Gene Expression Profile of Papillary Thyroid Cancer: Sources of Variability and Diagnostic Implications

Barbara Jarząb,[1] Małgorzata Wiench,[1] Krzysztof Fujarewicz,[5] Krzysztof Simek,[5] Michał Jarząb,[2] Małgorzata Oczko-Wojciechowska,[1] Jan Włoch,[3] Agnieszka Czarniecka,[3] Ewa Chmielik,[4] Dariusz Lange,[4] Agnieszka Pawlaczek,[1] Sylwia Szpak,[1] Elżbieta Gubała,[1] and Andrzej Świerniak[5]

Departments of [1]Nuclear Medicine and Endocrine Oncology, [2]Tumor Biology, [3]Oncological Surgery, and [4]Tumor Pathology, Maria Skłodowska-Curie Memorial Cancer Center and Institute of Oncology, Gliwice Branch; and [5]Automatic Control, Silesian University of Technology, Gliwice, Poland

Abstract

The study looked for an optimal set of genes differentiating between papillary thyroid cancer (PTC) and normal thyroid tissue and assessed the sources of variability in gene expression profiles. The analysis was done by oligonucleotide microarrays (GeneChip HG-U133A) in 50 tissue samples taken intraoperatively from 33 patients (23 PTC patients and 10 patients with other thyroid disease). In the initial group of 16 PTC and 16 normal samples, we assessed the sources of variability in the gene expression profile by singular value decomposition which specified three major patterns of variability. The first and the most distinct mode grouped transcripts differentiating between tumor and normal tissues. Two consecutive modes contained a large proportion of immunity-related genes. To generate a multigene classifier for tumor-normal difference, we used support vector machines-based technique (recursive feature replacement). It included the following 19 genes: *DPP4*, *GJB3*, *ST14*, *SERPINA1*, *LRP4*, *MET*, *EVA1*, *SPUVE*, *LGALS3*, *HBB*, *MKRN2*, *MRC2*, *IGSF1*, *KIAA0830*, *RXRG*, *P4HA2*, *CDH3*, *IL13RA1*, and *MTMR4*, and correctly discriminated 17 of 18 additional PTC/normal thyroid samples and all 16 samples published in a previous microarray study. Selected novel genes (*LRP4*, *EVA1*, *TMPRSS4*, *QPCT*, and *SLC34A2*) were confirmed by Q-PCR. Our results prove that the gene expression signal of PTC is easily detectable even when cancer cells do not prevail over tumor stroma. We indicate and separate the confounding variability related to the immune response. Finally, we propose a potent molecular classifier able to discriminate between PTC and nonmalignant thyroid in more than 90% of investigated samples. (Cancer Res 2005; 65(4): 1587-97)

Introduction

Tumor gene expression profiling by DNA microarrays has brought new important clues to our understanding of cancer pathophysiology and simultaneously has provided clinically valuable information on many malignant neoplasms. Gene expression studies in thyroid tumors also contribute to the growing experience on molecular differences between various types of thyroid disease. Differences in the gene expression of functional and nonfunctional benign thyroid tumors were described by Eszlinger et al. (1). In the first study of papillary thyroid cancer (PTC) by high density DNA microarrays, published by Huang et al. (2), its gene expression profile was highly consistent and some of the genes listed have been confirmed as valuable PTC markers (3).

Both papillary and follicular cancers, despite being evidently malignant, retain many properties of their cells of origin and in this way they are somewhat different from malignant tumors of other organs. From this point of view, thyroid cancer cells were expected to be less effectively distinguished by gene expression profiling from nontransformed tissue (1, 4, 5). However, the opposite has been proven (2, 6, 7), encouraging further studies on the clinical significance of microarray-based analyses.

An additional level of complexity is related to the fact that thyroid tumors consist of neoplastic cells intermingled irregularly with normal (connective tissue and vessels) and reactive (stromal and immune) cells (8). Quantitative relations between these components may vary between patients and even inside one tumor. Most microarray studies include tumor fragments containing more than 80% to 90% of tumor cells and some authors recommend investigation of microdissected cells (9). This step is indispensable for sound understanding of neoplastic transformation but precludes the use of microarrays for diagnostic purposes. Only when the expression signal is strong enough to be detected in biopsy specimens, diffuse infiltrates, etc., is microarray-based technology applicable in future diagnostics.

In the present study, we examine expression profiles of nonpreselected papillary tumor fragments taken intraoperatively on the basis of macroscopic judgment. First, we raise the question of what is the major source of variance in PTC expression profiles as compared with unchanged thyroid tissue—individual gene expression patterns, tumor-normal difference, or other factors which need identification. Next, we define the list of genes important for tumor-normal difference, obtained after comparison of different gene selection methods. Finally, we propose an optimal set of genes to differentiate between PTC and normal thyroid tissue. We also show preliminary data validating the proposed classifier in an independent set of thyroid tissues.

Materials and Methods

Patients and Tissue Samples. Fifty thyroid tissue samples were taken intraoperatively from 33 patients during primary thyroidectomy. All samples were collected after obtaining informed consent and with the approval of the Local Ethics Committee. According to WHO criteria, PTC was diagnosed in 23 patients. Remaining 10 were operated for other thyroid diseases.

Within the 23 PTC patients there were 17 females and 6 males, aged 5 to 71 years. The patients were euthyroid during surgery (thyrotropin range, 0.88–3.06 milliunits/L). The classic variant of PTC was diagnosed

Requests for reprints: Barbara Jarząb, Department of Nuclear Medicine and Endocrine Oncology, Maria Skłodowska-Curie Memorial Cancer Center and Institute of Oncology, Gliwice Branch, Wybrzeże Armii Krajowej 15, 44-100 Gliwice, Poland. Phone: 48-32-2789301; Fax: 48-32-2789325; E-mail: bjarzab@io.gliwice.pl.

in 15 cases, the follicular variant in 6 cases, the diffuse sclerosing variant once, and a Warthin-like variant once. Lymph node metastases were diagnosed during primary surgery in 13 (57%) patients whereas functional distant metastases were found by post-therapy ^{131}I whole-body scan in 6 of them (26%). Patients were followed up for 5 to 34 months without relapse. Among the 10 non-PTC patients there were 8 females and 2 males, aged 11 to 69 years.

Whenever possible, both tumor sample and macroscopically unchanged fragment of thyroid tissue were collected from the same PTC patient. Control fragments were taken from the opposite lobe. An initial group consisted of 16 pairs of tumors and respective benign/normal thyroid tissues (2 nodular goiters, 10 colloidal goiters, 2 cases of thyroiditis, and 2 normal thyroid tissues), which are further referred as normal ones for the sake of clarity, although a normal/benign designation would be probably more appropriate.

Eighteen thyroid tissues, among them 7 additional PTCs and 11 normal/benign thyroid samples (3 follicular adenomas, 1 colloidal goiter, and 7 normal thyroid tissues) taken from 10 non-PTC patients were included in a separate validation group.

Among the 23 papillary cancers included in both groups, information on the PTC cell content was available in 16 cases (13 in the initial and 3 in the validation group). It ranged between 20% and 100% with a median of 60%, according to the semiquantitative evaluation done by the pathologist (E.C.), through an estimation of the area covered by tumor cells in the largest section of the tumor fragment, adjacent to the sample taken for microarray analysis.

Isolation of RNA. Samples (100-150 mg) were ground in liquid nitrogen and homogenized in RLT buffer (Qiagen, Hilden, Germany). RNA was extracted and repurified using RNeasy Midi and Mini Kits (Qiagen), including a digestion step with DNase I set (Qiagen). RNA quantity was measured by UV spectrophotometry, and quality was assessed by the 260/280 ratio and 1% agarose gel electrophoresis.

Microarray Analysis. All the microarray preparation procedures were done according to recommendations of Affymetrix (Santa Clara, CA) using 8 μg of total RNA as a template. Fragmented cRNA was hybridized first to a control microarray (Test3) and then, after sample quality evaluation, to Human Genome U133A array (Affymetrix).

