BMC Bioinformatics

Research article

Meta-analysis of several gene lists for distinct types of cancer: A simple way to reveal common prognostic markers

Xinan Yang* and Xiao Sun

Address: State Key Laboratory of Bioelectronics, Southeast University, 210096 Nanjing, P.R.China

Email: Xinan Yang* - xnyang@seu.edu.cn; Xiao Sun - xsun@seu.edu.cn

* Corresponding author

Published: 6 April 2007

BMC Bioinformatics 2007, **8**:118 doi:10.1186/1471-2105-8-118

Received: 16 October 2006

Accepted: 6 April 2007

This article is available from: http://www.biomedcentral.com/1471-2105/8/118

Abstract

Background: Although prognostic biomarkers specific for particular cancers have been discovered, microarray analysis of gene expression profiles, supported by integrative analysis algorithms, helps to identify common factors in molecular oncology. Similarities of Ordered Gene Lists (SOGL) is a recently proposed approach to meta-analysis suitable for identifying features shared by two data sets. Here we extend the idea of SOGL to the detection of significant prognostic marker genes from microarrays of multiple data sets. Three data sets for leukemia and the other six for different solid tumors are used to demonstrate our method, using established statistical techniques.

Results: We describe a set of significantly similar ordered gene lists, representing outcome comparisons for distinct types of cancer. This kind of similarity could improve the diagnostic accuracies of individual studies when SOGL is incorporated into the support vector machine algorithm. In particular, we investigate the similarities among three ordered gene lists pertaining to mesothelioma survival, prostate recurrence and glioma survival. The similarity-driving genes are related to the outcomes of patients with lung cancer with a hazard ratio of 4.47 ($p = 0.035$). Many of these genes are involved in breakdown of EMC proteins regulating angiogenesis, and may be used for further research on prognostic markers and molecular targets of gene therapy for cancers.

Conclusion: The proposed method and its application show the potential of such meta-analyses in clinical studies of gene expression profiles.

Background

Changes in gene expression levels could reflect clinically distinct conditions. Genome-wide perspectives of gene expression can now be obtained, and these can be combined with other currently-used criteria to identify predictors of clinical outcome for specific cancers [1-7]. Also, distinct gene expression profiles can reportedly determine molecular treatment responses, e.g. in cancer [8]. Thus it is possible to discover biomarkers from gene expression profiles that help to predict outcomes, and this emphasizes the need in biomedical research to combine results from similar experiments in order to identify diagnostic or prognostic disease markers.

Much recent research has confirmed that microarray results are comparable among different laboratories, especially when a common platform and a set of procedures are used [9-13]. Integrative analysis that evaluates cancer transcriptome data in the context of data from other sources has received attention recently (reviewed by

Rhodes and Chinnaiyan [14]). An important emerging argument concerns the uniformity of cancer metastases as well as the evolution of malignancy in primary tumors [15-17]. Grutzmann et al. ran meta-analysis on four studies for pancreatic cancer, and validated their identified signatures using RT-PCR and immunohistochemistry [13]. In particular, Glinsky and colleagues innovatively published a 11-gene signature that is displayed consistently in stem cells self-renewal pathways, and this is a powerful predictor for prognosis in 11 distinct types of cancer [17]. These results exemplify the clinical application of meta-analysis signatures detected in different cancer stages or types. Rhodes et al. [18] presented a comprehensive investigation of 40 data sets. They identified a robust signature of a set of differentially expressed genes when cancer and normal tissues were compared. A recent study [19] identified lists of differentially regulated genes that also significantly overlap with genes regulated by the tumor suppressors p16 and pRB. This work helps to translate genome-wide expression analyses into clinically useful cancer markers. Meta-analysis is a powerful tool for identification and validation marker genes in above studies [13,18]. However, in these studies, meta-signatures are identified on the basis of the individual genes used for analysis. Segal et al. [20] divided genes into sets and reported that certain sets show coherent behavior across a diverse group of clinical conditions. Another recent publication compared gene expression in two conditions to generate a gene list for each study, and then detected significant Similarities of Ordered Gene Lists (SOGL) [21] from different studies. The above two approaches extend the determination of significance from single study analysis to meta-analysis.

However, none of the above studies involving multiple cancers mentions independent prediction, which is a key bridge between molecular knowledge and clinical application. In particular, the SOGL approach can detect similarities between two gene lists, irrespective of significant differences between them, because it does not rely on differential gene expression in each single list having strong effects, but rather on consistent changes across multiple lists. SOGL is similar to other non-parameter statistical tests, except that it uses different weighting schemes for ranks. The ideal is to give higher weights to the genes which expressed more differentially, and to sum all the weighted orders to quantify the similarity. This approach allows the significance of similarity to be decided during meta-analysis and identifies the genes responsible for the similarity. In contrast to previous methods, SOGL does not depend on the definition of a particular "significance" threshold for a single study. Thus it is superior to other methods for detecting signatures in studies with weak effects or small sample sizes.

However, the similarities among gene lists are not guaranteed to be transferable [21]. With the discovery of common cancer signatures, there is a need to extend the method to several rather than two lists. Therefore, to meta-analyze many microarray profiles together, and to analyze the problem of outcome in highly noisy data, we have developed and implemented the SOGL method in this paper, extending it from the comparison of two gene lists to the comparison of multiple gene lists, which is useful for meta-analysis of microarray data. When the gene lists show similarity, we ask whether the similarity-driving genes improve the predictive power of a single study. To this end, we implement SOGL in two ways. One is to compare the accuracy of prediction by meta-analysis with that of individual analysis, which has already been successfully demonstrated for multiple cancer microarray data sets [11]. The other is to compare the traditional classical highest t-score with SOGL in selecting variables for classification, which has not been used in the context of cross-validation and class prediction. Finally, we discuss the predictive capacity of the similarity-driving genes detected in three solid tumors, and prove its success on another independent cancer data set.

Results

Our major aim was to identify biological mechanisms, common to different kinds of cancer that involve genes and gene expression changes inducing poor outcomes, e.g. metastasis, recurrence and short-term survival. We assumed that such mechanisms may be revealed by gene expression profiles. We collected nine recently-published microarray data sets related to clinical outcomes (for details see Table 1). For meta-analysis, we developed SOGL from a test for two gene lists to a test for multiple gene lists, since the similarities among gene lists are not guaranteed to be transferable. In this section, we first performed a meta-analysis allowing common samples across data sets to generate artificial similarity and to identify it using SOGL. Then we turned to six data sets on solid tumor for discovery of similarity and its contributing genes. All the data sets were pre-processed independently for background correction, normalization, summarization and quality assessment using an Affymetrix platform pre-processing protocol. We adopted the methods for stabilizing variance to normalize these raw profiling files on an additive scale in the nine collected data sets, using the R package *compdiagTools*.

Using SOGL on Leukemia studies

The data set described by Ross [22] used a relatively newly designed microarray platform with 132 representative cases from another data set with 327 cases [23]. Therefore a significant similarity between the gene lists generated from these two data sets were expected. Adding Another data set on leukemia outcome, we applied SOGL to com-

Table 1: Clinical information about the microarray studies we collected

Studies			samples with outcome notation			
study ID	cancer	#sample	N	#good	#poor	ratio
A [3]	breast	37	37	19	18	0.49
B [3]	breast	52	52	34	18	0.35
L [27]	lung	203	126	117	9	0.07
M [5]	mesothelioma	31	17	8	9	0.47
P [38]	prostate	102	21	13	8	0.38
G [65]	glioma	42	18	8	10	0.44
L1 [64]	T-cell leukemia	30	13	7	6	0.46
L1 [22]	pediatric leukemia	132	93	71	12	0.13
L2 [23]	pediatric leukemia	327	245	201	44	0.18

parison of more than two gene lists. Thus we performed the meta-analysis allowing partially common samples to generate an "artificial" similarity. However, finding a similarity in gene lists between samples run on different platforms is not our interest as many programs would find this. The question we addressed here is to evaluate whether our method improves the accuracy of prediction from individual studies when there is significant similarity.

First, we separately analyzed all the data sets for differential gene expression between the good and the poor outcome groups. Differential expression was quantified using fold change and z-statistic respectively, and the result was used as the effect size for meta-analysis [24]. The later measurement is a moderated t-score with a fudge factor [25] and is expected to be more reliable. The significance of differential expression of a gene between outcome groups was estimated by comparison with the sizes of random effects in perturbed data. None of the three studies individually displayed strong evidence for differential expression; but while individual studies failed to identify signatures that might reliably distinguish between conditions, meta-analysis succeeded. All data sets ordered the genes, each beginning with the most markedly up-regulated genes in the poor outcome group and ending with the most markedly down-regulated ones. Matching of the probe sets between Affymetrix Hgu95av2 and Hgu133a, resulted in 10507 best-matched transcripts. These gene expression profiles revealed significant similarity in the outcome conditions of the three leukemia studies. Figure 1 shows the significance of this similarity. An empirical p-value = 0.004 (permutation times B = 1000, each based on permutation of gene ranks to estimate random similarity scores) was detected for an optimal α^* which focused only on the first 150 genes in the orders. The significant similarity (p-value = 0.002) could also be observed when our method focused on the first 100 genes in the orders using z-statistic as effect size.

This led us to expect that variable selection by SOGL would improve the predictive capacity when the gene orders are significantly similar. For each subset of samples, we kept the number of transcripts selected by the highest t-scores exactly the same as the number of major intersection transcripts identified by SOGL method, while letting α = 0.015 to count the highest and lowest 750 items in the sets. The range of genes in common between the sets reflects the degree of similarity. For the comparison of gene orders in the three leukemia outcomes, we iterated 3-fold cross-validation together with support vector machine (SVM) algorithm D (= 500) times. The 75th and 25th percentiles of the numbers of selected genes are 75 and 48. The median is 61. Any increase in sensitivity will be accompanied by decrease in specificity, so to evaluate the predictive accuracy of the SOGL-selected genes and that of the highest traditional t-score, we drew ROC curves for both comparisons. Figure 2 shows the ROC curves generated from the leukemia studies. The SOGL curve follows the left-hand border and then the top border of the ROC space more closely, suggesting that the test is more accurate. In contrast, the highest t-score curve comes closer to the 45-degree diagonal of the ROC space, implying a less accurate test. The area under the ROC curve (AUC) is 0.73 (95% confidence interval (CI) 0.64–0.76) for SOGL using z-statistic as effect size, while 0.69 (95% CI 0.63–0.71) for the highest t-score, indicating SOGL tends toward more accurate than the highest t-score if gene lists are significant similar. In the same way, we observed no different AUC between the results of SOGL using fold changed effect size and highest t-score, that was 0.64 (95% CI 0.64–0.68) for SOGL, 0.63 (95% CI 0.57–0.69) for the highest t-score, suggesting that the improvement of prediction by SOGL is limited to highly significant similarity.

