Gene-set reduction for analysis of major and minor Gleason scores based on differential gene expressions of biological pathways in prostate cancer

By

Surya Prasad Poudel

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

Master of Science

In

Epidemiology

School of Public Health Sciences

University of Alberta

Abstract

Introduction: Prostate cancer is a heterogeneous disease, and in spite of recent advances regarding understanding its biology, further discovery of the molecular events underlying prostate cancer is still needed. Gleason grading is an important predictor of prostate cancer outcomes. In current practices, patients with a total GS ≥7 are at greater risk but it is still unclear how prostate cancer outcomes differ for various distributions of the total GS between its major and minor components.

Objectives: Our goal is to identify genes and biological pathways differentiating between patients with various combinations of GS, while moving from a less aggressive combination (3,3) to a more aggressive combination (4,4).

Methods: The Swedish Watchful Waiting Cohort (n=255) consisting of mRNA expression of 6,100 genes in prostate tumor tissue has been used. Significance Analysis of Microarray for Gene Sets (SAM-GS) has been used to screen gene sets from C2 catalog of Molecular Signature Database (MSigDB) to identify those sets differentiating between patients who died from prostate cancer during follow-up (lethal prostate cancer) versus patients who survived at least 10 years after diagnosis (indolent prostate cancer). Those pathways not associated with both major and minor GS ≤ 3 versus both major and minor GS ≥ 4, based on SAM-GS method, has been discarded. Moving from a less aggressive GS combination of (3,3) to a more aggressive one of (4,4) via grey areas of (3,4) and (4,3), the reduced gene sets to their core subsets of genes contributing most to the association with the GS combinations has been obtained by using Significance Analysis of Gene Sets Reduction (SAM-GSR) method. Finally, these results to the gene sets and cores differentiating between GS of (3,4) vs (4,3) were compared.

Results: 1351 gene sets out of 1,892 MSigDB gene sets were found to be differentially expressed between 149 lethal and 106 indolent prostate cancer patients, using SAM-GS. Furthermore, 1,246 gene sets were found to be differentially expressed between 80 patients with major and minor GS ≤ 3 versus 68 patients with major and minor GS ≥ 4. SAM-GSR achieved a 91% reduction, averaged over the four GS combinations, starting from (3,3) and ending with (4,4). The numbers of significant gene sets and core set sizes decrease considerably when comparing patients with larger total GS, indicating a challenge in discriminating between higher risk groups of patients. Eight gene sets are differentially expressed between GS of (3,4) vs (4,4), and only one gene set differentiates between (4,3) and (4,4). At the gene level, none of the 13 core genes from comparing (3,4) vs (4,4) are represented among the 332 core genes comparing (3,3) vs (3,4), or among the 323 core genes comparing (3,3) vs (4,3). The set consisting of the 13 genes shows a marginal association with GS of (3,4) vs (4,3), with a SAM-GS p-value of 0.059.

Conclusions and Implications: Our comprehensive analysis of combinations of major and minor Gleason scores brings additional insights to the current practice based on the sum of the two components, especially for values of the total GS of 7 or 8, indicating patients at greater risk. Further studies are needed to validate our results at the gene and pathway levels.

Preface

This dissertation is original, unpublished, independent work by the author, Surya Poudel, under the supervision of committee members Drs. Irina Dinu, Sentil Senthilselvan and Saumyadipta Pyne. References to previous work are provided.

Drs. Saumyadipta Pyne and Irina Dinu identified the research project. Surya Poudel was responsible for the study design and data analyses, as well as a first draft of the thesis. Dr. Irina Dinu provided guidance to the data analyses. The final draft of the thesis was critically reviewed and approved by all committee members.

Dedication

This thesis is dedicated to my late father and family members. I am grateful to my family for their continuous support and encouragement. I remember their love, affection and inspiration throughout my life.

Acknowledgements

First and foremost, I want to thanks my supervisor Dr. Irina Dinu for her warm encouragement and thoughtful direction on this research project. It is a great opportunity for me to work with her. Without her supervision and suggestions, this project cannot go far. Her assistance and guidance helped me improve my analysis and writing skills. I learned a lot from working with her. In future, it is my cherished desired to contribute to science in the way that she is contributing right now.

I also want to thank to the committee members, Dr. Pyne and Dr. Senthilselvan who reviewed this thesis and provided critical feedback which helped me a lot to reshape the draft and present the arguments and thoughts in a coherent way to facilitate the understanding of the study findings.

Thanks to the faculties of the School of Public Health. I learned the core concepts of epidemiology and biostatics from the class lectures, and built confidence applying these principles in my thesis project.

My deepest thanks to my lovely parents for their love and encouragement. I am grateful to my father, Tika Ram Poudel and mother, Dil Maya Poudel for their inspiration and support all my life.

Completing my MSc thesis work would have never been a success without the support of my dear Wife, Goma Acharya. I was continually amazed by her sincere support and strong criticism throughout the study period.

Table of Contents

List of Table

List of Figures

List of Abbreviation

DRE	Digital Rectal Examination
GS	Gleason Score
NB	Needle Biopsy
PSA	Prostate Specific Antigen
SAM	Significance Analysis of Microarrays
SAM-GS	Significance Analysis of Microarrays for Gene Sets
SAM-GSR	Significance Analysis of Microarrays for Gene Sets Reduction

Chapter 1

Introduction

Prostate cancer has a high incidence as well as a high mortality, which makes it an important worldwide health issue; in fact it is the most commonly diagnosed male malignancy and second leading cause of cancer-related death in men worldwide[1,2]. Its etiology is complex, including many risk factors such as age, hormonal status, ethnic origin and family history of prostate cancer[3,4]. One in 7 men will develop prostate cancer during his life time and one in 27 will die of it in Canada[5].The death occurs even 20 years after diagnosis[6]. It is becoming an enormous health care burden and its early diagnosis is crucial for successful treatments, which will ultimately prolong and improve the quality of life in men[7], and reduce health care expenses[8]. Over 90% of prostate cancer cases are curable if detected and treated their earliest stage[9].

In current practice, the use of serum Prostate Specific Antigen (PSA) levels and prostate biopsy has increased the early detection rate of prostate cancer[10,11]. The primary factors guiding the treatment are the Gleason Score (GS) of needle biopsy (NB) specimens, serum PSA levels and the Digital Rectal Examination (DRE) findings[12,13]. The GS is the most important factor among these three[13] . The problem in GS of NB is under-grading, but it has gradually decreased since the early 1990s[12]. Another problem with GS is discrepancy on grading by lab technician. In Gleason Grading, the sample cells are taken from each side of prostate gland during the biopsy and then examined under a microscope by pathologist to determine whether cancer cells are present, and to evaluate the microscopic features of any cancer found. A Gleason Grade of 1 to 5 with decreasing differentiation is given to the prostate cancer based upon the microscopic appearance of cancer cells in the prostate gland. Gleason score is calculated as the sum of the major (primary) and minor (secondary) components, ranging from 2 to 10. Higher Gleason scores are more aggressive and have a worse prognosis. It has been long recognized that patients with a total GS $\geq$7 are at greater risk for prostate cancer outcomes[14]. Although this finding has influenced clinical practice, it is still unclear how prostate cancer outcomes differ for various distributions of the total GS between its major and minor components. For example, it has been recognized in the literature that within the GS of 7 patients, there are differences in outcomes between the patients with a combination of a

major Gleason Grade 3 and minor Gleason Grade 4 [GS (3,4)]; and patients with a major Gleason Grade 4 and a minor Gleason Grade 3 [GS (4,3)], with the former category exhibiting better outcomes; moreover, GS 7 (3,4) resembles closer to GS 6 (3,3) whereas GS 7 (4,3) resembles closely to GS 8 (4,4)[15]. Our goal is to identify genes and biological pathways differentiating between patients with various combinations of major and minor GS, while moving from a less aggressive combination (3,3) to a more aggressive combination (4,4).

This thesis consists of five chapters. The background on molecular biology relevant to the thesis topic is explained in Chapter 2. This chapter describes concepts of gene and gene expression, DNA and Microarray Technology used in genetic studies, as well as Gleason grading and Gleason Scores. We state the objectives and provide a description of the data. Important statistical challenges in analysis of microarray data are explained in Chapter 3. This chapter describes the adjustment procedures for multiple comparisons, and explains the methods used for gene set analysis and reduction in microarray studies. Chapter 4, we present the results of our analyses. Chapter 5 describes interpretations of our results, discussion, and conclusion.

Chapter 2

Background

2.1 Overview of Gleason score

2. 1.1 Prostate Gland

The prostate gland is a part of urinary system and reproductive system in men[16]. It is covered by a thin but firm fibrous capsule and separated from it by plexus of veins. It surrounds the urethra and is located just below the urinary bladder and in front of the rectum through which it may be distinctly felt, especially when enlarged[17]. The prostate is about the size of chestnut and has somewhat conical shape. It measures about 4 cm transversely at the base, 2 cm in its antero-posterior diameter, and 3 cm in its vertical diameter. Its weight is about 15- 20 grams for a male in his mid-20s.

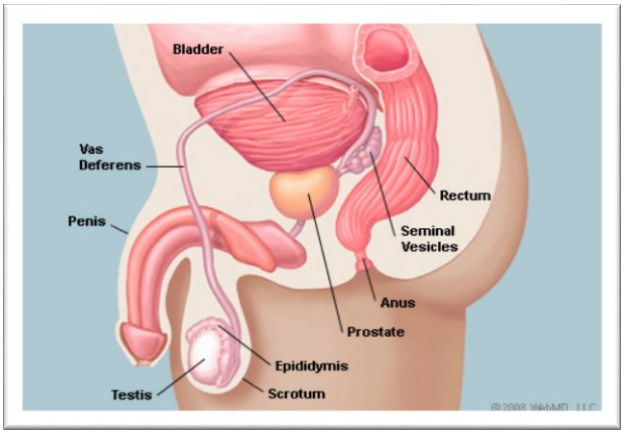


Figure 2.1: Side View of Prostate Gland

Urethra passes through its center; letting flow of urine from the urinary bladder to penis. The prostate secrets fluids that nourishes and protects sperm[18].

2.1.2 Gleason grading and Gleason score

Gleason grading system has proved to be a robust and durable method for the grading of prostate carcinoma[19]. The Gleason Grade of 1 to 5 with decreasing differentiation is given to the prostate

cancer based upon the microscopic appearance of cancer cells in prostate gland. This tumor grading system was developed by Dr. Donald Gleason[20].

Gleason grade and Gleason score, courtesy of Department of Urology, Stanford University, School of Medicine[21].

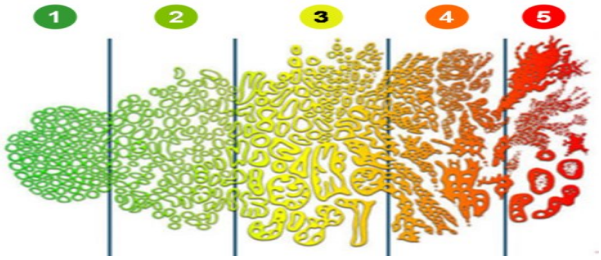


Figure 2.2: An updated version of Dr.Gleason's simplified drawing of the five Gleason grades of prostate cancer.