Real-Time Quantitative Reverse Transcription-PCR Validation of Microarray Data. Real-time quantitative PCR (Q-PCR) was done using an ABI PRISM 7700 Sequence Detection System (Applied Biosystems, Foster City, CA). Primers and Taqman probes were supplied by Applied Biosystems through the Assay-on-Demand program. A standard curve, used in all experiments, was prepared from serial dilutions of total RNA from a single sample of toxic goiter. The β-glucuronidase (*GUS*) was used as a reference gene, chosen due to the most stable expression in thyroid samples as assessed by Taqman Endogenous Control Plate (Applied Biosystems). All results were normalized to the reference gene expression. Correlations between expression levels detected by microarray and Q-PCR analyses were measured by Spearman coefficient.

Singular Value Decomposition. Singular value decomposition (SVD) was used to detect and extract internal structure existing in the data and corresponding to important relationships between expressions of different genes. The algorithm of matrix decomposition (see Web Appendix[6]) was used to obtain orthogonal vectors called characteristic modes (supergenes; refs. 10, 11) which represent major independent (not correlated) variability patterns in the analyzed data. As a result, for every gene in the array a set of coefficients was obtained, defining the contribution of the ith mode to the expression pattern of the kth gene. Each coefficient was compared with the cutoff value, equal to $W \cdot n^{-1/2}$, where n was the number of genes and W was a weight factor. Its value was set to 3 and, if greater, the corresponding gene was included in the set of genes related to each characteristic mode. Each gene was related only to one mode with the highest value of the coefficient. Analysis was done by K. Simek (ksimek@ia.polsl.gliwice.pl).

[6] http://www.genomika.pl/thyroidcancer.

Statistical Comparison of Gene Expression Levels. We did a paired sample analysis of corresponding tumor and normal tissues using the comparison analysis algorithm from MAS 5.0 based on Wilcoxon's signed rank test between probe level expression values (P value was set to 0.003 and no correction for multiple comparisons was applied during this analysis). In the analysis of SVD data, we used paired sample t test with Benjamini-Hochberg correction, which controls the False Discovery Rate.

Recursive Feature Replacement. Recursive feature replacement (RFR; ref. 12), an iterative method based on the support vector machines technique (13–15), aims to find an optimal gene subset in a leave-one-out cross-validation approach. RFR in successive steps modifies actual n-element gene subset (one gene is removed and one gene is introduced). RFR is our own modification of standard Recursive Feature Elimination algorithm (14) and it uses Recursive Feature Elimination to find starting gene subsets. In our data set, it showed superior quality of classification compared with Recursive Feature Elimination (see Web Appendix).

Before the analysis, genes were preselected to reduce computational load. We used modified Sebestyen Criterion and Neighbourhood Analysis (16). The first method selected genes based on quality of separation between tumor and normal tissues, simultaneously maximizing the differences between both sets and minimizing the distances inside each set. Neighbourhood Analysis evaluated the correlation between expression profile and an "ideal" differentiating profile. Both methods ordered genes according to the coefficients obtained. The top 250 genes were selected from each list and a sum of both sets was constructed (NA-SC set, 282 genes). Recursive Feature Elimination method was applied to sort genes in this set and the RFR algorithm was done. As a result we obtained 100 gene sets, each with the corresponding linear classification function, which was dependent on the expression of every gene in a set and was assumed to be positive for tumor samples and negative for normal ones. Further steps are described in Results. Analysis was done by K.F. (kfujarewicz@ia.polsl.gliwice.pl).

Data Preprocessing and Software. All data were obtained using MAS 5.0 software (Affymetrix). Arrays were scaled to a target value of 100 (scaling factor range, 0.44-1.7). We excluded all Affymetrix controls as well as all genes absent ($P > 0.06$) in all 32 samples from the initial group. The obtained list contained 16,502 probe sets (76.6% of HG-U133A genes) and was used in all analyses. Despite the fact that there was sometimes more than one probe set per transcript, in the subsequent text the probe sets will be referred to data as "genes". The whole preprocessed data set is given in Web Appendix.[6]

For SVD and RFR, original procedures were developed in the Matlab environment (MathWorks, Natick, MA). For these analyses, absolute gene expression values were log-transformed (base 10), then all columns and rows were normalized (subtraction of mean and division by a SD). For paired t test, clustering and other analyses we used GeneSpring 6.1 (Silicon Genetics, Redwood City, CA). Hierarchical clustering was done by centroid clustering method referred to as the "average-linkage" method, with Pearson correlation around zero as the distance metric (called "Standard correlation" in GeneSpring). Correlation analysis of Q-PCR and microarray expression values was carried out using SPSS 12 (SPSS, Chicago, IL).

Gene Ontology Analysis. Biological relevance of obtained sets was analyzed by Gene Ontology classification. We used Affymetrix annotations for HG-U133A (February 2004). Lists obtained were hand curated and genes lacking annotation were classified according to other properties.

Results

We first analyzed 16 pairs of PTC and respective normal/benign thyroid samples taken from the same patient (initial group) by an unsupervised method (SVD). Once we confirmed that the tumor-normal difference was the main source of variation in expression profiles obtained, the supervised methods of gene selection were applied.

Sources of Variability in the Gene Expression Profiles. To assess gene expression patterns in the initial group, we did SVD analysis. This method computes "modes" ("supergenes") which

Table 1. Gene ontology classification of genes from SVD analysis

Class of genes	First mode		Second mode		Third mode		All	
Proliferation and malignant transformation	105	33.9%	63	33.0%	44	22.5%	212	30.4%
Invasion and metastasis	53	17.1%	16	8.4%	10	5.1%	79	11.3%
Structure, metabolism and transport	60	19.3%	22	11.5%	17	8.7%	99	14.2%
Immunoglobulin genes	0	0%	7	3.7%	63	32.1%	70	10.0%
Other immune-related transcripts	30	9.7%	70	36.6%	14	7.1%	114	16.4%
Hemoglobin and hemostasis	18	5.8%	0	0%	0	0%	18	2.6%
Other or unknown	44	14.2%	13	6.8%	48	24.5%	105	15.1%
All	310		191		196		697	
Genes differentiating tumor and normal tissues (paired *t* test)	254	81.9%	37	19.4%	4	2.0%	295	42.3%

NOTE: Gene classification was created on the basis of Gene Ontology and other annotations provided by Affymetrix. The bottom row shows the results of paired *t* test analysis done on the genes from all three modes (SVD set, 697 genes) to detect genes significantly differentiating between tumor and normal tissues. It confirmed that they were found mainly in the first mode.

correspond to the most important trends within the examined data set without any information on the tissue origin. First, three modes were considered significant sources of variability and explained 40.4% of the total variation in the data. The analysis selected 310, 191, and 196 genes corresponding to the first, second, and third mode, respectively. The sum of these three sets, defined as the SVD set, contained 697 genes.

To analyze the meaning of the three identified patterns (modes), we did hierarchical clustering using the obtained genes (Fig. 1). The first and the strongest mode grouped genes determining the difference between tumor and normal thyroid tissue. All tumors clustered together were distinctly separated from all normal samples and the distance between the two groups was quite large. Genes corresponding to the second mode were also related to the tumor-normal difference. They determined two distinct subgroups, each subdivided into tumor and normal samples. The third SVD mode was not related to the tumor-normal difference at all but was at least partially related to the individual differences between patients, as 6 of 16 pairs from the same patient clustered together.

We classified all genes from the SVD set into functional groups on the basis of Gene Ontology annotations, relevant publications, and other annotations (Table 1). Genes related to neoplastic transformation and invasion (signal transduction, cell cycle, apoptosis, cell adhesion, extracellular matrix, etc.) constituted half of all genes within the first mode. Transformation, proliferation, and death genes were similarly abundant in the first and second mode and less frequent in the third mode. Cell adhesion and extracellular matrix-related transcripts (invasion and metastasis genes in Table 1) were more frequent in the first mode (17%) than in the second (8%) and third (5%) modes. Similar but less distinct patterns were obtained for structure, metabolism, and transport genes. The most striking differences were observed for blood (hemoglobin and clotting factors) and immunity-related genes, mainly immunoglobulin transcripts. With no exception, all 18 blood genes correlated to the first mode. Conversely, 90% of immunoglobulin transcripts were found within the third mode that constituted 32% of all genes within this mode. Remaining immunity-related genes occurred mainly in the second mode (36% of all genes in this mode).

