Ann Surg Oncol (2009) 16:487–498
DOI 10.1245/s10434-008-0251-1

Annals of
SURGICAL ONCOLOGY
OFFICIAL JOURNAL OF THE SOCIETY OF SURGICAL ONCOLOGY

ORIGINAL ARTICLE – ENDOCRINE TUMORS

Principal Component Analysis, Hierarchical Clustering, and Decision Tree Assessment of Plasma mRNA and Hormone Levels as an Early Detection Strategy for Small Intestinal Neuroendocrine (Carcinoid) Tumors

Irvin M. Modlin, MD, PhD, DSc, FRCS (Eng & Ed)[1], Björn I. Gustafsson[1,2,3], Ignat Drozdov[1], Boaz Nadler[4], Roswitha Pfragner[5], and Mark Kidd[1]

[1]Department of Gastroenterological Surgery, Yale University School of Medicine, 333 Cedar Street, P.O. Box 208062, New Haven, CT 06520-8062, USA; [2]St. Olav's Hospital HF, Trondheim University Hospital, Trondheim, Norway; [3]Department of Cancer Research and Molecular Medicine, Faculty of Medicine, Norwegian University of Science and Technology, N-7006 Trondheim, Norway; [4]Department of Computer Science and Applied Mathematics, Weizmann Institute of Science, Rehovot 76100, Israel; [5]Institute of Pathophysiology and Immunology, Centre for Molecular Medicine, Medical University of Graz, Graz, Austria

ABSTRACT Incidence of neuroendocrine tumors (NETs) is increasing (approximately 6%/year), but clinical presentation is nonspecific, resulting in delays in diagnosis (5–7 years; approximately 70% have metastases). This reflects absence of a sensitive plasma marker. The aim of this study is to investigate whether detection of circulating messenger RNA (mRNA) alone or in combination with circulating NET-related hormones and growth factors can detect gastrointestinal NET disease. The small intestinal (SI) NET cell line KRJ-I was used to define the sensitivity of real-time polymerase chain reaction (PCR) for mRNA detection in blood. *NSE*, *Tph-1*, and $VMAT_2$ transcripts were identified from one KRJ-I cell/ml blood. mRNA from the tissue and plasma of SI-NETs ($n = 12$) and gastric NETs ($n = 7$), and plasma from healthy controls ($n = 9$) was isolated and real-time PCR performed. *Tph-1* was a specific marker of SI-NETs (58%, $p < 0.03$) whereas *CgA* transcripts did not differentiate tumors from controls. Patients with metastatic disease expressed more marker transcripts than localized tumors (75% versus 18%, $p < 0.02$). Plasma 5-hydroxytryptamine (5-HT), chromogranin A (CgA), ghrelin, and connective tissue growth factor (CTGF) fragments were measured, combined with mRNA levels, and a predictive mathematical model for NET diagnosis developed using decision trees. The sensitivity and specificity to diagnose SI-NETs and gastric NETs were 81.2% and 100%, and 71.4% and 55.6%, respectively. We conclude that mRNA from one NET cell/ml blood can be detected. Circulating plasma *Tph-1* is a promising marker gene for SI-NET disease (specificity 100%) while an increased number of marker transcripts (>2) correlated with disease spread. Including NET-related circulating hormones and growth factors in the algorithm increased the sensitivity of detection of SI-NETs from 58 to 82%.

First Received: 4 June 2008;
Published Online: 3 December 2008

I. M. Modlin, MD, PhD, DSc, FRCS (Eng & Ed)
e-mail: imodlin@optonline.net

In the last 30 years, incidence of neuroendocrine tumor (NET) disease has increased dramatically. In the USA, incidence of NET was 6/100,000 in 2004.[1] The only curative treatment option for NETs is surgery, and outcome is critically dependent on the extent of spread at the time of diagnosis.[2] As the primary tumor usually is very small and asymptomatic, diagnosis is typically delayed many years and symptoms only become apparent once the tumor has spread to regional lymph nodes or has metastasized to the liver.[2] Epidemiologic studies clearly demonstrate that small intestinal (SI) NET 5-year survival rates drop from 74% for local disease to 40% with metastatic disease.[2] The requirement of a plasma test for surveillance and early diagnosis is thus obvious since 67% of SI-NET patients are first diagnosed when their disease is metastatic.[1]

Serum tumor markers such as alpha-fetoprotein (AFP) in liver cancer, carcinoembryonic antigen (CEA) in colorectal cancer, and prostate-specific antigen (PSA) in

prostate cancer have proven useful in clinical practice.[3] Unfortunately, due to the expression of these markers in benign conditions, their sensitivity and specificity are not satisfactory. Over the last few years, several techniques to detect circulating cancer cells or cancer-derived nucleic acids in peripheral blood have been developed. Using reverse transcriptase (RT) PCR towards tumor-specific genes overexpressed in neoplasia, circulating melanoma cells, breast, prostatic, colorectal, and gastric cancer cells have been detected in blood.[4–8] The presence and number of circulating tumor cells (CTC) is known to correlate with tumor size, progression, extent of metastases, and overall survival.[9] The rate at which cells are shed from tumor tissue appears to be relatively constant, and CTC counts at different time points from the same patient during a 24-h period yield similar results.[10] Differences in the vascularization of tumors as well as the sites of the primary and its metastases are factors that may affect the number of CTCs. It has been theorized that the relatively small number of CTCs seen in gastrointestinal carcinomas (approximately 4 CTC/7.5 mL blood) may be due to filtration via the portal circulation.[11]

Circulating DNA exhibits tumor-related alterations including point mutations, DNA hypermethylation, microsatellite instabilities, and loss of heterozygosity.[12] These are typically identical to alterations in the primary tumor tissue, and can thus be used when the primary tumor is not available for biopsy. Circulating tumor-derived RNA was first reported in patients with nasopharyngeal carcinoma and melanoma.[13,14] Later, a number of tumor-derived RNAs were detected in other cancer types including breast cancer, colorectal cancer, follicular lymphoma, and hepatocellular carcinoma.[15] Detection of circulating RNA has advantages over DNA as RNA is more tumor specific, although it can be less stable.[16] Circulating RNA is, however, often protected from degradation by inclusion within apoptotic bodies.[17]

Depending on the primary tumor location and the cell of origin, gastrointestinal (GI) NETs produce a variety of different amine and peptide hormones. Even though gastric enterochromaffin-like (ECL) tumors typically produce histamine and the small intestinal EC cell-derived tumors, serotonin, NETs may exhibit a variety of secretory products including gastrin, ghrelin, pancreatic polypeptide, and substance P.[2,18] In addition, chromogranin A (CgA), a water-soluble acidic glycoprotein stored in the secretory granules of neuroendocrine cells, can be detected in plasma and is widely utilized as a general NET marker.[19] Apart from hormonal agents, NETs produce tumor growth factor-β (TGF-β) and connective tissue growth factor (CTGF), which are both detectable in the circulation.[20] Nevertheless, none of these currently available markers are sensitive enough to detect early NET disease, and specificity is generally low.

NETs are highly vascularized tumors which may metastasize even when the primary tumor is <1 cm.[21] This propensity to metastasize indicates that tumors are shedding viable cells at a relatively high frequency. In this study, we investigated whether mRNA for specific NET-related genes could be detected by real-time PCR in blood spiked with the small intestinal neoplastic EC cell line, KRJ-I.[22] We further evaluated the sensitivity and specificity of this PCR technique for detecting circulating tumor-derived mRNA in plasma from patients with known NET disease. Thereafter, we examined serum levels of a variety of NET-related secretory products, and determined whether these in combination with circulating tumor-derived mRNA could be used in a predictive algorithm to facilitate the diagnosis of NETs.

