J Clin Endocrin Metab. First published ahead of print October 18, 2012 as doi:10.1210/jc.2012-1923

Analytical Performance Verification of a Molecular Diagnostic for Cytology-Indeterminate Thyroid Nodules

P. Sean Walsh, Jonathan I. Wilde, Edward Y. Tom, Jessica D. Reynolds, Daphne C. Chen, Darya I. Chudova, Moraima Pagan, Daniel G. Pankratz, Mei Wong, James Veitch, Lyssa Friedman, Robert Monroe, David L. Steward, Mark A. Lupo, Richard B. Lanman, and Giulia C. Kennedy

Veracyte, Inc. (P.S.W., J.I.W., E.Y.T., J.D.R., D.C.C., D.I.C., M.P., D.G.P., M.W., J.V., L.F., R.M., R.B.L., G.C.K.), South San Francisco, California 94080; Department of Otolaryngology—Head and Neck Surgery (D.L.S.), University of Cincinnati College of Medicine, Cincinnati, Ohio 45267; and the Thyroid and Endocrine Center of Florida (M.A.L.), Sarasota, Florida 34239

Objective: Our objective was to verify the analytical performance of the Afirma gene expression classifier (GEC) in the classification of cytologically indeterminate thyroid nodule fine-needle aspirates (FNAs).

Design: Analytical performance studies were designed to characterize the stability of RNA in FNAs during collection and shipment, analytical sensitivity as applied to input RNA concentration and malignant/benign FNA mixtures, analytical specificity (*i.e.* potentially interfering substances) as tested on blood and genomic DNA, and assay performance studies including intra-nodule, intra-assay, inter-assay, and inter-laboratory reproducibility.

Results: RNA content within FNAs preserved in FNAProtect is stable for up to 6 d at room temperature with no changes in RNA yield ($P = 0.58$) or quality ($P = 0.56$). FNA storage and shipping temperatures were found to have no significant effect on GEC scores ($P = 0.55$) or calls (100% concordance). Analytical sensitivity studies demonstrated tolerance to variation in RNA input (5–25 ng) and to the dilution of malignant FNA material down to 20%. Analytical specificity studies using malignant samples mixed with blood (up to 83%) and genomic DNA (up to 30%) demonstrated negligible assay interference with respect to false-negative calls, although benign FNA samples mixed with relatively high proportions of blood demonstrated a potential for false-positive calls. The test is reproducible from extraction through GEC result, including variation across operators, runs, reagent lots, and laboratories (SD of 0.158 for scores on a >6 unit scale).

Conclusions: Analytical sensitivity, analytical specificity, robustness, and quality control of the GEC were successfully verified, indicating its suitability for clinical use. (***J Clin Endocrinol Metab* 97: 0000–0000, 2012**)

Fine-needle aspirate (FNA) biopsy is the most widely used method for the clinical evaluation of potentially suspicious thyroid nodules. However, 15–30% of cases cannot be conclusively diagnosed by FNA cytology alone (1) and are considered indeterminate. The categories of indeterminate include atypia or follicular lesion of undetermined significance, follicular/Hürthle-cell neoplasm or suspicious for follicular/Hürthle-cell neoplasm, and suspicious for malignancy (2–5). Surgery is usually recommended for these patients to obtain a more definitive diagnosis (1, 4). Postoperatively, 66–80% of indeterminate cases are found to be benign, revealing a significant rate of unnecessary surgery, complications, and morbidity (6, 7).

ISSN Print 0021-972X ISSN Online 1945-7197
Printed in U.S.A.

doi: 10.1210/jc.2012-1923 Received April 9, 2012. Accepted September 24, 2012.

Abbreviations: CI, Confidence interval; CLIA, Clinical Laboratory Improvement Amendments; EGAPP, Evaluation of Genomic Applications in Practice and Prevention; FNA, fine-needle aspirate; GEC, gene expression classifier; LCT, lymphocytic thyroiditis; PTC, papillary thyroid carcinoma; RIN, RNA integrity number.

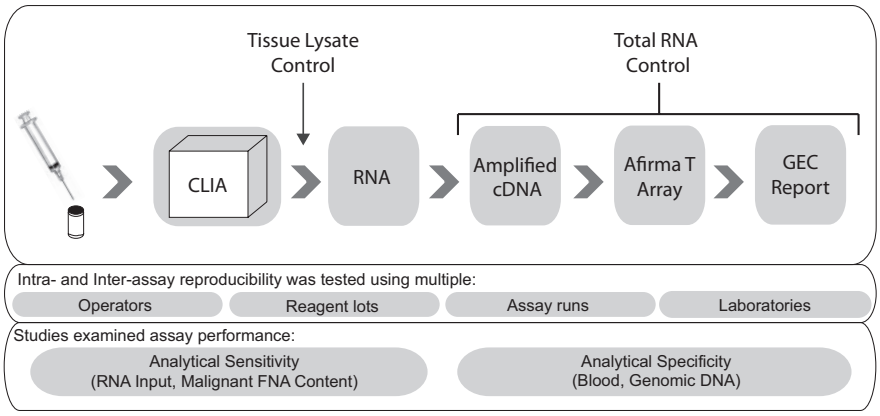


FIG. 1. Workflow of Afirma GEC. Total RNA is extracted and amplified to generate cDNA, which is subsequently labeled and hybridized to a custom Afirma-T microarray. Array signals are analyzed via a classification algorithm, producing a GEC report with either a benign or suspicious GEC call. Assay performance was measured by a series of reproducibility experiments as well as studies demonstrating analytical sensitivity and analytical specificity.

