CMLS, Cell. Mol. Life Sci. 61 (2004) 1372–1383
1420-682X/04/111372-12
DOI 10.1007/s00018-004-4069-0

CMLS Cellular and Molecular Life Sciences

Research Article

Selection and validation of differentially expressed genes in head and neck cancer

M. A. Kuriakose [a], W. T. Chen [b], Z. M. He [a], A. G. Sikora [a], P. Zhang [b], Z. Y. Zhang [b], W. L. Qiu [b], D. F. Hsu [c], C. McMunn-Coffran [c], S. M. Brown [d], E. M. Elango [e], M. D. Delacure [a], F. A. Chen [a,*]

[a] Department of Otolaryngology, Division of Head and Neck Surgery, New York School of Medicine, 411 Rusk Research Building, 400E, 34th Street, New York, New York 10016 (USA), Fax +12 12 2630972, e-mail: fangan.chen@med.nyu.edu
[b] Department of Oral and Maxillofacial Surgery, Affiliated Ninth People's Hospital, School of Stomatology, Shanghai Second Medical University, Shanghai 200011 (P. R. China)
[c] Department of Computer and Information Sciences, Fordham University, New York, New York 10023 (USA)
[d] Department of Research Computing, New York University School of Medicine, New York, New York 10016 (USA)
[e] Amrita Institute of Medical Science, Cochin (India)

Received 12 February 2004; received after revision 22 March 2004; accepted 14 April 2004

Abstract. We applied a robust combinatorial (multi-test) approach to microarray data to identify genes consistently up- or down-regulated in head and neck squamous cell carcinoma (HNSCC). RNA was extracted from 22 paired samples of HNSCC and normal tissue from the same donors and hybridized to the Affymetrix U95A chip. Forty-two differentially expressed probe sets (representing 38 genes and one expressed sequence tag) satisfied all statistical tests of significance and were selected for further validation. Selected probe sets were validated by hierarchical clustering, multiple probe set concordance, and target-subunit agreement. In addition, real-time PCR analysis of 8 representative (randomly selected from 38) genes performed on both microarray-tested and independently obtained samples correlated well with the microarray data. The genes identified and validated by this method were in comparatively good agreement with other rigorous HNSCC microarray studies. From this study, we conclude that combinatorial analysis of microarray data is a promising technique for identifying differentially expressed genes with few false positives.

Key words. Microarray analysis; differential gene expression; head and neck cancer; molecular profiling; squamous cell carcinoma.

High-throughput microarray analysis is now widely used to identify genes which are differentially expressed in tumors. However, despite its enormous potential, microarray analysis of gene expression in solid tumors has been plagued by inconsistent results. A number of previous reports [1–15] have utilized microarray analysis to determine genome-wide changes in gene expression associated with head and neck cancer. Some of these efforts have been limited by the use of a small number of normal/tumor tissue pairs, the use of cultured cells or cell lines instead of cancerous and normal mucosa acquired in vivo, limited statistical analysis, or the lack of validation of selected genes. Thus, these studies have yielded greatly varying lists of candidate genes involved in the progression to head and neck squamous cell carcinoma (HNSCC), with only a minority of genes being common in different studies.

Lists of differentially expressed genes generated by microarray studies are likely to contain both false-positive and

* Corresponding author.

false-negative results. False positives (genes which are inaccurately identified as differentially expressed) are more worrying than false negatives (failure to identify differentially expressed genes) because the use of false-positive genes as diagnostic, prognostic, or therapeutic targets will likely yield misleading or inconsistent results, while false negatives simply mean that the list of differentially expressed genes is incomplete, rather than inaccurate.

For this reason, we hypothesized that a combinatorial approach will minimize the selection of false positives by using multiple statistical tests, each applicable to the dataset, but with differing profiles of strengths and weaknesses. Forty-two differentially expressed probe sets [38 genes and one expressed sequence tag (EST)] satisfied all statistical tests and were validated further using multi-probe set concordance, target-subunit agreement, hierarchical clustering, and real-time PCR. The favorable validation results support our hypothesis that the combinatorial approach identifies differentially expressed genes with few false positives at a likely cost of increased false negatives.

Materials and methods

Tissue harvesting

Tumors and normal tissues from 22 patients with histologically confirmed HNSCC were harvested. Patients who had previous treatment (radiotherapy or chemotherapy) for the index tumor or another head and neck primary within the past 5 years were excluded. Tissue measuring about 1 cm^3 was harvested in the operating room at the time of tumor resection. It was immediately snap-frozen in liquid nitrogen and transferred to –80 °C storage until RNA extraction. Normal tissue was harvested from the contralateral mucosa or from an uninvolved site distant from the tumor. The tumors represented a variety of head and neck subsites and stages of disease (table 1).

Table 1. Stages and subsites of HNSCC tumors used for microarray analysis.

No.	Tumor		TNM	Institution
	site	subsite		
1	OC	hard palate	$T_4N_{2c}M_0$	NYU
2	OC	mandibular alveolus	$T_4N_{2b}M_0$	NYU
3	OC	oral tongue	$T_3N_0M_0$	NYU
4	OC	buccal mucosa	$T_4N_0M_0$	NYU
5	OC	oral tongue	$T_1N_0M_0$	NYU
6	OC	maxillary alveolus	$T_4N_1M_0$	NYU
7	OC	lower lip	$T_2N_0M_0$	NYU
8	OC	floor of mouth	$T_4N_{2a}M_0$	NYU
9	OC	oral tongue	$T_2N_0M_0$	SHC
10	OC	gingiva	$T_2N_0M_0$	SHC
11	OC	buccal mucosa	$T_2N_0M_0$	SHC
12	OC	gingiva	$T_2N_0M_0$	SHC
13	OC/OP	floor of mouth/ soft palate	$T_1N_0M_0$	NYU
14	L	supraglottic larynx	$T_4N_1M_0$	NYU
15	L	supraglottic larynx	$T_4N_{2a}M_0$	NYU
16	L	glottic larynx	$T_4N_0M_0$	NYU
17	L	glottic larynx	$T_4N_0M_0$	NYU
18	OP	posterial pharyngeal wall	$T_2N_0M_0$	NYU
19	OP	tonsillar fossa	$T_4N_3M_0$	NYU
20	OP	soft palate	$T_4N_1M_0$	SHC
21	HP	post-cricoid region	$T_1N_0M_0$	NYU
22	SNC	maxillary sinus	$T_4N_3M_0$	NYU

NYU, New York University; SHC, Shanghai Ninth People's Hospital; OC, oral cavity; OP, oropharynx, L, larynx; HP, hypopharynx; SNC, sinonasal cavity.

RNA extraction

Total RNA was initially isolated from the tissue samples using TRIzol reagent (Gibco BRL Life Technologies, Rockville, Md.). Pre-chilled Trizol was added to deep-frozen specimens for polytron (Brinkmann Instruments, Westbury, N. Y.) homogenization followed by chloroform extraction and isopropyl alcohol precipitation. Washed total RNA pellets were resuspended in RNAse-free water and passed through an RNeasy spin column (Qiagen, Chatsworth, Calif.) for further purification. Eluted total RNAs were quantified with a portion of the recovered total RNA adjusted to a final concentration of 1 µg/µl. Quality of all starting total RNA samples was assessed prior to target preparation and processing steps. A small amount of each sample (typically 25–250 ng/well) was tested on the RNA Lab-On-A-Chip (Caliper Technologies, Mountain View, Calif.) and evaluated on the Agilent Bioanalyzer 2100 (Agilent Technologies, Palo Alto, Calif.).