MATERIAL AND METHODS

Potential NET Marker Genes

Initially, we identified potential NET marker genes by screening GI-NET and KRJ-I transcriptome libraries.[23,24] Specifically, genes with moderate to high expression in nine primary SI-NETs and in the KRJ-1 cell line, compared with expression in normal small intestinal mucosa, were identified.[23,24] *CgA* and *Tph-1* were highly upregulated in both SI-NETs and KRJ-1, while $VMAT_1$ and *NSE* were upregulated in SI-NETs but not in KRJ-1. $VMAT_2$ and *DDC* were upregulated in KRJ-1 but not in the SI-NET transcriptome. These six genes were considered candidate NET marker genes; their expression was then evaluated in the blood spike-in and patient tissue and blood samples (institutional review board (IRB)-approved protocol HIC12589).

Blood and Tissue Samples

Five milliliters of blood from healthy donors ($n = 9$; mean age 45 years, range 32–68 years; M:F = 5:4) was collected in ethylenediamine tetraacetic acid (EDTA) tubes. Plasma was separated from the buffy coat following two spin cycles (5 min at 2,000 rpm), and stored at −80°C until nucleic acid isolation or hormone/growth factor analysis. A separate tube with 2 ml whole blood was frozen for spike-in analyses. Thereafter, we analyzed nucleic acid and hormone/growth factor levels in plasma samples from 12 patients with verified SI-NET disease (mean age 56 years, range 41–74 years; M:F = 6:6), and 7 patients with gastric NETs (mean age 60 years, range 52–85 years; M:F = 4:3). These samples were obtained from our blood databank of patients who have been treated for NET disease at Yale New Haven Hospital.[20] In

the biobank, tumor tissue samples were available from ten SI-NET patients and five gastric NETs. The following clinical data and staging information was available for the patients with SI-NETs: two patients with localized disease, two patients with invasive tumors and lymph node metastases (1/18 and 8/37 positive) but no hepatic metastases, and eight patients with liver involvement. For gastric NETs: six patients had type I tumors and one patient had a type II gastric NET. All gastric NETs were localized.

Nucleic Acid Isolation

Two protocols for isolation of RNA from plasma were compared: (1) a Trizol approach followed by RNA clean-up and (2) the QIAamp RNA Blood Mini Kit (QIAGEN). For buffy coat RNA isolation, the Trizol approach was used followed by RNA clean-up. RNA was dissolved in diethyl pyrocarbonate water and measured spectrophotometrically, and an aliquot analyzed by using Bioanalyzer (Agilent Technologies, Palo Alto, CA) to assess the quality of the RNA.[23] For tumor tissue ($n = 15$ samples), mRNA was isolated using the Trizol approach.[23]

cDNA Synthesis

Total RNA from each sample was subjected to reverse transcription with High-Capacity cDNA Archive Kit (ABI, Foster City, CA) following the manufacturer's directions. Briefly, 2 μg total RNA in 50 μL of water was mixed with 50 μL 2× RT mix containing Reverse Transcription Buffer, deoxynucleotide triphosphate solution, random primers, and Multiscribe Reverse Transcriptase. The RT reaction was performed in a thermal cycler for 10 min at 25°C followed by 120 min at 37°C.[23]

Real-Time PCR

Expression of three NET housekeeping genes (*ALG9*, *TFCP2*, and *ZNF410*) was measured in mRNA isolated from buffy coat with the Trizol approach, and from plasma with the QIAamp RNA Blood Mini Kit or from the tumor tissue, and message for *Tph-1*, *CgA*, *DDC*, *NSE*, $VMAT_1$, and $VMAT_2$ were investigated using real-time PCR by ABI 7900 Sequence Detection System (Applied Biosystems, Foster City, CA).[25] Briefly, complementary DNA in 7.2 μL water was mixed with 0.8 μL of 20·Assays-on-Demand primer, and probe mix and 8 μL of 2× TaqMan Universal Master mix in a 384-well optical reaction plate. The following PCR conditions were used: 50°C for 2 min and then 95°C for 10 min, followed by 50 cycles at 95°C for 15 min and 60° for 1 min.[23]

Hormone and Growth Factor Analyses

Enzyme-linked immunosorbent assays (ELISA) were used to measure serum 5-HT (Serotonin EIA, Rocky Mountain Diagnostics) and CgA (DAKO A/S, Glostrup, Denmark) as described.[22,26] Ghrelin-like immunoreactivity measurements were undertaken by Drs. S. Bloom and M. Ghatei (Imperial College London), while CTGF(W), and its N- and total fragments were measured by Dr. William Usinger (Fibrogen Inc.).[27,28]

Data Analysis

PCR ΔC_T levels of housekeeping genes were compared between each isolation protocol (plasma – Trizol versus plasma – QIAamp) to assess the most effective mRNA isolation procedure.

The ΔC_T values for each target gene were determined for each sample from each sample isolation procedure (whole blood, buffy coat, plasma – Trizol, plasma – QIAamp) and plotted (Y-axis $= \Delta C_T$, X-axis $=$ gene). Descriptive statistical analysis was undertaken [mean/standard deviation (SD) or median/range]. Genes with high expression ($\Delta C_T > 40$ cycles) were considered as negative.

Differences in plasma gene expression between groups (SI-NETs, gastric NETs, and healthy controls) were calculated using Fisher's exact (two-tailed) test; $p < 0.05$ was considered significantly different.

Hormone and Growth Factor Levels Levels of each hormone were compared between the three groups (SI-NETs, gastric NETs, and healthy controls) using a two-tailed t-test; $p < 0.05$ was considered significant.

Predictive Model Generation Raw data (mRNA transcript and serum hormone levels) were $\log_{10}$-transformed and imported into Partek® Genomic Suite for analysis.[29] Patterns in data were visualized using principal component analysis (PCA) and hierarchical clustering (HC). Decision trees (DT) with a pruning algorithm (parts not implicated in the decision process are excluded from the tree) were used to predict SI-NETs, gastric NETs, and healthy controls.[30]

Principal component analysis is an exploratory technique that is used to describe the structure of high-dimensional data by reducing its dimensionality into uncorrelated principal components (PCs) that explain most variation in the data.[31] PCA mapping was visualized in a three-dimensional space where the X-, Y-, and Z-axis represent first, second, and third PCs, respectively.

Hierarchical clustering is used to group similar objects into clusters; the algorithm produces a dendrogram representing the similarity between clusters. The methodology for this technique was described previously.[1] In our

analysis, a Euclidean distance metric was implemented with an average linkage clustering algorithm.

Decision trees are predictive models that map observations about an item to a conclusion about its target value.[32] We utilized a pruning algorithm to select the most relevant features for this classification model.[30] In these tree structures, leaves represent classifications and branches represent conjunctions of features that lead to those classifications. A ten-fold cross-validation was used to measure the efficiency of this technique.

RESULTS

Evaluation of mRNA Isolation Protocols

The expression of the three housekeeping genes *ALG9*, *TFCP2*, and *ZNF410* was determined in mRNA that had been isolated using the Trizol approach from buffy coat of five healthy donors. All three genes were detected in all samples with ΔC_T levels between 30 and 35. When comparing the isolation techniques for plasma mRNA, isolation of mRNA with the QIAamp RNA Blood Mini Kit was significantly better compared with the Trizol approach (Fig. 1), and was thus used for isolation of mRNA from NET patient plasma samples.

Identification of NET Cell Gene Expression (Selected Markers) in Buffy Coat: Spike-In Assay

NET gene expression of selected marker genes was examined by real-time PCR in mRNA isolated from buffy coat extracted from blood spiked with 0, 1, 10 or 100 KRJ-I cells. Using ΔC_T as a cutoff, potential markers were *DDC*, *NSE*, *Tph-1*, and $VMAT_2$, and all these genes had lower ΔC_T values after spiking in one KRJ-1 cell (Fig. 2).