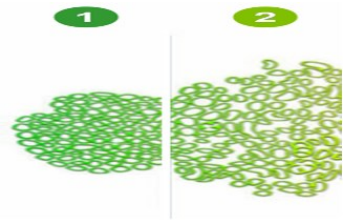


Figure 2.3: Grade 1 (left) and grade 2 (right) prostate adenocarcinoma. Both have pale cells and well formed, separate glands with lumens. Grade 1 is more compact (less invasive) than grade 2.

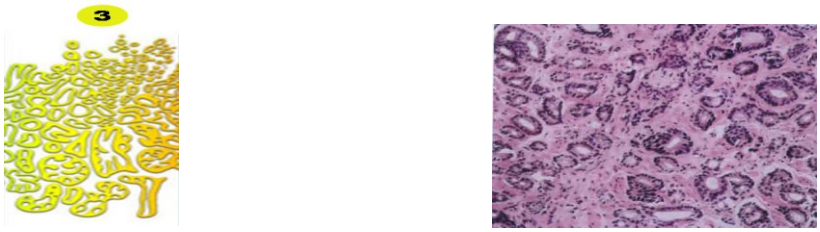


Figure 2.4: Grade 3 carcinoma with individual glands arranged randomly (invading), seen at low magnification.

Grade 3 carcinoma (same as shown in Figure 2.3) showing the usual single layer of cells around each lumen and showing almost all glands separated by muscle (stroma), seen at higher magnification.

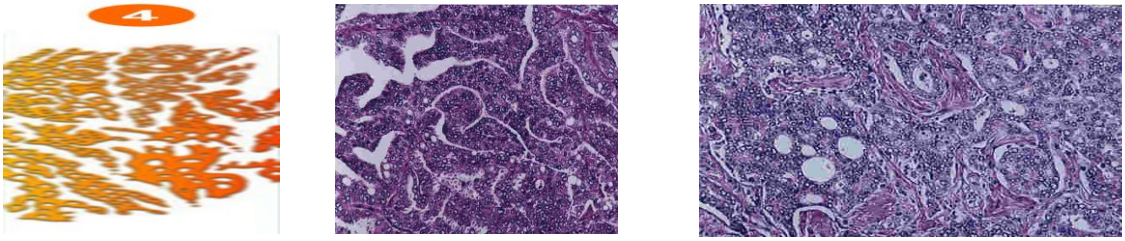


Figure 2.5: Grade 4 carcinomas with two different architectural patterns, each of which has lost the expression of complete "gland units," seen at higher magnification. There are sheets of cells randomly scattered

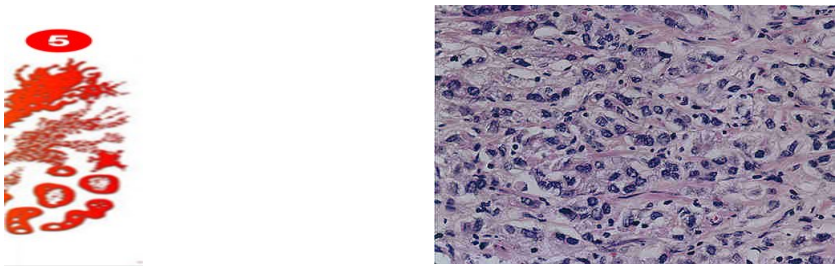


Figure 2.6: Grade 5 adenocarcinoma, consisting of sheets of cells whose lack of pattern in nuclear arrangement indicates total loss of architecture, seen at higher magnification.

A description of different Gleason Grades according to their tissues structure[22] is given below:

Figure 2.2 provides classical photomicrograph examples of Gleason grade of prostate cancers in needle core biopsy tissue sections.

Grade 1: The cancerous tissue will closely resemble the normal prostate tissues. They are identified after a type of surgery called a transurethral resection of prostate.

Grade 2: The cancerous tissue still has well advanced structures, such as the glands. However, they are much larger and tissues are present amongst them. They are also identified after a type of surgery called a transurethral resection of prostate.

Grade 3: The tissue still has the recognizable normal gland units. However, the cells are dimmer. This is the lowest Gleason grade identified by a prostate biopsy core.

Grade 4: The tissue has hardly any identifiable glands. It looks like branches of a large tree, reaching many directions from trunk.

Grade 5: There are no identifiable glands in the tissue. It is the indication of poor prognosis with no evidence of any attempt to form gland unit.

A pathologist examines the biopsy specimen and attempts to give a score to the two patterns. The primary grade, represents the majority of tumor (has to be greater than 50% of the total pattern seen). The secondary grade relates to the minority of the tumor (has to be less than 50%, but at least 5%, of the pattern of the total cancer observed)[23].

These scores are then added to obtain the final Gleason Score. The Gleason score is obtained from the sum of the major (primary) and minor (secondary) pattern. The Gleason score ranges from 2 to 10. Higher Gleason scores are more aggressive and have a worse prognosis. Gleason Score of 7 is more common during diagnosis and it is more important to understand its differential patterns for early treatment of disease[14,20].

It has been long recognized that patients with a total GS ≥7 are at greater risk for prostate cancer outcomes[24]. Although this finding has influenced clinical practice, it is still unclear how prostate cancer outcomes differ for various distributions of the total GS between its major and minor components. The Gleason Differential (GD) gives us the breakdown of the relative proportions or amount of Major Gleason and Minor Gleason. Moreover, GS 7 (4,3) has more advanced clinical and pathological stages, larger tumor volumes, higher preoperative PSA levels, older age and a higher proportion compared to GS 7(3,4) patients[14,25].

2.2 DNA and Gene Expression

2.2.1 DNA

DNA, or deoxyribonucleic acid, is a hereditary material in humans and almost all other organisms[26]. Most of the DNA are located in the nucleolus of cell (called nuclear DNA) and small amount of DNA can also be found in the mitochondria, called mitochondrial DNA or m(DNA). Human DNA consists of about 3 billion bases, and more than 99 percent of those bases are the

same in all the people[27]. Each nucleotide is composed of a nitrogen-containing nucleobase either guanine (G), adenine (A), thymine (T), or cytosine (C) as well as a monosaccharide sugar called deoxyribose and a phosphate group. The nucleotides are joined to one another in a chain by covalent bonds between the sugar of one nucleotide and the phosphate of the next, resulting in an alternating sugar-phosphate backbone. According to base pairing rules (A with T, and C with G), hydrogen bonds bind the nitrogenous bases of the two separate polynucleotide strands to make double-stranded DNA[28].

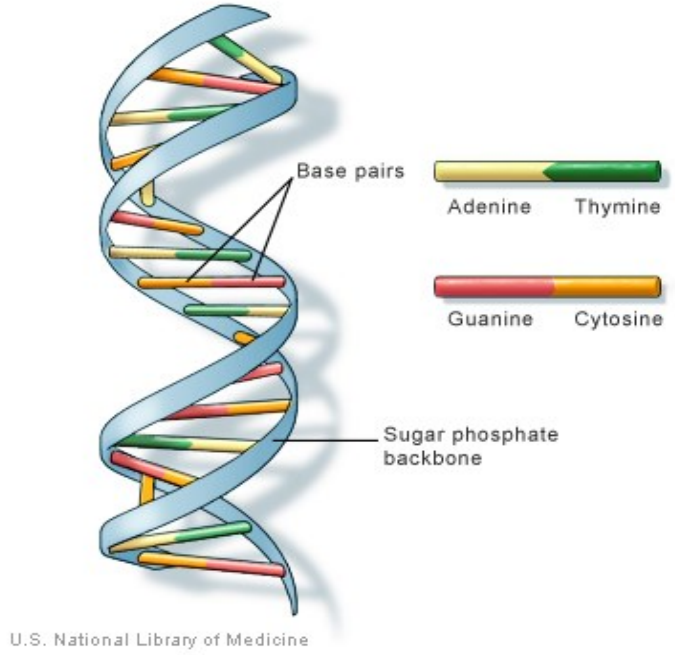


Figure 2.7: DNA Structure

DNA is a double helix formed by base pairs attached to a sugar-phosphate backbone[28]

2.2.2 Gene expression

Gene expression is the process where information from a gene is used in the synthesis of a functional gene product. These products are often proteins, but in non-protein coding genes such as transfer RNA (tRNA) or small nuclear RNA (snRNA) genes, the product is a functional RNA[29,30] .

Several steps in the gene expression process may be modulated, including thetranscription, RNA splicing, translation, and post-translational modification of a protein[29]. Gene regulation gives the cell control over structure and function, and is the basis for cellular differentiation, morphogenesis and the versatility and adaptability of any organism. Gene regulation may also serve as a substrate for evolutionary change, since control of the timing, location, and amount of gene expression can have a profound effect on the functions (actions) of the gene in a cell or in a multicellular organism.

In genetics, gene expression is the most fundamental level at which the genotype gives rise to the phenotype, i.e. observable trait. The genetic code stored in DNA is "interpreted" by gene expression, and the properties of the expression give rise to the organism's phenotype[30].

This means a phenotype (i.e. disease) and gene expression are correlated and the gene expression is useful to understand disease and their treatments. Gene expression plays an important role in scientific research[31].

Microarray technology

The development of microarray technology has been phenomenal in the past few years and, it has become a standard tool in many genomics research laboratories[32]. This technology has been used to understand various biological processes by allowing simultaneous study of the gene expressions of tens of thousands of genes at once [32,33]. It was first published in mid 1990s to monitor the expression of many genes in parallel[34]. The microarray technology has the potential to elucidate the molecular changes that occur in disease states. This method describes the gene expression DNA microarray as high-throughput 'dot-bolt' systems, where targets are fluorescently labeled, free floating amplified RNA or complementary DNA (cDNA) species originated from the samples[35]. Microarray technology involves placing of thousands of gene sequences from samples on a gene chip.The genes sequences on the chip produce light which is used to identify genes that are expressed in that sample[36,37].

Similarly, SNPs could be genotyped by a SNP chip, which is a type of DNA microarray designed to identify genetic variants associated with a phenotype of interest[38,39]. There are thousands of

probes attached on the surface of a chip, which represent the nucleotide sequences of the single-stranded DNA chain[40,41]. Although there are different microarray chips for SNP genotyping using different technologies, the "pair rule" still applies. During genotyping, the DNA samples are separated into two single stranded fragments and both are labelled by fluorescent substance[41,42]. Then, the DNA fragments will be attached onto the microchip and hybridized with the synthetic sequences on the chip following the "pair rule". After hybridization, specialized computer equipment is used to measure the fluorescent signal intensity contained in each probe[20]. Genotypes for the alleles of a locus can be inferred from the fluorescent signals.

