITGB7, KRT1, KRT17, LSM7, MAPK1, PTPRH, RAB1A, RABAC1, RAP1GA1, RHOA, RHOB, SERPINH1, SNRPD3, STAT1, TGFB1, WDR77	14	13	Organismal Survival, Cell Death, Neurological Disease
7	ADM, ARHGDIA, ATN1, CASP3, DAP, DHRS7, DNAJB6, ERBB2, GPAA1, GRB10, IGHMBP2, ITK, NDUFC1, NFYB, PHYH, PHYHIP, PTEN, RAC2, SETD7, SRC, TAF2, TAF4, TAF7, TAF10, TAF11, TAF12, TAF13, TAF15, TAF7L, TAX1BP3, TBN, TBP, TMEM87A, TNK2, TRAM1	14	13	Cellular Development, Cellular Growth and Proliferation, Connective Tissue Development and Function
8	ADAM12, ADM, APP, CALCRL, CAMKK2, CCL20, CCNF, CST3, CTSB, DRG2, EDN1, FGF10, GLI1, HDGF, HSPA5, HYOU1, IAPP, IRS1, M6PR, NME2, PIK3R3, PTN, RAMP3, RBL2, RELA, RETN, RPS6KB2, SDHA, SDHB, SDHC, SDHD, SERPINB2, TIMP3, VEGF, VEGFB	14	13	Cellular Growth and Proliferation, Cancer, Cell Cycle
9	ABCB1, ADH7, APOC3, APOD, ARF4, CDK5R1, CRAT, CTNNB1, DHCR24, G6PC, G6PD, GCLC, GCLM, HAX1, JUN, KARS, LLGL2, LSS, MITF, MTTP, MVD, NFE2L2, PARD6A, PKD1, PKD2, PPARGC1B, PRKCI, PXDN, SAA1, SLC1A4, SREBF1, SREBF2, TM7SF2, UBC, WT1	13	12	Gene Expression, Cancer, Genetic Disorder
10	ARCN1, ARG1, CDH1, CDH4, COPA, COPB, COPB2, COPE, COPG, COPG2, COPZ1, COX6C, CTSB, EIF4B, ENC1, FURIN, G6PD, GDF15, HIRA, IL15, IL1B, ITGB7, KRT19, LRPAP1, PAX7, PSG1, PTPRM, RB1, RPS6KA1, SORL1, STK11, TCF8, TGFB1, TGM2, USP19	13	12	Cellular Assembly and Organization, Cell Death, Neurological Disease

Supplementary Table 5.8 - Significant Networks identified by IPA associated to GS tumour class from the population of models of the molecular profile of tumour cells developed using the GA-MLHD procedure in Singh *et al.* dataset. Column identifiers as in Table 5.2.

NTW	*GENES*	*S*	*FG*	*FUNCTIONS*
1	BGN, BIN1, C7, C1R, CCR7, CD74, CSF1, DAXX, FHL2, FUS, GSK3A, HLA-DPB1, HLA-DRA, HMGN2, HSF1, IL1B, ING1, INHBB, ITGB1, KCNH2, KLK2, KLK3, MAZ, MDK, MYBPC1 (includes EG:4604), NPY, OAZ1, ODC1, PIM1, PSME2, RBP1, RPSA, SF1, TIMP3, TPT1	63	35	Cellular Growth and Proliferation, Cellular Movement, Hematological System Development and Function
2	ACPP, CD34, CDH1, CDKN1A, COL6A3, DDR1, DPT, ERBB2, GADD45G, HDAC1 (includes EG:3065), HYOU1, ID1, ID2, IER2, IGFBP3, KLF6, KRT18, LGALS3, LUM, MSX1, MYL9, NR4A1, PTPRF, PTPRM, RELA, SERPINF1, SPARCL1, TBX19, TFF3, TMSB4X, TNFRSF10B, VEGF, VIM, WFS1, ZFP36L2	63	35	Cancer, Cellular Growth and Proliferation, Cell Death
3	BAT1, CDKN2A, CHAF1A, CSDE1, CTCF, CTSD, DBI, DBN1, DDX17, G3BP, GSTP1 (includes EG:2950), GYPC (includes EG:2995), ID1, MAP4, MAZ, MBD2, MYC, NRG1, PRKCB1, PTBP1, RFX1, RFX2, RFX3, RPL8, RPL19, RPL30, RPL35, RPS19, S100A6, SGSH, STRAP, TFF1, TSPO, TUBB2A, ZFP36	22	18	Cancer, Tumor Morphology, Immunological Disease
4	ADAM10, BMP4, CBLC, CCNG1, CEACAM1 (includes EG:634), CRAT, CTNNB1, DLG5, EGF, EIF4A2, FLNA (includes EG:2316), FSHR, FST, GNAI2, GOLGA2, GORASP1, HSD11B1, ID2, IDI1, IGF1, IRS1, MC2R, PHLDA2, PLA2G2A, RAP1GAP (includes EG:5909), ROCK2, RPSA, SFRS3, SIM2, SLC1A5, SORBS3, SP7, STXBP1, TMED2, XRCC2	21	17	Connective Tissue Development and Function, Organ Development, Reproductive System Development and Function
5	ADCYAP1, AMACR, ARF5, AZIN1, BCAM, COX1, COX17, COX4I1, COX4I2, COX6A2, COX6B1, COX6B2, COX6C (includes EG:1345), COX7B, COX7B2, COX8A, COX8C, CSK, DEGS1, EGFR, FKBP4, GIPC1, GNRHR, IDH2, IGF1R, MAPK1, PEA15, PHYH, PHYHIP, PML, RPL24, RPL37A, SLC2A1, SUMO3, TCOF1 (includes EG:6949)	21	17	Cancer, Cell Cycle, Renal and Urological Disease
6	ANAPC2, ANK3, AOF2, ARID4B, ATRX, BRD2, CD63, E2F4, HBB (includes EG:15127), HDAC1 (includes EG:3065), HMG20B, IL3, KIAA0101, MT1G, MYCN, NFYB, PHB, PHF21A, POLR2L, PTTG1, PTTG1IP, RCOR1, RPL4, RPL11, RPL41, RPL27A, RPS5, RPS19, RPS20, SCN2A1, SLIT3, SMARCA3 (includes EG:6596), SRRM2, STMN1, UBB	19	16	Gene Expression, Protein Synthesis, Cellular Growth and Proliferation
7	ANXA11, B2M, CD74, CFD, CYP17A1, FOSL2, GBP1, GCH1, GPAM, GPX1, HSD11B1, IDH1, IFITM1, IL16, JUND, LEP, LTC4S, NR4A1, NTS (includes EG:57303), OPLAH, PPP1R12B, PTS, RBPMS, SEPP1, SFTPB, SMARCA4, SOD3, SPRR1B (includes EG:6699), TACSTD2, TNF, TREM2, UBE2H (includes EG:7328), UCP1, UCP3, ZFP36	17	15	Lipid Metabolism, Molecular Transport, Small Molecule Biochemistry
8	ADA, APP, CCNG1, CCT2, CTCF, CTSH, DHCR24, DIAPH1, DNASE1, ERCC1, ERCC5, HERPUD1, IL13, LTBP1, M6PR, MYB (includes EG:4602), MYBL2, MYOZ3, NGFR, NR2F2, PFKFB2, PHLDA1, PPP2R4, PPP3CA, PRKCG, RHOA, ROCK1, S100A2, SERPINF1, SLC9A1, TACSTD1, TNFAIP2 (includes EG:7127), TP53, TRIO, YWHAG	16	14	Cell Death, Cancer, Reproductive System Disease
9	ADRA1A, ADRA1B, ADRA1D, ALPP, ARF1, CAMK4, CAMK2G, CD46, DDB2, ECGF1, EGR2, EIF2AK2, FOS, HAS1, HINT1 (includes EG:3094), HSP90AB1, HTR2A, ITGAE, KARS, KLK3, MITF, MPZ, MST1R, MYLK, NR3C1, PMP22, PRKCG, PSMB1, RPS9, SELENBP1, SPTBN1, STAT1, TGFB1, TGTP, UNC45A	16	14	Immunological Disease, Inflammatory Disease, Respiratory Disease

Supplementary Table 5.9 - Significant Networks identified by IPA associated to CP tumour class from the population of models of the molecular profile of tumour cells developed using the GA-MLHD procedure in Lapointe *et al.* dataset. Column identifiers as in Table 5.2.

NTW	*GENES*	*S*	*FG*	*FUNCTIONS*
1	ADA, ANK3, CBLB, CISH, CYR61, DEFB103A, FBXO32, FLJ20701, FLOT1, GHR, GRB10, HSD11B1, HSD3B1, HTR2C, ID2, IFNGR2, IGF1, IL13, IL3RA, JAK2, LCP2, LPA, MAP3K5, MT1X, PDE3B, PDLIM2, PHLDA1, POU1F1, PPP3CA, PTD004, PTPRC, SOCS2, SORBS1, STS-1, TUB	23	16	Gene Expression, Cellular Development, Cellular Growth and Proliferation
2	APOC2 (includes EG:344), BTG2, CCL19, CEBPA, CFB, CLASP1, CYLN2, DDC, DSC3, GCH1, GHR, HMGN3, HSD11B1, IL1B, IL3RA, LCN2, MAPRE3, MGP, MMP8, MTPN, NALP2, OAS2 (includes EG:4939), OASL, PAPPA, PENK1, PGD, PTGES, RND1, SEMA3C, SLMAP, SPARC, THRB, TNF, TRIB1, UBXD5	21	15	Lipid Metabolism, Small Molecule Biochemistry, Amino Acid Metabolism
3	ACSL3, BTG2, CCNA1, CCNG1, CDKN2A, CKM, GH1, GPT, HLA-DQA1, HMMR, ID2, IFI16, IGFBP7, IGSF4 (includes EG:23705), INSM1, LPHN2, MDM2 (includes EG:246362), MELK, PCNA, PIAS4, POLD3, POLS, PRIM2A, PVRL3, RET, RNF19, SOD2, TBX3, TCF3, TFAP2C, TP53, TPX2, TWIST1, WRN (includes EG:7486), ZNF202	21	15	Cell Cycle, Cancer, Dermatological Diseases and Conditions
4	C1QBP, CACNA1A, CACNA1B, CACNA1C, CACNA1D, CTNNB1, CYR61, DVL3, F2, FZD8, GABRB1, GABRB3, GABRD, GNB2L1, GPX2, IL8, KCNMA1, KLK2, LRP6, MUC1, PLAT, PRKCB1, PRKCD, PRKCDBP, PRKD1, PTGS2, QPCT (includes EG:25797), RAB3B, RBP4, RIMS1, SERPINA5, SFRP4, SLC12A2, STK39, TGM1	20	14	Organismal Injury and Abnormalities, Cell-To-Cell Signaling and Interaction, Nervous System Development and Function
5	ACOX1, AGXT, BTG2, CDC6, CDC25B, CEACAM5, CRYZ, CYP1B1, EEF2, ESR2, GFAP, GLS, GNRH1, ID2, IL6, JUN, LGALS4, MAPK14, NRAP, PPARA, PRDX1, PREP, PTGES, PTGS2, REXO4, RHOH, SOCS2, SOD1, SYNPO2, TLN1, TPSD1, TXN, UCP1, WEE1, YWHAZ	18	13	Cell Cycle, Cancer, Cell Death
6	ADM, ANXA3, AREG, ARF1, CAMK1D, CD40LG, CFLAR, CXCL5, CXCL14, CYP2E1, FOS, GBF1, GCLC, GUSB, HRAS, ID2, IL8, IL18RAP, IVL, MPO, MT1H, MUC13, PRRX1, PTGER1, PTGER4, PTGS2, PTHLH, RGS1, RPS6, SH2B2, SPARC, TACR1, THBS4, TPSD1, TSLP	18	13	Cellular Growth and Proliferation, Cell-To-Cell Signaling and Interaction, Cardiovascular System Development and Function
7	AKT2, CD9, CDKN1A, DPT, EDN2, ERBB2, ERBB3, FAP, FYB, GHR, GMDS (includes EG:2762), IGFBP6, IGFBP7, IGSF8 (includes EG:93185), ITGA7, ITGB1, L1CAM (includes EG:3897), MAPK1, MKI67, PDHA1 (includes EG:5160), PDK1, PFKFB3, PIK4CA, PIM1, PRAME, PTEN, PTGS2, ROCK2, S100A4, SCAP1, SIAH1, SPARC, UBE2E3, VIM, WRN (includes EG:7486)	16	12	Cellular Movement, Cellular Growth and Proliferation, Cancer

Supplementary Table 5.10 - Significant Networks identified by IPA associated to GS tumour class from the population of models of the molecular profile of tumour cells developed using the GA-MLHD procedure in Lapointe *et al.* dataset. Column identifiers as in Table 5.2.

NTW	*GENES*	*S*	*FG*	*FUNCTIONS*
1	AHNAK, BTG2, CCND1, CD14, CEBPA, DGCR6, DUSP1, ELAVL2, EPOR, FCGRT, FOSL1, GADD45B, GADD45G, GDF15, H1FX, HNRPD, IGFBP3, IGFBP4, IL6, ILK, JUNB, KCNMB1, KRT18, LITAF, MEF2D, MYH11, NPY, NR4A1, PDPK1, RPS6KB2, SCARB1, SPINT2, TFF3, TNFRSF1B, YWHAQ	64	35	Cellular Growth and Proliferation, Cancer, Cell Death
2	ACP1, ANG, APOE, ASAH1, BTG2, CDKN2A, CFTR, CTSB, CXCL5, CYBA, CYP7A1, DHRS3, FBP1, FCGR3A, GNL1, HLA-C, HNF4A, ITGAL, LTBR, MTTP, PIGR, PTX3, RASA1, RPL8, RREB1, SLC4A4, SMPD1, SOD3, SSR4, TALDO1, TM4SF1, TNF, TRAF4, ZAP70, ZNF202	18	15	Lipid Metabolism, Molecular Transport, Small Molecule Biochemistry
3	ANXA2, CCNB1, CCNE1, CDC42, ECHS1, EEF2, GAPDH, HNRPK, ID2, JRK, MYCN, NCL, NME1, PARD6A, PDLIM5, RPL28, RPL30, RPL31, RPL34, RPL35, RPL27A, RPS7, RPS20, RPS24, SDC1, SNCA, SORD, STAU1, TIAM1, TIMP2, TMED9, TMED10, TUBB2A, WT1, YWHAZ	18	15	Cell Cycle, Cancer, Cellular Movement
4	ADM, ARG1, BSG, CD2, COL1A2, CSF2, CTSB, DDX5, DOK2, EIF3S6, IFNG, ING1, IREB2, ITK, KIR2DS2, LCK, LILRB1, MYC, PNN, PRPF8, PRPH, PURB, RFX1, RFX2, RFXAP, RPL35, RPS15A, SFRS4, SFRS12, SLAMF1, TREM3, TYROBP, YBX1, YWHAB, ZFP36	18	15	Cellular Growth and Proliferation, Hematological System Development and Function, Immune Response
5	APOE, BLM, COMT, DAXX, DDB2, DDT, ERCC1, ERCC5, EXO1, GDF15, GH1, HIPK1, HIPK2, KLK2, KLK3, MLH1, MSH2, MSH6, MUTYH, PA2G4, PLTP, PMAIP1, PMS2, POU4F1, RAD51L3, RFC1, S100A2, SPN, TERF2, TMPRSS2, TMSL8, TP53, UBB, XPC, XRCC2	14	13	DNA Replication, Recombination, and Repair, Cancer, Gastrointestinal Disease
6	ALAS2, BTG1, CALM2, CAMK2G, CD2, CLNS1A, CRKL, DIAPH1, DOK1, EDG2, EFNB1, FCGR2B, FHL1, FHL2, FHL3, INPP5D, ITGA7, ITGB5, ITGB6, ITGB7, KRT1, KRT17, LSM7, MAPK1, PTPRH, RAB1A, RABAC1, RAP1GA1, RHOA, RHOB, SERPINH1, SNRPD3, STAT1, TGFB1, WDR77	14	13	Organismal Survival, Cell Death, Neurological Disease
7	ADM, ARHGDIA, ATN1, CASP3, DAP, DHRS7, DNAJB6, ERBB2, GPAA1, GRB10, IGHMBP2, ITK, NDUFC1, NFYB, PHYH, PHYHIP, PTEN, RAC2, SETD7, SRC, TAF2, TAF4, TAF7, TAF10, TAF11, TAF12, TAF13, TAF15, TAF7L, TAX1BP3, TBN, TBP, TMEM87A, TNK2, TRAM1	14	13	Cellular Development, Cellular Growth and Proliferation, Connective Tissue Development and Function
8	ADAM12, ADM, APP, CALCRL, CAMKK2, CCL20, CCNF, CST3, CTSB, DRG2, EDN1, FGF10, GLI1, HDGF, HSPA5, HYOU1, IAPP, IRS1, M6PR, NME2, PIK3R3, PTN, RAMP3, RBL2, RELA, RETN, RPS6KB2, SDHA, SDHB, SDHC, SDHD, SERPINB2, TIMP3, VEGF, VEGFB	14	13	Cellular Growth and Proliferation, Cancer, Cell Cycle
9	ABCB1, ADH7, APOC3, APOD, ARF4, CDK5R1, CRAT, CTNNB1, DHCR24, G6PC, G6PD, GCLC, GCLM, HAX1, JUN, KARS, LLGL2, LSS, MITF, MTTP, MVD, NFE2L2, PARD6A, PKD1, PKD2, PPARGC1B, PRKCI, PXDN, SAA1, SLC1A4, SREBF1, SREBF2, TM7SF2, UBC, WT1	13	12	Gene Expression, Cancer, Genetic Disorder
10	ARCN1, ARG1, CDH1, CDH4, COPA, COPB, COPB2, COPE, COPG, COPG2, COPZ1, COX6C, CTSB, EIF4B, ENC1, FURIN, G6PD, GDF15, HIRA, IL15, IL1B, ITGB7, KRT19, LRPAP1, PAX7, PSG1, PTPRM, RB1, RPS6KA1, SORL1, STK11, TCF8, TGFB1, TGM2, USP19	13	12	Cellular Assembly and Organization, Cell Death, Neurological Disease

CHAPTER 6

Inference of Networks representing Cell to Cell Interaction

6.1 - Introduction

The previous chapter discussed how the interaction between different cell types is of fundamental importance in maintaining the normal tissue homeostasis and plays a major role in the development of pathological conditions such as cancer. Most of our understanding of cell to cell communication comes from studies addressing the role of stromal cells in shaping the microenvironment. However, our initial results suggest that the role of normal epithelial cells in influencing tumour physiology may not be negligible.

In particular it has been demonstrated that it is possible to predict tumour features from the molecular profile of adjacent normal cells. This chapter describes the application of network inference methods to deduce the structure of cell to cell communication networks in prostate cancer. A number of approaches for network inference have been developed and applied to varieties of biological systems. Information theoretical approaches have been demonstrated to be effective for inference from large scale datasets whereas probabilistic approaches, although potentially more powerful are only effective with relatively small datasets. Among these last class of methods, relevance networks (RN) are a simple but effective network inference approach that has been used to identify gene regulatory networks in a variety of experimental systems [173; 174]. Its performance is comparable with bayesian networks and mutual information based approached such as ARACNE.

Here, a methodology developed to infer cell to cell communication networks is described. This methodology is somehow derived from RN. By applying this methodology to a gene expression profiling dataset representing paired normal and tumour tissue, it will be shown that the identified genes provide insights into the complex molecular mechanisms implicated in normal-tumour interaction. The analysis of cell to cell interaction networks shows the existence of a relatively large number of genes that are associated to a directional effect (referred here as *polarization*). Of particular interest is the existence of signals generated by normal epithelial cells that are associated to transcriptional networks in the tumour cells. Experimental validation for one of the genes associated to polarization reveals a potentially new tumour suppressor gene in prostate cancer.

Although, this study is focused on a prostate dataset published by Singh *et al.*, analysis of other datasets suggests that the methodology is generally applicable.

6.2 - Results

6.2.1 - Inferring Gene Networks Representative of Cell to Cell Communication in Prostate Cancer

A RN based approach was applied to a dataset of paired normal and tumour tissue samples with the aim to infer the existence of significant connections between normal and tumour expressed genes. In the implementation used in this study, a Spearman correlation coefficient was employed. The significance of the interaction was established calculating a FDR [37] (see section 6.5.2 - in Methods) based on the distribution of correlation coefficients from a bootstrap dataset. Figure 6.1 (and Supplementary Figure 6.1) shows the distribution of correlation coefficients in normal cells, tumour cells, and across normal and tumour cells in relation to the distribution of

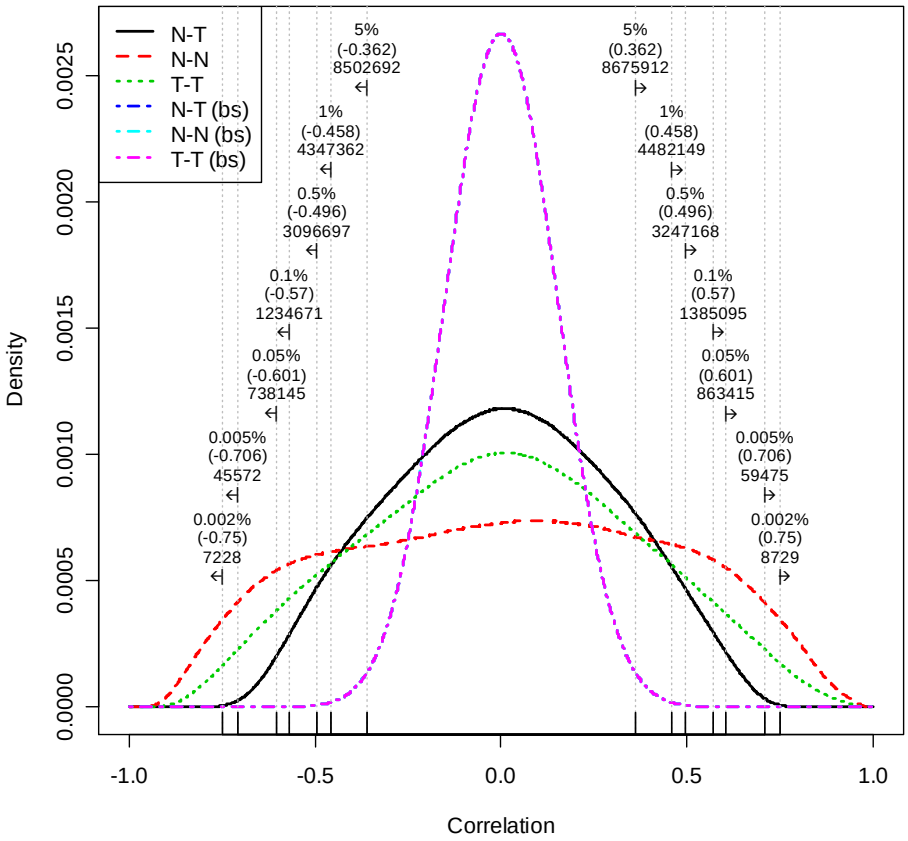


Figure 6.1 – Distribution of non-parametric correlations in Singh *et al.* dataset and their significance estimations. Distributions are shown by the type of gene-gene pairs (N-T, N-N, T-T). Their corresponding bootstrap distributions are also shown (the three are overlapped). Labels show FDR in percentage (%), the corresponding correlation cut-off value in brackets, and finally the number of correlations larger or smaller than the cut-off. The distributions in other datasets are shown in Supplementary Figure 6.1. The boostrap distributions (bs) N-T, N-N, and T-T are over imposed.

correlation coefficients in the randomized data. This simple characterization of the correlation structure that exist in the normal and tumour cells and in the interface between the two reveal that the degree of connectivity between the two adjacent cell types is sparser than the degree of connectivity within a normal or tumour tissue.

6.2.2 - The Definition of a Polarized Signal in Cell-to-Cell Communication Networks

The analysis and biological interpretation of a network representing gene interactions between normal and tumour epithelial cells is complex. Therefore, an approach was designed here to identify sub-networks that would be representative of a particularly interesting biological scenario. The study was focused in a scenario represented by a

directional effect. In this scenario, a given gene expressed in normal cells is characterized by a high number of connections with genes expressed in tumour cells whereas the same gene, expressed in tumour cells, would not have many connections with genes expressed in normal cells (a schema describing this scenario is shown in Figure 6.2). Such a scenario is biologically relevant and plausible. For example, any paracrine signalling, where released signals affects nearby cells, fits in this scenario.

The magnitude and direction of this effect is represented by defining a *polarization* index for a given gene *i* as:

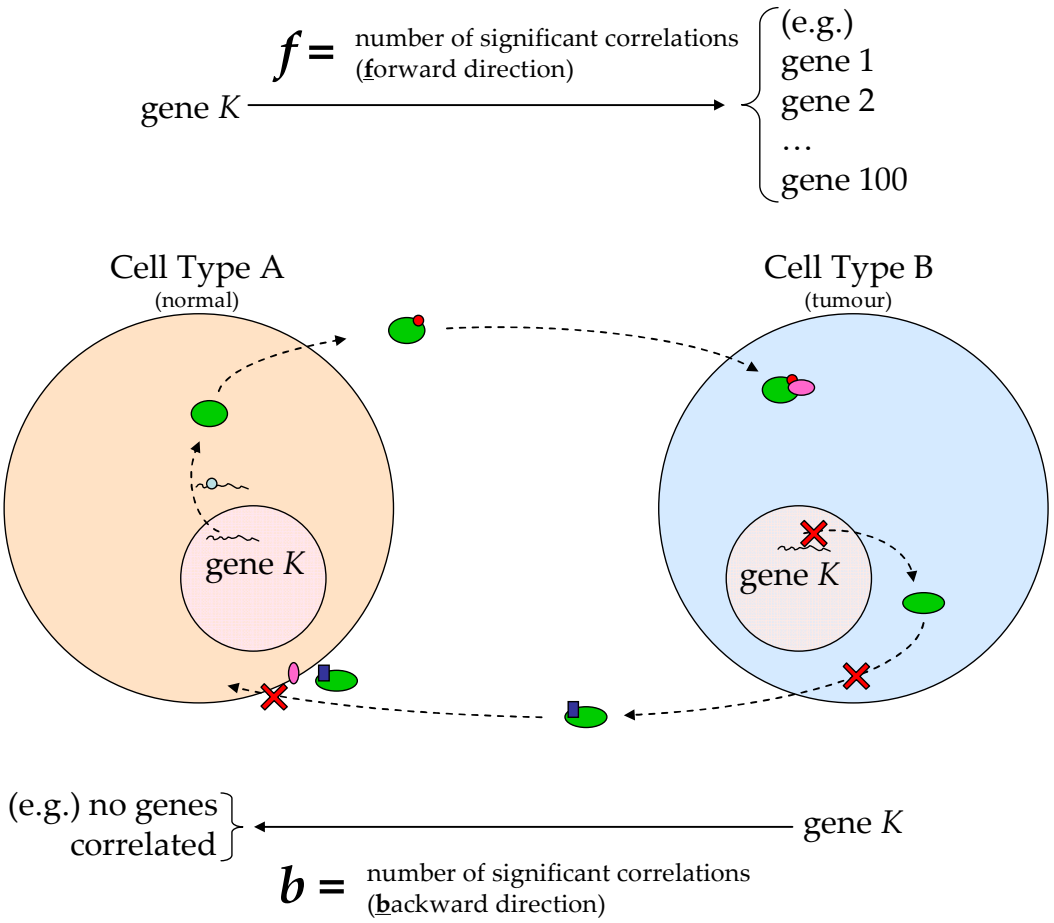


Figure 6.2 - Concept and estimation of polarization index (*pol*). A gene *k* expressed in cell type A (normal cells in this figure) could be correlated with several genes in cell type B (tumour cells in this figure), however, the same gene *k* expressed in cell type B has no correlations in the opposite direction, either because it is not expressed, not exported, not activated, defective, or receptor is not present. Candidate genes would have *pol* values close to +/- 1 depending on the direction of polarized correlations. Any other scenario would generate *pol* values far from +/- 1.

$$pol_i = \frac{f_i - b_i}{f_i + b_i + \varepsilon}$$

where f is the number of connections between gene i expressed in the normal tissue and genes expressed in the tumour tissue, b is the number of connections between the same gene i and genes expressed in the tumour tissue, and ε is a small constant designed to stabilize the *pol* ratio for small numbers of f and b. *pol* carry a number of desirable properties. The value of *pol* is proportional to the effect, the sign of *pol* gives the direction of the effect, and the metric tend to -1 and +1 for f or $b \gg 0$. As previously stated, the empirical distribution of correlations observed by random chance in bootstrap datasets is used to estimate an FDR value for every pair-wise connection in the network. Sparse networks are then generated by thresholding the adjacency matrix using an arbitrary FDR cut-off. Figure 6.3 shows the *pol* distribution for different FDR

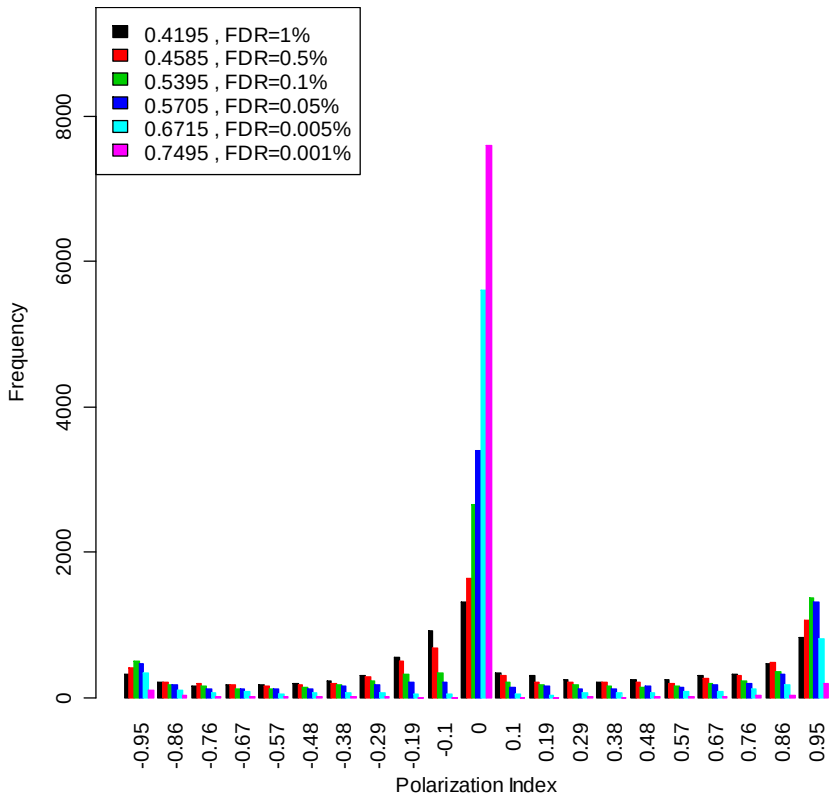


Figure 6.3 – *pol* distribution for Singh *et al.* dataset at various FDR correlation cut-offs. *pol* distributions for other datasets are shown in Supplementary Figure 6.2.

thresholds. The figure displays a tri-modal distribution with a relatively high frequency of highly polarized genes. The shape of the distribution is independent of the FDR threshold. A large proportion of these genes are therefore stable irrespective of the cut-off threshold (Figure 6.4). Similar results were observed in another two datasets though these peaks were not observed in three datasets due to the lack of significant correlations (Supplementary Figure 6.2). To select a cut-off point, the dependency on the number of genes having high *pol* with the correlation cut-off and FDR has been estimated for the datasets studied (Figure 6.5). As in any gene selection procedure, the choice of a good threshold is a trade-off, in this case, between the number of selected genes and the expected number of false correlations. For gene selection and experimental analysis derived from Singh *et al.* dataset, a correlation cut-off of 0.75 was choosen because at this point the number of polarized genes is maximum (5%) whilst the FDR is minimum (~0.001%, 1 false correlation out of 100,000, see Figure 6.5). In practice, to highlight only genes with large number of correlations, if the absolute value of *f-b* is less than δ=20, *pol* is set to zero.

