Gene-Set Reduction for Analysis of Major and Minor Gleason Scores Based on Differential Gene-Set Expressions and Biological Pathways in Prostate Cancer

Cancer Informatics
Volume 16: 1–11
© The Author(s) 2017
Reprints and permissions:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/1176935117730016

Irina Dinu[1], Surya Poudel[1] and Saumyadipta Pyne[2]

[1]School of Public Health, University of Alberta, Edmonton, AB, Canada. [2]Indian Institute of Public Health, Public Health Foundation of India, Hyderabad, India.

ABSTRACT: The Gleason score (GS) plays an important role in prostate cancer detection and treatment. It is calculated based on a sum between its major and minor components, each ranging from 1 to 5, assigned after examination of sample cells taken from each side of the prostate gland during biopsy. A total GS of at least 7 is associated with more aggressive prostate cancer. However, it is still unclear how prostate cancer outcomes differ for various distributions of GS between its major and minor components. This article applies Significance Analysis of Microarray for Gene-Set Reduction to a real microarray study of patients with prostate cancer and identifies 13 core genes differentially expressed between patients with a major GS of 3 and a minor GS of 4, or (3,4), vs patients with a combination of (4,3), starting from a less aggressive GS combination of (3,3), and moving toward a more aggressive one of (4,4) via gray areas of (3,4) and (4,3). The resulting core genes may improve understanding of prostate cancer in patients with a total GS of 7, the most common grade and most challenging with respect to prognosis.

KEYWORDS: DNA microarrays, gene set analysis, gene set reduction, core subsets, Gleason scores

RECEIVED: February 15, 2017. **ACCEPTED:** July 26, 2017.

PEER REVIEW: Four peer reviewers contributed to the peer review report. Reviewers' reports totaled 1087 words, excluding any confidential comments to the academic editor.

TYPE: Original Research

FUNDING: The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: S. Pyne is supported by Ramalingaswami Fellowship from the Department of Biotechnology, India.

DECLARATION OF CONFLICTING INTERESTS: The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

CORRESPONDING AUTHOR: Irina Dinu, School of Public Health, University of Alberta, Edmonton, AB T6G 1C9, Canada. Email: idinu@ualberta.ca

Introduction

Prostate cancer is a heterogeneous disease, where some prostate cells no longer function as healthy cells, by losing normal control of growth and division. Gleason grading plays an important role in detection and treatment of prostate cancer.[1] 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 on the microscopic appearance of cancer cells in the prostate gland. A pathologist examines the biopsy specimen and attempts to give a score to the 2 patterns. The primary grade represents most of the tumor; it has to be greater than 50% of the total pattern seen. This is also called the major component of the Gleason score (GS). The secondary grade relates to the minority of the tumor; it has to be less than 50%, but at least 5% of the total pattern seen.[2] This is also called the minor component of the GS.

Gleason score is calculated as the sum of the major (primary) and minor (secondary) components, therefore ranging from 2 to 10. Higher GSs are more aggressive and have a worse prognosis. It has been long recognized that patients with a total GS ≥7 are at greater risk for prostate cancer outcomes.[3] 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.[4] Our goal is to identify genes and biological pathways' expressions different between patients with a major GS of 3 and minor GS of 4, or (3,4), vs those with a major GS of 4 and minor GS of 3, or (4,3), starting from a less aggressive combination (3,3) and moving toward a more aggressive combination (4,4).

Our strategy for analyzing microarray gene expression data is to focus on biological pathways, ie, 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. Gene-set enrichment analysis was the first method proposed for analysis of sets of genes differentially expressed between 2 conditions. An intensive review and methodological discussions are given by Nam and Kim.[5] The methods are falling into 2 categories: competitive methods testing the strength of the association of a gene set with the phenotype against other sets of same sizes and self-contained methods testing the association of one set with the phenotype. Methods in both categories rely on a randomization testing approach to calculate significance and address the small sample size, large gene-set problem. Competitive methods use permutations based on gene sampling, whereas self-contained methods use permutations based on subject sampling. We prefer the latter because it preserves correlations across genes in a set.