A gene expression classifier (GEC) was developed to identify benign thyroid nodules in the subgroup of patients with indeterminate FNA cytopathology (8, 9). The GEC is intended for cytologically indeterminate FNAs collected in routine clinical practice and immediately stored in preservative solution. Processed samples are hybridized to a custom Afirma thyroid microarray and analyzed with a classification algorithm to produce either a benign or suspicious GEC call (Fig. 1).

The clinical validity of the GEC has been reported elsewhere (9) with results from a recently completed, large, independent, multicenter trial, confirming that a benign GEC result carries a risk of malignancy comparable to that of a benign cytopathology diagnosis. We expect that benign GEC results will enable a significant number of patients and physicians to consider clinical and sonographic follow-up in lieu of diagnostic surgery (8, 10), a finding supported by recent clinical studies (11).

Although the clinical validity of the GEC has been confirmed, it is equally important to demonstrate analytic validity of this newly developed molecular test. The Evaluation of Genomic Applications in Practice and Prevention (EGAPP) Working Group and the Centers for Disease Control's ACCE Project (analytic validity, clinical validity, clinical utility, and associated ethical, legal, and social implications) have defined parameters that should be used to evaluate analytical validity of novel genomic tests (12–14). Here we report the results of recommended studies designed to test the analytical performance of the GEC. Studies included evaluation of FNA stability during collection and shipment, analytical sensitivity to input RNA and FNA malignant content, analytical specificity in response to contaminating blood and genomic DNA, and several reproducibility studies (intra-nodule, intra- and inter-assay, and inter-laboratory) demonstrating robustness to changes across a range of technical variables. Quality control recommendations were extensively implemented and verified via the use of control materials and in-process quality checkpoints at key steps in the GEC procedure.

Materials and Methods

Specimens

Prospective FNA samples were obtained with patient informed consent through Institutional Review Board-approved protocols. Either one or two needle passes were 1) aspirated *in vivo* at outpatient clinical sites, 2) aspirated *in vivo* preoperatively, or 3) aspirated *ex vivo* immediately after surgical excision (Cureline, South San Francisco, CA) and placed into FNAprotect preservative solution (QIAGEN, Valencia, CA). Samples were shipped under controlled temperature conditions (chilled or frozen) and stored at −80 C upon receipt. Clinical sites and principal investigators are listed in the Supplemental Materials and Methods (published on The Endocrine Society's Journals Online web site at http:// jcem.endojournals.org).

RNA extraction, amplification, and microarray hybridization

RNA from clinical FNA specimens was extracted using the AllPrep Micro Kit (QIAGEN). Yield was determined using Quant-IT (Invitrogen, Carlsbad, CA), and quality was determined with a Bioanalyzer Picochip System (Agilent Technologies, Santa Clara, CA), generating an RNA integrity number (RIN); samples with concentration less than 1.5 ng/μl and/or RIN below 2 were stopped from further processing per prespecified quality control criteria. Positive (tissue lysate) and negative (water) controls were included in each extraction batch, and predefined yield and quality specifications were used as acceptance criteria to ensure the reliability of every run. For each sample, 15 ng total RNA was amplified using the WT-Ovation FFPE RNA amplification system (NuGEN, San Carlos, CA), followed by conversion to sense-strand cDNA using the WT-Ovation exon module (NuGEN). Samples were fragmented and labeled using the Encore biotin module (NuGEN), followed by overnight hybridization of 3.5 μg biotin-labeled cDNA to a proprietary Afirma-T custom microarray (Affymetrix, Santa Clara, CA). The arrays were then washed, stained, and scanned on a Gene Chip System GCS3000 or DXv2 (Affymetrix) following the manufacturer's protocols. Positive (total RNA) and negative (water) controls were included in each GEC batch starting from the amplification step. Predefined specifications for yield, quality, and GEC classification of control samples (one malignant and one benign per batch) were used as acceptance criteria.

Results

Control materials

Multiple lots of tissue lysate were manufactured and used as process controls during RNA extraction. Three

different lots of controls were tested over several weeks of independent runs, by three different operators. Testing of three lots is standard practice to verify the reproducibility of a manufacturing or laboratory process. Tissue lysate controls consistently produced the expected quantity and quality of total RNA, resulting in within-lot coefficients of variation ranging from 5–15% for yield and 4–5% for RIN.

Similarly, multiple lots of benign and malignant total RNA were manufactured and used as process controls for amplification and hybridization steps. The reproducible GEC results obtained from these controls enabled concurrent monitoring of assay performance for each run. All GEC tests and studies outlined below included one benign and one malignant total RNA control.

FNA stability

Standard FNA collection procedure for the GEC involves aspiration into a preservative, subsequent handling at room temperature before shipment (typically, same day), and shipment in chilled boxes (typically, overnight). To demonstrate the stability of the RNA content within FNA samples under room-temperature conditions, FNA samples preserved in FNAprotect were stored for up to 6 d at room temperature in the molecular laboratory. This length of time is required to account for sample collection, shipping, transport, and processing in the laboratory. Samples frozen immediately at −80 C served as controls. Total RNA was then extracted and evaluated for quantity and quality (Table 1 and Fig. 2A). There was no statistically significant difference between any of the test groups and the control group in RIN (0.3 RIN units, largest median difference, $P = 0.472$) or yield (<6 ng/μl, largest median difference, $P = 0.58$).

