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

Barbara Jarżab,1 Małgorzata Wiench,1 Krzysztof Fujarewicz,5 Krzysztof Simek,5 Michał Jarżab,2 Małgorzata Oczko-Wojciechowska,1 Jan Włoch,2 Agnieszka Czarniecka,2 Ewa Chmielińska,1 Dariusz Lange,1 Agnieszka Pawłaczek,2 Sylwia Szpak,1 Elżbieta Gubała,1 and Andrzej Świerniak1

Departments of Nuclear Medicine and Endocrine Oncology, Tumor Biology, Oncological Surgery, and Tumor Pathology, Maria Skłodowska-Curie Memorial Cancer Center and Institute of Oncology, Gliwice Branch; and 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 multicenter 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, MKN2, MR2, IGSF1, KIAA0830, RBX6, P4HA2, CDH3, IL13RA1, and MTMR4, and correctly discriminated 17 of 18 additional PTC/normal thyroid samples and all 16 samplespublished 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 non-preselected 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 milliuunits/L). The classic variant of PTC was diagnosed...
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. 13/1 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 Taqmán 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 Taqmán 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) 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 i-th mode to the expression pattern of the k-th gene. Each coefficient was compared with the cutoff value, equal to $W_0^{-1/2}$, where $W_0$ 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. Simiek (ksimek@ia.polsl.gliwice.pl).

**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 as ‘genes’. The whole preprocessed data set is given in Web Appendix.

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 pairwise $t$ 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
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

---

**Table 1. Gene ontology classification of genes from SVD analysis**

<table>
<thead>
<tr>
<th>Class of genes</th>
<th>First mode</th>
<th>Second mode</th>
<th>Third mode</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proliferation and malignant transformation</td>
<td>105 33.9%</td>
<td>63 33.0%</td>
<td>44 22.5%</td>
<td>212 30.4%</td>
</tr>
<tr>
<td>Invasion and metastasis</td>
<td>53 17.1%</td>
<td>16 8.4%</td>
<td>10 5.1%</td>
<td>79 11.3%</td>
</tr>
<tr>
<td>Structure, metabolism and transport</td>
<td>60 19.3%</td>
<td>22 11.5%</td>
<td>17 8.7%</td>
<td>99 14.2%</td>
</tr>
<tr>
<td>Immunoglobulin genes</td>
<td>0 0%</td>
<td>7 3.7%</td>
<td>63 32.1%</td>
<td>70 10.0%</td>
</tr>
<tr>
<td>Other immune-related transcripts</td>
<td>30 9.7%</td>
<td>70 36.6%</td>
<td>14 7.1%</td>
<td>114 16.4%</td>
</tr>
<tr>
<td>Hemoglobin and hemostasis</td>
<td>18 5.8%</td>
<td>0 0%</td>
<td>0 0%</td>
<td>18 2.6%</td>
</tr>
<tr>
<td>Other or unknown</td>
<td>44 14.2%</td>
<td>13 6.8%</td>
<td>48 24.5%</td>
<td>105 15.1%</td>
</tr>
<tr>
<td>All</td>
<td>310</td>
<td>191</td>
<td>196</td>
<td>697</td>
</tr>
</tbody>
</table>

**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.

---

**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.
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. 3A). 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