Detection of NET-Related Genes in Plasma Samples and Comparison with Tissue Expression

Real-time PCR for *Tph-1*, *CgA*, *DDC*, *NSE*, $VMAT_1$, and $VMAT_2$ in tissue and plasma of SI-NET and gastric NET patients identified that *Tph-1* was positive in 67% of all NET tissues and 37% of all plasma samples, *CgA* in 100% and 63%; *DDC* in 80% and 21%; *NSE* in 100% and 5%; $VMAT_1$ in 67% and 6%; and $VMAT_2$ in 33% and 0%, respectively (Fig. 3). Overall concordance of tissue expression and plasma identification of marker gene transcripts was 56% for *Tph-1*, 65% for *CgA*, 27% for *DDC*, 5% for *NSE*, 15% for $VMTA_1$, and 0% for $VMAT_2$. Within plasma, *Tph-1* was identified in 7/12 (58%) SI-NETs, 0/7 (0%) gastric NETs, and 0/9 (0%) controls, *CgA* in 8/12 (67%), 2/9 (22%), and 4/7 (57%), *DDC* in 4/12 (33%), 1/9 (11%), and 0/7 (0%), *NSE* in 1/12 (8.3%), 1/9 (11%), and 0/7 (0%), and $VMAT_1$ in 1/12 (8.3%), 1/9 (11%), and 2/7 (29%), respectively (Fig. 4). $VMAT_2$ was negative in all samples. Statistical analyses confirmed that *Tph-1* transcripts were identified more often in the plasma of SI-NETs than either gastric NETs ($p = 0.025$) and healthy controls ($p = 0.0075$). In contrast, the expression of *CgA* transcripts were not statistically different between SI-NETs and healthy controls ($p = 0.056$).

An examination of the number of positive transcripts in patients with localized and regional disease ($n = 7$ gastric NETs, $n = 4$ SI-NETs) compared with those with disseminated disease ($n = 8$ SI-NETs) identified that patients with localized disease were more likely to be positive for none or one transcripts (82%) compared with distantly staged disease, which were likely to be positive for two or more transcripts (75%, $p = 0.018$, unpaired *t*-test). Subanalysis of the SI-NET group identified that significantly more positive transcripts were identified in patients with

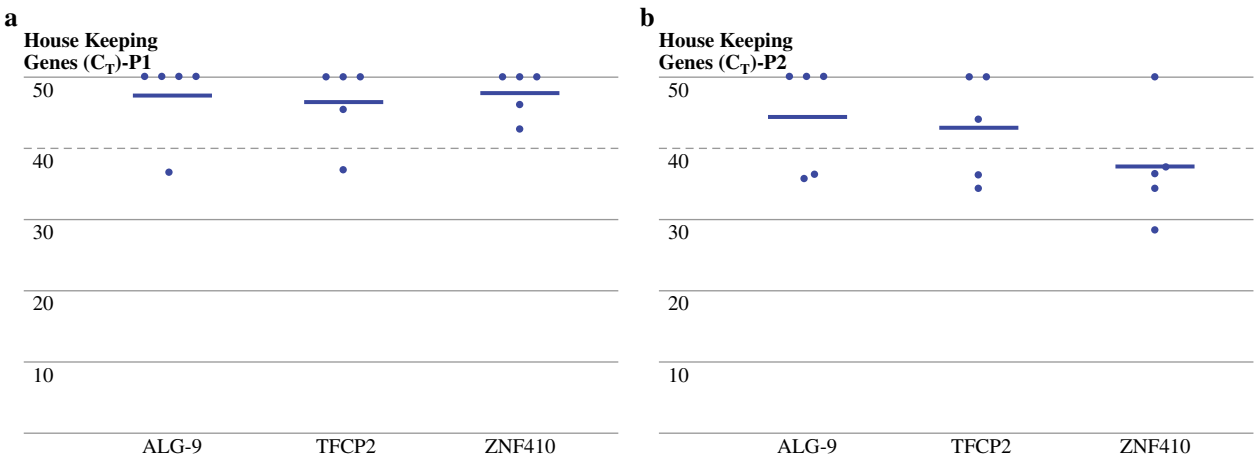

FIG. 1 Housekeeping genes identified in plasma after Trizol mRNA isolation (**a**), or the QIAamp RNA Blood Mini Kit approach (**b**). Significantly more housekeeping genes were identified (8/15 versus 2/15, $p = 0.05$) after isolation with protocol B

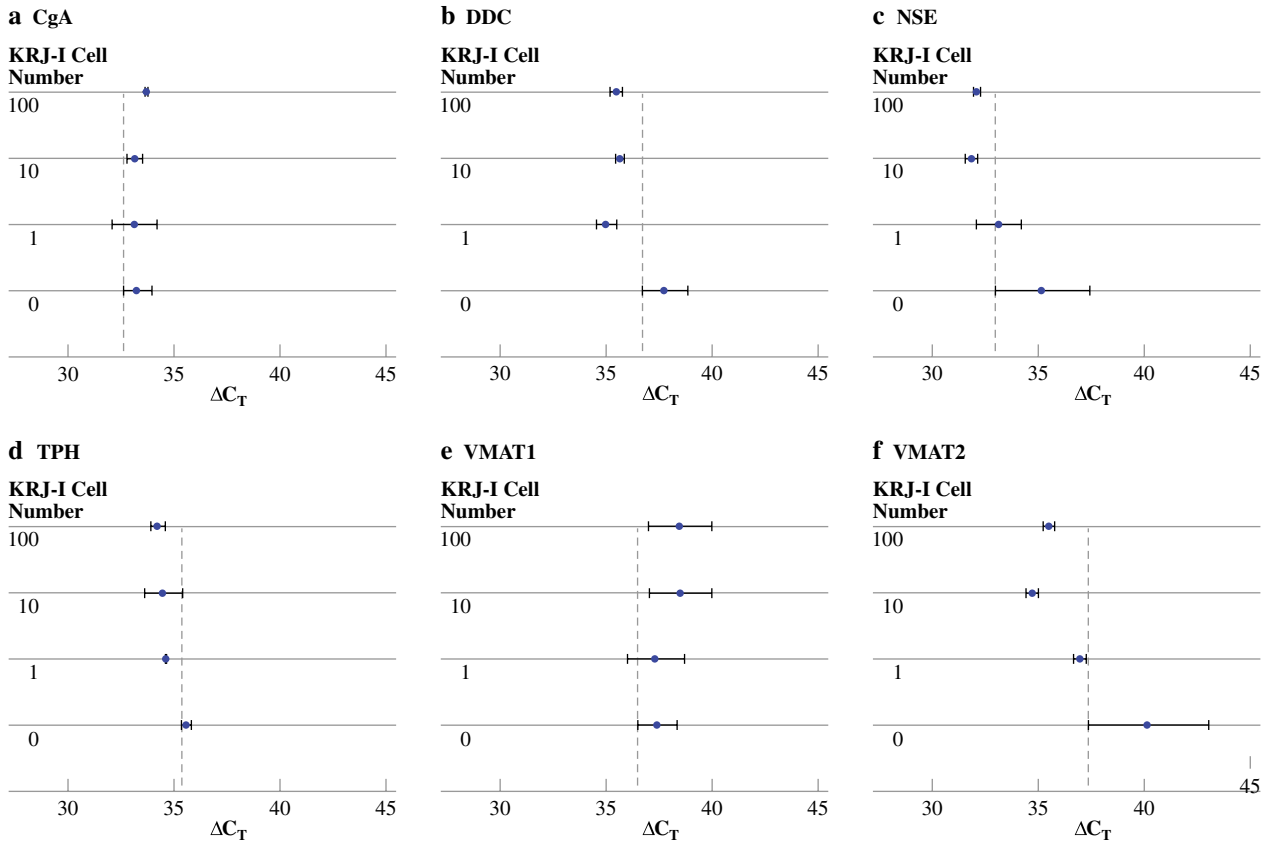

FIG. 2 Potential target genes identified in blood spiked in with 0, 1, 10 or 100 KRJ-1 cells. Using ΔC_T as a cutoff, potential markers are *DDC*, *NSE*, *Tph-1*, and *VMAT*$_2$