The standard FNA collection procedure was also evaluated along with an alternative (−20 C) storage condition

TABLE 1. Summary of all analytical verification studies

Study	Sample source	Design summary	GEC calls		GEC scores, pooled SD	Intensity R^2	
			Number of GEC calls	Concordance		Median	Range
Preanalytical							
Variability in shipping conditions	Clinical and preoperative FNA	24 samples tested in up to 3 different shipping conditions	69	24/24 (100%)[a]	0.118 (0.098–0.148)	0.984	0.970–0.993
Analytical sensitivity and specificity							
Variability in RNA input quantity	Clinical FNA	3 samples tested at four RNA input values (5, 10, 15, and 25 ng) in triplicate	36	35/36 (97.2%)[b]	0.129 (0.104–0.170)	0.984	0.923–0.993
Dilution with adjacent normal tissue	Ex *vivo* FNA, *in vitro* RNA mixtures	2 benign and 3 malignant FNAs mixed with ANT from malignant nodule, down to 20% FNA content	10 for mixtures + 5 for pure samples	10/10 (100%)[b]	NA	NA[d]	NA[d]
Dilution with whole blood	Clinical FNA and whole-blood samples, *in vitro* RNA mixtures	1 benign and 2 malignant FNAs mixed with one of 9 whole-blood samples, down to 17% FNA content	27 for mixtures + 12 for pure blood	M mixtures, 16/16 (100%)[b]; B mixtures, 2/11 (18%)[b]	NA	NA[d]	NA[d]
Genomic DNA contamination	Tissue controls	2 samples with and without 30% contamination and 6 replicates	24	23/24 (96%)[b]	0.115 (0.089–0.162)	0.981	0.971–0.988
Reproducibility							
Intra-assay	Clinical FNA and total RNA controls	33 samples from 81 experimental plates with up to 3 replicates per plate	243	31/33 (93.9%)[a]	0.121 (0.109–0.136)	0.988	0.945–0.994
Inter-assay	Clinical FNA	37 samples in up to 4 runs of reagents and operators, enriched near decision boundary	147	36/37 (97%)[a]	0.158 (0.140–0.182)	0.979	0.946–0.994
Inter-laboratory	Clinical FNA	20 samples run in 2 laboratories	40	20/20 (100%)[c]	0.138 (0.105–0.201)	0.981	0.953–0.989
Intra-nodule	*Ex vivo* FNA	9 nodules with up to 5 FNAs sampled per nodule	42	40/42 (95%)[a]	0.411 (0.241–0.702)[e]	0.952	0.548–0.985

Call concordance corresponds to designed analytical studies and may be biased by sample selection. ANT, Adjacent normal tissue; B, benign; M, malignant.

[a–c] GEC call concordance was calculated relative to each unique FNA sample as follows: [a] relative to the majority call for the sample; [b] relative to the call for the pure sample in standard condition; [c] relative to the GEC call obtained in the Research and Development Laboratory.

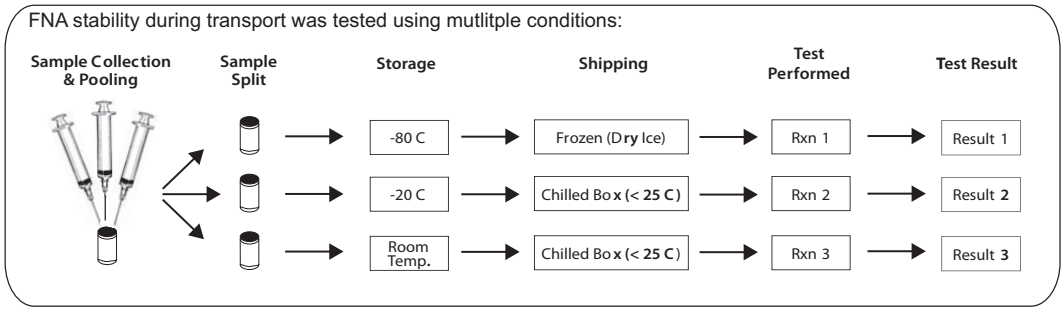
[d] Similarity of intensity profiles is not applicable (NA) for these studies because mixtures from different RNA sources are not expected to replicate original profiles.

[e] Robust estimate of pooled SD.

A

	Concentration, ng/ul				RIN			
Condition	Total	Median	Range	IQR	Total	Median	Range	IQR
-80 C	30	3.4	0.09 - 30.1	1 - 7.8	27	6.5	4.5 - 8.5	6.1 - 6.7
1 day	30	7.7	0.07 - 188.5	1.8 - 18.1	26	6.5	3.0 - 8.4	6.0 - 7.0
2 days	30	8.9	0.16 - 25.8	2.5 - 12.5	29	6.2	2.6 - 7.4	5.3 - 6.6
3 days	30	3.2	0.09 - 29.2	1.7 - 10.5	29	6.4	2.6 - 7.8	6.3 - 7.3
4 days	30	3.5	0.14 - 43.8	1 – 8	29	6.5	2.5 - 8.3	5.9 - 6.9
5 days	30	3.2	0.07 - 19.8	1.5 - 11.7	28	6.3	4.6 - 7.6	6.0 - 6.9
6 days	30	5.3	0.23 - 29.7	1.2 - 9.1	30	6.3	3.1 - 7.5	5.4 - 6.8