In this article, we use Significance Analysis of Microarray for Gene Sets (SAM-GS),[6] a method previously found to perform very well compared with 6 other self-contained methods. The performance was assessed in simulation studies of type I error and power, as well as applications to real data.[6,7] Another reason for using SAM-GS over other self-contained methods is its readily available extension. Significance Analysis of Microarray for Gene-Set Reduction (SAM-GSR)[8] is a method applied to extract core subsets, chiefly contributing to the significance of a set. The reasoning behind extracting core subsets is that not all the genes in a set contribute toward significance of a set. Significance Analysis of Microarray for Gene-Set Reduction identifies core subsets, by gradually retaining top-ranked genes and evaluating significance of the remaining subset. The ability of the method to identify core subsets was tested in simulations studies for a binary phenotype, as well as application to real microarray data.[8]

The rest of the article is organized as follows. In section "Methods," we describe the data from the Swedish Watchful Waiting Cohort, the gene sets and pathways catalog, as well as the 2 methods, SAM-GS and SAM-GSR. We also present our strategy of moving gradually from a less aggressive GS combination to a more aggressive one, to distinguish between patients with a major GS of 3 and minor GS of 4, vs patients with a major GS of 4 and a minor of 3. In sections "Results" and "Discussion," we present the results and discuss their implications.

Methods

Individual gene analysis

Individual gene analysis is a method for gene expression analyses focusing on identifying individual genes that exhibit difference between 2 states of interest. In response to challenging characteristics of microarray data, Significant Analysis of Microarray (SAM)[9] was proposed as an individual gene analysis method. Significant Analysis of Microarray is a moderated *t* test statistic, together with a false discovery rate (FDR) type of adjustment, calculated based on group-label (eg, 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 (eg, 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. Significant Analysis of Microarray is based on analyses of random fluctuations in the data and computes gene-specific *t*-like tests. Although 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 as follows:

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

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 as follows:

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

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; and 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 because 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.

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. Significance Analysis of Microarray for Gene Sets[6] combines the SAM *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 described above:

$$\text{SAM-GS} = \sum_{i=1}^{|S|} d_{(i)}^2$$

Statistical significance of *S* is obtained based on a phenotype label permutation test. The method can be summarized in a few 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 2 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 SAM-GS test statistic corresponding to set S:

$$\text{SAM-GS} = \sum_{i=1}^{|S|} d_{(i)}^2$$

3. Permute the labels of the phenotype and repeat steps (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 SAM-GS statistic from step (2) and its permutation distribution from step (3).

Gene-set reduction

Significance Analyses of Microarray for Gene-Set Reduction proposed by Dinu et al[8] was motivated by the fact that not all genes in a significant set are contributing to its significance. 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, $|d_i| > |d_j|$, suggest 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. Significance Analyses of Microarray for Gene-Set Reduction gradually partitions the entire set S, into 2 subsets, based on the principle above and evaluates their association with the phenotype. Significance Analyses of Microarray for Gene-Set Reduction 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, whereas $\bar{x}_2(i)$ is the average expression level for gene i in the control group; $s(i)$ is the pooled standard deviation of gene expression over the 2 groups of phenotype; s_0 is the small positive constant that adjusts for the small variability encountered in microarray data.

2. For $k = 1,\ldots,|S|-1$, 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, ie, $c_k > c$, we are protected against losing genes that are not significant by themselves, but they collectively form a set that is significant.[8]

Results

Data description

We used data from the Swedish Watchful Waiting Cohort with up to 30 years of clinical follow-up.[10,11] The data are nested in a cohort of men with localized prostate cancer diagnosed in the Örebro (1997-1994) and South East (1987-1999) Health Care Regions of Sweden. Eligible patients were identified through population-based prostate cancer quality databases maintained in these regions, which were described in detail in the study by Johansson et al[12] 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% of all potentially eligible cases. Messenger RNA expression of 6100 genes was measured on 255 patients, divided into 2 extreme groups: men who died of prostate cancer and men who survived more than 10 years of follow-up without metastases. These 2 groups are referred as lethal and indolent patients with prostate cancer. Clinical, pathological, and demographical characteristics of the 255 patients are given in Table 1. Prostate-specific antigen is not available in this cohort, as there were no screening programs in place at the time.

Biological pathways and gene sets from Molecular Signatures Database

An important aspect of microarray data analysis is accessing extensive collections of gene sets and properly linking them to gene expression data. Microarray studies typically result in long lists of genes, not always easy to interpret. Scientists put together lists of genes sharing a common biological function, ie, biological pathways. The analysis at the gene-set or pathway level improves on interpretation and reproducibility across studies. The Molecular Signatures Database (MSigDB)[13] available for download from http://www.broad.mit.edu/gsea is one of the most widely used repositories of knowledge expert–derived sets of genes and biological pathways. A growing number of databases store sets from gene expression signatures reported in the literature. Molecular Signatures Database differs from these resources in several aspects: (1) the catalog is formatted for gene-set analysis; (2) it covers a more diverse and wider range of gene-set resources and types, including original research publications and entire

Table 1. Clinical, pathological, and demographical characteristics of the 255 patients.