Target labeling and microarray processing

Isolated total RNA samples were processed as recommended by Affymetrix (Affymetrix GeneChip Expression Analysis Technical Manual; Affymetric, Santa Clara, Calif.). In brief, single-stranded, and then double-stranded (ds) cDNA was synthesized from the poly(A)+ mRNA present in the isolated total RNA (10 µg total RNA starting material for each sample reaction) using the SuperScript Double-Stranded cDNA Synthesis Kit (Invitrogen, Carlsbad, Calif.) and poly(T)-nucleotide primers that contained a sequence recognized by T7 RNA polymerase. A portion of the resulting ds cDNA was used as a template to generate biotin-tagged cRNA from an in vitro transcription reaction, using the BioArray High-Yield RNA Transcript Labeling Kit (T7) (Enzo Diagnostics, Farmingdale, N. Y.). Fifteen micrograms of the resulting biotin-tagged cRNA was fragmented to strands

of 35–200 bases in length following prescribed protocols (Affymetrix GeneChip Expression Analysis Technical Manual). Subsequently, 10 µg of this fragmented target cRNA was hybridized at 45 °C with rotation for 16 h (Affymetrix GeneChip Hybridization Oven 320) to probes present on an Affymetrix HG-U95Av2 array. The GeneChip arrays were washed and then stained (streptavidin-phycoerythrin) in an Affymetrix Fluidics Station 400, followed by scanning on a Hewlett-Packard GeneArray scanner. The scanned results were quantified from '.DAT' files to '.CEL' files using MicroArray Suite software (MAS 5.0; Affymetrix) and further normalized with the Robust Multi-Chip Analysis (RMA) program (Gene Traffic software).

Statistical analysis of microarray data

RMA interchip-normalized probe set intensity data were processed with the following seven individual analysis tests (a–g): (a) t-test: parametric, variance not assumed equal; (b) Wilcoxon rank-sum: non-parametric (these two tests were performed with Gene Spring software with Benjamini and Hochberg correction to reduce false discovery rate); (c) Paired t test: BRB ArrayTools (v3.0.2) by NIH/NCI; (d) SAM (v1.10): Significant Analysis of Microarray, provided by Stanford University (www.-stat-class.stanford.edu/SAM/SAMservlet) [16]; (e) PPV: Predict Parameter Value function of Gene Spring software; (f) MDMR: Mininum Distance to Modal Ranking, as previously described [17–19]; (g) WEPO: WEighted Punishment on Overlap, as previously described [20].

Individual selections from the tests were grouped into three interrelated panels (A), (B), and (C), which are described further in Results. The three panels were then pooled together to identify 42 probe sets (within the intersection of three panels) representing differentially expressed genes (fig. 1).

Real-time, reverse transcription-polymerase Chain- reaction

Primer oligonucleotides for each PCR target were designed with LightCycler (LC) Probe Design software, and based on the target sequence of corresponding probe sequences downloaded from the Affymetrix web site. Optimum amplification conditions for each target, and expected product sizes were characterized and confirmed in preliminary experiments (not shown). Samples were tested with two tenfold dilutions each, and compared with five serially (tenfold) diluted standards. At least four standard points were used to establish a regression curve, and the regression curves were used to calculate the corresponding starting concentration of mRNA for each sample.

RT-PCR was performed with the LC System (Roche Diagnostics, Indianapolis, Ind.) using sequence-independent RNA Amplification Kit SYBR Green I (Roche). The RNA Amplification Kit (SYBR Green I) was stored at –20 °C and reagents were thawed just before the experiment. A master mixture (appropriate primers, enzyme cocktails, and reaction mixture) was prepared according to the total number of samples, standards, and negative controls, and loaded into LC capillaries (16 µl master mix per capillary), both of which were precooled. Template RNA (4 µl) was then added. Stopper-sealed capillaries were carefully inserted into the carousel which was centrifuged at 700 × g for 5 s (3000 rpm) by an LC micro-centrifuge. The carousel was placed into the LC instrument and the samples were cycled as experiment protocol programmed. Cycle-by-cycle PCR amplification of target molecules was monitored simultaneously in real time and analyzed by LC software. PCR products were verified by melting curve analysis for specificity and quantified according to the fit points method for user contribution in terms of

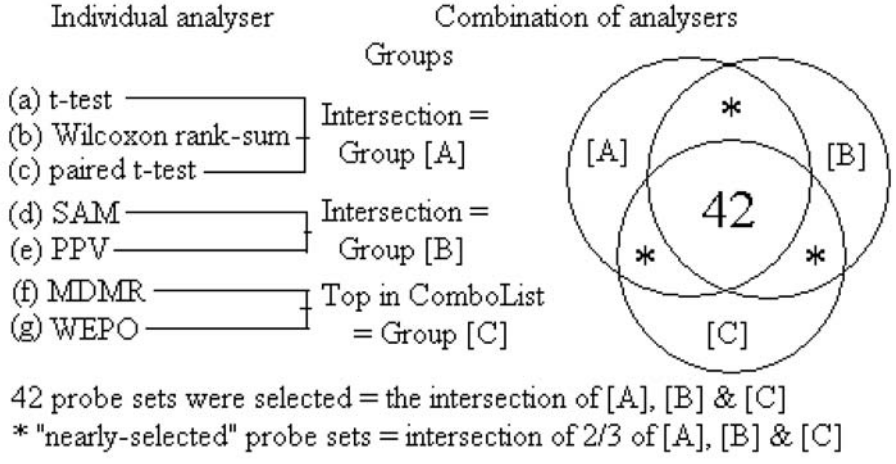

Figure 1. Combinatorial strategy for selection of differentially expressed genes. Forty-two probe sets were selected satisfying all statistical tests. Probe sets satisfying the next-lowest degree of stringency (satisfy two of three groups of tests) are marked as 'nearly-selected*'.

baseline, fit point number, p value, and standard curve optimization.

Results

Description of clinical samples

Paired samples consisted of HNSCC tumor tissue and a sample of clinically uninvolved mucosa from the same patient. All the patients were previously untreated for the index tumor. Tumor sites included the oral cavity, oropharynx, sinonasal cavities, hypopharynx, and larynx. The distribution of tumors among these sites is shown in table 1. Seventeen sample pairs were obtained from patients undergoing surgery at New York University Medical Center (NYU), and five pairs from the Ninth People's Hospital in Shanghai, China (SHC). All details of sample collection and processing were identical for samples collected at each institution.

Data analysis and selection of differentially expressed genes

The data derived from NYU and SHC were examined for comparability. When the 34 NYU samples were used as a training set for PPV, the resulting predictor probe sets correctly predicted tumor versus normal status for 31 of 34 NYU samples (3 unpredictable and 0 incorrectly predicted). When this panel of predictor probe sets was applied to the 10 SHC samples, 8 of 10 samples were correctly assigned (2 unpredictable and 0 incorrectly predicted). That is, the predictor genes selected from NYU samples also correctly predicted tumor versus normal status for SHC samples. Since tumor and normal samples can be distinguished in both the NYU and SHC samples by the same panel of predictor probe sets, we pooled samples from both institutions for the remainder of the analysis.