B



C

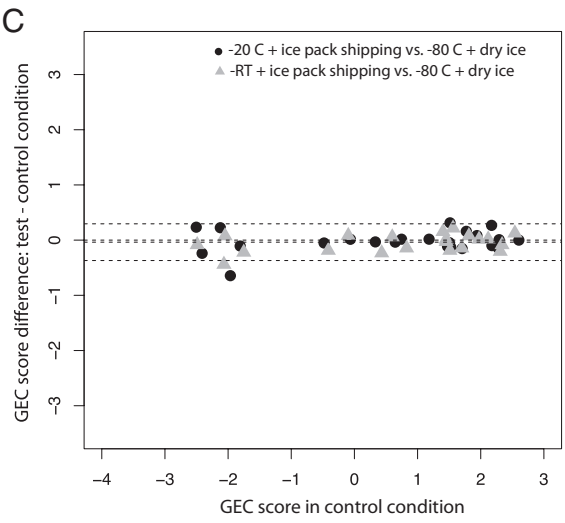


FIG. 2. A, RNA quality (RIN value) and quantity for control FNA samples kept at −80 C and FNA samples kept at 25 C for 1–6 d. Samples with RNA concentration no higher than 0.2 ng/μl were omitted from RIN analyses due to the technical limit of the Bioanalyzer method. B, Study design for testing FNA storage and shipping conditions. C, Intra-assay reproducibility of GEC scores across different shipping conditions starting from pooled and split FNA samples. IQR, Interquartile range; RT, room temperature.

and compared with the −80 C control condition. FNA samples from 28 different patient nodules were collected; for each patient nodule, a total of three FNA passes were combined into a single tube of FNAprotect (3× volume) and then divided equally into three different tubes of FNAprotect. Each of the three tubes was then subjected to different storage and shipping temperatures (Fig. 2B). RNA quality control results indicated no significant difference in total RNA concentration [<0.25 log2 (ng/μl) between the groups, P value = 0.076] but suggested small differences in RIN (< 0.4 RIN units between the groups, P value < 0.005). Such small differences in RIN value are within the claimed measurement error for the Bioanalyzer and were found to not be practically significant for this test, as seen from the analysis of GEC results described below. Sixty-nine samples from 24 nodules were processed and evaluated through to final GEC results. All samples from the same nodule produced concordant GEC calls irrespective of the shipping method. Analysis of the GEC results indicated no systematic difference in test conditions *vs.* the control condition (<0.04 score unit difference between groups with scores spanning a range of over

5 units, *P* value = 0.55, Fig. 2C). Pooled SD of GEC scores [SD = 0.118; 95% confidence interval (CI) = 0.098–0.148)] was comparable to standard intra-run reproducibility starting from total RNA. Signal intensities for transcripts used by the GEC were highly reproducible for each nodule across the three conditions tested (median R^2 = 0.984; range = 0.970–0.993), indicating that the sample-splitting procedure successfully produced three equivalent FNA samples. Thus, this study demonstrates a high level of technical reproducibility over the entire assay, from FNA collection, shipment, and RNA extraction to GEC results. Based on these data, room-temperature storage at the clinical site and chilled-box shipping was successfully verified for routine practice.

Analytical sensitivity: total RNA input quantity

Although the standard total RNA input quantity to the GEC assay is fixed (15 ng), some measurement variability around this nominal input amount can be expected in routine practice. Thus, a study was performed to characterize the tolerance of transcript array signal intensities and GEC results to variability in total RNA input, down to 5 ng. Total RNA was extracted from each of three different FNA samples and processed through the GEC in triplicate at varying total RNA input (5, 10, 15, and 25 ng). Samples were chosen to represent low, medium, and high ranges of the GEC score. GEC scores for each FNA did not differ significantly regardless of RNA input (<0.11 absolute mean GEC score difference to the standard amount of 15 ng, *P* value = 0.32). Overall, pooled SD of GEC scores across input amounts was 0.129 [95% CI = 0.104–0.170], consistent with intra-run expectations (Table 1). The transcript signal intensities were highly correlated within any set of sample triplicates and within each single group of RNA input [median R^2 coefficients of 0.973 (5 ng input), 0.985 (10 ng input), 0.986 (15 ng input), and 0.988 (25 ng input)]. A decrease in signal reproducibility at the 5 ng range was small but significant (*P* value <0.001). Transcript signal intensities from all three FNA samples were also highly correlated between triplicates processed at test input amounts and standard 15-ng condition [median R^2 coefficients of 0.980 (5 *vs.* 15 ng), 0.986 (10 *vs.* 15 ng), and 0.986 (25 *vs.* 15 ng)]. Overall, this study demonstrated high tolerance to RNA input variation within the tested range, showing that the 10 ng results were indistinguishable from the standard 15 ng input.

Analytical sensitivity: dilution of malignant FNA content

The content of malignant cells within an FNA sample obtained from a malignant nodule can vary from sample to sample. Tolerance of the GEC to dilution of malignant content was evaluated using *in vitro* total RNA mixtures derived from three papillary thyroid carcinoma (PTC) nodules from different patients and adjacent normal *ex vivo* FNAs from one of the patients with a malignant nodule. The pure adjacent normal tissue was called benign by the GEC, whereas all pure PTC samples and mixtures (with up to 80% adjacent normal content) resulted in suspicious GEC calls (Table 1 and Fig. 3). Tolerance of GEC results to dilution of benign content was evaluated in a similar manner for two benign nodules. All pure benign samples and mixtures tested resulted in benign GEC calls. GEC scores for the *in vitro* mixtures were in close agreement with an *in silico* mixture model as previously demonstrated with benign nodule mixtures (8), further demonstrating that the signature present in malignant PTC FNAs is sufficiently strong to withstand a wide range of dilution.

