A robust biomarker of differential correlations improves the diagnosis of cytologically indeterminate thyroid cancers

HUGO GOMEZ-RUEDA[1], REBECA PALACIOS-CORONA[2], HUGO GUTIÉRREZ-HERMOSILLO[3] and VICTOR TREVINO[1]

[1]Bioinformatics Research Group, Department of Research and Innovation, Medical School, Tecnológico de Monterrey, Colonia Los Doctores, 64710 Monterrey, Nuevo León; [2]Northeastern Biomedical Research Center, Instituto Mexicano del Seguro Social, Colonia Independencia, 64720 Monterrey, Nuevo León; [3]Department of Geriatrics, UMAE 1 CMN del Bajío, Instituto Mexicano del Seguro Social, Hospital Aranda de la Parra, Colonia Centro, 37000 León, Guanajuato, Mexico

Received October 3, 2015; Accepted February 23, 2016

DOI: 10.3892/ijmm.2016.2534

Abstract. The fine-needle aspiration of thyroid nodules and subsequent cytological analysis is unable to determine the diagnosis in 15 to 30% of thyroid cancer cases; patients with indeterminate cytological results undergo diagnostic surgery which is potentially unnecessary. Current gene expression biomarkers based on well-determined cytology are complex and their accuracy is inconsistent across public datasets. In the present study, we identified a robust biomarker using the differences in gene expression values specifically from cytologically indeterminate thyroid tumors and a powerful multivariate search tool coupled with a nearest centroid classifier. The biomarker is based on differences in the expression of the following genes: *CCND1*, *CLDN16*, *CPE*, *LRP1B*, MAGI3, *MAPK6*, *MATN2*, *MPPED2*, *PFKFB2*, *PTPRE*, *PYGL*, *SEMA3D*, *SERGEF*, *SLC4A4* and *TIMP1*. This 15-gene biomarker exhibited superior accuracy independently of the cytology in six datasets, including The Cancer Genome Atlas (TCGA) thyroid dataset. In addition, this biomarker exhibited differences in the correlation coefficients between benign and malignant samples that indicate its discriminatory power, and these 15 genes have been previously related to cancer in the literature. Thus, this 15-gene biomarker provides advantages in clinical practice for the effective diagnosis of thyroid cancer.

Correspondence to: Dr Victor Trevino, Bioinformatics Research Group, Department of Research and Innovation, Medical School, Tecnológico de Monterrey, Av. Morones Prieto 3000 Poniente, Colonia Los Doctores, 64710 Monterrey, Nuevo León, Mexico
E-mail: vtrevino@itesm.mx

Abbreviations: FNA, fine-needle aspiration; RT-PCR, real-time-polymerase chain reaction; NC, nearest centroid

Key words: biomarker, diagnostic, indeterminate, thyroid

Introduction

The incidence of thyroid cancer has been increased over the past few years (1-3). Thyroid nodules are one of most prevalent thyroid diseases, detectable by cervical echography in between 50 and 67% of healthy individuals (4). A confirmation study is required to verify the diagnosis as only 5% of these thyroid nodules are malignant (4).

Usually, diagnosing thyroid nodules as benign or malignant is performed by cytological evaluation (5). For this purpose, fine-needle aspiration (FNA) is the most commonly used sample extraction technique, since it is rapid, inexpensive and simple, as shown by Knezević-Usaj *et al* (6). Subsequent cytological analysis following FNA provides four possible results: non-diagnostic, positive to malignancy or suspicious, indeterminate and benign cytology (7). Indeterminate FNA cytological results are obtained in between 15 to 30% of cases (8-11). Moreover, only between 5 and 15% of cases are malignant, particularly in those with indeterminate cytological results (11). Consequently, patients with FNA indeterminate cytological results undergo diagnostic surgery, even though this has been proven to be unecessary in >50% of cases where patients are later found to have benign disease (5,12).

Microarray-based gene expression profiling studies of thyroid nodules have proposed molecular markers (11,13,14). However, it has been demonstrated in other types of cancer that some biomarkers identified in one cohort may fail to reproduce similar results with a high degree of accuracy in other cohorts (15,16). For example, the accuracy of the Afirma® genomic test for FNA thyroid samples, which is based on the expression of >170 genes and one of the most extensively studied, has been confirmed by some studies (17-20); however, it has been seriously questioned by more recent investigations in terms of its sensitivity, cost-effectiveness, or its ability to complement tests with a high specificity such as the BRAF mutation test (21-24).

Given that those patients with thyroid nodules of indeterminate FNA cytology may undergo unnecessary surgical intervention, and that previously proposed molecular

biomarkers cannot be used in these cases, or that the accuracy of a currently available test has been questioned, a molecular, robust, biomarker for FNA, that is simple, and cost-effective, is still required for clinical investigations and practice. In contrast to other authors, in this study, we propose a molecular biomarker designed specifically from FNA indeterminate thyroid samples identified by a bioinformatics approach, which has been validated in six datasets, including four from other authors. We demonstrate that the accuracy of the proposed biomarker is superior to other previously proposed biomarkers for thyroid tumors. The proposed biomarker is composed of 15 genes and has the potential to be easily implemented into clinical practice using common and cost-effective real-time-polymerase chain reaction (RT-PCR) assays.

Data collection methods

Datasets and processing. We used six gene expression microarray datasets from five different authors (Table I), which we obtained from large microarray repositories. The main inclusion criteria were that the number of samples was >40 and that the study contained histopathological diagnoses. To compare the results from the different datasets and microarray platforms, we transformed the gene expression data to a uniform distribution between 0 and 1, where 0 represents the lowest and 1 the highest expression. Multiple probes assigned to the same gene were averaged if they were correlated using a Pearson coefficient of ≥0.7. The probe with the highest expression was used if duplicate symbols remained. To facilitate future biomarker measurements in clinical practice using RT-PCR, which may use an internal control for normalization (25), we transformed the original Alexander dataset of 173 genes (11), which represent the previously identified Afirma® test, to a dataset of all combinations of gene-by-gene expression differences. This generated a dataset of 2,850 gene expression differences. In preliminary experiments, we observed that differences allowed better prediction than the raw expression measure (data not shown), which is consistent with other observations, where pairs of genes are more accurate predictors than separate genes, as shown by Grate (26).