Seven independent statistical methods falling into three groups were used in this study. These seven methods were divided into three groups: standard statistical analyses (group A), significance/predictor analyses (group B), and rank-based analyses (group C).

Group A comprised three conventional statistical tests. By filtering out those probe sets which had an absolute intensity <50 across all samples, or for which the relative difference between the highest and lowest values for any of the 44 samples was <2, a total of 9281 probe sets were selected for further analyses: paired and unpaired t tests as well as the Wilcoxon rank-sum test. Probe sets with a significant difference ($p < 0.05$ with multiple correction) were selected for each test and then combined. A total of 246 probe sets (the intersection of these three lists) satisfied all of these conventional statistical tests.

Group B comprised two well-established gene chip data analysis algorithms: Significant Analysis of Microarray (SAM) and Predict Parameter Value (PPV). SAM identified (response type = paired data, number of permutations = 100, delta value = 1.50940, fold change = 2, and false positive = 0.52927) 182 probe sets which were significantly over- or under-expressed in tumors in comparison to the normal tissues. PPV selected probe sets which distinguish tumors from normal tissues. While all 44 samples were used as a training set, PPV selected (neighbors = 10 and cutoff p value = 0.4) 93 probe sets which successfully predict 41 of 44 samples (3 unpredictable and 0 incorrectly predicted). The intersection of the SAM and PPV selections yielded 55 shared probe sets.

Group C comprised of a ranking list (M) created by combining the two lists M1 and M2 derived from MDMR and WEPO, respectively, according to a previously developed framework. It combined M1 and M2 by taking the average of rankings for each probe set. The MDMR ranking method first ranks all the sample intensity values for each probe set and then computes the minimum number of adjacent swaps between this ranking and a modal ranking. This minimal number is assigned as the score of the probe set on the ranking list M1 [18]. In calculating the minimal number of adjacent swaps, WEPO introduces a z-score into the swapping ranking scheme to avoid loss of information [20]. M1 and M2 are two rank-order lists of all probe sets. The combined ranking list M takes the average rankings for each probe set in M1 and M2. The 200 top-ranked probe sets in list M were selected [for more details on ranking methods and combination of ranking lists, see refs [17, 18].

Final selection of probe sets was performed by taking the intersection of significant differentially expressed probe sets selected from the three groups of statistical tests. This yielded 42 probe sets which satisfied all statistical criteria.

Subgroup analysis was performed on our sample set to analyze whether gene expression differed significantly according to clinical variables such as race, tumor stage, tumor site, nodal positivity, and originating institution (NYU versus SHC). None of these variables was found to be significant, probably because of the limited sample size for each category.

Clustering of normal and tumor samples

All tumor and normal mucosa samples were subjected to hierarchical clustering to determine how effectively our protocol was able to differentiate between tumors (T) and normal tissues (N). When samples were clustered using all 12,642 probe sets, segregation between Ts and Ns was incomplete (fig. 2A). However, using only the 42 selected probe sets, hierarchical clustering was capable of clustering all Ts closer to each other than to any N (fig. 2B). That is, the tumor samples are more similar to each other, with respect to expression of the 42 probe sets, than to any normal tissue samples; supporting our assertion that these 42 probe sets are reliably differentially expressed in

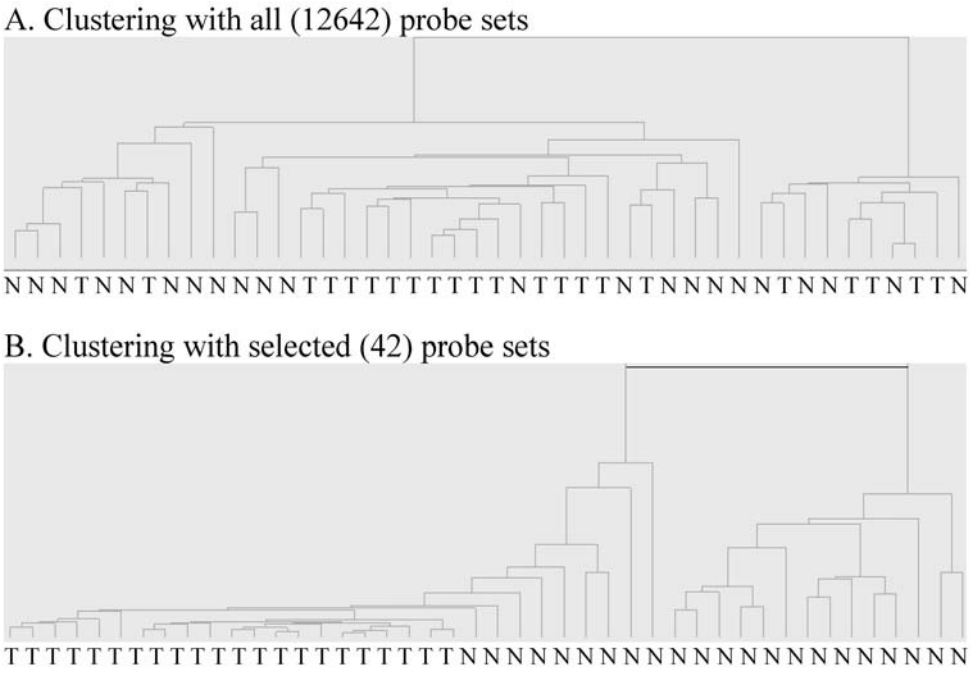

Figure 2. Hierarchical clustering analysis of all 22 pairs of samples. (*A*) Samples clustered using all 12,642 probe sets. (*B*) Samples clustered using only the 42 selected probe sets.

HNSCC. While some (13/22 = 59%) Ns were clustered into a completely separate group from the Ts, other (9/22 = 41%) Ns shared some similarity with Ts. This may be the result of field cancerization, a well-described phenomenon in HNSCC whereby exposure to environmental carcinogens (such as tobacco and alcohol) can cause precancerous genetic changes in clinically normal upper aerodigestive tract mucosa. Since the majority of patients in this study had a history of smoking and/or alcohol abuse, we were not able to perform subgroup analysis to determine whether the apparent effects of field cancerization were only observed in patients who smoked or drank.