To verify which genes selected by SVD were responsible for the tumor-normal difference, we applied a simple supervised method (paired *t* test with Benjamini-Hochberg correction). Two-hundred ninety-five genes (42.3%) from all 697 SVD set genes differed between normal and tumor tissue. As the False Discovery Rate was set to 0.01, on average three genes from this list are expected to be false positives. Differentiating genes constituted 81.9% of genes in the first mode and 19.4% in the second mode whereas the third mode contained only four such genes. To analyze the functional

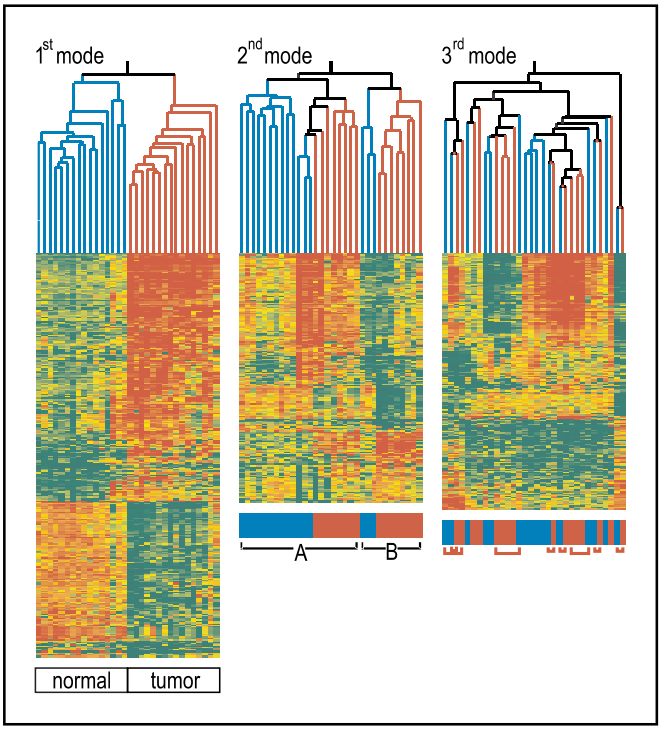
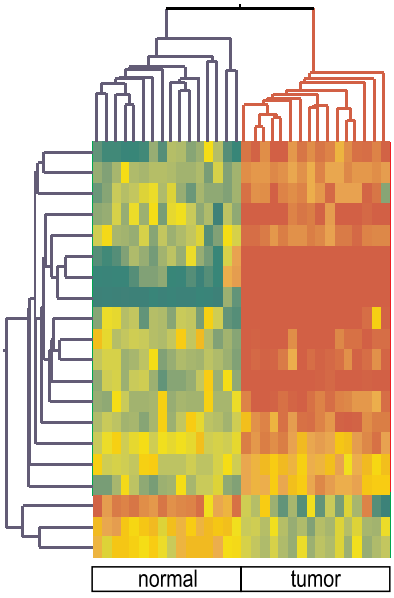


Figure 1. Hierarchical clustering of samples according to genes related to the first, second, and third mode. The first mode genes give a clear distinction of normal (*blue*) and tumor tissues (*red*). Clustering by the second mode genes gives two subgroups of samples (*A* and *B*), each containing normal and tumor tissues. Genes related to the third mode exhibit interindividual variability to some extent—a few pairs from the same patient cluster closely together (*red brackets*). Lists of the genes related to each mode as well as the genes differentiating sets *A* and *B* are given in the Web Appendix.



Gene	Affymetrix probe id	Normal thyroid tissues Median	Normal thyroid tissues Range	Papillary thyroid cancer Median	Papillary thyroid cancer Range
GJB3	205490_x_at	9.5	(4.9–34.8)	52.2	(6.7–233.2)
P4HA2	202733_at	109.3	(73.4–173.3)	295.3	(117.2–550.4)
MET	211599_x_at	87.3	(43.2–144.1)	208.9	(69.2–572.6)
RXRG	205954_at	17.5	(1.3–25.4)	88.6	(9.2–398.3)
LGALS3	208949_s_at	523.8	(234.7–708.4)	1564.3	(164.5–2580.9)
SERPINA1	202833_s_at	73.4	(31.9–279.5)	1311.2	(4.8–3044.6)
CDH3	203256_at	4.5	(2.7–36.8)	246.3	(3.3–668.5)
DPP4	203716_s_at	1.3	(0.7–5.7)	139.9	(1.0–312.5)
IGSF1	207695_s_at	54.4	(24.6–94.7)	553.5	(10.3–1305.6)
MET	203510_at	131.8	(64.6–221.3)	598.7	(66.4–1244.9)
EVA1	203780_at	49.3	(29.0–79.3)	266.9	(16.9–461.3)
LRP4	212850_s_at	22.1	(12.4–47.2)	220.9	(5.2–995.1)
SPUVE	202458_at	137.2	(106.8–255.5)	556.5	(43.3–1267.1)
MRC2	37408_at	85.0	(61.8–136.2)	269.3	(18.5–697.4)
KIAA0830	212573_at	208.7	(164.6–260.4)	325.3	(101.7–590.2)
IL13RA1	201887_at	212.7	(151.9–291.9)	297.9	(164.9–427.2)
ST14	202005_at	98.7	(50.7–138.8)	186.5	(91.4–410.3)
HBB	211696_x_at	4293.2	(1859.7–6878.0)	867.0	(220.8–4428.7)
MTMR4	212277_at	133.9	(86.0–199.9)	80.5	(39.2–153.7)
MKRN2	218071_s_at	119.9	(89.3–150.7)	90.5	(55.8–149.1)

Figure 2. Hierarchical clustering of samples from the initial group, based on the RFR-20 gene set. The upper 17 transcripts are up-regulated and the lower three are down-regulated in tumors. There are two probes for *MET* oncogene selected in this data set (both of them are important for proper classification of all samples), thus it encompasses 19 genes. In the table median and range of values are given; full annotations for these genes are provided in the Web Appendix.

significance of the second SVD mode, we first compared by t test the two main tissue subgroups within this mode (subgroup A and subgroup B, Fig. 1). In total, 129 genes were responsible for the major subdivision within the second supergene (False Discovery Rate, 0.05). Among them 62 genes (48%) were immunity-related ones (chemokines, T cell receptor transcripts, etc.); other genes were dispersed between various functional groups (signal transduction, metabolism, apoptosis, etc.). The majority of these genes exhibited lower expression in subgroup B than in subgroup A. The expression of immunoglobulin genes (belonging to third mode) was also diminished in subgroup B.

In summary, two major sources of gene expression profile variability were observed within the initial group of 32 thyroid samples: a very large difference between PTC tissues and normal thyroid tissues revealed in the first SVD mode by unsupervised analysis and confirmed by further supervised selection, and a confounding variability, due to the immune response-related transcripts. This variability was influenced at least partially by individual differences between patients.

Genes Characteristic for the Tumor-Normal Difference. We did paired sample analysis on probe-level expression values, using algorithms implemented in MAS 5.0 (Comparison Analysis). In 50% of tumor-normal pairs analyzed, 2,639 genes were changed; the most stringent criterion (i.e., change in all 16 pairs) was met by 110 transcripts. A list containing 957 genes significantly changed in one direction in at least 12 tumor-normal tissue pairs was used in further analysis.

Selection of the Best Set of Genes. For further reduction of the number of relevant genes, we applied Recursive Feature Replacement algorithm. Instead of creating a list of single genes, RFR allows obtaining a gene set, which constitutes the best possible combination of differentiating genes. This algorithm analyzed the discrimination power of consecutive sets with an increasing number of genes (ranging from 1 to 100 genes). The classification quality index increased quickly with sets composed of 3 to 10 genes, approached a plateau at a 20-gene set size, and then slowly declined (see Web Appendix). The redundancy was minor and all 100 sets included together 116 different gene probe sets detecting 102 various genes (RFR set, Table 2). The 20-element gene set (RFR-20 set, Fig. 2), chosen for further evaluation, contained 19 genes: 16 up-regulated and 3 down-regulated genes (*MET* gene was represented by two probes; one may notice that both the probes are important for correct classification).

Validation of the RFR-20 Gene Set. We have preliminarily validated the diagnostic power of the RFR-20 set by classifying 18 other thyroid tissue samples (validation group). Using the RFR-20 linear support vector machines-based classifier, unknown samples were assigned into tumor-normal ones. We found that the RFR-20 set classified properly 17 of 18 samples (94.4%; 95% CI: 72.7-99.9%). All seven normal thyroid samples and four benign lesions included were properly classified as normal/benign ones (Fig. 3*A*). From seven PTC samples, six were correctly classified, whereas one was misclassified and considered normal/benign. Thus, the sensitivity of PTC detection was 85.7% and the specificity was 100%. The misclassified cancer sample was an outlier as it contained 20% of tumor cells whereas only two other samples contained less than 40% of tumor cells (20% and 30%).