Biomarker identification. To the best of our knowledge, the Alexander dataset is the only data providing details of cytologically indeterminate thyroid samples [Alexander *et al* (11)]; therefore, we used this 'training' dataset as a gold standard in order to identify the biomarker. To discover combinations of gene differences that together yield the optimal classification of malignant and non-malignant samples, we used GALGO, a genetic algorithm for feature selection (27). GALGO is a feature selection approach based on genetic algorithms coupled with a classifier. Briefly, GALGO first generates a population of random combinations of features. Each combination of this population is evaluated using the accuracy of a classifier and the selected features. The genetic algorithm then selects those combinations with higher accuracy, which are subsequently re-combined and changed replacing a gene difference with another. The process is repeated until a predefined number of cycles yields a highly accurate feature combination. Since this process is stochastic, the specific features may change; thus, GALGO typically performs this procedure multiple times. Subsequently, a representative feature combination is selected based on the number of times each feature is present in the highly accurate combinations and a forward selection procedure. In this study, as proposed by GALGO tutorials (http://bioinformatica.mty.itesm.mx/GALGO), we used 300 combinations having five features to select the representative biomarker. For the classes, we used benign and malignant cytology as the sample class. For classification, we used the nearest centroid (NC) method shown in the study by Dabney (28). The NC method is based on centers per gene per class estimated as the mean of the gene expression values from the samples of the same class. The samples were classified as the class with the minimum Euclidean distance. The GALGO tutorial has further details of the genetic algorithm and the NC classifier (27). This procedure was performed using the subset of the Alexander dataset (GSE34289) that corresponded to indeterminate FNA cytology and post-surgery determinate, which is composed of 188 samples, 131 as benign and 57 as malignant. We used only 57 randomly selected benign samples for training to balance the number of benign samples with malignant samples.

Biomarker evaluation. To evaluate the performance of the proposed biomarkers and those proposed by other authors, we used an NC classifier learning the parameters from the gene expression measurements of their corresponding datasets. Given that the Alexander dataset is the only available data providing details of indeterminate thyroid samples (11), we used this dataset as the gold standard to evaluate the performance of biomarkers in the undetermined samples. For the cytologically-determined samples, we used the other five datasets shown in Table I as Test datasets.

Results

Previously proposed thyroid cancer biomarkers are not robust. To predict thyroid tumor malignancy, we compared the accuracy of four previously described biomarkers involving between 3 and 167 genes (10,11,14,29) evaluated in six datasets using an NC classifier. The average accuracy ranged between 73 and 78% (Fig. 1). However, we observed some issues. Firstly, none of these four biomarkers accurately predicted The Cancer Genome Atlas (TCGA) subtypes; the maximum was 55%. Secondly, the accuracies evaluated in the 265 indeterminate FNA samples (Alexander dataset) were poor. Thirdly, two of the biomarkers needed almost 100 genes or more, which would generate technical and economic difficulties in clinical practice.

Identification of a highly accurate and robust 15-gene biomarker. The application of previously proposed biomarkers was associated with several concerns: low accuracy, lack of robustness, need to screen of a high number of genes, as well as poor performance when used to analyze indeterminate samples. These biomarkers were all identified through the study of thyroid samples with a definitive cytological diagnosis. Therefore, we specifically selected thyroid tumors with indeterminate cytology from the Alexander GSE34289 dataset (11). To facilitate measurements in a clinical laboratory using RT-PCR and to improve accuracy, we used all combinations of gene differences (26,30-32) instead of the 173 gene expression profiles in GSE34289. To select a

Table I. Characteristics of datasets used.

Authors/(Refs.) dataset/(use)	ID/Platform	Sample characteristics	No. of benign/ malignant samples	Diagnosis
Alexander *et al* (11) indeterminate (training set)	GSE34289 Affymetrix Afirma-T (custom) 173 probes	265 Indeterminate FNA: 180 Benign after surgery (B) 85 Malignant after surgery	180/85	FNA cytology
Giordano *et al* (13) (test set)	GSE27155 Affymetrix HG_U133A 22,283 probes	89 Adenomas/carcinomas: 10 Follicular adenomas (B) 7 Oncocytic adenomas (B) 13 Follicular carcinomas 8 Oncocytic carcinomas 51 Papillary carcinomas	17/72	Surgical pathology
Borup *et al* (14) (test set)	E-MEXP-2442 Affymetrix HG U133 Plus 2.0 54,613 probes	69 Adenomas/carcinomas: 22 Follicular adenomas (B) 12 Microfollicular adenomas (B) 9 Nodular goiters (B) 18 Follicular carcinomas 4 Anaplastic carcinomas 2 Papillary carcinomas 2 Normal (B)	45/24	Surgical pathology
Alexander *et al* (11) determinate (test set)	GSE34289 Affymetrix Afirma-T (custom) 173 probes	99 Determinate FNA: 44 Benign after surgery (B) 55 Malignant after surgery	44/55	FNA cytology
TCGA (test set) (https://tcga-data.nci.nih.gov/tcga/)	Illumina Hi-Seq RNA-Seq 20,500 probes	547 Thyroid samples: 490 Papillary cancers 57 Benign tissues (B)	57/490	Surgical pathology
Tomás *et al* (51) Dom *et al* (52) (test set)	GSE33630 Affymetrix HG_U133 Plus 2.0 54,675 probes	105 Thyroid tumor/non-tumor: 11 Anaplastic carcinomas 49 Papillary carcinomas 45 Patient-matched non-tumor controls (B)	45/60	International Pathology Panel of the Chernobyl Tissue Bank

B, indicates benign samples; FNA, fine-needle aspiration.

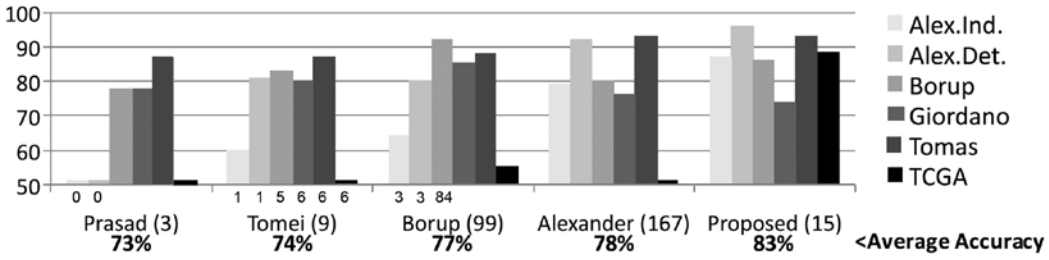

Figure 1. Evaluation of proposed and previous biomarkers for thyroid cancer in six datasets. The accuracy, shown on the vertical axis, was estimated using a nearest centroid classifier. Biomarkers are shown on the horizontal axis along with the different datasets indicated by the bars in different shades of gray. Average accuracy is shown below each biomarker. The number of genes per biomarker is shown in parenthesis. The number below some bars corresponds to the number of genes found in the corresponding dataset. For the Tomei biomarker, we did not use BRAF exon 15 status.

low number of genes, we used a multivariate search to identify optimal combinations (27). This strategy is based on genetic algorithms coupled with an NC classifier. Finally, to validate the proposed biomarker *in silico*, we used five additional datasets (Table I).