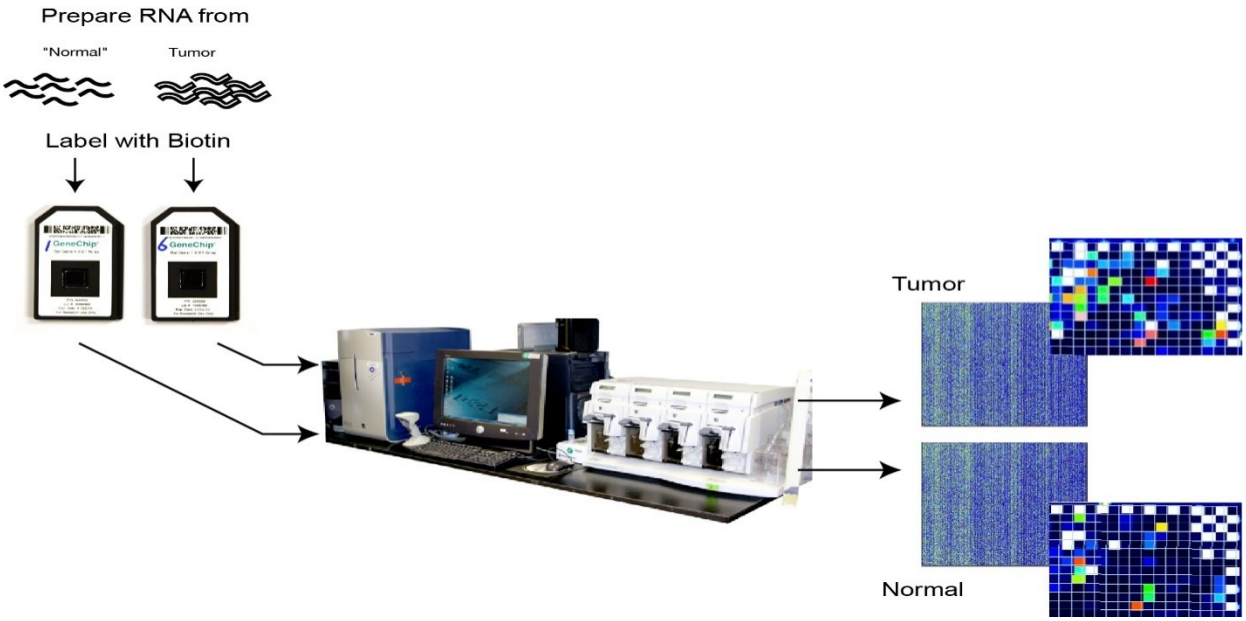


Figure 2.8: Microarray Technology.

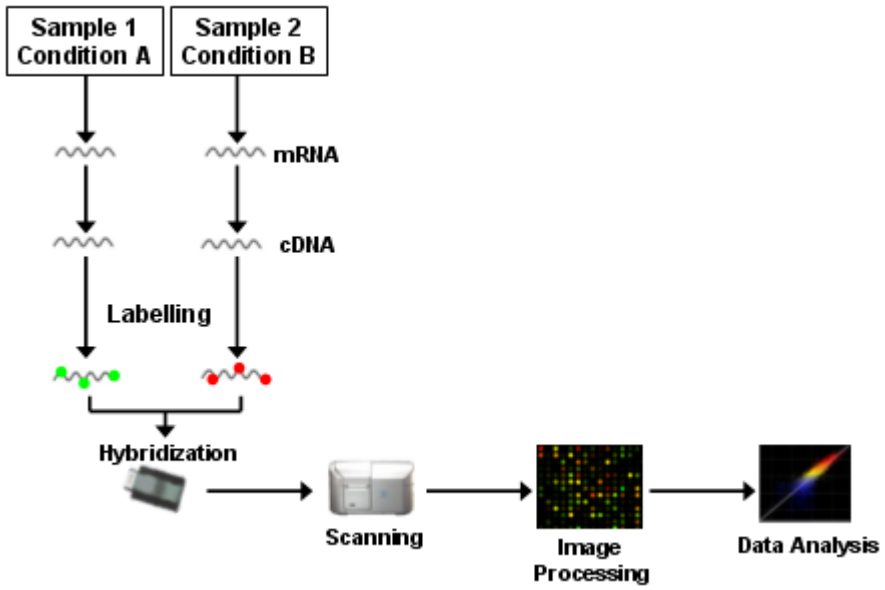


Figure 2.9: Microarray Experiment Workflow.

Red spots on the array correspond to genes found only in sample 2, green ones to genes found only in sample 1, and yellow spots to genes found in both samples[42].

2.2.3 Microarray for gene expression

There are three major applications of DNA microarrays: finding differences in expression levels between predefined groups of samples[43], class prediction[44], and analyzing a given set of gene expression profiles with the goal of discovering subgroups that share common features[45].

Gene expression is measured from the amount of complementary DNA (cDNA) to the mRNA[46]. cDNA is the product of reverse transcription which copies mRNA into DNA, pairing the RNA bases (A,T,G and C) to their corresponding DNA counterparts (T,A,C and G)[47].

Researchers collect the mRNA first to measure gene expression under certain set of circumstances. Then reverse transcriptases enzymes would generate a complementary DNA (cDNA) to the mRNA. Researchers can label and quantify the mRNA with fluorescent nucleotides attached to the cDNA[47,48]. If a gene is highly active, it produces more mRNA and more corresponding cDNA, than genes that are less active. Based on the pair rule, the fluorescent labelled cDNAs which

represent the mRNA of the gene, will match to their synthetic complementary DNAs on the microarray chip[48,49]. Researchers can use a special scanner to measure the fluorescent intensity for each gene and quantify their level of expression.

2.2.4 Biological pathways

Analyzing microarray data at an individual gene level usually leads to a list of many "significant" genes, even after multiple comparison adjustments have been made[45]. The process of trying to interpret such a large list of genes is difficult. Moreover, replication of the findings in different microarray experiments is another serious challenge with such individual gene level analysis. Molecular biologists have put together lists of genes grouped by function, such as biological *pathways*, or sets of genes. Various pathway and gene sets databases have been compiled, for example, Kyoto Encyclopedia of Genes and Genomes (KEGG)[50,51], Gene Ontology[52], Biocarta[53] and Molecular Signature Data Base[54]. There has been a shift in focus from gene level analysis to pathway level, or gene set level. Detailed descriptions on gene and pathway level analyses are given in Chapter 3.

2.3 Objectives and Study Overview

2.3.1 Objectives

Prostate cancer is a heterogeneous disease, and in spite of recent advances regarding understanding its biology, further discovery of the molecular events underlying prostate cancer is still needed. Gleason grading is an important predictor of prostate cancer outcomes. Higher Gleason scores are more aggressive and have a worse prognosis. It has been long recognized that patients with a total GS $\geq$7 are at greater risk for prostate cancer outcomes[14]. Although this finding has influenced clinical practice, it is still unclear how prostate cancer outcomes differ for various distributions of the total GS between its major and minor components. For example, within the GS of 7 patients, there are differences in outcomes between the patients with a combination of a major GS of 3 and minor of 4, and patients with a major GS of 4 and a minor of 3, with the former category exhibiting better outcomes[15,20]. Our goal is to identify genes and biological pathways differentiating between patients with various combinations of GS, while moving from a less aggressive combination (3,3) to a more aggressive combination (4,4). Our strategy for analyzing microarray gene-expression data is to focus on biological pathways, i.e., sets of genes sharing a biological function. Results of

gene-set analysis are easier to interpret than gene-level analysis, and more robust across similar studies.

2.3.2 Data description

We used data from the Swedish Watchful Waiting cohort with up to 30 years of clinical follow-up[55,56]. The data is nested in a cohort of men with localized prostate cancer diagnosed in the Orebro (1997 to 1994) and South East (1987 to 1999) Health Care Regions of Sweden. Eligible patients were identified through population-based prostate cancer quality databases maintained in these regions, and described in detail in Johansson et.al.[57] The study cohort was followed for cancer-specific and all-cause mortality until March 1, 2006 through record linkages to the Swedish Death Register, which provided date of death or migration. Information on causes of death was obtained through a complete review of medical records by a study end-point committee. Deaths were classified as cancer-specific when prostate cancer was the primary cause of death. Sboner et al were able to trace tumor tissue specimens from 92% if all potentially eligible cases. A total of mRNA expression of 6,100 genes expressions were measured on 255 patients, divided into two extreme groups: men who died of prostate cancer, and men who survived more than 10 years of follow off without metastases. These two groups are referred as lethal and indolent prostate cancer patients. Clinical, pathological and demographical characteristics of the 255 patients are given in Table 2.1. Prostate specific antigen is not available in this cohort, as there were no screening programs in place at the time.

Table2.1: Clinical, Pathological and Demographical characteristics

Characteristics	Counts (%)	Extreme groups		Chi-Square test p-value
Gleason:(score)		Indolent	Lethal	
<7	77(30.2)	52	25	
7	104(40.8)	46	58	
>7	74(29.0)	8	66	0
Age:(year)				
≤70	77 (30.2)	39	38	
>70	178(69.8)	67	111	0.05
Tumor area in biopsy :(%)				
≤5	82(32.2)	54	28	
>5-25	88(34.5)	39	49	
>25-50	45(17.6)	10	35	
>50	35(13.7)	2	33	0
Not assessable	5(2)			
ERG rearrangement status(fusion)				
Negative(0)	206(80.8)	96	110	
Positive(1)	40(15.7)	5	35	0
not assessable	9(3.5)			

2.3.3 Biological pathways from Molecular Signature Database

An important aspect of microarray data analysis, aside from deriving sound methodology, is accessing extensive collections of gene sets and properly linking them to gene expression data. We used the Molecular Signature Database C2 catalog[54] available for download from http://www.broad.mit.edu/gsea, and consisting of 1,892 gene sets, representing metabolic and signalling pathways from online pathway databases, gene sets from biomedical literature including 786 scientific publications, and gene sets compiled from published mammalian microarray studies.

Chapter 3

Methods

Although there are other kinds of data storing information on gene expressions, microarray data contains the largest, most complete information of gene expression. The large number (p) of genes measured on a relatively small number (N) of samples presents a difficult challenge in the analysis of DNA microarray data. This is referred to as the small N, large p problem, also called the high-dimensionality problem. Because of the high dimensionality problem, the classical analysis techniques, which consider the opposite situation, i.e. large N, small p, are no longer applicable to DNA microarray data. Another challenge in the analysis of microarray data is the small variability in the gene expression measures for some of the genes. As a result, the regular test statistic (e.g., two-sample t-test statistic) will have very large values because of the small standard deviation, calling genes whose expression means are not differentially expressed as 'statistically significant'. Another important challenge is inherent to the multiple hypothesis problem of testing tens of thousands of genes. For example, among 10,000 null genes, even if we set the threshold for p-values as low as 0.01, we will identify 100 of those as "significant" genes by chance. In this section we present statistical methods addressing the challenges described above.

3.1 Individual Gene Analysis

Individual gene analysis is a method for gene expression analyses focusing on identifying individual genes that exhibit difference between two states of interest[58]. In response to challenging characteristics of microarray data, Dr. Rob Tibshirani at Stanford University proposed Significant Analysis of Microarray (SAM)[43] , a moderated t-test statistic, together with a False Discovery Rate type of adjustment, calculated based on group-label (e.g., case-control label) permutation tests. The high dimensionality problem calls for permutation tests, which are the basis of calculating statistical significance of associations between a gene and the condition (e.g., disease) of interest. Once a test statistic is calculated for the original data, its significance is evaluated by calculating the test statistic for permuted versions of the data set. Under the null hypothesis of no association,

the group labels are interchangeable. The p-value is calculated based on the permutation distribution of the test statistic, as the proportion of times the permuted test statistic is as extreme, or more extreme than the observed test statistic. SAM is an example of a methodological development which had become a standard data analysis tool for microarray studies.