---

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

<table>
<thead>
<tr>
<th>Gene symbol</th>
<th>Gene name</th>
<th>Gene ID*</th>
<th>Median SLR (range)</th>
<th>Increased/decreased/not changed</th>
<th>References</th>
<th>Validated in Huang data set</th>
</tr>
</thead>
<tbody>
<tr>
<td>DPP4</td>
<td>Dipeptidylpeptidase 4 (CD26, adenosine deaminase complexing protein 2)</td>
<td>1803</td>
<td>6.855 (4.26–7.95)</td>
<td>16/0/0</td>
<td>(2, 18, 42–44)</td>
<td>Yes</td>
</tr>
<tr>
<td>GJB3</td>
<td>Gap junction protein, β-3, 31 kDa (connexin 31)</td>
<td>2707</td>
<td>2.99 (0.81–4.39)</td>
<td>10/0/6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ST14</td>
<td>Suppression of tumorigenicity 14 (colon carcinoma, matriptase, epithin)</td>
<td>6768</td>
<td>0.86 (0.35–1.35)</td>
<td>11/0/5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SERPINA1</td>
<td>Serine (or cysteine) proteinase inhibitor, clade A (α-1 antiproteinase, antitrypsin), member 1</td>
<td>5265</td>
<td>4.395 (2.83–5.45)</td>
<td>16/0/0</td>
<td>(2, 20, 33)</td>
<td>Yes</td>
</tr>
<tr>
<td>LRP4</td>
<td>Low density lipoprotein receptor-related protein 4</td>
<td>4038</td>
<td>4.28 (2.12–5.56)</td>
<td>16/0/0</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>MET</td>
<td>Met proto-oncogene (hepatocyte growth factor receptor)</td>
<td>4233</td>
<td>1.47 (0.1–2.95)</td>
<td>12/0/4</td>
<td>(2, 6, 17, 22, 48)</td>
<td>Yes</td>
</tr>
<tr>
<td>EVA1</td>
<td>Epithelial V-like antigen 1</td>
<td>10205</td>
<td>2.285 (1.54–2.94)</td>
<td>16/0/0</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>MYH10</td>
<td>Myosin, heavy polypeptide 10, nonmuscle</td>
<td>4628</td>
<td>1.445 (0.65–2.63)</td>
<td>16/0/0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SPP1VE</td>
<td>Protease, serine, 23</td>
<td>11098</td>
<td>2.24 (1.01–3.35)</td>
<td>16/0/0</td>
<td>(2)</td>
<td>Yes</td>
</tr>
<tr>
<td>LGALS3</td>
<td>Lectin, galactoside-binding, soluble, 3 (galectin 3)</td>
<td>3958</td>
<td>1.925 (1.02–2.98)</td>
<td>16/0/0</td>
<td>(2, 21, 42)</td>
<td>Yes</td>
</tr>
<tr>
<td>CD44</td>
<td>CD44 antigen (homing function and Indian blood group system)</td>
<td>960</td>
<td>1.165 (0.38–1.92)</td>
<td>15/0/1</td>
<td>(21, 26, 27, 45)</td>
<td></td>
</tr>
<tr>
<td>HBB</td>
<td>Hemoglobin, β</td>
<td>3043</td>
<td>−1.915 (−3.41–0.1)</td>
<td>0/15/1</td>
<td>(46)</td>
<td></td>
</tr>
<tr>
<td>MAP17</td>
<td>Membrane-associated protein 17</td>
<td>10158</td>
<td>3.715 (0.25–4.38)</td>
<td>15/0/1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MKRN2</td>
<td>Makorin, ring finger protein, 2</td>
<td>23609</td>
<td>−0.525 (−0.93–0.08)</td>
<td>0/10/5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TGFα</td>
<td>Transforming growth factor, α</td>
<td>7039</td>
<td>2.895 (1.19–3.99)</td>
<td>16/0/0</td>
<td>(6)</td>
<td></td>
</tr>
<tr>
<td>MRC2</td>
<td>Mannose receptor, C type 2</td>
<td>9902</td>
<td>1.525 (0.49–2.51)</td>
<td>14/0/1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IG5F1</td>
<td>Immunoglobulin superfamily, member 1</td>
<td>3547</td>
<td>4.25 (1.04–5.07)</td>
<td>15/0/1</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>KIA0830</td>
<td>KIAA0830 protein</td>
<td>23052</td>
<td>1.01 (0.16–1.64)</td>
<td>15/0/1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RXRγ</td>
<td>Retinoid X receptor γ</td>
<td>6258</td>
<td>3.53 (0.95–6.78)</td>
<td>16/0/0</td>
<td>(35)</td>
<td></td>
</tr>
<tr>
<td>FBN1</td>
<td>Fibronectin 1</td>
<td>2335</td>
<td>3.2 (2.07–4.44)</td>
<td>16/0/0</td>
<td>(2, 17)</td>
<td>Yes</td>
</tr>
<tr>
<td>MTMR4</td>
<td>Myotubularin-related protein 4</td>
<td>9110</td>
<td>−0.875 (−1.33–0.29)</td>
<td>0/14/2</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>IL13RA1</td>
<td>Interleukin 13 receptor, α1</td>
<td>3597</td>
<td>0.565 (−0.01–1.09)</td>
<td>8/0/7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>P4HA2</td>
<td>Procollagen-proline, 2-oxoglutarate 4-dioxygenase (proline 4-hydroxylase), α polypeptide II</td>
<td>8974</td>
<td>1.55 (0.7–2.27)</td>
<td>16/0/0</td>
<td>(2)</td>
<td>Yes</td>
</tr>
<tr>
<td>CDH3</td>
<td>Cadherin 3, type 1, P-cadherin (placental)</td>
<td>1001</td>
<td>5.61 (2.99–8.16)</td>
<td>16/0/0</td>
<td>(2, 31, 47)</td>
<td></td>
</tr>
<tr>
<td>TLE4</td>
<td>Transducin-like enhancer of split 4 [E(spl) homologue, Drosophila]</td>
<td>7091</td>
<td>−1.26 (−2.43–0.19)</td>
<td>0/12/3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PAXIP1L</td>
<td>PAX transcription activation domain interacting protein 1-like</td>
<td>22976</td>
<td>−0.465 (−0.77–0.11)</td>
<td>0/6/8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HSBP1</td>
<td>Heat shock factor binding protein 1</td>
<td>3281</td>
<td>−0.52 (−0.95–0.05)</td>
<td>0/14/2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NRCAM</td>
<td>Neuronal cell adhesion molecule</td>
<td>4897</td>
<td>1.95 (1.4–3.49)</td>
<td>16/0/0</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>TSTA3</td>
<td>Tissue specific transplantation antigen P35B</td>
<td>7264</td>
<td>0.68 (0.31–1.32)</td>
<td>9/0/7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RNF24</td>
<td>Ring finger protein 24</td>
<td>11237</td>
<td>0.63 (0.11–1)</td>
<td>2/0/13</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(Continued on the following page)
Table 2. All genes obtained by RFR for gene sets of 1 to 100 genes (Cont’d)