CHARACTERISTICS	COUNTS (%)	EXTREME GROUPS		FISHER EXACT TEST *P* VALUE	ODDS RATIO (95% CI)
		INDOLENT	LETHAL		
Gleason score					
<7	77 (30.2)	52	25		
7	104 (40.8)	46	58		
>7	74 (29.0)	8	66	$1.14*10^{-12}$	
Gleason combinations					
(3,3)	77 (37.5)	52	25		
(3,4)	71 (34.6)	36	35		
(4,3)	33 (16)	10	23		
(4,4)	24 (11.7)	7	17	$2.2*10^{-16}$	
Age					
≤70	77 (30.2)	39	38		
>70	178 (69.8)	67	111	.07	1.7 (0.95-3.02)
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	$9.02*10^{-11}$	
Not assessable	5 (2)				
ERG rearrangement status (fusion)					
Negative (0)	206 (80.8)	96	110		
Positive (1)	40 (15.7)	5	35	$3.64*10^{-5}$	6.07 (2.24-20.65)
Not assessable	9 (3.5)				
Extreme groups					
Lethal	149 (58.4)				
Indolent	106 (41.6)				
Survival status					
Alive	71 (27.8)				
Dead	184 (72.2)				

collections of sets derived from specialized resources; (3) MSigDB is built both through manual curation and by automatic computational means, whereas other databases emphasize only one of these approaches; and (4) the collection contains the largest number of gene sets overall. For our analyses, we used the MSigDB C2 catalog consisting of 1892 gene sets, representing metabolic and signaling pathways from online pathway databases, gene sets from biomedical literature including 786 scientific publications, gene sets compiled from published mammalian microarray studies, and gene sets defined by mining large collections of cancer-oriented microarray data.

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 vs nonlethal patients with prostate cancer. In total, 1351 genes out of 1892 MSigDB gene sets were

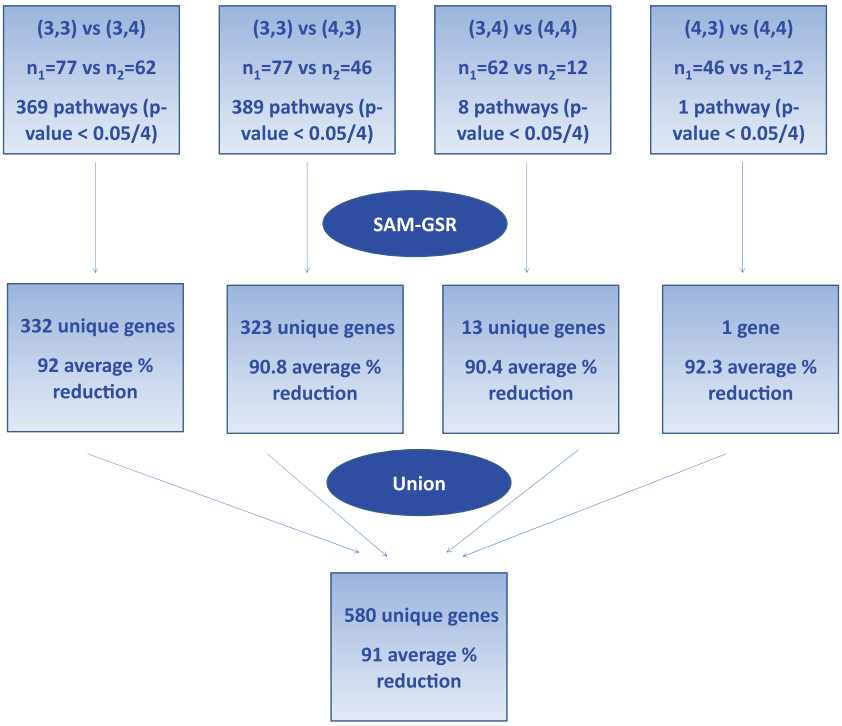


Figure 1. Gene-set reduction flowchart. SAM-GSR indicates Significance Analyses of Microarray for Gene-Set Reduction.