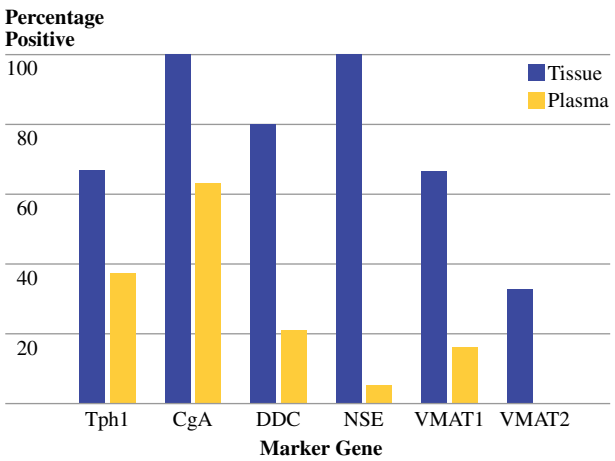

FIG. 3 Real-time PCR for NET marker mRNA in tumor tissue ($n = 15$) and plasma ($n = 19$) of SI-NETs and gastric NETs. Transcript expression of *Tph-1* was identified in 67% of all NET tissues and 37% of all plasma samples, *CgA* in 100% and 63%, *DDC* in 80% and 21%, *NSE* in 100% and 5%, *VMAT1* in 67% and 6%, and *VMTA2* in 33% and 0%, respectively

distant disease compared with those with localized/regional disease (75% vs. 25%, $p = 0.0005$, Yates corrected chi-square test). This was reflected in an increase in the number of *CgA*-positive samples from 50% to 75% and in the other four neuroendocrine genes (from 12.5% to 62.5%, $p = 0.011$).

Detection of NET-Related Hormones in Plasma Samples

ELISAs for 5-HT, CgA, ghrelin, and three forms of CTGF in plasma of healthy controls, and SI-NET and gastric NET patients identified that 5-HT and CgA were specific marker for SI-NETs versus both controls ($p < 0.0002$) and gastric NETs ($p < 0.04$) and that ghrelin was also positive in SI-NETs versus controls ($p = 0.03$). CTGF-W was elevated in both SI and gastric NETs compared with controls ($p < 0.04$) whereas CTGF N+W was elevated only in gastric NETs ($p = 0.005$) (Fig. 5).

Development of a SI-NET Predictive Model Based on Plasma Nucleic Acid and Circulating Hormone and CTGF Levels

To determine whether marker expression in blood samples from healthy controls, G-NETs or SI-NETs could

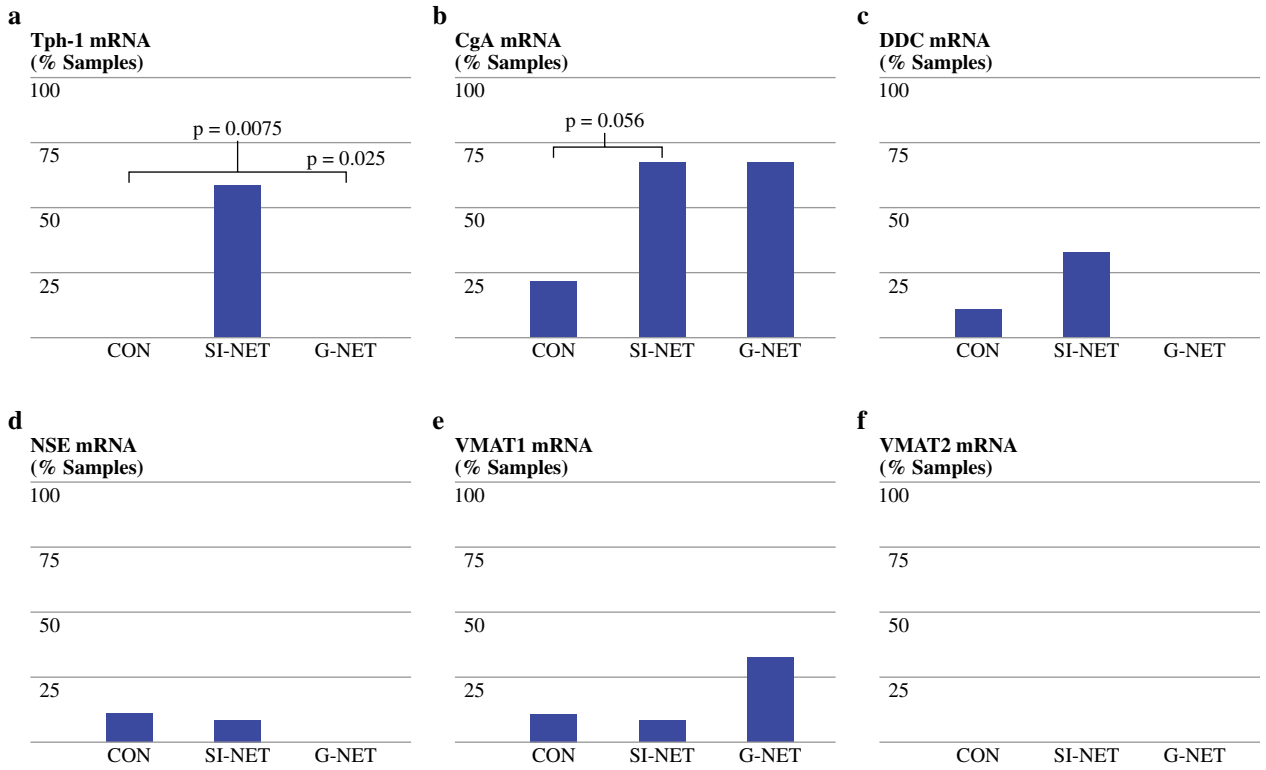

FIG. 4 Real-time PCR for NET marker mRNA in plasma of healthy controls, SI-NETs, and gastric NETs. Transcript expression of *Tph-1* was statistically significant in SI-NETs compared with either gastric NETs ($p < 0.05$) or healthy controls (**a**). A trend toward significance ($p = 0.056$) was noted for *CgA* in SI-NETs compared with healthy controls, but no significant differences were noted for *DDC*, *NSE*, $VMAT_1$, and $VMAT_2$. CON, healthy controls; SI-NETs, small intestinal NETs; G-NET, gastric NETs

be differentiated, PCA was used to reduce presence/absence of mRNA (*Tph-1*, *CgA*, *DDC*, *NSE*, $VMAT_1$, and $VMAT_2$) as well as serum levels of 5-HT, CgA, Ghrelin, CTGF-W, CTGF-N+W, and CTGF(T) into three principal components that capture the variance in the data set (Fig. 6a). Thirty-five percent of the variance was represented by PC#1, 14.7% by PC#2, and 12.9% by PC#3 with an overall 63.4% represented by all three PCs. The distance between the centroids (i.e., the centers of mass) for each sample subtype (healthy control, gastric NET, and SI-NET) in the three dimensions represents the relative similarity of their regulatory signatures. As assessed by PCA, the gastric NETs, SI-NETs, and healthy controls have distinct regulatory signatures.

Hierarchical clustering of healthy controls, gastric NETs, and SI-NETs considering only the levels of 5-HT, CgA, ghrelin, CTGF-W, CTGF-N+W, and CTGF(T) revealed a similar pattern that confirmed the PCA analysis (Fig. 6b). Using this approach, SI-NETs could be differentiated from other samples while gastric NETs were associated with the healthy control group.