Study on different solid tumors

We then set out to determine the significant similarity among gene lists of different tumor outcomes. We needed

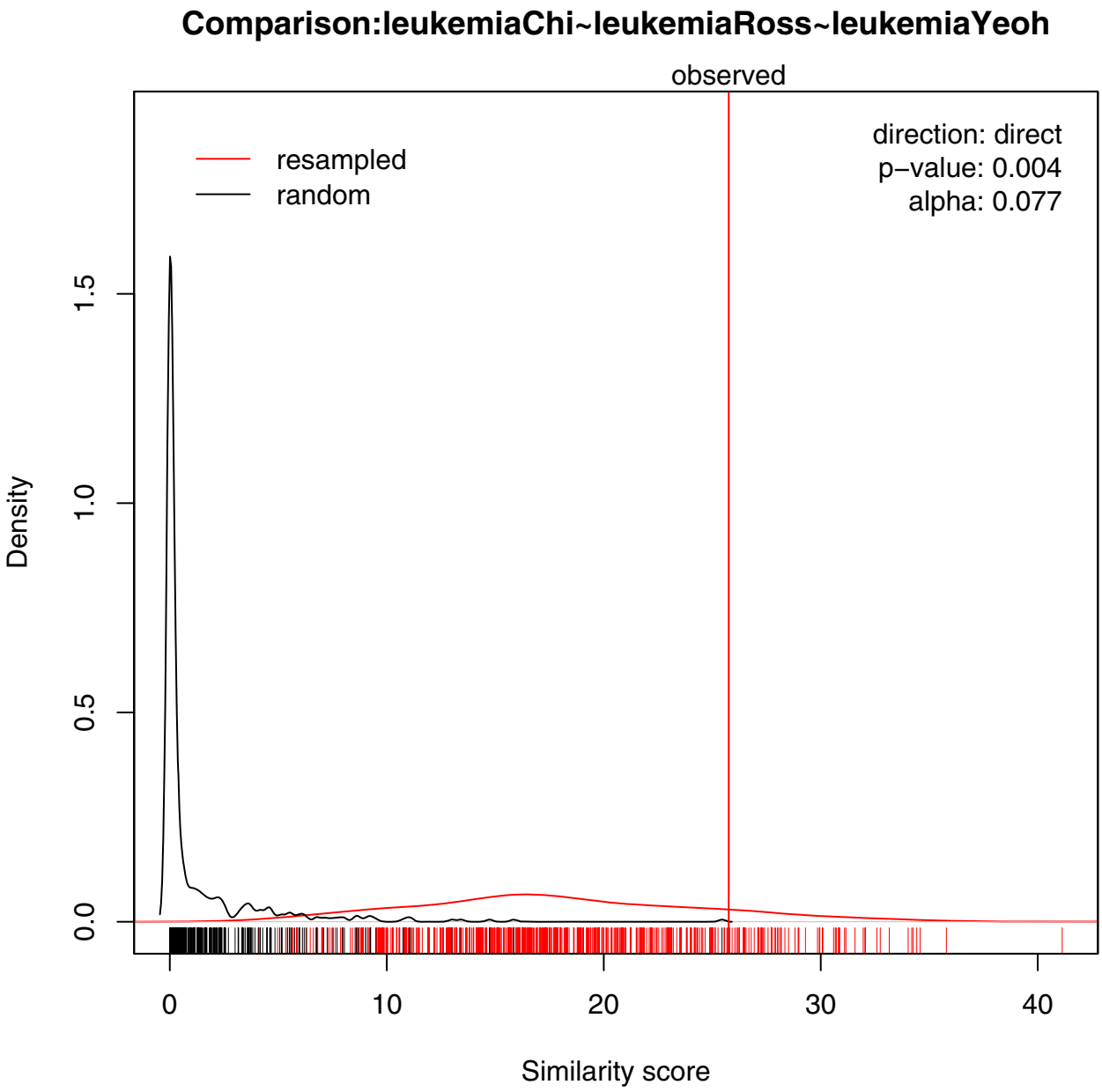

Figure 1
Similarity scores for leukemia outcome. The similarity of three gene orders for leukemia studies. In the plot, the red curve corresponds to estimated scores and the black curve to simulated random scores. These are kernel density estimates of the two-score distributions underlying the pAUC-score for optimal α^*. The vertical red line denotes the observed similarity score. The bottom rugs mark the simulated values.

confirmation first that the clinical diagnostic problem addressed here in regard to different kinds of cancer achieves similarity and improves the accuracy of prediction. To address this problem, we investigated six gene lists for comparing cancer outcomes, which are labeled A, B, L, M, P and G in Table 1. Figure 3 shows that 21 of the 57 possible comparisons from these gene lists show significant ($p < 0.05$) similarity for a pre-defined finite grid

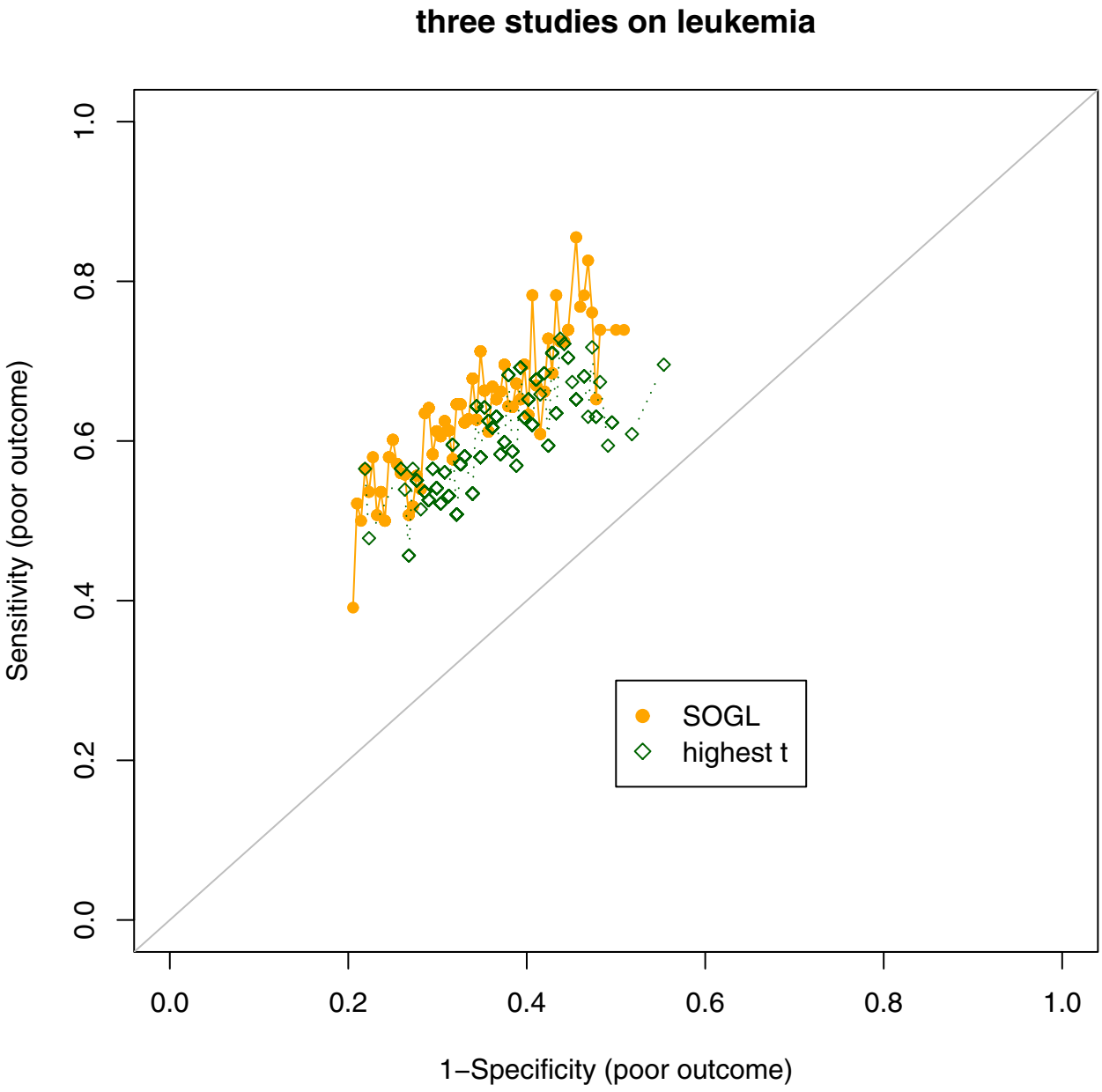

Figure 2
Comparison of methods using leukemia data. The ROC points for 500 prediction runs. Two points are generated for each time: the solid circle is the result from SOGL, and the diamond is the result from the same number of highest t-scores.

of parameter choices $\alpha \in [0.3, 0.01]$ spanning the leading 400–1500 items in order. These comparisons include the gene lists referring to:

• Recurrence of breast cancer and lymph node status of breast cancer;

• The same two lists, and neuroendocrine of lung cancer;

• Survival of mesothelioma and glioma, and recurrence of prostate;

• The above set of lists and the lymph node status of breast cancer or neuroendocrine of lung cancer;