The average accuracy (83%) of the proposed biomarker was superior to the other biomarkers (Fig. 1). The proposed biomarker was the most accurate in four of the six datasets, including the TCGA dataset and the cytologically indeterminate samples; it was also highly competitive in the remaining two datasets [Borup (14) and Giordano (13)]. The biomarker identified using GALGO was based on 15 gene differences covering 15 genes (*CCND1*, *CLDN16*, *CPE*, *LRP1B*, *MAGI3*, *MAPK6*, *MATN2*, *MPPED2*, *PFKFB2*, *PYGL*, *PTPRE*, *SEMA3D*,

Table II. Differences in centroids between benign and malignant samples across datasets.

	Giordano		Borup		Alex. Ind		Alex. Det		TCGA		Tomás		
Gene difference	Diff	p-value[a]	Diff	p-value[a]	Diff	p-value[a]	Diff	p-value[a]	Diff	p-value[a]	B	p-value[a]	Highly[b] significant
MPPED2 - CPE	0.89	5	4.15	6	0.35	10	3.67	20	-0.05	8	0.29	20	6
LRP1B - CPE	0.45	2	2.91	10	0.39	13	3.53	23	-0.09	31	-0.26	32	5
PYGL - TIMP1	0.29	2	0.98	2	0.19	12	2.61	19	0.17	14	-0.26	NS	3
SLC4A4 - CPE	0.73	3	3.19	4	0.29	8	3.14	18	-0.1	20	-0.21	28	6
PFKFB2 - CLDN16	0.66	6	0.94	NS	0.44	13	4.47	19	0.17	37	0.00	15	5
MATN2 - CPE	0.66	3	1.53	2	0.21	6	2.55	13	0.24	13	0.01	6	5
MAGI3 - CLDN16	-	-	1.18	2	0.51	12	3.44	19	0.08	15	0.33	20	4
PFKFB2 -CCND1	0.25	NS	-0.22	NS	0.17	8	2.04	19	0.18	15	-0.35	10	4
PFKFB2 - CPE	-0.65	NS	-1.23	2	0.13	2	2.13	9	-0.11	23	-0.47	NS	2
SERGEF - PFKFB2	0.04	NS	0.44	NS	-0.14	7	-1.06	7	-0.19	14	0.26	3	4
CPE - MAPK6	-0.26	NS	0.22	NS	-0.4	10	-4.7	24	0.09	26	0.03	4	4
CPE - SEMA3D	-0.34	NS	-1.17	2	-0.38	9	-4.06	17	-0.01	2	0.47	21	3
CCND1 - LRP1B	-0.5	5	-2.66	9	-0.41	15	-3.43	26	-0.1	13	0.14	35	6
PTPRE - PYGL	-0.37	7	-0.4	NS	-0.1	8	-0.96	9	-0.07	14	0.08	6	5
CLDN16 - SEMA3D	-0.79	11	-2.08	3	-0.67	12	-5.8	18	-0.25	41	-0.0	29	6

[a]p-values are expressed as -log10 (p-values); [b]number of datasets with a p-value <0.001. Ind, indeterminate; Det, determinate; Diff, differences in the centroids; NS, not significant.

SERGEF, *SLC4A4* and *TIMP1*). This signature seems to be preserved across the six datasets, exhibiting clear differences between malignant and benign samples (Fig. 2). Twelve of these gene differences were statistically altered in four, five or six datasets (Table II). We also tested the signature across available strata. We observed a high degree of accuracy which was independent of gender, age, tumor size and ethnicity (Table III).

Genes in biomarker play important roles in cancer. Remarkably, the majority of the 15 genes that compose the biomarker have been previously associated with cancer. *CLDN16* has been shown to be elevated in patients with thyroid papillary cancer (33). *LRP1B* inactivation has been shown to influence the tumor environment, thereby increasing the growth and invasiveness of thyroid cancer cells (34). *SLC4A4* is expressed in low levels in papillary thyroid carcinoma (35). *TIMP1*, an inhibitor of the metalloproteinases in the extracellular matrix (36), has been shown to be highly expressed in thyroid cancer (37,38). *CCND1*, which is involved in the inactivation of the retinoblastoma (RB) protein, as well as in the G1-S phase transition within the cell cycle, has been shown to be associated with many tumors (39), including thyroid papillary carcinomas and follicular adenomas and carcinomas, as shown by Seybt *et al* (40). *SEMA3D*, a semaphorin that guides migrating cells during developmental morphogenesis and in adult tissues (41), has been shown to have anti-tumorigenic properties (42). The expression levels of *CLDN16*, *LRP1B*, *SLC4A4*, *TIMP1*, *CCND1* and *SEMA3D* across subtypes in the six datasets analyzed in the present study were consistent with the findings of the above-mentioned studies (Fig. 3). *PFKFB2*, which is involved in the control of glycolysis, has been shown to be highly expressed in patients with papillary thyroid cancer aged >40 years compared with younger patients (43). However, in the present study, *PFKFB2* appeared to be more highly expressed in the benign tumors across the datasets. *CPE* mutations have been related to deficiencies in thyrotropin-releasing hormone (44), suggesting that it plays important roles in the thyroid gland. *CPE* has been shown to be associated with tumor growth and metastases in pheochromocytomas and others types of cancer (45). In this study, we found a consistently high expression of *CPE* in malignant tumors; however, *CPE* expression levels varied in the benign samples. *MATN2* and *MPPED2* seem to be highly correlated (r=0.7 in benign samples in Alexander dataset) and highly expressed in the benign thyroid gland (Fig. 3). It is well known that the former is involved in the formation of filamentous networks in the extracellular matrix and the latter displays low metallophosphoesterase activity. *MPPED2* has been proposed to play an important role in neuroblastoma tumorigenesis (46) and the increased expression of this gene has been shown to be associated with a good prognosis (46), which is consistent with the higher expression observed in the benign thyroid tumors in the present study. By contrast, *MATN2* overexpression has been observed in pilocytic astrocytoma (47). *MAPK6* is a member of the Ser/Thr protein kinase family that has been found to be associated with tumor invasion in lung cancer (48). Polymorphisms in *MAGI3* and *PYGL* have been associated with various disorders, *MAGI3* with hypothyroidism (49) and *PYGL* with relapse in leukemia (50). This literature review of the involved genes suggests that the majority play or may play an important role in thyroid tumors.