Analytical specificity: blood

FNA samples may contain varying amounts of blood due to variation in the needle collection procedure. To test the impact of blood on the GEC results, *in vitro* mixtures were created using RNA from malignant or benign FNAs mixed into a background of RNA derived from fresh

Adjacent normal mixtures			
	Specimen	Percent (%) Mixing proportion	GEC Call
Pure samples	ANT	100	Benign
	BFN	100	Benign
	FA	100	Benign
	PTC	100	Suspicious
	PTC	100	Suspicious
	PTC	100	Suspicious
Benign mixtures	BFN + ANT	50/50	Benign
	FA + ANT	20/80	Benign
		60/40	Benign
Malignant mixtures	PTC + ANT*	20/80	Suspicious
		40/60	Suspicious
		50/50	Suspicious
		60/40	Suspicious
	PTC + ANT	20/80	Suspicious
		50/50	Suspicious
	PTC + ANT	20/80	Suspicious

FIG. 3. Classification results for *in vitro* mixtures of malignant FNA and adjacent normal tissue. *, Paired mixtures of malignant and adjacent normal samples obtained from the same patient. ANT, Adjacent normal tissue; BFN, benign follicular nodule; FA, follicular adenoma.

TABLE 2. GEC results from *in vitro* mixtures of thyroid FNA and blood

Sample	Thyroid FNA (%)	Blood (%)	Undiluted FNA	Whole-blood sample								
				WB-01	WB-02	WB-03	WB-04	WB-05	WB-06	WB-07	WB-08	WB-09
PTC-1	100	0	S[a]									
	50	50							S[a]		S	
	33	66							S[a]		S	
	17	83		S	S				S[a]		S	
PTC-2	100	0	S[a]									
	50	50									S	
	33	66									S	
	17	83				S	S				S	
LCT	100	0	B[a]									
	50	50							S[a]		B	
	33	66							S[a]		B	
	17	83						S	S[a]	S	S	
Pure blood	0	100		S	S	S	B	S	S[a]	S	S[a]	B[a]

Mixtures were done with total RNA from three FNA samples and nine whole-blood samples from independent patients. B, Benign; S, suspicious.

[a] GEC results obtained on two technical replicates of the same mixture, and were concordant in all cases.

whole blood. GEC calls for pure whole blood were suspicious in seven of nine samples; malignant FNA/blood mixtures were correctly classified as suspicious for all tested samples, even those with up to 83% blood content (Tables 1 and 2). This included a mixture of PTC-2 with WB-04, where pure blood was classified as benign, demonstrating that 17% malignant FNA content is sufficient to correctly classify the mixture. Additional *in silico* mixing experiments with signals from pure blood samples indicated that 80% of all malignant samples, including PTC and non-PTC indeterminate FNAs, maintained a correct suspicious GEC call up to at least 80% blood content (data not shown). A benign sample [lymphocytic thyroiditis (LCT)] mixed with blood sample WB-08 resulted in a correct benign GEC call at 50 and 66% blood content but not at 83%. However, this same benign sample (LCT) resulted in a suspicious GEC call with blood sample WB-06 at 50% blood content, demonstrating that false-positive results may occur with some FNA samples that are dominated by blood.

Analytical specificity: genomic DNA

Genomic DNA was tested as a potentially interfering substance, because the presence of DNA can occur from inadvertent deviations from the RNA extraction process. Routine in-process quality control methods using the Bioanalyzer are capable of detecting at least 30% genomic DNA content in total RNA isolates, preventing such samples from additional processing. Thus, assay testing was only necessary for up to 30% genomic DNA contamination (*i.e.* 15 ng total RNA + 6.4 ng genomic DNA from the same sample). Benign and malignant total RNA control samples were tested in a standard and test condition with six replicates per condition. GEC scores for samples contaminated with worst case 30% genomic DNA had a small systematic bias of −0.11 (P value <0.02) toward suspicious GEC calls, resulting in a slight potential false-positive rate increase in the highly unlikely case of inadvertent contamination with genomic DNA (Table 1). Importantly, the data show that this type of potential interference does not affect the false-negative characteristics of the GEC, the most important factor in clinical validity.

Intra-nodule reproducibility

Thyroid FNA sampling variability presents a potential challenge in accurate FNA interpretation. To evaluate the reproducibility of GEC results for different double-pass FNA samplings from the same nodule, we processed 42 samples collected *ex vivo* from nine independent nodules, with up to five FNA samplings per nodule. Six of nine nodules tested had cytopathology and surgical histopathology classifications of malignant, and all replicates from each of these samples classified correctly in the GEC as suspicious (Table 1 and Fig 4A). A robust estimate of within-nodule pooled SD in GEC scores for all nine nodules was 0.411 (95% CI = 0.241–0.702). One nodule had significantly higher within-nodule SD in GEC scores compared with the other eight nodules (1.36 SD, P value <0.001), yet each of its FNA samplings was correctly classified. The transcript signal intensities from different samplings of the same nodule had median R^2 coefficients of 0.952 (range 0.548–0.985). These data suggest that biological variability accounts for a larger component of variation in GEC scores compared with technical/assay variability ($P < 0.001$, Table 1 and Fig. 4B).