Concordance between multiple probe sets for single-gene targets

In the Affymetrix U95A chip, some genes are represented by multiple probe sets whose sequences are derived from different regions of a same gene (multi-probe set genes) (table 2). Since changes in target expression should be reflected in all probe sets representing those genes, analysis of common expression patterns of probe sets for multi-probe set genes serves as a measure of internal consistency for the dataset. Among the 42 selected probe sets, 15 probe sets represent 12 separate genes with multiple probe sets on the U95A chip. These 12 multi-probe set genes are represented on the chip by a total of 27 probe sets; although 15 of these probe sets were selected in this study, the remaining 12 were not, because they failed to satisfy at least one of the three groups of statistical tests. Eight of the 12 unselected probe sets satisfied two out of three groups of statistical tests (referred to as 'nearly selected*' probe sets in figure 1, since they would have been chosen had the stringency of selection been slightly reduced). The remaining 4 unselected probe sets satisfied none of the three groups of statistical tests. The original expression patterns of each of those four unselected probe sets are shown in figure 3. While probe sets 32307_s_at and 41266_at were unselected, their pattern of expression was quite similar to the partner probe sets which were selected for their respective genes, suggesting that they were borderline cases, excluded because of the stringency of the combinatorial method of analysis. Two other unselected probe sets (32308_r_at and 35473_at) both differ substantially in sequence from their selected partner probe sets. For COL1A2, while three probe sets match all three potential COL1A2 sequences (pre-pro-COL1A2, pro-COL1A2, and COL1A2) with 100% homology, probe set 32308_r_at is only 94%, 87%, and 87% homologous to these sequences, respectively. For COL1A1, sequence homology between probe sets 35474_s_at (unselected) and 35473_at (selected) is 98%. Since the stringency of hybridization to the gene chip is so high that a single base pair mismatch can abolish the signal, it is not surprising that the results for these two probe sets were discordant. Thus all multi-probe sets involved in the selection were either validated or their failure to be validated was well-explained.

Table 2. Multi-probe sets and corresponding genes in the selected list.

No.	Target gene	Multi-probe sets		
		42 list	nearly selected*	unselected
1	APCC	32200_at 617_at		
2	OPN	2092_s_at 34342_s_at		
3	COL1A1	35474_s_at		35473_at
4	COL1A2	32306_g_at 32305_at		32307_s_at 32308_r_at
5	EMP-1	1321_s_at	37762_at	
6	FN1	31719_at	31720_s_at	
7	HPGD	32570_at	37322_s_at	
8	CYP3A5	37125_f_at	37124_i_at	
9	IL1RN	37603_at	31343_at	
10	SERPINH2	39166_s_at	39167_r_at	
11	CEACAM1	988_at	36082_at	
12	ITGA6	33410_at	33411_g_at	41266_at
Sum of probes		15	8	4

A total of 27 probe sets are involved with selected genes. For explanation of nearly selected see figure 1 legend.

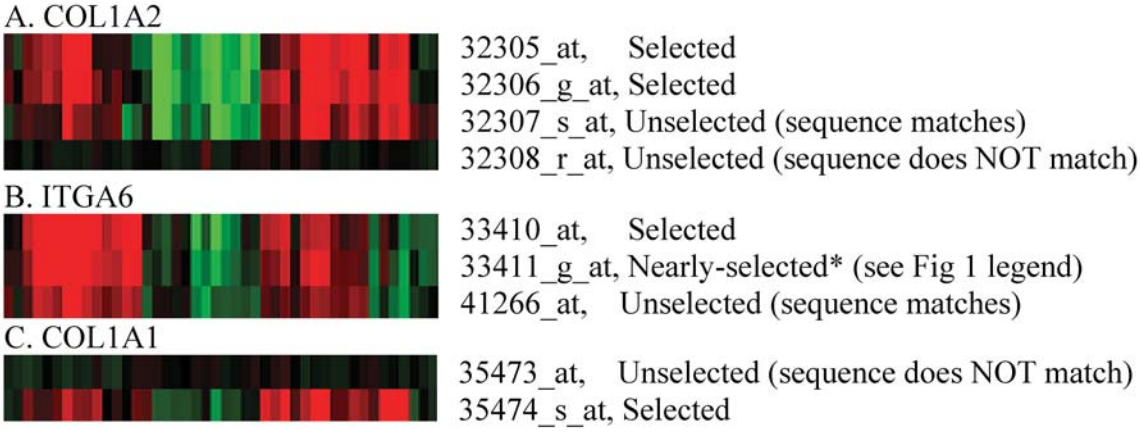

Figure 3. Ratio grids showing selected and unselected probe sets for multi-probe set genes. (*A*) COL1A2. (*B*) ITGA6. (*C*) COL1A1.

Target-subunit concordance

Type I collagen (COL1) and type IV collagen (COL4) are both heterotrimers of two subunits, $\alpha 1$ and $\alpha 2$ in a 2:1 ratio [21]. Both $\alpha 1$ and $\alpha 2$ of COL1 and COL4 were selected in the list of 42 highly significant genes. Such co-selection of structurally related gene targets strongly validates these results. ITGA6 is the $\alpha 6$ chain of integrin VLA6 ($\alpha_6\beta_1$) and has been demonstrated to be the controlling subunit for this heterodimer expression. Thus, the appearance of ITGA6 alone in the 42 list is consistent with the known mechanism of regulation of integrin VLA6. Although COL5A2 was selected, its partner subunit COL5A1 was not. Thus COL5A2 is the only likely false-positive result, barring an alternative explanation, such as the existence of an A2 heterotrimer, or A2 being the sole expression-controlling subunit. In summary, both subunits of COL1, COL4, and ITGA6 were well-validated, while the status of COL5A2 is questionable. However, if COL5A2 is truly a false positive, then COL5A1 must be a true negative, yielding a total of seven of eight involved subunits validated.

Real-time PCR validation of selected differentially expressed genes

Genes found to be differentially expressed by microarray analysis should be validated with other well-established technologies for analyzing gene expression. We used real-time PCR, which allows quantitative analysis of mRNA expression, for this purpose. From 42 differentially expressed probe sets selected by microarray analysis, 8 (4 up-regulated and 4 down-regulated) target genes were chosen randomly from 33 probe sets with lower stringency in the 42 list. That is, 9 probe sets (6 genes) which were well-validated at the in silico level (double selection of probe sets or subunits for the same gene) were excluded from consideration to ensure that real-time

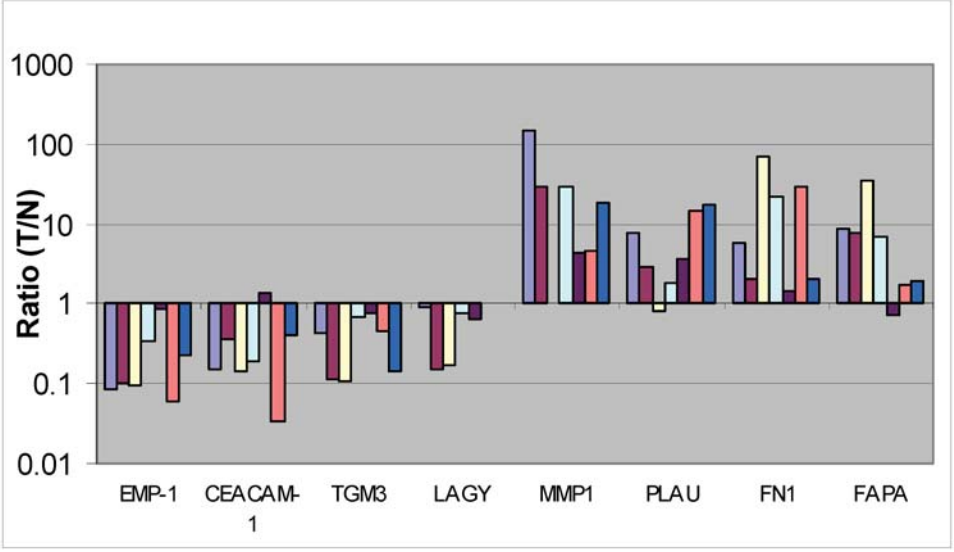

Figure 4. Real-time PCR analysis of 8 from 38 selected genes. Four up-regulated in HNSCC (MMP1, PLAU, FN1, and FAPA) and 4 down-regulated (EMP1, CEACAM1, TGM3, and LAGY). Data are presented on a logarithmic scale. Each column stands for one pair of samples. The three columns on the left represent samples analyzed by microarray, whereas the right columns represent independently obtained samples.