Alterations in correlation coefficients characterize differences in gene expression. Eight genes are included in only one difference and seven in more than one difference (Fig. 2). Notably, the *CPE* gene was found in seven gene differences indicating an important contribution within the biomarker. We

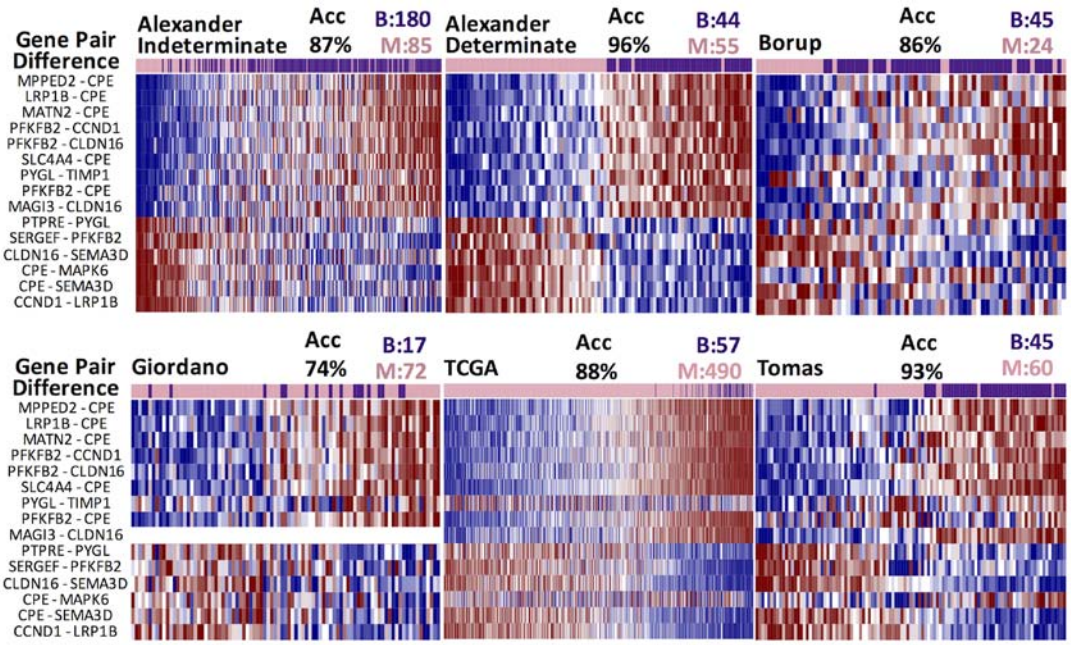

Figure 2. Profiles of the proposed biomarker in the six datasets. The heatmaps show samples in columns and gene differences in rows. Purple and pink columns represent benign (B) and malignant (M) samples, respectively. The total number of samples per subtype is shown as 'B' and 'M'. Accuracy (Acc) is indicated. Cells within the heatmap are shaded in colors; red for high expression, white for median expression, and blue for low expression. The difference involving the *MAGI3* gene that was not found in Giordano was removed.

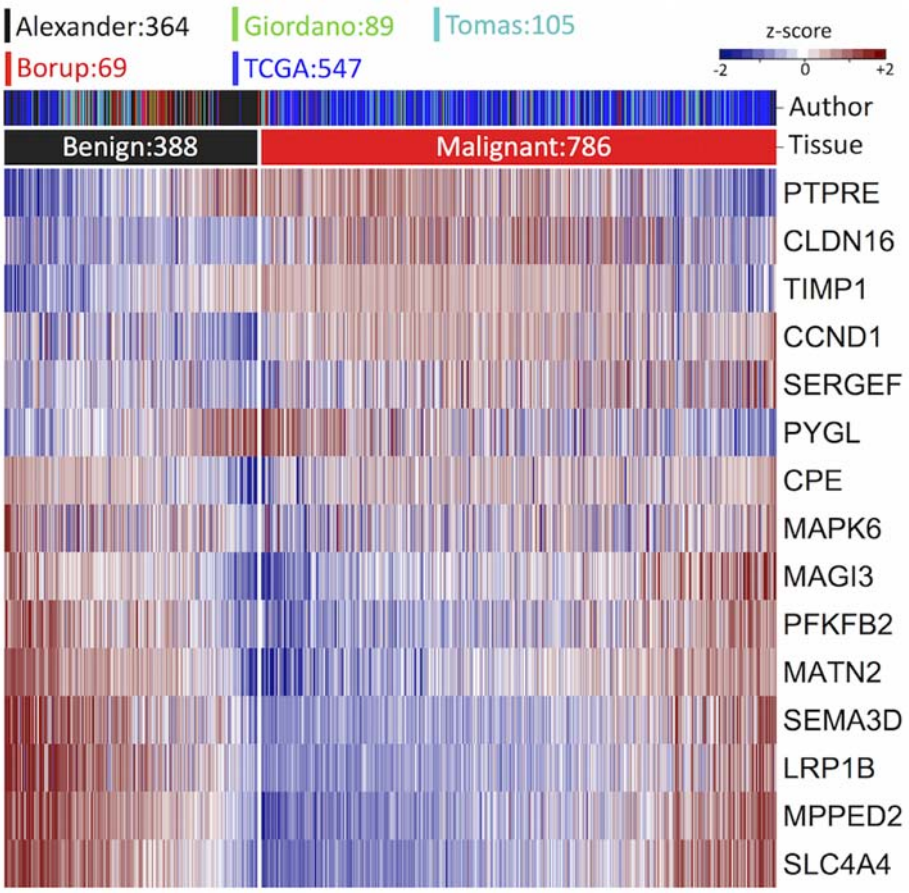

Figure 3. Expression levels across six datasets of the genes included in the proposed biomarker. Cells within the heatmap are shaded in colors; red for high expression, white for median expression, and blue for low expression.