<table>
<thead>
<tr>
<th>Gene symbol</th>
<th>Gene name</th>
<th>Gene ID*</th>
<th>Median SLR (range)</th>
<th>Increased/decreased/not changed</th>
<th>References Validated in Huang data set</th>
</tr>
</thead>
<tbody>
<tr>
<td>KIAA0937</td>
<td>KIAA0937 protein</td>
<td>23220</td>
<td>2.49 (1.1–3.69)</td>
<td>16/0/0 (2)</td>
<td>Yes</td>
</tr>
<tr>
<td>TMPRSS4</td>
<td>Transmembrane protease, serine 4</td>
<td>56649</td>
<td>4.305 (2.3–6.12)</td>
<td>14/0/2</td>
<td></td>
</tr>
<tr>
<td>WWOX</td>
<td>WW domain containing oxidoreductase</td>
<td>51741</td>
<td>–1.51 (–2.07–0.05)</td>
<td>0/15/1</td>
<td></td>
</tr>
<tr>
<td>LOCI157567</td>
<td>Hypothetical protein LOCI157567</td>
<td>157567</td>
<td>–1.005 (–1.54–0.52)</td>
<td>0/15/1</td>
<td></td>
</tr>
<tr>
<td>CITED1</td>
<td>Cbp/p300-interacting transactivator, with Glu/Asp-rich carboxyl-terminal domain, 1</td>
<td>4435</td>
<td>4.42 (1.86–7.04)</td>
<td>14/0/2 (2, 3)</td>
<td>Yes</td>
</tr>
<tr>
<td>Hs.297681.2</td>
<td>Homo sapiens PRO2275 mRNA, complete cds (SERPINA1 homologue)</td>
<td>Hs.513816</td>
<td>3.92 (2.4–4.67)</td>
<td>16/0/0</td>
<td>Yes</td>
</tr>
<tr>
<td>HBA2</td>
<td>Hemoglobin, α2</td>
<td>3040</td>
<td>–2.24 (–3.72–0.21)</td>
<td>0/14/2 (2)</td>
<td></td>
</tr>
<tr>
<td>SLIT1</td>
<td>Slit homologue 1 (Drosophila)</td>
<td>6585</td>
<td>0.03 (0.38–0.45)</td>
<td>3/0/21</td>
<td></td>
</tr>
<tr>
<td>PAM</td>
<td>Peptidylglycine α-amidating monoxygenase</td>
<td>5066</td>
<td>0.755 (0.27–1.24)</td>
<td>15/0/1</td>
<td></td>
</tr>
<tr>
<td>FLJ12541</td>
<td>Stimulated by retinoic acid gene 6</td>
<td>64220</td>
<td>1.21 (0.04–2.4)</td>
<td>9/0/7</td>
<td></td>
</tr>
<tr>
<td>SLC34A2</td>
<td>Solute carrier family 34 (sodium phosphate), member 2</td>
<td>10568</td>
<td>4.215 (1.17–5.63)</td>
<td>16/0/0</td>
<td></td>
</tr>
<tr>
<td>AGR2</td>
<td>Anterior gradient 2 homologue (Xenopus laevis)</td>
<td>10551</td>
<td>3.8 (0.89–6.86)</td>
<td>15/0/0</td>
<td></td>
</tr>
<tr>
<td>MINA53</td>
<td>Myc-induced nuclear antigen, 53 kDa</td>
<td>84864</td>
<td>–1.16 (–1.78–0.16)</td>
<td>0/14/2</td>
<td></td>
</tr>
<tr>
<td>ADPRTL1</td>
<td>ADP-ribosyltransferase (NAD+; poly(ADP-ribose) polymerase)-like 1</td>
<td>143</td>
<td>1.05 (0.23–1.92)</td>
<td>16/0/0</td>
<td></td>
</tr>
<tr>
<td>FLJ10748</td>
<td>Hypothetical protein FLJ10748</td>