PCR was used to evaluate 'average' selected genes, rather than the best candidates. Seven paired (tumor and normal) samples were tested for the eight representative genes. Three pairs of samples were among the 22 pairs already subjected to microarray analysis. The other four sample pairs were obtained from four additional patients who were not part of the microarray analysis.

As shown in figure 4, while there was considerable sample-to-sample variability in the magnitude of differential expression, the overall direction of differential expression was remarkably consistent. Among the eight genes assayed, three (EMP1, TGM3 and FN1) yielded concordance for 7/7 sample pairs tested. The other five genes produced discordant results (either no significant change or a slight change in the opposite direction) for one (four genes) or three (one gene: LAGY) sample pairs. When analysis was restricted to only the three pairs of samples which were also tested by microarray, real-time PCR exactly paralleled the microarry results, including identification of the same sample pairs as discordant for genes MMP-1 and PLAU. The gross discordance rate for RT-PCR is about 1 out of 7 (14.3%). When heterogeneity within the microarray data was reviewed, for each of the 42 probe sets selected, among 22 paired samples, an average of three pairs (13.6%) were found to show no significant change or a slight change in the opposite direction. Thus the amount of heterogeneity within the microarray data was found to be consistent with the heterogeneity seen in real-time PCR. Thus, our selection of these eight genes in the microarray experiment was well-validated by real-time PCR.

Discussion

Study design and overall strategy

In this study, we chose to analyze gene expression of the entire cell population resident in HNSCC tumors, rather than using laser capture microscopy (LCM) to isolate HNSCC cells from other tumor components. Our rationale is that changes in any tumor constituents (including stromal, immune/inflammatory, and other components) may potentially be exploited for diagnostic or therapeutic purposes. For example, the gene MAL was among the list of genes selected in this study. MAL expression is limited to mature T lymphocytes [18], and changes in MAL expression are most likely related to tumor-infiltrating lymphocytes rather than carcinoma cells. However, as long as the expression of MAL within tumor tissue differs from that of normal mucosa, it may potentially serve as a candidate prognostic marker.

Although each of the statistical methods used in this study possesses its own unique profile of advantages and disadvantages, they are all appropriate for application to our dataset. The choice of a combinatorial approach to analysis of microarray data reflects our desire to minimize the number of false-positive results, even at the expense of increased false-negative genes. To prove this hypothesis, a series of rigorous validation tests were applied. The results of validation strongly support our hypothesis: genes selected in this study were well-validated by multiple approaches; however, some oncogenes that have been repeatedly associated with HNSCC (such as cyclin D1 and EGF-R) were missed. The lack of false

positives (selected probe sets which failed validation) confirms the stringency of the combinatorial method of selection. The failure to select several genes with strong associations to HNSCC (false negatives) may be the result of two possibilities: excessive stringency of the combinatorial method, or heterogeneity of our HNSCC samples.

While all HNSCCS share many clinical and pathologic characteristics, HNSCC at different subsites can be quite heterogeneous with respect to treatment response and prognosis. This clinical heterogeneity implies co-existing genetic heterogeneity between tumors from different subsites, as well as tumors with varying levels of differentiation. While 22 is a comparatively large number of sample pairs, the number is too small to perform subgroup analysis according to tumor subsite. Common oncogenes and tumor supressor genes which do not appear on our final list were possibly expressed in too few samples to satisfy the high degree of stringency required by combinatorial analysis.

Rationale of the approach to selection of differentially expressed genes

The combinatorial approach used in this study includes seven independent statistical methods. Each individual method is applicable to our dataset, but has a different profile of strengths and weaknesses.

Parametric tests such as the t test are based on differences of the means in the two groups of normal and tumor tissues (unpaired tests) or on the means of the differences (paired t test). Non-parametric approaches such as the Wilcoxin rank-sum test (equivalent to the Mann-Whitney test) are based on the difference of rank sums in the two groups. Parametric approaches work better for variables which follow a normal distribution. Non-parametric methods are better suited for analysis if the rank sums or calculated differences better reflect the nature of the samples, the type of disease and meaning of biological and physiological systems. SAM is a powerful and frequently used tool developed specifically for microarray data analysis. It uses repeated permutations to measure the strength of the relationship between gene expression and the response variables. A main advantage of SAM is that it allows using permutations on a small number of samples to increase the significance of the analysis. However, when the sample size is large and the distributions of the variables can not be pre-determined, this advantage becomes less attractive. PPV is a supervised learning method which uses the k-nearest neighbor as a discrimination method; it is appropriate for identifying genes that discriminate tumors from normals. PPV is limited by so-called over-fitting, which means the number of parameters of the model is too large relative to the number of cases of specimens available. MDMR and WEPO are ranking systems which assign a score to each probe. Both methods rank the samples within two separate groups for each probe. Instead of calculating a p value using the rank-sum of samples in the tumor group, they identify differentially expressed genes by computing the distance between this ranking and a modal ranking. Both MDMR and WEPO are non-parametric ranking systems which do not assume any particular pattern of distribution. However, the appropriate choice of the 'distance' and the 'ideal ranking' to be used depends closely on the experiment and the objective of the microarray study.

A combinatorial approach was used to take advantage of the diverse strengths and to balance the respective weaknesses of each individual test. The 246 probe sets on the group A list satisfied a threshold of p value in all of the three conventional statistical tests (a), (b) and (c). Group B contained 55 probe sets which lie in the intersection of the (d) list (satisfied several SAM conditions) and (e) list (satisfied PPV test). The combination of the MDMR and WEPO lists used the concept of data fusion and rank combination (by taking the average of rankings for each probe set). This combination method is different from those employed in group A and B. It considers each rank list as a member of a rank space. By combining two or more rank lists, one aims to obtain an overall ranking that yields an improved result no matter how inconsistent the individual rankings may be [17, 18, 22].

The proof of our hypothesis that the genes selected by this stringent strategy contain fewer false positives at the cost of more false negatives lies in our validation results.

Validation of selected genes

We validated microarray-selected genes by two methods: *in silico* analysis (to assess internal consistency of selected probe sets) and real-time PCR (a well-established quantitative assay of gene expression). *In silico* analysis supported our probe set selections, since all 27 involving multi-probe sets, and 7 of 8 genes amenable to analysis of target-subunit concordance were validated. Eight genes, representing 24% (8/33) of selected genes, were validated by real-time PCR analysis, using seven pairs of samples. While the results from three sample pairs (which have been tested with chips) validated the consistency between microarray and RT-PCR, that from four independent sample pairs (which were not used in the microarray analysis, see figure 4 legend) further verify that genes selected by microarray analysis are reliably over- or under-expressed in HNSCC tissue.