Validation of the Microarray-Derived Expression Data by Quantitative PCR. To validate the microarray expression data by an independent method, we did for 10 genes the real-time PCR expression analysis on 37 samples. For eight genes, the correlation coefficient was higher than 0.75, indicating a high correspondence of results (Table 3).

Comparative Analysis with Other Papillary Thyroid Cancer Microarray Data Sets. Comparison with the high density microarray data was possible for eight PTC and eight

Table 2. All genes obtained by RFR for gene sets of 1 to 100 genes

Gene symbol	Gene name	Gene ID*	Median SLR (range)	Increased/ decreased/ not changed	References	Validated in Huang data set
DPP4	**Dipeptidylpeptidase 4 (CD26, adenosine deaminase complexing protein 2)**	**1803**	**6.855 (4.26–7.95)**	**16/0/0**	**(2, 18, 42–44)**	**Yes**
GJB3	Gap junction protein, β-3, 31 kDa (connexin 31)	2707	2.99 (0.81–4.39)	10/0/6		
ST14	Suppression of tumorigenicity 14 (colon carcinoma, matriptase, epithin)	6768	0.86 (0.35–1.35)	11/0/5		
SERPINA1	**Serine (or cysteine) proteinase inhibitor, clade A (α-1 antiproteinase, antitrypsin), member 1**	**5265**	**4.395 (2.83–5.45)**	**16/0/0**	**(2, 20, 33)**	**Yes**
LRP4	Low density lipoprotein receptor-related protein 4	4038	4.28 (2.12–5.56)	16/0/0		Yes
MET	**Met proto-oncogene (hepatocyte growth factor receptor)**	**4233**	**1.47 (0.1–2.95)**	**12/0/4**	**(2, 6, 17, 22, 48)**	**Yes**
EVA1	Epithelial V-like antigen 1	10205	2.285 (1.54–2.94)	16/0/0		Yes
MYH10	Myosin, heavy polypeptide 10, nonmuscle	4628	1.445 (0.65–2.63)	16/0/0		
SPUVE	**Protease, serine, 23**	**11098**	**2.24 (1.01–3.35)**	**16/0/0**	**(2)**	**Yes**
LGALS3	**Lectin, galactoside-binding, soluble, 3 (galectin 3)**	**3958**	1.925 (1.02–2.98)	**16/0/0**	**(2, 21, 42)**	**Yes**
CD44	CD44 antigen (homing function and Indian blood group system)	960	1.165 (0.38–1.92)	15/0/1	(21, 26, 27, 45)	
HBB	Hemoglobin, β	3043	−1.915 (−3.41–0.1)	0/15/1	(46)	
MAP17	Membrane-associated protein 17	10158	3.715 (0.25–4.38)	15/0/1		
MKRN2	Makorin, ring finger protein, 2	23609	−0.525 (−0.93–−0.08)	0/10/5		
TGFA	Transforming growth factor, α	7039	2.895 (1.19–3.99)	16/0/0	(6)	
MRC2	Mannose receptor, C type 2	9902	1.525 (0.49–2.51)	14/0/1		
IGSF1	Immunoglobulin superfamily, member 1	3547	4.25 (1.04–5.07)	15/0/1		Yes
KIAA0830	KIAA0830 protein	23052	1.01 (0.16–1.64)	15/0/1		
RXRG	Retinoid X receptor γ	6258	3.53 (0.95–6.78)	16/0/0	(35)	
FN1	**Fibronectin 1**	**2335**	**3.2 (2.07–4.44)**	**16/0/0**	**(2, 17)**	**Yes**
MTMR4	Myotubularin-related protein 4	9110	−0.875 (−1.33–−0.29)	0/14/2		Yes
IL13RA1	Interleukin 13 receptor, α1	3597	0.565 (−0.01–1.09)	8/0/7		
P4HA2	**Procollagen-proline, 2-oxoglutarate 4-dioxygenase (proline 4-hydroxylase), α polypeptide II**	**8974**	1.55 (0.7–2.27)	16/0/0	**(2)**	Yes
CDH3	Cadherin 3, type 1, P-cadherin (placental)	1001	5.61 (2.99–8.16)	16/0/0	(2, 31, 47)	
TLE4	Transducin-like enhancer of split 4 [E(sp1) homologue, *Drosophila*]	7091	−1.26 (−2.43–−0.19)	0/12/3		
PAXIP1L	PAX transcription activation domain interacting protein 1-like	22976	−0.465 (−0.77–0.11)	0/6/8		
HSBP1	Heat shock factor binding protein 1	3281	−0.52 (−0.95–−0.05)	0/14/2		
NRCAM	Neuronal cell adhesion molecule	4897	1.95 (1.4–3.49)	16/0/0		Yes
TSTA3	Tissue specific transplantation antigen P35B	7264	0.68 (0.31–1.32)	9/0/7		
RNF24	Ring finger protein 24	11237	0.63 (0.11–1)	2/0/13		

(Continued on the following page)

Table 2. All genes obtained by RFR for gene sets of 1 to 100 genes (Cont'd)

Gene symbol	Gene name	Gene ID*	Median SLR (range)	Increased/ decreased/ not changed	References	Validated in Huang data set
KIAA0937	**KIAA0937 protein**	**23220**	**2.49 (1.1–3.69)**	**16/0/0**	**(2)**	**Yes**
TMPRSS4	Transmembrane protease, serine 4	56649	4.305 (2.3–6.12)	14/0/2		
WWOX	WW domain containing oxidoreductase	51741	−1.51 (−2.07–0.05)	0/15/1		
LOC157567	Hypothetical protein LOC157567	157567	−1.005 (−1.54–−0.52)	0/15/1		
CITED1	**Cbp/p300-interacting transactivator, with Glu/Asp-rich carboxyl-terminal domain, 1**	**4435**	**4.42 (1.86–7.04)**	**14/0/2**	**(2, 3)**	**Yes**
Hs.297681.2	*Homo sapiens* PRO2275 mRNA, complete cds (SERPINA1 homologue)	Hs.513816	3.92 (2.4–4.67)	16/0/0		Yes
HBA2	**Hemoglobin, α2**	**3040**	**−2.24 (−3.72–0.21)**	**0/14/2**	**(2)**	
SLIT1	Slit homologue 1 (*Drosophila*)	6585	2.38 (−0.48–4.35)	12/0/4		Yes
PAM	Peptidylglycine α-amidating monooxygenase	5066	0.755 (0.27–1.24)	15/0/1		
FLJ12541	Stimulated by retinoic acid gene 6	64220	1.21 (0.04–2.4)	9/0/7		
SLC34A2	Solute carrier family 34 (sodium phosphate), member 2	10568	4.215 (1.17–5.63)	16/0/0		
AGR2	Anterior gradient 2 homologue (*Xenopus laevis*)	10551	3.8 (0.89–6.86)	15/0/0		
MINA53	Myc-induced nuclear antigen, 53 kDa	84864	−1.16 (−1.78–−0.16)	0/14/2		
ADPRTL1	ADP-ribosyltransferase (NAD+; poly(ADP-ribose) polymerase)-like 1	143	1.05 (0.23–1.92)	16/0/0		
FLJ10748	Hypothetical protein FLJ10748	55220	2.225 (0.47–4.68)	15/0/1		
BMP1	Bone morphogenetic protein 1	649	0.965 (0.46–1.85)	6/0/9		
LAMB3	Laminin, β3	3914	3.695 (2.25–5.92)	16/0/0		
FLJ10178	Hypothetical protein FLJ10178	55086	−3.03 (−5.33–0.07)	0/13/3		
EPS8	**Epidermal growth factor receptor pathway substrate 8**	**2059**	**1.6 (0.23–2.34)**	**15/0/1**	**(2)**	**Yes**
NRP2	Neuropilin 2	8828	1.03 (−0.17–2.78)	12/0/4		
COL13A1	Collagen, type XIII, α 1	1305	2.45 (0.34–3.19)	15/0/1		
SOSTDC1	Sclerostin domain containing 1	25928	−3.365 (−4.7–−0.18)	0/11/3		
FLRT3	Fibronectin leucine-rich transmembrane protein 3	23767	1.51 (0.3–3.74)	15/0/1		
SCEL	**Sciellin**	**8796**	**3.91 (1.04–5.88)**	**15/0/1**	**(2)**	**Yes**
FLII	Flightless I homologue (*Drosophila*)	2314	0.54 (0.35–1.26)	13/0/3		
IKBKE	Inhibitor of light polypeptide gene enhancer in B-cells, kinase epsilon	9641	0.315 (0.13–1.1)	3/0/11		
N33	**Putative prostate cancer tumor suppressor**	**7991**	**2.885 (1.12–4.09)**	**16/0/0**	**(2)**	**Yes**
EGFL5	Epidermal growth factor-like-domain, multiple 5	1955	1.41 (0.46–1.99)	16/0/0		Yes
FRCP1	Likely orthologue of mouse fibronectin type III repeat containing protein 1	64838	3.075 (0.11–4.61)	11/0/5		
FDFT1	Farnesyl-diphosphate farnesyltransferase 1	2222	−0.525 (−0.98–−0.09)	0/12/4		
SFN	Stratifin	2810	1.655 (0.29–3.26)	12/0/4	(24)	