observed a higher correlation of genes combined with *CPE* in the benign samples compared with those correlations in the malignant tumor samples from the Alexander indeterminate dataset (Fig. 4A). By contrast, a higher correlation between the

A Details of the Spearman Correlations in Alexander Indeterminate Dataset

SLC4A4-CPE	MATN2-CPE	LRP1B-CPE	MPPED2-CPE	CPE-SEMA3D	CPE-MAPK6	PFKFB2-CPE	PFKFB2-CLDN16	PFKFB2-CCND1	PYGL-TIMP1	PTPRE-PYGL	MAGI3-CLDN16	SERGEF-PFKFB2	CCND1-LRP1B	CLDN16-SEMA3D	
0.71	0.55	0.57	0.70	0.74	0.58	0.37	0.08	0.42	0.72	0.90	0.27	0.32	0.32	0.14	Benign
↓		↓	↓	↓	↓			↓	↓	↓			↓		
0.09	0.30	-0.05	0.13	0.16	0.01	0.10	-0.13	-0.06	-0.06	0.66	0.18	0.06	-0.45	-0.14	Malignant

B Spearman Correlations in All Datasets

	SLC4A4-CPE	MATN2-CPE	LRP1B-CPE	MPPED2-CPE	CPE-SEMA3D	CPE-MAPK6	PFKFB2-CPE	PFKFB2-CLDN16	PFKFB2-CCND1	PYGL-TIMP1	PTPRE-PYGL	MAGI3-CLDN16	SERGEF-PFKFB2	CCND1-LRP1B	CLDN16-SEMA3D	
Alex.Det.	0.76	0.58	0.74	0.73	0.75	0.79	0.35	0.30	0.50	0.69	0.85	0.10	0.37	0.65	0.19	Benign
Alex.Ind.	0.71	0.55	0.57	0.70	0.74	0.58	0.37	0.08	0.42	0.72	0.90	0.27	0.32	0.32	0.14	
Borup	-0.11	0.41	0.63	0.60	0.53	-0.56	-0.26	-0.27	-0.14	0.37	0.27	0.18	0.01	0.48	0.21	
Giordano	0.47	0.02	0.48	0.68	0.38	-0.43	-0.30	-0.19	-0.48	0.13	0.07	0.00	0.02	0.49	0.45	
TCGA	0.75	0.76	0.78	0.67	0.68	0.22	0.56	0.13	0.03	0.08	0.04	0.11	-0.13	0.26	0.20	
Tomas	0.55	0.67	0.75	0.50	0.33	-0.15	-0.05	0.25	0.00	0.00	-0.03	0.21	0.21	0.16	-0.04	
Alex.Det.	⇩ 0.85	⇩ 0.39	⇩ 0.89	⇩ 0.81	⇩ 0.85	⇩ 0.84	⇩ 0.51	0.14	⇩ 0.58	⇩ 1.16	⇩ 0.54	⇧ -0.30	⇩ 0.62	⇩ 0.82	0.19	Δ Malignant
Alex.Ind.	⇩ 0.62	0.25	⇩ 0.62	⇩ 0.57	⇩ 0.58	⇩ 0.57	0.27	0.21	⇩ 0.48	⇩ 0.78	0.24	0.09	0.26	⇩ 0.77	0.28	
Borup	-0.23	0.23	⇩ 0.62	⇩ 0.33	⇩ 0.40	⇧ -0.98	-0.06	-0.19	-0.18	-0.28	-0.13	-0.10	⇧ -0.45	⇩ 0.40	⇩ 0.61	
Giordano	⇩ 0.85	⇩ 0.31	⇩ 0.80	⇩ 0.88	⇩ 0.43	⇧ -0.54	0.14	⇩ 0.36	0.01	-0.13	-0.25	0.00	0.05	⇩ 0.68	⇩ 0.62	
TCGA	⇩ 0.87	⇩ 0.80	⇩ 0.81	⇩ 0.66	⇩ 0.63	0.10	⇩ 0.79	⇩ 0.44	⇩ 0.32	-0.16	-0.27	⇩ 0.38	-0.04	⇩ 0.64	⇩ 0.50	
Tomas	⇩ 0.31	⇩ 0.39	⇩ 0.56	0.14	0.03	-0.11	⇧ -0.41	0.04	-0.20	-0.16	0.00	-0.06	-0.03	0.08	-0.10	
Conserved	5	4	6	5	5	4	3	2	3	2	1	2	2	5	3	

Figure 4. Differences in correlation coefficients between benign and malignant samples. (A) Spearman correlation coefficients in benign and malignant samples of the genes involved in each difference for the Alexander Indeterminate dataset. Arrows mark differences >0.3. (B) Spearman correlation coefficients (horizontal axis) across datasets (vertical axis) in benign samples (top section of table) and the decrease in spearman correlation coefficients in malignant samples (bottom section of table). Columns sorted as in (A). Downward gray arrows mark a decrease >0.3 in correlation coefficients, while upward arrows show an increase >0.3 in correlation coefficients (negative decrease). The 'conserved' cells show the number of datasets whose difference in correlation coefficients is >0.3. Ind, indeterminate; Det, determinate.

Table III. Accuracy of the biomarker in groups of samples.

Authors/(Refs.) Dataset	Groups	Samples	Accuracy
Alexander *et al* (11) (indeterminate)	Data available	265	0.82
	Male	61	0.82
	Female	204	0.82
	Age ≤47	103	0.79
	Age >47	162	0.83
Alexander *et al* (11) (determinate)	Data available	102	0.94
	Male	27	0.91
	Female	75	0.94
	Age ≤47	38	0.91
	Age >47	64	0.95
Borup *et al* (14)	Data available	69	0.86
	Male	23	0.71
	Female	46	0.89
TCGA (https://tcga-data.nci.nih.gov/tcga/)	Data available	555	0.88
	Male	151	0.89
	Female	404	0.87
	Tumor size ≤3 cm	323	0.90
	Tumor size >3 cm	196	0.86
	Asian	54	0.94
	African-American	25	0.86
	Latino	42	0.94
	Caucasian	330	0.90
Tomás *et al* (51)	Data available	105	0.93
Dom *et al* (52)	(no strata available)	-	-
Giordano *et al* (13)	Data available	89	0.74
	(no strata available)	-	-

malignant samples was not observed. A similar analysis of all gene pairs across the datasets confirmed this trend (Fig. 4B).