found to be differentially expressed between 149 lethal and 106 nonlethal patients with prostate cancer, using SAM-GS. Furthermore, 1246 gene sets were found to be differentially expressed between 80 patients with major and minor GSs ≤3 vs 68 patients with major and minor GS ≥4. Our goal was to compare biological pathways and gene sets across various combinations of major and minor GS components. There might be some overlapping sets differentiating across various combinations. However, we did not hypothesize that the same groups of genes would differentiate across all the combinations. Therefore, a union of unique core genes from all the combinations analyses is reported in Figure 1. The number of significant gene sets and core set sizes decreased 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) vs 62 patients with GS of (3,4) gives 369 gene sets significant at a *P* value of .05/4 = .0125. The Bonferroni adjustment corresponds to a total of 4 GS combinations, as described in Figure 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[14] is provided as a measure of adjustment for testing a large number of genes and is given by the expected proportion of false positives among all tests called significant. The FDR cutoffs for the 4 combinations are 0.006, 0.004, 0.27, and 0.95.

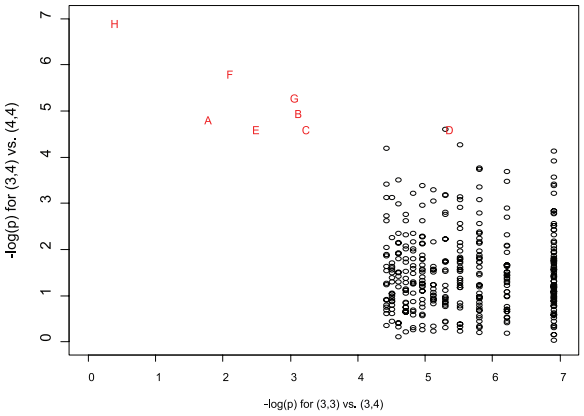
Significance Analyses of Microarray for Gene-Set Reduction achieved a 91% reduction, averaged over the 4 GS combinations, starting from (3,3) and ending with (4,4). The 369 gene sets differentiating between (3,3) and (3,4) were reduced to 332 unique genes shared across the core gene sets. The percent 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%. Moving from a less aggressive GSs combination (3,3) to a more aggressive combination (4,4), 580 unique genes were identified.

At the gene set–level analysis, only 1 of the 8 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 2 analyses are shown in Figure 2. The 8 pathways are represented as letters of the alphabet from A to H. Similarly, only 1 of the 8 pathways differentiating between (3,4) vs (4,4) is represented among the 389 pathways differentiating between (3,3) vs (4,3) (Figure 3).

There were 179 gene sets overlapping across the analyses of (3,3) vs (3,4) and (3,3) vs (4,3). At the gene level, there were 84 overlapping genes across the core genes differentiating between (3,3) vs (3,4) and (3,3) vs (4,3). These results are presented as Supplementary Material.

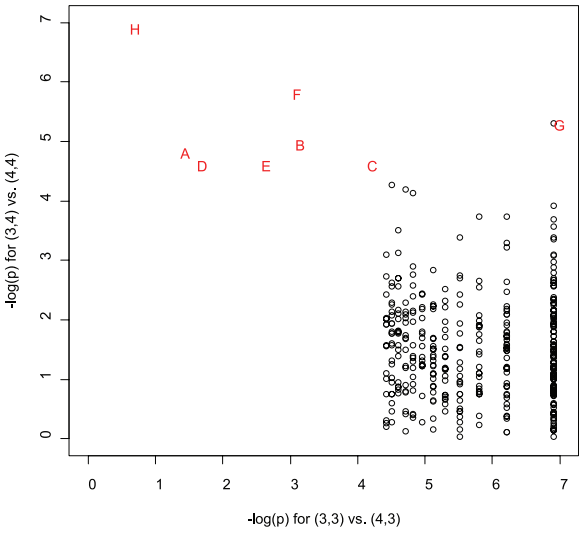
Gene-set reduction results for GS of (3,4) vs (4,3)

We performed a gene-set analysis and reduction for 62 patients with GS of (3,4) vs 46 patients with GS of (4,3). In total, 32 gene sets were identified at .05 significance level, with an FDR value of 0.75. The core sets of the 32 gene sets are presented in Table 2.



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 2. Negative log *P* values for gene sets differentially expressed between (3,4) vs (4,4) or (3,3) vs (3,4).
The 8 gene sets differentiating between (3,4) vs (4,4) are denoted as letters of alphabet as shown 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 3. Negative log *P* values for gene sets differentially expressed between (3,4) vs (4,4) or (3,3) vs (4,3).
The 8 gene sets differentiating between (3,4) vs (4,4) are denoted as letters of alphabet as shown below.