<td>55220</td>
<td>2.225 (0.47–6.68)</td>
<td>15/0/1</td>
<td></td>
</tr>
<tr>
<td>BMP1</td>
<td>Bone morphogenetic protein 1</td>
<td>649</td>
<td>0.965 (0.46–1.85)</td>
<td>6/0/9</td>
<td></td>
</tr>
<tr>
<td>LAMB3</td>
<td>Laminin, β3</td>
<td>3914</td>
<td>3.695 (2.25–5.92)</td>
<td>16/0/0</td>
<td></td>
</tr>
<tr>
<td>FLJ10178</td>
<td>Hypothetical protein FLJ10178</td>
<td>55086</td>
<td>–3.03 (–5.33–0.07)</td>
<td>0/13/3</td>
<td></td>
</tr>
<tr>
<td>EP88</td>
<td>Epidermal growth factor receptor pathway substrate 8</td>
<td>2059</td>
<td>1.6 (0.23–2.34)</td>
<td>15/0/1 (2)</td>
<td>Yes</td>
</tr>
<tr>
<td>NRP2</td>
<td>Neuruplin 2</td>
<td>8828</td>
<td>1.03 (–0.17–2.78)</td>
<td>12/0/4</td>
<td></td>
</tr>
<tr>
<td>COL13A1</td>
<td>Collagen, type XIII, α 1</td>
<td>1305</td>
<td>2.45 (0.34–3.19)</td>
<td>15/0/1</td>
<td></td>
</tr>
<tr>
<td>SOSTDC1</td>
<td>Sclerostin domain containing 1</td>
<td>25928</td>
<td>–3.365 (–4.7–0.18)</td>
<td>0/11/3</td>
<td></td>
</tr>
<tr>
<td>FLRT3</td>
<td>Fibronectin leucine-rich transmembrane protein 3</td>
<td>23767</td>
<td>1.51 (0.3–3.74)</td>
<td>15/0/1</td>
<td></td>
</tr>
<tr>
<td>SCEL</td>
<td>Sciellin</td>
<td>8796</td>
<td>3.91 (1.04–5.88)</td>
<td>15/0/1 (2)</td>
<td>Yes</td>
</tr>
<tr>
<td>FLII</td>
<td>Flightless 1 homologue (Drosophila)</td>
<td>2314</td>
<td>0.54 (0.35–0.26)</td>
<td>13/0/3</td>
<td></td>
</tr>
<tr>
<td>IKBKE</td>
<td>Inhibitor of light polypeptide gene enhancer in B-cells, kinase epsilon</td>
<td>9641</td>
<td>0.315 (0.13–1.1)</td>
<td>3/0/11</td>
<td></td>
</tr>
<tr>
<td>N33</td>
<td>Putative prostate cancer tumor suppressor</td>
<td>7991</td>
<td>2.885 (1.12–4.09)</td>
<td>16/0/0 (2)</td>
<td>Yes</td>
</tr>
<tr>
<td>EGFL5</td>
<td>Epidermal growth factor-like-domain, multiple 5</td>
<td>1955</td>
<td>1.41 (0.46–1.99)</td>
<td>16/0/0</td>
<td></td>
</tr>
<tr>
<td>FRCP1</td>
<td>Likely orthologue of mouse fibronectin type III repeat containing protein 1</td>
<td>64838</td>
<td>3.075 (0.11–4.61)</td>
<td>11/0/5</td>
<td></td>
</tr>
<tr>
<td>FDF1</td>
<td>Farnesyl-diphosphatase farnesyltransferase 1</td>
<td>2222</td>
<td>–0.525 (–0.98–0.09)</td>
<td>0/12/4</td>
<td></td>
</tr>
<tr>
<td>SFN</td>
<td>Stratifin</td>
<td>2810</td>
<td>1.655 (0.29–3.26)</td>
<td>12/0/4 (24)</td>
<td></td>
</tr>
</tbody>
</table>