(Continued on the following page)

Table 2. All genes obtained by RFR for gene sets of 1 to 100 genes (Cont'd)

Gene symbol	Gene name	Gene ID*	Median SLR (range)	Increased/ decreased/ not changed	References	Validated in Huang data set
SLPI	Secretory leukocyte protease inhibitor (antileukoproteinase)	6590	2.705 (−0.15–4.4)	14/0/2		
TREM1	Triggering receptor expressed on myeloid cells 1	54210	2.21 (0.73–3.75)	12/0/4		
DAT1	Neuronal specific transcription factor DAT1	55885	1.205 (0.32–2.01)	15/0/1		
S100A14	S100 calcium-binding protein A14 (calgizzarin)	Hs.247697.0	1.215 (0.58–1.71)	16/0/0		
FLJ21511	Hypothetical protein FLJ21511	80157	−2.215 (−5.87–−0.8)	0/16/0		
CCND2	Cyclin D2	894	1.31 (0.07–2.84)	13/0/3		
CCND1	Cyclin D1 (PRAD1: parathyroid adenomatosis 1)	595	1.315 (0.56–2.84)	16/0/0	(25)	Yes
SLC27A6	Solute carrier family 27 (fatty acid transporter), member 6	28965	3.455 (−0.55–8.7)	14/0/2		
EBAG9	Estrogen receptor binding site associated antigen 9	9166	−0.65 (−1.19–−0.13)	0/13/3	(30)	
KRT19	**Keratin 19**	**3880**	**3.225 (1.17–6.61)**	**16/0/0**	**(2)**	
ETV5	Ets variant gene 5 (ets-related molecule)	2119	1.575 (0.7–2.51)	16/0/0		Yes
COMP	Cartilage oligomeric matrix protein	1311	3.835 (1.15–8.41)	15/0/1		
CLDN10	Claudin 10	9071	2.32 (−0.34–4.32)	13/0/3		
CCL15	Chemokine (C-C motif) ligand 15	6359	−1.305 (−2.5–0.52)	0/14/2		
LCN2	Lipocalin 2 (oncogene 24p3)	3934	2.8 (−0.21–5.56)	13/0/3		
TIMP1	Tissue inhibitor of metalloproteinase 1 (erythroid potentiating activity, collagenase inhibitor)	7076	2.325 (0.82–3.72)	16/0/0	(2, 17, 32)	
SLC25A15	Solute carrier family 25 (mitochondrial carrier; ornithine transporter) member 15	10166	−1.95 (−3.64–−0.44)	0/15/0		
DUSP6	**Dual specificity phosphatase 6**	**1848**	**1.88 (0.65–2.91)**	**16/0/0**	**(2)**	**Yes**
BC008967	Hypothetical gene BC008967	89927	1.035 (0.36–2.67)	13/0/3		
COL8A2	Collagen, type VIII, α2	1296	1.99 (−0.81–2.82)	15/1/0		
GATM	Glycine amidinotransferase (L-arginine:glycine amidinotransferase)	2628	−1.735 (−2.59–−0.41)	0/15/1		
QPCT	Glutaminyl-peptide cyclotransferase (glutaminyl cyclase)	25797	3.545 (0.63–4.59)	15/0/1		
NDP52	Nuclear domain 10 protein	10241	−0.61 (−1.08–−0.01)	0/13/2		
PHF10	PHD finger protein 10	55274	−0.665 (−1.09–0.22)	0/11/5		
S100A6	S100 calcium binding protein A6 (calcyclin)	6277	0.945 (0.37–1.7)	16/0/0	(29)	
CTSC	Cathepsin C	1075	2.24 (0.87–3.5)	16/0/0		Yes
DKFZp761K1423	Hypothetical protein DKFZp761K1423	55358	1.915 (1.49–2.92)	16/0/0		
NPC2	Niemann-Pick disease, type C2	10577	1.4 (0.76–2.23)	16/0/0		Yes
FLJ39207	C219-reactive peptide	348477	−0.75 (−1.47–0.26)	0/12/3		
MLF1	Myeloid leukemia factor 1	4291	−1.135 (−2.04–−0.31)	0/13/2		
KIAA1006	Tomosyn-like	9515	−0.585 (−1.3–−0.06)	0/4/12		
CaMKIINα	Calcium/calmodulin-dependent protein kinase II	55450	2.41 (0.97–4.05)	16/0/0		
CHI3L1	**Chitinase 3-like 1 (cartilage glycoprotein-39)**	**1116**	**4.45 (−0.08–6.76)**	**14/0/2**	**(2)**	**Yes**
HTCD37	TcD37 homologue	58497	−0.88 (−1.71–−0.16)	0/11/4	(7)	
GPR51	G protein-coupled receptor 51	9568	2.445 (0.07–3.55)	15/0/1		
CTH	Cystathionase (cystathionine γ-lyase)	1491	−1.365 (−2.33–−0.53)	0/15/1		
S100A11	S100 calcium binding protein A11 (calgizzarin)	6282	1.335 (0.72–2.42)	15/0/0		Yes
PLAU	Plasminogen activator, urokinase	5328	2.56 (0.93–4.17)	16/0/0		

(Continued on the following page)

Table 2. All genes obtained by RFR for gene sets of 1 to 100 genes (Cont'd)

Gene symbol	Gene name	Gene ID*	Median SLR (range)	Increased/ decreased/ not changed	References	Validated in Huang data set
MVP	Major vault protein	9961	1.37 (0.63–2.49)	16/0/0		
TCFL5	Transcription factor-like 5 (basic helix-loop-helix)	10732	−1.01 (−1.46–−0.31)	0/14/2		
FLJ13946	Hypothetical protein FLJ13946	92104	−1.295 (−2.67–−0.61)	0/12/3		

NOTE: Genes are listed according to size of the smallest set in which they appear. Median Signal Log Ratio (SLR) and SLR range, as well as results of paired sample analysis by MAS 5 algorithm, are given. Last two columns provide the references to other studies in thyroid cancer [genes given by Huang et al. (2) are in boldface] and results of RFR analysis done by us on Huang's data.
*LocusLink (UniGene where not available).

normal tissues analyzed by Huang et al. (2) and available from http://thinker.med.ohio-state.edu. This analysis was done on older (HG-U95) chips. SVD analysis and RFR were done on this data set in analogy to the evaluation done in this study (see Web Appendix). Interestingly, our analysis specified 13 new genes within their data set, which were not found by the authors themselves on the basis of univariate analysis, but were also present in our RFR set (Table 2). Simultaneously, there was a wide overlap between genes selected by RFR in both data sets.

We also tested our RFR-20 gene classifier for its discrimination ability on Huang's tissues (Fig. 3*B*). From the 20 probes selected from the HG-U133A microarray, 19 were present on their chip (the one missing probe was replaced by median value). The discrimination between normal thyroid and PTC was correct in all cases.