Considering that molecular-level differentiations between gastric NETs, SI-NETs, and healthy controls exist, as demonstrated by the PCA and hierarchical clustering, a decision tree learning algorithm was implemented primed in order: (1) to construct a predictive model, and (2) to select variables which greatly increase performance of the classifier. Using only presence or absence of the circulating mRNA (*Tph-1*, *CgA*, *DDC*, *NSE*, $VMAT_1$, and $VMAT_2$), a decision tree was constructed with *Tph-1* and *CgA* at the nodes as determined by the pruning algorithm (Fig. 7a). This method predicted 77.8% of the healthy controls, 63.6% of the SI-NETs, and 71.4% of the gastric NETs. SI-NETs were predicted with 100% precision (Table 1). When gastric NETs were excluded from the classifier, only *Tph-1* mRNA was identified as a regulator of the decision process (Fig. 7b). Under these conditions, the sensitivity and specificity for detecting SI-NETs remained the same, but specificity for predicting healthy tissue increased from 63.6% to 69.2% and sensitivity from 77.8% to 100% (Table 2). From the serum levels of 5-HT, CgA, Ghrelin, CTGF-W, CTGF-N+W, and CTGF(T), only 5-HT was selected by the pruning algorithm (Fig. 7c). This model was successful in validating 88.9% of the healthy control samples, 81.8% of the SI-NETs, and 71.4% of the gastric NETs. SI-NETs were predicted with 100% precision (Table 3). From the merger of the two models (Fig. 7d), the pruning algorithm selected serum levels of 5-HT and *CgA* mRNA. Using these criteria, blood samples from healthy controls were validated with 77.8% accuracy,

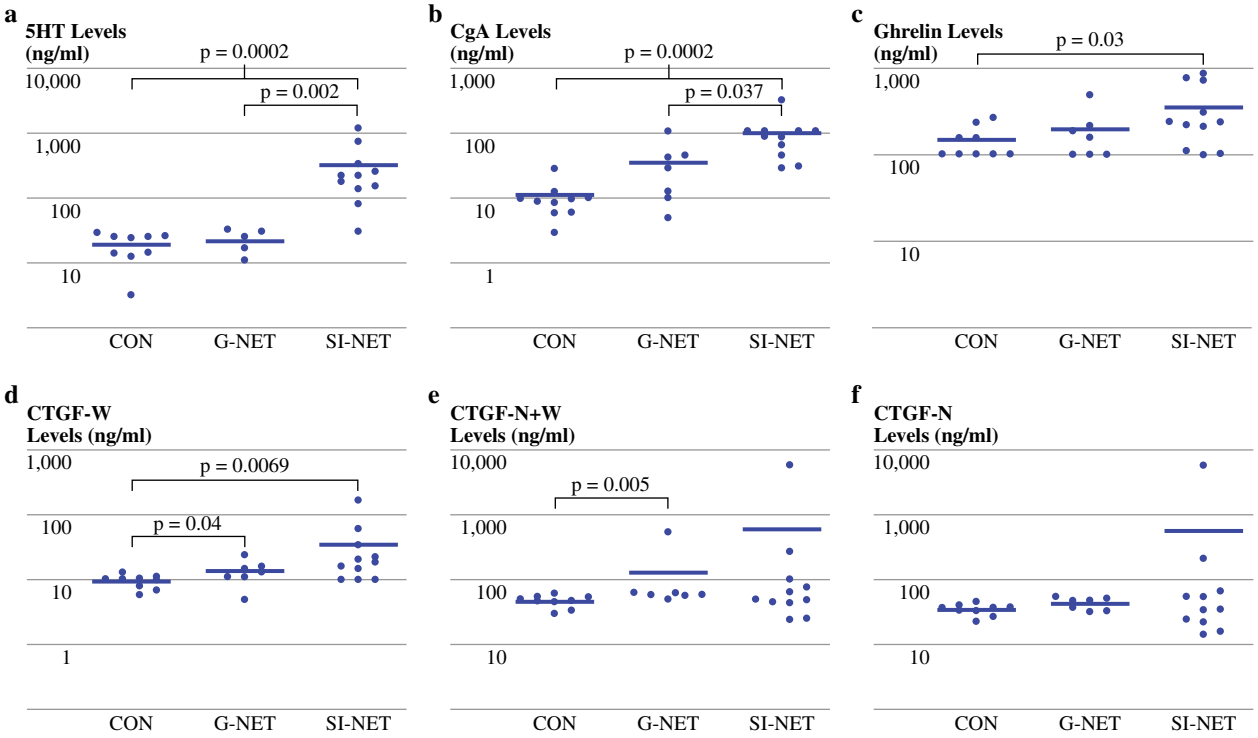

FIG. 5 NET marker hormone levels in plasma of healthy controls, SI-NETs, and gastric NETs. Expression of 5-HT and CgA was statistically significant in SI-NETs compared to either gastric NETs or healthy controls (A/B). Ghrelin and CTGF-W were also significantly elevated in SI-NETs compared with healthy controls, while CTGF-W and CTGF-N+W were elevated in gastric NETs compared with healthy controls. No significant differences were noted for CTGF-N. CON, healthy control; G-NET, gastric NET; SI_NET, small intestinal NET

SI-NETs with 81.8% accuracy, and gastric NETs with 71.4% accuracy. The algorithm was 100% precise in predicting SI-NETs (Table 4).

DISCUSSION

In this study, we demonstrated that real-time PCR is a sensitive approach for the detection of circulating NET-derived mRNA in blood, particularly in SI-NETs. Among candidate genes, measurements of *Tph-1* mRNA in plasma were 100% specific and had 58% sensitivity for detecting SI-NET disease. Despite having a higher sensitivity (67%), *CgA* had a low specificity as it was positive in both SI and gastric NETs as well as in 2/9 (28%) of controls; differences in its expression patterns did not achieve statistical significance. The low specificity for detection of *CgA* mRNA reflects the low specificity for detection of CgA itself in plasma.[33] Overall, this approach is more sensitive than identifying proopiomelanocortin (POMC) transcripts in patients with Cushing's disease (0% identification) and tyrosinase mRNA in melanoma patients (20.9%) while the relationship between tumor tissue expression and plasma levels is better than that of cathepsin D in glioblastomas (100% and 42%, respectively).[4,34,35]

In addition, the number of circulating positive transcripts identified also correlated with the stage of disease. Thus, patients with localized disease were more likely to be positive for none or one of the marker transcripts (80%) compared with distant disease where two or more positive transcripts was identified. This suggests that metastasis has a significant impact on number of circulating tumor transcripts and indicates that identification of more than one transcript may identify metastatic disease. This was particularly highlighted in SI-NETs where significantly more positive transcripts (increased detection of *CgA* as well as the other four marker genes) were identified in patients with distant disease compared with those with localized/regional disease (75% versus 25%). The sensitivity and specificity for identifying limited versus advanced disease was 75% and 82%, respectively. This is higher than the sensitivity of measuring plasma CgA (43% for localized disease and 57% for disseminated disease) and is similar to that using somatostatin receptor scintigraphy (72% specificity).[2,36] This suggests that a PCR-based method examining a panel of five marker genes is more effective than an ELISA approach and may be as effective as ^{111}In-octreotide scanning. However, it should be noted that, as a general screening test (identification of two or more