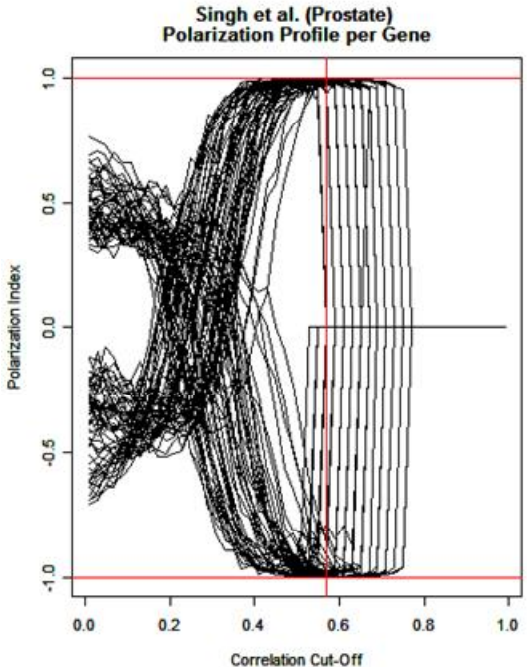


Figure 6.4 – Examples of genes whose polarization is stable in a wide range. Genes were selected by high area under the polarization curve (vertical axis) depending on the correlation cut-off (horizontal axis).

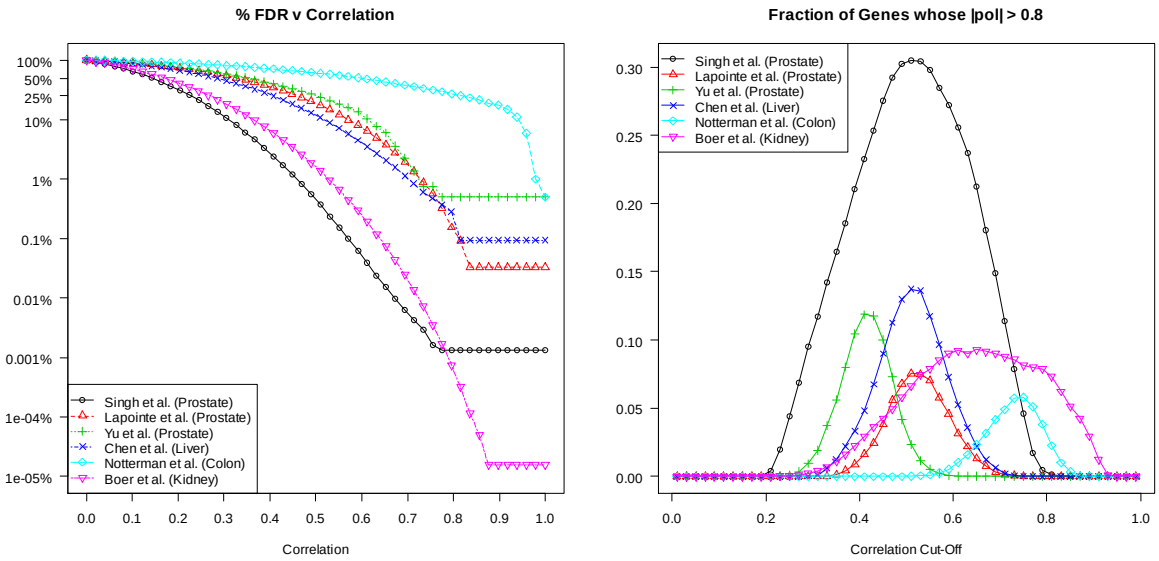


Figure 6.5 – Number of highly polarized genes and their FDR estimations. Left panel shows the dependency of FDR on the correlation cut-off. Right panel shows the fraction of the total number of genes whose absolute *pol* value is greater than 0.9 at various correlation cut-offs.

6.2.3 - High Frequency of Highly Polarized Genes is Independent of Experimental Noise and Dependent on the Normal-Tumour Connections

The analysis of the distribution of the polarization metric has revealed a high frequency of highly polarized genes. Although intriguing, this interesting property may not be associated by biological phenomena but by the result of random chance or by some interaction between the properties of the two expression matrices. In order to acquire confidence in the biological relevance of the high frequency of polarization, two

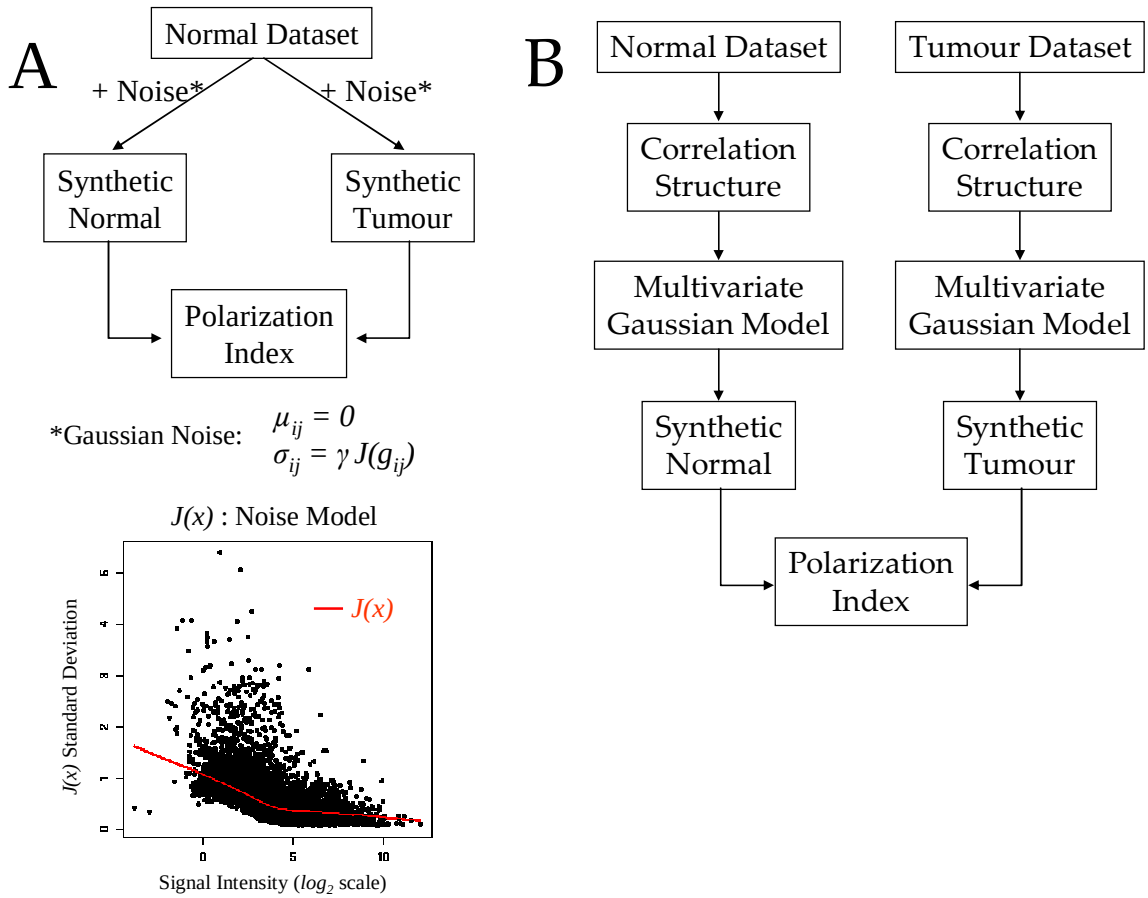


Figure 6.6 – Simulation Experiments. (A) Noise Simulation. The original dataset from normal cells is used to add noise depending on signal levels (bottom plot) multiplied by a scaling factor γ (see Methods section). The observed levels of polarization index computed from these simulated datasets would be due therefore to random experimental noise. (B) Correlation Structure Simulation. The correlation structure within normal or tumour (but not both) is resembled by a multivariate Gaussian model which generates random data with similar correlation structures. Thus, the calculated degree of polarization index from these synthetic datasets is due by unconnected correlation structures.

different simulations were used. These datasets are representative of two scenarios where expression data are simulated in the absence of any interaction between normal and tumour tissues. Thus, these datasets are meant to test whether the connection structure between genes expressed in the normal and tumour tissues or whether experimental and technical variability would be responsible of such a high frequency of polarized genes. A schema for the procedure of these simulations is shown in Figure 6.6.

Expression data acquired with microarray technology is subject to experimental and technical variability [20; 175; 176]. The properties of this noise have been investigated by a number of publications and error models have been proposed [32]. In the first case, the expression data associated to the normal tissue samples has been used to generate two separated datasets by adding noise using a well defined error model [32]. In this model (see Figure 6.6A and Methods section), noise has been added with intensity depending on a scaling factor γ (see methods sections for details). In the simulation, a value of this parameter was choosen (γ=3) to match the distribution of correlations between a given gene *i* in the simulated datasets with the observed in the real data

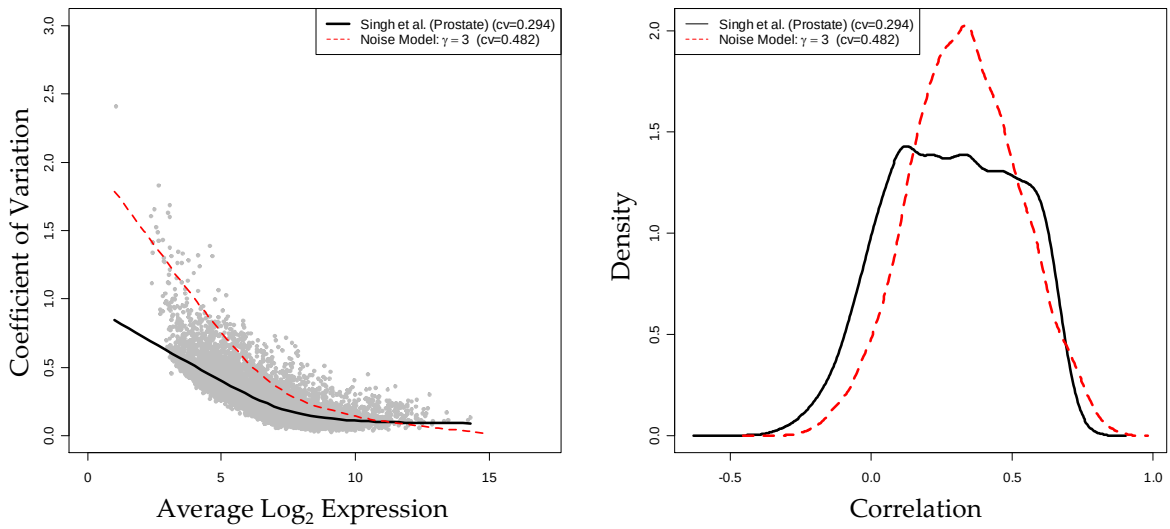


Figure 6.7 – γ parameter chosen for the noise model simulations. Left panel show the level of noise injected depending on the gene expression (dotted line). Right panel shows the correlation distribution of the same gene in both datasets (see Methods section).

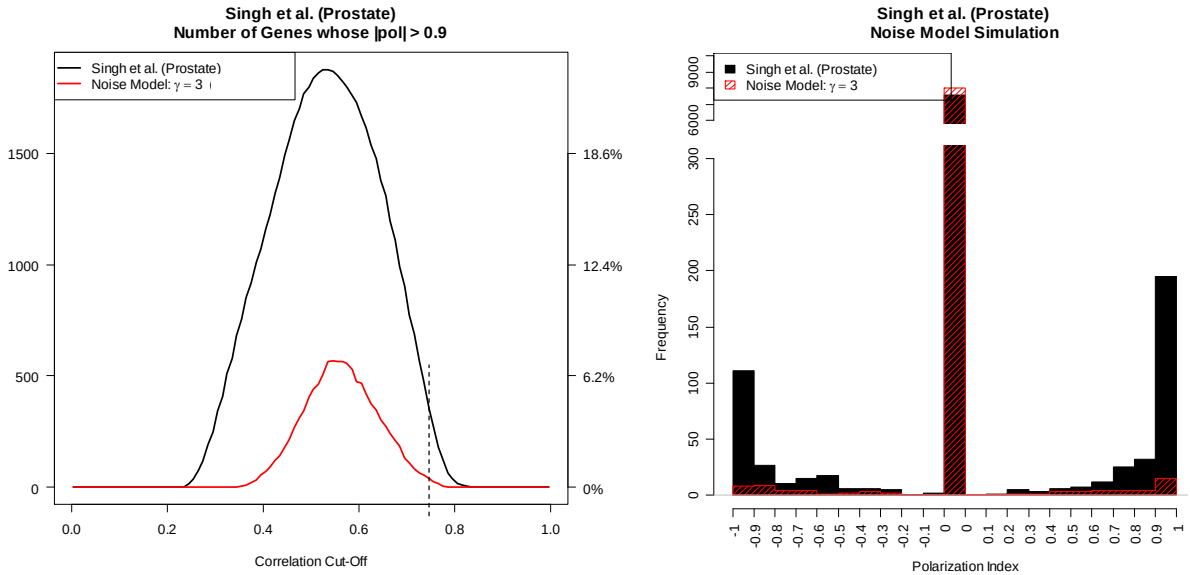


Figure 6.8 – Comparison of *pol* in real and noise-model simulated datasets. Left panel shows the number of highly polarized genes independent of the correlation cut-off. Right panel shows a comparison of *pol* distribution for the chosen correlation cut-off (0.75).

between normal and tumour tissue samples as a measure of the overall similitude between the two tissues (Figure 6.7 and Figure 6.21). The polarization index was subsequently computed for this synthetic dataset (Figure 6.8).

Figure 6.8 shows that the *pol* distribution observed in the noise-model-generated dataset is very low compared to the observed in the real dataset independently of the FDR correlation cut-off chosen. Furthermore, the number of synthetic genes whose absolute polarization index is greater than 0.9 (at 0.75 cut-off) was always less than the observed in the real dataset independently of noise scaling factor (left panel in Supplementary Figure 6.3). Moreover, the shape of these distributions is symmetrical and more populated around zero compared to the distribution in the real dataset (right panel in Supplementary Figure 6.3). Similar results were obtained for other datasets (Supplementary Figure 6.4). Altogether, these results suggest that the observed level of polarized genes is not due to experimental and technical random noise.

Reasoning about the correction structure, it is possible that the observed connection structure might be the result of the interaction of the underlying networks between normal and tumour. Hence, to evaluate whether unconnected but similar connection structures could generate comparable polarization indices, another simulation experiment (depicted in Figure 6.6B) was performed. Using normal and tumour separately, the observed correlation matrix from 2,000 randomly selected genes was employed to fit a multivariate Gaussian correlation matrix. The fitted parameters were then used to generate 2,000 synthetic genes whose correlation distribution is similar to the observed one to finally estimate and compare the polarization index (Figure 6.9, and Supplementary Figure 6.5-7). In this generated dataset, the number of synthetic genes whose absolute polarization index is higher that 0.9 was 0, which is opposite to 255 genes estimated under the same methodological conditions in the real dataset (middle panel in Figure 6.9). A similar trend was observed in other datasets (Supplementary Figure 6.6). Despite the number of polarized genes in the synthetic dataset was not always zero along correlation cut-off (right panel in Figure 6.9), the trend of observing

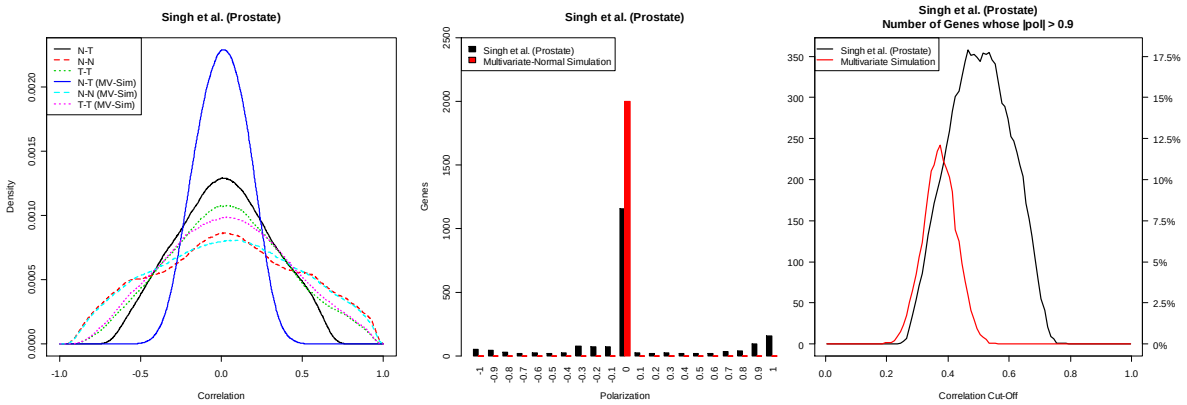


Figure 6.9 – Comparison of *pol* in real and multivariate Gaussian simulated datasets. Left panel shows that the distribution of correlations in the simulated dataset resembles that of the original dataset for N-N and T-T. However, the unconnected N-T correlation distribution is largely different from the observed in the Singh *et al.* dataset. Middle panel shows the distribution of *pol* values in both datasets under the same methodological conditions (using the same 2,000 randomly chosen genes, see also Methods section). For other datasets see Supplementary Figure 6.6. Right panel shows a comparison on the dependency of the number of highly polarized genes to the correlation cut-off under the same methodological conditions. For other datasets see Supplementary Figure 6.7.

fewer polarized synthetic genes still holds for sensible correlation cut-offs (Supplementary Figure 6.7). These multivariate simulations suggest therefore that the generation of high polarization indices, at the observed correlations, are not due to unconnected correlation structures.

6.2.4 - Relationship of Differential Expression and Polarization

The analysis was then focused on whether highly polarized genes tend to be among the differentially expressed genes (DEG). In order to test this hypothesis, the overlap between genes with high *pol* and those DEG generated between normal and tumour gene levels resulted by appling a t-Test, a Wilcoxon-Mann-Whitney rank sum test, and SAM analysis [43] (see Methods section) has been inspected. Results are shown in

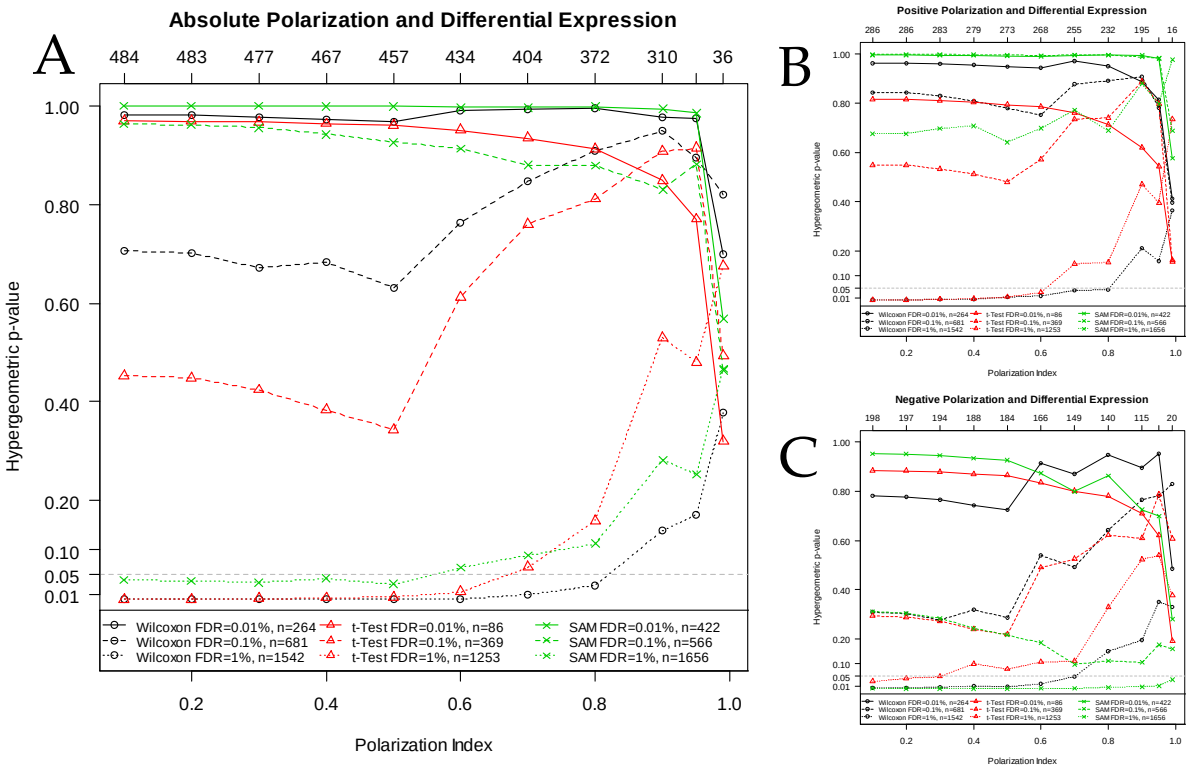


Figure 6.10 – Overlap between genes differentially expressed and polarized genes. Genes differentially expressed (n in legend) were tested for overlap with those genes whose *pol* was higher than a threshold (horizontal axis). Raw p-value is shown in the vertical axis. Numbers in the top axis show the number of genes whose *pol* was higher than the corresponding threshold (in bottom axis). (A) Ignoring *pol* sign. (B) Considering only genes having positive *pol*. (C) Considering only genes having negative *pol*. Data estimated from Singh *et al.*

Figure 6.10. No significant overlaps were found between highly polarized genes (*pol* > *0.9 or pol* < *-0.9*) and DEG even though the number of differential expressed genes were three times more than those highly polarized. On the contrary, a significant overlap exists for |*pol*| between *0.2~0.6*. In this condition the degree of polarization is very low and the number of DEG is about 1,500 out of 8,059 [18%]. This analysis supports the conclusion that highly polarized genes are not more frequently differentially expressed between normal and tumour samples.

6.2.5 - A Link between Dramatic Over-Expression and Polarization

Among the causes in cancer that could induce very extensive changes in gene expression are genetic mutations, such as translocations. A common consequence of

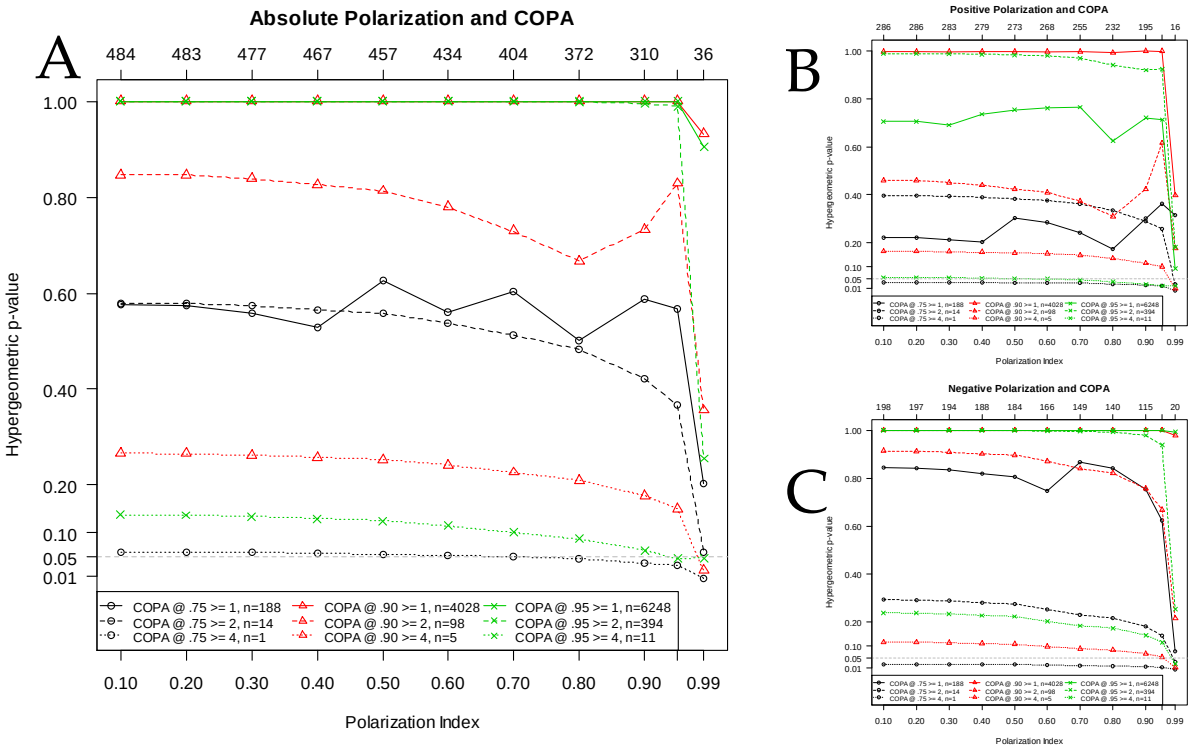


Figure 6.11 – Overlap between COPA genes and polarized genes. COPA genes (n in legend) were tested for overlap with those genes whose *pol* was higher than a threshold (horizontal axis). The p-value is shown in vertical axis. Numbers in top axis show the number of genes whose *pol* (in absolute value) was higher than the corresponding threshold (in bottom axis). Legend "COPA @ X >= Y" stands for COPA genes whose X quantile is larger Y (see Methods). (A) Ignoring *pol* sign. (B) Considering only genes having positive *pol*. (C) Considering only genes having negative *pol*. Data estimated from Singh *et al.*

such rearrangement is the dramatic change in the expression of genes associated to this translocation. Tomlins *et al.* developed a method named cancer outlier profile analysis (COPA) to identify genes likely to be a target of this genetic rearrangement [49]. COPA, which was successfully used to detect gene fusions in prostate cancer [49], is essentially a non-parametric standardization coupled with quantile filtering to detect genes that are over-expressed in a considerable subset of samples (see section 2.5.1.3 in Chapter 2 for details). Thus, highly polarized genes were compared to genes highlighted by COPA. The results shown in Figure 6.11 suggest that there is a significant overlap only for the topmost ranked COPA genes (n <= 11) with the topmost polarized genes (n < 36). No other significant associations were apparent. These results propose therefore that a small proportion of highly polarized genes are associated to an effect on gene expression that is compatible with gene translocations.

6.2.6 - Effect of Gene Silencing on Polarization

Since silencing may affect the gene expression, an obvious hypothesis is whether silencing could be a cause for large polarization indexes. If a gene is silenced during tumorigenesis or tumour progression, it is likely that the normal counterpart gene still correlates with a subset of genes in tumour cells that were previously correlated before the silencing event. This would generate high polarization indexes. In order to verify if silencing is a plausible mechanism that generates, systematically, high polarization indexes, it has been defined for this study that a gene is silenced when it has been 'expressed' in a significantly larger number of normal samples than in tumour samples (or vice versa). A gene was called 'expressed' if the normalized expression is larger than a threshold (details are described in section 6.5.7 -). Results are shown in Figure 6.12. Although there is an abrupt decrease in the random probability for very high polarized genes ($|pol| >= 0.99$), a statistically significant overlap (α=0.05) between the genes called

expressed and those highly polarized for a sensible range of expression thresholds was not observed.

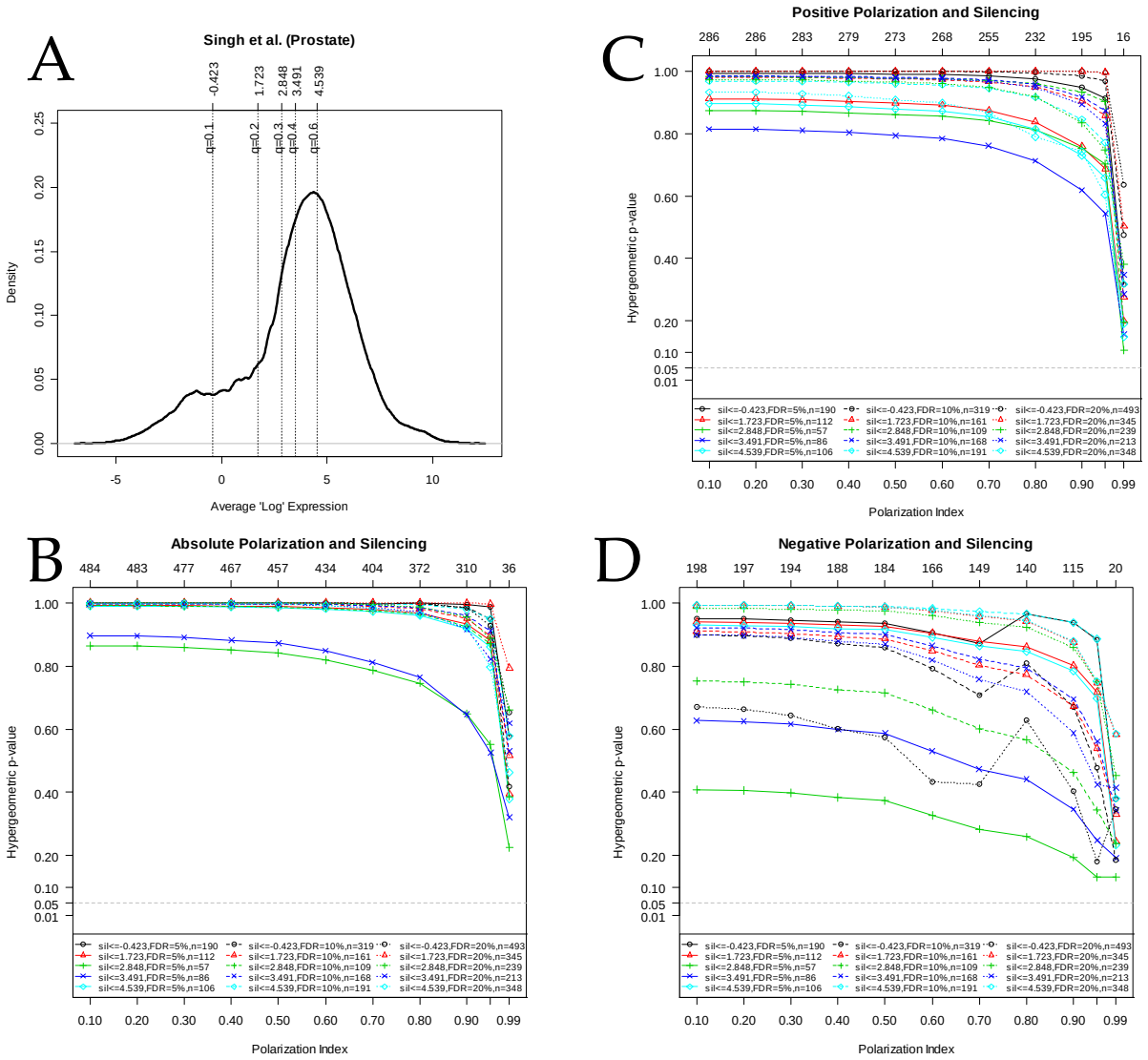


Figure 6.12 – Overlap between lack of expression and polarization. (A) Overall data distribution showing the threshold values tested. For (B), (C), and (D), not expressed genes (n in legend) were tested for overlap with those genes whose *pol* was higher than a threshold (horizontal axis). The p-value is shown in vertical axis. Numbers in top axis show the number of genes whose *pol* (in absolute value) was higher than the corresponding threshold (in bottom axis). Legend "sil <= X, FDR=Y%" stands for genes whose number of genes <= X (not expressed genes) was not due by chance at a FDR=Y (see Methods). (B) Ignoring *pol* sign. (C) Considering only genes having positive *pol*. (D) Considering only genes having negative *pol*. Data estimated from Singh *et al.*

6.2.7 - A Significant Proportion of Genes Methylated in Tumour Cells Are Heavily Polarized

In previous section, it has been demonstrated that genes that are expressed mainly in one tissue are not associated to any particular degree of polarization. However, the analysis was then directed to verify whether specific silencing mechanisms that have a very important role in the development of cancer may be associated to polarization. The most obvious candidate is methylation which is one of the best studied mechanisms for silencing [177] in which upstream gene promoters (or first exon) in nuclear DNA are methylated in the cytosine residue of CpG islands impeding the binding of activators and depleting gene expression [178; 179]. In tumour cells, several genes are methylated and not expressed in opposite to their corresponding normal cells in the same tissue where those genes are expressed and not methylated [178; 180-183]. If silencing is occurring in the tumour side, more genes with high Polarization Index in normal than in tumour would be expected because the tumour correlations would be depleted while normal correlations would be maintained. This agrees with the observation that there are more genes at the +1 end than at the -1 end in Singh *et al.* dataset (Figure 6.3) using the sign convention like in Figure 6.2. This same trend seems to be present in other datasets though in different degrees (Supplementary Figure 6.2). To further support the hypothesis that silencing by promoter methylation could be a driven force for high polarization indexes, two literature mining approaches were performed. First, highly polarized genes where compared with the number of Pubmed abstracts that relates each gene with "methylation" and "cancer" keywords. Significant p-values ($p <= 0.03$, hyper-geometric test) were obtained for the overlap of highly positive polarized genes and Pubmed abstracts, but not for negative polarized genes (Figure 6.13). This result suggests that methylation, cancer and highly (positive) polarized genes are, somehow, related.