When validating microarray results, considering the consistency of results between studies may be constructive. We compared our panel of selected genes with those selected by four other rigorous microarray studies of HNSCC [6–8, 12]. Those four studies met the following criteria: (i) they were genome-wide microarray analyses of gene expression in primary HNSCC versus normal mucosa; (ii) genes were selected by using multiple sta-

tistical tests, or other comparatively stringent criteria; (iii) they validated microarray results with real-time PCR. A total of 26 genes were selected in two or more studies, with three genes (KRT4, NMU, and TGM3) selected in three or more studies. All three of these genes are in our selection list. Of the 39 genes identified in our analysis, 15 (39%) were also identified in at least one other study. Considering the variability in tumor types, acquisition techniques, and statistical methods used in each of these five studies, the degree of concordance across analyses is notable.

Annotation and relevance of selected genes to malignancy

The identities of the selected genes are listed and annotated in table 3. These 42 probe sets correspond to 38 previously characterized genes (three genes were selected twice) and one EST sequence. The 38 differentially expressed genes were divided into four broad categories: tumor-associated antigens, enzymes, structural molecules, and miscellaneous molecules. In all, more than half of the selected genes play established roles in oncogenesis or are otherwise cancer related, as discussed below.

The tumor-associated antigen group consists of five genes with clear relevance to oncogenesis. Over-expression of osteopontin (OPN, with two of two probe sets selected) has been found to be inversely correlated with tumor pO_2 and an indicator of tumor regression. Serum levels of OPN have thus been reported to be a candidate prognostic marker for HNSCC [23]. CEACAM1 and CEACAM5 are both members of the carcinoembryonic antigen (CEA) family of adhesion molecules. CEACAM1 is better described, and has been characterized as a tumor suppressor gene [24, 25] involved in prostate, breast, and colon tumors. Lung cancer-associated gene Y (LAGY), which has been shown to be significantly down-regulated in lung cancer [26] is also shown in our study to be down-regulated in HNSCC. EMP1 has been shown to be a marker of cancer progression in mammary carcinoma cell lines, with a well-established correlation between its expression and invasive and metastatic potential [27].

The enzyme group contains 12 genes, of which 9 have established relevance to malignancy. ACPP (which had two probe sets both selected in this study) has already shown clinical utility as a serum marker for prostate cancer [28]. Several other tumor-expressed proteolytic enzymes are known for their ability to degrade extracellular matrix and facilitate tumor invasion and metastasis: MMPs (metalloproteinases), PLAU (or uPA, urokinase-type plasminogen activator), and serine proteases [29] are major proteolytic enzymes involved in degradation of extracellular matrix by tumors. MMP1, also known as collagenase 1, has been found to be consistently over-expressed in a number of malignancies [30, 31] and has been found to be up-regulated in prior gene chip analyses of HNSCC [12]. PLAU has been described as an independent prognostic marker for a variety of malignancies [32]. The serine protease PP11, as well as SERPINH2 and SPINK5 (both serine protease inhibitors) are reported to be up- or down-regulated during carcinogenesis and tumor invasion [33]. Fibroblast activation protein alpha (FAPA) contains a catalytic domain which is highly conserved in serine proteases, and its expression correlates with the invasiveness of human melanoma and carcinoma cells [34]. High expression of GPX3 has been reported in ovarian cancer [35]. Finally, transglutaminase 3 (TGM3) is involved in assembly of structural proteins into the cornified cell envelope [36], a structure characteristic of terminally differentiated stratified squamous epithelium. This role in normal epithelium is consistent with its under-expression in HNSCC in this study

Five of 11 genes in the structural molecule group have previously established oncogenic significance. Expression of fibronectin 1 (FN1) is thought to accompany tumor growth, and its proteolytic breakdown products have been shown to potentiate malignant transformation [37]. SCEL, down-regulated in HNSCC in this study, is a precursor of the cornified cell envelope [38], and periplakin (PPL) is another component of the cornified cell envelope [39] which is cross-linked by transglutaminase 3 (TGM3, also down-regulated in HNSCC in this study). CRK4 and CRK13 are both cytokeratins known to be present in well-differentiated epithelium. All these markers of terminally differentiated epithelium (SCEL, PPL, CRK4, and CRK13), as well as the related enzyme TGM3, were found to be down-regulated in HNSCC in this study, as would be predicted by their roles in normal mucosa.

Finally, several other miscellaneous genes found to be differentially expressed in this analysis also have direct or indirect relevance to oncogenesis. C5ORF13, also known as P311, has been reported to be over-expressed by invasive glioblastoma cells, and in vitro studies have confirmed this role [40]. Allelic variation of the parathyroid hormone-like hormone (PTHLH) gene has been reported to be associated with increased risk and worse prognosis of lung cancer [41]. Over expression of PTHLH has also been shown to cause serum hypercalcemia associated with malignancy [42]. Polymorphism of IL1RA has been reported to be involved in the induction of several solid tumors [43]. The transcription factor PITX1 has been reported to be down-regulated in pituitary adenomas [44].

Due to the stringency of our selection, some well-known oncogenes were not identified, while the selected genes include several keratinocyte differentiation markers. However, while we do not believe that our final list of differentially expressed genes captures all genes which play a pathogenic role in HNSCC, it is likely to contain genes which do. For example, support in the literature

Table 3. List of selected genes (with brief annotation and fold change information).

Category	Up-regulated in tumors			Down-regulated in tumors			Category
	probe ID	target	FC*	probe ID	target	FC*	
[I]	35474_s_at	COL1A1	2.7	39657_at	KRT 4	–16.9	[I]
Structural	32305_at	COL1A2	3.1	36883_at	KRT 13	– 7.6	see left
or	32306_g_at	COL1A2	3.3	36890_at	PPL	– 4.6	
associated	39333_at	COL4A1	2.6	35105_at	SCEL	– 5.0	
	36659_at	COL4A2	2.5	32200_at	ACPP	– 2.4	
	38420_at	COL5A2	2.9	617_at	ACPP	– 2.6	
	31719_at	FN1	3.5	37125_f_at	CYP3A5	– 3.5	[II]
	38442_at	MFAP2	1.9	529_at	DUSP5	– 2.3	Enzyme
				770_at	GPX3	– 2.7	or
[II]	39945_at	FAPA	2.9	32570_at	HPGD	– 3.0	inhibitor
see right	38428_at	MMP1	4.7	37093_at	PP11	– 6.3	
	37310_at	PLAU	2.8	40315_at	SPINK5	– 9.7	
	39166_s_at	SERPINH2	2.0	32868_at	TGM3	–15.9	
				988_at	CEACAM1	– 4.3	[III]
[III]	2092_s_at	OPN	5.8	1582_at	CEACAM5	– 3.9	Tumor-
see right	34342_s_at	OPN	6.0	39698_at	LAGY	– 6.6	associated
				1321_s_at	EMP1	– 3.9	antigen
[IV]	39710_at	C5ORF13	1.9	38242_at	BLNK	– 2.9	
Other	33410_at	ITGA6	2.8	37603_at	IL1RN	– 4.9	
	1837_at	NA	1.8	41644_at	KIAA0790	– 2.3	[IV]
	615_s_at	PTHLH	2.5	38051_at	MAL	–14.4	Other
				33483_at	NMU	– 3.6	
				37920_at	PITX1	– 3.2	
				32139_at	ZNF185	– 3.4	