Assay reproducibility

The within-run repeatability of the GEC was evaluated using total RNA from 33 FNA samples and controls, pro-

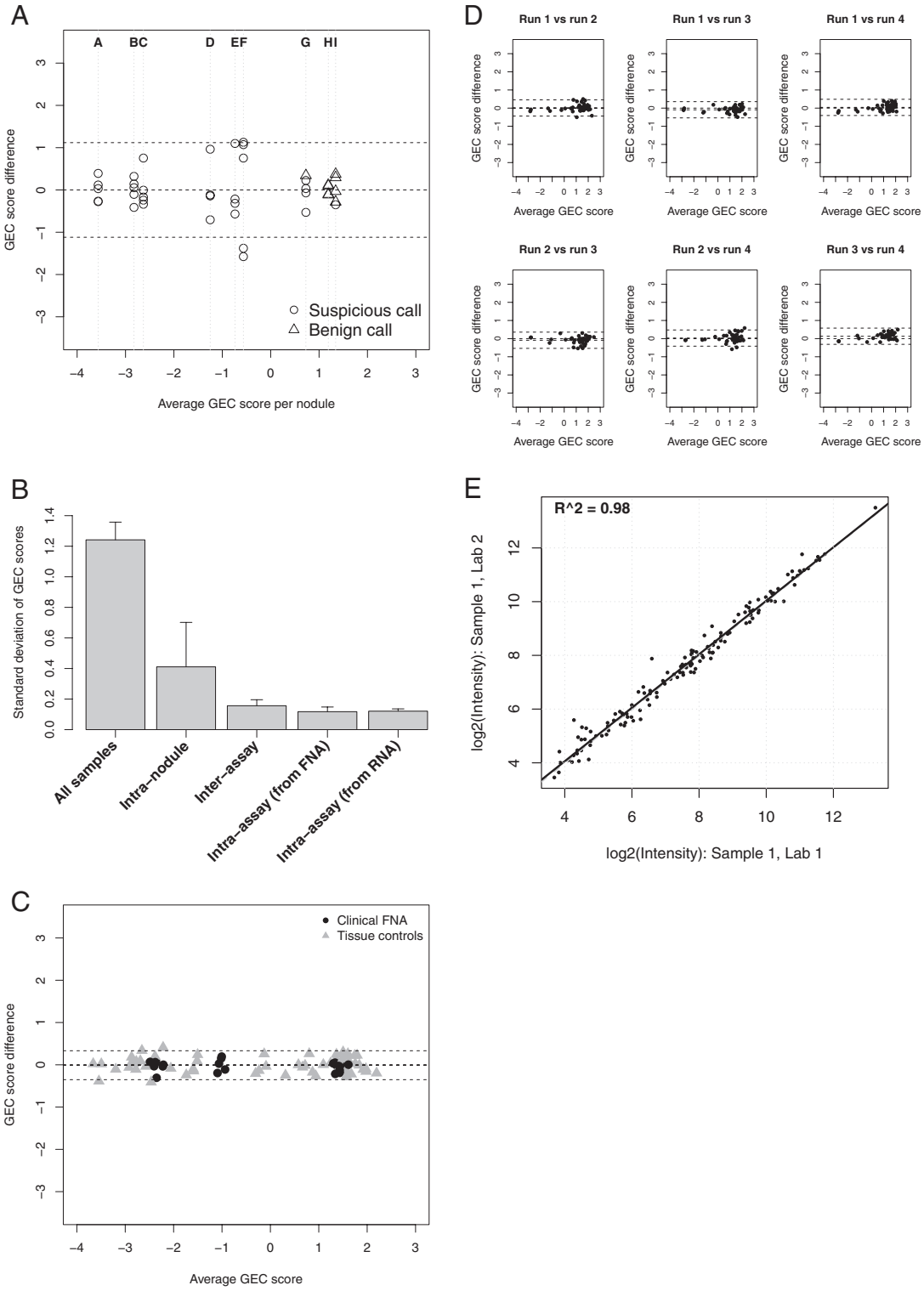


FIG. 4. Panel A, Intra-nodule reproducibility of GEC scores. Each *vertical column* represents different samplings of the same nodule (nodules A–I). Within-nodule GEC calls were concordant for all nodules except those labeled G and I. Panel B, Comparison of GEC score SD demonstrating low variability in inter- and intra-assay studies. Panel C, Intra-assay (*i.e.* within-run) reproducibility of GEC scores demonstrating similar technical variability across the range of GEC scores. Panel D, Inter-assay reproducibility of GEC scores across four runs. Panel E, Scatter plot of normalized signal intensity for transcripts used by the classifier for a representative sample in the inter-laboratory reproducibility study.

cessed in triplicate in a series of 81 experimental runs (243 GEC results), with varying reagent lots and operators and spanning more than 15 months. The pooled within-run SD of GEC scores was estimated to be 0.121 (95% CI = 0.109–0.1364; Table 1 and Fig. 4C). Variation of GEC scores was similar across the range of GEC scores, as measured by the dependence of absolute residuals of the scores on the mean scores (P value = 0.86). The within-run SD of

GEC score for total RNA controls [0.130 (95% CI = 0.115–0.149), estimated from 59 triplicates of 28 unique tissue control lots] was not smaller than the variation in triplicate FNA samples [0.092 (95% CI = 0.077–0.117), estimated from 22 triplicates of five unique FNA samples]. The transcript signal intensities from within-run replicates had median R^2 coefficients of 0.988 (range 0.945–0.994).

In a formal study of inter-run reproducibility, total RNA from 37 different FNAs were tested in four different runs corresponding to four different prequalified lots of critical reagents, with each run performed by one of three different operators. Benign and malignant FNA samples with GEC scores concentrated around the clinical decision boundary were chosen for this study to increase the statistical power to detect changes in this range, and these had RIN scores in the range of 4.1–9.0. Of 37 samples tested, 36 resulted in concordant GEC calls across all four runs (97% concordance). The GEC scores were estimated to have an inter-run pooled SD of 0.158 (95% CI = 0.140–0.182) across all FNAs in this study (Table 1 and Fig. 4D) and no dependence on RIN score (*P* value 0.20). The transcript signal intensities from across-run replicates had median R^2 coefficients of 0.979 (range 0.946–0.994). An additional study using eight samples, representing four malignant subtypes (follicular thyroid carcinoma, follicular variant of papillary carcinoma, PTC, and lymphoma) across 47 technical replicates, showed similar inter-run reproducibility performance, pooled SD = 0.138 (Supplemental Table 1). Thus, GEC call concordance demonstrated high reproducibility across reagent lots, operators, and processing runs.