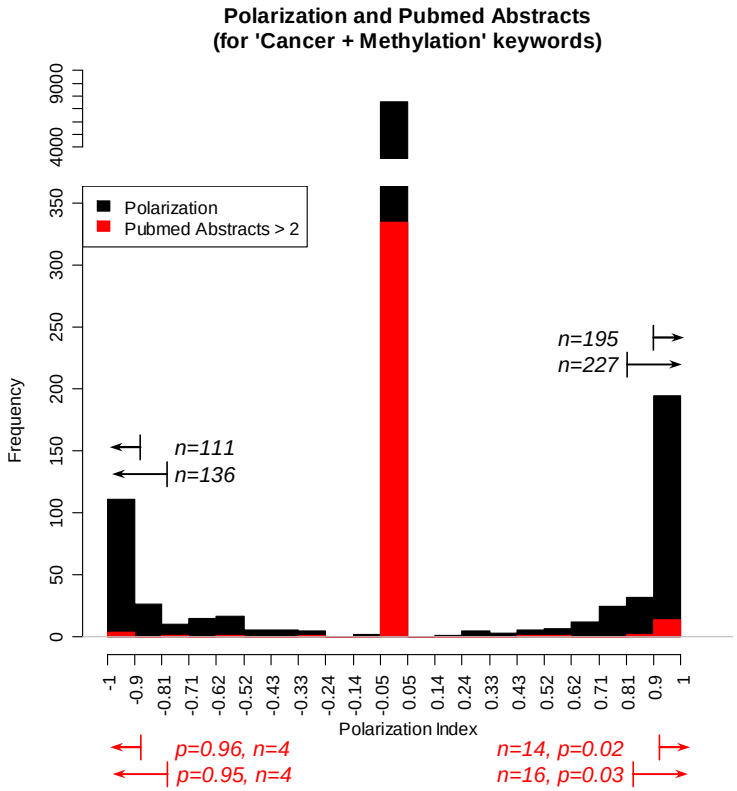


Figure 6.13 - Comparison of *pol* and the number of Pubmed abstracts. Black bars in vertical axis show the number of polarized genes at a certain polarization index (in horizonal axis). Red bars represent the number of genes whose Pubmed abstracts is larger than 2 in queries including the gene symbol, "cancer", and "methylation" as keywords. Indicated p-values corresponds to hyper-geometric p-values testing the null hyphotesis that the overlap of highly polarized genes (pol >= 0.8, pol >= 0.9, pol <= -0.8, and pol <= -0.9) and Pubmed abstracts per gene are due by chance. Data from *Singh et al.* dataset.

To further stress the association of methylation and polarization specifically for prostate cancer, genes affected directly or indirectly by methylation in prostate cancer cell lines reported in the literature derived from genome-wide studies [184-186] were compared to corresponding polarization indexes estimated in Singh *et al.* (Table 6.1). In these methylation studies nevertheless prostate cancer cell lines were used whose methylation patterns may vary slightly with those in clinical samples. In addition, one of the studies has treated cell lines with demethylation agents (5'-aza-2' deoxycytidine and trichostatin A) which may lead to the detection of genes affected by indirect de-methylation regulation. Thus, it is expected that a fraction of reported methylated genes

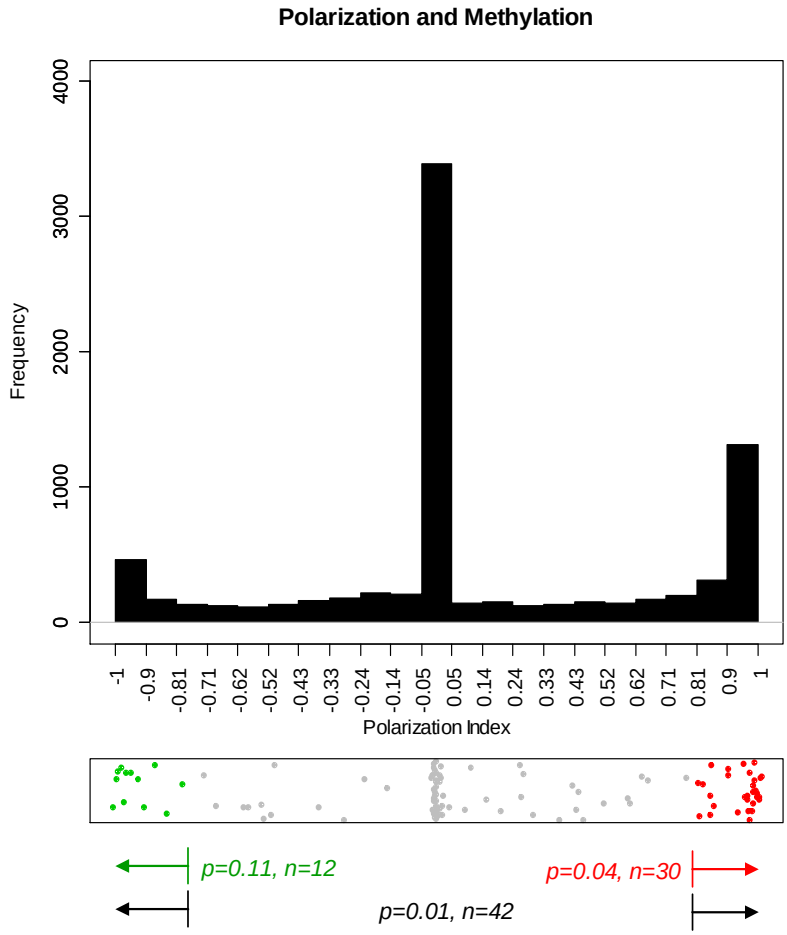


Figure 6.14 – Comparison of *pol* and methylated genes reported in the literature for prostate cancer. P-values were estimated by the hypergeometric probability of observing an overlap between the methylated genes (dots, n=12+30) and highly polarized genes (*pol* >= *0.8 or pol* <= *-0.8*). *pol* values were estimated at 0.57 correlation cut-off in Singh *et al.* dataset (see Methods section).

in these studies could not be methylated in clinical samples. From the 111 genes reported and included in the Singh *et al.* dataset, 30 genes had *pol* greater than 0.8 which is significantly higher than the expected by random chance (p=0.035, hyper-geometric test) whereas only 12 genes had *pol* less than -0.8, which is not significant (p=0.105). These results agree with the hypothesis that silencing and specifically promoter methylation in tumour cells could be, partially, a source of high polarization indexes.

Table 6.1 – *pol* in methylated genes reported in the literature for prostate cancer. Original reference is shown in headings. U column marks unique genes with a star, the total is 111. +/- column specifies whether the gene is considered highly positive polarized (30 in total) or highly negative polarized (12 in total). *pol* values were estimated at 0.57 correlation cut-off in Singh *et al.* dataset (see Methods section).

Lodygin et al.					*Wang et al.*					*Yu et al.*				
Symbol	*Affymetrix Id*	*pol*	*U*	*+/-*	*Symbol*	*Affymetrix Id*	*pol*	*U*	*+/-*	*Symbol*	*Affymetrix Id*	*pol*	*U*	*+/-*
APAF1	37227_at	-0.918*		-	ADK	38301_at	0.000*			ABCC4	34955_at	0.000*		
APOD	36681_at	0.000*			BBS4	33174_s_at	0.999*		+	ACTB	32318_s_at	-0.225*		
BIN1	459_s_at	0.112*			BCL2	2038_g_at	0.994*		+	ANXA2	769_s_at	0.000*		
BRCA2	1503_at	0.629*			BRCA1	33724_at	-0.560*			AOX1	37599_at	0.000*		
BTG1	37294_at	0.042*			CCNB2	32263_at	0.439*			CKS2	40690_at	0.000*		
BTG3	37218_at	0.974*		+	CD44	40493_at	0.967*		+	CNN3	40953_at	0.906*		+
CASP7	38281_at	0.000*			CDC25C	1584_at	-0.725*			GAPDH	35905_s_at	0.998*		+
CDKN1A	2031_s_at	0.000*			CDKN1A	2031_s_at	0.000			GSN	32612_at	0.391*		
CDKN1C	1787_at	0.989*		+	CDKN2C	36053_at	0.973*		+	GSTP1	829_s_at	0.976*		+
CDKN2D	41497_at	0.000*			DMXL1	33271_r_at	0.000*			GYPC	38119_at	0.831*		+
CITED2	33113_at	0.000*			ESR1	1681_at	-0.374*			HPS1	38467_at	0.000*		
CTGF	36638_at	0.000*			FAS	1440_s_at	-0.694*			IGFBP4	1737_s_at	0.528*		
CUTL2	37812_at	-0.866*		-	FBN1	32535_at	0.812*		+	IL6ST	35842_at	0.999*		+
CYLD	39582_at	0.000*			FEM1C	39976_at	0.000*			LRRFIP1	41320_s_at	-0.909*		-
DDB2	1243_at	0.000*			FOXD1	33203_s_at	0.815*		+	MARCKSL1	36174_at	0.999*		+
DKK1	35977_at	0.000*			GADD45A	1911_s_at	0.776			MGC5576	38655_at	0.000*		
DKK3	31454_f_at	-0.542*			GRK6	1392_at	-0.989*		-	MYO6	33375_at	0.000*		
DLC1	37951_at	0.422*			HOXD1	39476_at	0.988*		+	NR3C1	36690_at	0.000*		
DUSP1	1005_at	-0.603*			K6HF	33058_at	-0.969*		-	OXSR1	39136_at	0.000*		
GADD45A	1911_s_at	0.776*			KRTHB6	32329_at	-0.281*			PLAGL1	36943_r_at	0.988*		+
GAS2L1	31874_at	0.253*			LAMA4	37671_at	0.000*			PLEKHC1	36577_at	0.861*		+
GPX3	770_at	0.823*		+	MAPK7	35617_at	-0.962*		-	RAB31	33372_at	0.309*		
GSTM1	39054_at	0.000*			MTAP	38150_at	0.000*			SAT	34304_s_at	0.875*		+
HPGD	37323_r_at	0.433*			MYC	37724_at	-1.000*		-	SGCE	41449_at	0.000*		
ID3	37043_at	0.913*		+	OSMR	39277_at	0.995*		+	SOD2	34666_at	0.000*		
IRF1	669_s_at	-0.980*		-	PAX9	34933_at	0.196*			TGFB3	1767_s_at	0.970*		+
IRF7	36412_s_at	0.000*			PLAU	37310_at	-0.800*		-	TUBB4	429_f_at	-0.512*		
JUNB	32786_at	0.997*		+	RAB11A	36660_at	0.433*			TXNIP	31508_at	0.923*		+
MRE11A	32869_at	0.000*			RPL17	32440_at	0.000*			VAMP5	32534_f_at	0.000*		
NGFR	1673_at	0.285*			RPS4Y1	41214_at	0.972*		+					
PMS2	526_s_at	-0.496*			RTEL1	33727_r_at	0.991*		+					
PTGER4	1118_at	0.000*			SFN	33323_r_at	0.000							
PTGS2	1069_at	-0.565*			SLC26A4	36376_at	0.000*							
RBL2	32596_at	0.868*		+	SPARC	671_at	-0.152*							
RIS1	35692_at	0.000*			STC1	41354_at	0.000*							
SFN	33323_r_at	0.000*			STK4	36294_at	0.961*		+					
SFRP1	32521_at	0.603*			SYK	36885_at	-0.996*		-					
SGK	973_at	0.000*			TJP2	36655_at	-0.837*		-					
SMARCA1	40213_at	0.667*			TP53	1974_s_at	0.080*							
SQSTM1	40898_at	-0.974*		-	TSPY1	35929_s_at	0.997*		+					
THBS1	866_at	0.156*			WIT-1	1946_at	0.000*							
TNFRSF10B	34892_at	0.982*		+										
XPC	1873_at	0.260*												
ZFP36	40448_at	0.583*												

6.2.8 - Functional Analysis of Polarized Genes

In previous sections polarization index, *pol*, was defined as a heuristic measure of directional cell interaction. It was also demonstrated that the relatively high percentage of polarized genes cannot be explained as a property derived from experimental noise nor independent correlation structure of normal and tumour gene expression. It was

Table 6.2 – Functional analysis of polarized genes in BABELOMICS.

Polarization Index	*Database*	*Term*	*Inner*	*Outer*	*p-value*	*FDR*
Positive	InterPro	ATP-dependent helicase, DEAD-box	5/141	2/2310	0.0000132	0.0252
		DEAD/DEAH box helicase	6/141	8/2310	0.0000819	0.0784
		DEAD/DEAH box helicase, N-terminal	6/141	9/2310	0.0001297	0.0827
	SwissProt Keywords	Differentiation	10/123	33/2084	0.0001332	0.0608
		Hydroxylation	6/123	11/2084	0.0002473	0.0608
Negative	Gene Ontology Molecular Function	RNA binding	16/67	86/1833	0.0000020	0.0030
		Nucleic acid binding	38/74	446/2047	0.0000885	0.0667
	InterPro	RNA-binding region RNP-1/ RNA recognition motif	8/85	29/2322	0.0000505	0.0956
	SwissProt Keywords	RNA-binding	13/81	54/2099	0.0000021	0.0010

also shown that a specific gene silencing mechanism (methylation), unlike differential expression, seems to be to some extent associated to gene polarization. All these results strongly suggest that the polarization index is associated to biological properties of the system and is therefore likely to be a useful tool to identify interesting candidate genes involved in cell to cell interactions.

In order to further support this claim, a bioinformatics functional analysis of highly polarized genes was performed looking for an association between polarization and functional properties of the genes. This analysis was performed using the functional analysis tools implemented in the web-based toolset BABELOMICS [152]. The results revealed a strong association (FDR<1%) between highly polarized genes and the terms *nucleic acid binding* (in particular *RNA binding*) and *cell differentiation* (Table 6.2). No other functional terms were significant, even if the FDR threshold would be relaxed to FDR< 25%.

Although these results were encouraging, they were of limited validity. By simply comparing two or more lists of genes in functional terms, a large amount of biological information concerning the interaction between gene products is ignored. It is of interest to address the question whether the polarized genes form complex networks of interacting genes. For this reason, a more sophisticated analysis using the Ingenuity Pathways Analysis (IPA) was used [187]. Ingenuity knowledge base stores curated information on the interactions between genes, maps of canonical functional pathways,

and functional relationships supported by published literature and by protein-protein interaction data. Pathways of highly interconnected predictive genes are identified in this database by statistical likelihood and can be used to formulate hypotheses on the biological framework underlying the statistical models (refer to sections 5.5.4 in Chapter 5 for details about IPA). The list of positively and negatively polarized genes was overlaid onto a global molecular network developed from information contained in the IPA (see Methods section 6.5.8 -).

Table 6.3 – Top functional networks for positive polarized genes.

Ntw	*Top Functions*	*Score*	*Focus Genes*
1	Skeletal and Muscular System Development and Function, Tissue Morphology, **Cellular Movement**	63	35
2	**Cell Death**, Cancer, **Cellular Growth and Proliferation**	41	27
3	**Cellular Movement**, Hematological System Development and Function, Immune Response	22	18
4	Cancer, Cell Morphology, Connective Tissue Disorders	20	17
5	Cancer, Cell Morphology, Reproductive System Disease	19	16
6	Cancer, Cell Morphology, Connective Tissue Disorders	19	16
7	Lipid Metabolism, Molecular Transport, Small Molecule Biochemistry	15	14
8	Cell-To-Cell Signaling and Interaction, Hematological System Development and Function, Immune Response	15	14
9	Cell Morphology, Cellular Compromise, Connective Tissue Disorders	14	13
10	Cellular Development, Gene Expression, Organ Development	6	7

Table 6.4 – Top functional networks for negative polarized genes.

Ntw	*Top Functions*	*Score*	*Focus Genes*
1	Cellular Development, **Cellular Growth and Proliferation**, Hematological System Development and Function	28	18
2	**RNA Post-Transcriptional Modification, Cancer, Tumor Morphology**	24	16
3	Organ Development, Reproductive System Development and Function, Tissue Development	22	15
4	Cell Death, Cancer, Reproductive System Disease	20	14
5	Gene Expression, Digestive System Development and Function, Hepatic System Development and Function	18	13
6	Cellular Development, Skeletal and Muscular System Development and Function, Gene Expression	14	11

This analysis identified a number of interesting high scoring networks (Table 6.3 and Table 6.4). The two highest scoring networks for the positively polarized genes are respectively associated to *cell mobility/morphogenesis* and *cell growth*, and *survival* (Table 6.3). Interestingly, 15 genes were secreted factors or membrane proteins indicating a consistent large component of cell communication processes (Figure 6.15). The two highest scoring networks for the negatively polarized genes represent other functional components. The top scoring network is associated to *cellular growth* and *proliferation* whereas the second network is associated to *RNA post-transcriptional modifications*, *cancer* and *tumour morphology* (Table 6.4). It is interesting that the most significant functional category represented in the second network is RNA post-transcriptional

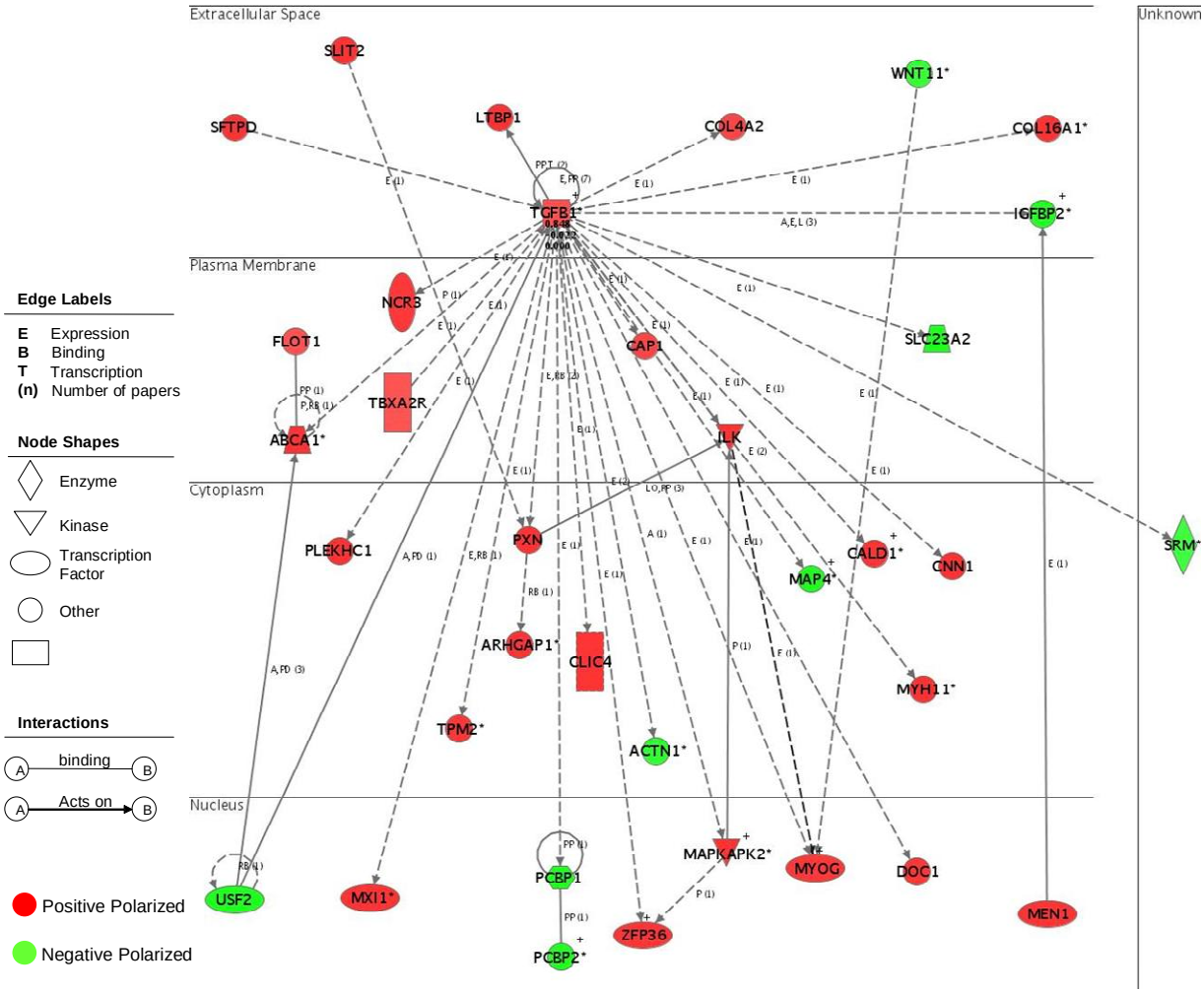


Figure 6.15 – TGF-beta network, a large component of secreted factors or membrane proteins are highly polarized.

modifications (Figure 6.16). This result is consistent with the initial analysis suggesting an association between negatively polarized genes and RNA binding. Further insights into the biology of polarized genes come from the analysis of the largest scoring network for the positively polarized genes. The analysis of the highest scoring network for the positively polarized genes shows that the majority of the genes are somehow linked to TGF beta (Figure 6.15). Some genes are not directly linked to TGF beta but are somehow related to one of the effects of TGF beta, such as Slit-2 that interacts with paxillin, an important component of the actin remodelling pathway.

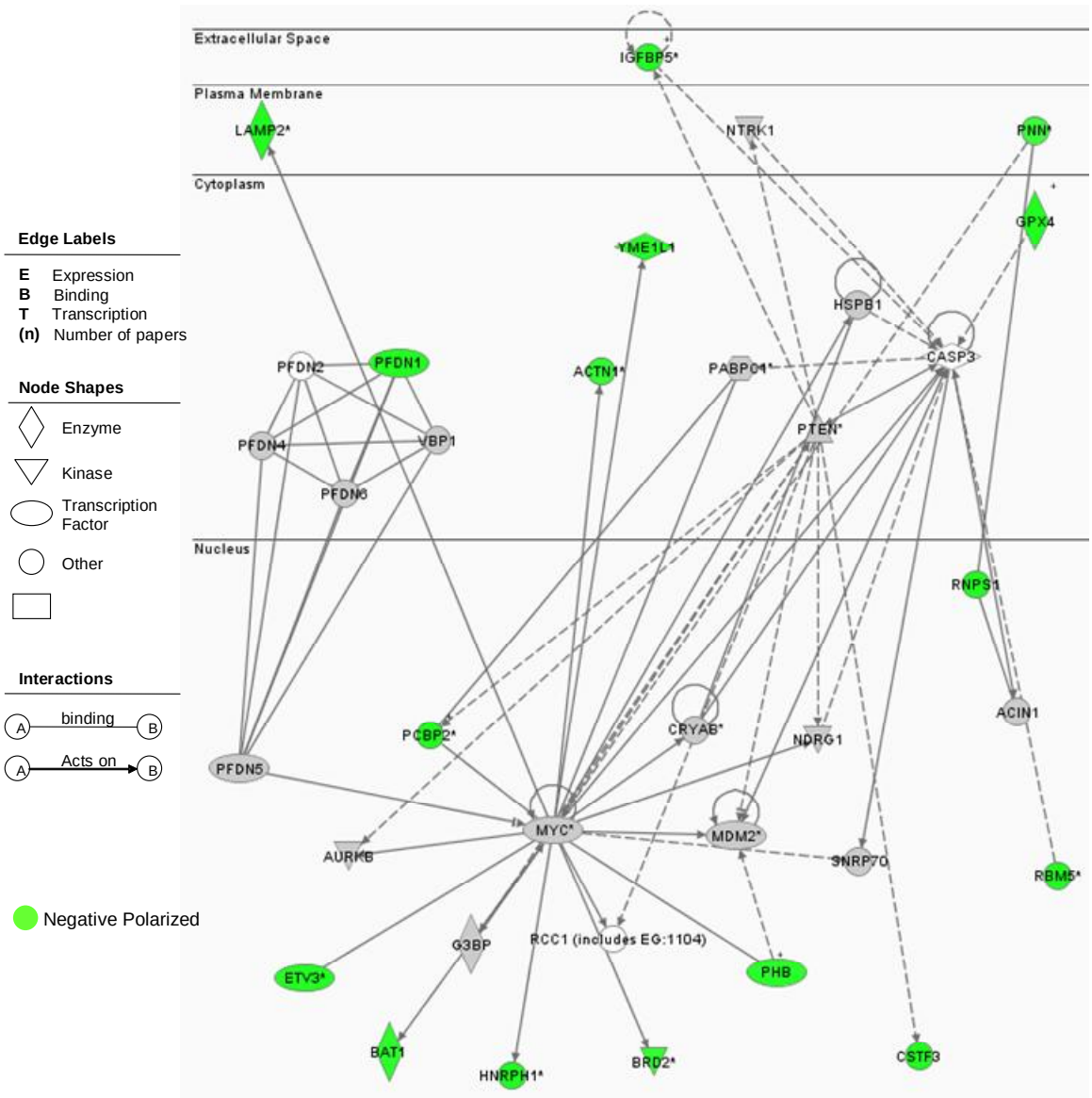


Figure 6.16 - RNA Post-Transcriptional Modification, Cancer, and Tumour Morphology Network and negatively polarized genes.

In addition to the general properties of the polarized genes that have been identified using network analysis, it was interesting that the most polarized genes (top 20) were characterized by being important regulators of key processes involved in cell communication. The most polarized gene SEC23A is indeed a key gene involved in controlling protein secretion [188], DBN1 play a role in cell migration , AHSG promotes endocytosis [189; 190], LPP has been implicated in cell adhesion, motility, and signalling events [191; 192], SLIT2 has been related to cell migration [193-195], MAML1 (*pol ~= -1*) has been involved in hematopoietic development by regulating Notch-mediated lymphoid cell fate decisions [196; 197], and PRKCI (*pol ~= -1*)has been related to epithelial tight junctions [198].

6.2.9 - Slit-2, One of the Most Polarized Genes, is Methylated and Control Survival in Prostate Cancer

The network analysis described above revealed a very interesting pattern. The Transforming Growth Factor Beta 1 (TGFB1) pathway, primarily involved in the survival of cancer cells, is connecting a significant number of genes selected by the approach described in this Chapter. Moreover, Slit-2 has the potential to interfere with this survival pathway. Slit-2 is also interesting because it is one of the top polarized genes regardless of the choice of parameters (Supplementary Table 6.1). Slit-2 is a secreted factor and is known to be methylated in tumours such as colon, glioma, lung, breast, Wilms, neuroblastoma, cervix, and renal cell cancer [199-203] but the methylation status in prostate cancer remains unknown so far. Furthermore, when genes highly correlated with Slit-2 ($|r| >= 0.75, n=151$) are used for searching functional networks in IPA, the terms observed are similar to those obtained using only polarized genes. In addition, Slit-2 is significantly differentially expressed between normal and tumour in Singh *et al.* dataset, and seems to be more expressed in normal than in

tumour (Figure 6.17). This raises the interesting hypothesis that Slit-2, and perhaps all other members of the slit gene family secreted by normal epithelial cells, may be involved in controlling the molecular and physiological state of prostate cancer cells. In order to test this hypothesis and to provide an indication of the mechanism involved in the polarized nature of this relationship, some experiments were performed.

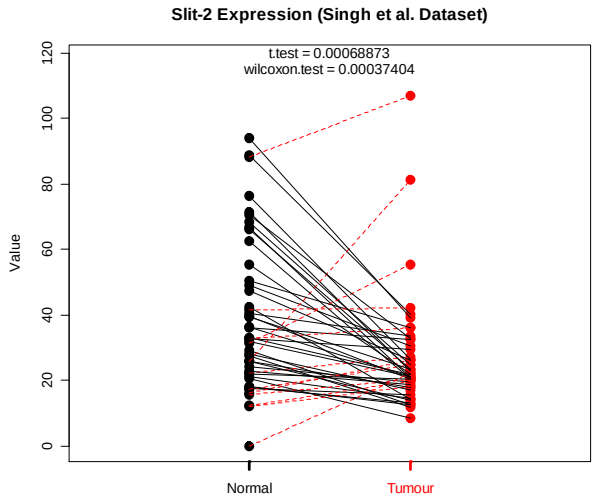


Figure 6.17 – Slit-2 is differentially expressed in normal and tumour. p-values are shown in legend. Vertical axis represents linear expression values. Lines connecting dots represent the expression of the same individual in both samples. Black lines represent larger expression in normal whereas red lines represent larger expression in tumour.

In particular, we wanted to first identify the biological effect that is associated to the relationship identified between the expression of Slit-2 and the transcriptional state of tumour cells. Second, we wanted to identify the molecular mechanism behind the directionality of the signal. Considering that genes with high *pol* that are secreted factors and whose methylation status is known in cancers others than prostate could be potential targets to support the observations derived from our methodology.

6.2.9.1 - Slit-2 is Methylated in Prostate Cancer Cell-lines

Having shown that slit proteins could be able to inhibit survival in prostate cancer cells, we wondered if the discovery that Slit-2 was a highly polarized gene could be related

with the methylation status of its promoter in tumour cells. In order to verify this, a methylation specific PCR assay was performed (Figure 6.22) from PC-3 and DU-145 cell lines derived from prostate cancer, and RWPE-1, a non-tumorigenic human prostate epithelial cell line (Figure 6.18). The methylation of the Slit-2 promoter was positive in PC-3 and DU-145 but not in RWPE-1. This result strongly suggests that Slit-2 is methylated in prostate cancer.

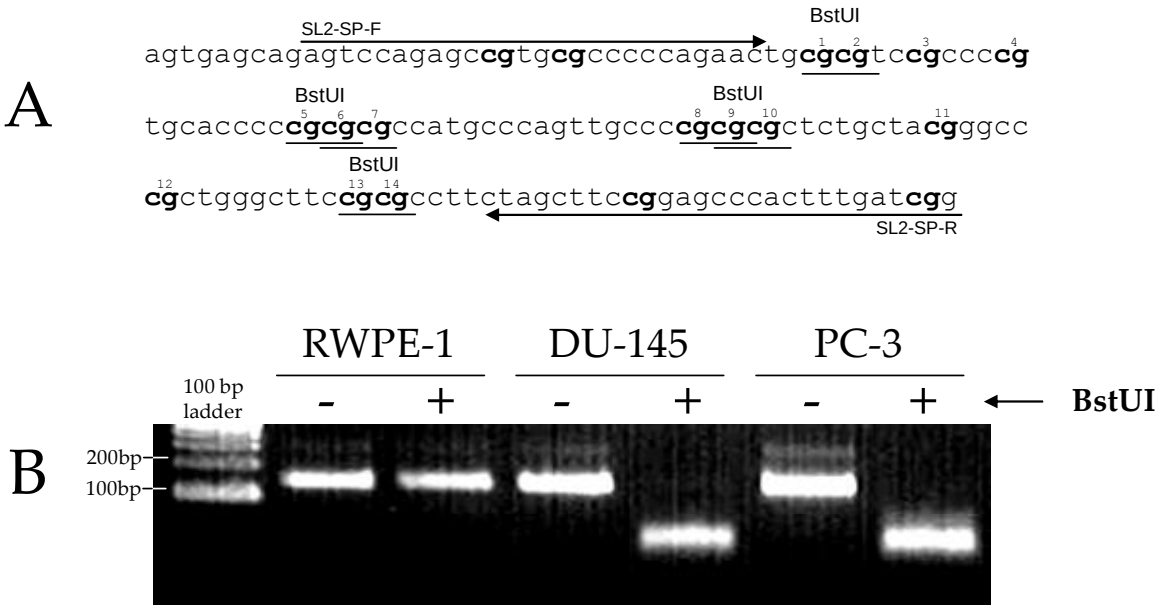


Figure 6.18 – Methylation assay for SLIT2 promoter. (A) The *SLIT2* putative promoter region was predicted by Promoter Inspector software (http://www.genomatix.de). This region is from -761 to -212 relative to the translation start site. "CpG" sites are highligthed in bold. The region was amplified from cell lines using the primers Sli2MOD4F (5′-GGGAGGTGGGATTGTTTAGATAT TT-3′) and primer Sli2MOD4R2 (5′-CAAAAACTCCTTAAACAACTTTAAATCCTAAAA-3′) as described previously (Dallol *et al.* Cancer Res. 2003 Mar 1;63(5):1054-8). 1/50 volume of the PCR reaction (with primers Sli2MOD4F and Sli2MOD4R2) was used in a nested PCR reaction with 30 cycles using primer SL2-SP-F (5′-AGTTTAGAGTYGTGYGTTTTTAGAAT-3′) and the primer SL2-SP-R (5′-CCRATCAAAATAAACTCCRTAAACTAA-3′) where Y is C+T and R is A+G. These primers amplify a region where most of the methylated CpGs in the putative *SLIT2* promoter are concentrated. The PCR conditions for both the first and second PCR were 95°C for 10 min, followed by 30-40 cycles of 1 min denaturation at 95°C, 1 min annealing at 52-54°C, and 2 min extension at 74°C. The PCR products were concentrated and purified using QIAquick PCR Purification columns (Qiagen). (B) The PCR products were then digested with 10 U of BstUI for 2 hours at 60ºC. The restriction enzyme digestion products were then visualized by separation in a 3% agarose gel. Experiments performed by Dr. Ashraf Dallol in Dr. Farida Latif's Laboratory. Data and image kindly provided by Dr. Latif.