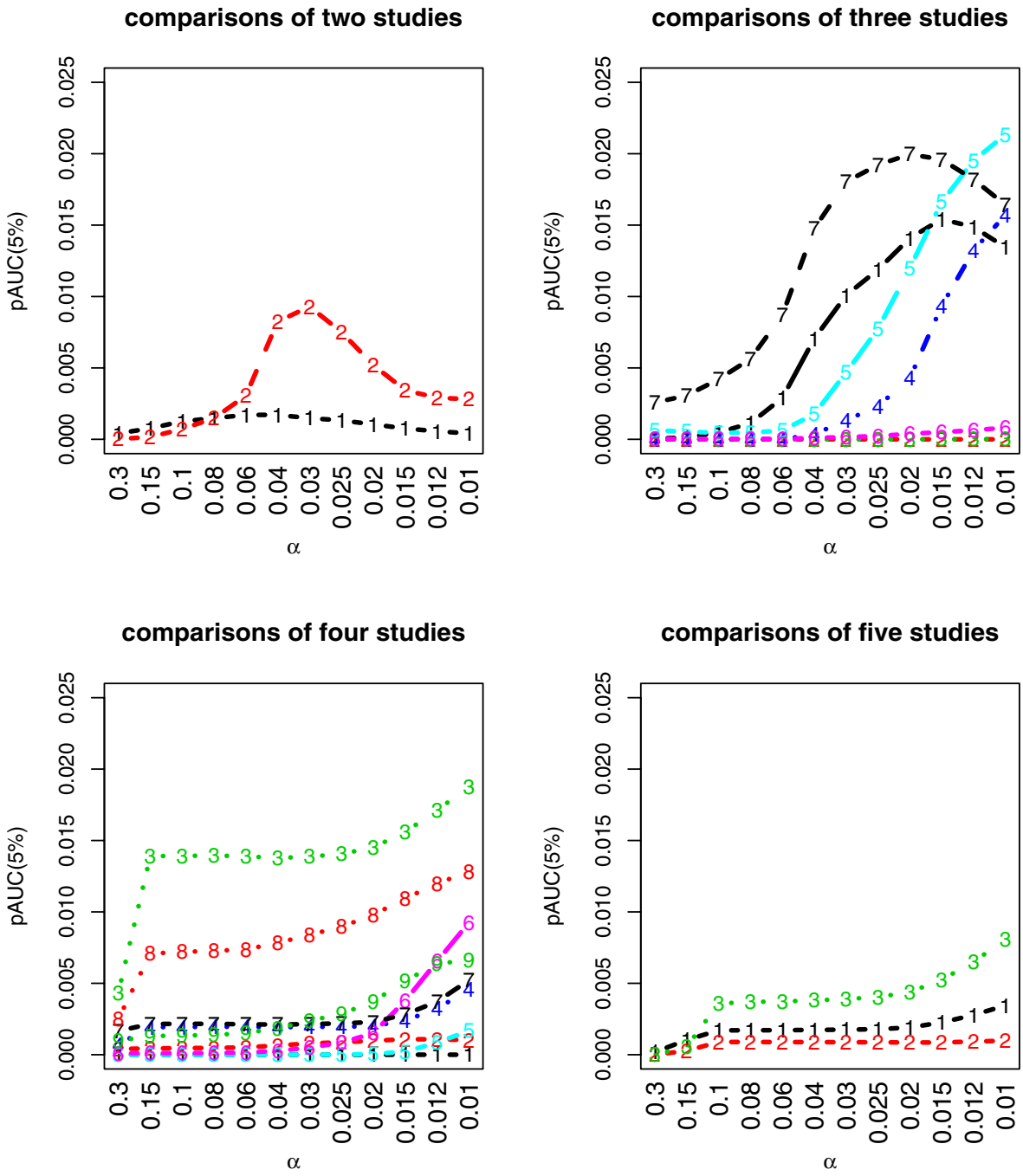

Figure 3

Similar comparisons among 6 solid tumors. 21 comparisons of gene lists show similar with separation between signal and noise. In the plot, α is given on the x axis and the pAUC-score for the randomized and alternative scores on the y axis. The pAUC test detects the difference between the distributions of alternative scores and random scores to select an optimized α^*, which reaches a highest value for each comparison. We iterate a sub-sample strategy C (= 500) times to obtain an estimation of the variability of the similarity score and the random score. Each time, by bootstrapping 80% the labels of patients, we obtain the alternative effect size (signals). And by shuffling these labels of patients, we calculate the background noise of the same size. The details of the similarities are given in Table 4.

• and others.

All the above sets of gene lists achieved higher pAUC (partial area under curve) scores [26] than most other comparisons. A pAUC-score evaluates the degree of overlap between two distributions. Note that a higher pAUC-score shows a greater likelihood that the estimated SOGL scores exceed chance in our method, and a larger α indicates more similarities at the higher ends of the gene lists. This finding supports the emerging notion that when prognosis is poor, there are commonalities among distinct types of cancer in the dysregulation of gene expression, implying that poor prognosis is sometimes independent of the original cancer type. In contrast, this kind of similarity was not so significant when more than 4 of the studies we collected were compared, demonstrating that the similarities spanning tumor tissues are limited.

The similarity among gene lists for glioma, prostate and mesothelioma outcomes

Comparison of the ordered gene lists generated from the outcomes for glioma, prostate and mesothelioma typically shows significant similarity; and significance (B = 1000, $p < 0.05$) is found for all the pre-defined finite grids of observed orderings [100, 1500]. It means that even for the highest orderings (biggest α values), the numbers of genes common to these three orders are not due to chance. Figure 4 shows the significance of the similarity. An empirical p-value = 0.024 is obtained for an optimal α^* focus on the highest 750 items in the order. To compare the accuracy of prediction by using SOGL as variable selection method to traditional highest-t-statistic, we iterated 3-fold cross validation D (= 500) times. The resulting 75th and 25th percentiles of the number of selected transcripts are 35 and 20. The median is 26. The superiority of SOGL is observed when the three solid tumor studies are integrated (for details see Figure 5). The area under the ROC curve is 0.747 (95% CI 0.709–0.774) for SOGL, 0.665 (95% CI 0.634–0.702) for the highest t-score. This proves that adopting SOGL for variable selection improves the predictive capacity when the gene lists involved are significantly similar. A similar improvement was observed when we examined a range of observed orderings [100, 1500].

We then turned to investigate the genes contributing to this similarity that were relevant to the survival of mesothelioma and glioma and the recurrence of prostate cancer. Table 2 shows the ranks and the symbols of these similarity-driving genes. The definition of "effect size" will affect the SOGL results and the identified genes. The genes identified by fold-change as effect size of SOGL yielded in 17 transcripts; 5 transcripts were reported if a moderated t-score with a fudge factor (also called as z-statistic) [25] was adopted as SOGL effect size. Four of these were identified by both approaches. Since our fold-change statistic is based on variance-stabilized data, it should generate a result similar to the t-statistic.

Nevertheless, the z-statistic puts less weight on variances than a classical t-statistic. These genes contain a high proportion of known prognostic marker genes and represent biological processes involved in tumor progression and metastasis. To evaluate over-representation of GO annotations from gene lists that were calculated from specific microarray (Affymetrix Hgu95av2), we ran hypergeometric tests to compute p-values. It evaluates the likelihood that the corresponding number of annotations is occurring in a random list of genes of the same size. Interestingly, 4 of them are genes for the human extracellular matrix (ECM)-receptor interaction pathway (hypergeometric test p = 1e-6), namely COL4A1, COL1A2, COL5A2 and FN1. Moreover, 7 of our short-list of 13 genes encode ECM proteins and regulators of ECM assembly, namely FN1, BGN, POSTN, COL4A1, COL11A1, COL1A2 and COL5A2. The other 5 genes have roles in angiogenesis: ANXA2, CPE, MDK, IGFBP3, and 3 transcripts of PTGDS. Although ANXA2 (annexin A2) is a substrate for a variety of protein kinases, and plays an important role in plasmin regulation and in cancer cell invasiveness and metastasis, ANXA2P3 (annexin A2 pseudogene 3) is a novel marker not being previously reported. We discuss these genes in more detail in a later section.

Validation of similarity-driving genes in the outcomes of three cancers on lung cancer data

We have found that the neuroendocrine differentiation was significantly similar to the three gene lists M, P and G (comparison ID 4_9 in the Table 4 and Figure 3). And Bhattacharjee et al. reported that the C2 neuroendocrine differentiation was associated with good outcome [27]. We therefore tried to establish the utility of the 5-transcript signature for M-P-G similarity on the outcome of patients with lung cancer. We expected that the 5-transcript signature was related to the lung cancer interpreted as neuroendocrine if its change statistically relevants to cancer development. On the other hand, we did not expect a strong power to predict the outcomes, because many non-C2 adenocarcinoma patients have short survival times. To this end, we divided the 125 lung cancer patients into two outcome groups. We employed a robust K-means classification method, *Pam* [28] calling R package *cluster*, to partitions (clusters) the data into 2 clusters around medians using the 5-transcript signature. We then used a Cox proportional hazards regression model (calling the R package *survival*) to explore the relationship between the pam-predicted conditions and clinical survival. The estimated hazard ratio defined by our 5-transcript signature was 4.47 (p = 0.035). As Figure 6 shows, the median survival after therapy in the poor-prognosis