SAM is based on analyses of random fluctuations in the data and computes gene-specific t-like tests. While SAM is used for a wide variety of phenotypes, we focus on the binary phenotype here. The statistic $d(i)$ measuring the relative difference in gene expression for gene i, is given by:

$$d(i) = \frac{\bar{x}_1(i) - \bar{x}_2(i)}{s(i) + s_0} \qquad (3.1)$$

Where $\bar{x}_1(i)$ is defined as the average level of expression for gene i in the case group and $\bar{x}_2(i)$ is the average expression level for gene i in the control group. The pooled standard deviation "gene-specific scatter" $s(i)$ is:

$$s(i) = \sqrt{a\{\sum[x_1(i) - \bar{x}_1(i)]^2 + \sum[x_2(i) - \bar{x}_2(i)]^2\}} \qquad (3.2)$$

Where $a = (1/n_1 + 1/n_2)/(n_1 + n_2 - 2)$, n_1 and n_2 are the numbers of cases and controls, respectively, the small positive constant s_0 is added to adjust for the "small variability problem" in microarray measurements. The adjustment makes the variance of $d(i)$ independent of the mean level of gene expression: at lower expression levels, since values of $d(i)$ could become very high due to very small values of $s(i)$. Adding a small positive constant s_0 to the denominator ensures that the variance of $d(i)$ is independent of the mean level of gene expression.

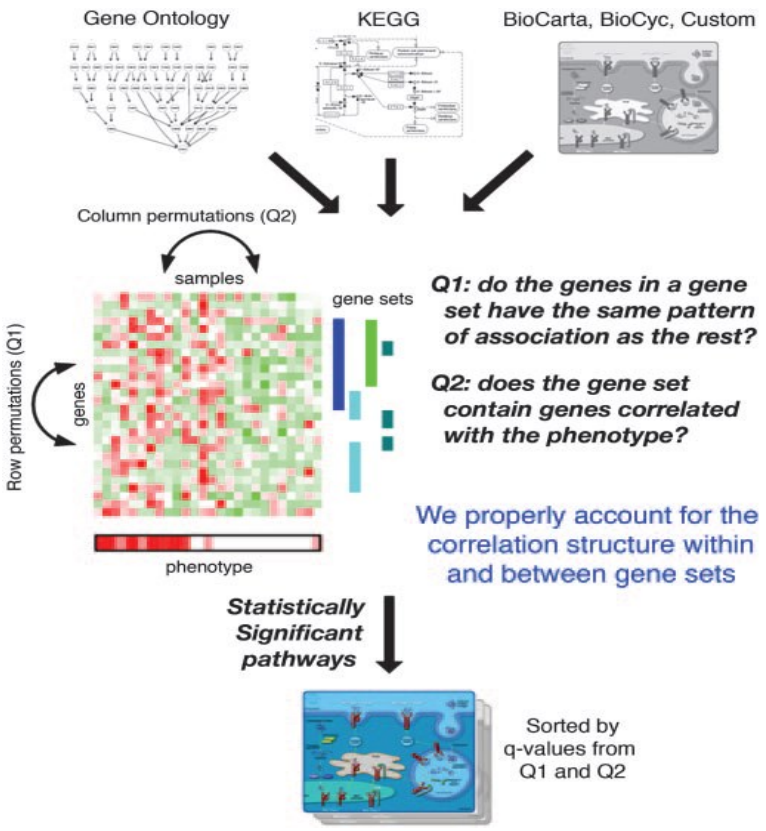
Permutation method is used to calculate p-value for each gene i. Samples in the case group are exchanged randomly with the samples in the control group to get the permuted test statistic $d'(i)$. The significance value for gene i is obtained by comparing the original $d(i)$ with the permuted set of test statistic $d'(i)$'s.

SAM is implemented as both R function and user-friendly Excel Add-On, free for download. It comes with a detailed documentation manual presenting the method, as well as several applications using real microarray datasets.

3.2 Gene-Set Analysis

Analyzing microarray data at an individual gene level usually leads to a list of many "significant" genes, even after multiple comparison adjustments have been made. The process of trying to interpret such a large list of genes is difficult. Moreover, replication of the findings in different microarray experiments is another serious challenge with such individual gene level analysis. Molecular biologists have put together lists of genes grouped by function, such as biological *pathways*, or sets of genes. Various pathway and gene sets databases have been compiled, for example, Kyoto Encyclopedia of Genes and Genomes (KEGG)[50,51], Gene Ontology[52], Biocarta[53] and Molecular Signature Data Base[54]. There has been a shift in focus from gene level analysis to pathway level, or gene set level. Many Gene Set Analysis (GSA) methods for a binary outcome have been proposed in the past decade. The most popular one is Gene Set Enrichment Analysis (GSEA)[56]. A typical GSA works as follows. Gene expression data are used in conjunction with a disease-relevant collection of gene sets (e.g., biological pathways) in the analysis. The gene sets are a-priori determined and based on biological knowledge. A GSA assigns each gene a statistical significance value. An interpretation of this list of p-values leads to identification of gene sets that are differentially expressed by the condition of interest, leading to biologically relevant pathways and other information that can be used for early diagnosis, or tailored treatment, of a disease. A schematic illustration of GSA is given in Figure 3.1, courtesy of Tian *et al*.[59].

Figure 3 1: Outline of the Gene Set Analysis methodology, from Tian et. al., PNAS 2005.



An extensive collection of pathway information is assembled from various *databases*; a statistical test is applied to find relationships between the expression levels and the phenotype, and then two different testing procedures are used to find statistically significant pathways. Proper adjustments for correlation structure and multiple testing are critical.

Many GSA methods have been proposed, with extensive reviews and methodological discussions given by Goeman and Buhlmann[60], and Nam and Kim[61]. An important methodological aspect consists of understanding the difference between competitive or self-contained GSA methods[60,62]. A competitive method employs gene permutation to test whether the association between a gene set and the outcome is equal to those of the other gene sets (so-called "Q1 hypothesis"). A self-contained method employs subject permutation to test the equality of the two mean vectors of gene-set expressions corresponding to the two groups (so-called "Q2 hypothesis"). Goeman and Buhlmann[60] strongly recommended against the testing of Q1 hypothesis using competitive methods with gene sampling, on the grounds of its untenable statistical independence assumption

across genes. Delongchamp *et al.*[63] also commented on how ignoring the correlations within the sets can overstate statistical significance, and proposed meta-analysis methods for combining p-values with a modification to adjust for correlation. Chen *et al.*[64] argue their preference for Q2 over Q1, because the p-values computed under Q2 are consistent with the principle of statistical significance testing, while the p-values computed under Q1 do not take into account correlations among genes. Our focus here is on self-contained methods testing the Q2 hypothesis.

SAM-GS combines the t-like statistics of individual genes into a measure of association of the gene set with the phenotype. For a gene set *S*, it is the L_2 norm of the t-like statistics from equation (3.1):

$$SAMGS = \sum_{i=1}^{|s|} d_{(i)}^2$$

Statistical significance of *S* is obtained based on a phenotype label permutation test[653].

SAM-GS Steps

1) For each of the *N* genes, calculate the statistic *d* as in SAM for an individual-gene analysis:

$$d(i) = \frac{\bar{x}_1(i) - \bar{x}_2(i)}{s(i) + s_0}$$

where the 'gene-specific scatter' $s(i)$ is a pooled standard deviation over the two groups of the phenotype, and s_0 is a small positive constant that adjusts for the small variability encountered in microarray data.

2) Compute the *SAMGS* test statistic corresponding to set *S*:

$$SAMGS = \sum_{i=1}^{|s|} d_{(i)}^2$$

3) permute the labels of the phenotype *and* repeat 1) and (2). Repeat until all (or a large number of) permutations are considered.

4) Statistical significance for the association of *S* and the phenotype *is* obtained by comparing the observed value of the *SAMGS* statistic from (2) and its permutation distribution from (3).

3.3 Gene Set Reduction

Significance Analyses of Microarray for Gene-Set Reduction (SAM-GSR) proposed by Dinu et al.[65] has established a new direction of finding core subsets from gene sets differentially expressed. SAM-GSR was motivated by the fact that not all genes in a significant set are contributing to its significance.

SAM-GSR

Given a statistically significant association of the gene set *S* with the phenotype, SAM-GSR applies SAM-GS sequentially to subsets of the significant gene set *S* and identifies a core set of genes that mostly contribute to the statistical significance of *S*. In reducing the gene set *S*, we used the following principle: for a pair of genes in *S*, genes *i* and *j*, $|d_i| > |d_j|$ suggests that gene *j* belongs to a subset only if gene *i* belongs to the subset. This principle is motivated by the fact that d_i^2 represents each gene's contribution to the test statistic SAM-GS and the core subset must consist of genes with larger contributions. SAM-GSR gradually partitions the entire set *S*, into two subsets, based on the principle above and evaluates their association with the phenotype. SAM-GSR can be summarized in a few steps:

1) For each of N genes, calculate the statistic *d(i)* as in SAM for an individual gene analyses :

$$d(i) = \frac{\bar{x}_1(i) - \bar{x}_2(i)}{s(i) + s_0}$$

where $\bar{x}_1(i)$ is the average level of expression for gene i in the case group while $\bar{x}_2(i)$ is the average expression level for gene i in the control group, $s(i)$ is the pooled standard deviation of gene expression over the two groups of phenotype, s_0 is the small positive constant that adjusts for the small variability encountered in microarray data.

2) For k = 1,… ,|S| − 1, select the first *k* genes with largest statistic |*d*| to form a reduced set R_k. Let c_k be the *SAM-GS* p-value of the complement of R_k in *S*.

3) The reduced set R_k corresponds to the least *k* such that c_k is larger than a threshold *c*, chosen by the analyst.

By removing genes with joint statistical significance, as a set, above a threshold, i.e. $c_k > c$, we are protected against losing genes that are not significant by themselves, but collectively, they form a set that is significant[65].

3.4 Multiple Hypothesis Testing in Microarray Studies

Adjustments for multiple hypothesis testing need to be made in the analysis of microarray data, as thousands of genes are being tested. Multiple hypothesis adjustments are also needed for gene set analysis, as a large number of gene sets are being tested.

Table2: Property of Multiple Hypothesis test

	Number not rejected	Number rejected	Total
True Null hypothesis	U	V	m_0
True Alternative hypothesis	T	S	m_1
	m-R	R	M

Consider the problem of testing simultaneously m null hypotheses H_j ; j = 1, 2, . . . ,m, and denote by R the number of rejected hypotheses. This situation can be summarized in the Table 3.1:

The specific m hypotheses are assumed to be known in advance, m_0 and $m_1 = m - m_0$ represent the numbers of true and false null hypotheses, respectively, and are unknown parameters. R is an observable random variable and S, T, U and V are unobservable random variables. A variety of generalizations of the Type I error are possible. The Family-wise error rate (FWER) is defined as;

The probability of at least one Type I error, i.e., FWER=Pr(V ≥ 1). The false discovery rate (FDR) of Benjamini and Hochberg[66] is the expected proportion of Type I errors among the rejected hypothesis, i.e., FDR = E (Q), where

$$Q = \begin{cases} V/R, \text{ if } R > 0 \\ 0, \text{ if } R = 0 \end{cases}$$

A multiple testing procedure is said to control a particular Type I error rate at level α, if this error rate is less than or equal to α when the given procedure is applied to produce a list of R rejected hypotheses.