We compared the results of analysis of GS (3,4) vs (4,3) with results of analysis of GS ranging from (3,3) to (4,4). Significance Analysis of Microarray for Gene Sets *P* values of the 8 gene sets differentiating between (3,4) and (4,4) are presented in Table 3.

At the gene set–level analysis, only 1 of the 8 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 2 analyses are shown in Figure 4. The 8 pathways are represented as letters of the alphabet from A to H.

At the gene-level analysis, 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. Boxplots of some of these core gene expressions are presented in Figure 5. Although the boxplots show small differences, we need to keep in mind that the concept of gene set analysis was developed to address small but coordinated changes in gene expressions, across the set. The correlations across a gene set or biological pathway drive the association with the phenotype, even if the changes at the individual gene level are small.[5,6] Biological process and cellular component from Gene Ontology for core genes are presented in Table 5. 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 .059.

We also performed a global analysis of GSs of 6, 7, and 8. The results of the global analysis are shown in Tables 2 and 3. The global analysis resulted in 66% of the gene sets with *P* values less than or equal to .05 and 25% less than or equal to .001. This supports previous knowledge that GS of 7 and above are significantly different from GS of 6. However, the breakdown by major and minor components is needed to sort out the groups of patients where the differences occur. Most of the gene sets show significant *P* values in the global analysis, in agreement with differences occurring at some of the major and minor combinations. However, we note that some of the gene sets show lack of significance in the global analysis, despite significance in some of the analysis of the major and minor combinations. In Table 2, 9 out of 32 gene sets that are significantly different between (3,4) and (4,3) are not significant in the global analysis of multiple GSs. In Table 3, 2 out of 8 gene sets did not reach significance level in the global analysis, although they appear different in some

Table 2. Results of SAM-GS and SAM-GSR analyses for 62 patients with GS of (3,4) vs 46 patients with GS of (4,3), together with *P* values from a global gene set analysis of GSs of 6, 7, and 8.

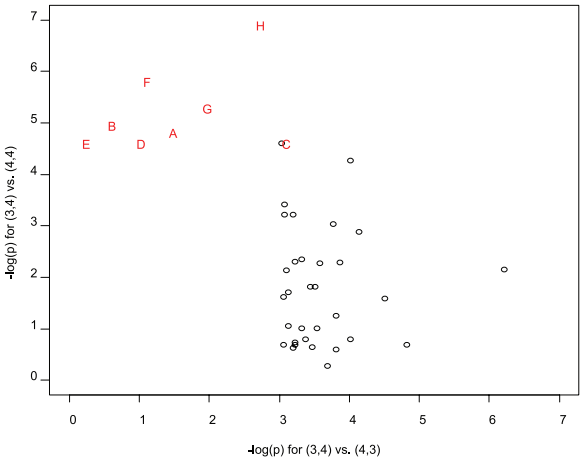
GENE-SET NAME	GENE-SET SIZE	SAM-GS, *P* VALUE	GLOBAL ANALYSIS OF GSS OF 6, 7, AND 8	SAM-GSR CORE GENES
AGED_MOUSE_HYPOTH_DN	28	.002	.002	*DNM1 FSTL1 APOE*
CD40PATHWAY[a]	9	.008	.365	*IKBKAP*
HSA05110_CHOLERA_INFECTION	23	.011	.027	*SEC61A1*
HEATSHOCK_YOUNG_UP	9	.016	<.001	*ANXA1*
NOUZOVA_CPG_METHLTD	22	.018	<.001	*EFNA5 EPHA5*
VEGF_HUVEC_2HRS_UP[a]	25	.018	.274	*APOE PPY*
HYPOPHYSECTOMY_RAT_DN	39	.021	<.001	*COL3A1 NPPA*
PENG_GLUCOSE_UP	32	.022	<.001	*OCLN*
LIAN_MYELOID_DIFF_TF	31	.022	.015	*BHLHB2 MYB NFKB1*
HSA00330_ARGININE_AND_ PROLINE_METABOLISM[a]	25	.023	.212	*ARG2*
ADIPOGENESIS_HMSC_ CLASS5_UP[a]	6	.025	.285	*MYB*
ONE_CARBON_POOL_BY_FOLATE	15	.028	.041	*SHMT2*
TNFR2PATHWAY	14	.029	.019	*IKBKAP*
UVC_HIGH_D9_DN	20	.03	.039	*NAP1L1*
HDACI_COLON_CLUSTER6[a]	24	.031	.317	*NAP1L1*
NDKDYNAMINPATHWAY[a]	15	.032	.112	*DNM1*
TYPE_III_SECRETION_SYSTEM	14	.034	.011	*ATP6V1C1*
ANDROGEN_GENES	43	.036	.013	*NR1I3*
GH_HYPOPHYSECTOMY_RAT_UP	10	.036	.042	*COL3A1*
ARGININE_AND_PROLINE_ METABOLISM	42	.04	.001	*MAOA*
FMLPPATHWAY[a]	30	.04	.103	*NFATC3*
HSA00670_ONE_CARBON_ POOL_BY_FOLATE	13	.04	.031	*SHMT2*
PHOTOSYNTHESIS	15	.041	.016	*ATP6V1C1*
HSA00051_FRUCTOSE_AND_ MANNOSE_METABOLISM	28	.041	.005	*MTMR6*
KIM_TH_CELLS_UP[a]	31	.044	.124	*ETS1*
GCRPATHWAY	16	.044	.001	*ANXA1*
HEARTFAILURE_ATRIA_UP	20	.045	.051	*FKBP8*
ALZHEIMERS_INCIPIENT_DN	88	.046	<.001	*UROS*
GAMMA.UV_FIBRO_UP	25	.046	.005	*IL10RB*
AGUIRRE_PANCREAS_CHR8	28	.047	.002	*HAS2*
GH_GHRHR_KO_24HRS_DN	73	.047	.013	*IFNAR1*
FERRANDO_CHEMO_ RESPONSE_PATHWAY[a]	9	.048	.329	*DTYMK*