6.2.9.2 - Slit-2 Inhibit Survival in Prostate Cancer Cell-lines

In order to test the possible role of Slit-2 protein in prostate cancer, Dr. Nicholas Davies and Dr. Moray Campbell tested the ability of conditioned supernatants from cells transfected with Slit2, Slit3, and Slit-like-2[1] in a clonogenic assay using PC-3 and DU-145 (Figure 6.19). All conditioned supernatants significantly inhibited the survival of both cells lines at 1:50 dilution.

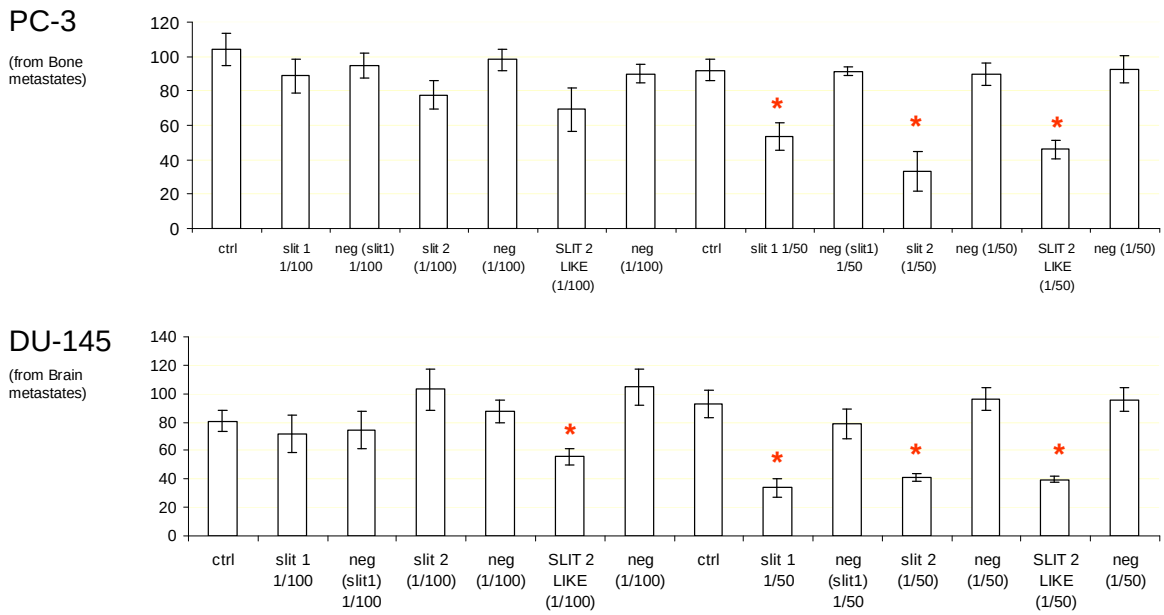


Figure 6.19 – Clonogenic assay in prostate cancer cell lines. PC-3 and DU-145 cell lines were treated with dilutions (1/100 or 1/50) of slit-1, slit-2, and slit-2-like (horizontal axis). A significant reduction in the number of colonies (vertical axis, marked with stars "*") was observed in both cell lines.

6.3 - Discussion

In this work, a polarization index was defined that is based on highly directional gene-gene connections between two cell types to emphasize genes hypothetically involved in

[1] Kindly provided by Dr. Heiner who determined that transfected products of the Slit-2-like construction fused with MYC is anchored in the membrane of HEK-293T cells by fluorescence microscopy assays.

cell to cell communication. The fact that only significant connections are used and that several genes show a stable index over a wide range of cut-off values suggests that *pol* measure is robust. High *pol* values were detectable at varied frequencies among datasets suggesting that the level of detectable interactions depends highly on the biological systems and the experimental setup. For instance, three prostate cancer datasets were used in which the levels of polarization detected were very high for Singh *et al.*, small for Lapointe *et al.*, and none for Yu *et al.*. Since Singh *et al.* and Yu *et al.* used the same microarray technology and Lapointe *et al.* used a different one, the observed polarization differences might be due mostly to differences in the biological samples used and to their relatively proximity (see Table 6.5) rather than experimental issues.

Polarization indexes were estimated based on significant connections based on a FDR approach. Thus, high values of *f* or *b* (e.g. > 100) would result in robust *pol* estimations if the chosen FDR is kept low (< 10%). However, a proper FDR associated to *pol*, rather than to each individual connection, is needed. Although rather laborious, perhaps the most important and fair FDR estimation for *pol* would be the one found empirically, testing a number of highly polarized genes experimentally. In this context, the FDR would be the fraction of those genes not involved in cell to cell communication.

Data that could be used as positive and negative controls for cell to cell communication is lacking. Therefore, the results of the simulations showing that high polarization indexes are not due to random noise neither to unconnected data are, at least, the first approaches to validate that the observed *pol* are not artefacts. The establishment of proper controls (mainly negative), however, is rather challenging because cells in contact or cells in shared media would inevitably affect each other.

Interestingly, differentially expressed genes do not tend to be highly polarized suggesting that some components of gene-gene connections between cells are regulated by subtle changes in expression. On the contrary, COPA results suggest that minor components of gene-gene connections between cells are due to drastic over-expression. In addition, another component seems to be related to silencing by methylation. Thus, although systematic changes in expression (silencing and DEG) were not significantly related to cell to cell communication processes under the scenario considered, specific mechanisms for under-expression (methylation) and over-expression (COPA) were detected to be, at certain degrees, related to cell to cell communication.

Functional and pathways analysis support the hypothesis that selected genes might be involved in the communication process. This is based on the fact that several functional terms and networks are related to cellular growth, proliferation, mobility, and signalling.

Based on polarization, the extra-cellular protein Slit2 was selected to be tested experimentally. The experiments performed in prostate cancer cell-lines show that Slit2 affects survival that its promoter is methylated. This agrees with the hypothesis that Slit2 is not beneficial for the tumour, which methylate Slit2 promoter in the course of tumour progression to neutralize its effects. However, Slit2 is being secreted by normal cells to control tumour expansion.

6.4 - Conclusions

A gene index named polarization has been defined to select genes important for a specific cell to cell communication scenario. Paracrine signalling are clear examples of this scenario. The index is based on highly directional gene-gene connections between

two cell types. Simulation experiments show that polarization is not a data artefact. On the contrary, *in-silico* analysis shows that highly polarized genes are related to meaningful biological phenomena such as methylation, over-expression, signalling, proliferation, cell-growth, and cell mobility. Experimental results of a selected gene show that polarization successfully predicted its role in cell to cell communication events. The evidence shown here supports the use of the defined polarization index to identify components of cell to cell communication from large scale functional genomics data.

6.5 - Materials and Methods

6.5.1 - Datasets

Six public datasets were used. Three prostate datasets originally published by Singh *et al.*[8], Lapointe *et al.*[9], and Yu *et al.*[204], a liver dataset published by Chen *et al.*[205], a colon dataset published by Notterman *et al*.[206], and a kidney dataset published by Boer *et al.*[207]. Only paired data corresponding to tumour and normal from the same tissue were used. Only one pair of samples per individual was used. In general, normal tissue was adjacent to the tumour. Datasets were normalized and processed before analysis to remove low variant and low expressed genes. Further details are included in Table 6.5.

Table 6.5 – Summary of datasets used.

Reference	*Microarrays*	*Tissue*	*Samples*	*Probes*	*Author's definition of "Normal"*	*Pre-processing*	*Probes Used*
Singh *et al.* [8]	Affymetrix U95Av2	Prostate	47	12,600	Adjacent Prostate Tissue	Log scale; Quantile Normalization (in simulations); Mean > 1	8,079
Lapointe *et al.* [9]	Custom cDNA 2-dyes	Prostate	39	11,490	Prostate Non-cancerous region	Median Imputed-NA; Quantile Normalization; SD >= 66% Quantile OR Mean >= 66% Quantile	6,269
Yu *et al.* [204]	Affymetrix U95Av2	Prostate	56	12,625	Adjacent Prostate Tissue	Quantile Normalization; SD >= 66% Quantile OR Mean >= 66% Quantile	6,741
Chen *et al.* [205]	Custom cDNA 2-dyes	Liver	43	22,645	Adjacent non-Tumour Tissue	Array removal: #NA > 3000; KNN Imputed-NA; Quantile Normalization; SD >= 66% Quantile	7,550
Notterman *et al.* [206]	Affymetrix Hum6500	Colon	18	7,464	Normal Tissue	-	7,464
Boer *et al.* [207]	Custom Nylon membranes	Kidney	34	36,864	Normal Tissue	Signal-to-Noise; Quantile Normalization; t-test FDR < 0.2 OR SD >= 80% Quantile	7,743

6.5.2 - Correlations

Non-linear Spearman ranking correlation was used to make the estimations robust to any monotone transformation. In order to estimate the number of significant correlations *f* and *b*, 100 bootstrap versions were used for each dataset to draw the null distribution of Spearman ranking correlation coefficients expected by chance. The bootstrap distribution was used to estimate a p-value which was subsequently corrected for multiple-test using an FDR correction [37]. Thus, FDR is meant as the expected number of false correlations. This ensures that estimations of *pol* are robust for genes displaying a large number of significant correlations. Scripts were written in R (http://www.r-project.org/).

6.5.3 - Noise Model Simulations

The normal datasets were used to generate synthetic normal and tumour datasets by adding random Gaussian noise with zero mean and standard deviation derived from an error model. Jain *et al.* proposed a model to estimate the error based on replicated microarray experiments [32]. This approach was adapted, for all non-replicated datasets used here, to estimate the experimental error by considering only the 50% centred data for each gene. This was based on empirical observations that the smoothed loess (non-

linear) fitted curve displays a similar behaviour to that exhibited by Jain *et al.* Figure 6.20 shows that the adapted error estimation for Singh *et al.* dataset is similar to that of Jain *et al.*. In order to control the level of noise that would conserve the overall similitude to the observed Normal and Tumour correlations, more ad hoc deviances were introduced by a scaling factor γ controlling the observed same-gene correlation distribution as a measure of similarity between Normal and Tumour gene expression. The selection of γ obeys to the distribution of same-gene correlations in the synthetic datasets that are closer to the observed ones (Figure 6.21). Therefore, γ=3 was used in this study. Synthetic datasets were then employed to compute the correlation matrix followed by the estimation of the polarization index.

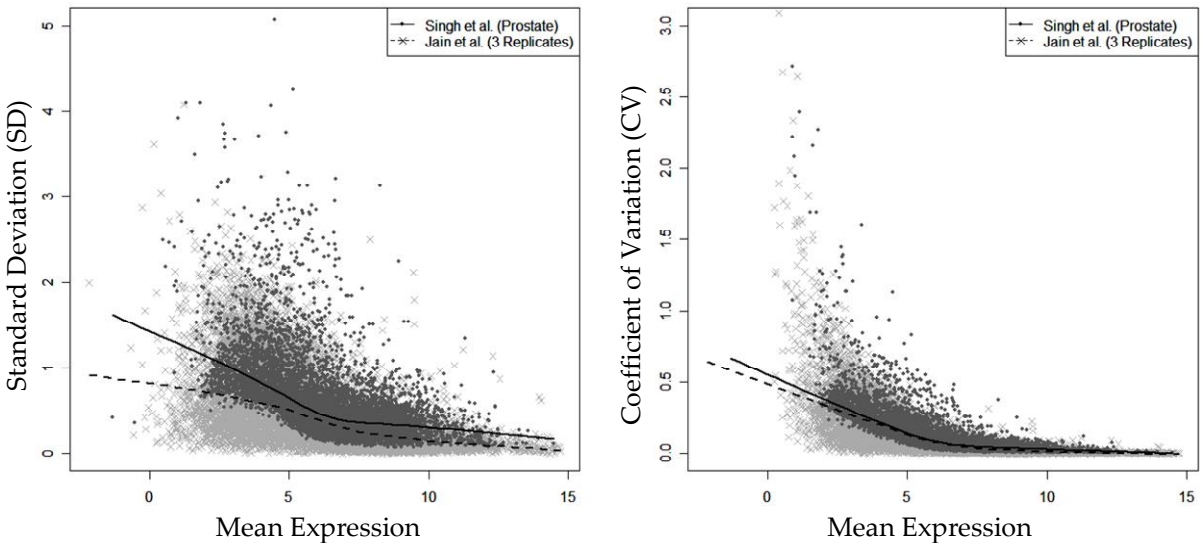


Figure 6.20 – Error model comparisons for Singh *et al.* and Jain *et al.* datasets. The figure shows similar behaviours of error models in SD (left) and CV (right).

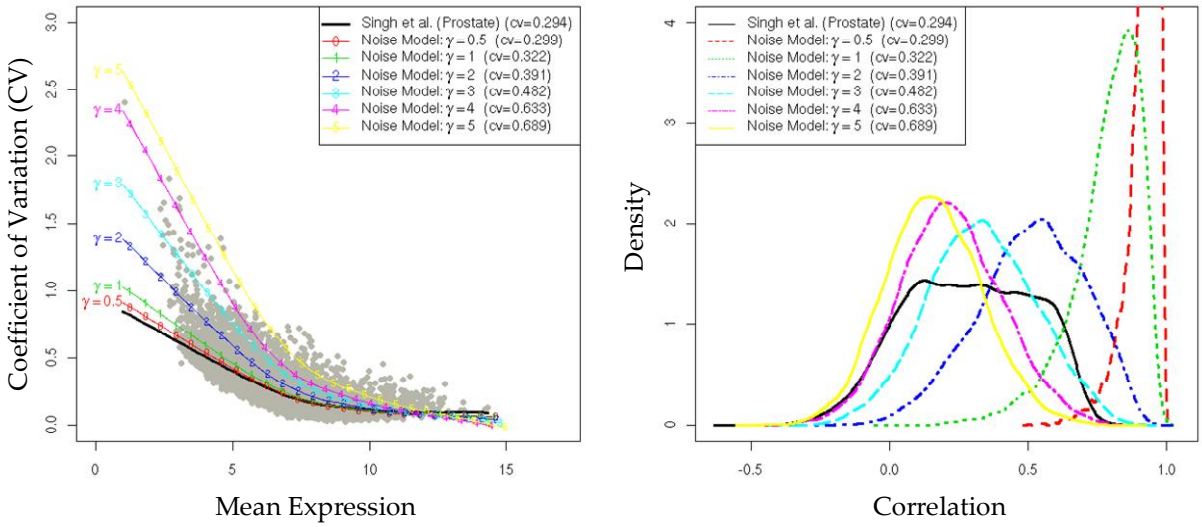


Figure 6.21 – Adjusting same-gene distribution in error model. γ factor controls the level of noise (left panel). Same-gene distributions depend therefore on γ factor (right panel). γ=0.5 is the right-most curve whereas γ=5 is the left-most curve.

6.5.4 - Multivariate Gaussian Simulation

Two thousand randomly selected genes from each dataset were used to compute normal and tumour correlation matrices. The choice of 2,000 genes follows to computational resources restrictions. The observed distribution of polarized genes is not affected by this choice due to the random gene selection. Similar results are observed even selecting randomly 500 genes (data not shown). Each matrix was fitted by a multivariate Gaussian [208] model to generate a synthetic dataset. Synthetic datasets were then used to compute the correlation matrix (supplementary figure 5). Subsequently, the polarization index was estimated from this correlation matrix. For comparisons with real data, 2,000 randomly selected genes from real data were used. The multivariate fitting and subsequent random dataset generation was performed using the function *rmvnorm* within *mvtnorm* package in R (http://www.r-project.org/).

6.5.5 - Differential Expression

Genes differentially expressed were estimated by a t-Test, a Wilcoxon-Mann-Whitney rank sum test, and SAM analysis in R. The p-values were corrected for multiple testing using a false discovery rate approach [37] or empirical FDR in the case of SAM[43]. Log-transformed and quantile-normalized data were used. The overlap between polarized genes and DEG were assessed by estimating the Hypergeometric probability. Genes whose *pol* were larger than a threshold were then tested.

6.5.6 - COPA analysis

Log-transformed and quantile-normalized data were used. 0.75, 0.90, and 0.95 quantiles of COPA-transformed data were used as suggested by Tomlins *et al.* [49] to select COPA genes. A gene was called COPA if the COPA-transformed value was larger than selected thresholds (75%, 90%, and 95%, see section 2.5.1.3 in Chapter 2 for COPA mathematical formula). The overlap between polarized genes and COPA genes were assessed by estimating the Hypergeometric probability. Genes whose *pol* were larger than a threshold were then tested.

6.5.7 - Gene Silencing

For a gene, silencing was defined as the significant difference in the number of normal expressed samples minus the number of tumour expressed samples (or vice versa). A gene was called 'expressed' if the expression value, log-transformed and quantile normalized, was larger than a threshold. Explored thresholds ranged from 10% to 60% of the overall observed distribution. To determine statistical significance in calling a gene silenced, 100 bootstrap versions for each threshold were used to draw the null distribution of differences. The p-value for an observed difference was obtained by counting the number of bootstrapped differences equal or larger than the observed

were then divided by the total number of bootstrapped differences. Raw p-values were corrected using a FDR approach [37]. Results for 5%, 10%, and 20% FDR are reported.

6.5.8 - Functional Gene Annotation

For the annotation of genes, OMIM, GeneCards, HubMed, AmiGO, PubGene, and iHOP were used by querying the current gene symbol given by HUGO ([209-215]). FatiGO+ within BABELOMICS [152] was used for the functional association of positive or negative polarized genes. The search included Gene Ontology, InterPro motifs, SwissProt keywords, KEGG Pathways, and Transcription Factors. Fisher exact test was used to assess the significance of annotation ratios between polarized genes (*195: pol > 0.9, 111: pol < -0.9*) and non-polarized genes (*3241: pol = 0* in correlation cut-off 0.57 and 0.75). p-values were corrected by an FDR approach [37]. Only significant terms after this multiplicity test correction are reported (FDR <= 1%). IPA [187] was used for the systematic analysis of sets of highly polarized genes in association to functional networks. Three gene lists were generated, genes whose *pol* was positive (232 genes >= 0.8 at 0.75 correlation cut-off), negative (140 genes <= -0.8), and consistently zero between 0.57 and 0.75 correlation cut-off (3241 genes). Only those IPA networks and canonical pathways that contained, at least, 7 polarized genes were used. The functional analysis of a network identified the biological functions that were most significantly associated to the genes in the network. Fisher's exact test was used to calculate a p-value determining the probability that each biological function and/or disease assigned to that network is due to chance alone. See details in section 5.5.4 in Chapter 5. For the analysis of genes highly correlated with Slit-2, 151 genes whose absolute Spearman correlation was higher than 0.74 were selected.

6.5.9 - Methylation Experiments

The detection of the methylation status was determined by comparing the *BstUI* digestion pattern of the PCR products from prostate cancer cell-lines DNA with and without sodium bisulphate treatment which leads to the conversion of cytosine to uracil. Methyl-cytosines are not converted by this treatment thus methylated CGCG patterns are conserved. This pattern is then recognized and digested by BstUI enzyme (Figure 6.22). To confirm that methylation is somehow related to genes identified by the method described in this chapter, Slit-2 promoter was investigated in three prostate cancer cell-lines using the procedure described. The experiment was performed by Dr. Ashraf Dallol in Dr. Farida Latif Laboratory at the University of Birmingham. Dr. Latif kindly provided data and images in Figure 6.18. Details of the PCR, primers, and experimental conditions are given in [199-201] and Figure 6.18.

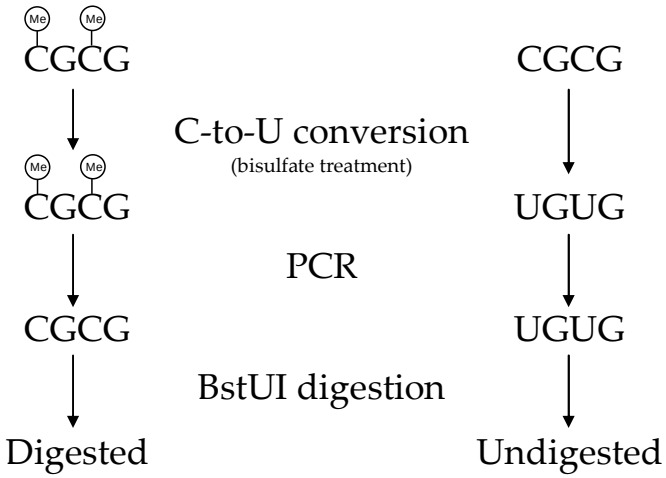


Figure 6.22 – Detection of methylated CpG promoter sites. Methylated cytosines (left) are finally digested because they are not sensitive to bisulfate treatment. Unmethylated cytosines are converted to uracil avoiding BstUI digestion. Figure inspired and adapted from Akagi *et al*. [216].

For comparison of methylated genes reported in the literature, Singh *et al.* dataset was additionally annotated using HCNetDat (http://www.hartwellcenter.org /hcnetdat). A correlation cut-off of 0.57 was used to estimate the *pol* in order to obtain a representative larger overlap between the genes present in our dataset and those genes

reported in the literature affected directly or indirectly by methylation in prostate cancer cell-lines. Gene-names reported to be methylated in papers were sought in Singh *et al.* dataset (111 genes in total). For gene symbols appearing more than once, the highest polarization value was considered. For the text-mining association of gene-cancer-methylation, systematic queries to PubMed using the gene name and "Cancer Methylation" as keywords were used.

6.6 - Supplementary Material

Supplementary Table 6.1 – Positive polarized genes. Sorted by *pol*. For a gene expressed in normal, *f* is the number of significant correlations with tumour genes whereas *b* is the number of significant correlations with normal genes for the same gene expressed in tumour.

Affy Id	*Symbol*	*Gene Name*	*pol(+)*	*f*	*b*
35749_at	TADA3L	transcriptional adaptor 3 (NGG1 homolog, yeast)-like	0.997	374	0
39099_at	SEC23A	Sec23 homolog A (S. cerevisiae)	0.997	357	0
37981_at	DBN1	drebrin 1	0.996	232	0
36621_at	AHSG	alpha-2-HS-glycoprotein	0.995	206	0
41195_at	LPP	LIM domain containing preferred translocation partner in lipoma	0.995	195	0
37042_at	HYAL2	hyaluronoglucosaminidase 2	0.994	166	0
38506_at	TCF2	transcription factor 2, hepatic, LF-B3, variant hepatic nuclear factor	0.994	163	0
34203_at	CNN1	calponin 1, basic, smooth muscle	0.994	163	0
39634_at	**SLIT2**	**slit homolog 2 (Drosophila)**	**0.993**	**151**	**0**
31440_at	TCF7	transcription factor 7 (T-cell specific, HMG-box)	0.993	145	0
32464_at	DEFB4	defensin, beta 4	0.993	133	0
40945_at	KLF11	Kruppel-like factor 11	0.992	127	0
37606_at	KHK	ketohexokinase (fructokinase)	0.991	116	0
34539_at	OR7E37P	olfactory receptor, family 7, subfamily E, member 37 pseudogene	0.991	115	0
32531_at	GJA1	gap junction protein, alpha 1, 43kDa (connexin 43)	0.991	108	0
888_s_at	GDF1	growth differentiation factor 1	0.990	103	0
40474_r_at	KPNA1	karyopherin alpha 1 (importin alpha 5)	0.990	98	0
35365_at	ILK	integrin-linked kinase	0.989	93	0
38725_s_at	DPM2	dolichyl-phosphate mannosyltransferase polypeptide 2, regulatory subunit	0.989	90	0
38643_at	LOC92689		0.989	90	0
40359_at	RASSF7	Ras association (RalGDS/AF-6) domain family 7	0.989	86	0
35263_at	EIF4EBP2	eukaryotic translation initiation factor 4E binding protein 2	0.988	85	0
37279_at	GEM	GTP binding protein overexpressed in skeletal muscle	0.988	84	0
37065_f_at	GAGE5	G antigen 5	0.988	83	0
36762_at	GABRG2	gamma-aminobutyric acid (GABA) A receptor, gamma 2	0.988	82	0
36577_at	PLEKHC1	pleckstrin homology domain containing, family C (with FERM domain) member 1	0.988	80	0
39145_at	MYL9	myosin, light polypeptide 9, regulatory	0.986	73	0
40592_at	IDS	iduronate 2-sulfatase (Hunter syndrome)	0.986	73	0
38230_at	EPAS1	endothelial PAS domain protein 1	0.986	73	0
32026_s_at	RAPGEF2	Rap guanine nucleotide exchange factor (GEF) 2	0.986	72	0
36650_at	CCND2	cyclin D2	0.986	69	0
37697_s_at	VDAC2	voltage-dependent anion channel 2	0.986	69	0
41856_at	UNC5B	unc-5 homolog B (C. elegans)	0.985	66	0
37732_at	RYBP	RING1 and YY1 binding protein	0.985	65	0
32000_g_at	ABCA1	ATP-binding cassette, sub-family A (ABC1), member 1	0.985	64	0

36159_s_at	PRNP	prion protein (p27-30) (Creutzfeld-Jakob disease, Gerstmann-Strausler-Scheinker syndrome, fatal familial insomnia)	0.984	63	0
2057_g_at	FGFR1	fibroblast growth factor receptor 1 (fms-related tyrosine kinase 2, Pfeiffer syndrome)	0.984	62	0
36613_at	IFRD2	interferon-related developmental regulator 2	0.984	61	0
32582_at	MYH11	myosin, heavy polypeptide 11, smooth muscle	0.983	59	0
40352_at	YPEL1	yippee-like 1 (Drosophila)	0.983	59	0
1495_at	LTBP1	latent transforming growth factor beta binding protein 1	0.983	58	0
33091_at	HOXD13	homeo box D13	0.983	58	0
37318_at	ETF1	eukaryotic translation termination factor 1	0.983	58	0
38863_at	RFC5	replication factor C (activator 1) 5, 36.5kDa	0.983	57	0
37628_at	MAOB	monoamine oxidase B	0.982	55	0
39511_at	MLLT4	myeloid/lymphoid or mixed-lineage leukemia (trithorax homolog, Drosophila), translocated to, 4	0.982	55	0
38559_at	MGC22014		0.982	55	0
1469_at	MAPKAPK2	mitogen-activated protein kinase-activated protein kinase 2	0.982	55	0
32367_at	SIRPB1	signal-regulatory protein beta 1	0.982	54	0
38644_at	PXN	paxillin	0.982	54	0
34355_at	MECP2	methyl CpG binding protein 2 (Rett syndrome)	0.982	54	0
41402_at	DKFZP564O0823		0.981	53	0
33092_at	FPRL2	formyl peptide receptor-like 2	0.981	52	0
33017_at	GPLD1	glycosylphosphatidylinositol specific phospholipase D1	0.981	52	0
527_at	CENPA	centromere protein A, 17kDa	0.981	52	0
32420_at	GPR6	G protein-coupled receptor 6	0.981	52	0
41191_at	KIAA0992		0.981	51	0
1922_g_at			0.980	49	0
31321_at			0.980	49	0
37875_at	GPA33	glycoprotein A33 (transmembrane)	0.980	49	0
31793_at	DEFA1	defensin, alpha 1	0.980	49	0
1859_s_at	MDM2	Mdm2, transformed 3T3 cell double minute 2, p53 binding protein (mouse)	0.980	48	0
33876_at	WWTR1	WW domain containing transcription regulator 1	0.979	47	0
34677_f_at	LOC220594		0.979	47	0
39225_at	AGPS	alkylglycerone phosphate synthase	0.979	140	1
33891_at	CLIC4	chloride intracellular channel 4	0.979	46	0
32457_f_at	PRB4	proline-rich protein BstNI subfamily 4	0.979	46	0
41137_at	PPP1R12B	protein phosphatase 1, regulatory (inhibitor) subunit 12B	0.979	46	0
41738_at	CALD1	caldesmon 1	0.978	45	0
654_at	MXI1	MAX interactor 1	0.978	45	0
41161_at	DAXX	death-associated protein 6	0.978	45	0
760_at	DYRK2	dual-specificity tyrosine-(Y)-phosphorylation regulated kinase 2	0.978	45	0
40468_at	FNBP1	formin binding protein 1	0.978	44	0
755_at	ITPR1	inositol 1,4,5-triphosphate receptor, type 1	0.977	43	0
41207_at	C9orf3	chromosome 9 open reading frame 3	0.977	43	0
448_s_at	MEN1	multiple endocrine neoplasia I	0.977	43	0
37360_at	LY6E	lymphocyte antigen 6 complex, locus E	0.977	43	0
35168_f_at	COL16A1	collagen, type XVI, alpha 1	0.977	42	0
31618_at			0.976	41	0
36149_at	DPYSL3	dihydropyrimidinase-like 3	0.976	41	0
40841_at	TACC1	transforming, acidic coiled-coil containing protein 1	0.976	40	0
40146_at	RAP1B	RAP1B, member of RAS oncogene family	0.975	39	0
41382_at	DMBT1	deleted in malignant brain tumors 1	0.974	38	0
37972_at	DNASE1L3	deoxyribonuclease I-like 3	0.974	38	0
40927_at	SLC6A8	solute carrier family 6 (neurotransmitter transporter, creatine), member 8	0.974	38	0
34097_at	NME7	non-metastatic cells 7, protein expressed in (nucleoside-diphosphate kinase)	0.974	37	0
41566_at	TCF15	transcription factor 15 (basic helix-loop-helix)	0.974	37	0
39700_at	ARHGAP1	Rho GTPase activating protein 1	0.973	36	0
31775_at	SFTPD	surfactant, pulmonary-associated protein D	0.973	109	1
32542_at	FHL1	four and a half LIM domains 1	0.971	34	0
35330_at	FLNC	filamin C, gamma (actin binding protein 280)	0.971	34	0
648_at	AVPR1B	arginine vasopressin receptor 1B	0.971	34	0
34849_at	SARS	seryl-tRNA synthetase	0.971	34	0
38607_at	TM4SF5	transmembrane 4 L six family member 5	0.971	34	0
38991_at	LOC441773		0.971	103	1
938_at			0.971	33	0
716_at	GGTLA1	gamma-glutamyltransferase-like activity 1	0.971	33	0
34063_at	RECQL5	RecQ protein-like 5	0.971	33	0
35595_at	RCP9		0.971	33	0
41488_at	LOC57149		0.970	32	0
32045_at	DATF1	death associated transcription factor 1	0.970	32	0
36956_at	SLC20A2	solute carrier family 20 (phosphate transporter), member 2	0.970	32	0
36785_at	HSPB1	heat shock 27kDa protein 1	0.969	31	0