Bonferroni adjustment is a popular statistical method to control for multiple hypotheses testing. This adjustment is straightforward and consists of dividing the Type I error for each individual hypothesis by the total number of hypotheses.

Besides its ability to control the overall Type I error, Bonferroni method has some limitations. Firstly, when Bonferroni adjustment is applied, it is assumed that all the tests are independent, which is untenable in microarray studies, as it is well understood that many of the genes or sets are correlated. Secondly, Bonferroni adjustment is a conservative adjustment method leading to omissions of truly significant associations[67,68].

Some methods proposed corrections to Bonferroni adjustment to make the approach less conservative[69,70]. They reduce the number of false positive as well as the number of discoveries by attempting to assign an adjusted p-value to each test, or similarly, decrease the p-value threshold. While there are a number of approaches to address limitations of Bonferroni adjustment, false discovery rate (FDR) is a widely used statistical method to control multiple testing hypotheses in microarray studies[66].

FDR is an expected proportion of false positives among all the tests called significant and is given by:

$$FDR = E(\frac{V}{R})$$

Benjamini and Hochberg[66] noticed that the FDR estimation reduces to estimating the proportion of null hypotheses.

While FDR is a useful measurement of the overall error rate for a set of tests declared significant, q-value is a measure given to each single test. The q-value of an individual hypothesis test is an estimated measure of the probability of false discovery when this test is declared significant[67]. Benjamini and Hochberg proposed the following algorithm for calculating individual FDR values, or q-values, for each gene, in a microarray study:

1. The p-values for each of the genes are ranked from the smallest to the largest
2. The largest p-value remains as it is.
3. The second largest p-value, p_{N-1}, is adjusted as:
Corrected p-value = $p_{N-1} \times N/N-1$
4. The third largest p-value, p_{N-2}, is adjusted as:
Corrected p-value = $p_{N-2} \times N/N-2$
5. The adjustments are made for the entire list of genes and the smallest p-value, p1, is adjusted as:
Corrected p-value = $p_1 \times N$

Benjamini and Hochberg method controls the false discovery rate. If the error rate is 0.05, then 5% of the genes declared significant are truly null genes.

Chapter 4

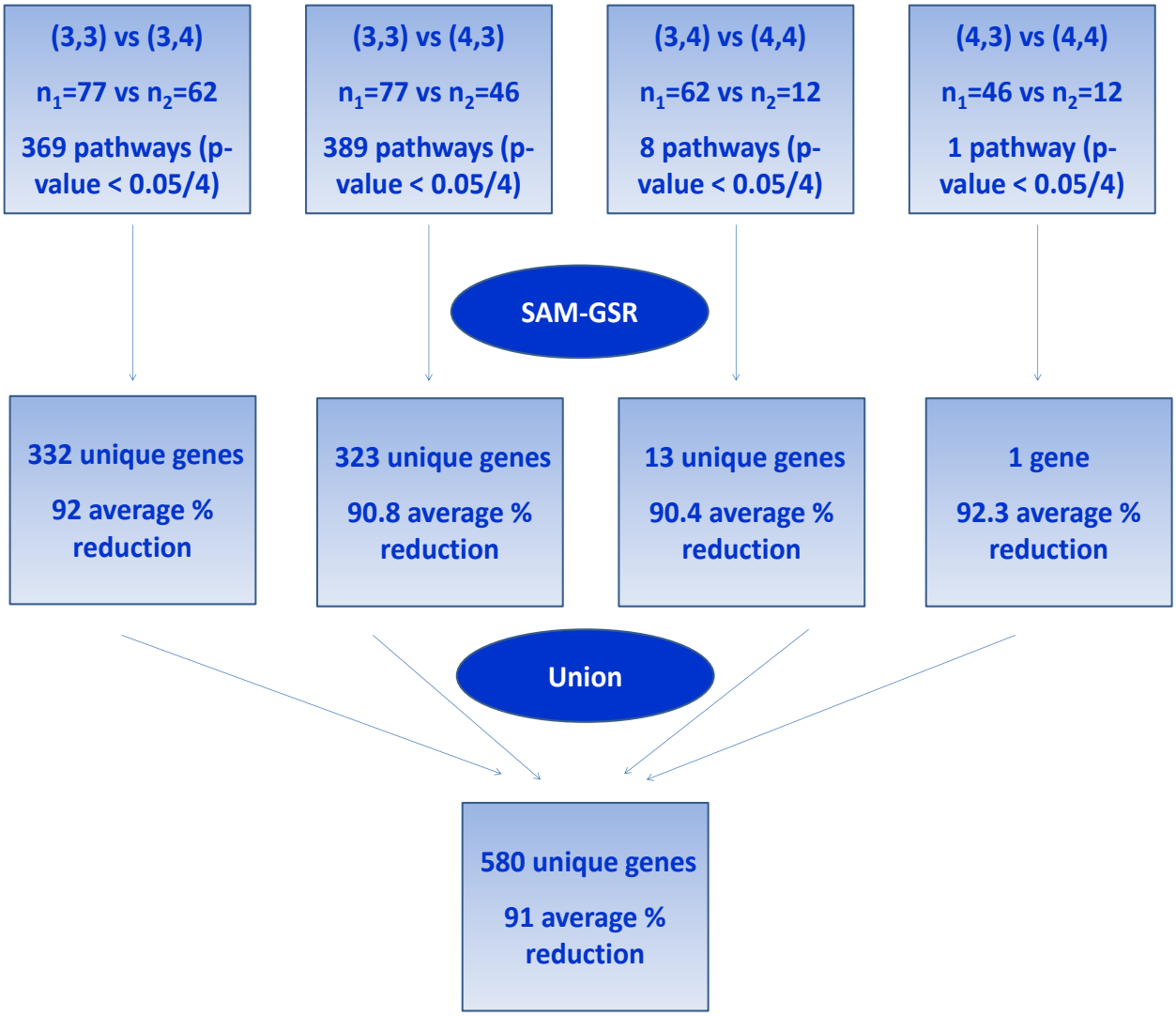
Results

4.1 Gene Set Reduction results for GS ranging from (3,3) to (4,4)

Data analyses started by validating a strong signal in our data at the level of lethal versus non-lethal prostate cancer patients. One thousand three hundred and fifty one genes out of 1,892 MSigDB gene sets were found to be differentially expressed between 140 lethal and 117 non-lethal prostate cancer patients, using SAM-GS. Furthermore, 1,246 gene sets were found to be differentially expressed between 80 patients with major and minor GS ≤ 3 versus 68 patients with major and minor GS ≥ 4. The number of significant gene sets and core set sizes decrease considerably when comparing patients with larger total GS, indicating a challenge in discriminating between higher risk groups of patients. For example, a comparison of 77 patients with GS of (3,3) versus 62 patients with GS of (3,4) gives 369 gene sets significant at a p-value of 0.05/4=0.0125. The adjustment corresponds to a total of 4 GS combinations, as described in Figure 4.1. Eight gene sets are differentially expressed between GS of (3,4) vs (4,4), and only one gene set differentiates between (4,3) and (4,4). The FDR cut-offs for the four combinations are 0.006, 0.004, 0.27, and 0.95.

SAM-GSR achieved a 91% reduction, averaged over the four GS combinations, starting from (3,3) and ending with (4,4). The 369 gene sets differentiating between (3,3) and (3,4) were reduced to 332 unique genes shared across the core gene sets. The percent reduction was calculated for each gene-set as the number of genes outside the core set divided by the size of the gene set, and multiplied by 100. The percent reduction is averaged over the significant gene-sets. The overall average percent reduction across combinations ranging from (3,3) to (4,4) was 91%.

Figure 4.1: Gene-set reduction flow chart.

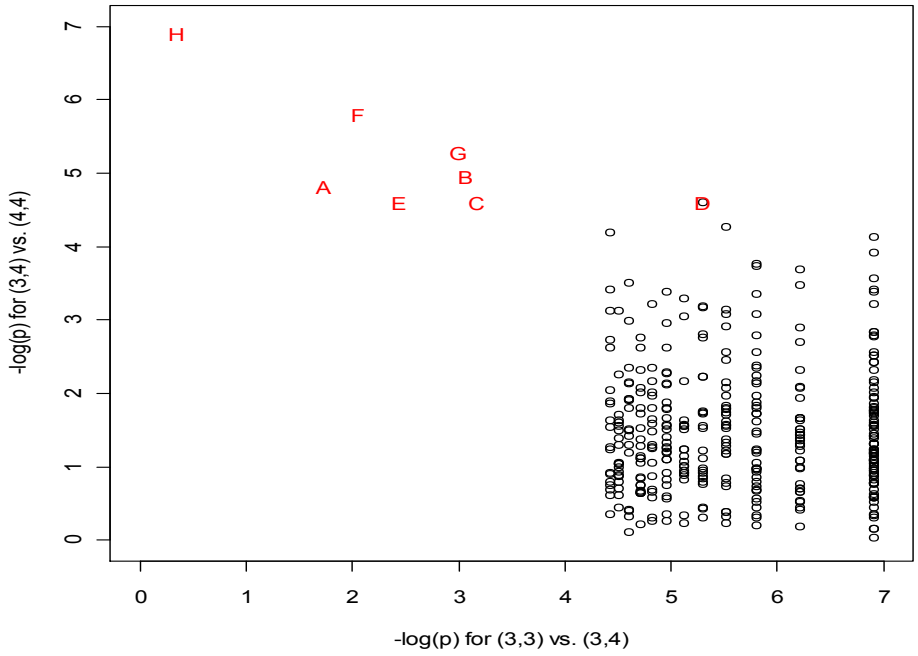


Moving from a less aggressive gleason scores combination (3,3) to a more aggressive combination (4,4), 580 unique genes were identified.

At the gene set level analysis, only one of the eight pathways differentiating between (3,4) vs (4,4) is represented among the 369 pathways differentiating between (3,3) vs (3,4). Negative log p-values according to the two analyses are shown in Figure 4.2. The eight pathways are represented as letters of the alphabet from A to H. Similarly, only one out of the eight pathways differentiating

between (3,4) vs (4,4) is represented among the 389 pathways differentiating between (3,3) vs (4,3), Figure 4.3.