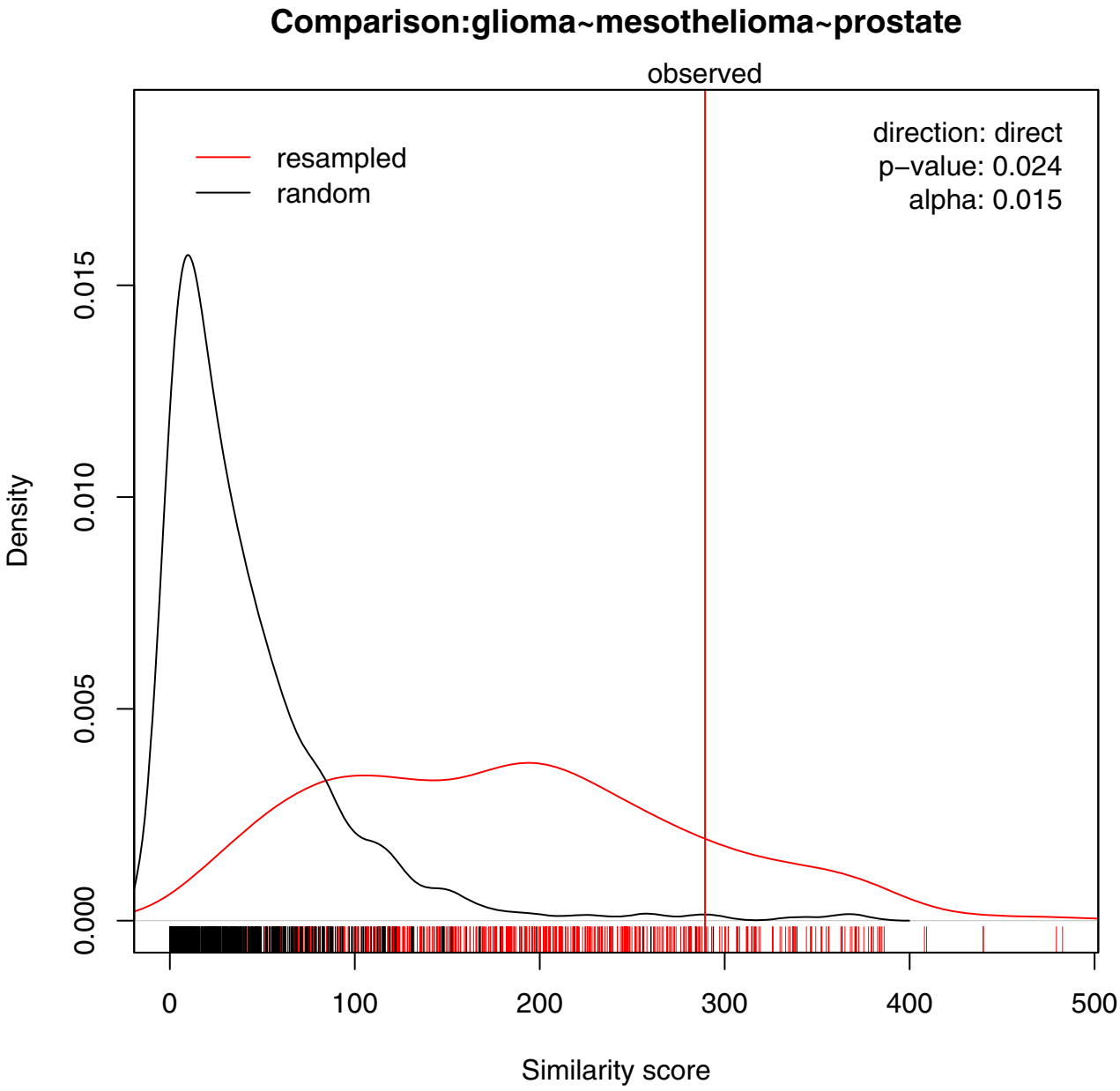

Figure 4
Similarity score for solid tumors. The similarities among three gene orders for different solid tumors. In the plot, the red curve corresponds to simulated observed scores and the black curve to simulated random scores. These are kernel density estimates of the two score distributions underlying the pAUC-score for optimal α^*. The vertical red line denotes the observed similarity score. The bottom rugs mark the simulated values.

subgroup was 26.7 months, compared to 71.5 months for patients in the good-prognosis subgroup. The estimated hazard ratio generated by the 17 transcripts was not significant (p = 0.25). It might due to the sub-optimal measurement of fold-change for gene expression studies. Here we relied on the remaining default settings of the R-package *cluster* and *survival*, though other classifier arguments may yield better results after sophisticated fine tuning. This

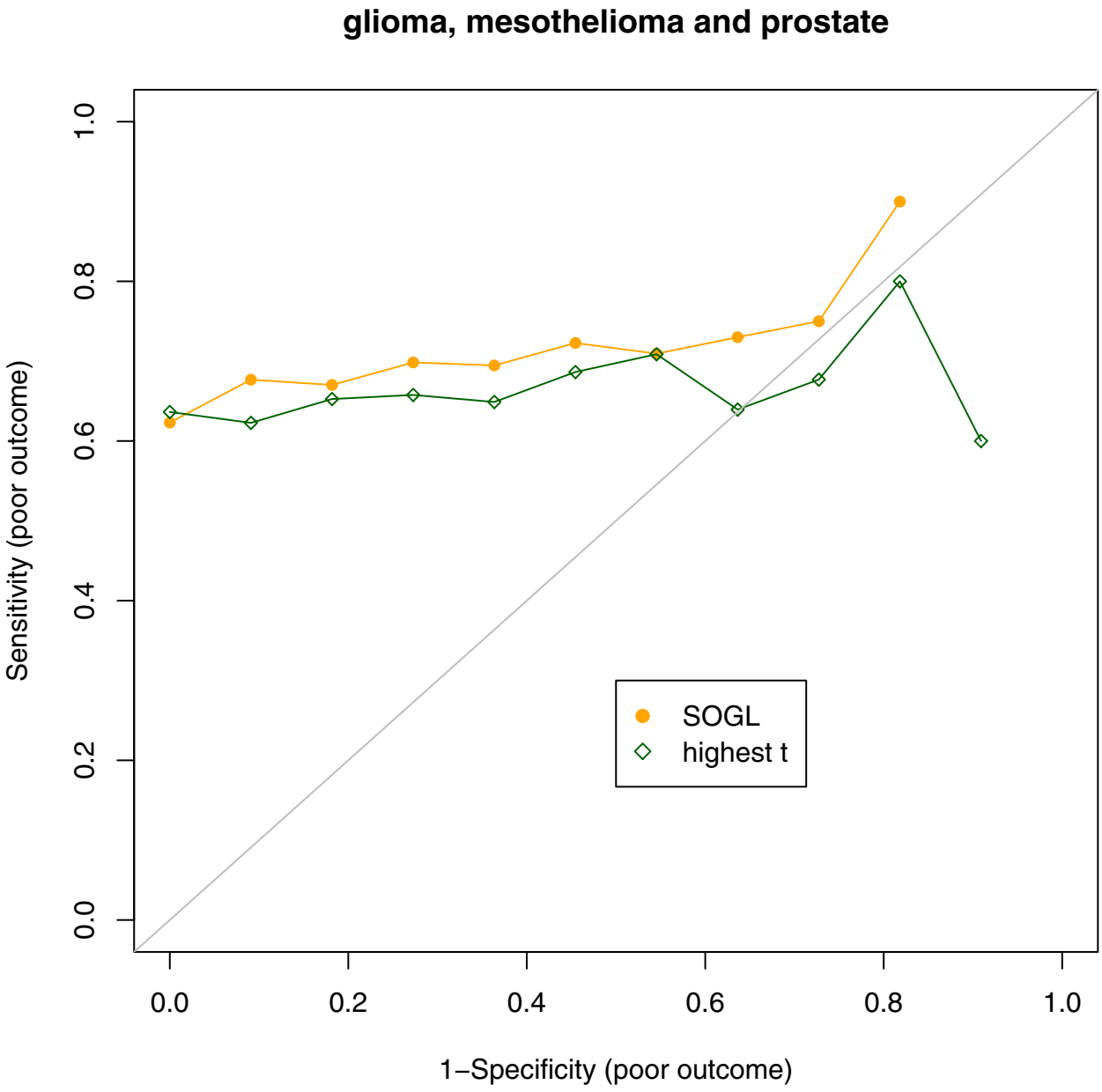

Figure 5
Comparison of methods using solid tumor data. The ROC points for 500 prediction runs. Each time, the solid circle is the result from SOGL, and the diamond is the result from the same number of highest t-scores.

result provides insights into the application of our microarray analysis in clinical settings and could help to identify novel targets for molecular pharmacodynamics.

We want to emphasize that we did not test the statistical significance of the identified genes with the survival outcomes by fitting the Cox proportional-hazards model to each gene [29]. We believe it contains information that the consensus change of these genes in a group, and this information is of critical importance in elucidating the complex genetic architecture of tumor progression, e.g. certain biochemical path. In fact, two of the small set of transcripts are insulin-like growth factor binding protein-3 (IGFBP3), over-expression of which has already anno-

Table 2: The similarity-driving genes found in the G, M, and P studies

gene		rank(fold-change)			rank(z-statistic)		
Symbol	probeID	G	M	P	G	M	P
IGFBP3	37319_at	-11	-41	-11	-68	-53	-22
	1586_at	-35	-177	-52	-90	-128	-19
COL4A2	36659_at	-27	-4	-111			
COL4A1	39333_at	-17	-11	-58			
COL1A2	32306_g_at	-23	-86	-1			
PTGDS	38406_f_at	24	62	55			
	216_at	28	87	50	446	407	9
	38407_r_at	34	196	52			
ANXA2	769_s_at	-28	-16	-57			
ANXA2P3	31444_s_at	-31	-17	-72			
CPE	36606_at	23	147	137			
FN1	31719_at	-39	-61	-53			
BGN	38126_at	-34	-113	-41			
MDK	577_at	-85	-63	-37			
	38124_at	-96	-50	-78			
COL5A2	38420_at	-80	-51	-151	-84	-227	-30
POSTN	1451_s_at	-130	-52	-56			
PTTG1	40412_at				-69	-234	-86

A negative value indicates a relatively high expression level in the poor-outcome patients than in the good-outcome patients, and a positive value indicates a relatively low expression level in the poor-outcome patients.

tated as apoptosis promoter of cancer cells, activated by p53 [30,31]. Moreover, it has recently been independently detected by other studies in vivo or in vitro that the increased expression of COL5A2 in colorectal cancer [32], the increased expression of PTTG1 with correlation to poor prognosis in glioma [33], and the down-regulation of the PTGDS as an important variable in liver and blad-

Table 4: 21 Significantly similar comparisons of the ordered gene lists with the same labels used by the Figure 3

comparison ID	studies	α.opt	# up	#down	#genes (0.03)
2_1	A B	0.06	162	154	43
2_2	A P	0.03	166	164	46
3_1	A B L	0.015	267	176	16
3_2	A B M	0.3	142	271	8
3_3	A B P	0.012	135	252	4
3_4	A B G	0.01	140	255	7
3_5	A L P	0.01	198	213	12
3_6	B M G	0.01	0	169	5
3_7	M P G	0.02	206	187	17
4_1	A B L M	0.3	350	296	4
4_2	A B L P	0.012	0	340	0
4_3	A B L G	0.01	0	235	0
4_4	A B M G	0.01	314	395	7
4_5	A B P G	0.01	448	243	3
4_6	A L M P	0.01	483	354	6
4_7	A L M G	0.01	483	354	6
4_8	A M P G	0.01	463	251	7
4_9	L M P G	0.01	232	396	6
5_1	A B L M G	0.01	0	481	3
5_2	A B M P G	0.01	666	419	2
5_3	A L M P G	0.01	466	424	2

The column "# up" records how many orders to count for up-regulated genes [66], and "# down" records how many orders to count for down-regulated genes in poor outcomes.