Discussion

Huang et al. (2) reported a very distinct difference in gene expression profile between papillary thyroid carcinoma and "normal" thyroid tissue. Our results confirm that the PTC expression signal is very prominent and easily detectable even when cancer cells do not prevail over tumor stroma. Unlike in the previous studies which did not check the quality of classification or used less advanced methods, our main goal was not to list genes with the largest fold-change (2, 17) but rather to specify a most powerful set of genes. By the bioinformatical methods which have not been used in thyroid cancer before, we eliminated gene redundancy and applied a full leave-one-out cross-validation procedure. Simultaneously, the analysis used proved that the difference between tumor and normal samples was the major source of variability in the gene expression pattern of thyroid tissues.

Multigene Molecular Classifier of Papillary Thyroid Cancer. The difference in expression profile between PTC and uninvolved thyroid tissue encompassed 700 to 3,000 genes depending on the method and the cutoff level used. These numbers illustrate the complexity of the gene networks involved in PTC transformation as well as stromal cell response. Despite the very distinct changes in expression, none of the several genes has proved to be an ideal single marker of PTC in an independent set of PTCs analyzed by Q-PCR (data not shown).[7] Even *DPP4* (18), indicated in the previous study of Huang et al. as the most up-regulated gene in PTC and clearly confirmed in this study, did not fulfill these expectations, although the distance between normal and tumor values was particularly large. To overcome this lack of accuracy, we looked for the gene sets where genes complement each other rather than for collections of transcripts selected by univariate approaches. The RFR algorithm applied in our study optimized the gene sets selected by the standard Recursive Feature Elimination algorithm and this effect was particularly visible in smaller sets (for comparison, see Web Appendix). Finally, we present a gene set (RFR-20) which is an optimal molecular classifier for discriminating between PTC and normal/benign thyroid tissue. The validation step was done by classification of an independent group of 18 thyroid tissues which gave correct results for all normal tissues and benign tumors and failed only in one case of seven papillary cancers analyzed. The most probable reason for this misclassification was the low content of tumor cells in this sample. The additional confirmation was obtained when we applied our molecular classifier on gene expression data published by the earlier PTC study (2).

Genes Selected by Recursive Feature Replacement Analysis. The RFR-20 set, considered in our study as "the best set of genes," includes some genes with a very distinct change in expression signal, previously known for their up-regulation in PTC. Among them are dipeptidylpeptidase 4 (*DPP4*; refs. 19–22), α-1 antitrypsin (*SERPINA1*; ref. 23), galectin 3 (*LGALS3*; ref. 24), and *MET* oncogene (25, 26). Six of these (including the four mentioned above) were also selected by previous studies on PTC expression profile (2, 17, 27). Other known genes which were also very heavily overexpressed in our PTC samples were not included in the RFR-20 set: fibronectin 1 (*FN1*), tissue inhibitor of metalloproteinase 1 (*TIMP1*), and keratin 19 (*KRT19*) were the most prominent examples (2, 17). However, all of them were included in the larger set of 102 genes selected by the RFR algorithm. Genes involved in signal transduction like Cbp/p300-interacting transactivator (*CITED1*) or calcyclin (*S100A6*), the cell cycle regulators stratifin (*SFN*) and cyclin D1 (*CCND1*), the estrogen-responsive gene *EBAG9*, or CD44 antigen constitute other examples of genes in the large RFR set which were already indicated in previous PTC studies (see Table 2; refs. 24, 28–35). On the other side, there was no overlap between our genes and genes proposed recently by Mazzanti et al. (7) for differential diagnosis of PTC. Nearly all genes specified by them were confirmed in our study by *t* test analysis, but only one (*HTCD37*) was included in the RFR set.

[7] Manuscript in preparation.

Concerning new genes, not described until now in PTC, the most distinct changes in expression were encountered for the retinoid X receptor gene (*RXRG*), the epithelial V-like antigen (*EVA1*), and the low density lipoprotein receptor-related protein gene (*LRP4*). Some genes included in the RFR-20 exhibited less distinct changes in expression (*GJB3*, *IL13RA1*, *ST14*, *KIAA8300*, *MKRN2*, or *MTMR4*). They complement information covered by the stronger genes and thus increase diagnostic power of the proposed set.

The data for some new genes were validated by Q-PCR; they exhibited a rather good correlation with microarray-derived values (Table 3). In the whole RFR gene set (Table 2), 17 other transcripts (corresponding to 16 genes) were indicated previously by Huang et al. (2) and, thus, we did not perform additional validation for them. Further confirmation was obtained when we did RFR analysis on Huang's data set. By this approach we found in their data 13 genes indicated by our study and not mentioned by the authors themselves (Table 2), among them *EVA1* and *LRP4*.

Biological Relevance of the Selected Genes. The importance of cell adhesion genes has been already indicated by the previous microarray study (2). This group of genes, possibly related to invasion and metastasis processes, constituted the most numerous gene ontology class both in the RFR set (22%) and in first SVD mode (17%). Both *EVA1* and *LRP4*, which are among novel genes specified by us, are involved in cell adhesion/extracellular matrix regulation. The expression of *CDH3*, responsible for calcium-dependent cell-cell adhesion, which showed the second most distinct difference in tumor and normal values, after *DPP4* signal, has been indicated previously by immunochemistry (36, 37).

Other molecules related to cell adhesion are proteins with metalloendopeptidase inhibitor activity. *TIMP1* overexpression in PTC has been indicated in many studies (38) and also in recent microarray analyses (2, 17, 27). *SERPINA1* was reported previously to be overexpressed in PTC (23, 39). *TMPRSS4*, a novel serine protease which may be important for metastasis and tumor invasion, was included in the RFR-20 set. Among other invasion-related molecules we should also indicate urokinase plasminogen activator receptor-associated protein (*MRC2*), involved in collagen matrix degradation and remodeling, as well as urokinase plasminogen activator (*PLAU*) itself (ref. 40; Table 2).

In the whole RFR set, signal transduction genes were moderately abundant (10 of 102 genes), similarly to apoptosis/cell cycle genes (8 of 102). Concerning signal transduction-related genes, the very distinct up-regulation of *MET* is a consistent feature found in nearly all PTC genomic studies (2, 17, 27). There were also five transcription factors in the RFR set, among which the *RXRG* (Fig. 2) deserves special attention. It was considered as a novel one by us; thus, we did Q-PCR validation of its overexpression. A recently published study by Haugen et al. (41) indicated its up-regulation in PTC and related it to the response to retinoids in thyroid cancer.

Other new PTC genes confirmed by Q-PCR include *QPCT*, a very poorly known glutaminyl cyclase, and *SLC34A2* gene (NaPi3B), a sodium-dependent phosphate transporter expressed in several human tissues of epithelial origin.

Despite the variability in the genes mentioned above, the expression of genes related to DNA replication, cell cycle, mRNA splicing, and protein biosynthesis showed much less variation than could be expected. These genes constituted only a minority of RFR genes and their distribution between main SVD modes was nearly equal.

We should bear in mind that the functions of selected genes should be considered not only in the aspect of their role in PTC, as the observed changes in gene expression may also be related to the tumor stromal component. For some genes (e.g., fibronectin 1 or metalloproteinases and their inhibitors, like *TIMP1*) an increased expression may be observed in both tumor and stromal cells (42, 43) and for diagnostic purposes the evaluation of their expression in the whole tumor may be more informative. Blood genes constitute an example of gene group expressed outside of thyrocytes expression of which was important for differentiating between tumor and normal tissue both in our study as well as in other PTC microarray papers (2, 17, 44). Thus, even genes which are regarded as marginal for the investigation of transformation mechanisms (45) may be of diagnostic value as a part of a multigene molecular classifier. Not only hemoglobin β (*HBB*) and α chains (*HBA1* and *HBA2*) but also 16 other genes characteristic for the blood expression profile, among them many clotting factors, constituted strong differentiating signal and were found in the first SVD mode. As all blood genes exhibited concordantly decreased expression in tumors, blood supply in PTC is clearly diminished in comparison to surrounding thyroid parenchyma.