Abbreviations: GS, Gleason score; SAM-GS, Significance Analysis of Microarray for Gene Sets; SAM-GSR, Significance Analyses of Microarray for Gene-Set Reduction.
[a]Gene sets not significant in the global analysis of GS of 6, 7, and 8, although significant in the analysis of (3,4) vs (4,3) combinations.

Table 3. SAM-GS *P* values for various distributions of Gleason scores, together with *P* values from a global gene set analysis of Gleason scores of 6, 7, and 8.

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)	GLOBAL ANALYSIS OF GLEASON SCORES 6, 7, AND 8
BUT_TSA_UP	18	.179	.254	.008	.174	.24	.047
CMV_HCMV_	36	.047	.046	.007	.069	.574	.005
TIMECOURSE_14HRS_DN							
FERRANDO_CHEMO_ RESPONSE_PATHWAY	9	.042	.016	.01	.045	.048	.329
HDACI_COLON_ CUR24HRS_UP	27	.005	.2	.01	.069	.383	.024
LEE_CIP_UP	50	.088	.076	.01	.066	.834	.002
TSA_PANC50_UP	29	.128	.048	.003	.029	.346	.001
UEDA_MOUSE_SCN	58	.05	.001	.005	.228	.15	.011
UREACYCLEPATHWAY	7	.721	.536	.001	.016	.07	.155

Abbreviation: SAM-GS, Significance Analysis of Microarray for Gene Sets.



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. Negative log *P* values for gene sets differentially expressed between (3,4) vs (4,4) or (3,4) vs (4,3).
The 8 gene sets differentiating between (3,4) vs (4,4) are denoted as letters of alphabet as shown below.

of the analyses of the major and minor combinations. These differences may be caused by the fact that the overall scores are collapsed over major and minor combinations and, after further validation, may provide some insights into the $3+4 \neq 4+3$ prostate cancer hypothesis.

Discussion

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 components 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 GSs, moving from a less aggressive combination of (3,3) and toward 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 Microarray for Gene Sets 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. Significance Analysis of Microarray for Gene Sets was found to perform well in comparative studies of 7 self-contained gene-set analysis methods.[8] 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 Microarray for Gene Set Reduction was designed to extract core genes that contribute to

Table 4. 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 GENES	
BUT_TSA_UP	18	.008	1	*GADD45A*	
CMV_HCMV_	36	.007	2	*ETV1 APEX1*	
TIMECOURSE_14HRS_DN					
FERRANDO_CHEMO_ RESPONSE_PATHWAY	9	.01	1	*CDA*	
HDACI_COLON_	27	.01	3	*RPN2 ALDOA CCND1*	
CUR24HRS_UP					
LEE_CIP_UP	50	.01	2	*ETV1*	*COL4A2*
TSA_PANC50_UP	29	.003	2	*BIK*	*NOTCH3*
UEDA_MOUSE_SCN	58	.005	2	*GADD45A*	*SMPDL3A*
UREACYCLEPATHWAY	7	.001	2	*CPS1*	*ASL*

Abbreviations: SAM-GS, Significance Analysis of Microarray for Gene Sets; SAM-GSR, Significance Analyses of Microarray for Gene-Set Reduction.