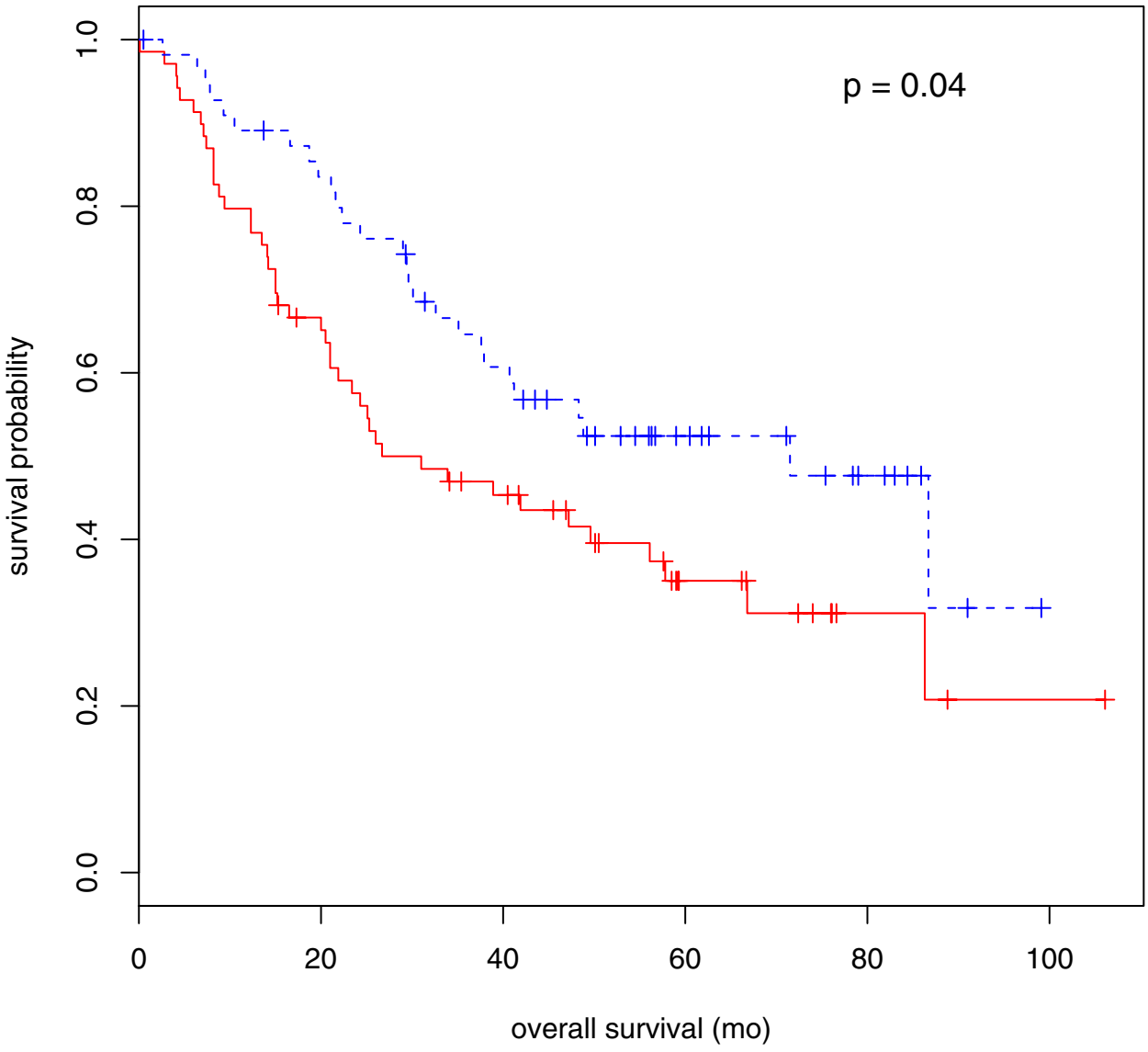

Figure 6
Survival analysis of lung cancer patients. Kaplan-Meier survival analysis of individual outcomes defined by 5 similarity-driving transcripts in the three solid tumors.

der cancer cell and in malignant progression forms of oral tissue [34-36].

Discussion

Treatment of cancer patients is known to impact in several ways on prognosis. For an identical tumor, prognosis may be good if the condition has been diagnosed in good time but hopeless otherwise. Also, the set of genes that show significant changes of expression in one specific tumor includes genes that are significant for prognosis. Genes that are recognized statistically, especially in small data sets, might be of little value for new patients. In contrast, the genes that show consistent changes across all prognostic gene-lists have key roles in cancer development and

progression. Therefore, to detect universal prognostic markers, integrated analysis based on large patient groups is required, and significance needs to be judged at the meta-analysis stage. SOGL quantifies and tests the similarities between two or more gene lists. The genes driving the similarity are those with prominent ranks in all the lists compared. Notwithstanding personal and other influences, these genes may genuinely indicate molecular alterations common among neoplasias. Another serious concern for bioinformatics researchers is the arbitrary or over-fitted choice of statistical approach that yields far-from-reliable gene sets. Information about clinical outcomes is unstable and weak because the differences among individuals might be large, and the challenge is to overcome this problem. Our results show that the SOGL method complements previous methods and is robust. The marker genes identified on the basis of one effect size concur with those based on another in our limited data. Though without strongly superiority, SOGL is tend to be more accurate than highest t-score for variable selection by meta-analysis. Studies that in isolation do not provide solid evidence for differential gene expression may present striking similarities in their gene lists. Thus SOGL can identify consensus signals from either strong or weak effects, independently of the arbitrary threshold. Moreover, it would be of greater interest to apply SOGL to the exploration of disease mechanisms based on these commonly changed genes in consensus. It is different from the approaches targeting only the "best" marker, result of SOGL might include genes that are so-called "redundant" by certain "threshold" of significance or correlation in individual study. Co-regulating genes in a biological path, genes in a parallel path, and genes having epistatic actions are in fact genes of critical importance in elucidating the complex genetic architecture of a complex disease [37]. Thus SOGL might be used to uncover the hidden pattern of genes on microarrays. Instead of distinctions of significance or correlation, it focuses on the genes relevant to the condition of interest that are consistently changed across multiple studies.

Biologists usually compare independent studies addressing the same research question to confirm findings. It is also possible to compare studies from slightly different but related contexts in order to discover common markers. This is an attempt to revolutionize cancer data sets to screen for common molecular features shared among phenotypically different types of cancer involving distinct biological underpinnings, disease progression, diagnosis and prognosis. We detected and confirmed that significant similarities span several kinds of cancer. This result supports the emerging notion that different types of tumors for which prognosis is poor share common disorders in the regulation of gene expression. This implies that poor prognosis sometimes develops independently of original cancer type.

A substantial literature suggests that the similarity-driving genes are promising as tumor markers and as targets for tumor therapy. The genes common to the top ends of the lists for the outcomes of the three cancers studied here include those originally used by Singh and Gordon [5,38] for outcome prediction, such as IGFBP3. FN1 has also been used in a real-time PCR-based multigene outcome predictive model for lymphoma [39] and prostate cancer [40]. Expression of POSTN is reportedly a bone metastasis from breast cancer [41] and is proposed as a prognostic marker in lung tumor invasion [42]. Dysregulation of ANXA2 has been reported in human bone cancer metastases [43] and is correlated with the clinical prognosis of prostate cancer [44]. Additional supportive evidence of the prognostic value of the genes in Table 2 from experiment in vitro and in vivo has been cited in the last section of result.

Our most striking finding, however, is the over-representation of genes detected from fold changes (MDK, CPE, POSTN, COL4A1, COL11A1, COL1A2, COL5A2, IGFBP3, FN1, ANXA2, BGN and PTGDS) and all 4 genes detected from the z-statistic as effect size (PTTG1, COL5A2, IGFBP3 and PTGDS) are associated with angiogenesis. Angiogenesis leads to the formation of a large anastomosing vascular network, allowing tumor growth, intravasation and the spread of metastases. MDK, which plays an important role in the intercellular interactions involved in angiogenesis, is reported to be strongly correlated with poor prognosis in a large number of cases irrespective of tissue type [45-50]. Another gene, CPE, is relatively down-regulated in the three poor-outcome samples of carcinoid tumors [51], and takes part in producing angiogenic factors upon the maturation of follicle stimulating hormone [52]. Generally, the breakdown of ECM proteins, which correlates with angiogenesis, is an essential step in cancer invasion and metastasis [53]. We found that up-regulation of 7 genes involved with the ECM is associated with poor cancer outcomes. ECM-related genes that promoted the strongest proliferation, including POSTN [54], BGN [55] type I collagen [56] and type IV collagen [56], have already been identified as cancer markers, and might be molecular targets for gene therapy. In addition, BGN and PTGDS have recently been reported in an in vitro angiogenesis system [57]. The oncogenic potential of PTTG1 has been well characterized in mouse fibroblast (NIH3T3) cells, in which it induces proliferation and promotes tumor formation and angiogenesis [58]. It has been reported as a prognostic marker for tumor invasiveness and metastasis [59] and is suggested to be a potent human oncogene [60]. These findings suggest that by inhibiting angiogenesis, it may be possible to restrict the blood sup-

ply to tumors and limit their ability to grow and metastasize. Our results support the anti-angiogenic hypothesis concerning polymeric FN1 [40] and ANXA2 [54] and suggest more candidate markers. Because the similarities among multiple tumor tissues can not be identified by speculation, we believe that further meta-analysis on more data will aid further research on prognostic markers of many cancers.