Table 3. Validation of microarray expression data by quantitative real-time PCR

Gene title	Affymetrix ID	Q-PCR assay ID	Microarray fold change	Q-PCR fold change	Spearman correlation coefficient
CDH3	203256_at	Hs00354998_m1	52.0	13.7	0.873
SLC34A2	204124_at	Hs00197519_m1	18.4	48.9	0.907
TMPRSS4	218960_at	Hs00212669_m1	17.1	100.9	0.770
LRP4	212850_s_at	Hs00323496_m1	16.0	28.5	0.792
RXRG	205954_at	Hs00199455_m1	11.3	38.5	0.925
QPCT	205174_s_at	Hs002002680_m1	9.2	15.1	0.931
EVA1	203780_at	Hs00170684_m1	8.0	5.5	0.856
MRC2	37408_at	Hs00195862_m1	2.6	6.1	0.867
MTMR4	212277_at	Hs00608316_m1	1.9	0.6	0.570
ST14	202005_at	Hs00222707_m1	1.7	2.1	0.527

NOTE: All correlations are statistically significant with $P < 0.0001$.

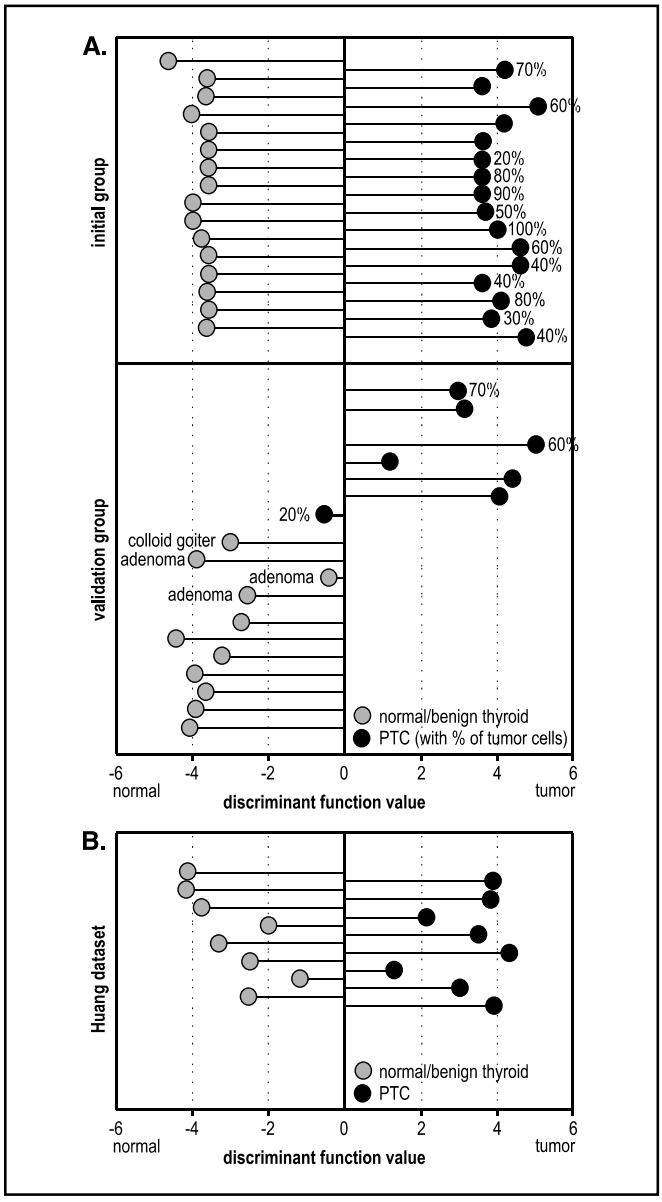


Figure 3. Classification of samples using the RFR-20 gene set. The positive values of discriminant function correspond to papillary cancer diagnosis, the negative values to normal/benign tissue. *Black dots,* PTCs are marked; *grey dots,* normal/benign tissues. For PTC tissues, the percent of tumor cells is given where available. *A, top,* the initial group is shown to illustrate the classification principle; *bottom,* the validation group of tissues are classified. Only one (from 18) tissue is misclassified, a papillary cancer with low (20%) tumor cell content. *B,* results of the classification with our RFR-20 classifier done on the data published by Huang et al. (2). All eight tumors and eight normal tissues were correctly classified.

Variation in the Gene Expression Profiles and Its Main Sources. The use of gross tumor fragments precluded the necessity of a wide analysis of the gene expression variability sources. SVD used in our study does not take tissue class into consideration and looks for the most prominent differences in the obtained expression patterns in an unsupervised manner. It is similar to principal component analysis used by others (7). In our study SVD confirmed that the main source of variability was related to the difference between PTC and normal/benign thyroid tissue. Additionally, SVD revealed that immunity-related genes provided the most intensive confounding signal, possibly related to the tumor infiltration by lymphocytes (46). Lymphocytic infiltration of PTC and its surroundings is a well-known phenomenon, often related to a favorable prognosis (47). In this context, we should point out that in our study "normal thyroid" tissue specimens were taken from thyroid region as distant as possible from the macroscopic tumor and usually located in the opposite lobe.

In the third mode, we observed a large number of immunoglobulin genes of which expression was highly differentiated among individuals. It is to be stressed that interindividual differences were rather weak in this study and were only partially visible in the third SVD mode. They influenced the PTC profiles obtained by Huang et al. (2) to a much stronger degree, which was seen in their own analysis and in our SVD analysis on their data set (see Web Appendix[6]). Immunoglobulin genes were important source of variability also in Huangs' data set. The tumor-normal difference, the main source of variability in our data set, was within the second mode in Huang's data. It cannot be excluded that the variance in gene expression among individuals was due to artifacts of sample preparation or biological and clinical factors. We assured that all patients in our study were euthyroid before surgery, were operated during the same time of the day, and received the same type of anesthesia.

Although our data confirm a highly consistent expression profile of papillary thyroid carcinoma, we refrain from definitely accepting this statement until a larger group of PTC was studied by microarray analysis in relation to the disease outcome. From the clinical point of view, 10% to 15% of patients with this carcinoma exhibit poor prognosis, related to still insufficiently identified features of tumor biology which may be uncovered by further expression profiling (48). Thus, we support the more conservative conclusion that the gene expression profile of PTC is stable enough to be used for diagnostic purposes and is easily detectable even when cancer cells do not prevail over tumor stroma. Simultaneously, we indicate the confounding variability related to the immune response in thyroid gland, which needs further investigation.

Acknowledgments

Received 8/25/2004; revised 11/15/2004; accepted 11/29/2004.

Grant support: Polish Committee of Scientific Research grant PBZ/KBN/040/P04/2001 and EU Program MRTN-CT-2004-503661. M. Wiench is a Foundation of Polish Science Fellowship recipient.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked advertisement in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

We thank Jarosław Szary, Ph.D., for preparation of poly(A) spike controls; Ron Hancock, Ph.D., and Aleksander Sochanik, Ph.D., for the thorough language revision of the manuscript; and the members of our Institute faculty for the valuable discussions.

References

1. Eszlinger M, Krohn K, Paschke R. Complementary DNA expression array analysis suggests a lower expression of signal transduction proteins and receptors in cold and hot thyroid nodules. J Clin Endocrinol Metab 2001;86:4834–42.
2. Huang Y, Prasad M, Lemon WJ, et al. Gene expression in papillary thyroid carcinoma reveals highly consistent profiles. Proc Natl Acad Sci U S A 2001;98:15044–9.
3. Prasad ML, Pellegata NS, Kloos RT, Barbacioru C, Huang Y, de la Chapelle A. CITED1 protein expression suggests papillary thyroid carcinoma in high throughput tissue microarray-based study. Thyroid 2004;14:169–75.