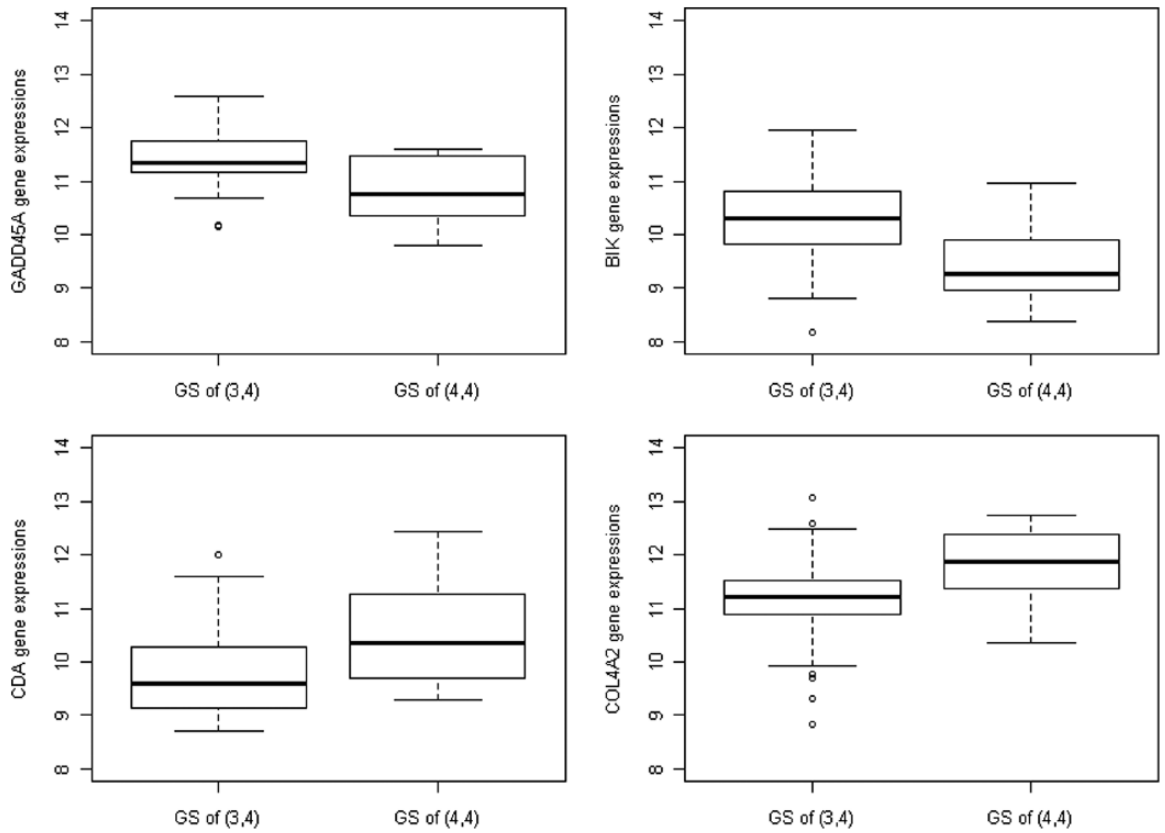


Figure 5. Boxplots of core gene expressions among the 13 genes differentiating between GS of (3,4) and GS of (4,4). GS indicates Gleason score.

the significance of the whole set. We reason that these 2 methods are appropriate for analyzing differences at gene and gene-set levels across various combinations of GSs.

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.[15] The CD40 pathway has also been shown to play a crucial role in production of cytokines, which modulate the function of T lymphocytes in antitumor responses.[16] TNFR2 pathway was also differentially expressed between GS of (3,4) vs (4,3). TNFR2 is a receptor of tumor necrosis factor, a multifunctional pro-inflammatory cytokine.