Inter-laboratory reproducibility

Total RNA from 20 different patient FNA samples was processed through the GEC in the laboratory where the test was developed (Veracyte Research and Development Laboratory). A second aliquot of RNA from each of these samples was later tested in a different, Clinical Laboratory Improvement Amendments (CLIA)-certified reference laboratory using different operators, reagent lots, and equipment (same model equipment, different by serial number) (Veracyte CLIA laboratory). The GEC calls for all samples were 100% concordant between the two laboratories and with available surgical pathology diagnosis, demonstrating inter-laboratory reproducibility and accuracy of GEC results. Inter-laboratory pooled SD of GEC scores was estimated to be 0.138 (95% CI = 0.105–0.201), which is in agreement with the 0.158 SD calculated for within-lab inter-assay reproducibility. Similarly, transcript signal intensities were highly correlated between laboratories across all samples (median R^2 = 0.981, range = 0.953–0.989), consistent with expectations for interassay results (Fig. 4E).

Discussion

Analytical and clinical validity are important factors in the evaluation of any new molecular test. We have previously reported the clinical validity of the Afirma GEC classifier as a useful tool in the clinical evaluation of cytologically indeterminate FNAs (9). Here we set out to verify the analytical validity of this test. The entire process of collection, storage, shipping, sample processing, and classification was evaluated (Table 1). We demonstrate that nucleic acids extracted from clinical FNAs are stable and yield reproducible results across a variety of conditions. Thyroid FNAs are complex samples comprised of a heterogeneous mixture of cells and colloid, the exact proportion of which cannot be readily determined. Hence, FNAs are by nature dilute samples that contain varied malignant content. To construct a model system to evaluate the analytical sensitivity of the GEC to decreasing amounts of informative cellular content, PTC samples were used to represent as much as possible pure malignant material as the starting point for *in vitro* mixtures with adjacent normal tissue. We previously reported the response of the GEC to simulated dilution using mixtures of malignant and benign FNA (8). In both studies, the observed *in vitro* GEC scores fit the *in silico* model. This suggests that the mixing model can be generalized to other samples, and from this analysis using PTC samples, we determined that FNAs with malignant content can tolerate significant further dilution. FNA samples with heterogeneous malignant content (*i.e.* due to sampling or biology) may have less tolerance to *in vitro* dilution than the highly malignant PTC samples used in the model system described here. The analytical verification results with PTC samples represented in the studies described in this report may not extend to all thyroid and nonthyroid cancers. However, multiple clinical sensitivity studies demonstrate that 92% of malignant samples are correctly classified (9).

Analytical specificity was also evaluated. In the case of malignant samples, the GEC was robust in the presence of blood, maintaining correct classification up to 83% blood RNA. Benign samples mixed with very high proportions of blood had the potential for false-positive results, because the majority of pure blood samples tested were called GEC suspicious. FNAs with large amounts of obscuring blood are generally classified as nondiagnostic by

a cytopathologist (15), and corresponding dedicated FNA molecular FNA passes from the same patient would not be tested by the GEC, which is intended for use only in cytology indeterminate cases. Although the dedicated FNA passes tested by the GEC are not prescreened for blood or epithelial cell content as in previous studies (16), the impact of these factors is already reflected in the clinical specificity of this test as measured in a large prospective clinical validity study (9). As separate needle passes are used for cytological examination and for GEC molecular testing, different cellular compositions in these two fractions may occur as a result. This factor represents a limitation of the GEC test as well as other molecular tests using separate dedicated needle passes.

Analytical reproducibility was evaluated following technical assessment criteria outlined by EGAPP, Centers for Disease Control's ACCE Project, and Agency for Healthcare Research and Quality, using clinical samples with GEC scores covering the entire range and concentrated around the decision boundary of the assay (17). It has been argued that accuracy studies for multigene molecular tests are often impossible due to the absence of reference methods (18). To establish accuracy of the test offered at the CLIA-certified laboratory, we demonstrated with an inter-laboratory reproducibility study that the results in this lab are identical to those generated in the laboratory where the test was developed. When taken together with our clinical validation study, the GEC successfully achieves EGAPP level I analytic validity criteria. Namely, technical validation involved the extensive use of well-characterized samples with multiple reference standard comparison methods including cytopathology, histopathology, and reference laboratory. We also evaluated the role of intra-nodule heterogeneity. Our data highlight that biological variability within a nodule accounts for a larger component of GEC score variation than technical factors.

The robustness of the GEC to induced variables, including those that may be encountered in clinical samples, indicates that routine testing of FNA specimens is feasible at high confidence from the standpoint of analytical performance and reproducibility.

Acknowledgments

Address all correspondence and requests for reprints to: P. Sean Walsh, M.P.H., VP Product Development, Veracyte, Research and Development, 7000 Shoreline Court, Suite 250, South San Francisco, California 94080. E-mail: sean@veracyte.com.

Disclosure Summary: P.S.W., J.I.W., E.Y.T., J.D.R., D.G.P., M.W., L.F., R.M., R.B.L., and G.C.K. are employed by Veracyte. D.I.C. and M.P. are consultants for Veracyte. J.V. and D.C.C. were previously employed by Veracyte. D.L.S. and M.A.L. have received research support from Veracyte.