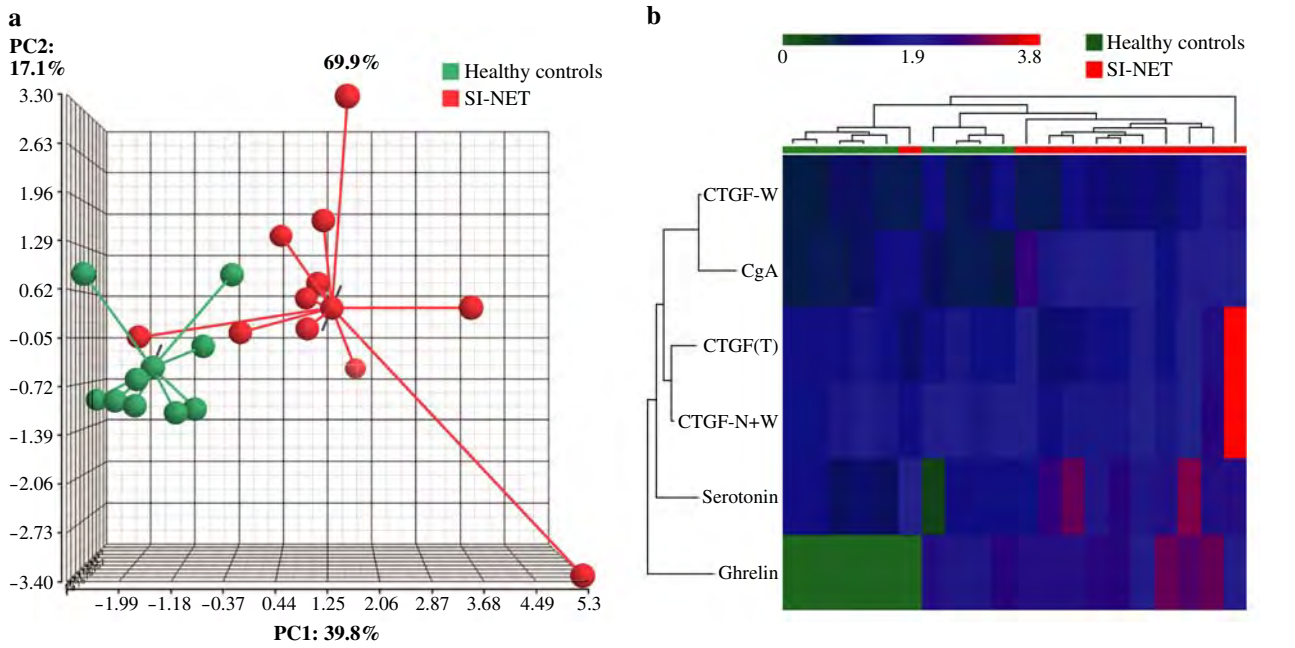

FIG. 6 Principal component analysis and hierarchical cluster of samples. (**a**) PCA visualizing the variance in healthy control samples (green), circulating gastric NETs (G-NETs: blue), and circulating SI-NETs (red). Presence or absence of mRNA (*Tph-1*, *CgA*, *DDC*, *NSE*, $VMAT_1$, and $VMAT_2$) as well as serum levels of 5-HT, CgA, Ghrelin, CTGF-W, CTGF-N+W, and CTGF(T) were reduced to $n = 3$ principal components visualizing 63.4% of the variance in the data. SI-NETs were distinct from healthy controls. (**b**) Dendrogram of hierarchically clustered samples of healthy controls (green), circulating gastric NETs (G-NETs: blue), and circulating SI-NETs (red). Log_{10}-normalized expression values ranged from 0 to 3.8 with median value of 1.9. Samples with similar expression profiles were grouped together generating distinct clusters of healthy controls and circulating G-NETs and SI-NETs. These analyses indicate that presence/absence of mRNA as well as serum hormone levels are enough to establish distinct regulatory signatures for healthy controls and circulating tumor cells

positive NET transcripts), the PCR approach is not applicable as the specificity drops to 44%. The utility of this approach is therefore in identifying both the presence and extent of disease in GI-NETs and specifically SI-NETs. In the case of gastric NETs, the correlation of clinical investigation [upper gastrointestinal endoscopy (UGI) and biopsy] would remove this lesion from consideration.

A major problem in the detection of CTC is their low number in circulating blood, estimated to be about 1 per 10^6 peripheral blood mononuclear cells.[37] Enrichment methods using tumor antibodies and flow cytometry, immunomagnetic separation, density gradient centrifugation, and filtration have been used. Systems using EpCAM antibody (targeting epithelial cell adhesion molecules) based immunomagnetic capture have proven especially efficient in finding small number of CTCs with high accuracy. In a study on 63 patients with metastatic prostate cancer, immunomagnetic capture followed by flow cytometry identified that 65% had five or more CTCs per 7.5 ml blood, and using molecular profiling the identity was classified as prostate cancer cells.[38] In another study using immunomagnetic cell capture, CTC $\geq$2 in 7.5 mL blood was identified in 52% of prostate cancer, 31% of gastric cancer, and in 30% of colorectal cancer, and by spiking blood samples with tumor cells, the technique was shown to have approximately 100% accuracy in detecting CTCs.[11] In the current study, using spike-in with KRJ-I cells (derived from a SI-NET), we found that mRNA for specific NET genes from just one KRJ-1 cell/ml blood could be detected by real-time PCR, thus confirming the high sensitivity of the method.[39] We further used this PCR approach to detect circulating mRNA in plasma samples from patients treated for NETs in our institution. As this study was performed retrospectively on frozen plasma samples, an optimal procedure to collect blood and extract mRNA from buffy coat, as we demonstrated in the spike-in experiments, could not be followed. Despite this, expression of *Tph-1* appeared to be a promising plasma marker gene for SI-NETs.

NETs produce and secrete amine and peptide hormones together with CgA from granule stores within the tumor cells.[2] Although increased plasma CgA levels are sensitive (>90%) as markers of GI-NETs, they are nonspecific as they are also elevated in other NETs, including pancreatic and small cell lung neoplasia, and even in some prostate carcinomas.[33] False positives also occur in patients with renal impairment and atrophic gastritis and individuals receiving proton-pump inhibitor therapy.[40] In addition to

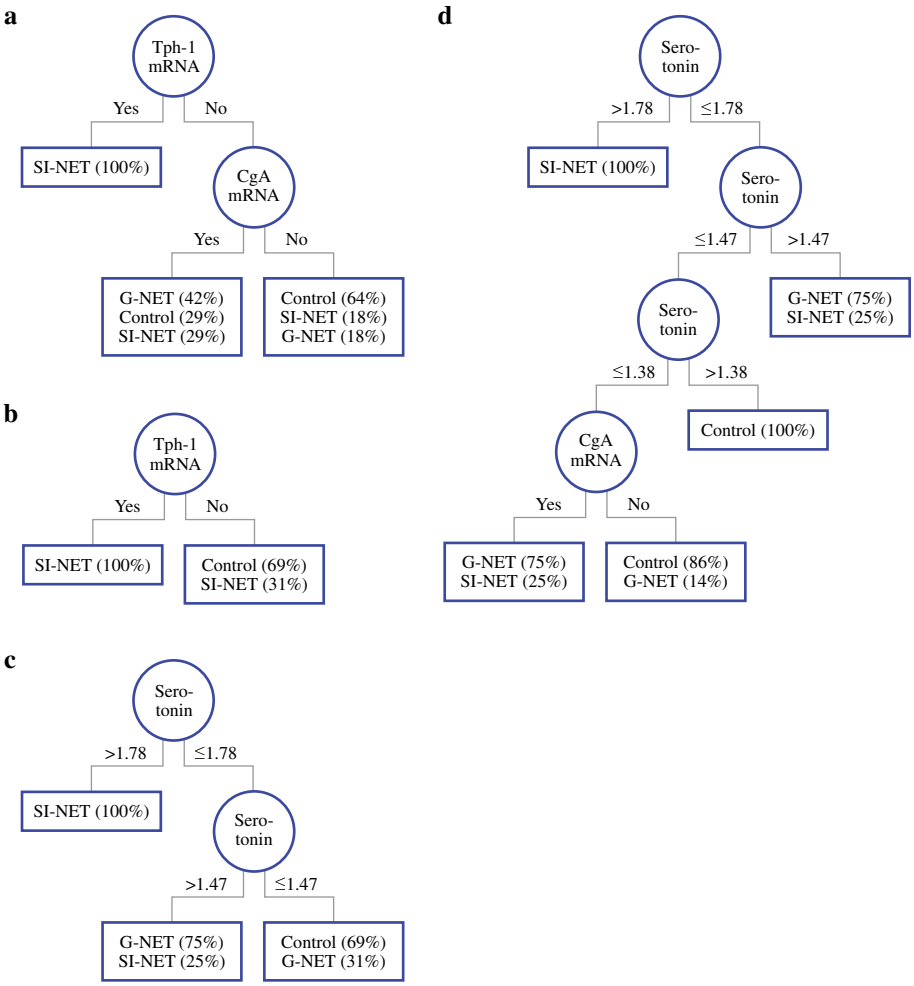

FIG. 7 Decision trees supplemented with pruning algorithm was used to predict and classify SI-NETs and G-NETs (**a**) or SI-NETs alone (**b**) based on presence/absence of circulating marker transcript levels. A similar classification strategy was undertaken to predict G-NETs and SI-NETs using marker serum levels (**c**) while the utility of all factors in all tumor types is included in (**d**). The presence or absence of *Tph-1* mRNA and serum levels of serotonin are the key differentiators in the decision-making process