Table 5. Biological process and cellular component from Gene Ontology for core genes from SAM-GSR analyses for 62 patients with GS of (3,4) vs 12 patients with GS 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[a]	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[a]	Angiogenesis, endodermal cell differentiation, cellular response to transforming growth factor β 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[a]	Translation, cellular protein modification process, cellular protein metabolic process, response to drug, posttranslational protein modification	Autophagosome membrane, nucleus, integral component of membrane
SMPDL3A[a]	Sphingomyelin catabolic process	Extracellular space, extracellular exosome

Abbreviation: GS, Gleason score; SAM-GSR, Significance Analyses of Microarray for Gene-Set Reduction.
[a]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.

Members of the tumor necrosis factor receptor superfamily can send both survival and death signals to cells.[17]

Urea cycle pathway was differentially expressed between GS of (3,4) vs (4,4), *P* value of .001, and GS of (4,3) vs (4,4), *P* value of .016; marginally significant for GS of (3,4) vs (4,3), *P* value of .07; and not significant for GS of (3,3) vs (3,4), *P* value of .721, or (3,3) vs (4,3), *P* value of .536. In urea cycle pathway, the enzyme ornithine decarboxylase converts the metabolite ornithine to putrescine. Ornithine decarboxylase has previously been found as overexpressed in prostate cancer[18] and is the target of the chemotherapeutic agent difluoromethylornithine.[19]

Acknowledgements

The authors thank the reviewers and editors for their thoughtful comments and suggestions, which improved the manuscript substantially.

Author Contributions

ID and SPy identified the research gap and formulated the research question and provided biological interpretations of the analysis results. ID and SPo performed the data analysis and programming. The manuscript was primarily written by ID and critically reviewed by all authors. All authors read and approved the final manuscript.

Availability

Free R-codes to perform gene-set analysis and reduction for binary phenotypes are available at http://www.ualberta.ca/~yyasui/homepage.html.

REFERENCES

1. Poudel SP, Pyne S, Dinu I. Analysis of major and minor Gleason scores based on differential gene expressions of biological pathways in prostate cancer. Poster presentation at: Annual meeting of the INSIGHTS: A focus on public health research; 2013. http://citation.allacademic.com/meta/p549887_index.html.
2. 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 Prostatic Dis*. 2012;15:75–86.
3. Pierorazio PM, Walsh PC, Partin AW, Epstein JI. Prognostic Gleason grade grouping: data based on the modified Gleason scoring system. *BJU Int*. 2013;111:753–760.
4. 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;167:2440–2442.
5. Nam D, Kim SY. Gene-set approach for expression pattern analysis. *Brief Bioinform*. 2008;9:189–197.
6. Dinu I, Potter JD, Mueller T, et al. Improving gene set analysis of microarray data by SAM-GS. *BMC Bioinformatics*. 2007;8:242
7. Liu Q_, Dinu I, Adewale AJ, Potter JD, Yasui Y. Comparative evaluation of gene-set analysis methods. *BMC Bioinformatics*. 2007;8:431.
8. Dinu I, Potter JD, Mueller T, et al. Gene-set analysis and reduction. *Brief Bioinform*. 2008;10:24–34.
9. Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A*. 2001;98:5116–5121.

10. Sboner A, Demichelis F, Calza S, et al. Molecular sampling of prostate cancer: a dilemma for predicting disease progression. *BMC Med Genomics*. 2010;3:8.
11. Demichelis F, Fall K, Perner S, et al. TMPRSS2:ERG gene fusion associated with lethal prostate cancer in a watchful waiting cohort. *Oncogene*. 2007;26:4596–4599.
12. Johansson J, Andrén O, Andersson S, et al. Natural history of early, localized prostate cancer. *JAMA*. 2004;291:2713–2719.
13. Liberzon A, Subramanian A, Pinchback R., Thorvaldsdóttir H, Tamayo P, Mesirov JP. Molecular Signatures Database (MSigDB) 3.0. *Bioinformatics*. 2011;27:1739–1740.
14. Storey JD, Tibshirani R. Statistical significance for genome wide studies. *Proc Natl Acad Sci U S A*. 2003;100:9440–9445.
15. Huang J, Jochems C, Talaie T, et al. Elevated serum soluble CD40 ligand in cancer patients may play an immunosuppressive role. *Blood*. 2012;120:3030–3038.
16. Brunda MJ, Luistro L, Warrier RR, et al. Antitumor and antimetastatic activity of interleukin 12 against murine tumors. *J Exp Med*. 1993;178:1223–1230.
17. 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;20:376–380.
18. Dhanasekaran SM, Barrette TR, Ghosh D, et al. Delineation of prognostic biomarkers in prostate cancer. *Nature*. 2001;412:822–826.
19. 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;101:811–816.