* FC, fold change (expressed as mean of all 22 pairs). Positive values indicate up-regulation (tumors > normals) and negative values indicate down-regulation (normal > tumors).
ACPP, acid phosphatase, prostate; BLNK, B-cell linker; C5ORF13, chromosome 5 open reading frame 13; CEACAM#, carcinoembryonic antigen-related cell adhesion molecule #; COL#A#, collagen, type #, alpha # ; CYP3A5, cytochrome P450, subfamily IIIA; DUSP5, dual-specificity phosphotase 5; EMP1, epithelial membrane protein 1 (tumor-associated membrane protein); FAPA, fibroblast activation protein α; FN1, fibronectin 1; GPX3, glutathione peroxidase 3; HPGD, hydroxyprostaglandin dehydrogenase 15-(NAD); IL1RN, interleukin 1 receptor antagonist; ITGA6, integrin alpha 6; KRT#, keratin #; LAG, lung cancer-associated Y protein; MAL, T-cell differentiation protein MAL isoform; MFAP2, microfibrillar-associated protein 2; MMP1, matrix metalloproteinase 1; NA, not applied, i.e., an EST; NMU, neuromedin U; OPN, osteopontin; PITX1, paired-like homeodomain transcription factor 1; PLAU, plasminogen activator, urokinase; PP11, placental protein 11, protease, serine, 22; PPL, periplakin; PTHLH, parathyroid hormone like hormone; SCEL, sciellin; SERPINH2, serine (or cycteine) proteinase inhibitor, clade H2; SPINK5, serine protease inhibitor, Kazal type, 5; TGM3, transglutaminase 3 ; ZNF185, zinc finger protein 185.

for a role of OPN in HNSCC is strong [23]. This gene passed multiple levels of validation to be selected in our study, including 'double picking' of two of its two probe sets. While other selected genes may also play role(s) in HNSCC pathogenesis, detailed functional analysis of selected genes was beyond the scope of this study.

Conclusions

Combinatorial analysis of microarray data from 22 paired tumor/normal samples from HNSCC patients yielded 42 probe sets (representing 38 genes plus one EST) which satisfy highly stringent criteria for significant over- or under-expression in HNSCC tissue. A subset of selected genes was validated by *in silico* analysis and real-time PCR of both chip-tested and independently obtained sample pairs to confirm their differential expression in HNSCC versus normal tissues. Several genes identified in the present study were also found to be concordant with genes identified in other HNSCC microarray analyses of similar rigor. The strength of validation of these targets suggests that the intersection of mathematically distinct statistical methods can yield a reliable list of differentially expressed genes containing fewer false positives than a single method. This approach may have broad utility in the analysis of gene expression in HNSCC and other malignancies.

Acknowledgements. Supported by the 'George E. Hall Endowment Fund for head and neck cancer research' and 'The National Natural Science Foundation of China, grant No. 30330580,30171014.' We thank the UCI DNA MicroArray Facility (University of California, Irvine) for chip processing, Dr. J. Z. Xiang's group (Cornell University) for instructions in data analysis, and Dr. J. C. Sok (University of Pittsburg) for his contributions to the initiation of this study.

1 Ginos M. A., Page G. P., Michalowicz B. S., Patel K. J., Volker S. E., Pambuccian S. E. et al. (2004) Identification of a gene expression signature associated with recurrent disease in squamous cell carcinoma of the head and neck. Cancer Res. **64:** 55–63

2 Mendez E., Cheng C., Farwell D. G., Ricks S., Agoff S. N., Futran N. D. et al. (2002) Transcriptional expression profiles of oral squamous cell carcinomas. Cancer **95:** 1482–1494

3 Hartmann K. A., Modlich O., Prisack H. B., Gerlach B. and Bojar H. (2002) Gene expression profiling of advanced head and neck squamous cell carcinomas and two squamous cell carcinoma cell lines under radio/chemotherapy using cDNA arrays. Radiother. Oncol. **63:** 309–320

4 Dimitroulakos J., Marhin W. H., Tokunaga J., Irish J., Gullane P., Penn L. Z. et al. (2002) Microarray and biochemical analysis of lovastatin-induced apoptosis of squamous cell carcinomas. Neoplasia **4:** 337–346

5 Yoo G. H., Piechocki M. P., Ensley J. F., Nguyen T., Oliver J., Meng H. et al. (2002) Docetaxel induced gene expression patterns in head and neck squamous cell carcinoma using cDNA microarray and PowerBlot. Clin. Cancer Res. **8:** 3910–3921

6 Hwang D., Alevizos I., Schmitt W. A., Misra J., Ohyama H., Todd R. et al. (2003) Genomic dissection for characterization of cancerous oral epithelium tissues using transcription profiling. Oral Oncol. **39:** 259–268

7 El-Naggar A. K., Kim H. W., Clayman G. L., Coombes M. M., Le B., Lai S. et al. (2002) Differential expression profiling of head and neck squamous carcinoma: significance in their phenotypic and biological classification. Oncogene. **21:** 8206–8219

8 Gonzalez H. E., Gujrati M., Frederick M., Henderson Y., Arumugam J., Spring P. W. et al. (2003) Identification of 9 genes differentially expressed in head and neck squamous cell carcinoma. Arch. Otolaryngol. – Head Neck Surg. **129:** 754–759

9 Squire J. A., Bayani J., Luk C., Unwin L., Tokunaga J., MacMillan C. et al. (2002) Molecular cytogenetic analysis of head and neck squamous cell carcinoma: by comparative genomic hybridization, spectral karyotyping, and expression array analysis. Head Neck **24:** 874–887

10 Villaret D. B., Wang T., Dillon D., Xu J., Sivam D., Cheever M. A. et al. (2000) Identification of genes overexpressed in head and neck squamous cell carcinoma using a combination of complementary DNA subtraction and microarray analysis. Laryngoscope **110:** 374–381

11 Zhang X., Liu Y., Gilcrease M. Z., Yuan X. H., Clayman G. L., Adler-Storthz K. et al. (2002) A lymph node metastatic mouse model reveals alterations of metastasis-related gene expression in metastatic human oral carcinoma sublines selected from a poorly metastatic parental cell line. Cancer **95:** 1663–1672

12 Alevizos I., Mahadevappa M., Zhang X., Ohyama H., Kohno Y., Posner M. et al. (2001) Oral cancer in vivo gene expression profiling assisted by laser capture microdissection and microarray analysis. Oncogene **20:** 6196–6204

13 Hasina R., Hulett K., Bicciato S., Di Bello C., Petruzzelli G. J. and Lingen M. W. (2003) Plasminogen activator inhibitor-2: a molecular biomarker for head and neck cancer progression. Cancer Res. **63:** 555–559

14 Al Moustafa A. E., Alaoui-Jamali M. A., Batist G., Hernandez-Perez M., Serruya C., Alpert L. et al. (2002) Identification of genes associated with head and neck carcinogenesis by cDNA microarray comparison between matched primary normal epithelial and squamous carcinoma cells. Oncogene **21:** 2634–2640