Conclusion

For a small clinical trial, it is important to summarize all the evidence obtained and combine it with evidence from other trials or laboratory studies. Meta-analysis enables general conclusions to be drawn, develops support for hypotheses, and produces an estimate of the overall effects of a program, combining with the developed multiple statistical algorithm. This study suggests that our meta-analysis of gene lists for different clinical or physiological phenotypes provides a golden opportunity for detecting biologically relevant gene dysregulations between different phenotypes and possibly leading to improved diagnostic accuracy, or generating insightful molecular mechanisms to build the underlying bridges between different phenotypes. To this end, SOGL is superior to other measurements of gene selection for meta-analysis of clinical microarrays for handling study-to-study differences. It focuses on the genes relevant to the condition of interest that are consistently changed across multiple studies, rather than on distinctions of significance or correlation. Our study has assessed its potential for identifying prognostic markers of multiple cancer types from studies of different laboratories, especially for studies with large inter-individual variations or small sample size. The proposed method is a complementarity and enlargement algorithm for research on gene expression.

In addition, our results suggest and confirm that a common molecular mechanism underlies the poor outcomes of several kinds of cancer. The genes we detected have important implications for our understanding of the potential involvement of angiogenesis in the malignant progression of primary tumors. It suggests that meta-analysis has considerable potential in clinical studies of gene expression profiles, which is a focus of active research for computer-assisted diagnosis. To ensure reproducibility of our biological findings, larger numbers representing a greater percentage of disease is required. It is expected that further studies incorporating more data sets with larger number of samples will identify universal prognostic markers in cancer.

Methods

Transcript expression data and outcome

In transcriptional research, the raw data have to be corrected for different conditions by normalization. We normalized all raw profiling files on an additive scale by preprocessing methods for stabilizing variance [61]. "An additive scale" means transforming the intensities to a scale where the variance is approximately independent of the mean intensity. This can be achieved by calibrating for sample-to-sample variations through shifting and scaling, or by log-transforming the data. For simplicity, we focused on the published microarray studies of cancer outcomes based on Affymetrix chips, which have sufficient data and have gained acceptance in recent years because of the reliable annotation and identification and the good hybridization characteristics of oligonucleotides with wide-ranging expression levels [62]. Only the best-matched transcripts [63] were used to compare studies based on different chips.

The definitions of outcomes for all the studies we collected strictly followed those of the original papers. To evaluate the power of signature detection in transcript expression and the accuracy of prediction by our adopted method, we integrated all the relatively non-malignant outcomes as "good". In contrast, the patients were grouped as "poor" if they suffered shorter survival or if there was recurrence within the observed time. The data sets were:

• **Leukemia C:** The data came from research on adult T-cell acute lymphoblastic leukemia (ALL) [64]. The good prognosis group consisted of 7 patients in complete clinical remission (CCR) and 2 patients who had not relapsed within two years. The poor prognosis group consisted of 6 refractory patients and 12 who had relapsed within two years.

• **Leukemia Y:** The data included 327 children suffering leukemia [23]. Excluding the patients without outcome information, The good group consisted of 201 CCR patients, while the poor responder group consisted of 44 patients with different types of relapse.

• **Leukemia R:** 93 patients with prognostic information from above study were examined the gene expression profiling by Ross et al. using another microarray chip [22]. The good prognosis group consisted of 71 CCR patients. The poor prognosis group consisted of 16 relapsed patients and six 2nd AML patients.

• **Mesothelioma:** A prognostic study on mesothelioma, a lethal neoplasia of the pleura [5]. The good responder group consisted of 8 patients who survived more than sev-

Table 3: Illustration of the cardinality of the intersection $O_n(G_D)$

n	G_1	G_2	G_3	$O_n(G_{1,2,3})$
1	a	**h**	**k**	0
2	**k**	w	z	0
3	**h**	b	**h**	1
4	m	**K**	b	2
5	**t**	a	**t**	2
6	w	**t**	i	3
...	...	...	...	...

Rows correspond to the orderings (n), columns to the different studies. $G_d(n)$ denotes the n'th most strongly up-regulated transcript in study d that changes when poor-outcome is compared with good-outcome.

enteen months, while the 10 patients in the poor responder group survived less than six months.

• **Prostate:** This comparison was constructed from 21 prostate tumor samples with respect to recurrence following surgery [38]. The good prognosis group consisted of 13 patients who had shown no relapse for at least four years, and the poor outcome groups consisted of 8 relapse patients.

• **Glioma:** This comparison was based on the data of Shai et al. [65]. The good prognosis group consisted of 8 primary (not secondary) glioblastoma multiforme (GBM) patients of various pathological types and grades with a survival time of more than three years, while the poor responder group consisted of 10 malignant glioma patients who survived less than one year.

• **Breast 1:** The data were taken from a prognostic study of primary breast tumors by Huang et al. [3]. In total, 37 patients were included. The good prognosis group consisted of 19 "low-risk" patients, and the poor responder group consisted of 18 patients identified as "high-risk" by their lymph-node status.

• **Breast 2:** These data were also described by Huang et al. [3]. Here, however, the prognostic groups were defined directly by clinical outcome. The good responder group consisted of 34 patients who were recurrence-free over three years, while the poor responder group consisted of 18 patients who suffered recurrent disease within the first three years after surgery.

• **Lung:** The data included 126 adenocarcinoma (one subtype of lung cancer) cases without metastases reported by Bhattacharjee et al. [27]. The lung cancer data set did not define the outcome classification for each case. However, the author reported that the neuroendocrine C2 adenocarcinoma were associated with a less favorable survival outcome. Therefore the poor responder group consisted of 9 neuroendocrine C2 adenocarcinoma patients, while the good responder group consisted of all the other 117 adenocarcinoma patients.

Detecting similarities amongst ordered gene lists and their contributing genes

SOGL introduces a comparison between two states [21,66]. Preferably, one state relates to a good outcome or prognosis and the other to a bad outcome. Let D^* be the collection of studies. Applying a standard statistic to each study, $d \in D^*$, we can obtain a gene list g_i^d representing the differences in expression between samples in the poor- and good-outcome classes. The original similarity score, S_n, is based on the number of overlapping genes in the top n ranks deriving from $k = 2$ gene lists [21,66]. We can assign a more general similarity score to a comparison of several gene lists as Table 3 shows. Thus the extended SOGL score is here defined as a summation of weighted partial intersection sizes on k ends of ordered gene lists:

$$S_\alpha(G_1, G_2, ..., G_k) = \sum_n w_n^\alpha O_n$$

where decreasing weights (w) are used as: $w_n^\alpha = e^{-\alpha n}$. In this way, we strengthen the two ends of the integrated transcript orders. By setting the parameter α, one can calibrate the weight to decide that how deeply these gene orders are to be investigated.

To calibrate an adaptive α^* to the gene lists of interest, we partially (80%) resampled the class labels of patients in the original raw data [21,66]. Class-balanced resampling from the good- and poor-outcome groups estimates the signal (alternative score), and class-shuffled resampling in each study simulates background noise of the same size (noise score). This resampling for estimation step was iterated C (= 500) times. To evaluate the separation of these two score distributions, we applied the pAUC-score [26] resulting from a comparison of signal and noise. Fixing a

maximally acceptable false positive rate w_0, we measured separability as the area under ROC(w) with $w < w_0$ as

$$pauc_{\alpha_i}(w_0) = \int_0^{w_0} \mathrm{ROC}(w)\mathrm{d}w,$$

where w was the false positive rate, and ROC(w) was the true positive rate. A high pAUC-score indicates good separation. Given a parameter α_i, the separation of alternative scores and noise scores indicates the similarity between the leading genes in these gene orders. For a predefined finite grid of parameters, then we can pick the value providing the best discrimination between signal and noise. The significance was then evaluated for a given α. To this end, we simulated the distribution of similarity score under assumption of unrelated lists and generated B (= 1000) set of ternately random ranks to calculate the random scores. Significance was evaluated by computing an empirical p-value for the observed scores from the B random scores.

The similarity-driving genes should be consistently represented among the leading items in the gene orders. One can count a cutoff value n^* such as $\sum_{n=1}^{n^*} e^{-\alpha^* n} O_n$ to accounts for 95% of the score S_{α^*}, given an identical α^*. Note that SOGL is the sum of the scores for the two ends. Thus we identify the similarity scores for up- and down-regulation, ignoring genes for which the isolated up- or down-regulation yields scores no higher than the 99th percentile of the random scores. The expected random scores are given by B (= 1000) shuffled orderings. For example, if a certain significance is due to the most strongly down-regulated genes but not to the most strongly up-regulated genes, we ignore the intersection of up-regulated genes.

Estimating the accuracy of prediction

We expected that combined studies will predict the outcome for single patients better than a single study can, assuming that there is commonality in the dysregulation of gene expression for certain malignant processes. To validate this assumption, we calculated the number of correct predictions for each study via two steps. (1) All patients from three similar studies were mixed into one integrated data set to cross-validate the outer and inner loops. This resulted in a vote matrix containing the number of times each sample was assigned to each class in the outer cross-validation loop. We counted the coincidences between true class and consensus class for samples study by study to obtain three tables. (2) The same cross-validation was run on single data to obtain independent tables for each study. For both steps, we repeated the cross-validation with the same stratified strategy (class-balanced folds [67]) and adopted the identified variable selection method and the classification method. We assume that we can combine data from different studies into one replicated data set, if the gene lists are significantly similar for a certain two-condition test.