(Continued on the following page)
Table 2. All genes obtained by RFR for gene sets of 1 to 100 genes (Cont’d)

<table>
<thead>
<tr>
<th>Gene symbol</th>
<th>Gene name</th>
<th>Gene name*</th>
<th>Gene ID*</th>
<th>Median SLR (range)</th>
<th>Increased/decreased/not changed</th>
<th>References</th>
<th>Validated in Huang data set</th>
</tr>
</thead>
<tbody>
<tr>
<td>SLPI</td>
<td>Secretory leukocyte protease inhibitor (antileukoproteinase)</td>
<td>6590</td>
<td>2.705 (−0.15–4.4)</td>
<td>14/0/2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TREM1</td>
<td>Triggering receptor expressed on myeloid cells 1</td>
<td>54210</td>
<td>2.21 (0.73–3.75)</td>
<td>12/0/4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DAT1</td>
<td>Neuronal specific transcription factor DAT1</td>
<td>55885</td>
<td>1.205 (0.32–2.01)</td>
<td>15/0/1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S100A14</td>
<td>S100 calcium-binding protein A14 (calgizzarin)</td>
<td>Hs.247697.0</td>
<td>1.215 (0.58–1.71)</td>
<td>16/0/0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FLJ21511</td>
<td>Hypothetical protein FLJ21511</td>
<td>80157</td>
<td>−2.215 (−5.87–0.8)</td>
<td>0/16/0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CCND2</td>
<td>Cyclin D2</td>
<td>276</td>
<td>1.315 (0.56–2.84)</td>
<td>16/0/0</td>
<td>(25) Yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SLC27A6</td>
<td>Solute carrier family 27 (fatty acid transporter), member 6</td>
<td>28965</td>
<td>3.455 (0.55–8.7)</td>
<td>14/0/2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EBAG9</td>
<td>Estrogen receptor binding site associated antigen 9</td>
<td>9166</td>
<td>−0.65 (−1.19–0.13)</td>
<td>0/13/3</td>
<td>(30)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>KRT19</td>
<td>Keratin 19</td>
<td>3880</td>
<td>3.225 (1.17–6.61)</td>
<td>16/0/0</td>
<td>(2)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ETV5</td>
<td>Ets variant gene 5 (ets-related molecule)</td>
<td>2119</td>
<td>1.575 (0.7–2.51)</td>
<td>16/0/0</td>
<td>Yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>COMP</td>
<td>Cartilage oligomeric matrix protein</td>
<td>1311</td>
<td>3.835 (1.15–8.41)</td>
<td>15/0/1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CLDN10</td>
<td>Claudin 10</td>
<td>9071</td>
<td>2.32 (0.34–4.32)</td>
<td>13/0/3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CCL15</td>
<td>Chemokine (C-C motif) ligand 15</td>
<td>6359</td>
<td>1.305 (2.5–0.52)</td>
<td>0/14/2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LCN2</td>
<td>Lipocalin 2 (oncogene 24p3)</td>
<td>3934</td>
<td>2.825 (0.82–3.72)</td>
<td>16/0/0</td>
<td>(2, 17, 32)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TIMP1</td>
<td>Tissue inhibitor of metalloproteinase 1 (erythroid potentiating activity, collagenase inhibitor)</td>
<td>25797</td>
<td>3.545 (0.63–4.59)</td>
<td>15/0/1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SLC25A15</td>
<td>Solute carrier family 25 (mitochondrial carrier; ornithine transporter) member 15</td>
<td>10166</td>
<td>1.95 (3.64–0.44)</td>
<td>0/15/0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DUSP6</td>
<td>Dual specificity phosphatase 6</td>
<td>1848</td>
<td>1.88 (0.65–2.91)</td>
<td>16/0/0</td>
<td>(2) Yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BC008967</td>
<td>Hypothetical gene BC008967</td>
<td>89927</td>
<td>1.035 (0.36–2.67)</td>