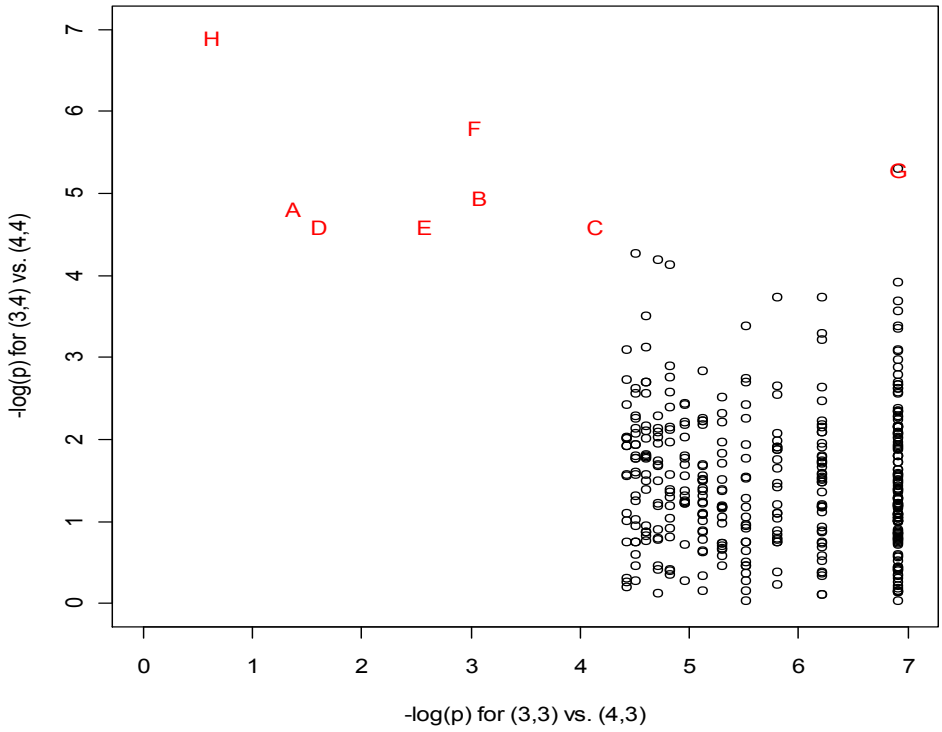
Figure 4.2: Negative log p-values for gene-sets differentially expressed between (3,4) vs (4,4) or (3,3) vs (3,4).



The eight gene-sets differentiating between (3,4) vs (4,4) are denoted as letters of alphabet as below.

A	BUT_TSA_UP
B	CMV_HCMV_TIMECOURSE_14HRS_DN
C	FERRANDO_CHEMO_RESPONSE_PATHWAY
D	HDACI_COLON_CUR24HRS_UP
E	LEE_CIP_UP
F	TSA_PANC50_UP
G	UEDA_MOUSE_SCN
H	UREACYCLEPATHWAY

Figure 4.3: Negative log p-values for gene-sets differentially expressed between (3,4) vs (4,4) or (3,3) vs (4,3).



The eight gene-sets differentiating between (3,4) vs (4,4) are denoted as letters of alphabet as below.

A	BUT_TSA_UP
B	CMV_HCMV_TIMECOURSE_14HRS_DN
C	FERRANDO_CHEMO_RESPONSE_PATHWAY
D	HDACI_COLON_CUR24HRS_UP
E	LEE_CIP_UP

F	TSA_PANC50_UP
G	UEDA_MOUSE_SCN
H	UREACYCLEPATHWAY

4.2 Gene Set Reduction results for GS of (3,4) vs (4,3)

We performed a gene-set analysis and reduction for 62 patients with GS of (3,4) versus 46 patients with GS of (4,3). Thirty two gene sets were identified at 0.05 significance level, with a FDR value of 0.75. The core sets of the thirty two gene sets are presented in Table 4.1.

Table 4.3: Results of SAM-GS and SAM-GSR analyses for 62 patients with Gleason Score of (3,4) vs 46 patients with Gleason Score of (4,3).

Gene-Set Name	**Gene-Set Size**	**P-value**	**Core Set Size**	**Core gene 1**	**Core gene 2**	**Core gene 3**
AGED_MOUSE_HYPOTH_DN	28	0.002	3	*DNM1*	*FSTL1*	*APOE*
CD40PATHWAY	9	0.008	1	*IKBKAP*		
HSA05110_CHOLERA_INFECTION	23	0.011	1	*SEC61A1*		
HEATSHOCK_YOUNG_UP	9	0.016	1	*ANXA1*		
NOUZOVA_CPG_METHLTD	22	0.018	2	*EFNA5*	*EPHA5*	
VEGF_HUVEC_2HRS_UP	25	0.018	2	*APOE*	*PPY*	
HYPOPHYSECTOMY_RAT_DN	39	0.021	2	*COL3A1*	*NPPA*	
PENG_GLUCOSE_UP	32	0.022	1	*OCLN*		
LIAN_MYELOID_DIFF_TF	31	0.022	3	*BHLHB2*	*MYB*	*NFKB1*
HSA00330_ARGININE_AND_ PROLINE_METABOLISM	25	0.023	1	*ARG2*		
ADIPOGENESIS_HMSC_ CLASS5_UP	6	0.025	1	*MYB*		
ONE_CARBON_POOL_BY_FOLATE	15	0.028	1	*SHMT2*		

TNFR2PATHWAY	14	0.029	1	*IKBKAP*
UVC_HIGH_D9_DN	20	0.03	1	*NAP1L1*
HDACI_COLON_CLUSTER6	24	0.031	1	*NAP1L1*
NDKDYNAMINPATHWAY	15	0.032	1	*DNM1*
TYPE_III_SECRETION_SYSTEM	14	0.034	1	*ATP6V1C1*
ANDROGEN_GENES	43	0.036	1	*NR1I3*
GH_HYPOPHYSECTOMY_RAT_UP	10	0.036	1	*COL3A1*
ARGININE_AND_PROLINE_ METABOLISM	42	0.04	1	*MAOA*
FMLPPATHWAY	30	0.04	1	*NFATC3*
HSA00670_ONE_CARBON_ POOL_BY_FOLATE	13	0.04	1	*SHMT2*
PHOTOSYNTHESIS	15	0.041	1	*ATP6V1C1*
HSA00051_FRUCTOSE_AND_ MANNOSE_METABOLISM	28	0.041	1	*MTMR6*
KIM_TH_CELLS_UP	31	0.044	1	*ETS1*
GCRPATHWAY	16	0.044	1	*ANXA1*
HEARTFAILURE_ATRIA_UP	20	0.045	1	*FKBP8*
ALZHEIMERS_INCIPIENT_DN	88	0.046	1	*UROS*
GAMMA.UV_FIBRO_UP	25	0.046	1	*IL10RB*
AGUIRRE_PANCREAS_CHR8	28	0.047	1	*HAS2*
GH_GHRHR_KO_24HRS_DN	73	0.047	1	*IFNAR1*
FERRANDO_CHEMO_ RESPONSE_PATHWAY	9	0.048	1	*DTYMK*

We compared the results of analysis of GS (3,4) versus (4,3) with results of analysis of GS ranging from (3,3) to (4,4). SAM-GS p-values of the eight gene sets differentiating between (3,4) and (4,4) are presented in Table 4.2.

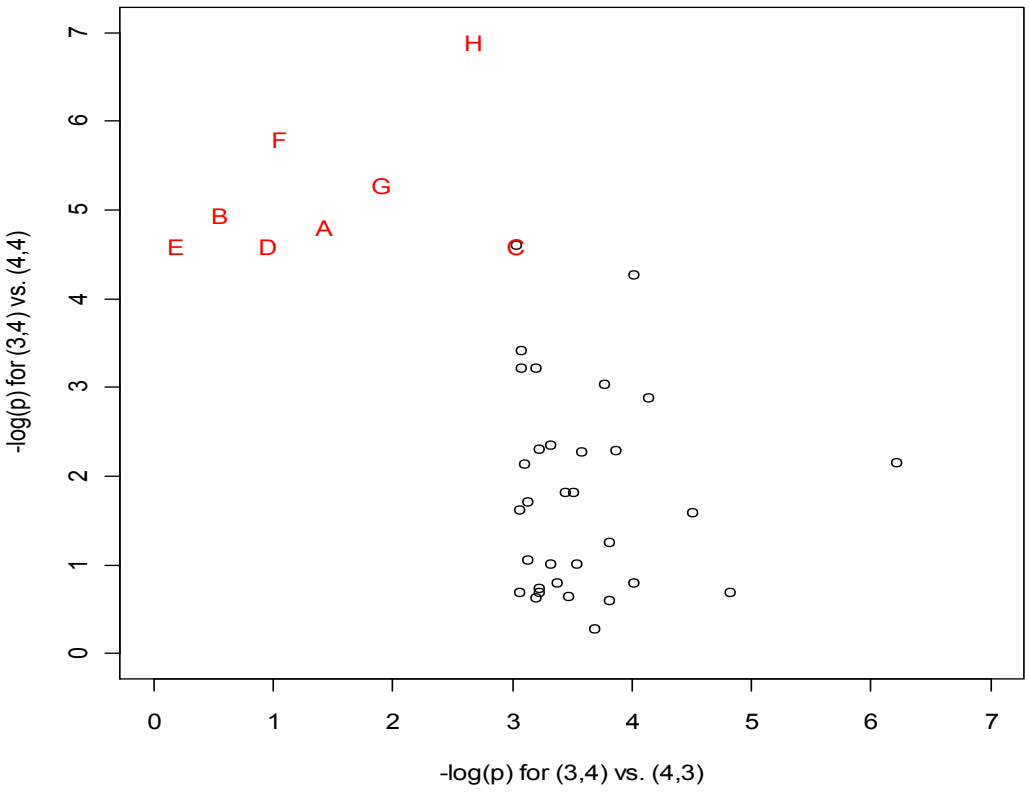
At the gene set level analysis, only one of the eight pathways differentiating between (3,4) vs (4,4) is represented among the 32 pathways differentiating between (3,4) vs (4,3). Negative log p-values according to the two analyses are shown in Figure 4.3. The eight pathways are represented as letters of the alphabet from A to H.

Table 4.4: SAM-GS p-values for various distributions of Gleason Scores.

Gene-Set Name	**Gene-Set Size**	**(3,3) vs (3,4)**	**(3,3) vs (4,3)**	**(3,4) vs (4,4)**	**(4,3) vs (4,4)**	**(3,4) vs (4,3)**
BUT_TSA_UP	18	0.179	0.254	0.008	0.174	0.24
CMV_HCMV_ TIMECOURSE_14HRS_DN	36	0.047	0.046	0.007	0.069	0.574
FERRANDO_CHEMO_RESPONSE_PA THWAY	9	0.042	0.016	0.01	0.045	0.048
HDACI_COLON_CUR24HRS_UP	27	0.005	0.2	0.01	0.069	0.383
LEE_CIP_UP	50	0.088	0.076	0.01	0.066	0.834
TSA_PANC50_UP	29	0.128	0.048	0.003	0.029	0.346
UEDA_MOUSE_SCN	58	0.05	0.001	0.005	0.228	0.15
UREACYCLEPATHWAY	7	0.721	0.536	0.001	0.016	0.07

At the gene level, none of the 13 core genes from comparing (3,4) vs (4,4) are represented among the 332 core genes comparing (3,3) vs (3,4), or among the 323 core genes comparing (3,3) vs (4,3). The 13 core genes are shown in Table 4.3. Biological process and cellular component from Gene Ontology for core genes are presented in Table 4.4. The set consisting of the 13 genes shows a marginal association with GS of (3,4) vs (4,3), with a SAM-GS p-value of 0.059.

Figure 4.4: Negative log p-values for gene-sets differentially expressed between (3,4) vs (4,4) or (3,4) vs (4,3).



The eight gene-sets differentiating between (3,4) vs (4,4) are denoted as letters of alphabet as below.

A	BUT_TSA_UP
B	CMV_HCMV_TIMECOURSE_14HRS_DN
C	FERRANDO_CHEMO_RESPONSE_PATHWAY
D	HDACI_COLON_CUR24HRS_UP
E	LEE_CIP_UP
F	TSA_PANC50_UP
G	UEDA_MOUSE_SCN
H	UREACYCLEPATHWAY

Table 4.5: SAM-GS and SAM-GSR analyses for 62 patients with Gleason Score of (3,4) vs 12 patients with Gleason Score of (4,4).