We next compared SOGL with the traditional highest t-score to select variables for prediction, carrying out the same classification and patient clustering strategy before meta-analysis. To avoid study-to-study bias or prevalence of smaller sample sizes, we randomly employed class-balanced [67] and study-balanced training sets. "Class-balanced" means that we guarantee the combined training set comprises approximately half poor-outcome and half good-outcome patients. "Study-balance" means that we guarantee the training set contains all the different tumors, and keeps more or less the same proportion of each. Patients not used in the combined training set were used for validation. For SOGL variable selection, we focused on a fixed number of orderings to calculate the similarity score, and selected the intersection to account for 95% of the score. The resulting variables were used to predict the outcomes for patients in the associated validation set after tuning hyperparameters of SVM. After this step, we recorded the number of selected genes, then picked the same number of genes with the highest t-statistic to estimate the accuracy of prediction in the validation set. The above training/validation step was iterated D (= 500) times carrying SVM algorithm performing linear kernel by R package *e1071*. To compare the two variable-selection methods, we drew a Receiver Operating Characteristic (ROC) curve from the correct error metrics generated from the D repeats of training/validation step for each test. ROC is a plot of the true positive rate (TPR) on the y axis against the false positive rate (FPR) on the x axis for the different possible cut-off points of a diagnostic test. Thus, for every observed FPR, we calculated the mean value of the corresponding TPR to plot the point on the ROC curve. Let u be the good prognostic for the true good-outcome patients; ve the bad prognostic for the true good-outcome patients; t be the good prognostic for the true bad-outcome patients; and s be the bad prognostic for the true bad-outcome patients. For the null hypothesis that all patients are poor outcome, the sensitivity and specificity are:

$$TPR(poor\text{ - }outcome) = s/(s + t);\ FPR(poor\text{ - }outcome) = v/(u + v).$$

We measured the area under the ROC curves to evaluate the difference between SOGL and t-scores.

Authors' contributions

XY carried out the data collection, performed the statistical analysis, and drafted the manuscript. XS participated in the design of the study. All authors read and approved the final manuscript.

Acknowledgements

We thank Dr. C Lottaz from Max Planck Institute for Molecular Genetics (Berlin) and Dr. J Jaeger from Swiss company Hamilton for helpful discussions about the statistical analysis. We are grateful to Prof. Z Ai from University of Illinois at Chicago for careful reading of draft. This research has been supported by the Natural Science Foundation 60671018, 60121101, National High Technology Research and Development Program of China 863-2005AA231070 and Southeast University Foundation XJ0711279.

References

1. van 't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AAM, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT, Schreiber GJ, Kerkhoven RM, Roberts C, Linsley PS, Bernards R, Friend SH: **Gene expression profiling predicts clinical outcome of breast cancer.** *Nature* 2002, **415(6871):**530-536.
2. van de Vijver MJ, He YD, van't Veer LJ, Dai H, Hart AAM, Voskuil DW, Schreiber GJ, Peterse JL, Roberts C, Marton MJ, Parrish M, Atsma D, Witteveen A, Glas A, Delahaye L, van der Velde T, Bartelink H, Rodenhuis S, Rutgers ET, Friend SH, Bernards R: **A gene-expression signature as a predictor of survival in breast cancer.** *N Engl J Med* 2002, **347(25):**1999-2009.
3. Huang E, Cheng SH, Dressman H, Pittmane J, Tsou MH, Horng CF, Bild A, Iversen ES, Liao M, Chen CM, West M, Nevins JR, Huang AT: **Gene expression predictors of breast cancer outcomes.** *Lancet* 2003, **361(9369):**1590-1596.
4. Nevins JR, Huang ES, Dressman H, Pittman J, Huang AT, West M: **Towards integrated clinico-genomic models for personalized medicine: combining gene expression signatures and clinical factors in breast cancer outcomes prediction.** *Hum Mol Genet* 2003, **12(Spec No 2):**R153-157.
5. Gordon GJ, Jensen RV, Hsiao LL, Gullans SR, Blumenstock JE, Richards WG, Jaklitsch MT, Sugarbaker DJ, Bueno R: **Using Gene Expression Ratios to Predict Outcome Among Patients With Mesothelioma.** *Journal of the National Cancer Institute* 2003, **95(8):**598-605.
6. Futschik ME, Sullivan M, Reeve A, Kasabov N: **Prediction of clinical behaviour and treatment for cancers.** *Appl Bioinformatics* 2003, **2(3 Suppl):**S53-58.
7. Choi JK, Choi JY, Kim DG, Choi DW, Kim BY, Lee KH, Yeom YI, Yoo HS, Yoo OJ, Kim S: **Integrative analysis of multiple gene expression profiles applied to liver cancer study.** *FEBS Lett* 2004, **565(1–3):**93-100.
8. Cario G, Stanulla M, Fine BM, Teuffel O, Neuhoff NV, Schrauder A, Flohr T, SchACURfer BW, Bartram CR, Welte K, Schlegelberger B, Schrappe M: **Distinct gene expression profiles determine molecular treatment response in childhood acute lymphoblastic leukemia.** *Blood* 2005, **105(2):**821-826.
9. Bloom G, Yang IV, Boulware D, Kwong KY, Coppola D, Eschrich S, Quackenbush J, Yeatman TJ: **Multi-platform, multi-site, microarray-based human tumor classification.** *Am J Pathol* 2004, **164:**9-16.
10. Larkin JE, Frank BC, Gavras H, Sultana R, Quackenbush J: **Independence and reproducibility across microarray platforms.** *Nature Methods* 2005, **2(5):**337-344.
11. Warnat P, Eils R, Brors B: **Cross-platform analysis of cancer microarray data improves gene expression based classification of phenotypes.** *BMC Bioinformatics* 2005, **6:**265.
12. Bammler T, Beyer RP, Bhattacharya S, Boorman GA, Boyles A, Bradford BU, Bumgarner RE, Bushel PR, Chaturvedi K, Choi D, Cunningham ML, Deng S, Dressman HK, Fannin RD, Farin FM, Freedman JH, Fry RC, Harper A, Humble MC, Hurban P, Kavanagh TJ, Kaufmann WK, Kerr KF, Jing L, Lapidus JA, Lasarev MR, Li J, Li YJ, Lobenhofer EK, Lu X, Malek RL, Milton S, Nagalla SR, O'malley JP, Palmer VS, Pattee P, Paules RS, Perou CM, Phillips K, Qin LX, Qiu Y, Quigley SD, Rodland M, Rusyn I, Samson LD, Schwartz DA, Shi Y, Shin JL, Sieber SO, Slifer S, Speer MC, Spencer PS, Sproles DI, Swenberg JA, Suk WA, Sullivan RC, Tian R, Tennant RW, Todd SA, Tucker CJ, Houten BV, Weis BK, Xuan S, HZ : **Standardizing global gene expression analysis between laboratories and across platforms.** *Nature Methods* 2005, **2(5):**351-356.
13. Grutzmann R, Boriss H, Ammerpohl O, Luttges J, Kalthoff H, Schackert HK, Kloppel G, Saeger HD, Pilarsky C: **Meta-analysis of microarray data on pancreatic cancer defines a set of commonly dysregulated genes.** *Oncogene* 2005, **24(32):**5079-5088.
14. Rhodes DR, Chinnaiyan AM: **Integrative analysis of the cancer transcriptome.** *Nature Genet* 2005, **37(Suppl):**S31-37.
15. Bernards R, Weinberg RA: **Metastasis genes: A progression puzzle.** *Nature* 2002, **418(6900):**823.
16. Yang X, Bentink S, Spang R: **Detecting common gene expression patterns in multiple cancer outcome entities.** *Biomed Microdevices* 2005, **7(3):**247-251.
17. Glinsky GV, Berezovska O, Glinskii AB: **Microarray analysis identifies a death-from-cancer signature predicting therapy failure in patients with multiple types of cancer.** *J Clin Invest* 2005, **115(6):**1503-1521.
18. Rhodes DR, Yu J, Shanker K, Deshpande N, Varambally R, Ghosh D, Barrette T, Pandey A, Chinnaiyan AM: **Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression.** *Proc Natl Acad Sci USA* 2004, **101(25):**9309-9314.
19. Finocchiaro G, Mancuso F, Muller H: **Mining published lists of cancer related microarray experiments: identification of a gene expression signature having a critical role in cell-cycle control.** *BMC Bioinformatics* 2005, **6(Suppl 4):**S14.
20. Segal E, Friedman N, Koller D, Regev A: **A module map showing conditional activity of expression modules in cancer.** *Nature Genet* 2004, **36(10):**1090-1098.
21. Yang X, Bentink S, Scheid S, Spang R: **Similarities of ordered gene lists.** *J Bioinform Comput Biol* 2006, **4(3):**693-708.
22. Ross ME, Zhou X, Song G, Shurtleff SA, Girtman K, Williams WK, Liu HC, Mahfouz R, Raimondi SC, Lenny N, Patel A, Downing JR: **Classification of pediatric acute lymphoblastic leukemia by gene expression profiling.** *Blood* 2003, **102(8):**2951-2959.
23. Yeoh EJ, Ross ME, Shurtleff SA, Williams WK, Patel D, Mahfouz R, Behm FG, Raimondi SC, Relling MV, Patel A, Cheng C, Campana D, Wilkins D, Zhou X, Li J, Liu H, Pui CH, Evans WE, Naeve C, Wong L, Downing JR: **Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling.** *Cancer Cell* 2002, **1(2):**133-143.
24. Scheid S, Spang R: **Twilight; a Bioconductor package for estimating the local false discovery rate.** *Bioinformatics* 2005, **21(12):**2921-2922.
25. Storey JD, Tibshirani R: **Statistical significance for genome-wide studies.** *Proc Natl Acad Sci USA* 2003, **100(16):**9440-9445.
26. Pepe MS, Longton G, Anderson GL, Schummer M: **Selecting Differentially Expressed Genes from Microarray Experiments.** *Biometrics* 2003, **59:**133-142.
27. Bhattacharjee A, Richards WG, Staunton J, Li C, Monti S, Vasa P, Ladd C, Beheshti J, Bueno R, Gillette M, Loda M, Weber G, Mark EJ, Lander ES, Wong W, Johnson BE, Golub TR, Sugarbaker DJ, Meyerson M: **Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses.** *Proc Natl Acad Sci USA* 2001, **98(24):**13790-13795.
28. Kaufman L, Rousseeuw PJ: *Finding Groups in Data: An Introduction to Cluster Analysis* New York: John Wiley and Sons, Inc; 1990.
29. Xu JZ, Guo Z, Zhang M, Li X, Li YJ, Rao SQ: **Peeling off the hidden genetic heterogeneities of cancers based on disease-relevant functional modules.** *Mol Med* 2006, **12(1–3):**25-33.
30. Harms KL, Chen X: **The C terminus of p53 family proteins is a cell fate determinant.** *Mol Cell Biol* 2005, **25(5):**2014-2030.
31. Silha JV, Sheppard PC, Mishra S, Gui Y, Schwartz J, Dodd JG, Murphy LJ: **Insulin-like growth factor (IGF) binding protein-3 attenuates prostate tumor growth by IGF-dependent and IGF-independent mechanisms.** *Endocrinology* 2006, **147(5):**2112-2121.
32. Fischer H, Stenling R, Rubio C, Lindblom A: **Colorectal carcinogenesis is associated with stromal expression of COL11A1 and COL5A2.** *Carcinogenesis* 2001, **22(6):**875-878.
33. Genkai N, Homma J, Sano M, Tanaka R, Yamanaka R: **Increased expression of pituitary tumor-transforming gene (PTTG)-1 is correlated with poor prognosis in glioma patients.** *Oncol Rep* 2006, **15(6):**1569-1574.