15 Leethanakul C., Patel V., Gillespie J., Pallente M., Ensley J. F., Koontongkaew S. et al. (2000) Distinct pattern of expression of differentiation and growth-related genes in squamous cell carcinomas of the head and neck revealed by the use of laser capture microdissection and cDNA arrays. Oncogene **19:** 3220–3224

16 Tusher V. G., Tibshirani R. and Chu G. (2001) Significance analysis of microarrays applied to the ionizing radiation response. Proc. Nat. Acad. Sci. USA **98:** 5116–5121 [erratum in Proc. Natl. Acad. Sci. USA (2001) **98:** 10515

17 Hsu D. F., Shapiro J. and Taksa I. (2002) Methods of data fusion in information retrieval: rank vs score combination. DIMACS Tech. Rep. 58 (www.dimacs.rutgers.edu)

18 Chuang H. Y., Liu H. F., Brown S., McMunn-Coffran C., Kao C. Y. and Hsu D. F. (in press) Identifying significant genes from microarray data. Proc. IEEE Bioinform. Bioeng.'04, IEEE Computer Society Press, Los Alamitos, USA

19 Park P. J., Pagano M. and Bonetti M. (2001) A nonparametric scoring algorithm for identifying informative genes from microarray data. Pacific Symp. Biocomput.: 52–63

20 Chuang H. Y., Tsai H. K., Tsai Y. F. and Kao C. Y. (2003) Ranking genes for discriminability on microarray data. J. Inform. Sci. Eng. **19:** 47–58

21 Burgeson R. E. and Nimni M. E. (1992) Collagen types: molecular structure and tissue distribution. Clin. Orthopaed. Rel. Res. **282:** 250–72

22 Marden J. I. (1995) Analysing and Modeling Rank Data, Chapman & Hall, London, UK

23 Le Q. T., Sutphin P. D., Raychaudhuri S., Yu S. C., Terris D. J., Lin H. S. et al. (2003) Identification of osteopontin as a prognostic plasma marker for head and neck squamous cell carcinomas. Clin. Cancer Res. **9:** 59–67

24 Estrera V. T., Chen D. T., Luo W., Hixson D. C. and Lin S. H. (2001) Signal transduction by the CEACAM1 tumor suppressor: phosphorylation of serine 503 is required for growth-inhibitory activity. J. Biol. Chem. **276:** 15547–15553

25 Volpert O., Luo W., Liu T. J., Estrera V. T., Logothetis C. and Lin S. H. (2002) Inhibition of prostate tumor angiogenesis by the tumor suppressor CEACAM1. J. Biol. Chem. **277:** 35696–35702

26 Chen Y., Petersen S., Pacyna-Gengelbach M., Pietas A. and Petersen I. (2003) Identification of a novel homeobox-containing gene, LAGY, which is downregulated in lung cancer. Oncology **64:** 450–458

27 Gnirke A. U. and Weidle U. H. (1998) Investigation of prevalence and regulation of expression of progression associated protein (PAP). Anticancer Res. **18:** 4363–4369

28 Fishman W. H. (1995) The 1993 ISOBM Abbott Award Lecture: Isozymes, tumor markers and oncodevelopmental biology. Tumour Biol. **16:** 394–402

29 Noel A., Gilles C., Bajou K., Devy L., Kebers F., Lewalle J. M. et al. (1997) Emerging roles for proteinases in cancer. Invasion Metastasis **17:** 221–239

30 Mueller M. M. and Fusenig N. E. (2002) Tumor-stroma interactions directing phenotype and progression of epithelial skin tumor cells. Differentiation **70:** 486–497

31 Schiemann S., Schwirzke M., Brunner N. and Weidle U. H. (1998) Molecular analysis of two mammary carcinoma cell lines at the transcriptional level as a model system for progression of breast cancer. Clin. Exp. Metastasis **16:** 129–139

32 Reuning U., Magdolen V., Wilhelm O., Fischer K., Lutz V., Graeff H. et al. (1998) Multifunctional potential of the plasminogen activation system in tumor invasion and metastasis. Int. J. Oncol. **13:** 893–906

33 Yudate T., Isaka K., Suzuki Y., Takada J., Hosaka M., Shiraishi K. et al. (1995) mRNA expression and protein localization of placental tissue protein 11, 12, 19 in gynecologic malignant tumors. Nippon Sanka Fujinka Gakkai Zasshi – Acta Obstet. Gynaecol. Jpn **47:** 1248–1254

34 Goldstein L. A., Ghersi G., Pineiro-Sanchez M. L., Salamone M., Yeh Y., Flessate D. et al. (1997) Molecular cloning of seprase: a serine integral membrane protease from human melanoma. Biochim. Biophy. Acta. **1361:** 11–19

35 Joncourt F., Buser K., Altermatt H., Bacchi M., Oberli A. and Cerny T. (1998) Multiple drug resistance parameter expression in ovarian cancer. Gynecol. Oncol. **70:** 176–182
36 Candi E., Paci M., Oddi S., Paradisi A., Guerrieri P and Melino G. (2000) Ordered structure acquisition by the N- and C-terminal domains of the small proline-rich 3 protein. J. Cell. Biochem. **77:** 179–185
37 Labat-Robert J. (2002) Fibronectin in malignancy. Semin. Cancer Biol. **12:** 187–195
38 Champliaud M. F., Burgeson R. E., Jin W., Baden H. P. and Olson P. F. (1998) cDNA cloning and characterization of sciellin, a LIM domain protein of the keratinocyte cornified envelope. J. Biol. Chem. **273:** 31547–31554
39 Leung C. L., Green K. J. and Liem R. K. (2002) Plakins: a family of versatile cytolinker proteins. Trends Cell Biol. **12:** 37–45
40 Mariani L., McDonough W. S., Hoelzinger D. B., Beaudry C., Kaczmarek E., Coons S. W. et al. (2001) Identification and validation of P311 as a glioblastoma invasion gene using laser capture microdissection. Cancer Res. **61:** 4190–4196
41 Manenti G., Nomoto T., De Gregorio L., Gariboldi M., Stefania Falvella F., Nagao M. et al. (2000) Predisposition to lung tumorigenesis. Toxicol. Lett. **112–113:** 257–263
42 Suva L. J., Winslow G. A., Wettenhall R. E., Hammonds R. G., Moseley J. M., Diefenbach-Jagger H. et al. (1987) A parathyroid hormone-related protein implicated in malignant hypercalcemia: cloning and expression. Science. **237:** 893–896
43 Sehouli J., Mustea A., Konsgen D., Katsares I. and Lichtenegger W. (2002) Polymorphism of IL-1 receptor antagonist gene: role in cancer. Anticancer Res. **22:** 3421–3424
44 Skelly R. H., Korbonits M., Grossman A., Besser G. M., Monson J. P., Geddes J. F. et al. (2000) Expression of the pituitary transcription factor Ptx-1, but not that of the trans-activating factor prop-1, is reduced in human corticotroph adenomas and is associated with decreased alpha-subunit secretion. J. Clin. Endocrinol. Metab. **85:** 2537–2542