Gene-Set Name	**Gene-Set Size**	**P-value**	**Core Set Size**	**Core gene 1**	**Core gene 2**	**Core gene 3**
BUT_TSA_UP	18	0.008	1	*GADD45A*		
CMV_HCMV_ TIMECOURSE_14HRS_DN	36	0.007	2	*ETV1*	*APEX1*	
FERRANDO_CHEMO_RESPONSE_PATHWAY	9	0.01	1	*CDA*		
HDACI_COLON_ CUR24HRS_UP	27	0.01	3	*RPN2*	*ALDOA*	*CCND1*
LEE_CIP_UP	50	0.01	2	*ETV1*	*COL4A2*	
TSA_PANC50_UP	29	0.003	2	*BIK*	*NOTCH3*	
UEDA_MOUSE_SCN	58	0.005	2	*GADD45A*	*SMPDL3A*	
UREACYCLEPATHWAY	7	0.001	2	*CPS1*	*ASL*	

Table 4.6: Biological process and cellular component from Gene Ontology for core genes from SAM-GSR analyses for 62 patients with Gleason Score of (3,4) vs 12 patients with Gleason Score of (4,4).

Core Gene Name	Biological process	Cellular component
ETV1	cell growth, angiogenesis, migration, proliferation and differentiation	nucleus
GADD45A	cell cycle arrest	nucleus, cytoplasm
*ALDOA**	fructose and glucose metabolic process	nucleus, cytosol
APEX1	mitotic cell cycle	nucleus, cytoplasm
ASL	urea cycle, cellular nitrogen compound metabolic process	cytoplasm, cytosol
BIK	apoptotic	endomembrane system
CCND1	transition of mitotic cell cycle	nucleus, cytosol
CDA	pyrimidine nucleobase metabolic process, cell surface receptor signaling pathway	extracellular region, cytosol
*COL4A2**	angiogenesis, endodermal cell differentiation, cellular response to transforming growth factor beta stimulus	extracellular region
CPS1	urea cycle, glutamine metabolic process	nucleus, cytoplasm, mitochondrial inner membrane
NOTCH3	notch signaling pathway, negative regulation of neuron differentiation	nucleoplasm, cytoplasm, extracellular region
*RPN2**	translation, cellular protein modification process, cellular protein metabolic process, response to drug, post-translational protein modification	autophagosome membrane, nucleus, integral component of membrane
*SMPDL3A**	sphingomyelin catabolic process	extracellular space, extracellular exosome

*Indicates genes not identified as significant in SAM-GSR analysis of patients with GS of 6 vs GS of 7, or GS of 7 vs GS of 8.

Chapter 5.

5.1 Discussion and Conclusion

Gleason score plays an important role in prostate cancer diagnostic and treatment. The current practice indicates patients with a total GS of 7 or larger to be at higher risk. It has been recognized in the literature that the representation of the total GS into its major and minor component plays an important role in understanding severity of the disease, with patients exhibiting a GS combination of (4,3) being at higher risk than those with a GS combination of (3,4). We studied differences at the gene and gene-set levels between patients with various combinations of major and minor Gleason Scores, moving from a less aggressive combination of (3,3) and towards a more aggressive combination of (4,4). We note that groups of patients within this GS range are expected to exhibit subtle changes, especially at the gene level. Significance Analysis of Microarrays for Gene Sets (SAM-GS) is a powerful method for detecting subtle and coordinated changes in microarray gene expression data. Gene-set analysis was developed in response to moderate to weak signal at the gene level. The key element in gene-set analysis is to take advantage of correlations across genes in a set, therefore boosting the analysis power. SAM-GS was found to perform well in comparative studies of seven self-contained gene-set analysis methods[65]. One of the weaknesses of self-contained methods is that only a few genes in a set can drive the significance of the whole set. Significance Analysis of Microarrays for Gene Set Reduction (SAM-GSR) was designed to extract core genes that contribute to the significance of the whole set. We reason that these two methods are appropriate for analysing differences at gene and gene-set levels across various combinations of Gleason Scores.

Some of the gene-sets and pathways identified significant in our analyses have been previously found to play various roles in cancer progression and identification of novel therapeutic strategies. For example, the CD40 pathway differentially expressed between GS of (3,4) vs (4,3), has been shown to play an immunosuppressive role[71]. The CD40 pathway has been shown to play a crucial role in production of cytokines, which modulate the function of T lymphocytes in antitumor responses[72]. TNFR2 pathway was also differentially expressed between GS of (3,4) vs (4,3).

TNFR2 is a receptor of Tumor Necrosis Factor (TNF), a multifunctional proinflamatory cytokine. Members of the TNFR superfamily can send both survival and death signals to cells[73].

Urea cycle pathway was differentially expressed between GS of (3,4) vs (4,4), p-value of 0.001, and GS of (4,3) vs (4,4), p-value of 0.016, marginally significant for GS of (3,4) vs (4,3), p-value of 0.07, and not significant for GS of (3,3) vs (3,4), p-value of 0.721, or (3,3) vs (4,3), p-value of 0.536. In urea cycle pathway, the enzyme ornithine decarboxylase (ODC) converts the metabolite ornithine to putrescene. ODC has previously been found as over-expressed in prostate cancer[74] and is the target of the chemotherapeutic agent diflourmethylornithine (DFMO)[75] .

5.2 Strength and Limitation

Our comprehensive analysis of combinations of major and minor Gleason scores brings additional insights to the current practice based on the sum of the two components, especially for values of the total GS of 7 or 8, indicating patients at greater risk. There are a plethora of possible area in which our study can be applied, from public health science to genetics. My particular interest would be to use the method to find core gene-set responsible for other cancer diseases and non-cancer diseases as well. The result of our research can play an important role as a source of information to improve personalized medicine and intervention therapy by interpreting the biological association of our obtained core gene. Further studies are needed to validate our results at the gene and pathway levels.

References

1. King AJ, Evans M, Moore TH, et al. Prostate cancer and supportive care: A systematic review and qualitative synthesis of men's experiences and unmet needs. *Eur J Cancer Care (Engl)*. 2015. doi: 10.1111/ecc.12286; 10.1111/ecc.12286.

2. Rocca BJ, Ginori A, Barone A, et al. Translationally controlled tumor protein in prostatic adenocarcinoma: Correlation with tumor grading and treatment-related changes. *BioMed Research International*. 2014;2015:985950. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4312572/. doi: 10.1155/2015/985950.

3. Mimeault M, Batra SK. Recent advances on multiple tumorigenic cascades involved in prostatic cancer progression and targeting therapies. *Carcinogenesis*. 2006;27(1):1-22. doi: 10.1093/carcin/bgi229.

4. Mordukhovich I, Reiter PL, Backes DM, et al. A review of african american-white differences in risk factors for cancer: Prostate cancer. *Cancer causes & control : CCC*. 2010;22(3):341-357. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3443558/. doi: 10.1007/s10552-010-9712-5.

5. Fradet Y, Klotz L, Trachtenberg J, Zlotta A. The burden of prostate cancer in canada. *Canadian Urological Association Journal*. 2009;3(3):S92-S100. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2698782/.

6. Mucci LA, Pawitan Y, Demichelis F, et al. Testing a multigene signature of prostate cancer death in the swedish watchful waiting cohort. *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American*

Society of Preventive Oncology. 2008;17(7):1682-1688. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2536630/. doi: 10.1158/1055-9965.EPI-08-0044.

7. Hall PS, Hamilton P, Hulme CT, et al. Costs of cancer care for use in economic evaluation: A UK analysis of patient-level routine health system data. *Br J Cancer*. 2015. http://dx.doi.org/10.1038/bjc.2014.644.

8. Mohler J, Bahnson RR, Boston B, et al. Prostate cancer. *Journal of the National Comprehensive Cancer Network*. 2010;8(2):162-200.

9. Prostate cancer foundation. http://www.pcf.org/site/c.leJRIROrEpH/b.5802045/k.6D36/What_Is_Prostate_Cancer.htm. Accessed 02/13, 2015.

10. DeMarzo AM, Nelson WG, Isaacs WB, Epstein JI. Pathological and molecular aspects of prostate cancer. *Lancet*. 2003;361(9361):955-964. doi: 10.1016/S0140-6736(03)12779-1.

11. Kang DY, Li HJ. The effect of testosterone replacement therapy on prostate-specific antigen (PSA) levels in men being treated for hypogonadism: A systematic review and meta-analysis. *Medicine (Baltimore)*. 2015;94(3):e410. doi: 10.1097/MD.0000000000000410; 10.1097/MD.0000000000000410.

12. Shaik R, Ramakrishna W. Genes and co-expression modules common to drought and bacterial stress responses in arabidopsis and rice. *PLoS ONE*. 2013;8(10):e77261. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3795056/. doi: 10.1371/journal.pone.0077261.

13. Miyamoto S, Ito K, Miyakubo M, et al. Impact of pretreatment factors, biopsy gleason grade volume indices and post-treatment nadir PSA on overall survival in patients with metastatic prostate cancer treated with step-up hormonal therapy. *Prostate Cancer and Prostatic Diseases*. 2012;15(1):75-86.

14. Pierorazio PM, Walsh PC, Partin AW, Epstein JI. Prognostic gleason grade grouping: Data based on the modified gleason scoring system. *BJU Int*. 2013;111(5):753-760. doi: 10.1111/j.1464-410X.2012.11611.x; 10.1111/j.1464-410X.2012.11611.x.

15. MAKAROV DV, SANDERSON H, PARTIN AW, EPSTEIN JI. Gleason score 7 prostate cancer on needle biopsy: Is the prognostic difference in gleason scores 4 3 and 3 4 independent of the number of involved cores? *J Urol*. 2002;167(6):2440-2442. http://www.sciencedirect.com/science/article/pii/S0022534705650008. doi: http://dx.doi.org/10.1016/S0022-5347(05)65000-8.

16. Anatomy and physiology of prostate. Gleason DF. Histologic grading of prostate cancer: A perspective. Hum Pathol. 1992;23(3):273-279. http://www.sciencedirect.com/science/article/pii/004681779290108F. doi: http://dx.doi.org/10.1016/0046-8177(92)90108-F. Updated 2015. Accessed 03/21, 2015.

17. Prostate enlargement: Benign prostatic hyperplacia. http://www.kidneyurology.org/Library/Urologic_Health.php/Prostate_Enlargement.php. Updated 2006.

18. Immage collection: Human anatomy. http://www.webmd.com/urinary-incontinence-oab/picture-of-the-prostate. Updated 2015.

19. Cullen J, Rosner IL, Brand TC, et al. A biopsy-based 17-gene genomic prostate score predicts recurrence after radical prostatectomy and adverse surgical pathology in a racially diverse population of men with clinically low- and intermediate-risk prostate cancer. LID - S0302-2838(14)01213-5 pii] LID - 10.1016/j.eururo.2014.11.030 doi]. *European urology JID - 7512719 OTO - NOTNLM*. 1203.

20. Humphrey PA. Gleason grading and prognostic factors in carcinoma of the prostate. *Mod Pathol*. 2004;17(3):292-306. http://dx.doi.org/10.1038/modpathol.3800054.