34. Rasmussen HH, Orntoft TF, Wolf H, Celis JE: **Towards a comprehensive database of proteins from the urine of patients with bladder cancer.** *J Urol* 1996, **155(6):**2113-2119.
35. Banerjee AG, Bhattacharyya I, Vishwanatha JK: **Identification of genes and molecular pathways involved in the progression of premalignant oral epithelia.** *Mol Cancer Ther* 2005, **4(6):**865-875.
36. Cho-Rok J, Yoo J, Jang YJ, Kim S, Chu IS, Yeom YI, Choi JY, Im DS: **Adenovirus-mediated transfer of siRNA against PTTG1 inhibits liver cancer cell growth in vitro and in vivo.** *Hepatology* 2006, **43(5):**1042-1052.
37. Li X, Rao S, Wang Y, Gong B: **Gene mining: a novel and powerful ensemble decision approach to hunting for disease genes using microarray expression profiling.** *Nucleic Acids Res* 2004, **32(9):**2685-2694.
38. Singh D, Febbo PG, Ross K, Jackson DG, Manola J, Ladd C, Tamayo P, Renshaw AA, D'Amico AV, Richie JP, Lander E, Loda M, Kantoff PW, Golub TR, Sellers WR: **Gene expression correlates of clinical prostate cancer behavior.** *Cancer Cell* 2002, **1:**203-209.
39. Lossos IS, Czerwinski DK, Alizadeh AA, Wechser MA, Tibshirani R, Botstein D, Levy R: **Prediction of survival in diffuse large-B-cell lymphoma based on the expression of six genes.** *N Engl J Med* 2004, **350(18):**1828-1837.
40. Ifon ET, Pang ALY, Johnson W, Cashman K, Zimmerman S, Muralidhar S, Chan WY, Casey J, Rosenthal LJ: **U94 alters FN1 and ANGPTL4 gene expression and inhibits tumorigenesis of prostate cancer cell line PC3.** *Cancer Cell Int* 2005, **5:**19.
41. Sasaki H, Yu CY, Dai M, Tam C, Loda M, Auclair D, Chen LB, Elias A: **Elevated serum periostin levels in patients with bone metastases from breast but not lung cancer.** *Breast Cancer Res Treat* 2003, **77(3):**245-252.
42. Sasaki H, Lo KM, Chen LB, Auclair D, Nakashima Y, Moriyama S, Fukai I, Tam C, Loda M, Fujii Y: **Expression of Periostin, homologous with an insect cell adhesion molecule, as a prognostic marker in non-small cell lung cancers.** *Jpn J Cancer Res* 2001, **92(8):**869-873.
43. Gillette JM, Chan DC, Nielsen-Preiss SM: **Annexin 2 expression is reduced in human osteosarcoma metastases.** *J Cell Biochem* 2004, **92(4):**820-832.
44. Banerjee AG, Liu J, Yuan Y, Gopalakrishnan VK, Johansson SL, Dinda AK, Gupta NP, Trevino L, Vishwanatha JK: **Expression of biomarkers modulating prostate cancer angiogenesis: differential expression of annexin II in prostate carcinomas from India and USA.** *Mol Cancer* 2003, **2:**34.
45. Sumi Y, Muramatsu H, Takei Y, Hata KI, Ueda M, Muramatsu T: **Midkine, a heparin-binding growth factor, promotes growth and glycosaminoglycan synthesis of endothelial cells through its action on smooth muscle cells in an artificial blood vessel model.** *J Cell Sci* 2002, **115:**2659-2667.
46. Muramatsu T: **Midkine and pleiotrophin: two related proteins involved in development, survival, inflammation and tumorigenesis.** *J Biochem (Tokyo)* 2002, **132(3):**359-371.
47. Kadomatsu K, Muramatsu T: **Midkine and pleiotrophin in neural development and cancer.** *Cancer Lett* 2004, **204(2):**127-143.
48. Ikematsu S, Nakagawara A, Nakamura Y, Sakuma S, Wakai K, Muramatsu T, Kadomatsu K: **Correlation of elevated level of blood midkine with poor prognostic factors of human neuroblastomas.** *Br J Cancer* 2003, **88(10):**1522-1526.
49. Shimada H, Nabeya Y, Tagawa M, Okazumi S, Matsubara H, Kadomatsu K, Muramatsu T, Ikematsu S, Sakuma S, Ochiai T: **Preoperative serum midkine concentration is a prognostic marker for esophageal squamous cell carcinoma.** *Cancer Sci* 2003, **94(7):**628-632.
50. Roversi G, Pfundt R, Moroni RF, Magnani I, van Reijmersdal S, Pollo B, Straatman H, Larizza L, Schoenmakers EFPM: **Identification of novel genomic markers related to progression to glioblastoma through genomic profiling of 25 primary glioma cell lines.** *Oncogene* 2006, **25(10):**1571-1583.
51. He P, Varticovski L, Bowman ED, Fukuoka J, Welsh JA, Miura K, Jen J, Gabrielson E, Brambilla E, Travis WD, Harris CC: **Identification of carboxypeptidase E and gamma-glutamyl hydrolase as biomarkers for pulmonary neuroendocrine tumors by cDNA microarray.** *Hum Pathol* 2004, **35(10):**1196-1209.
52. Sasson R, Dantes A, Tajima K, Amsterdam A: **Novel genes modulated by FSH in normal and immortalized FSH-responsive cells: new insights into the mechanism of FSH action.** *FASEB J* 2003, **17(10):**1256-1266.
53. Zhang D, Samani AA, Brodt P: **The role of the IGF-I receptor in the regulation of matrix metalloproteinases, tumor invasion and metastasis.** *Horm Metab Res* 2003, **35(11–12):**802-808.
54. Shao R, Bao S, Bai X, Blanchette C, Anderson RM, Dang T, Gishizky ML, Marks JR, Wang XF: **Acquired expression of periostin by human breast cancers promotes tumor angiogenesis through up-regulation of vascular endothelial growth factor receptor 2 expression.** *Mol Cell Biol* 2004, **24(9):**3992-4003.
55. Chen WB, Lenschow W, Tiede K, Fischer JW, Kalthoff H, Ungefroren H: **Smad4/DPC4-dependent regulation of biglycan gene expression by transforming growth factor-beta in pancreatic tumor cells.** *J Biol Chem* 2002, **277(39):**36118-36128.
56. Grzesiak JJ, Clopton P, Chalberg C, Smith K, Burton DW, Silletti S, Moossa AR, Deftos LJ, Bouvet M: **The extracellular matrix differentially regulates the expression of PTHrP and the PTH/PTHrP receptor in FG pancreatic cancer cells.** *Pancreas* 2004, **29(2):**85-92.
57. Sun XT, Zhang MY, Shu C, Li Q, Yan XG, Cheng N, Qiu YD, Ding YT: **Differential gene expression during capillary morphogenesis in a microcarrier-based three-dimensional in vitro model of angiogenesis with focus on chemokines and chemokine receptors.** *World J Gastroenterol* 2005, **11(15):**2283-2290.
58. Kakar SS, Jennes L: **Molecular cloning and characterization of the tumor transforming gene (TUTR1): a novel gene in human tumorigenesis.** *Cytogenet Cell Genet* 1999, **84(3–4):**211-216.
59. Ramaswamy S, Ross KN, Lander ES, Golub TR: **A molecular signature of metastasis in primary solid tumors.** *Nature Genet* 2003, **33:**49-54.
60. Hamid T, Malik MT, Kakar SS: **Ectopic expression of PTTG1/securin promotes tumorigenesis in human embryonic kidney cells.** *Mol Cancer* 2005, **4:**3.
61. Huber W, von Heydebreck A, Sultmann H, Poustka A, Vingron M: **Variance stabilization applied to microarray data calibration and to the quantification of differential expression.** *Bioinformatics* 2002, **18(Suppl 1):**S96-104.
62. Woo Y, Affourtit J, Daigle S, Viale A, Johnson K, Naggert J, Churchill G: **A comparison of cDNA, oligonucleotide, and Affymetrix GeneChip gene expression microarray platforms.** *J Biomol Tech* 2004, **15(4):**276-284.
63. **affymetrix comparison_spreadsheets** [http://www.affymetrix.com/support/technical/comparison_spreadsheets.affx]
64. Chiaretti S, Li X, Gentleman R, Vitale A, Vignetti M, Mandelli F, Ritz J, Foa R: **Gene expression profile of adult T-cell acute lymphocytic leukemia identifies distinct subsets of patients with different response to therapy and survival.** *Blood* 2004, **103(7):**2771-2778.
65. Shai R, Shi T, Kremen TJ, Horvath S, Liau LM, Cloughesy TF, Mischel PS, Nelson SF: **Gene expression profiling identifies molecular subtypes of gliomas.** *Oncogene* 2003, **22(31):**4918-4923.
66. Lottaz C, Yang X, Scheid S, Spang R: **OrderedList – a bioconductor package for detecting similarity in ordered gene lists.** *Bioinformatics* 2006, **22(18):**2315-2316.
67. Michiels S, Koscielny S, Hill C: **Prediction of cancer outcome with microarrays: a multiple random validation strategy.** *Lancet* 2005, **365(9458):**488-492.