Discussion

Previously proposed biomarkers were not robust across datasets or indeterminate FNA samples when evaluated under similar conditions *in silico*. This may be due to characteristics of the samples, microarray technology, or the methodology used for biomarker identification. Whereas other studies have focused on determinate samples in order to identify biomarkers (10,11,13,14,29,51,52), we specifically used indeterminate samples as the training set, and therefore, we captured the particular expression signatures in these samples. We then showed that the signatures were also conserved in five studies of determinate tumors, which validates the proposed signature. For this purpose, we used gene expression differences between pairs of genes and a multivariate search methodology. Notably, the proposed biomarker is more compact and more accurate than other previously proposed biomarkers.

The proposed biomarker was found to be robust when evaluated in other databases and across patient characteristics (tumor size, age, gender and ethnicity). These results suggest that differences in expression are independent of the cohort, methodology, genomic technology and particular characteristics of the cohort and thus, it is highly likely to represent true biological alterations. In the Giordano (13) and Borup (14) databases, the biomarker was capable of classifying, with high accuracy, many cellular types of thyroid cancer. In the case of determinate FNA samples in the Alexander database (11), the performance of the biomarker was higher (96%) than that of the indeterminate ones (87%) even though the latter was used to identify the biomarker. This result

suggests that indeterminate samples may contain transitional stages between benign and malignant subtypes.

The differences in gene expression allow for the easier measurement in widely used technologies, such as RT-PCR, thereby facilitating implementation in clinical practice. Surprisingly, most of the gene pairs in differences were highly correlated in the benign tumors and poorly correlated in the malignant tumors. This concurs with observations in prostate (53), colon, lung, pancreatic, cervical and gastric cancers (54) where the tumor correlation distribution is different than in normal counterparts, generally sharper around zero. Notably, these results suggest that differences in correlations may be an important characteristic of tumor transformation, which may be exploited for biomarker identification, cancer prognosis and gene targeting. We hypothesized that these differences between malignant and non-maligant samples were important for the multivariate search and the classifier to select the genes involved in the proposed biomarker. This may explain the high number of occurrences of the CPE gene, which showed the largest differences in correlation coefficients (ten in benign and one in malignant samples).

From the 15 genes identified in our biomarker, which is a subset of those from the study by Alexander *et al* (11), none are similar to those proposed by Prasad *et al* (*HMGA2*, *MRC2*, and *SFN*) (55) and Tomei *et al* (*KIT*, *C21orf4*, *PDK3/Hs.296031*, *DDI2*, *CDH1*, *LSM7* and *TC1*) (29), and only two (*MATN2* and *MPPED2*) are included in the genes from the study by Borup *et al* (14). Thus, the proposed signature combined with the use of gene differences and a NC classifier appears to be distinctive.

The contribution of the multivariate search was important since the other methodologies tested, such as PAM-R (56) and support vector machine recursive feature elimination (SVM-RFE) (57), generated lower accuracies (65 and 75%, respectively) or higher numbers of gene differences (85 and 15, respectively). Besides the multivariate search, we believe that the use of the gene-pair difference was an important factor which enabled us to identify the highly accurate marker. We then showed that the gene-pair difference may also associated with the difference in correlation coefficients between genes and tumor subtypes. Nevertheless, this approach is almost prohibited in large datasets since a dataset of 20,000 genes would generate 200 million differences combinations. Thus, the use of the 173 genes (or a stringent gene filter) was also a critical factor.

As the proposed biomarker was only tested *in silico*, a validation study is warranted to confirm the potential use of this biomarker in clinical practice. Although we aim to explore this line of research in the near future, the availability of the proposed biomarker and the methodology used may encourage other research groups to test the biomarker or to design better ones.

In conclusion, the proposed biomarker is composed of 15 gene differences involving 15 genes. The majority of the genes have been associated with cancer and some specifically with thyroid cancer in the research literature. Our analysis suggests that the proposed biomarker is more accurate and robust than previous thyroid biomarkers in tumors and indeterminate FNA samples. Measuring the biomarker may be made relatively easy by RT-PCR facilitating implementation. Changes in the gene expression correlations between benign and malignant samples may be associated with tumor progression and may explain the presence and robustness of the gene differences that compose the proposed biomarker.

Acknowledgements

The present study was supported by Grupo de Investigación con Enfoque Estratégico en Bioinformática of the Instituto Tecnológico y de Estudios Superiores of Monterrey, CONACyT (Posgrado Nacional 002087 and grant scholarship 339770). We thank the Instituto Tecnológico y de Estudios Superiores of Monterrey, Hospital San José de Monterrey, and the Instituto Mexicano del Seguro Social for supporting this study.