TABLE 1 Class predictions produced by the decision tree classification model using presence or absence of circulating mRNA (*Tph-1*, *CgA*, *DDC*, *NSE*, $VMAT_1$, and $VMAT_2$). Pruning identified *Tph-1* and *CgA* mRNA as the most relevant variables in the decision process

	True CON	True SI-NET	True G-NET	Class precision (specificity)
Pred. CON	7	2	2	63.6%
Pred. SI-NET	0	7	0	100.0%
Pred. G-NET	2	2	5	55.6%
Class recall (sensitivity)	77.8%	63.6%	71.4%	

TABLE 2 Class predictions produced by the decision tree classification model using presence or absence of circulating mRNA (*Tph-1*, *CgA*, *DDC*, *NSE*, $VMAT_1$, and $VMAT_2$). Pruning algorithm determined *Tph-1* transcript alone to be the differentiator

	True CON	True SI-NET	Class precision (specificity)
Pred. CON	9	4	69.2%
Pred. SI-NET	0	7	100.0%
Class recall (sensitivity)	100.0%	63.6%	

amines and peptide hormones, NETs also produce growth factors, including CTGF, which may be a marker of fibrotic disease.[20]

In the current study, we identified that 5-HT and CgA were specific serum markers for SI-NETs versus both controls and gastric NETs, while ghrelin was also more

TABLE 3 Class predictions produced by the decision tree classification model using serum levels of 5-HT, CgA, Ghrelin, CTGF-W, CTGF-N+W, and CTGF(T). Only serum 5-HT was identified by pruning as the most relevant variable in the decision process

	True CON	True SI-NET	True G-NET	Class precision (specificity)
Pred. CON	8	0	4	66.7%
Pred. SI-NET	0	9	0	100.0%
Pred. G-NET	1	2	3	50.0%
Class recall (sensitivity)	88.9%	81.8%	42.9%	

TABLE 4 Class predictions produced by the decision tree classification model using presence or absence of mRNA (*Tph-1*, *CgA*, *DDC*, *NSE*, $VMAT_1$, and *VMAT2*) as well as serum levels of 5-HT, CgA, Ghrelin, CTGF-W, CTGF-N+W, and CTGF(T). Pruning identified *CgA* mRNA and serum 5-HT levels as the most relevant variables in the decision process

	True CON	True SI-NET	True G-NET	Class precision (specificity)
Pred. CON	7	0	2	77.8%
Pred. SI-NET	0	9	0	100.0%
Pred. G-NET	2	2	5	55.6%
Class recall (sensitivity)	77.8%	81.8%	71.4%	

commonly positive in SI-NETs versus controls. CTGF-W was elevated in both SI- and gastric NETs compared with controls, while CTGF N+W were elevated only in gastric NETs. To further refine this, we utilized principal component analysis and hierarchical clustering to examine whether SI-NETs and gastric NETs could be differentiated by measurement of circulating mRNA and plasma levels of NET secretory products. PCA was used as this has been successfully undertaken in sample sizes with subject-to-item ratios of less than 5:1.[41] In the current study, the subject-to-item ratio was 4:1 (12 variables of mRNA and serum levels and three tissue types). Similarly, hierarchical clustering was used as this is an unsupervised pattern recognition technique where sample size has no well-documented exclusion effect on analyses. Furthermore, we utilized a decision tree model to identify disease-associated classifiers. As decision-tree-based data classifiers are subject to overfitting as the size of the data set increases, smaller sample sizes are important and do not affect the classifier(s) identified.[42,43] Using these approaches, we identified a classifier that has mathematical capacity to diagnose SI-NETs with 81.8% sensitivity and 100% specificity. For gastric NETs, however, both sensitivity and specificity were low (42.9% and 50%, respectively). These results obviously reflect the fact that the markers we examined are more commonly associated with SI-NETs. Furthermore, gastric NETs do not represent a dangerous disease since the majority (65%) are type I lesions associated with atrophic gastritis and behave in a benign fashion (survival does not differ from the general population).[44,45] Nevertheless, by combining the results from the detection of circulating mRNA with those for tumor secretory products, we could improve the overall sensitivity for gastric NETs to 71.4%, although the specificity remained relatively low (55.6%).

The key issue in surveillance remains identification of SI-NETs since routine clinical surveillance of the small bowel does not exist as there are no identified small-intestinal-specific symptoms. Indeed, most SI-NETS are identified serendipitously on surveillance colonoscopy or in the assessment of an occult GI bleed or surgery for a perforation.[44] In contrast, the presence of dyspepsia commonly leads to an UGI endoscopy.[46]

In conclusion, measuring circulating mRNA represents a promising technique for the detection of NET disease since it is possible to detect the presence of a single cell/ml of blood. Furthermore, an amplification of this strategy by including several mRNA markers may be useful for real-time PCR-based detection of RNA in circulating tumor cells and free circulating RNA. Refinements of the methodology such as the use of multimarker analysis (including NET secretory products) may further increase the sensitivity.[47] Given the current lack of a sensitive and early technique for detecting NETS and the fact that approximately 70% of patients present with metastatic disease, a prospective study, where blood collection and mRNA extraction can be optimized, is warranted.

ACKNOWLEDGEMENTS Financial support for these studies comes from NIH R01-CA115285 (I. Modlin).

REFERENCES

1. Cleary S, Phillips JK, Huynh TT, et al. Chromogranin a expression in phaeochromocytomas associated with von Hippel-Lindau syndrome and multiple endocrine neoplasia type 2. *Horm Metab Res.* 2007;39:876–83.