34164_at			0.969	31	0
38526_at	PDE4D	phosphodiesterase 4D, cAMP-specific (phosphodiesterase E3 dunce homolog, Drosophila)	0.969	31	0
38282_at	ADAM15	ADAM metallopeptidase domain 15 (metargidin)	0.969	31	0
35112_at	RGS9	regulator of G-protein signalling 9	0.969	31	0
33007_at	LOC63928		0.968	30	0
1305_s_at	CYP4F3	cytochrome P450, family 4, subfamily F, polypeptide 3	0.967	29	0
40817_at	NUCB1	nucleobindin 1	0.967	29	0
35756_at	GIPC1	GIPC PDZ domain containing family, member 1	0.967	29	0
35177_at	DDHD2	DDHD domain containing 2	0.967	29	0
2066_at	BAX	BCL2-associated X protein	0.967	29	0
40448_at	ZFP36	zinc finger protein 36, C3H type, homolog (mouse)	0.967	29	0
36805_s_at	NTRK1	neurotrophic tyrosine kinase, receptor, type 1	0.967	29	0
348_at	KIFC1	kinesin family member C1	0.967	29	0
40074_at	MTHFD2	methylenetetrahydrofolate dehydrogenase (NADP+ dependent) 2, methenyltetrahydrofolate cyclohydrolase	0.966	28	0
1915_s_at	FOS	v-fos FBJ murine osteosarcoma viral oncogene homolog	0.966	28	0
39046_at	H2AFV	H2A histone family, member V	0.966	28	0
33756_at	AOC3	amine oxidase, copper containing 3 (vascular adhesion protein 1)	0.964	27	0
37049_g_at	TOMM34	translocase of outer mitochondrial membrane 34	0.964	27	0
38026_at	FBLN1	fibulin 1	0.964	27	0
484_at	NCOA1	nuclear receptor coactivator 1	0.964	27	0
34647_at	DDX5	DEAD (Asp-Glu-Ala-Asp) box polypeptide 5	0.964	27	0
40527_at	KCNQ1	potassium voltage-gated channel, KQT-like subfamily, member 1	0.964	27	0
35255_at	IPO7	importin 7	0.964	27	0
41331_at	LRIG2	leucine-rich repeats and immunoglobulin-like domains 2	0.964	27	0
37418_at	POU2F2	POU domain, class 2, transcription factor 2	0.964	27	0
32313_at	TPM2	tropomyosin 2 (beta)	0.963	26	0
31897_at	DOC1		0.963	26	0
1071_at	GATA2	GATA binding protein 2	0.963	26	0
33442_at	KIAA0367	KIAA0367	0.963	26	0
41598_at	SEC22L1	SEC22 vesicle trafficking protein-like 1 (S. cerevisiae)	0.963	26	0
35324_at	SLIT3	slit homolog 3 (Drosophila)	0.963	26	0
32242_at	CRYAB	crystallin, alpha B	0.963	26	0
32469_at	CEACAM3	carcinoembryonic antigen-related cell adhesion molecule 3	0.963	26	0
1276_g_at	RBPMS	RNA binding protein with multiple splicing	0.962	25	0
39175_at	PFKP	phosphofructokinase, platelet	0.962	25	0
32749_s_at	FLNA	filamin A, alpha (actin binding protein 280)	0.962	25	0
37664_at	DRG2	developmentally regulated GTP binding protein 2	0.962	25	0
38786_at	SVEP1	sushi, von Willebrand factor type A, EGF and pentraxin domain containing 1	0.962	25	0
38714_at	GYPA	glycophorin A (includes MN blood group)	0.960	24	0
39797_at	UBR2	ubiquitin protein ligase E3 component n-recognin 2	0.960	24	0
35438_at	PLXNA3	plexin A3	0.960	24	0
36283_at	DDX6	DEAD (Asp-Glu-Ala-Asp) box polypeptide 6	0.960	24	0
36443_at	DNAH9	dynein, axonemal, heavy polypeptide 9	0.960	24	0
33418_at	RAB3GAP1	RAB3 GTPase activating protein subunit 1 (catalytic)	0.959	71	1
37230_at	KLHL21	kelch-like 21 (Drosophila)	0.958	23	0
affx-murfas_at			0.958	23	0
33907_at	EIF4G3	eukaryotic translation initiation factor 4 gamma, 3	0.958	23	0
38736_at	WDR1	WD repeat domain 1	0.958	23	0
32971_at	C9orf61	chromosome 9 open reading frame 61	0.957	22	0
33388_at	TEX261	testis expressed sequence 261	0.957	22	0
32314_g_at	TPM2	tropomyosin 2 (beta)	0.955	21	0
39661_s_at	SLC29A2	solute carrier family 29 (nucleoside transporters), member 2	0.955	21	0
40776_at	DES	desmin	0.955	21	0
32458_f_at	PRB4	proline-rich protein BstNI subfamily 4	0.955	21	0
720_at	HSF4	heat shock transcription factor 4	0.955	21	0
32410_at	MYOG	myogenin (myogenic factor 4)	0.955	21	0
32239_at	MATN2	matrilin 2	0.952	20	0
39370_at	MAP1LC3B	microtubule-associated protein 1 light chain 3 beta	0.952	20	0
35413_s_at	ZNF22	zinc finger protein 22 (KOX 15)	0.952	20	0
39544_at	DMN	desmuslin	0.952	20	0
33643_at	MYCL1	v-myc myelocytomatosis viral oncogene homolog 1, lung carcinoma derived (avian)	0.952	20	0
32662_at	MDC1	mediator of DNA damage checkpoint 1	0.952	20	0
37968_at	NCR3	natural cytotoxicity triggering receptor 3	0.952	20	0
33472_at	FMO4	flavin containing monooxygenase 4	0.952	20	0
32624_at	GARNL1	GTPase activating Rap/RanGAP domain-like 1	0.952	20	0
1652_at	PIM2	pim-2 oncogene	0.952	20	0
41546_at	CDK6	cyclin-dependent kinase 6	0.951	59	1
41153_f_at	CTNNA1	catenin (cadherin-associated protein), alpha 1, 102kDa	0.949	57	1
34906_g_at	GRIK5	glutamate receptor, ionotropic, kainate 5	0.948	94	2

36423_at	P8		0.948	94	2
32255_i_at	TERF1	telomeric repeat binding factor (NIMA-interacting) 1	0.947	55	1
38355_at	DDX3Y	DEAD (Asp-Glu-Ala-Asp) box polypeptide 3, Y-linked	0.946	89	2
34788_at	C9orf132	chromosome 9 open reading frame 132	0.945	53	1
34933_at	PAX9	paired box gene 9	0.945	53	1
34306_at	MBNL1	muscleblind-like (Drosophila)	0.943	51	1
38205_at	NEUROD2	neurogenic differentiation 2	0.939	47	1
1992_at	FHIT	fragile histidine triad gene	0.936	106	3
37790_at	VENTX2	VENT-like homeobox 2	0.935	44	1
31535_i_at			0.933	43	1
1419_g_at	NOS2A	nitric oxide synthase 2A (inducible, hepatocytes)	0.932	42	1
31317_r_at			0.931	97	3
1775_at	POLA2	polymerase (DNA directed), alpha 2 (70kD subunit)	0.925	140	5
37680_at	AKAP12	A kinase (PRKA) anchor protein (gravin) 12	0.922	61	2
34507_s_at	LMTK2	lemur tyrosine kinase 2	0.921	232	9
38443_at	PTPN11	protein tyrosine phosphatase, non-receptor type 11 (Noonan syndrome 1)	0.921	36	1
36952_at	HADHA	hydroxyacyl-Coenzyme A dehydrogenase/3-ketoacyl-Coenzyme A thiolase/enoyl-Coenzyme A hydratase (trifunctional protein), alpha subunit	0.919	59	2
100_g_at	RABGGTA	Rab geranylgeranyltransferase, alpha subunit	0.919	35	1
39744_at	DDX3X	DEAD (Asp-Glu-Ala-Asp) box polypeptide 3, X-linked	0.914	33	1
38711_at	CLASP2	cytoplasmic linker associated protein 2	0.911	53	2
36946_at	DYRK1A	dual-specificity tyrosine-(Y)-phosphorylation regulated kinase 1A	0.909	31	1
33520_at	F7	coagulation factor VII (serum prothrombin conversion accelerator)	0.909	31	1
41097_at	TERF2	telomeric repeat binding factor 2	0.907	51	2
39639_s_at	TNP1	transition protein 1 (during histone to protamine replacement)	0.894	44	2
935_at	CAP1	CAP, adenylate cyclase-associated protein 1 (yeast)	0.893	26	1
41001_at	RPH3A	rabphilin 3A homolog (mouse)	0.892	61	3
36659_at	COL4A2	collagen, type IV, alpha 2	0.889	25	1
38221_at	CNKSR1	connector enhancer of kinase suppressor of Ras 1	0.889	42	2
37939_at	APOBEC3C	apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like 3C	0.889	25	1
37814_g_at	DDX51	DEAD (Asp-Glu-Ala-Asp) box polypeptide 51	0.887	91	5
40326_at	CBLN1	cerebellin 1 precursor	0.885	24	1
34702_f_at	psiTPTE22		0.885	24	1
32897_at	MTHFR	5,10-methylenetetrahydrofolate reductase (NADPH)	0.884	40	2
34559_at	LOC388818		0.884	40	2
39110_at	EIF4B	eukaryotic translation initiation factor 4B	0.884	40	2
39539_at	ZBTB7A	zinc finger and BTB domain containing 7A	0.875	112	7
32340_s_at	YBX1	Y box binding protein 1	0.875	22	1
37517_at	GARNL4	GTPase activating Rap/RanGAP domain-like 4	0.870	21	1
39346_at	KHDRBS1	KH domain containing, RNA binding, signal transduction associated 1	0.870	21	1
33565_at	TSHB	thyroid stimulating hormone, beta	0.865	34	2
32007_at			0.865	34	2
32569_at	PAFAH1B1	platelet-activating factor acetylhydrolase, isoform Ib, alpha subunit 45kDa	0.863	47	3
35536_at	ECE2	endothelin converting enzyme 2	0.848	30	2
41445_at	TGFB1	transforming growth factor, beta 1 (Camurati-Engelmann disease)	0.848	42	3
36517_at	U2AF1	U2(RNU2) small nuclear RNA auxiliary factor 1	0.844	29	2
37082_at	ZNF96	zinc finger protein 96	0.844	29	2
1793_at	CDC2L5	cell division cycle 2-like 5 (cholinesterase-related cell division controller)	0.841	40	3
41703_r_at	AKAP7	A kinase (PRKA) anchor protein 7	0.839	28	2
37284_at	SEMA4D	sema domain, immunoglobulin domain (Ig), transmembrane domain (TM) and short cytoplasmic domain, (semaphorin) 4D	0.833	60	5
1709_g_at	MAPK10	mitogen-activated protein kinase 10	0.829	37	3
40636_at	FLOT1	flotillin 1	0.828	79	7
336_at	TBXA2R	thromboxane A2 receptor	0.821	25	2
34457_at	SLC30A3	solute carrier family 30 (zinc transporter), member 3	0.821	25	2
39263_at	OAS2	2'-5'-oligoadenylate synthetase 2, 69/71kDa	0.817	54	5
34018_at	COL19A1	collagen, type XIX, alpha 1	0.816	34	3
41190_at	TNFRSF25	tumor necrosis factor receptor superfamily, member 25	0.805	78	8
33080_s_at	RAP1GA1	RAP1, GTPase activating protein 1	0.800	22	2
32119_at			0.800	31	3
774_g_at	MYH11	myosin, heavy polypeptide 11, smooth muscle	0.800	22	2
39423_f_at	SPOP	speckle-type POZ protein	0.800	22	2
35339_at	RAB8A	RAB8A, member RAS oncogene family	0.795	65	7
40089_at	WBSCR22	Williams Beuren syndrome chromosome region 22	0.794	30	3
35598_at	HIST1H3E	histone 1, H3e	0.792	47	5
37620_at	TAF12	TAF12 RNA polymerase II, TATA box binding protein (TBP)-associated factor, 20kDa	0.786	37	4
38062_at	RAPGEF5	Rap guanine nucleotide exchange factor (GEF) 5	0.785	70	8
32263_at	CCNB2	cyclin B2	0.780	36	4
1742_at	ERBB3	v-erb-b2 erythroblastic leukemia viral oncogene homolog 3 (avian)	0.776	43	5
34467_g_at	HTR4	5-hydroxytryptamine (serotonin) receptor 4	0.771	42	5

33865_at	ZMYND11	zinc finger, MYND domain containing 11	0.771	42	5
37294_at	BTG1	B-cell translocation gene 1, anti-proliferative	0.767	26	3
1782_s_at	STMN1	stathmin 1/oncoprotein 18	0.759	25	3
31889_at	MLANA	melan-A	0.759	25	3
40890_at	MTX1	metaxin 1	0.750	24	3
37122_at	PLIN	perilipin	0.748	93	13
38942_r_at	SPBC25	spindle pole body component 25 homolog (S. cerevisiae)	0.743	30	4
37551_at	ZNF592	zinc finger protein 592	0.743	30	4
36355_at	IVL	involucrin	0.743	30	4
32490_at	CEACAM4	carcinoembryonic antigen-related cell adhesion molecule 4	0.741	23	3
36098_at	SFRS1	splicing factor, arginine/serine-rich 1 (splicing factor 2, alternate splicing factor)	0.729	41	6
33246_at	MAPK13	mitogen-activated protein kinase 13	0.725	122	19
40832_s_at	TOR1AIP1	torsin A interacting protein 1	0.711	38	6
41840_r_at	ANTXR1	anthrax toxin receptor 1	0.710	26	4
41752_at	GHITM	growth hormone inducible transmembrane protein	0.710	26	4
31429_at			0.690	73	13
41276_at	SAP18	sin3-associated polypeptide, 18kDa	0.688	40	7
39103_s_at	DHRS1	dehydrogenase/reductase (SDR family) member 1	0.683	50	9
33069_f_at	UGT2B15	UDP glucuronosyltransferase 2 family, polypeptide B15	0.679	44	8
34606_s_at	ATF7	activating transcription factor 7	0.676	28	5
35596_at	MGC11271		0.651	68	14
39790_at	ATP2A2	ATPase, Ca++ transporting, cardiac muscle, slow twitch 2	0.648	58	12
40904_at	PRPF31	PRP31 pre-mRNA processing factor 31 homolog (yeast)	0.641	125	27
1468_at	TRAP1	TNF receptor-associated protein 1	0.639	29	6
38928_r_at	TYR	tyrosinase (oculocutaneous albinism IA)	0.618	27	6
1567_at	FLT1	fms-related tyrosine kinase 1 (vascular endothelial growth factor/vascular permeability factor receptor)	0.618	27	6
36865_at	ANGEL1	angel homolog 1 (Drosophila)	0.615	52	12
35323_at	EIF3S9	eukaryotic translation initiation factor 3, subunit 9 eta, 116kDa	0.615	31	7
32710_at	KCNAB1	potassium voltage-gated channel, shaker-related subfamily, beta member 1	0.595	33	8
37779_at	SMPDL3B	sphingomyelin phosphodiesterase, acid-like 3B	0.543	35	10
37270_at	ATP1B2	ATPase, Na+/K+ transporting, beta 2 polypeptide	0.539	58	17
728_at			0.515	51	16
38091_at	LGALS9	lectin, galactoside-binding, soluble, 9 (galectin 9)	0.500	34	11
39730_at	ABL1	v-abl Abelson murine leukemia viral oncogene homolog 1	0.494	57	19
34558_at	OPRL1	opiate receptor-like 1	0.490	144	49
41799_at	DNAJC7	DnaJ (Hsp40) homolog, subfamily C, member 7	0.479	35	12
1339_s_at	BCR	breakpoint cluster region	0.444	32	12
40336_at	FDXR	ferredoxin reductase	0.427	58	23
39855_at	FZR1	fizzy/cell division cycle 20 related 1 (Drosophila)	0.407	41	17
396_f_at	EPOR	erythropoietin receptor	0.359	43	20
32560_s_at	LSM4	LSM4 homolog, U6 small nuclear RNA associated (S. cerevisiae)	0.321	51	26
36894_at	CBX7	chromobox homolog 7	0.319	47	24
1409_at	SRF	serum response factor (c-fos serum response element-binding transcription factor)	0.304	51	27
36991_at	SFRS4	splicing factor, arginine/serine-rich 4	0.296	109	59
33686_at	SPINT3	serine peptidase inhibitor, Kunitz type, 3	0.244	53	32
31819_at	MGC34821		0.209	95	62

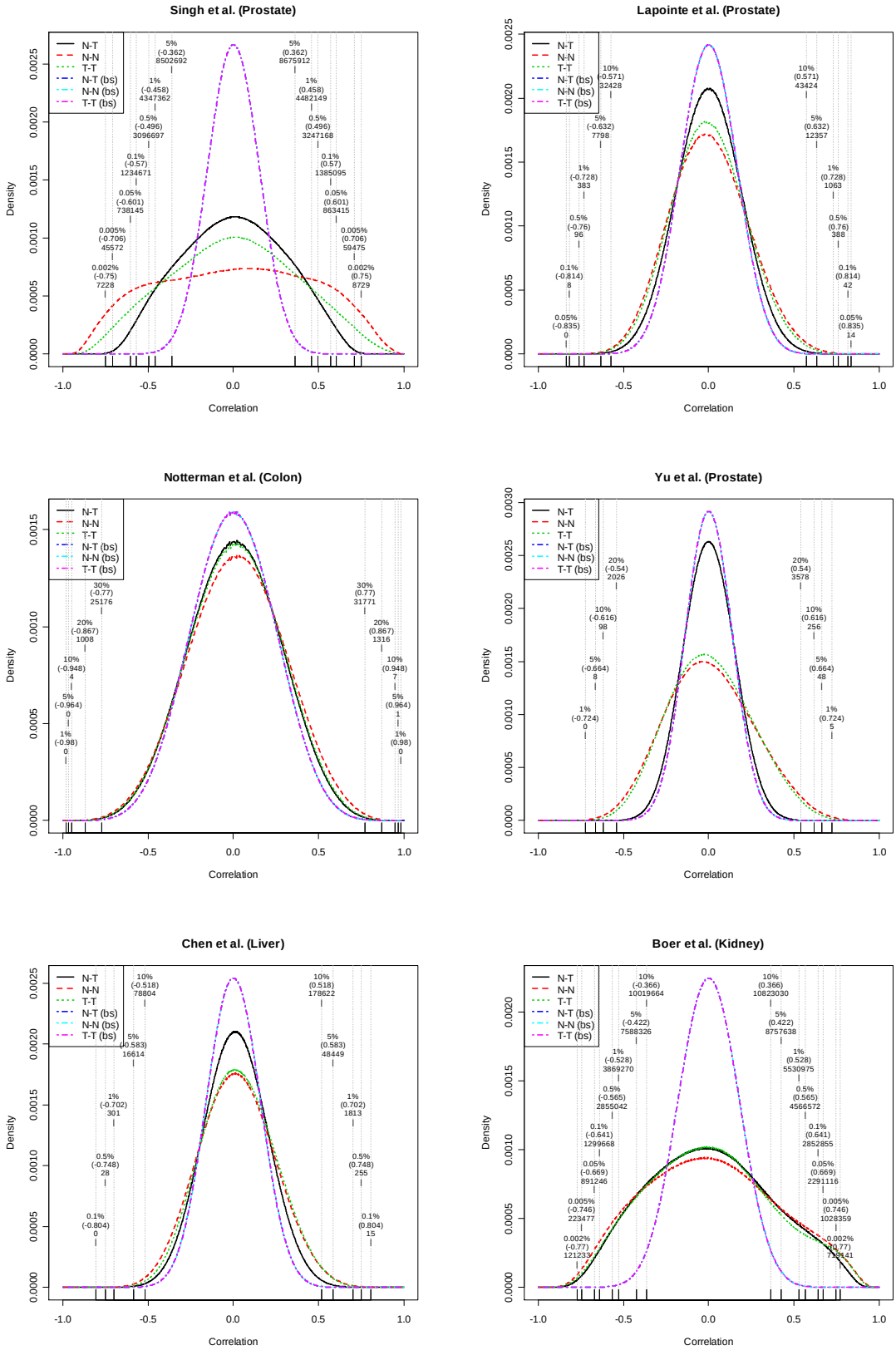
Supplementary Table 6.2 - – Negative polarized genes. Sorted by *pol*. For a gene expressed in normal, *f* is the number of significant correlations with tumour genes whereas *b* is the number of significant correlations with normal genes for the same gene expressed in tumour.

Affy Id	*Symbol*	*Gene Name*	*pol(-)*	*F*	*b*
31311_at			-0.998	0	545
37281_at	FAM38A	family with sequence similarity 38, member A	-0.998	0	467
41302_at	AHCYL1	S-adenosylhomocysteine hydrolase-like 1	-0.997	0	345
34652_at	NPAS1	neuronal PAS domain protein 1	-0.997	0	311
41766_at	MAN2A2	mannosidase, alpha, class 2A, member 2	-0.996	0	241
37292_at	MAML1	mastermind-like 1 (Drosophila)	-0.996	0	235
41292_at	HNRPH1	heterogeneous nuclear ribonucleoprotein H1 (H)	-0.996	0	225
39180_at	FUS	fusion (involved in t(12,16) in malignant liposarcoma)	-0.995	0	189
1602_at	PRKCI	protein kinase C, iota	-0.995	0	189
32085_at	PIP5K3	phosphatidylinositol-3-phosphate/phosphatidylinositol 5-kinase, type III	-0.995	0	186
32230_at	EIF3S2	eukaryotic translation initiation factor 3, subunit 2 beta, 36kDa	-0.994	0	176
39024_at	NUP98	nucleoporin 98kDa	-0.994	0	158
39117_at	PHF2	PHD finger protein 2	-0.994	0	156
38122_at	SLC23A2	solute carrier family 23 (nucleobase transporters), member 2	-0.993	0	145
36186_at	RNPS1	RNA binding protein S1, serine-rich domain	-0.993	0	142

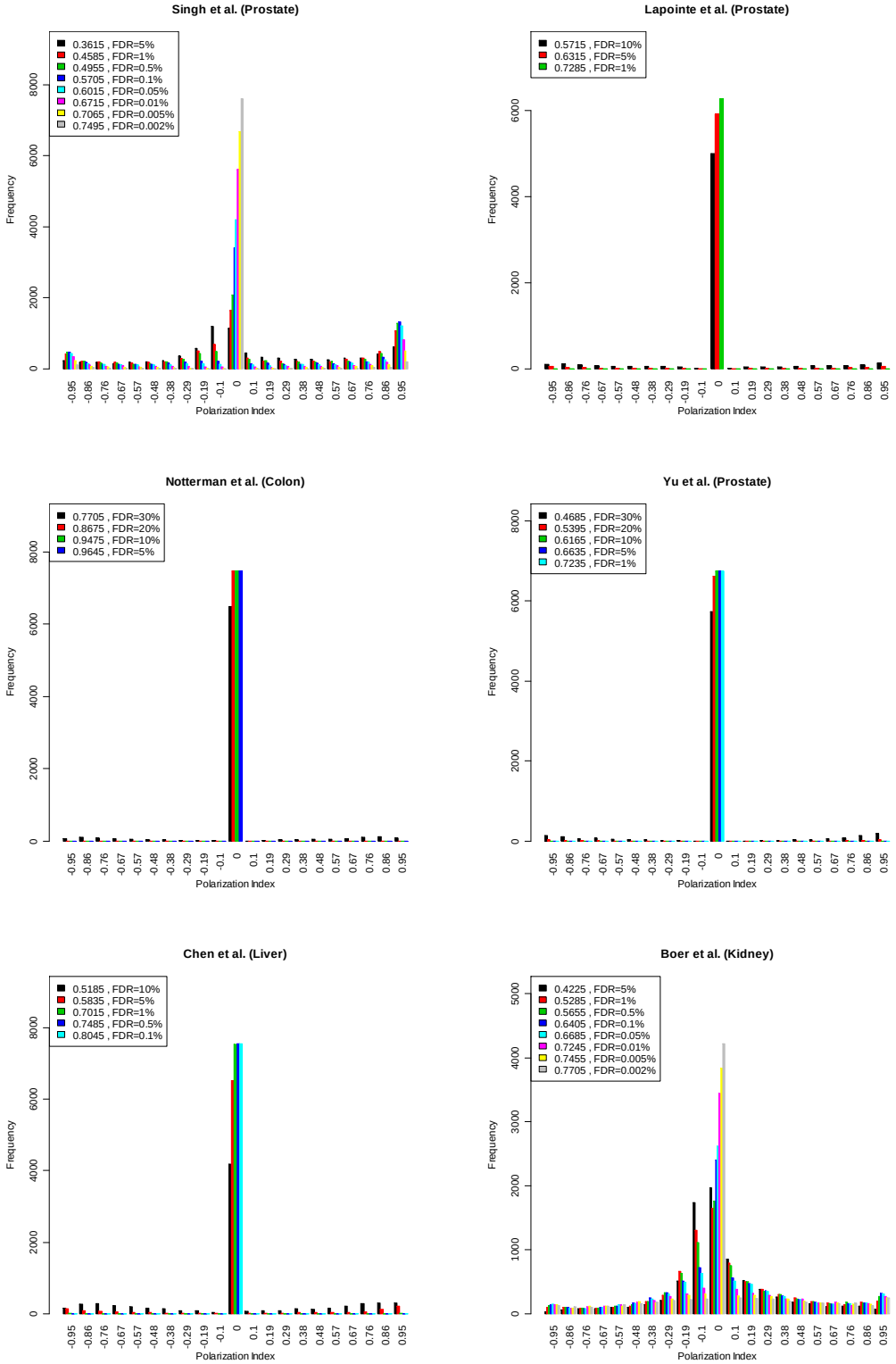
34169_s_at	OCRL	oculocerebrorenal syndrome of Lowe	-0.993	0	137
39435_at	PFDN1	prefoldin 1	-0.992	0	132
40491_at	ARID4B	AT rich interactive domain 4B (RBP1- like)	-0.991	0	116
32804_at	RBM5	RNA binding motif protein 5	-0.991	2	542
1556_at	RBM5	RNA binding motif protein 5	-0.990	0	99
40998_at	MED12	mediator of RNA polymerase II transcription, subunit 12 homolog (yeast)	-0.990	0	97
242_at	MAP4	microtubule-associated protein 4	-0.990	0	97
32508_at	BAT2D1	BAT2 domain containing 1	-0.989	1	265
34305_at	PCBP1	poly(rC) binding protein 1	-0.988	0	85
34754_at	SCYL3	SCY1-like 3 (S. cerevisiae)	-0.987	0	75
37676_at	PDE8A	phosphodiesterase 8A	-0.986	0	69
39405_at	UTP14C	UTP14, U3 small nucleolar ribonucleoprotein, homolog C (yeast)	-0.984	0	62
40870_g_at	RBM6	RNA binding motif protein 6	-0.984	0	61
31447_at	SCC-112		-0.984	0	61
36576_at	H2AFY	H2A histone family, member Y	-0.983	1	177
37450_r_at	GNAS	GNAS complex locus	-0.981	0	53
41183_at	CSTF3	cleavage stimulation factor, 3' pre-RNA, subunit 3, 77kDa	-0.981	1	159
36153_at	DHX9	DEAH (Asp-Glu-Ala-His) box polypeptide 9	-0.981	1	154
40826_at	MARK3	MAP/microtubule affinity-regulating kinase 3	-0.981	0	51
41399_at	PHF8	PHD finger protein 8	-0.980	0	50
33218_at	ERBB2	v-erb-b2 erythroblastic leukemia viral oncogene homolog 2, neuro/glioblastoma derived oncogene homolog (avian)	-0.980	0	50
34767_at	MOAP1	modulator of apoptosis 1	-0.980	2	251
41258_at	WBSCR20C		-0.980	0	49
33931_at	GPX4	glutathione peroxidase 4 (phospholipid hydroperoxidase)	-0.980	0	49
35242_at	PCTK3	PCTAIRE protein kinase 3	-0.980	0	49
39112_at	USF2	upstream transcription factor 2, c-fos interacting	-0.980	0	48
36971_at	RW1		-0.980	0	48
41841_at	BRD3	bromodomain containing 3	-0.980	0	48
32654_g_at	BRD8	bromodomain containing 8	-0.979	0	46
41836_at	CHERP	calcium homeostasis endoplasmic reticulum protein	-0.978	0	45
40869_at	RBM6	RNA binding motif protein 6	-0.978	0	45
37449_i_at	GNAS	GNAS complex locus	-0.978	0	44
34192_at	VPS13B	vacuolar protein sorting 13B (yeast)	-0.977	0	43
35477_at	ZNF79	zinc finger protein 79 (pT7)	-0.977	0	43
329_s_at			-0.976	0	41
40835_at	MTA2	metastasis associated 1 family, member 2	-0.976	0	40
36784_at	CSHL1	chorionic somatomammotropin hormone-like 1	-0.976	0	40
32209_at	FAM89B	family with sequence similarity 89, member B	-0.975	7	601
1863_s_at	ATM	ataxia telangiectasia mutated (includes complementation groups A, C and D)	-0.975	0	39
35745_f_at	PCBP2	poly(rC) binding protein 2	-0.975	1	116
39866_at	USP22	ubiquitin specific peptidase 22	-0.974	2	193
41528_at	LOC130074		-0.974	0	38
1616_at	FGF9	fibroblast growth factor 9 (glia-activating factor)	-0.973	0	36
32439_at	ATP4B	ATPase, H+/K+ exchanging, beta polypeptide	-0.972	0	35
40506_s_at	PABPC4	poly(A) binding protein, cytoplasmic 4 (inducible form)	-0.972	0	35
38402_at	LAMP2	lysosomal-associated membrane protein 2	-0.972	0	35
40988_at	YME1L1	YME1-like 1 (S. cerevisiae)	-0.972	4	314
33885_at	KIAA0907	KIAA0907	-0.971	0	33
34148_at	SIX3	sine oculis homeobox homolog 3 (Drosophila)	-0.971	0	33
319_g_at	H1FX	H1 histone family, member X	-0.970	0	32
41553_at	C8orf1	chromosome 8 open reading frame 1	-0.970	0	32
39240_at	NRXN1	neurexin 1	-0.970	0	32
35436_at	GOLGA2	golgi autoantigen, golgin subfamily a, 2	-0.970	3	226
33360_at	FBXL11	F-box and leucine-rich repeat protein 11	-0.967	0	29
38650_at	IGFBP5	insulin-like growth factor binding protein 5	-0.966	0	28
37744_r_at	FEZ1	fasciculation and elongation protein zeta 1 (zygin I)	-0.964	0	27
39520_at	KIAA0692		-0.964	0	27
1741_s_at	IGFBP2	insulin-like growth factor binding protein 2, 36kDa	-0.964	0	27
35030_i_at	PDZK10	PDZ domain containing 10	-0.963	0	26
1644_at	EIF3S2	eukaryotic translation initiation factor 3, subunit 2 beta, 36kDa	-0.963	0	26
1908_at	ETV3	ets variant gene 3	-0.962	2	130
37597_s_at	SEC6L1	SEC6-like 1 (S. cerevisiae)	-0.962	0	25
1728_at	PCGF4	polycomb group ring finger 4	-0.962	0	25
33120_at	RGS10	regulator of G-protein signalling 10	-0.960	0	24
32171_at	EIF5	eukaryotic translation initiation factor 5	-0.960	0	24
32164_at	EXT1	exostoses (multiple) 1	-0.960	0	24
35618_at	HELZ	helicase with zinc finger	-0.958	0	23
40845_at	ILF3	interleukin enhancer binding factor 3, 90kDa	-0.958	0	23
36962_at	COPA	coatomer protein complex, subunit alpha	-0.956	2	111