References

1. **Cooper DS, Doherty GM, Haugen BR, Hauger BR, Kloos RT, Lee SL, Mandel SJ, Mazzaferri EL, McIver B, Pacini F, Schlumberger M, Sherman SI, Steward DL, Tuttle RM** 2009 Revised American Thyroid Association management guidelines for patients with thyroid nodules and differentiated thyroid cancer. Thyroid 19:1167–1214
2. **Yang J, Schnadig V, Logrono R, Wasserman PG** 2007 Fine-needle aspiration of thyroid nodules: a study of 4703 patients with histologic and clinical correlations. Cancer 111:306–315
3. **Raber W, Kaserer K, Niederle B, Vierhapper H** 2000 Risk factors for malignancy of thyroid nodules initially identified as follicular neoplasia by fine-needle aspiration: results of a prospective study of one hundred twenty patients. Thyroid 10:709–712
4. **Gharib H, Papini E, Paschke R, Duick DS, Valcavi R, Hegedüs L, Vitti P; AACE/AME/ETA Task Force on Thyroid Nodules** 2010 American Association of Clinical Endocrinologists, Associazione Medici Endocrinologi, and European Thyroid Association medical guidelines for clinical practice for the diagnosis and management of thyroid nodules: Executive Summary of recommendations. J Endocrinol Invest 33:287–291
5. **Baloch ZW, Fleisher S, LiVolsi VA, Gupta PK** 2002 Diagnosis of "follicular neoplasm": a gray zone in thyroid fine-needle aspiration cytology. Diagn Cytopathol 26:41–44
6. **Baloch ZW, LiVolsi VA, Asa SL, Rosai J, Merino MJ, Randolph G, Vielh P, DeMay RM, Sidawy MK, Frable WJ** 2008 Diagnostic terminology and morphologic criteria for cytologic diagnosis of thyroid lesions: a synopsis of the National Cancer Institute Thyroid Fine-Needle Aspiration State of the Science Conference. Diagn Cytopathol 36:425–437
7. **Wang CC, Wang CP, Tsai TL, Liu SA, Wu SH, Jiang RS, Shiao JY, Su MC** 2011 The basis of preoperative vocal fold paralysis in a series of patients undergoing thyroid surgery: the preponderance of benign thyroid disease. Thyroid 21:867–872
8. **Chudova D, Wilde JI, Wang ET, Wang H, Rabbee N, Egidio CM, Reynolds J, Tom E, Pagan M, Rigl CT, Friedman L, Wang CC, Lanman RB, Zeiger M, Kebebew E, Rosai J, Fellegara G, LiVolsi VA, Kennedy GC** 2010 Molecular classification of thyroid nodules using high-dimensionality genomic data. J Clin Endocrinol Metab 95:5296–5304
9. **Alexander EK, Kennedy GC, Baloch ZW, Cibas ES, Chudova D, Diggans J, Friedman L, Kloos RT, Livolsi VA, Mandel SJ, Raab SS, Rosai J, Steward DL, Walsh PS, Wilde JI, Zeiger MA, Lanman RB, Haugen BR** 2012 Preoperative Diagnosis of Benign Thyroid Nodules with Indeterminate Cytology. N Engl J Med 387:705–715
10. **Li H, Robinson KA, Anton B, Saldanha IJ, Ladenson PW** 2011 Cost-effectiveness of a novel molecular test for cytologically indeterminate thyroid nodules. J Clin Endocrinol Metab 96:E1719–E1726
11. **Duick DS, Klopper J, Diggans JC, Friedman L, Kennedy GC, Lanman RB, McIver B** 8 August 2012 The impact of benign gene expression classifier test results on the endocrinologist-patient decision to operate on patients with thyroid nodules with indeterminate FNA cytopathology. Thyroid 10.1089/thy.2012.0180
12. **Teutsch SM, Bradley LA, Palomaki GE, Haddow JE, Piper M, Calonge N, Dotson WD, Douglas MP, Berg AO** 2009 The Evaluation of Genomic Applications in Practice and Prevention (EGAPP) Initiative: methods of the EGAPP Working Group. Genet Med 11:3–14
13. **Sun F, Bruening W, Uhl S, Ballard R, Tipton R, Schoelles K** 2010 Quality regulation and clinical utility of laboratory-developed molecular tests. AHRQ Technology Assesment Program. Washington, DC: Department of Health and Human Services

14. **Project A** 2007 Evaluation of genetic testing. Atlanta, GA: Centers for Disease Control and Prevention
15. **Romitelli F, Di Stasio E, Santoro C, Iozzino M, Orsini A, Cesareo R** 2009 A comparative study of fine needle aspiration and fine needle non-aspiration biopsy on suspected thyroid nodules. Endocr Pathol 20:108–113
16. **Nikiforov YE, Ohori NP, Hodak SP, Carty SE, LeBeau SO, Ferris RL, Yip L, Seethala RR, Tublin ME, Stang MT, Coyne C, Johnson JT, Stewart AF, Nikiforova MN** 2011 Impact of mutational testing on the diagnosis and management of patients with cytologically indeterminate thyroid nodules: a prospective analysis of 1056 FNA samples. J Clin Endocrinol Metab 96:3390–3397
17. **Dimech W, Bowden DS, Brestovac B, Byron K, James G, Jardine D, Sloots T, Dax EM** 2004 Validation of assembled nucleic acid-based tests in diagnostic microbiology laboratories. Pathology 36:45–50
18. **Cronin M, Sangli C, Liu ML, Pho M, Dutta D, Nguyen A, Jeong J, Wu J, Langone KC, Watson D** 2007 Analytical validation of the Oncotype DX genomic diagnostic test for recurrence prognosis and therapeutic response prediction in node-negative, estrogen receptor-positive breast cancer. Clin Chem 53:1084–1091