References

1. Aschebrook-Kilfoy B, Ward MH, Sabra MM and Devesa SS: Thyroid cancer incidence patterns in the United States by histologic type, 1992-2006. Thyroid 21: 125-134, 2011.
2. Sadowski SM, Köhler BB, Meyer P, Pusztaszeri M, Robert JH and Triponez F: Treatment of differentiated thyroid cancer. Rev Med Suisse 8: 1321-1325, 2012 (In French).
3. Kazaure HS, Roman SA and Sosa JA: Aggressive variants of papillary thyroid cancer: Incidence, characteristics and predictors of survival among 43,738 patients. Ann Surg Oncol 19: 1874-1880, 2012.
4. Liénart F: Thyroid nodule: Benign or malignant? Rev Med Brux 33: 254-262, 2012 (In French).
5. Lew JI, Snyder RA, Sanchez YM and Solorzano CC: Fine needle aspiration of the thyroid: correlation with final histopathology in a surgical series of 797 patients. J Am Coll Surg 213: 188-195, 2011.
6. Knezević-Usaj S, Eri Z, Panjković M, Klem I, Petrović T, Ivković-Kapicl T, Karapandzić A and Jelić J: Diagnostic relevance of fine needle aspiration cytology in nodular thyroid lesions. Vojnosanit Pregl 69: 555-561, 2012 (In Serbian).
7. American Thyroid Association (ATA) Guidelines Taskforce on Thyroid Nodules and Differentiated Thyroid Cancer; Cooper DS, Doherty GM, Haugen BR, Kloos RT, Lee SL, Mandel SJ, Mazzaferri EL, McIver B, Pacini F, Schlumberger M, *et al*; Revised American Thyroid Association management guidelines for patients with thyroid nodules and differentiated thyroid cancer. Thyroid 19: 1167-1214, 2009.
8. Mehanna R, Murphy M, McCarthy J, O'Leary G, Tuthill A, Murphy MS and Sheahan P: False negatives in thyroid cytology: impact of large nodule size and follicular variant of papillary carcinoma. Laryngoscope 123: 1305-1309, 2013.
9. Duick DS: Overview of molecular biomarkers for enhancing the management of cytologically indeterminate thyroid nodules and thyroid cancer. Endocr Pract 18: 611-615, 2012.
10. Prasad NB, Kowalski J, Tsai HL, Talbot K, Somervell H, Kouniavsky G, Wang Y, Dackiw AP, Westra WH, Clark DP, *et al*: Three-gene molecular diagnostic model for thyroid cancer. Thyroid 22: 275-284, 2012.
11. Alexander EK, Kennedy GC, Baloch ZW, Cibas ES, Chudova D, Diggans J, Friedman L, Kloos RT, LiVolsi VA, Mandel SJ, *et al*: Preoperative diagnosis of benign thyroid nodules with indeterminate cytology. N Engl J Med 367: 705-715, 2012.
12. Tamez-Pérez HE, Gutiérrez-Hermosillo H, Forsbach-Sánchez G, Gómez-de Ossio MD, González-González G, Guzmán-López S, Tamez-Peña AL, Mora-Torres NE and González-Murillo EA: Nondiagnostic thyroid fine needle aspiration cytology: outcome in surgical treatment. Rev Invest Clin 59: 180-183, 2007.
13. Giordano TJ, Kuick R, Thomas DG, Misek DE, Vinco M, Sanders D, Zhu Z, Ciampi R, Roh M, Shedden K, *et al*: Molecular classification of papillary thyroid carcinoma: Distinct BRAF, RAS, and RET/PTC mutation-specific gene expression profiles discovered by DNA microarray analysis. Oncogene 24: 6646-6656, 2005.
14. Borup R, Rossing M, Henao R, Yamamoto Y, Krogdahl A, Godballe C, Winther O, Kiss K, Christensen L, Høgdall E and Nielsen FC: Molecular signatures of thyroid follicular neoplasia. Endocr Relat Cancer 17: 691-708, 2010.
15. Issaq HJ, Waybright TJ and Veenstra TD: Cancer biomarker discovery: opportunities and pitfalls in analytical methods. Electrophoresis 32: 967-975, 2011.