31594_at	KRTHA3A	keratin, hair, acidic, 3A	-0.955	1	64
affx-dapx-m_at			-0.955	0	21
34778_at	LRRC15	leucine rich repeat containing 15	-0.955	0	21
40946_at	ISG20L2	interferon stimulated exonuclease gene 20kDa-like 2	-0.955	0	21
35746_r_at	PCBP2	poly(rC) binding protein 2	-0.952	0	20
40928_at	WSB1	WD repeat and SOCS box-containing 1	-0.952	1	61
34634_s_at	HTR7	5-hydroxytryptamine (serotonin) receptor 7 (adenylate cyclase-coupled)	-0.947	3	127
40780_at	CTBP2	C-terminal binding protein 2	-0.945	1	53
33325_at	RPS6KA2	ribosomal protein S6 kinase, 90kDa, polypeptide 2	-0.944	1	52
919_at			-0.943	1	51
1161_at	HSPCB	heat shock 90kDa protein 1, beta	-0.943	1	51
36592_at	PHB	prohibitin	-0.940	1	48
35450_s_at	GTF2I	general transcription factor II, i	-0.933	1	43
34493_at	TNFRSF10C	tumor necrosis factor receptor superfamily, member 10c, decoy without an intracellular domain	-0.930	13	374
156_s_at	GPR19	G protein-coupled receptor 19	-0.929	7	202
35514_at	MCF2L2	MCF.2 cell line derived transforming sequence-like 2	-0.928	14	387
37448_s_at	GNAS	GNAS complex locus	-0.928	3	93
41316_s_at	SAFB	scaffold attachment factor B	-0.925	1	38
33539_at	MYEF2	myelin expression factor 2	-0.923	2	62
36226_r_at	SFPQ	splicing factor proline/glutamine-rich (polypyrimidine tract binding protein associated)	-0.919	1	35
35292_at	BAT1	HLA-B associated transcript 1	-0.919	14	342
31536_at	RTN4	reticulon 4	-0.918	2	58
41358_at	CNNM2	cyclin M2	-0.914	1	33
40553_at	LOC441079		-0.907	4	92
2056_at	FGFR1	fibroblast growth factor receptor 1 (fms-related tyrosine kinase 2, Pfeiffer syndrome)	-0.906	1	30
32474_at	PAX7	paired box gene 7	-0.906	2	50
39229_at	SDCCAG1	serologically defined colon cancer antigen 1	-0.906	2	50
39330_s_at	ACTN1	actinin, alpha 1	-0.904	5	109
36388_at	PTHR2	parathyroid hormone receptor 2	-0.902	2	48
34026_at			-0.902	6	125
36209_at	BRD2	bromodomain containing 2	-0.900	1	28
33683_at	TI-227H		-0.897	1	27
38393_at	KIAA0247	KIAA0247	-0.893	1	26
33027_at	SYNPO2L	synaptopodin 2-like	-0.889	1	25
2030_at	WNT11	wingless-type MMTV integration site family, member 11	-0.885	3	57
31452_at	SMNP		-0.880	1	23
40322_at	IL1RL1	interleukin 1 receptor-like 1	-0.875	2	37
37415_at	ATP10B	ATPase, Class V, type 10B	-0.873	3	51
1008_f_at	EIF2AK2	eukaryotic translation initiation factor 2-alpha kinase 2	-0.870	3	50
41529_g_at	LOC130074		-0.870	60	867
39097_at	SON	SON DNA binding protein	-0.870	7	107
36702_at	TBX19	T-box 19	-0.870	1	21
1643_g_at	MTA1	metastasis associated 1	-0.870	1	21
41366_at	KIAA1002		-0.868	2	35
31961_r_at			-0.865	2	34
33543_s_at	PNN	pinin, desmosome associated protein	-0.861	2	33
32284_at	TBX1	T-box 1	-0.857	3	45
38860_at	PDE4C	phosphodiesterase 4C, cAMP-specific (phosphodiesterase E1 dunce homolog, Drosophila)	-0.848	2	30
38854_at	CEP4	centrosomal protein 4	-0.845	4	53
31923_f_at	DKFZP586A0522		-0.844	2	29
240_at	SRM	spermidine synthase	-0.839	2	28
38527_at	NONO	non-POU domain containing, octamer-binding	-0.828	5	58
37425_g_at	CCHCR1	coiled-coil alpha-helical rod protein 1	-0.821	3	35
1396_at	IGFBP5	insulin-like growth factor binding protein 5	-0.815	2	24
37226_at	BNIP1	BCL2/adenovirus E1B 19kDa interacting protein 1	-0.806	6	60
39295_s_at	ARGBP2		-0.800	2	22
37675_at	SLC25A3	solute carrier family 25 (mitochondrial carrier, phosphate carrier), member 3	-0.796	60	532
526_s_at	PMS2	PMS2 postmeiotic segregation increased 2 (S. cerevisiae)	-0.781	3	28
31463_s_at	HNRPA1	heterogeneous nuclear ribonucleoprotein A1	-0.776	22	178
36026_at	PGAM2	phosphoglycerate mutase 2 (muscle)	-0.769	4	34
40790_at	BHLHB2	basic helix-loop-helix domain containing, class B, 2	-0.767	3	26
32303_at	ETV3	ets variant gene 3	-0.750	4	31
38624_at	SLC12A4	solute carrier family 12 (potassium/chloride transporters), member 4	-0.745	6	44
35956_s_at	PSG7	pregnancy specific beta-1-glycoprotein 7	-0.744	5	37
affx-hsac07/x00351_m_at			-0.701	11	65
34841_at	EIF3S8	eukaryotic translation initiation factor 3, subunit 8, 110kDa	-0.690	4	24
834_at	ZNFN1A1	zinc finger protein, subfamily 1A, 1 (Ikaros)	-0.683	9	50
36338_at	AOF2	amine oxidase (flavin containing) domain 2	-0.681	7	39
38901_at	USP19	ubiquitin specific peptidase 19	-0.676	5	28

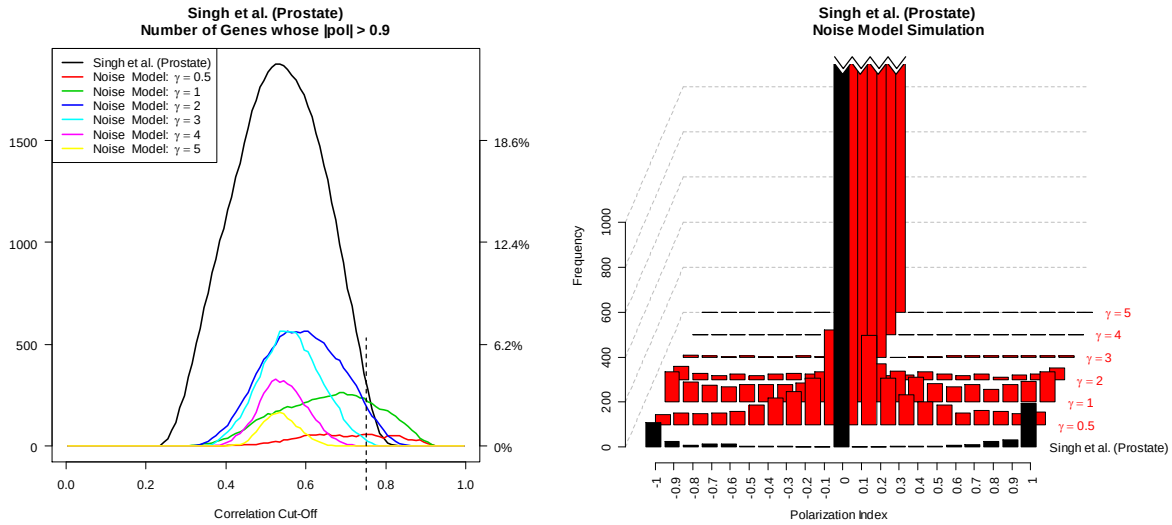
36285_at	KCNJ4	potassium inwardly-rectifying channel, subfamily J, member 4	-0.676	5	28
33943_at	FTH1	ferritin, heavy polypeptide 1	-0.672	10	53
40984_at	76P		-0.667	5	27
32588_s_at	ZFP36L2	zinc finger protein 36, C3H type-like 2	-0.664	125	620
37031_at	C9orf10	chromosome 9 open reading frame 10	-0.658	6	31
36004_at	IKBKG	inhibitor of kappa light polypeptide gene enhancer in B-cells, kinase gamma	-0.655	19	93
32152_at	ANK1	ankyrin 1, erythrocytic	-0.643	7	34
40723_at	SIT1	signaling threshold regulating transmembrane adaptor 1	-0.638	12	56
33221_at	PAXIP1	PAX interacting (with transcription-activation domain) protein 1	-0.632	19	86
41155_at	CTNNA1	catenin (cadherin-associated protein), alpha 1, 102kDa	-0.627	9	41
37195_at	CYP11A1	cytochrome P450, family 11, subfamily A, polypeptide 1	-0.618	10	44
37053_at	ATP2B2	ATPase, Ca++ transporting, plasma membrane 2	-0.614	24	102
36328_at	SHBG	sex hormone-binding globulin	-0.612	9	39
40725_at	GOSR1	golgi SNAP receptor complex member 1	-0.590	45	176
41225_at	DUSP3	dual specificity phosphatase 3 (vaccinia virus phosphatase VH1-related)	-0.589	11	44
39692_at	CREB3L2	cAMP responsive element binding protein 3-like 2	-0.578	9	35
36888_at	KIAA0841	KIAA0841	-0.578	9	35
37839_at			-0.566	16	59
39508_at	NDRG2	NDRG family member 2	-0.564	8	30
34727_at	AHCYL1	S-adenosylhomocysteine hydrolase-like 1	-0.561	14	51
38400_at	FAM61A	family with sequence similarity 61, member A	-0.557	13	47
31936_s_at	LKAP		-0.548	9	32
32247_at	GTF2I	general transcription factor II, i	-0.544	33	113
32297_s_at	KLRC2	killer cell lectin-like receptor subfamily C, member 2	-0.540	14	48
32328_at	KRTHB5	keratin, hair, basic, 5	-0.538	24	81
32576_at	EIF3S5	eukaryotic translation initiation factor 3, subunit 5 epsilon, 47kDa	-0.536	48	160
38066_at	NQO1	NAD(P)H dehydrogenase, quinone 1	-0.528	83	270
34302_at	EIF3S4	eukaryotic translation initiation factor 3, subunit 4 delta, 44kDa	-0.512	10	32
39200_s_at	CIP29		-0.508	14	44
39839_at	CSDA	cold shock domain protein A	-0.502	81	245
1998_i_at	BAX	BCL2-associated X protein	-0.500	25	76
40203_at	SUI1		-0.497	42	126
39065_s_at	TTC3	tetratricopeptide repeat domain 3	-0.465	11	31
39845_at	HTRA2	HtrA serine peptidase 2	-0.426	15	38
31668_f_at	EPB41L2	erythrocyte membrane protein band 4.1-like 2	-0.424	28	70
33817_at	HNRPA3P1	heterogeneous nuclear ribonucleoprotein A3 pseudogene 1	-0.392	15	35
34538_at	LOC145678		-0.372	29	64
32855_at	LDLR	low density lipoprotein receptor (familial hypercholesterolemia)	-0.369	26	57
1019_g_at	WNT10B	wingless-type MMTV integration site family, member 10B	-0.350	19	40
41156_g_at	CTNNA1	catenin (cadherin-associated protein), alpha 1, 102kDa	-0.326	46	91
31324_at	ATP8A2	ATPase, aminophospholipid transporter-like, Class I, type 8A, member 2	-0.308	31	59
33856_at	CXX1	CAAX box 1	-0.277	49	87
36298_at	PRPH	peripherin	-0.267	42	73
32814_at	IFIT1	interferon-induced protein with tetratricopeptide repeats 1	-0.264	33	57
38558_at	MAG	myelin associated glycoprotein	-0.127	140	181
32505_at	GRINA	glutamate receptor, ionotropic, N-methyl D-asparate-associated protein 1 (glutamate binding)	-0.064	160	182



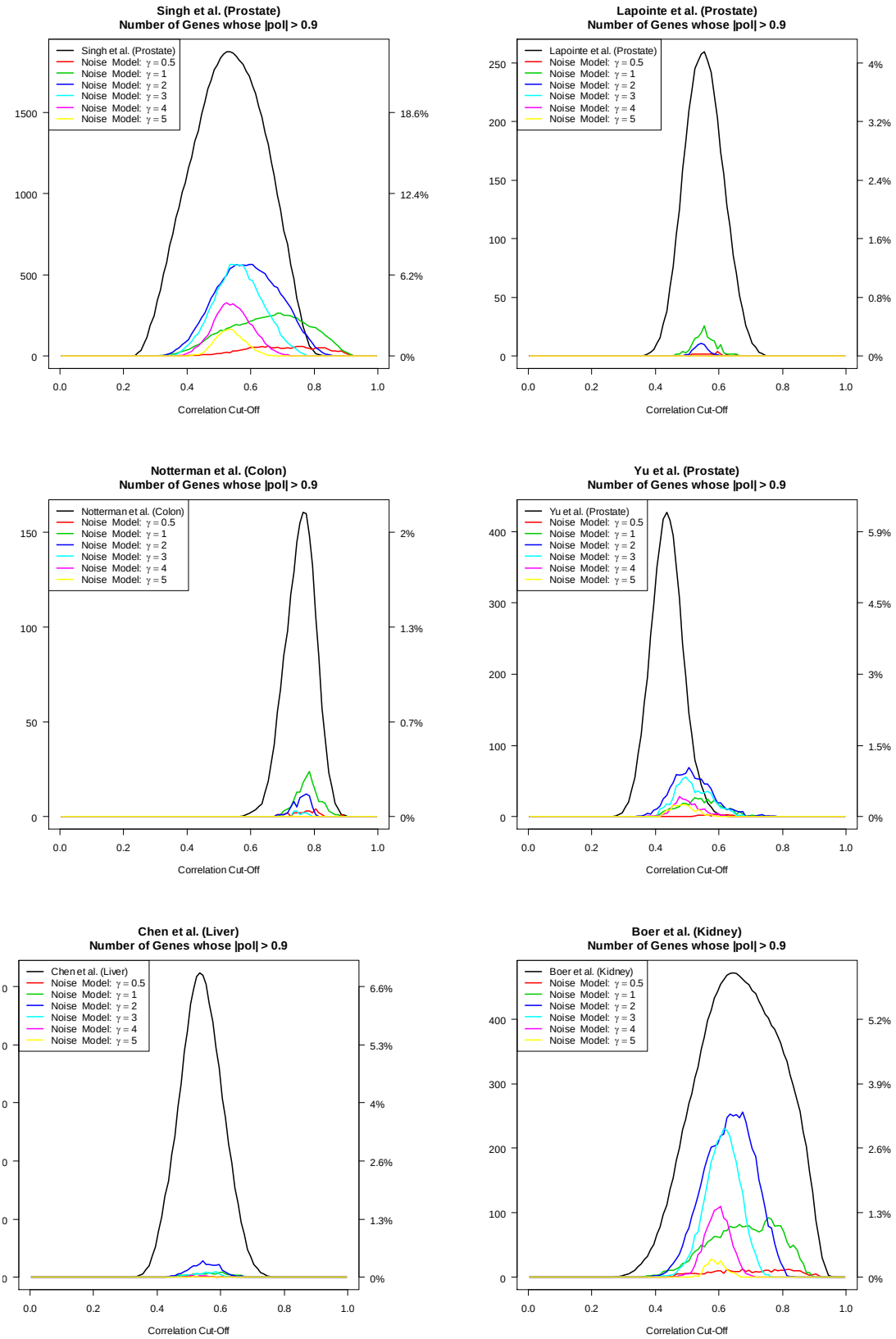
Supplementary Figure 6.1 – Correlation distributions in all datasets studied.



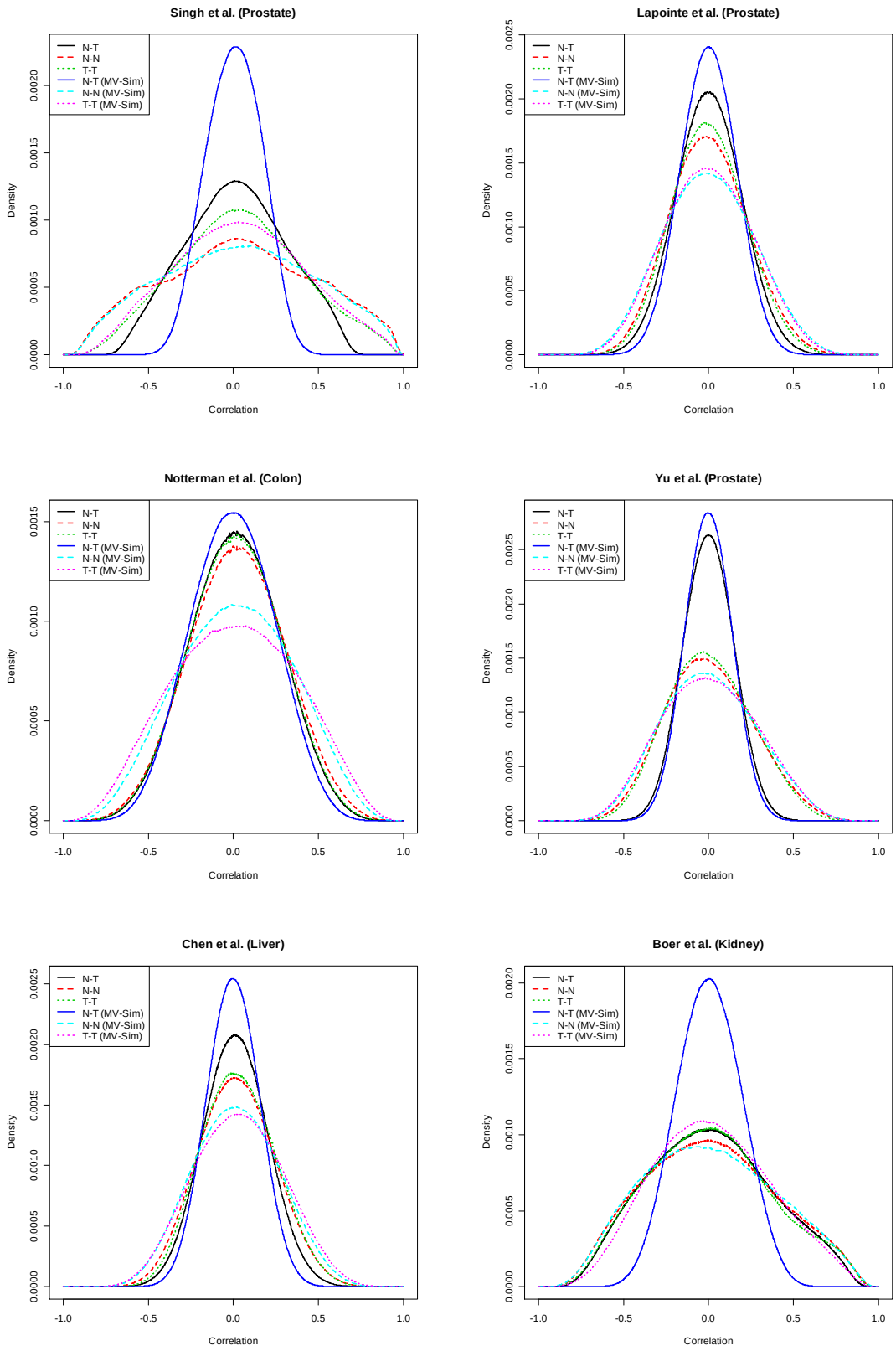
Supplementary Figure 6.2 – Polarization for all datasets studied at selected FDR correlation cut-offs.



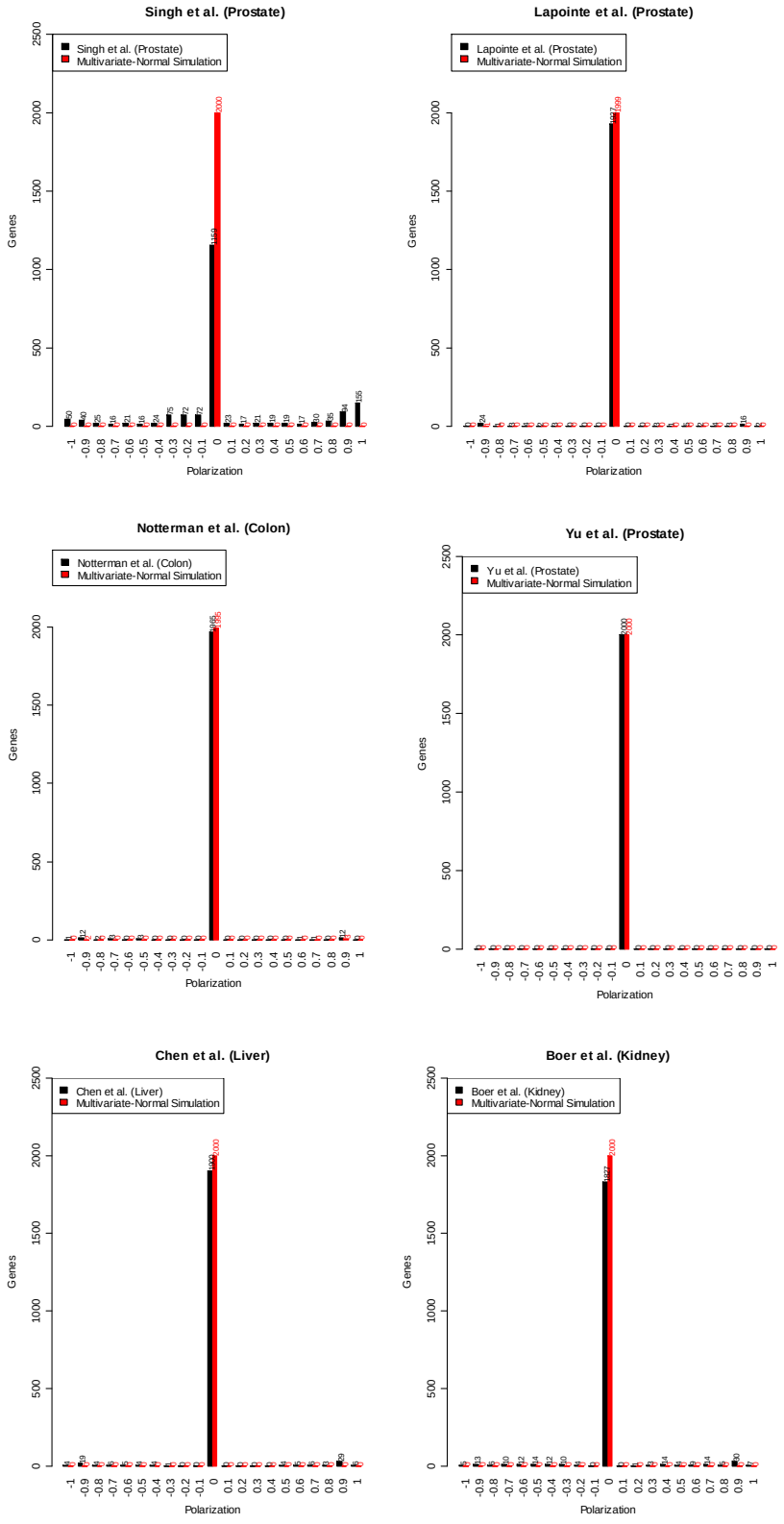
Supplementary Figure 6.3 – Dependence of *pol* to noise level factor γ. Left panel shows the number of genes with high *pol* (absolute value greater than 0.9). Right panel shows the distribution of *pol* across noise levels at the correlation cut-off chosen (0.75). In the right panel, the top section of the histograms has been omitted to clarify the comparison.



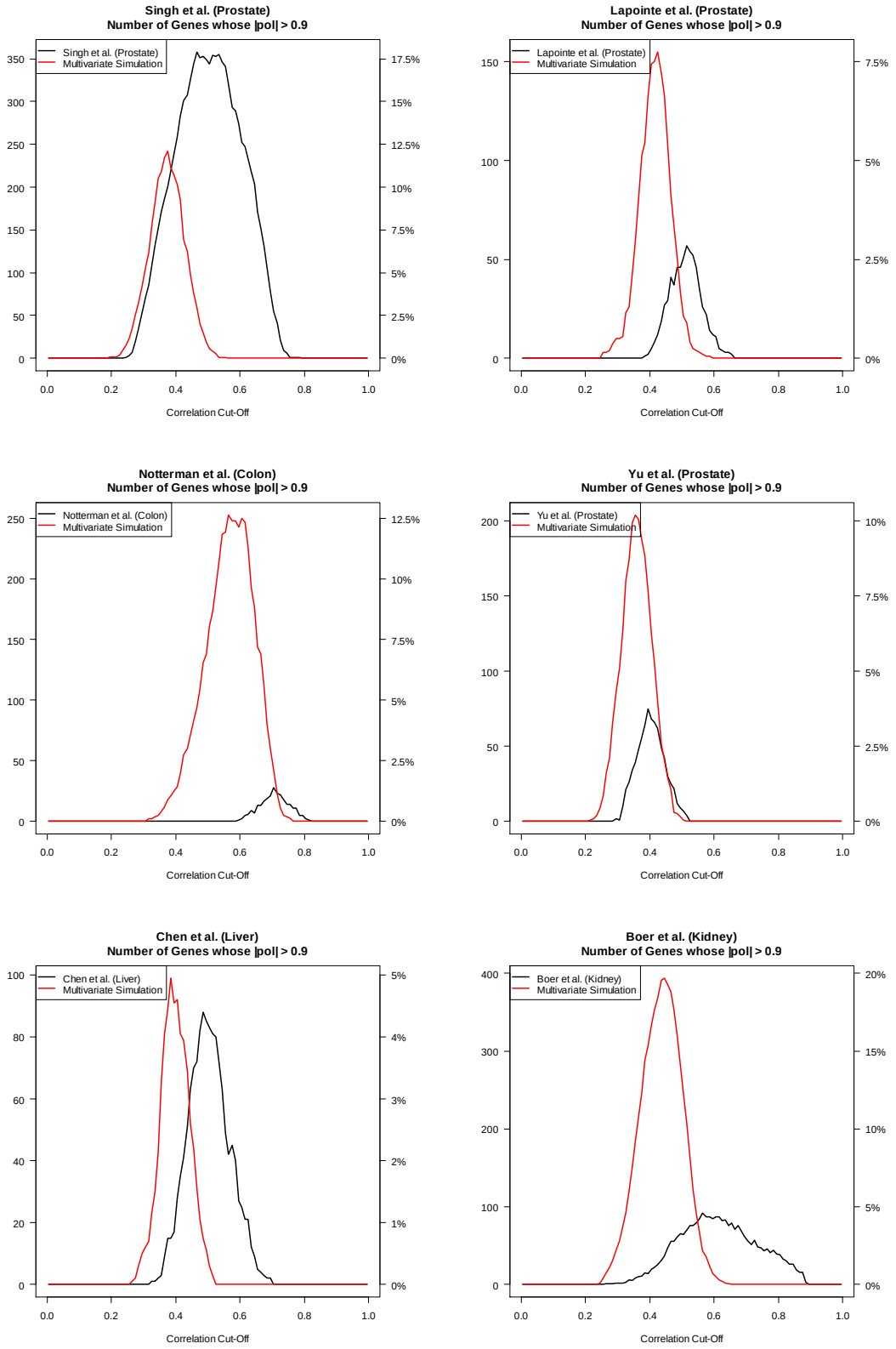
Supplementary Figure 6.4 – Number of genes with high values of *pol* in all datasets studied.



Supplementary Figure 6.5 – Comparison of the correlation distributions of the multivariate Gaussian generated data for all datasets studied.



Supplementary Figure 6.6 – Comparison of *pol* distribution for the multivariate Gaussian generated datasets.



Supplementary Figure 6.7 – *pol* dependency to correlation cut-off in the multivariate-Gaussian generated datasets.

CHAPTER 7

Conclusions and Discussions on the Biological Findings

7.1 - Multivariate Variable Selection

An important part of this thesis concerns the development and validation of a methodology for multivariate variable selection based on Genetic Algorithms. Although an approach based on GA has been previously proposed for the analysis of microarray data, our approach is more general, it allows the application of these methodologies to regression and survival analysis, it is capable of using several classification machines or coupling a user-specific one, it offers more options in error estimation procedures, and it provides a powerful graphical framework. The importance of parameters such as chromosome size in influencing model predictivity has been shown. A series of functions to analyze and optimize populations of models discovered using GA has also been developed. It is also important to stress the fact that these methodologies have been implemented in the statistical modelling environment R making it a very flexible system and indeed accessible to the average Bioinformatician.

Many questions remain open. An important issue is whether other methods for multivariate variable selection (for example the Bayesian Variable Selection methods described briefly in Chapter 2 and applied in Chapter 5) offer any advantage with respect to evolutionary algorithms. Do they explore a similar space of solutions? Because of the diverse strategy employed to search in the variable space we may be tempted to hypothesize that this is the case. The results described in Chapter 5 suggest

however that the genes most frequently included in the models discovered with GA are in part overlapping with the models discovered with the BVS methodology that have highest posterior probability. As a matter of fact more research is needed to address this important issue.

In Chapter 5, the analysis of models developed with GA and BVS strategies suggested that the expression of genes encoding for a number of membrane bound and extra-cellular proteins may be predictive of tumour physiology. In the analysis of other datasets however we have noticed that models may not reflect plausible biological scenarios. For example, models that are predictive of the likelihood of developing metastases in breast cancer appear not to represent solutions that include genes involved in the interaction between extra-cellular matrix components and membrane receptors. Since we know that these processes are of fundamental importance for metastases formation it is reasonable to wonder whether the search strategy used is capturing models that are representative of important biological components of the biological process of interest. We suspect that, because of the limited number of combinations that can be effectively explored with random searches, the scenario described is possible. A possible solution to the problem could be to use biological functional descriptors to "inform" the search. By doing this, other genes or models that could not be obtained in an unbiased search may be obtained. This can be achieved for GA based searches by assigning higher chromosome inclusion weights to genes in predefined categories. As described in Chapter 3, this is easy to do using the functionality that already exists in GALGO. As a result, genes with higher weights are more likely to be included in explored and final models. In the context of a BVS approach this could be achieved in a more principled way using biological knowledge to define a prior probability distribution. Indeed, the GA strategy is being investigated by Russell Compton, a PhD student in our research group. His preliminary results are

so far confirming our expectation that using biological knowledge to drive model search allow the identification of biologically relevant models.

In the context of cell to cell communication, GALGO can be easily adapted to extend the network of previously selected genes. In Chapter 5 several genes were selected that may be involved in arresting tumour expansion. This response should hypothetically be derived from sensing external factors produced by the tumour. In this context, GALGO could be used to explain the expression profile of selected genes by means of other gene profiles using a regression model. This further selection of genes could extend the network in normal cells of genes related to the detection mechanism of tumour signals.

Chapter 3 and 4 have demonstrated that the methodology can be used successfully to identify biomarkers from large scale datasets. Examples have been presented using datasets from expression profiling, NMR metabolomics, and proteomics that collaborators have developed in the context of collaboration with our group. It has been also stressed along the thesis the fact that statistical models can be used for formulating hypotheses on the molecular basis of a biological process. To explore this possibility we have performed an extensive analysis of public domain datasets focussing on the problem of identifying genes involved in the interaction between normal and epithelial cells. In the next few paragraphs discussions of the implications of the findings is provided.

7.2 - The Rational behind Developing a Methodology to Identify Molecular Components Likely to be involved in the Communication Between Adjacent Cell Types

All living organisms sense their surroundings. These can contain nutrients, diffusible factors, and, of course other organisms including pathogens and parasites. Depending on the detected stimulus, cells activate specific pathways whose function is to orchestrate the response of the cell to the changing environment. A cell can integrate different stimuli via interaction between different signalling pathways. Another form of interaction between different components of response is the interaction between different cells in a tissue. Insulin, hormones, cytokines, and neurotransmitters are examples of factors involved in these interactions. To determine genes involved in the response to these factors, experiments are carried out where cells are exposed to controlled titration of these factors and the cell response is analyzed. This methodology has been successfully used for years and has provided most of the cellular knowledge we have today. Recently, microarray technology has expanded this knowledge by providing the response of thousands of genes in single experiments. Microarray technology gives also the opportunity to study the interaction between different cells at the genome level because mRNA from each cell type can be isolated and assayed in specific microarrays. However, there are no data analysis tools that can study this kind of dataset. Therefore, my goal was to propose a framework to study cell crosstalk from functional genomics data.

7.3 - Bioinformatics Approaches to Studying Cell to Cell Interactions

Along this thesis, two novel methods for studying cell crosstalk have been proposed. First, in Chapter 5, a multivariate variable selection approach was designed explaining

features of one cell type (tumour) by means of data from a surrounding cell type (normal epithelial cells). This method is detailed in Chapter 3, which was proven to be useful for other functional genomics data in Chapter 4. Second, in Chapter 6, a method based on the difference of significant correlations for the same gene in both cell types was also proposed. For a gene expressed in one cell type, the number of significant correlations with genes in the opposite cell type are counted. Interesting genes are those whose difference in the number of significant correlations in both cell types is high. In other work, it has been shown in this work that simple differential expression analysis commonly applied to microarray data is also interesting to study cell to cell interaction.

Because the first method provides which genes are predictive, the second selects genes that are correlated to others genes expressed in other cells, and the third gives those genes whose expression has changed, these three methods should be considered complementary.

7.4 - Overall Biological Results

It has been shown that the molecular state of normal cells surrounding the tumour is predictive of tumour physiology and clinical outcome. Predictive models were characterized, mainly, by extra-cellular factors with demonstrated biological activity on tumour cells and by factors associated to cancer predisposition or regulation of oncogenes.