2. Gustafsson BI, Kidd M, Modlin IM. Neuroendocrine tumors of the diffuse neuroendocrine system. *Curr Opin Oncol.* 2008;20: 1–12.
3. Perkins GL, Slater ED, Sanders GK, Prichard JG. Serum tumor markers. *Am Fam Phys.* 2003;68:1075–82.
4. Visus C, Andres R, Mayordomo JI, et al. Prognostic role of circulating melanoma cells detected by reverse transcriptase-polymerase chain reaction for tyrosinase mRNA in patients with melanoma. *Melanoma Res.* 2007;17:83–9.
5. Cristofanilli M, Broglio KR, Guarneri V, et al. Circulating tumor cells in metastatic breast cancer: biologic staging beyond tumor burden. *Clin Breast Cancer.* 2007;7:471–9.
6. Pfitzenmaier J, Ellis WJ, Hawley S, et al. The detection and isolation of viable prostate-specific antigen positive epithelial cells by enrichment: a comparison to standard prostate-specific antigen reverse transcriptase polymerase chain reaction and its clinical relevance in prostate cancer. *Urol Oncol.* 2007;25: 214–20.
7. Zitt M, Zitt M, Muller HM, et al. Disseminated tumor cells in peripheral blood: a novel marker for therapy response in locally advanced rectal cancer patients undergoing preoperative chemoradiation. *Dis Colon Rectum.* 2006;49:1484–91.
8. Wu CH, Lin SR, Hsieh JS, et al. Molecular detection of disseminated tumor cells in the peripheral blood of patients with gastric cancer: evaluation of their prognostic significance. *Dis Markers.* 2006;22:103–9.
9. Dawood S, Cristofanilli M. Integrating circulating tumor cell assays into the management of breast cancer. *Curr Treat Options Oncol.* 2007;8:89–95.
10. Moreno JG, O'Hara SM, Gross S, et al. Changes in circulating carcinoma cells in patients with metastatic prostate cancer correlate with disease status. *Urology.* 2001;58:386–92.
11. Allard WJ, Matera J, Miller MC, et al. Tumor cells circulate in the peripheral blood of all major carcinomas but not in healthy subjects or patients with nonmalignant diseases. *Clin Cancer Res.* 2004;10:6897–904.
12. Anker P, Mulcahy H, Stroun M. Circulating nucleic acids in plasma and serum as a noninvasive investigation for cancer: time for large-scale clinical studies? *Int J Cancer.* 2003;103:149–52.
13. Lo KW, Lo YM, Leung SF, et al. Analysis of cell-free Epstein–Barr virus associated RNA in the plasma of patients with nasopharyngeal carcinoma. *Clin Chem.* 1999;45:1292–4.
14. Kopreski MS, Benko FA, Kwak LW, Gocke CD. Detection of tumor messenger RNA in the serum of patients with malignant melanoma. *Clin Cancer Res*. 1999;5:1961–5.
15. Goebel G, Zitt M, Zitt M, Muller HM. Circulating nucleic acids in plasma or serum (CNAPS) as prognostic and predictive markers in patients with solid neoplasias. *Dis Markers.* 2005;21:105–20.
16. YK, Lo YM. Diagnostic developments involving cell-free (circulating) nucleic acids. *Clin Chim Acta* 2006;363:187–96.
17. El-Hefnawy T, Raja S, Kelly L, et al. Characterization of amplifiable, circulating RNA in plasma and its potential as a tool for cancer diagnostics. *Clin Chem.* 2004;50:564–73.
18. de Herder WW. Biochemistry of neuroendocrine tumours. *Best Pract Res Clin Endocrinol Metab.* 2007;21:33–41.
19. Stridsberg M, Oberg K, Li Q, et al. Measurement of chromogranin A, chromogranin B (secretogranin I), chromogranin C (secretogranin II) and pancreastatin in plasma and urine from patients with carcinoid tumours and endocrine pancreatic tumours. *J Endocrinol.* 1995;144:49–59.
20. Kidd M, Modlin IM, Shapiro MD, et al. CTGF, intestinal stellate cells and carcinoid fibrogenesis. *World J Gastroenterol.* 2007;13:5208–16.
21. Soga J. Early-stage carcinoids of the gastrointestinal tract: an analysis of 1914 reported cases. *Cancer.* 2005;103:1587–95.
22. Modlin IM, Kidd M, Pfragner R, et al. The functional characterization of normal and neoplastic human enterochromaffin cells. *J Clin Endocrinol Metab.* 2006;91:2340–8.
23. Kidd M, Modlin IM, Mane SM, et al. The role of genetic markers—NAP1L1, MAGE-D2, and MTA1—in defining small-intestinal carcinoid neoplasia. *Ann Surg Oncol.* 2006;13:253–62.
24. Kidd M, Eick GN, Modlin IM, et al. Further delineation of the continuous human neoplastic enterochromaffin cell line, KRJ-I, and the inhibitory effects of lanreotide and rapamycin. *J Mol Endocrinol.* 2007;38:181–92.
25. Kidd M, Nadler B, Mane S, et al. GeneChip, geNorm, and gastrointestinal tumors: novel reference genes for real-time PCR. *Physiol Genomics.* 2007;30:363–70.
26. Stridsberg M, Eriksson B, Oberg K, Janson ET. A comparison between three commercial kits for chromogranin A measurements. *J Endocrinol* 2003;177:337–41.
27. Patterson M, Murphy KG, le Roux CW, et al. Characterization of ghrelin-like immunoreactivity in human plasma. *J Clin Endocrinol Metab.* 2005;90:2205–11.
28. Jaffa AA, Usinger WR, McHenry MB, et al. Connective tissue growth factor and susceptibility to renal and vascular disease risk in type 1 diabetes. *J Clin Endocrinol Metab.* 2008.
29. Partek. Partek® genomics suite™. St. Louis: Partek Inc.; 2008.
30. Mohlig M, Floter A, Spranger J, et al. Predicting impaired glucose metabolism in women with polycystic ovary syndrome by decision tree modelling. *Diabetologia*. 2006;49:2572–9.
31. Jolliffe IT. Principle component analysis. New York: Springer; 1986.
32. Zhang H, Singer B. Recursive partitioning in the health sciences (statistics for biology and health). New York: Springer; 1999.
33. Sciarra A, Monti S, Gentile V, et al. Chromogranin A expression in familial versus sporadic prostate cancer. *Urology.* 2005; 66:1010–4.
34. Bondioni S, Mantovani G, Polentarutti N, et al. Evaluation of proopiomelanocortin mRNA in the peripheral blood from patients with Cushing's syndrome of different origin. *J Endocrinol Invest.* 2007;30:828–32.
35. Fukuda ME, Iwadate Y, Machida T, et al. Cathepsin D is a potential serum marker for poor prognosis in glioma patients. *Cancer Res.* 2005;65:5190–4.
36. Cimitan M, Buonadonna A, Cannizzaro R, et al. Somatostatin receptor scintigraphy versus chromogranin A assay in the management of patients with neuroendocrine tumors of different types: clinical role. *Ann Oncol.* 2003;14:1135–41.
37. Ross AA, Cooper BW, Lazarus HM, et al. Detection and viability of tumor cells in peripheral blood stem cell collections from breast cancer patients using immunocytochemical and clonogenic assay techniques. *Blood.* 1993;82:2605–10.
38. Shaffer DR, Leversha MA, Danila DC, et al. Circulating tumor cell analysis in patients with progressive castration-resistant prostate cancer. *Clin Cancer Res.* 2007;13:2023–9.
39. Pfragner R WG, Niederle B, Behmel A, Rinner I, Mandl A, Wawrina F, et al. Establishment of a continuous cell line from a human carcinoid of the small intestine (KRJ-I): characterization and effects of 5-azacytidine on proliferation. *Int J Oncol.* 1996;8:513–20.
40. Syversen U, Ramstad H, Gamme K, et al. Clinical significance of elevated serum chromogranin A levels. *Scand J Gastroenterol.* 2004;39:969–73.
41. Ford JK, MacCallum RC, Tait M. The application of exploratory factor analysis in applied psychology: a critical review and analysis. *Personnel Psychol.* 1986;39:291–314.
42. Domingos P. Occam's two razors: the sharp and the blunt. In: Agrawal R, Stolorz P (editors) Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining. New York: AAAI; 1998.

43. Oates T, Jensen D. The Effects of training set size on decision tree complexity. In: Proceedings of the fourteenth international conference on machine learning. Nashville: Morgan Kaufmann; 1997.
44. Modlin IM, Kidd M, Latich I, et al. Current status of gastrointestinal carcinoids. *Gastroenterology.* 2005;128:1717–51.
45. Borch K, Ahren B, Ahlman H, et al. Gastric carcinoids: biologic behavior and prognosis after differentiated treatment in relation to type. *Ann Surg.* 2005;242:64–73.
46. Modlin I, Sachs G. Acid related diseases-biology and treatment. Philadelphia: Lippincott Williams and Wilkins; 2004.
47. Xi L, Nicastri DG, El-Hefnawy T, et al. Optimal markers for real-time quantitative reverse transcription PCR detection of circulating tumor cells from melanoma, breast, colon, esophageal, head and neck, and lung cancers. *Clin Chem.* 2007;53:1206–15.