4. Takano T, Hasegawa Y, Matsuzuka F, et al. Gene expression profiles in thyroid carcinomas. Br J Cancer 2000;83:1495–502.
5. Baris O, Savagner F, Nasser V, et al. Transcriptional profiling reveals coordinated up-regulation of oxidative metabolism genes in thyroid oncocytic tumors. J Clin Endocrinol Metab 2004;89:994–1005.
6. Barden CB, Shister KW, Zhu B, et al. Classification of follicular thyroid tumors by molecular signature: results of gene profiling. Clin Cancer Res 2003;9:1792–800.
7. Mazzanti C, Zeiger MA, Costourous N, et al. Using gene expression profiling to differentiate benign versus malignant thyroid tumors. Cancer Res 2004; 64:2898–903.
8. Kroll TG. Molecular rearrangements and morphology in thyroid cancer. Am J Pathol 2002;160:1941–4.
9. Fusco A, Chiappetta G, Hui P, et al. Assessment of RET/PTC oncogene activation and clonality in thyroid nodules with incomplete morphological evidence of papillary carcinoma: a search for the early precursors of papillary cancer. Am J Pathol 2002;160:2157–67.
10. Simek K, Kimmel M. A note on estimation of dynamics of multiple gene expression based on singular value decomposition. Math Biosci 2003;182: 183–99.
11. Holter NS, Maritan A, Cieplak M, Fedoroff NV, Banavar JR. Dynamic modeling of gene expression data. Proc Natl Acad Sci U S A 2001;98:1693–8.
12. Fujarewicz K, Wiench M. Selecting differentially expressed genes for colon tumor classification. Int J Applied Math Comp Science 2003;13:327–35.
13. Christianini N, Shawe-Taylor J. An introduction to support vector machines and other kernel-based learning methods. Cambridge University Press; 2000.
14. Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using Support Vector Machines. Machine Learning 2002;64:389–422.
15. Fujarewicz K, Kimmel M, Rzeszowska-Wolny J, Swierniak A. A note on classification of gene expression data using support vector machines. J Biol Systems 2003;11:43–56.
16. Golub TR, Slonim DK, Tamayo P, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 1999;286:531–7.
17. Wasenius VM, Hemmer S, Kettunen E, Knuutila S, Franssila K, Joensuu H. Hepatocyte growth factor receptor, matrix metalloproteinase-11, tissue inhibitor of metalloproteinase-1, and fibronectin are up-regulated in papillary thyroid carcinoma: a cDNA and tissue microarray study. Clin Cancer Res 2003;9: 68–75.
18. Kholova I, Ryska A, Ludvikova M, Cap J, Pecen L. Dipeptidyl peptidase IV expression in thyroid cytology: retrospective histologically confirmed study. Cytopathology 2003;14:27–31.
19. Chen WT, Kelly T. Seprase complexes in cellular invasiveness. Cancer Metastasis Rev 2003;22:259–69.
20. Aratake Y, Umeki K, Kiyoyama K, et al. Diagnostic utility of galectin-3 and CD26/DPPIV as preoperative diagnostic markers for thyroid nodules. Diagn Cytopathol 2002;26:366–72.
21. Kehlen A, Lendeckel U, Dralle H, Langner J, Hoang-Vu C. Biological significance of aminopeptidase N/CD13 in thyroid carcinomas. Cancer Res 2003;63:8500–6.
22. Umeki K, Tanaka T, Yamamoto I, et al. Differential expression of dipeptidyl peptidase IV (CD26) and thyroid peroxidase in neoplastic thyroid tissues. Endocr J 1996;43:53–60.
23. Poblete MT, Nualart F, del Pozo M, Perez JA, Figueroa CD. Alpha 1-antitrypsin expression in human thyroid papillary carcinoma. Am J Surg Pathol 1996; 20:956–63.
24. Bartolazzi A, Gasbarri A, Papotti M, et al. Application of an immunodiagnostic method for improving preoperative diagnosis of nodular thyroid lesions. Lancet 2001;357:1644–50.
25. Inaba M, Sato H, Abe Y, Umemura S, Ito K, Sakai H. Expression and significance of c-met protein in papillary thyroid carcinoma. Tokai J Exp Clin Med 2002;27:43–9.
26. Ramirez R, Hsu D, Patel A, et al. Overexpression of hepatocyte growth factor/scatter factor (HGF/SF) and the HGF/SF receptor (cMET) are associated with a high risk of metastasis and recurrence for children and young adults with papillary thyroid carcinoma. Clin Endocrinol (Oxf) 2000;53:635–44.
27. Finley DJ, Arora N, Zhu B, Gallagher L, Fahey TJ III. Molecular profiling distinguishes papillary carcinoma from benign thyroid nodules. J Clin Endocrinol Metab 2004;89:3214–23.
28. Ito Y, Miyoshi E, Uda E, et al. 14-3-3 sigma possibly plays a constitutive role in papillary carcinoma, but not in follicular tumor of the thyroid. Cancer Lett 2003; 200:161–6.
29. Bohm JP, Niskanen LK, Pirinen RT, et al. Reduced CD44 standard expression is associated with tumour recurrence and unfavourable outcome in differentiated thyroid carcinoma. J Pathol 2000;192:321–7.
30. Bieche I, Franc B, Vidaud D, Vidaud M, Lidereau R. Analyses of MYC, ERBB2, and CCND1 genes in benign and malignant thyroid follicular cell tumors by real-time polymerase chain reaction. Thyroid 2001; 11:147–52.
31. Castellone MD, Celetti A, Guarino V, et al. Autocrine stimulation by osteopontin plays a pivotal role in the expression of the mitogenic and invasive phenotype of RET/PTC-transformed thyroid cells. Oncogene 2004; 23:2188–96.
32. Kim JY, Cho H, Rhee BD, Kim HY. Expression of CD44 and cyclin D1 in fine needle aspiration cytology of papillary thyroid carcinoma. Acta Cytol 2002;46:679–83.
33. Bohm J, Niskanen L, Kiraly K, et al. Expression and prognostic value of α-, β-, and γ-catenins in differentiated thyroid carcinoma. J Clin Endocrinol Metab 2000;85:4806–11.
34. Nagy N, Decaestecker C, Dong X, et al. Characterization of ligands for galectins, natural galactoside-binding immunoglobulin G subfractions and sarcolectin and also of the expression of calcyclin in thyroid lesions. Histol Histopathol 2000;15:503–13.
35. Ito Y, Yoshida H, Nakano K, et al. Overexpression of human tumor-associated antigen, RCAS1, is significantly linked to dedifferentiation of thyroid carcinoma. Oncology 2003;64:83–9.
36. Rocha AS, Soares P, Seruca R, et al. Abnormalities of the E-cadherin/catenin adhesion complex in classic papillary thyroid carcinoma and in its diffuse sclerosing variant. J Pathol 2001;194:358–66.
37. Rocha AS, Soares P, Machado JC, et al. Mucoepidermoid carcinoma of the thyroid: a tumour histotype characterised by P-cadherin neoexpression and marked abnormalities of E-cadherin/catenins complex. Virchows Arch 2002;440:498–504.
38. Patel A, Straight AM, Mann H, et al. Matrix metalloproteinase (MMP) expression by differentiated thyroid carcinoma of children and adolescents. J Endocrinol Invest 2002;25:403–8.
39. Lai ML, Rizzo N, Liguori C, Zucca G, Faa G. α-1-antichymotrypsin immunoreactivity in papillary carcinoma of the thyroid gland. Histopathology 1998;33: 332–6.
40. Ito Y, Takeda T, Kobayashi T, et al. Plasminogen activation system in active even in thyroid tumors; an immunohistochemical study. Anticancer Res 1996;16: 81–9.
41. Haugen BR, Larson LL, Pugazhenthi U, et al. Retinoic acid and retinoid X receptors are differentially expressed in thyroid cancer and thyroid carcinoma cell lines and predict response to treatment with retinoids. J Clin Endocrinol Metab 2004;89:272–80.
42. Scarpino S, Stoppacciaro A, Pellegrini C, et al. Expression of EDA/EDB isoforms of fibronectin in papillary carcinoma of the thyroid. J Pathol 1999;188: 163–7.
43. Shi Y, Parhar RS, Zou M, et al. Tissue inhibitor of metalloproteinases-1 (TIMP-1) mRNA is elevated in advanced stages of thyroid carcinoma. Br J Cancer 1999;79:1234–9.
44. Aldred MA, Ginn-Pease ME, Morrison CD, et al. Caveolin-1 and caveolin-2, together with three bone morphogenetic protein-related genes, may encode novel tumor suppressors down-regulated in sporadic follicular thyroid carcinogenesis. Cancer Res 2003;63: 2864–71.
45. Yano Y, Uematsu N, Yashiro T, et al. Gene expression profiling identifies platelet-derived growth factor as a diagnostic molecular marker for papillary thyroid carcinoma. Clin Cancer Res 2004;10:2035–43.
46. Whitney AR, Diehn M, Popper SJ, et al. Individuality and variation in gene expression patterns in human blood. Proc Natl Acad Sci U S A 2003;100:1896–901.
47. Tamimi DM. The association between chronic lymphocytic thyroiditis and thyroid tumors. Int J Surg Pathol 2002;10:141–6.
48. Jarząb B, Włoch J, Wiench M. Molecular changes in thyroid neoplasia. Folia Histochem Cytobiol 2001;39 Suppl 2:26–7.