It has been shown also that genes involved in cell communication can be identified from microarray data independently of the technology and tissue. One of the genes selected, Slit-2, has been experimentally validated to be related in killing tumour cells.

7.5 - Discussion and Scope of Methods

Our main purpose was to design computational methods to study cel to cell interactions. The biological results, reviewed in previous section, are interesting and successful for our purposes. These results together with their biological implications are mainly discussed along this thesis. Therefore, in the following paragraphs, the discussion will be focused on some methodological issues of the procedures used to study cell to cell interactions.

7.5.1 - Polarization Hypothesis

The polarization hypothesis described and characterized in Chapter 6 may be important in discovering mechanisms and genes involved in cell communications. For experimental designs equivalent to that detailed in Chapter 6, *pol* can be used "as is", without any modification. However, to use *pol* or a *pol*-like metric in slightly different experimental designs or contexts, parts of the process should be inevitably adapted. In order to discuss changes and adaptations, the overall gene-selection procedure will be summarized as follow:

1. Estimation of the null distribution of correlations.
2. Computation of the observed correlations.
3. Estimation of the number of expected false correlations along null distribution.
4. Selection of a significant threshold.
5. Determination of significant correlations (*f* and *b*).
6. Computation of *pol*.
7. Gene selection based on *pol*.

The entire process is based on a measure of correlation between the expression profiles of any two genes in both cell types. Spearman ranking coefficient was used in order to

detect any monotone correlation. However, any other correlation measure could be used.

Our metric is based on "significant" correlations which impose the task that a significance threshold should be carefully picked. A FDR approach was used to choose this threshold. The threshold value chosen was around 0.75 whose FDR is estimated to 0.000001. In our case, FDR means "the expected number of false correlations", thus the expected number of false correlations is around 1 in 10,000. This selection was conservative considering that highly polarized genes (whose absolute *pol* value is high) is around 200 versus 1 correlations (*pol=(200-1)/(200+1+1)=0.985*). Thus, these genes should not be sensitive to a few false correlations. However, the use of dissimilar significance thresholds could generate that *pol* for some genes may be substantially different. This instability may produce false positives which may diminish the chances

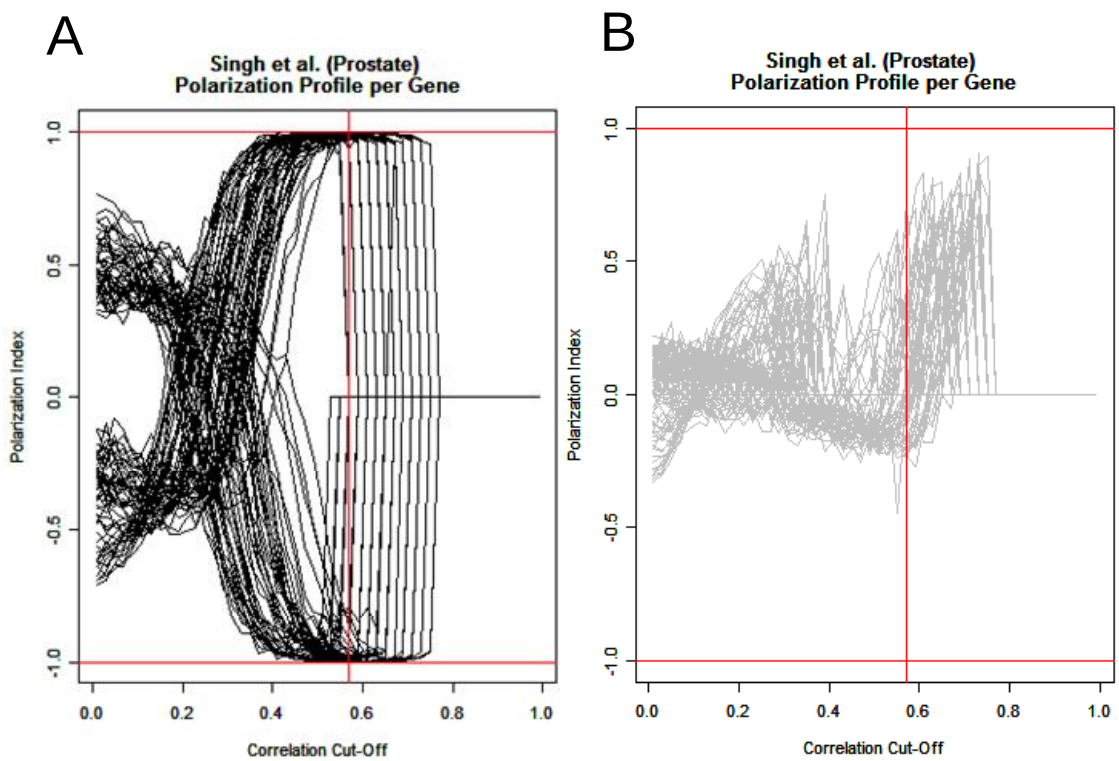


Figure 7.1 – *pol* profile for random selected genes whose *pol* area under the curve is high (panel A) and low (panel B). A vertical line used as reference is shown around 0.57.

to select genes genuinely involved in cell interactions. To avoid this issue, other more stable metrics could be used. For this, a simple strategy could be averaging *pol* measurements for a broad range of significance thresholds (e.g. 0.6, 0.61, 0.62, … 0.80). Another strategy could use, for example, the area under the curve of *pol* as shown in Figure 7.1. Genes with a high area would tend to have high *pol* for a broad range (Figure 7.1A) whereas genes a with low area would tend to have low *pol* (Figure 7.1B).

The threshold to determine "significant" correlations explained above also produce that a huge amount of correlations are removed from the analysis (those smaller than a significance threshold). For example, a threshold of 0.6 would use only 0.05% of the total number of correlations. This is mainly due to the use of a general threshold. To circumvent this issue, a new procedure considering more correlations per gene could be designed. For instance, the distribution of correlations of Slit-2 gene expression could be modelled by three additive Gaussians distributions: one centred close to zero similar to the null hypothesis and the other two for the highest negative and positive correlations respectively (Figure 7.2). The underlying hypothesis is that the observed correlations are a random and noisy sample of true and random correlations. True correlations should be, by far, different from zero whereas random correlations should be around zero. Thus, only correlations within the central Gaussian should be considered as random correlations hence removed from the analysis. Indeed, this is the assumption used in the procedure detailed in Chapter 6. However, in the case just described, two other Gaussian curves has been added whose mean and standard deviation depends on the observed correlations for each gene.

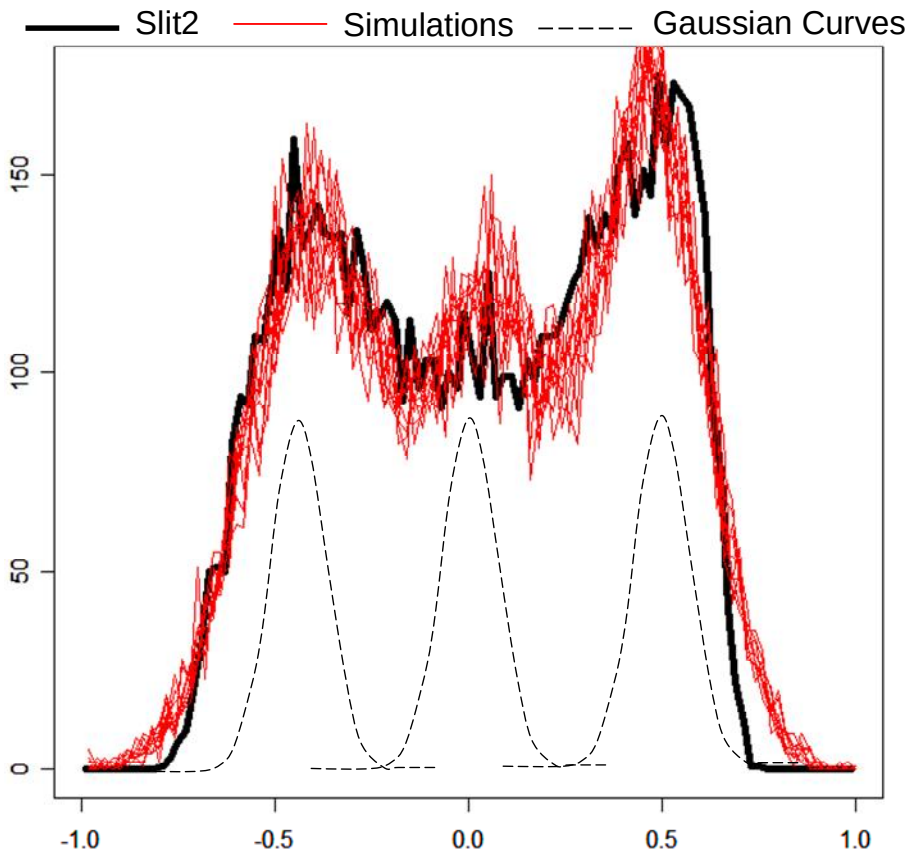


Figure 7.2 – Modelling correlation distributions by three Gaussians distributions. The horizontal axis shows the correlation coefficient whereas vertical axis shows the frequency (in genes) for Slit-2 gene (expressed in normal correlated with genes in tumour). Dotted lines represent three hypothetical Gaussian distributions. Red lines correspond to the distribution of simulations of random values generated by the same number of genes sampled from the three fitted Gaussian distributions.

As an example, the parameters of three Gaussian distributions were estimated using a K-Means algorithm with 3 centres that initially set to -1, 0, and +1. Correlations predicted to pertain to any of the three resulted groups were then used to estimate the mean and standard deviation of the Gaussian distributions. In addition, a 1.25 was used as a scaling factor for the standard deviation to compensate for this kind of biased estimation. Figure 7.2 shows simulations of samples obtained from the three fitted Gaussian curves. The figure demonstrates a good fit of the proposed approach for the observed correlation distribution of the Slit-2 gene. Consequently, this procedure, or a similar procedure may allow choosing a specific correlation threshold for each gene which may allow the usage of much larger total number of correlations decreasing the chance, perhaps, of selecting false positive genes.

7.5.1.1 - Identifying interaction networks in a Host-Pathogen interaction system

The polarization metric presented in Chapter 6 assumes that both interacting cell types are of the same species therefore the genes measured in the two cell types are the same. There are however extremely interesting biological systems where cells from different species interact. One of these systems is the interaction between bacterial pathogens and human host cells. During the course of this thesis, work has been done for the characterization of this cell interaction system [38]. Unfortunately the experimental system has not yet been fully developed to allow the simultaneous analysis of cell interaction. However, it is important to discuss here how gene networks involved in cell to cell interaction in Host-Pathogen interaction can be identified.

Our approach was based on around 40 paired tumour and adjacent normal samples taken from biopsies. This approach was designed, in part, to the fact that obtaining dynamical data from patients is not practical. The use of this experimental design implies that time was not considered. That is, the observed transcriptional state of both cell types is assumed to be the result of continuous interaction with very slow dynamics or in different equilibrium states. This assumption is realistic because tumours may take years to develop and a clinical sample represents a temporary equilibrium between the two cell types. However, the observed dynamic responses in a host-pathogen interaction system are in the order of hours [217]. This has implications in the experimental design. A more appropriate experimental design could be designed by taking a sample from the host and the pathogen just before infection (time = 0) and then at time intervals after infection (time > 0) [217]. In this experimental design, in which our research group has made remarkable progress [217], the first approach would be perhaps the detection of differentially expressed genes in all time intervals in both, the

host [218] and Pathogen. Then, to identify gene-gene connections between cell species, a method dealing with time-course data could be used, for example, Bayesian networks [219], state-space models [220], ordinary differential equations [221], or mutual information [222].

7.5.1.2 - Polarization metric for multiple cell types

In the work presented in Chapter 6, it has been assumed that the interaction is composed of two cell types only. Theoretically, however, more complex interaction systems could be analysed. For instance, in prostate cancer the role of stomal cells should be considered or immune cells in the Host-Pathogen interaction model. In a three cell interaction system where all cell types come from the same species, *pol* can be used in a pair-wise manner. Polarization indexes can still be used in a host-pathogen interaction model if more than one cell type is considered on the host side, however, specifically for the host-pathogen interaction, more appropriate methods mentioned in previous sections should be considered.

REFERENCES

1. Fujimori, F., Gunji, W., Kikuchi, J., Mogi, T., Ohashi, Y., Makino, T., Oyama, A., Okuhara, K., Uchida, T., and Murakami, Y. (2001), 'Crosstalk of prolyl isomerases, Pin1/Ess1, and cyclophilin A', *Biochemical and Biophysical Research Communications,* **289** (1), 181-90.
2. Genoud, T. and Metraux, J. P. (1999), 'Crosstalk in plant cell signaling: structure and function of the genetic network', *Trends Plant Sci.,* **4** (12), 503-07.
3. Nilsson, J. A. and Cleveland, J. L. (2003), 'Myc pathways provoking cell suicide and cancer', *Oncogene,* **22** (56), 9007-21.
4. Ogawa, M., Hanada, A., Yamauchi, Y., Kuwalhara, A., Kamiya, Y., and Yamaguchi, S. (2003), 'Gibberellin biosynthesis and response during Arabidopsis seed germination', *Plant Cell,* **15** (7), 1591-604.
5. Amsterdam, A., Sasson, R., Keren-Tal, I., Aharoni, D., Dantes, A., Rimon, E., Land, A., Cohen, T., Dore, Y., and Hirsh, L. (2003), 'Alternative pathways of ovarian apoptosis: death for life', *Biochemical Pharmacology,* **66** (8), 1355-62.
6. Vodovar, N., Vinals, M., Liehl, P., Basset, A., Degrouard, J., Spellman, P., Boccard, F., and Lemaitre, B. (2005), 'Drosophila host defense after oral infection by an entomopathogenic Pseudomonas species', *Proceedings of the National Academy of Sciences of the United States of America,* **102** (32), 11414-19.
7. Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., and Lander, E. S. (1999), 'Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring', *Science,* **286** (5439), 531-37.
8. Singh, D., Febbo, P. G., Ross, K., Jackson, D. G., Manola, J., Ladd, C., Tamayo, P., Renshaw, A. A., D'Amico, A. V., Richie, J. P., Lander, E. S., Loda, M., Kantoff, P. W., Golub, T. R., and Sellers, W. R. (2002), 'Gene expression correlates of clinical prostate cancer behavior', *Cancer Cell,* **1** (2), 203-09.
9. Lapointe, J., Li, C., Higgins, J. P., van de Rijn, M., Bair, E., Montgomery, K., Ferrari, M., Egevad, L., Rayford, W., Bergerheim, U., Ekman, P., DeMarzo, A. M., Tibshirani, R., Botstein, D., Brown, P. O., Brooks, J. D., and Pollack, J. R. (2004), 'Gene expression profiling identifies clinically relevant subtypes of prostate cancer', *Proceedings of the National Academy of Sciences of the United States of America,* **101** (3), 811-16.
10. Trevino, V., Falciani, F., and Barrera-Saldana, H. A. (2007), 'DNA microarrays: a powerful genomic tool for biomedical and clinical research', *Mol Med*.
11. Westermeier, R. and Naven, T. (2002), *Proteomics in Practice: A Laboratory Manual of Proteome Analysis* (Wiley-VCH Verlag GmbH).
12. Liebler, D. C. (2002), *Introduction to Proteomics: Tools for the New Biology* (Totowa, NJ: Humana Press).

13. Albala, J. S. and Humphery-Smit, I. (2003), *Protein Arrays, Biochips, and Proteomics: The Next Phase of Genomic Discovery* (Besel, Switzerland: Marcel Dekker, Inc.).
14. Vignali, D. (2000), 'Multiplexed particle-based flow cytometric assays', *Journal of Immunological Methods,* **243**, 243-55.
15. Dettmer, K., Aronov, P. A., and Hammock, B. D. (2007), 'Mass spectrometry-based metabolomics', *Mass Spectrometry Reviews,* **26** (1), 51-78.
16. Lindon, J. C., Holmes, E., and Nicholson, J. K. (2001), 'Pattern recognition methods and applications in biomedical magnetic resonance', *Progress in Nuclear Magnetic Resonance Spectroscopy,* **39** (1), 1-40.
17. Barash, Y., Dehan, E., Krupsky, M., Franklin, W., Geraci, M., Friedman, N., and Kaminski, N. (2004), 'Comparative analysis of algorithms for signal quantitation from oligonucleotide microarrays', *Bioinformatics,* **20** (6), 839-46.
18. Seo, J. and Hoffman, E. P. (2006), 'Probe set algorithms: is there a rational best bet?' *Bmc Bioinformatics,* **7**, -.
19. Leung, Y. F. and Cavalieri, D. (2003), 'Fundamentals of cDNA microarray data analysis', *Trends in Genetics,* **19** (11), 649-59.
20. Stekel, D. (2003), *Microarray bioinformatics* (Cambridge; New York: Cambridge University Press) xiv, 263 p.
21. Affymetrix (2002), 'MAS5 Statistical Algorithms Description Document', **Affymetrix**, http://www.affymetrix.com/products/software/index.affx.
22. Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U., and Speed, T. P. (2003), 'Exploration, normalization, and summaries of high density oligonucleotide array probe level data', *Biostatistics,* **4** (2), 249-64.
23. Bolstad, B. M., Irizarry, R. A., Astrand, M., and Speed, T. P. (2003), 'A comparison of normalization methods for high density oligonucleotide array data based on variance and bias', *Bioinformatics,* **19** (2), 185-93.
24. Axon_Instruments 'GenePix® Pro 6.0 Microarray Image Analysis', (Axon Instruments. http://www.axon.com/GN_GenePixSoftware.html.).
25. PerkinElmer 'Quantarray from Packard BioChip Technologies', (Now part Cyclone of PerkinElmer Life company. http://las.perkinelmer.com/).
26. ScanAnalyze. 'SuperArray. ' (SuperArray Bioscience Corporation. http://www.superarray.com/.).
27. Huber, W., Heydebreck, A. v., Sueltmann, H., Poustka, A., and Vingron, M. (2003), 'Parameter estimation for the calibration and variance stabilization of microarray data', *Statistical Applications in Genetics and Molecular Biology,* **2** (1), Article 3.
28. Quackenbush, J. (2002), 'Microarray data normalization and transformation', *Nature Genetics,* **32**, 496-501.

29. Li, C. and Wong, W. H. (2001), 'Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection', *Proceedings of the National Academy of Sciences of the United States of America,* **98** (1), 31-36.
30. Rajagopalan, D. (2003), 'A comparison of statistical methods for analysis of high density oligonucleotide array data', *Bioinformatics,* **19** (12), 1469-76.
31. Tukey, J. W. (1977), *Exploratory data analysis* (Reading, Mass.: Addison-Wesley Pub. Co.) xvi, 688 p.
32. Jain, N., Thatte, J., Braciale, T., Ley, K., O'Connell, M., and Lee, J. K. (2003), 'Local-pooled-error test for identifying differentially expressed genes with a small number of replicated microarrays', *Bioinformatics,* **19** (15), 1945-51.
33. Yoon, D., Yi, S. G., Kim, J. H., and Park, T. (2004), 'Two-stage normalization using background intensities in cDNA microarray data', *Bmc Bioinformatics,* **5**, -.
34. Reiner-Benaim, A. (2007), 'FDR control by the BH procedure for two-sided correlated tests with implications to gene expression data analysis', *Biom J,* **49** (1), 107-26.
35. Bretz, F., Landgrebe, J., and Brunner, E. (2005), 'Multiplicity issues in microarray experiments', *Methods Inf Med,* **44** (3), 431-7.
36. Benjamini, Y. and Hochberg, Y. (1995), 'Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing', *Journal of the Royal Statistical Society Series B,* **57** (1), 289-300.
37. Storey, J. D. (2002), 'A direct approach to false discovery rates', *Journal of the Royal Statistical Society Series B,* **64**, 479-98.
38. Stekel, D. J., Sarti, D., Trevino, V., Zhang, L., Salmon, M., Buckley, C. D., Stevens, M., Pallen, M. J., Penn, C., and Falciani, F. (2005), 'Analysis of host response to bacterial infection using error model based gene expression microarray experiments', *Nucleic Acids Research,* **33** (6), -.
39. Speed, T. P. 'Statistical analysis of gene expression microarray data'.
40. Yue, H., Eastman, P. S., Wang, B. B., Minor, J., Doctolero, M. H., Nuttall, R. L., Stack, R., Becker, J. W., Montgomery, J. R., Vainer, M., and Johnston, R. (2001), 'An evaluation of the performance of cDNA microarrays for detecting changes in global mRNA expression', *Nucleic Acids Res,* **29** (8), E41-1.
41. Mutch, D. M., Berger, A., Mansourian, R., Rytz, A., and Roberts, M. A. (2001), 'Microarray data analysis: a practical approach for selecting differentially expressed genes', *Genome Biol,* **2** (12), PREPRINT0009.
42. Kim, S. Y., Lee, J. W., and Sohn, I. S. (2006), 'Comparison of various statistical methods for identifying differential gene expression in replicated microarray data', *Statistical Methods in Medical Research,* **15** (1), 3-20.
43. Tusher, V. G., Tibshirani, R., and Chu, G. (2001), 'Significance analysis of microarrays applied to the ionizing radiation response', *Proc Natl Acad Sci U S A. Jan 20,* **98** (9), 5116-21.

44. Smyth, G. K. (2004), 'Linear models and empirical bayes methods for assessing differential expression in microarray experiments', *Stat Appl Genet Mol Biol,* **3**, Article3.
45. Broberg, P. (2002), 'Ranking genes with respect to differential expression', *Genome Biol,* **3** (9), preprint0007.
46. Zhao, Y. and Pan, W. (2003), 'Modified nonparametric approaches to detecting differentially expressed genes in replicated microarray experiments', *Bioinformatics,* **19** (9), 1046-54.
47. Baldi, P. and Long, A. D. (2001), 'A Bayesian framework for the analysis of microarray expression data: regularized t -test and statistical inferences of gene changes', *Bioinformatics,* **17** (6), 509-19.
48. Pan, W., Lin, J., and Le, C. T. (2003), 'A mixture model approach to detecting differentially expressed genes with microarray data', *Funct Integr Genomics,* **3** (3), 117-24.
49. Tomlins, S. A., Rhodes, D. R., Perner, S., Dhanasekaran, S. M., Mehra, R., Sun, X. W., Varambally, S., Cao, X. H., Tchinda, J., Kuefer, R., Lee, C., Montie, J. E., Shah, R. B., Pienta, K. J., Rubin, M. A., and Chinnaiyan, A. M. (2005), 'Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer', *Science,* **310** (5748), 644-48.
50. Tibshirani, R. and Hastie, T. (2007), 'Outlier sums for differential gene expression analysis', *Biostatistics,* **8** (1), 2-8.
51. Getz, G., Levine, E., and Domany, E. (2000), 'Coupled two-way clustering analysis of gene microarray data', *Proceedings of the National Academy of Sciences of the United States of America,* **97** (22), 12079-84.
52. Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D., and Levine, A. J. (1999), '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,* **96** (12), 6745-50.
53. Azuaje, F. and Dopazo, J. (2005), *Data analysis and visualization in genomics and proteomics* (Hoboken, NJ: John Wiley) xv, 267 p.
54. Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. (1998), 'Cluster analysis and display of genome-wide expression patterns', *Proceedings of the National Academy of Sciences of the United States of America,* **95** (25), 14863-68.
55. Johnson, D. E. (1998), *Applied multivariate methods for data analysts* (Pacific Grove, Calif.: Duxbury) xiv, 567 p.
56. Berrar, D. P., Dubitzky, W., and Granzow, M. (2003), *A practical approach to microarray data analysis* (Boston: Kluwer Academic Publishers) xv, 368 p.
57. Hubert, M. and Engelen, S. (2004), 'Robust PCA and classification in biosciences', *Bioinformatics,* **20** (11), 1728-36.

58. Bicciato, S., Luchini, A., and Di Bello, C. (2003), 'PCA disjoint models for multiclass cancer analysis using gene expression data', *Bioinformatics,* **19** (5), 571-78.
59. Hsu, A. L., Tang, S. L., and Halgamuge, S. K. (2003), 'An unsupervised hierarchical dynamic self-organizing approach to cancer class discovery and marker gene identification in microarray data', *Bioinformatics,* **19** (16), 2131-40.
60. Vaquerizas, J. M., Conde, L., Yankilevich, P., Cabezon, A., Minguez, P., Diaz-Uriarte, R., Al-Shahrour, F., Herrero, J., and Dopazo, J. (2005), 'GEPAS, an experiment-oriented pipeline for the analysis of microarray gene expression data', *Nucleic Acids Research,* **33**, W616-W20.
61. Saeed, A. I., Sharov, V., White, J., Li, J., Liang, W., Bhagabati, N., Braisted, J., Klapa, M., Currier, T., Thiagarajan, M., Sturn, A., Snuffin, M., Rezantsev, A., Popov, D., Ryltsov, A., Kostukovich, E., Borisovsky, I., Liu, Z., Vinsavich, A., Trush, V., and Quackenbush, J. (2003), 'TM4: A free, open-source system for microarray data management and analysis', *Biotechniques,* **34** (2), 374-8.
62. Grewal, A. and Conway, A. (2000), 'Tools for Analyzing Microarray Expression Data', *Journal of Lab Automation,* **5** (5), 62-64.
63. Sturn, A., Quackenbush, J., and Trajanoski, Z. (2002), 'Genesis: cluster analysis of microarray data', *Bioinformatics,* **18** (1), 207-8.
64. Simon, R., Radmacher, M. D., Dobbin, K., and McShane, L. M. (2003), 'Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification', *Journal of the National Cancer Institute,* **95** (1), 14-18.
65. Simon, R. (2003), 'Diagnostic and prognostic prediction using gene expression profiles in high-dimensional microarray data', *British Journal of Cancer,* **89** (9), 1599-604.
66. Theodoridis, S. and Koutroumbas, K. (1999), *Pattern recognition* (San Diego: Academic Press) xiv, 625 p.
67. Sá, J. P. M. d. (2001), *Pattern recognition: concepts, methods, and applications* (Berlin; New York: Springer) xix, 318 p.
68. Gordon, A. D. (1999), *Classification* (2nd edn., Monographs on statistics and applied probability; 82; Boca Raton: Chapman & Hall/CRC) x, 256 p.
69. Duda, R. O., Hart, P. E., and Stork, D. G. (2001), *Pattern classification* (2nd edn.; New York: Wiley) xx, 654 p.
70. Fukunaga, K. (1990), *Introduction to statistical pattern recognition* (2nd Ed. edn., Computer science and scientific computing.; Boston: Academic Press) xiii, 591 p.
71. Webb, A. R. (2002), *Statistical pattern recognition* (2a edn.; West Sussex, England; New Jersey: Wiley) xviii, 496 p.
72. Qiu, P., Wang, Z. J., and Liu, K. J. R. (2005), 'Ensemble dependence model for classification and prediction of cancer and normal gene expression data', *Bioinformatics,* **21** (14), 3114-21.

73. Tibshirani, R., Hastie, T., Narasimhan, B., and Chu, G. (2002), 'Diagnosis of multiple cancer types by shrunken centroids of gene expression', *Proc Natl Acad Sci U S A. Jan 20,* **99** (10), 6567-72.
74. Li, L. P., Weinberg, C. R., Darden, T. A., and Pedersen, L. G. (2001), 'Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the GA/KNN method', *Bioinformatics,* **17** (12), 1131-42.
75. Breiman, L. (2001), 'Random Forest. January 2001'.
76. Diaz-Uriarte, R. and Alvarez de Andres, S. (2006), 'Gene selection and classification of microarray data using random forest', *BMC Bioinformatics,* **7**, 3.
77. Lattin, J. M., Carroll, J. D., Green, P. E., and Green, P. E. (2003), *Analyzing multivariate data* (Duxbury applied series.; Pacific Grove, CA: Thomson Brooks/Cole) xxiv, 556 p.
78. Tabachnick, B. G. and Fidell, L. S. (2001), *Using multivariate statistics* (4th edn.; Boston, MA: Allyn and Bacon) xxvi, 966 p.
79. Ooi, C. H. and Tan, P. (2003), 'Genetic algorithms applied to multi-class prediction for the analysis of gene expression data', *Bioinformatics,* **19** (1), 37-44.
80. Mcculloch, W. S. and Pitts, W. (1990), 'A Logical Calculus of the Ideas Immanent in Nervous Activity (Reprinted from Bulletin of Mathematical Biophysics, Vol 5, Pg 115-133, 1943)', *Bulletin of Mathematical Biology,* **52** (1-2), 99-115.
81. Rumelhart, D., Hinton, G., and Williams, R. (1986), 'Learning Internal Representations by Error Propagation.' in M. JL (ed.), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition.* (1: MIT Press).
82. Wang, Z. Y., Wang, Y., Xuan, J. H., Dong, Y. B., Bakay, M., Feng, Y. J., Clarke, R., and Hoffman, E. P. (2006), 'Optimized multilayer perceptrons for molecular classification and diagnosis using genomic data', *Bioinformatics,* **22** (6), 755-61.
83. Smola, A. J. (2000), *Advances in large margin classifiers* (Cambridge, Mass.: MIT Press) vi, 412 p.
84. Komura, D., Nakamura, H., Tsutsumi, S., Aburatani, H., and Ihara, S. (2005), 'Multidimensional support vector machines for visualization of gene expression data', *Bioinformatics,* **21** (4), 439-44.
85. Dudoit, S., Fridlyand, J., and Speed, T. P. (2002), 'Comparison of discrimination methods for the classification of tumors using gene expression data', *Journal of the American Statistical Association,* **97** (457), 77-87.
86. Efron, B. and Tibshirani, R. (1993), *An introduction to the bootstrap* (Monographs on statistics and applied probability; 57.; New York: Chapman & Hall) xvi, 436 p.
87. Ye, N. (2003), *The handbook of data mining* (Human factors and ergonomics; Mahwah, N.J.; London: Lawrence Erlbaum Assoc.) xxx, 689 p.
88. Sima, C., Braga-Neto, U., and Dougherty, E. R. (2005), 'Superior feature-set ranking for small samples using bolstered error estimation', *Bioinformatics,* **21** (7), 1046-54.

89. Wessels, L. F. A., Marcel J.T. Reinders, Augustinus A. M. Hart, Cor J. Veenman, Hongyue Dai, Yudong D. He, and Veer, L. J. v. t. (2005), 'A protocol for building and evaluating predictors of disease state based on microarray data.' *Bioinformatics*.
90. Antonov, A. V., Tetko, I. V., Mader, M. T., Budczies, J., and Mewes, H. W. (2004), 'Optimization models for cancer classification: extracting gene interaction information from microarray expression data', *Bioinformatics,* **20** (5), 644-U145.
91. Cox, D. R. (1972), 'Regression Models and Life-Tables', *Journal of the Royal Statistical Society Series B-Statistical Methodology,* **34** (2), 187-&.
92. Sha, N., Tadesse, M. G., and Vannucci, M. (2006), 'Bayesian variable selection for the analysis of microarray data with censored outcomes', *Bioinformatics,* **22** (18), 2262-8.
93. Montaner, D., Tarraga, J., Huerta-Cepas, J., Burguet, J., Vaquerizas, J. M., Conde, L., Minguez, P., Vera, J., Mukherjee, S., Valls, J., Pujana, M. A., Alloza, E., Herrero, J., Al-Shahrour, F., and Dopazo, J. (2006), 'Next station in microarray data analysis: GEPAS', *Nucleic Acids Res,* **34** (Web Server issue), W486-91.
94. Subramani, P., Sahu, R., and Verma, S. (2006), 'Feature selection using Haar wavelet power spectrum', *Bmc Bioinformatics,* **7**, -.
95. Lio, P. (2003), 'Wavelets in bioinformatics and computational biology: state of art and perspectives', *Bioinformatics,* **19** (1), 2-9.
96. Hardle, W., Kerkyacharian, G., Picard, D., and Tsybakov, A. (2000), *Wavelets, Approximation, and Statistical Applications (Lecture Notes in Statistics)* (Lecture Notes in Statistics: Springer-Verlag New York).
97. Fryzlewicz, P. (2003), 'Wavelet Techniques for Time Series and Poisson Data', (University of Bristol UK).
98. Jeffery, I. B., Higgins, D. G., and Culhane, A. C. (2006), 'Comparison and evaluation of methods for generating differentially expressed gene lists from microarray data', *BMC Bioinformatics,* **7**, 359.
99. Fröhlich H (2002), 'Feature Selection for Support Vector Machines by Means of Genetic Algorithms', (Max-Planck-Institute).
100. Ho, S. Y., Hsieh, C. H., Chen, H. M., and Huang, H. L. (2006), 'Interpretable gene expression classifier with an accurate and compact fuzzy rule base for microarray data analysis', *Biosystems,* **85** (3), 165-76.
101. Paul, T. and Iba, H. 'Linear and Combinatorial Optimizations by Estimation of Distribution Algorithms', <http://www.iba.k.u-tokyo.ac.jp/english/EDA.htm>.
102. Holland, J. H. (1975), *Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence* (Ann Arbor: University of Michigan Press) viii, 183 p.
103. Goldberg, D. E. (1989), *Genetic algorithms in search, optimization, and machine learning* (Reading, Mass.: Addison-Wesley Pub. Co.) xiii, 412 p.