16. Drucker E and Krapfenbauer K: Pitfalls and limitations in translation from biomarker discovery to clinical utility in predictive and personalised medicine. EPMA J 4: 7, 2013.
17. Walsh PS, Wilde JI, Tom EY, Reynolds JD, Chen DC, Chudova DI, Pagan M, Pankratz DG, Wong M, Veitch J, *et al*: Analytical performance verification of a molecular diagnostic for cytology-indeterminate thyroid nodules. J Clin Endocrinol Metab 97: E2297-E2306, 2012.
18. Duick DS, Klopper JP, Diggans JC, Friedman L, Kennedy GC, Lanman RB and McIver B: The impact of benign gene expression classifier test results on the endocrinologist-patient decision to operate on patients with thyroid nodules with indeterminate fine-needle aspiration cytopathology. Thyroid 22: 996-1001, 2012.
19. Ali SZ, Fish SA, Lanman R, Randolph GW and Sosa JA: Use of the Afirma® gene expression classifier for preoperative identification of benign thyroid nodules with indeterminate fine needle aspiration cytopathology. PLoS Curr 5: 5, 2013.
20. Alexander EK, Schorr M, Klopper J, Kim C, Sipos J, Nabhan F, Parker C, Steward DL, Mandel SJ and Haugen BR: Multicenter clinical experience with the Afirma gene expression classifier. J Clin Endocrinol Metab 99: 119-125, 2014.
21. Ward LS and Kloos RT: Molecular markers in the diagnosis of thyroid nodules. Arq Bras Endocrinol Metabol 57: 89-97, 2013.
22. Harrell RM and Bimston DN: Surgical utility of Afirma: effects of high cancer prevalence and oncocytic cell types in patients with indeterminate thyroid cytology. Endocr Pract 20: 364-369, 2014.
23. Vriens D, Adang EM, Netea-Maier RT, Smit JW, de Wilt JH, Oyen WJ and de Geus-Oei LF: Cost-effectiveness of FDG-PET/CT for cytologically indeterminate thyroid nodules: a decision analytic approach. J Clin Endocrinol Metab 99: 3263-3274, 2014.
24. McIver B, Castro MR, Morris JC, Bernet V, Smallridge R, Henry M, Kosok L and Reddi H: An independent study of a gene expression classifier (Afirma) in the evaluation of cytologically indeterminate thyroid nodules. J Clin Endocrinol Metab 99: 4069-4077, 2014.
25. Bustin SA: Absolute quantification of mRNA using real-time reverse transcription polymerase chain reaction assays. J Mol Endocrinol 25: 169-193, 2000.
26. Grate LR: Many accurate small-discriminatory feature subsets exist in microarray transcript data: biomarker discovery. BMC Bioinformatics 6: 97, 2005.
27. Trevino V and Falciani F: GALGO: an R package for multivariate variable selection using genetic algorithms. Bioinformatics 22: 1154-1156, 2006.
28. Dabney AR: Classification of microarrays to nearest centroids. Bioinformatics 21: 4148-4154, 2005.
29. Tomei S, Marchetti I, Zavaglia K, Lessi F, Apollo A, Aretini P, Di Coscio G, Bevilacqua G and Mazzanti C: A molecular computational model improves the preoperative diagnosis of thyroid nodules. BMC Cancer 12: 396, 2012.
30. Meng H, Murrelle EL and Li G: Identification of a small optimal subset of CpG sites as bio-markers from high-throughput DNA methylation profiles. BMC Bioinformatics 9: 457, 2008.
31. Price ND, Trent J, El-Naggar AK, Cogdell D, Taylor E, Hunt KK, Pollock RE, Hood L, Shmulevich I and Zhang W: Highly accurate two-gene classifier for differentiating gastrointestinal stromal tumors and leiomyosarcomas. Proc Natl Acad Sci USA 104: 3414-3419, 2007.
32. Wang X and Gotoh O: Accurate molecular classification of cancer using simple rules. BMC Med Genomics 2: 64, 2009.
33. Fluge Ø, Bruland O, Akslen LA, Lillehaug JR and Varhaug JE: Gene expression in poorly differentiated papillary thyroid carcinomas. Thyroid 16: 161-175, 2006.
34. Prazeres H, Torres J, Rodrigues F, Pinto M, Pastoriza MC, Gomes D, Cameselle-Teijeiro J, Vidal A, Martins TC, Sobrinho-Simões M and Soares P: Chromosomal, epigenetic and microRNA-mediated inactivation of LRP1B, a modulator of the extracellular environment of thyroid cancer cells. Oncogene 30: 1302-1317, 2011.
35. Kim HS, Kim H, Kim JY, Jeoung NH, Lee IK, Bong JG and Jung ED: Microarray analysis of papillary thyroid cancers in Korean. Korean J Intern Med 25: 399-407, 2010.
36. Kenney MC, Chwa M, Atilano SR, Tran A, Carballo M, Saghizadeh M, Vasiliou V, Adachi W and Brown DJ: Increased levels of catalase and cathepsin V/L2 but decreased TIMP-1 in keratoconus corneas: evidence that oxidative stress plays a role in this disorder. Invest Ophthalmol Vis Sci 46: 823-832, 2005.
37. Hawthorn L, Stein L, Varma R, Wiseman S, Loree T and Tan D: TIMP1 and SERPIN-A overexpression and TFF3 and CRABP1 underexpression as biomarkers for papillary thyroid carcinoma. Head Neck 26: 1069-1083, 2004.
38. Kebebew E, Peng M, Reiff E, Duh QY, Clark OH and McMillan A: Diagnostic and prognostic value of angiogenesis-modulating genes in malignant thyroid neoplasms. Surgery 138: 1102-1110, 2005.
39. Fu M, Wang C, Li Z, Sakamaki T and Pestell RG: Minireview: Cyclin D1: normal and abnormal functions. Endocrinology 145: 5439-5447, 2004.
40. Seybt TP, Ramalingam P, Huang J, Looney SW and Reid MD: Cyclin D1 expression in benign and differentiated malignant tumors of the thyroid gland: diagnostic and biologic implications. Appl Immunohistochem Mol Morphol 20: 124-130, 2012.
41. Cagnoni G and Tamagnone L: Semaphorin receptors meet receptor tyrosine kinases on the way of tumor progression. Oncogene 33: 4795-4802, 2014.
42. Kigel B, Varshavsky A, Kessler O and Neufeld G: Successful inhibition of tumor development by specific class-3 semaphorins is associated with expression of appropriate semaphorin receptors by tumor cells. PLoS One 3: e3287, 2008.
43. Vriens MR, Moses W, Weng J, Peng M, Griffin A, Bleyer A, Pollock BH, Indelicato DJ, Hwang J and Kebebew E: Clinical and molecular features of papillary thyroid cancer in adolescents and young adults. Cancer 117: 259-267, 2011.
44. Nillni EA, Xie W, Mulcahy L, Sanchez VC and Wetsel WC: Deficiencies in pro-thyrotropin-releasing hormone processing and abnormalities in thermoregulation in Cpefat/fat mice. J Biol Chem 277: 48587-48595, 2002.
45. Murthy SR, Pacak K and Loh YP: Carboxypeptidase E: elevated expression correlated with tumor growth and metastasis in pheochromocytomas and other cancers. Cell Mol Neurobiol 30: 1377-1381, 2010.
46. Liguori L, Andolfo I, de Antonellis P, Aglio V, di Dato V, Marino N, Orlotti NI, De Martino D, Capasso M, Petrosino G, *et al*: The metallophosphodiesterase Mpped2 impairs tumorigenesis in neuroblastoma. Cell Cycle 11: 569-581, 2012.
47. Sharma MK, Watson MA, Lyman M, Perry A, Aldape KD, Deák F and Gutmann DH: Matrilin-2 expression distinguishes clinically relevant subsets of pilocytic astrocytoma. Neurology 66: 127-130, 2006.
48. Long W, Foulds CE, Qin J, Liu J, Ding C, Lonard DM, Solis LM, Wistuba II, Qin J, Tsai SY, *et al*: ERK3 signals through SRC-3 coactivator to promote human lung cancer cell invasion. J Clin Invest 122: 1869-1880, 2012.
49. Medici M, Porcu E, Pistis G, Teumer A, Brown SJ, Jensen RA, Rawal R, Roef GL, Plantinga TS, Vermeulen SH, *et al*: Identification of novel genetic Loci associated with thyroid peroxidase antibodies and clinical thyroid disease. PLoS Genet 10: e1004123, 2014.
50. Yang JJ, Cheng C, Devidas M, Cao X, Campana D, Yang W, Fan Y, Neale G, Cox N, Scheet P, *et al*: Genome-wide association study identifies germline polymorphisms associated with relapse of childhood acute lymphoblastic leukemia. Blood 120: 4197-4204, 2012.
51. Tomás G, Tarabichi M, Gacquer D, Hébrant A, Dom G, Dumont JE, Keutgen X, Fahey TJ III, Maenhaut C and Detours V: A general method to derive robust organ-specific gene expression-based differentiation indices: application to thyroid cancer diagnostic. Oncogene 31: 4490-4498, 2012.
52. Dom G, Tarabichi M, Unger K, Thomas G, Oczko-Wojciechowska M, Bogdanova T, Jarzab B, Dumont JE, Detours V and Maenhaut C: A gene expression signature distinguishes normal tissues of sporadic and radiation-induced papillary thyroid carcinomas. Br J Cancer 107: 994-1000, 2012.
53. Treviño Alvarado VM and Falciani F: Edinburgh DTo and Mexico CNdCyTC: Identifying the molecular components that matter: a statistical modelling approach to linking functional genomics data to cell physiology. PhD thesis. School of Biosciences, University of Birmingham, England, 2007.
54. Anglani R, Creanza TM, Liuzzi VC, Piepoli A, Panza A, Andriulli A and Ancona N: Loss of connectivity in cancer co-expression networks. PLoS One 9: e87075, 2014.
55. Prasad NB, Somervell H, Tufano RP, Dackiw AP, Marohn MR, Califano JA, Wang Y, Westra WH, Clark DP, Umbricht CB, *et al*: Identification of genes differentially expressed in benign versus malignant thyroid tumors. Clin Cancer Res 14: 3327-3337, 2008.
56. Tibshirani R, Hastie T, Narasimhan B and Chu G: Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proc Natl Acad Sci USA 99: 6567-6572, 2002.
57. Guyon I, Weston J, Barnhill S and Vapnik V: Gene selection for cancer classification using support vector machines. Mach Learn 46: 389-422, 2002.