21. Understanding gleason grading. http://prostatecancerinfolink.net/treatment/staging-grading/gleason-grading/. Updated November 29, 2008. Accessed 02/02, 2015.

22. Gleason DF. Histologic grading of prostate cancer: A perspective. *Hum Pathol*. 1992;23(3):273-279. http://www.sciencedirect.com/science/article/pii/004681779290108F. doi: http://dx.doi.org/10.1016/0046-8177(92)90108-F.

23. FURUBAYASHI N, NAKAMURA M, NISHIYAMA K, HASEGAWA Y. Original and infiltrating patterns of prostatic carcinoma. *Oncology Letters*. 2009;1(1):41-44. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3436381/. doi: 10.3892/ol_00000007.

24. Stangelberger A, Waldert M, Djavan B. Prostate cancer in elderly men. *Reviews in Urology*. 2008;10(2):111-119. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2483315/.

25. Pan CC, Potter SR, Partin AW, Epstein JI. The prognostic significance of tertiary gleason patterns of higher grade in radical prostatectomy specimens: A proposal to modify the gleason grading system. *Am J Surg Pathol*. 2000;24(4):563-569.

26. The herediatary material of life. http://www.nlm.nih.gov/medlineplus/magazine/issues/summer13/articles/summer13pg11-12.html. Updated 2013.

27. What is DNA? http://ghr.nlm.nih.gov/handbook/basics/dna. Updated June 29, 2015.

28. Genetic home reference. http://ghr.nlm.nih.gov/handbook/basics/dna. Updated February 9, 2015. Accessed 02/13, 2015.

29. Shu Y, Hong-Hui L. Transcription, translation, degradation, and circadian clock. *Biochem Biophys Res Commun*. 2004;321(1):1-6. doi: 10.1016/j.bbrc.2004.06.093.

30. What is gene expression? http://www.news-medical.net/health/What-is-Gene-Expression.aspx. Updated Jun 24, 2014. Accessed 02/17, 2015.

31. Guo L, Liu Y, Bai Y, Sun Y, Xiao F, Guo Y. Gene expression profiling of drug-resistant small cell lung cancer cells by combining microRNA and cDNA expression analysis. *Eur J Cancer*. 2010;46(9):1692-1702. doi: 10.1016/j.ejca.2010.02.043; 10.1016/j.ejca.2010.02.043.

32. Leung YF, Cavalieri D. Fundamentals of cDNA microarray data analysis. *Trends in Genetics*. 2003;19(11):649-659. http://www.sciencedirect.com/science/article/pii/S0168952503002695. doi: http://dx.doi.org/10.1016/j.tig.2003.09.015.

33. Singh NK, Repsilber D, Liebscher V, Taher L, Fuellen G. Identifying genes relevant to specific biological conditions in time course microarray experiments. *PLoS ONE*.

2013;8(10):e76561. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3795718/. doi: 10.1371/journal.pone.0076561.

34. Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*. 1995;270(5235):467-470.

35. Mirnics K, Pevsner J. Progress in the use of microarray technology to study the neurobiology of disease. *Nat Neurosci*. 2004;7(5):434-439. http://dx.doi.org/10.1038/nn1230.

36. Microarray technology. http://www.genome.gov/dmd/img.cfm?node=Photos/Graphics&id=85200. Updated 2010. Accessed 03/29, 2015.

37. Gao X. Multiple testing corrections for imputed SNPs. *Genet Epidemiol*. 2011;35(3):154-158. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3055936/. doi: 10.1002/gepi.20563.

38. McDonald JH. *Handbook of biological statistics.* Vol 2. Sparky House Publishing Baltimore; 2009.

39. Kutalik Z, Inwald J, Gordon SV, et al. Advanced significance analysis of microarray data based on weighted resampling: A comparative study and application to gene deletions in mycobacterium bovis. *Bioinformatics*. 2004;20(3):357-363. doi: 10.1093/bioinformatics/btg417.

40. Barrett JC, Fry B, Maller J, Daly MJ. Haploview: Analysis and visualization of LD and haplotype maps. *Bioinformatics*. 2005;21(2):263-265. doi: 10.1093/bioinformatics/bth457.

41. Gao X, Becker LC, Becker DM, Starmer JD, Province MA. Avoiding the high bonferroni penalty in genome-wide association studies. *Genet Epidemiol*. 2010;34(1):100-105. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2796708/. doi: 10.1002/gepi.20430.

42. Microarray technology. http://www.genomic.ch/techno_array.php. Updated 2008. Accessed 03/29, 2015.

43. Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences*. 2001;98(9):5116-5121. http://www.pnas.org/content/98/9/5116.abstract.

44. Tibshirani R, Hastie T, Narasimhan B, Chu G. Class prediction by nearest shrunken centroids, with applications to DNA microarrays. *Statistical Science*. 2003;18(1):104-117. http://www.jstor.org/stable/3182873.

45. Tarca AL, Romero R, Draghici S. Analysis of microarray experiments of gene expression profiling. *Am J Obstet Gynecol*. 2006;195(2):373-388. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2435252/. doi: 10.1016/j.ajog.2006.07.001.

46. Lipes BD, Keene JD. Autoimmune epitopes in messenger RNA. *RNA*. 2002;8(6):762-771. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1370295/.

47. National human genome research institute (NHGRI). http://www.genome.gov/10000533. Updated November 15, 2011. Accessed 02/01, 2015.

48. Sealfon S, Chu T. RNA and DNA microarrays. In: Khademhosseini A, Suh K, Zourob M, eds. Vol 671. Humana Press; 2011:3-34. http://dx.doi.org/10.1007/978-1-59745-551-0_1. 10.1007/978-1-59745-551-0_1.

49. Wang DG, Fan JB, Siao CJ, et al. Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science*. 1998;280(5366):1077-1082.

50. Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res*. 2011;40:D109-D114. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3245020/. doi: 10.1093/nar/gkr988.

51. Kanehisa M, Goto S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*. 1999;28(1):27-30. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC102409/.

52. The Gene OC, Ashburner M, Ball CA, et al. Gene ontology: Tool for the unification of biology. *Nat Genet*. 2000;25(1):25-29. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3037419/. doi: 10.1038/75556.

53. Nishimura D. *BioCarta:Biotech software & internet report .june 2001, 2(3): 117-120. .* ; Volume: 2 Issue 3: July 5, 2004. http://online.liebertpub.com/doi/abs/10.1089/152791601750294344. doi:10.1089/152791601750294344.

54. Liberzon A, Subramanian A, Pinchback R, ThorvaldsdÃ³ttir H, Tamayo P, Mesirov JP. Molecular signatures database (MSigDB) 3.0. *Bioinformatics*. 2011;27(12):1739-1740. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3106198/. doi: 10.1093/bioinformatics/btr260.

55. Sboner A, Demichelis F, Calza S, et al. Molecular sampling of prostate cancer: A dilemma for predicting disease progression. *BMC Medical Genomics*. 2010;3:8-8. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2855514/. doi: 10.1186/1755-8794-3-8.

56. Demichelis F, Fall K, Perner S, et al. TMPRSS2:ERG gene fusion associated with lethal prostate cancer in a watchful waiting cohort. *Oncogene*. 2007;26(31):4596-4599. http://dx.doi.org/10.1038/sj.onc.1210237.

57. Johansson J, Andrén O, Andersson S,et al. NAtural history of early, localized prostate cancer. *JAMA*. 2004;291(22):2713-2719. doi: 10.1001/jama.291.22.2713.

58. Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*. 2005;102(43):15545-15550. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1239896/. doi: 10.1073/pnas.0506580102.

59. Tian L, Greenberg SA, Kong SW, Altschuler J, Kohane IS, Park PJ. Discovering statistically significant pathways in expression profiling studies. *Proc Natl Acad Sci U S A*. 2005;102(38):13544-13549. doi: 0506577102 [pii].

60. Goeman JJ, Buhlmann P. Analyzing gene expression data in terms of gene sets: Methodological issues. *Bioinformatics*. 2007;23(8):980-987. doi: 10.1093/bioinformatics/btm051.

61. Nam D, Kim SY. Gene-set approach for expression pattern analysis. *Brief Bioinform*. 2008;9(3):189-197. doi: 10.1093/bib/bbn001; 10.1093/bib/bbn001.

62. Dinu I, Potter JD, Mueller T, et al. Gene-set analysis and reduction. *Brief Bioinform*. 2009;10(1):24-34. doi: 10.1093/bib/bbn042 [doi].

63. Delongchamp R, Lee T, Velasco C. A method for computing the overall statistical significance of a treatment effect among a group of genes. *BMC Bioinformatics*. 2006;7:S11-S11. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1683577/. doi: 10.1186/1471-2105-7-S2-S11.

64. Chen JJ, Lee T, Delongchamp RR, Chen T, Tsai C. Significance analysis of groups of genes in expression profiling studies. *Bioinformatics*. 2007;23(16):2104-2112. doi: 10.1093/bioinformatics/btm310.

65. Dinu I, Potter JD, Mueller T, et al. Gene-set analysis and reduction. *Briefings in Bioinformatics*. 2008;10(1):24-34. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2638622/. doi: 10.1093/bib/bbn042.

66. Benjamini Y, Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B (Methodological)*. 1995;57(1):289-300. http://www.jstor.org/stable/2346101. doi: 10.2307/2346101.

67. Zeggini E, Weedon MN, Lindgren CM, et al. Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. *Science*. 2007;316(5829):1336-1341. doi: 10.1126/science.1142364.

68. National Human Genome Research Institute. http://www.genome.gov/10000533. Updated November 15, 2011. Accessed 02/01, 2015.

69. Holm S. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*. 1979;6(2):65-70. http://www.jstor.org/stable/4615733.

70. Westfall PH, Young SS. Resampling-based multiple testing. In: New York: John Wiley & Sons; 1993:2-4. https://books.google.ca/books?hl=en&lr=&id=nuQXORVGI1QC&oi=fnd&pg=PR17&dq=info:PtihG052vgUJ:scholar.google.com&ots=Xmb_CP45OM&sig=W5Jnl3l5L090IyqP0FbsMrrLkuU#v=onepage&q&f=false.

71. Huang J, Jochems C, Talaie T, et al. Elevated serum soluble CD40 ligand in cancer patients may play an immunosuppressive role. *Blood*. 2012;120(15):3030-3038. doi: 10.1182/blood-2012-05-427799 [doi].

72. Brunda MJ, Luistro L, Warrier RR, et al. Antitumor and antimetastatic activity of interleukin 12 against murine tumors. *J Exp Med*. 1993;178(4):1223-1230.

73. Kawasaki H, Onuki R, Suyama E, Taira K. Identification of genes that function in the TNF-alpha-mediated apoptotic pathway using randomized hybrid ribozyme libraries. *Nat Biotechnol*. 2002;20(4):376-380. doi: 10.1038/nbt0402-376 [doi].

74. Dhanasekaran SM, Barrette TR, Ghosh D, et al. Delineation of prognostic biomarkers in prostate cancer. *Nature*. 2001;412(6849):822-826. doi: 10.1038/35090585 [doi].

75. Lapointe J, Li C, Higgins JP, et al. Gene expression profiling identifies clinically relevant subtypes of prostate cancer. *Proc Natl Acad Sci U S A*. 2003;101(3):811-816. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC321763/. doi: 10.1073/pnas.0304146101.