104. Xiao, Y. H., Frisina, R., Gordon, A., Klebanov, L., and Yakovlev, A. (2004), 'Multivariate search for differentially expressed gene combinations', *Bmc Bioinformatics,* **5**, -.
105. Lu, Y., Liu, P. Y., Xiao, P., and Deng, H. W. (2005), 'Hotelling's T-2 multivariate profiling for detecting differential expression in microarrays', *Bioinformatics,* **21** (14), 3105-13.
106. Grate, L. R. (2005), 'Many accurate small-discriminatory feature subsets exist in microarray transcript data: biomarker discovery', *Bmc Bioinformatics,* **6**, -.
107. Choudhary, A., Brun, M., Hua, J. P., Lowey, J., Suh, E., and Dougherty, E. R. (2006), 'Genetic test bed for feature selection', *Bioinformatics,* **22** (7), 837-42.
108. Li, J. Y., Liu, H. Q., Downing, J. R., Yeoh, A. E. J., and Wong, L. S. (2003), 'Simple rules underlying gene expression profiles of more than six subtypes of acute lymphoblastic leukemia (ALL) patients', *Bioinformatics,* **19** (1), 71-78.
109. Paul, T. K. and Iba, H. (2005), 'Gene selection for classification of cancers using probabilistic model building genetic algorithm', *Biosystems,* **82** (3), 208-25.
110. Sha, N. J., Vannucci, M., Tadesse, M. G., Brown, P. J., Dragoni, I., Davies, N., Roberts, T. R. C., Contestabile, A., Salmon, M., Buckley, C., and Falciani, F. (2004), 'Bayesian variable selection in multinomial probit models to identify molecular signatures of disease stage', *Biometrics,* **60** (3), 812-19.
111. Gamerman, D. (1997), *Markov chain Monte Carlo: stochastic simulation for Bayesian inference* (1a edn.; New York: Chapman & Hall) xiii, 245 p.
112. Hastings, W. (1970), 'Carlo sampling methods using Markov chains and their applications', *Biometrika,* **57**, 97-109.
113. O'Neill, M. C. and Song, L. (2003), 'Neural network analysis of lymphoma microarray data: prognosis and diagnosis near-perfect', *Bmc Bioinformatics,* **4**, -.
114. Bicciato, S., Pandin, M., Didone, G., and Di Bello, C. (2001), 'Analysis of an associative memory neural network for pattern identification in gene expression dataSIGKDD01, Conference).' *BIOKDD01: Workshop on Data Mining in Bioinformatics (with SIGKDD01, Conference)*.
115. Gunther, E. C., Stone, D. J., Gerwien, R. W., Bento, P., and Heyes, M. P. (2003), 'Prediction of clinical drug efficacy by classification of drug-induced genomic expression profiles in vitro', *Proceedings of the National Academy of Sciences of the United States of America,* **100** (16), 9608-13.
116. Alter, O., Brown, P. O., and Botstein, D. (2000), 'Singular value decomposition for genome-wide expression data processing and modeling', *Proceedings of the National Academy of Sciences of the United States of America,* **97** (18), 10101-06.
117. Hyvarinen, A., Karhunen, J., and Oja, E. (2001), *Independent component analysis* (New York: J. Wiley) xxi, 481 p.
118. Saidi, S. A., Holland, C. M., Kreil, D. P., MacKay, D. J. C., Charnock-Jones, D. S., Print, C. G., and Smith, S. K. (2004), 'Independent component analysis of microarray data in the study of endometrial cancer', *Oncogene,* **23** (39), 6677-83.

119. Huang, D. S. and Zheng, C. H. (2006), 'Independent component analysis-based penalized discriminant method for tumor classification using gene expression data', *Bioinformatics,* **22** (15), 1855-62.
120. Lee, S. I. and Batzoglou, S. (2003), 'Application of independent component analysis to microarrays', *Genome Biology,* **4** (11), -.
121. Frigyesi, A., Veerla, S., Lindgren, D., and Hoglund, M. (2006), 'Independent component analysis reveals new and biologically significant structures in micro array data', *Bmc Bioinformatics,* **7**, -.
122. Guyon, I., Weston, J., Barnhill, S., and Vapnik, V. (2002), 'Gene selection for cancer classification using support vector machines', *Machine Learning,* **46** (1-3), 389-422.
123. Jong K, Marchiori E, Sebagy M, and A., v. d. V. 'Feature Selection in Proteomic Pattern Data with Support Vector Machines', *IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology.*
124. Liu, C. C., Chen, W. S. E., Lin, C. C., Liu, H. C., Chen, H. Y., Yang, P. C., Chang, P. C., and Chen, J. J. W. (2006), 'Topology-based cancer classification and related pathway mining using microarray data', *Nucleic Acids Research,* **34** (14), 4069-80.
125. Alizadeh, A. A., Eisen, M. B., Davis, R. E., Ma, C., Lossos, I. S., Rosenwald, A., Boldrick, J. G., Sabet, H., Tran, T., Yu, X., Powell, J. I., Yang, L. M., Marti, G. E., Moore, T., Hudson, J., Lu, L. S., Lewis, D. B., Tibshirani, R., Sherlock, G., Chan, W. C., Greiner, T. C., Weisenburger, D. D., Armitage, J. O., Warnke, R., Levy, R., Wilson, W., Grever, M. R., Byrd, J. C., Botstein, D., Brown, P. O., and Staudt, L. M. (2000), 'Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling', *Nature,* **403** (6769), 503-11.
126. Ein-Dor, L., Kela, I., Getz, G., Givol, D., and Domany, E. (2005), 'Outcome signature genes in breast cancer: is there a unique set?' *Bioinformatics,* **21** (2), 171-78.
127. Keller, A. D., Schummer, M., Hood, L., and Ruzzo, W. L. (2000), 'Bayesian Classification of DNA Array Expression Data.
' *Technical Report UW-CSE-(2000)-08-01.* (Department of Computer Science and Engineering, University of Washington, Seattle.).
128. Hwang, D., Rust, A. G., Ramsey, S., Smith, J. J., Leslie, D. M., Weston, A. D., de Atauri, P., Aitchison, J. D., Hood, L., Siegel, A. F., and Bolouri, H. (2005), 'A data integration methodology for systems biology', *Proc Natl Acad Sci U S A,* **102** (48), 17296-301.
129. Hwang, D., Smith, J. J., Leslie, D. M., Weston, A. D., Rust, A. G., Ramsey, S., de Atauri, P., Siegel, A. F., Bolouri, H., Aitchison, J. D., and Hood, L. (2005), 'A data integration methodology for systems biology: experimental verification', *Proc Natl Acad Sci U S A,* **102** (48), 17302-7.

130. Gevaert, O., De Smet, F., Timmerman, D., Moreau, Y., and De Moor, B. (2006), 'Predicting the prognosis of breast cancer by integrating clinical and microarray data with Bayesian networks', *Bioinformatics,* **22** (14), e184-90.
131. Dudoit, S. and Fridlyand, J. (2003), 'Classification in Microarray Experiments', in T. P. Speed (ed.), *Statistical Analysis of Gene Expression Microarray Data* (Boca Raton, Florida: Chapman & Hall/CRC), 93-158.
132. Lee, J. W., Lee, J. B., Park, M., and Song, S. H. (2005), 'An extensive comparison of recent classification tools applied to microarray data', *Computational Statistics & Data Analysis,* **48** (4), 869-85.
133. Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y. C., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li, C., Maechler, M., Rossini, A. J., Sawitzki, G., Smith, C., Smyth, G., Tierney, L., Yang, J. Y. H., and Zhang, J. H. (2004), 'Bioconductor: open software development for computational biology and bioinformatics', *Genome Biology,* **5** (10), R80.
134. Yeoh, E. J., Ross, M. E., Shurtleff, S. A., Williams, W. K., Patel, D., Mahfouz, R., Behm, F. G., Raimondi, S. C., Relling, M. V., Patel, A., Cheng, C., Campana, D., Wilkins, D., Zhou, X. D., Li, J. Y., Liu, H. Q., Pui, C. H., Evans, W. E., Naeve, C., Wong, L. S., and Downing, J. R. (2002), 'Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling', *Cancer Cell,* **1** (2), 133-43.
135. van't Veer, L. J., Dai, H. Y., van de Vijver, M. J., He, Y. D. D., Hart, A. A. M., Mao, M., Peterse, H. L., van der Kooy, K., Marton, M. J., Witteveen, A. T., Schreiber, G. J., Kerkhoven, R. M., Roberts, C., Linsley, P. S., Bernards, R., and Friend, S. H. (2002), 'Gene expression profiling predicts clinical outcome of breast cancer', *Nature,* **415** (6871), 530-36.
136. Raza, K., Scheel-Toellner, D., Lee, C. Y., Pilling, D., Curnow, S. J., Falciani, F., Trevino, V., Kumar, K., Assi, L. K., Lord, J. M., Gordon, C., Buckley, C. D., and Salmon, M. (2006), 'Synovial fluid leukocyte apoptosis is inhibited in patients with very early rheumatoid arthritis', *Arthritis Research & Therapy,* **8** (4), -.
137. Rosenbaum, J. T., Deodhar, A., Suhler, E. B., and Smith, J. R. (2004), 'How do you know?' *British Journal of Ophthalmology,* **88** (8), 980-81.
138. Wallace, G. R., Farmer, I., Church, A., Graham, E. M., and Stanford, M. R. (2003), 'Serum levels of chemokines correlate with disease activity in patients with retinal vasculitis', *Immunology Letters,* **90** (1), 59-64.
139. Wallace, G. R., Curnow, S. J., Wloka, K., Salmon, M., and Murray, P. I. (2004), 'The role of chemokines and their receptors in ocular disease', *Progress in Retinal and Eye Research,* **23** (4), 435-48.
140. Chen, Y., Vaughan, R. W., Kondeatis, E., Fortune, F., Graham, E. M., Stanford, M. R., and Wallace, G. R. (2004), 'Chemokine gene polymorphisms associate with gender in patients with uveitis', *Tissue Antigens,* **63** (1), 41-45.

141. Kotter, I., Koch, S., Vonthein, R., Ruckwaldt, U., Amberger, M., Gunaydin, I., Zierhut, M., and Stubiger, N. (2005), 'Cytokines, cytokine antagonists and soluble adhesion molecules in patients with ocular Behcet's disease treated with human recombinant interferon-alpha 2a. Results of an open study and review of the literature', *Clinical and Experimental Rheumatology,* **23** (4), S20-S26.
142. Witmer, A. N., Vrensen, G. F. J. M., Van Noorden, C. J. F., and Schlingemann, R. O. (2003), 'Vascular endothelial growth factors and angiogenesis in eye disease', *Progress in Retinal and Eye Research,* **22** (1), 1-29.
143. Bas, S., Genevay, S., Meyer, O., and Gabay, C. (2003), 'Anti-cyclic citrullinated peptide antibodies, IgM and IgA rheumatoid factors in the diagnosis and prognosis of rheumatoid arthritis', *Rheumatology,* **42** (5), 677-80.
144. Nishimura, K., Sugiyama, D., Kogata, Y., Tsuji, G., Nakazawa, T., Kawano, S., Saigo, K., Morinobu, A., Koshiba, M., Kuntz, K. M., Kamae, I., and Kumagai, S. (2007), 'Meta-analysis: Diagnostic accuracy of anti-cyclic citrullinated peptide antibody and rheumatoid factor for rheumatoid arthritis', *Annals of Internal Medicine,* **146** (11), 797-808.
145. Young, S. P., Nessim, M., Falciani, F., Trevino, V., Siviraj, R. R., Banerjee, S. P., Savant, V., Levine, B. A., Scott, R. A., Murray, P. I., and Wallace, G. R. (2007), 'Metabolomic analysis of vitreous humour from patients with vitreoretinal disease: a new research tool?' *Manuscript in Preparation*.
146. Trevino, V. and Falciani, F. (2006), 'GALGO: an R package for multivariate variable selection using genetic algorithms', *Bioinformatics,* **22** (9), 1154-56.
147. Raza, K., Lee, C. Y., Pilling, D., Heaton, S., Situnayake, R. D., Carruthers, D. M., Buckley, C. D., Gordon, C., and Salmon, M. (2003), 'Ultrasound guidance allows accurate needle placement and aspiration from small joints in patients with early inflammatory arthritis', *Rheumatology,* **42** (8), 976-79.
148. Arnett, F. C., Edworthy, S. M., Bloch, D. A., McShane, D. J., Fries, J. F., Cooper, N. S., Healey, L. A., Kaplan, S. R., Liang, M. H., Luthra, H. S., and et al. (1988), 'The American Rheumatism Association 1987 revised criteria for the classification of rheumatoid arthritis', *Arthritis Rheum,* **31** (3), 315-24.
149. Quackenbush, J. (2006), 'Microarray analysis and tumor classification', *N Engl J Med,* **354** (23), 2463-72.
150. Alberti, C. (2006), 'Prostate cancer progression and surrounding microenvironment', *Int J Biol Markers,* **21** (2), 88-95.
151. Rollins, B. J. (2006), 'Inflammatory chemokines in cancer growth and progression', *Eur J Cancer,* **42** (6), 760-7.
152. Al-Shahrour, F., Minguez, P., Tarraga, J., Montaner, D., Alloza, E., Vaquerizas, J. M., Conde, L., Blaschke, C., Vera, J., and Dopazo, J. (2006), 'BABELOMICS: a systems biology perspective in the functional annotation of genome-scale experiments', *Nucleic Acids Res,* **34** (Web Server issue), W472-6.

153. Antoniou, A. C. and Easton, D. F. (2006), 'Models of genetic susceptibility to breast cancer', *Oncogene,* **25** (43), 5898-905.
154. Shand, R. L. and Gelmann, E. P. (2006), 'Molecular biology of prostate-cancer pathogenesis', *Curr Opin Urol,* **16** (3), 123-31.
155. Shipp, M. A., Vijayaraghavan, J., Schmidt, E. V., Masteller, E. L., D'Adamio, L., Hersh, L. B., and Reinherz, E. L. (1989), 'Common acute lymphoblastic leukemia antigen (CALLA) is active neutral endopeptidase 24.11 ("enkephalinase"): direct evidence by cDNA transfection analysis', *Proc Natl Acad Sci U S A,* **86** (1), 297-301.
156. Osman, I., Yee, H., Taneja, S. S., Levinson, B., Zeleniuch-Jacquotte, A., Chang, C., Nobert, C., and Nanus, D. M. (2004), 'Neutral endopeptidase protein expression and prognosis in localized prostate cancer', *Clin Cancer Res,* **10** (12 Pt 1), 4096-100.
157. Redner, A., Melamed, M. R., and Andreeff, M. (1986), 'Detection of central nervous system relapse in acute leukemia by multiparameter flow cytometry of DNA, RNA, and CALLA', *Ann N Y Acad Sci,* **468**, 241-55.
158. Dawson, L. A., Maitland, N. J., Turner, A. J., and Usmani, B. A. (2004), 'Stromal-epithelial interactions influence prostate cancer cell invasion by altering the balance of metallopeptidase expression', *Br J Cancer,* **90** (8), 1577-82.
159. Zhang, X., Wei, H., Wang, H., and Tian, Z. (2006), 'Involvement of interaction between Fractalkine and CX3CR1 in cytotoxicity of natural killer cells against tumor cells', *Oncol Rep,* **15** (2), 485-8.
160. Limesand, K. H., Barzen, K. A., Quissell, D. O., and Anderson, S. M. (2003), 'Synergistic suppression of apoptosis in salivary acinar cells by IGF1 and EGF', *Cell Death Differ,* **10** (3), 345-55.
161. Harman, S. M., Metter, E. J., Blackman, M. R., Landis, P. K., and Carter, H. B. (2000), 'Serum levels of insulin-like growth factor I (IGF-I), IGF-II, IGF-binding protein-3, and prostate-specific antigen as predictors of clinical prostate cancer', *J Clin Endocrinol Metab,* **85** (11), 4258-65.
162. Sakamoto, S., Yokoyama, M., Zhang, X., Prakash, K., Nagao, K., Hatanaka, T., Getzenberg, R. H., and Kakehi, Y. (2004), 'Increased expression of CYR61, an extracellular matrix signaling protein, in human benign prostatic hyperplasia and its regulation by lysophosphatidic acid', *Endocrinology,* **145** (6), 2929-40.
163. Kaplan, S. A. (2005), 'Induction and function of CYR61 (CCN1) in prostatic stromal and epithelial cells: CYR61 is required for prostatic cell proliferation', *J Urol,* **174** (3), 1012.
164. Sakamoto, S., Yokoyama, M., Aoki, M., Suzuki, K., Kakehi, Y., and Saito, Y. (2004), 'Induction and function of CYR61 (CCN1) in prostatic stromal and epithelial cells: CYR61 is required for prostatic cell proliferation', *Prostate,* **61** (4), 305-17.
165. Pilarsky, C. P., Schmidt, U., Eissrich, C., Stade, J., Froschermaier, S. E., Haase, M., Faller, G., Kirchner, T. W., and Wirth, M. P. (1998), 'Expression of the

extracellular matrix signaling molecule Cyr61 is downregulated in prostate cancer', *Prostate,* **36** (2), 85-91.
166. Brenner, V., Nyakatura, G., Rosenthal, A., and Platzer, M. (1997), 'Genomic organization of two novel genes on human Xq28: compact head to head arrangement of IDH gamma and TRAP delta is conserved in rat and mouse', *Genomics,* **44** (1), 8-14.
167. Dong, J. T. (2006), 'Prevalent mutations in prostate cancer', *J Cell Biochem,* **97** (3), 433-47.
168. Ellwood-Yen, K., Graeber, T. G., Wongvipat, J., Iruela-Arispe, M. L., Zhang, J., Matusik, R., Thomas, G. V., and Sawyers, C. L. (2003), 'Myc-driven murine prostate cancer shares molecular features with human prostate tumors', *Cancer Cell,* **4** (3), 223-38.
169. Bernard, D., Pourtier-Manzanedo, A., Gil, J., and Beach, D. H. (2003), 'Myc confers androgen-independent prostate cancer cell growth', *J Clin Invest,* **112** (11), 1724-31.
170. Lin, X., Asgari, K., Putzi, M. J., Gage, W. R., Yu, X., Cornblatt, B. S., Kumar, A., Piantadosi, S., DeWeese, T. L., De Marzo, A. M., and Nelson, W. G. (2001), 'Reversal of GSTP1 CpG island hypermethylation and reactivation of pi-class glutathione S-transferase (GSTP1) expression in human prostate cancer cells by treatment with procainamide', *Cancer Res,* **61** (24), 8611-6.
171. Hochberg, Y. and Benjamini, Y. (1990), 'More powerful procedures for multiple significance testing', *Stat Med,* **9** (7), 811-8.
172. Maere, S., Heymans, K., and Kuiper, M. (2005), 'BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks', *Bioinformatics,* **21** (16), 3448-9.
173. Margolin, A. A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Dalla Favera, R., and Califano, A. (2006), 'ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context', *BMC Bioinformatics,* **7 Suppl 1**, S7.
174. Butte, A. J. and Kohane, I. S. (2000), 'Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements', *Pac Symp Biocomput*, 418-29.
175. Tu, Y., Stolovitzky, G., and Klein, U. (2002), 'Quantitative noise analysis for gene expression microarray experiments', *Proceedings of the National Academy of Sciences of the United States of America,* **99** (22), 14031-36.
176. Klebanov, L. and Yakovlev, A. (2007), 'How high is the level of technical noise in microarray data?' *Biol Direct,* **2**, 9.
177. Issa, J. P. (2004), 'CpG island methylator phenotype in cancer', *Nat Rev Cancer,* **4** (12), 988-93.
178. Laird, P. W. and Jaenisch, R. (1996), 'The role of DNA methylation in cancer genetic and epigenetics', *Annu Rev Genet,* **30**, 441-64.

179. Li, E., Beard, C., and Jaenisch, R. (1993), 'Role for DNA Methylation in Genomic Imprinting', *Nature,* **366** (6453), 362-65.
180. Zeschnigk, M., Tschentscher, F., Lich, C., Brandt, B., Horsthemke, B., and Lohmann, D. R. (2003), 'Methylation analysis of several tumour suppressor genes shows a low frequency of methylation of CDKN2A and RARB in uveal melanomas', *Comparative and Functional Genomics,* **4** (3), 329-36.
181. Hayslip, J. and Montero, A. (2006), 'Tumor suppressor gene methylation in follicular lymphoma: a comprehensive review', *Mol Cancer,* **5**, 44.
182. Esteller, M. (2007), 'Cancer epigenomics: DNA methylomes and histone-modification maps', *Nat Rev Genet,* **8** (4), 286-98.
183. Agrawal, S., Unterberg, M., Koschmieder, S., zur Stadt, U., Brunnberg, U., Verbeek, W., Buchner, T., Berdel, W. E., Serve, H., and Muller-Tidow, C. (2007), 'DNA methylation of tumor suppressor genes in clinical remission predicts the relapse risk in acute myeloid leukemia', *Cancer Res,* **67** (3), 1370-7.
184. Wang, Y. P., Yu, Q. J., Cho, A. H., Rondeau, G., Welsh, J., Adamson, E., Mercola, D., and McClelland, M. (2005), 'Survey of differentially methylated promoters in prostate cancer cell lines', *Neoplasia,* **7** (8), 748-60.
185. Yu, Y. P., Paranjpe, S., Nelson, J., Finkelstein, S., Ren, B., Kokkinakis, D., Michalopoulos, G., and Luo, J. H. (2005), 'High throughput screening of methylation status of genes in prostate cancer using an oligonucleotide methylation array', *Carcinogenesis,* **26** (2), 471-79.
186. Lodygin, D., Epanchintsev, A., Menssen, A., Diebold, J., and Hermeking, H. (2005), 'Functional epigenomics identifies genes frequently silenced in prostate cancer', *Cancer Res,* **65** (10), 4218-27.
187. Ingenuity-Pathways-Analysis *Ingenuity(c) Systems*, www.ingenuity.com.
188. Paccaud, J. P., Reith, W., Carpentier, J. L., Ravazzola, M., Amherdt, M., Schekman, R., and Orci, L. (1996), 'Cloning and functional characterization of mammalian homologues of the COPII component Sec23', *Mol Biol Cell,* **7** (10), 1535-46.
189. Xu, W. and Stamnes, M. (2006), 'The actin-depolymerizing factor homology and charged/helical domains of drebrin and mAbp1 direct membrane binding and localization via distinct interactions with actin', *J Biol Chem,* **281** (17), 11826-33.
190. Peitsch, W. K., Grund, C., Kuhn, C., Schnolzer, M., Spring, H., Schmelz, M., and Franke, W. W. (1999), 'Drebrin is a widespread actin-associating protein enriched at junctional plaques, defining a specific microfilament anchorage system in polar epithelial cells', *Eur J Cell Biol,* **78** (11), 767-78.
191. Petit, M. M., Meulemans, S. M., and Van de Ven, W. J. (2003), 'The focal adhesion and nuclear targeting capacity of the LIM-containing lipoma-preferred partner (LPP) protein', *J Biol Chem,* **278** (4), 2157-68.
192. Petit, M. M., Meulemans, S. M., Alen, P., Ayoubi, T. A., Jansen, E., and Van de Ven, W. J. (2005), 'The tumor suppressor Scrib interacts with the zyxin-related

protein LPP, which shuttles between cell adhesion sites and the nucleus', *BMC Cell Biol,* **6** (1), 1.
193. Fernandis, A. Z. and Ganju, R. K. (2001), 'Slit: a roadblock for chemotaxis', *Sci STKE,* **2001** (91), PE1.
194. Liu, D., Hou, J., Hu, X., Wang, X., Xiao, Y., Mou, Y., and De Leon, H. (2006), 'Neuronal chemorepellent Slit2 inhibits vascular smooth muscle cell migration by suppressing small GTPase Rac1 activation', *Circ Res,* **98** (4), 480-9.
195. Niclou, S. P., Jia, L., and Raper, J. A. (2000), 'Slit2 is a repellent for retinal ganglion cell axons', *J Neurosci,* **20** (13), 4962-74.
196. Wu, L., Aster, J. C., Blacklow, S. C., Lake, R., Artavanis-Tsakonas, S., and Griffin, J. D. (2000), 'MAML1, a human homologue of Drosophila mastermind, is a transcriptional co-activator for NOTCH receptors', *Nat Genet,* **26** (4), 484-9.
197. Nam, Y., Sliz, P., Song, L., Aster, J. C., and Blacklow, S. C. (2006), 'Structural basis for cooperativity in recruitment of MAML coactivators to Notch transcription complexes', *Cell,* **124** (5), 973-83.
198. Suzuki, A., Yamanaka, T., Hirose, T., Manabe, N., Mizuno, K., Shimizu, M., Akimoto, K., Izumi, Y., Ohnishi, T., and Ohno, S. (2001), 'Atypical protein kinase C is involved in the evolutionarily conserved par protein complex and plays a critical role in establishing epithelia-specific junctional structures', *J Cell Biol,* **152** (6), 1183-96.
199. Dallol, A., Krex, D., Hesson, L., Eng, C., Maher, E. R., and Latif, F. (2003), 'Frequent epigenetic inactivation of the SLIT2 gene in gliomas', *Oncogene,* **22** (29), 4611-6.
200. Dallol, A., Da Silva, N. F., Viacava, P., Minna, J. D., Bieche, I., Maher, E. R., and Latif, F. (2002), 'SLIT2, a human homologue of the Drosophila Slit2 gene, has tumor suppressor activity and is frequently inactivated in lung and breast cancers', *Cancer Res,* **62** (20), 5874-80.
201. Dallol, A., Forgacs, E., Martinez, A., Sekido, Y., Walker, R., Kishida, T., Rabbitts, P., Maher, E. R., Minna, J. D., and Latif, F. (2002), 'Tumour specific promoter region methylation of the human homologue of the Drosophila Roundabout gene DUTT1 (ROBO1) in human cancers', *Oncogene,* **21** (19), 3020-8.
202. Morris, M. R., Hesson, L. B., Wagner, K. J., Morgan, N. V., Astuti, D., Lees, R. D., Cooper, W. N., Lee, J., Gentle, D., Macdonald, F., Kishida, T., Grundy, R., Yao, M., Latif, F., and Maher, E. R. (2003), 'Multigene methylation analysis of Wilms' tumour and adult renal cell carcinoma', *Oncogene,* **22** (43), 6794-801.
203. Narayan, G., Goparaju, C., Arias-Pulido, H., Kaufmann, A. M., Schneider, A., Durst, M., Mansukhani, M., Pothuri, B., and Murty, V. V. (2006), 'Promoter hypermethylation-mediated inactivation of multiple Slit-Robo pathway genes in cervical cancer progression', *Mol Cancer,* **5**, 16.
204. Yu, Y. P., Landsittel, D., Jing, L., Nelson, J., Ren, B. G., Liu, L. J., McDonald, C., Thomas, R., Dhir, R., Finkelstein, S., Michalopoulos, G., Becich, M., and Luo, J. H.

(2004), 'Gene expression alterations in prostate cancer predicting tumor aggression and preceding development of malignancy', *Journal of Clinical Oncology,* **22** (14), 2790-99.

205. Chen, X., Cheung, S. T., So, S., Fan, S. T., Barry, C., Higgins, J., Lai, K. M., Ji, J. F., Dudoit, S., Ng, I. O. L., van de Rijn, M., Botstein, D., and Brown, P. O. (2002), 'Gene expression patterns in human liver cancers', *Molecular Biology of the Cell,* **13** (6), 1929-39.
206. Notterman, D. A., Alon, U., Sierk, A. J., and Levine, A. J. (2001), 'Transcriptional gene expression profiles of colorectal adenoma, adenocarcinoma, and normal tissue examined by oligonucleotide arrays', *Cancer Research,* **61** (7), 3124-30.
207. Boer, J. M., Huber, W. K., Sultmann, H., Wilmer, F., von Heydebreck, A., Haas, S., Korn, B., Gunawan, B., Vente, A., Fuzesi, L., Vingron, M., and Poustka, A. (2001), 'Identification and classification of differentially expressed genes in renal cell carcinoma by expression profiling on a global human 31,500-element cDNA array', *Genome Research,* **11** (11), 1861-70.
208. Genz, A. (1992), 'Numerical computation of multivariate normal probabilities', *J. Comput. Graph. Statist,* (1), 141-50.
209. McKusick-Nathans (2006), 'Online Mendelian Inheritance in Man, OMIM (TM)', *Institute of Genetic Medicine, Johns Hopkins University (Baltimore, MD) and National Center for Biotechnology Information, National Library of Medicine (Bethesda, MD). URL: http://www.ncbi.nlm.nih.gov/omim/*.
210. Rebhan, M., Chalifa-Caspi, V., Prilusky, J., and Lancet, D. (2006), 'GeneCards: encyclopedia for genes, proteins and diseases', *Weizmann Institute of Science, Bioinformatics Unit and Genome Center (Rehovot, Israel), 1997. URL: http://www.genecards.org/*.
211. Eaton, A. D. (2006), 'HubMed: a web-based biomedical literature search interface', *Nucleic Acids Res,* **34** (Web Server issue), W745-7.
212. Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. (2000), 'Gene ontology: tool for the unification of biology. The Gene Ontology Consortium', *Nat Genet,* **25** (1), 25-9.
213. Jenssen, T. K., Laegreid, A., Komorowski, J., and Hovig, E. (2001), 'A literature network of human genes for high-throughput analysis of gene expression', *Nat Genet,* **28** (1), 21-8.
214. Wain, H. M., Lush, M. J., Ducluzeau, F., Khodiyar, V. K., and Povey, S. (2004), 'Genew: the Human Gene Nomenclature Database, 2004 updates', *Nucleic Acids Res,* **32** (Database issue), D255-7.
215. Fernandez, J. M., Hoffmann, R., and Valencia, A. (2007), 'iHOP web services', *Nucleic Acids Res*.

216. Akagi, T., Yin, D., Kawamata, N., Bartram, C. R., Hofmann, W. K., Wolf, I., Miller, C. W., and Koeffler, H. P. (2006), 'Methylation analysis of asparagine synthetase gene in acute lymphoblastic leukemia cells', *Leukemia,* **20** (7), 1303-06.
217. Sarti, D. (2005), 'Host-Pathogen Interaction in a E.Coli Model of Infection. PhD Thesis', (University of Birmingham).
218. Carzaniga*, T., Sarti*, D., Trevino*, V., Buckley, C., Salmon, M., Wild, D., Deho', G., and Falciani, F. (2007), 'The analysis of Cellular Transcriptional Response at the Genome Level: Two Case Studies with Relevance in Bacterial Pathogenesis', in F. Falciani (ed.), *Microarrays Technology Through Applications (ISBN: 0415378532)* (TAYLOR & FRANCIS).
219. Hartemink, A. J., Gifford, D. K., Jaakkola, T. S., and Young, R. A. (2001), 'Using graphical models and genomic expression data to statistically validate models of genetic regulatory networks', *Pac Symp Biocomput*, 422-33.
220. Rangel, C., Angus, J., Ghahramani, Z., Lioumi, M., Sotheran, E., Gaiba, A., Wild, D. L., and Falciani, F. (2004), 'Modeling T-cell activation using gene expression profiling and state-space models', *Bioinformatics,* **20** (9), 1361-72.
221. Guthke, R., Moller, U., Hoffmann, M., Thies, F., and Topfer, S. (2005), 'Dynamic network reconstruction from gene expression data applied to immune response during bacterial infection', *Bioinformatics,* **21** (8), 1626-34.
222. Margolin, A. A., Wang, K., Lim, W. K., Kustagi, M., Nemenman, I., and Califano, A. (2006), 'Reverse engineering cellular networks', *Nat Protoc,* **1** (2), 662-71.