<td>13/0/3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>COL8A2</td>
<td>Collagen, type VIII, a2</td>
<td>1296</td>
<td>1.99 (−0.81–2.82)</td>
<td>15/1/0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GATM</td>
<td>Glycine amidotransferase</td>
<td>2628</td>
<td>−1.735 (−2.59–0.41)</td>
<td>0/15/1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>QPCT</td>
<td>Glutaminyl-peptide cyclotransferase (glutaminyl cyclase)</td>
<td>25797</td>
<td>3.545 (0.63–4.59)</td>
<td>15/0/1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NDP52</td>
<td>Nuclear domain 10 protein</td>
<td>10241</td>
<td>−0.61 (−1.08–0.01)</td>
<td>0/13/2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PTHF10</td>
<td>PHD finger protein 10</td>
<td>55274</td>
<td>−0.665 (−1.09–0.22)</td>
<td>0/11/5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S100A6</td>
<td>S100 calcium binding protein A6 (calcycin)</td>
<td>6277</td>
<td>0.945 (0.37–1.7)</td>
<td>16/0/0</td>
<td>(29)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CTSC</td>
<td>Cathepsin C</td>
<td>7076</td>
<td>1.98 (0.87–3.5)</td>
<td>16/0/0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DKKZp761K1423</td>
<td>Hypothetical protein DKKZp761K1423</td>
<td>55338</td>
<td>1.915 (1.49–2.92)</td>
<td>16/0/0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NPC2</td>
<td>Niemann-Pick disease, type C2</td>
<td>10577</td>
<td>1.4 (0.76–2.23)</td>
<td>16/0/0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FLJ39207</td>
<td>C219-reactive peptide</td>
<td>348477</td>
<td>−0.75 (−1.47–0.26)</td>
<td>0/12/3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MLF1</td>
<td>Myeloid leukemia factor 1</td>
<td>4291</td>
<td>−1.135 (−2.04–0.31)</td>
<td>0/13/2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>KIAA1006</td>
<td>Tomosyn-like</td>
<td>9515</td>
<td>−0.585 (−1.3–0.06)</td>
<td>0/4/12</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CaMKIIIma</td>
<td>Calcium/calmodulin-dependent protein kinase II</td>
<td>55450</td>
<td>2.41 (0.97–4.05)</td>
<td>16/0/0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CHI3L1</td>
<td>Chitinase 3-like 1 (cartilage glycoprotein-39)</td>
<td>1116</td>
<td>4.45 (−0.08–6.76)</td>
<td>14/0/2</td>
<td>(2) Yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HTCD37</td>
<td>Tcd37 homologue</td>
<td>58497</td>
<td>−0.88 (−1.71–0.16)</td>
<td>0/11/4</td>
<td>(7)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GPB51</td>
<td>G protein-coupled receptor 51</td>
<td>9568</td>
<td>2.445 (0.07–3.55)</td>
<td>15/0/1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CTH</td>
<td>Cystathionase (cystathionine γ-lyase)</td>
<td>1491</td>
<td>−1.365 (−2.33–0.53)</td>
<td>0/15/1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S100A11</td>
<td>S100 calcium binding protein A11 (calgizzarin)</td>
<td>6282</td>
<td>1.335 (0.72–2.42)</td>
<td>15/0/0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PLAU</td>
<td>Plasminogen activator, urokinase</td>
<td>3328</td>
<td>2.56 (0.93–4.17)</td>
<td>16/0/0</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(Continued on the following page)
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. 3B). 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). 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 (SERPIN1; ref. 23), galec tin 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 calcrexin (SI100A6), 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.

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

<table>
<thead>
<tr>
<th>Gene Symbol</th>
<th>Gene Name</th>
<th>Gene ID*</th>
<th>Median SLR (range)</th>
<th>Increased/decreased/not changed</th>
<th>References</th>
<th>Validated in Huang data set</th>
</tr>
</thead>
<tbody>
<tr>
<td>MVP</td>
<td>Major vault protein</td>
<td>9961</td>
<td>1.37 (0.63–2.49)</td>
<td>16/0/0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TCF5</td>
<td>Transcription factor-like 5</td>
<td>10732</td>
<td>–1.01 (–1.46–0.31)</td>
<td>0/14/2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FLJ13946</td>
<td>Hypothetical protein</td>
<td>92104</td>
<td>–1.295 (–2.67–0.61)</td>
<td>0/12/3</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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).
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). 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 SLCA42 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

<table>
<thead>
<tr>
<th>Gene title</th>
<th>Affymetrix ID</th>
<th>Q-PCR assay ID</th>
<th>Microarray fold change</th>
<th>Q-PCR fold change</th>
<th>Spearman correlation coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>CDH3</td>
<td>203256_at</td>
<td>Hs00354998_m1</td>
<td>52.0</td>
<td>13.7</td>
<td>0.875</td>
</tr>
<tr>
<td>SLC34A2</td>
<td>200124_at</td>
<td>Hs00197519_m1</td>
<td>18.4</td>
<td>48.9</td>
<td>0.907</td>
</tr>
<tr>
<td>TMPRSS4</td>
<td>218960_at</td>
<td>Hs00012369_m1</td>
<td>17.1</td>
<td>100.9</td>
<td>0.770</td>
</tr>
<tr>
<td>LRP4</td>
<td>212850_s_at</td>
<td>Hs00323496_m1</td>
<td>16.0</td>
<td>28.5</td>
<td>0.792</td>
</tr>
<tr>
<td>RXRG</td>
<td>205954_at</td>
<td>Hs00199455_m1</td>
<td>11.3</td>
<td>38.5</td>
<td>0.925</td>
</tr>
<tr>
<td>QPCT</td>
<td>205174_s_at</td>
<td>Hs0002002680_m1</td>
<td>9.2</td>
<td>1.5</td>
<td>0.931</td>
</tr>
<tr>
<td>EVA1</td>
<td>2003780_at</td>
<td>Hs00170684_m1</td>
<td>8.0</td>
<td>5.5</td>
<td>0.856</td>
</tr>
<tr>
<td>MRC2</td>
<td>37408_at</td>
<td>Hs001995862_m1</td>
<td>2.6</td>
<td>6.1</td>
<td>0.867</td>
</tr>
<tr>
<td>MTMB4</td>
<td>212277_at</td>
<td>Hs000608316_m1</td>
<td>1.9</td>
<td>0.6</td>
<td>0.570</td>
</tr>
<tr>
<td>ST14</td>
<td>202005_at</td>
<td>Hs00222707_m1</td>
<td>1.7</td>
<td>2.1</td>
<td>0.527</td>
</tr>
</tbody>
</table>

**NOTE:** All correlations are statistically significant with $P < 0.0001.$
Negative values to normal/benign tissue. The values of the discriminant function correspond to papillary cancer diagnosis, 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 Appendix6). Immunoglobulin genes were important source of variability also in Huang's 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.

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 Appendix6). Immunoglobulin genes were important source of variability also in Huang's 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**


**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 discussions.

**References**

1. Eszlinger M, Krohn K, Paschke R. Complementary DNA expression array analysis suggests a lower expression of signal transduction proteins and recep-


3. Prasad ML, Pellegrati NS, Kloos RT, Barbacioru C, Huang Y, de la Chapelle A. CITED1 protein expression suggests papillary thyroid carcinoma in high through-

put tissue microarray-based study. Thyroid 2004;14: 169–75.
Gene Expression Profile of Papillary Thyroid Cancer


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

Barbara Jarzab, Małgorzata Wiench, Krzysztof Fujarewicz, et al.


Updated version
Access the most recent version of this article at:
http://cancerres.aacrjournals.org/content/65/4/1587

Cited articles
This article cites 45 articles, 10 of which you can access for free at:
http://cancerres.aacrjournals.org/content/65/4/1587.full#ref-list-1

Citing articles
This article has been cited by 21 HighWire-hosted articles. Access the articles at:
http://cancerres.aacrjournals.org/content/65/4/1587.full#related-urls

E-mail alerts
Sign up to receive free email-alerts related to this article or journal.

Reprints and Subscriptions
To order reprints of this article or to subscribe to the journal, contact the AACR Publications Department at pubs@aacr.org.

Permissions
To request permission to re-use all or part of this article, contact the AACR Publications Department at permissions@aacr.org.