Supplementary Materials and Methods

Patients and samples: De-identified archived samples representing various histologies of benign and malignant tumors were obtained from several sources, as follows: FFPE tissue from resected tumors from Rabin Medical Center, Petach Tikva, Israel, FNA samples preserved as cell blocks from Temple University, Philadelphia, PA, USA, and FNA smears from 1) Sourasky Medical Center, Tel Aviv, Israel (139 samples in training set), 2) Rabin Medical Center, Petach Tikva, Israel (100 samples in training set), 3) Palacky University, Czech Republic (52 samples in training set. 33 samples in validation set, June 2011-January 2015; rate of malignancy: 54.6%), 4) H. Lee Moffitt Cancer Center and Research Institute, FL, USA (26 samples in training set. 66 in validation set, June 2010-November 2013; rate of malignancy: 30.3%), 5) Meir Medical Center, Kfar Saba, Israel (22 samples in training set. 37 in validation set, January 2006-January 2013; rate of malignancy: 35.1%), 6) National BioService LLC, St. Petersburg, Russia (16 samples in training set. 58 in validation set, April 2014-October 2014; rate of malignancy: 13.8%), and 7) The Johns Hopkins University School of Medicine, MD, USA (20 samples in training set. Seven in validation set, August 2003-December 2003; rate of malignancy: 28.6%). Institutional Review Board (IRB) approvals were obtained for all samples in accordance with each institute's institutional review board or IRB-equivalent guidelines; a waiver of consent was obtained from the IRB. The validation set was independent of the training set and included different patients. We did not gather information regarding whether the patients were at initial diagnosis or were returning patients. For FFPE and cell block samples, representative blocks were sectioned into 1.5 ml microcentrifuge tubes.

RNA extraction: Total RNA was extracted from FFPEs and cell blocks as previously described [1 , 2]. Samples were deparaffinized with xylene, washed in ethanol, and digested with proteinase K. RNA was extracted with acid phenol:chloroform followed by ethanol precipitation and DNase. Cytological slides with a cover slip were soaked in xylene prior to extraction. For all cytological slides, the material was scraped from the slide and total RNA extracted as described above. RNA amounts were quantified using the NanoDrop 3300 Fluorospectrometer. In the training set, for several samples, the RNA extracted from 2-3 slides was combined and profiled.

MicroRNA microarrays: Agilent custom-designed arrays containing eight identical subarrays were used (Agilent Technologies, Santa Clara, CA, USA). The arrays included probes for more than 2000 known (miRBase version 19 [3 , 4]) and predicted human miRNAs printed in triplicate, as well as control probes. 40-200ng of total RNA was labeled by ligation of an RNA linker, p-rCrU-Cy/dye (IBA, Germany; Cy3 or Cy5) to the 3' end. Synthetic small RNA controls were spiked before labeling. Slides were incubated with the labeled RNA for 12-16 hours at 55°C and washed according to the Agilent protocol. Arrays were scanned using the Agilent DNA Microarray Scanner Bundle at a resolution of 5 um, dual pass at 100%, and 10% laser power. Array images were analyzed using Agilent Feature Extraction software version 10.7.1.1. Triplicate spots were combined to produce one signal by taking the logarithmic mean of reliable spots. Signals were normalized as previously described [5].

Next generation sequencing: 500ng of RNA from 11 FFPE thyroid follicular resection samples were used for small RNA deep sequencing (miRSeq). Libraries were loaded on two lanes of the Illumina HiSeq2000 DNA sequence analyzer. An average of

approximately 6.3 million reads per library were obtained. To find novel microRNAs, miRDeep2[6] was run on the raw sequencing data (primer-adapter sequences were trimmed).

qRT-PCR: miRNA amounts were quantified using a quantitative Real-Time PCR method described previously [7 , 8]. Briefly, up to 20ng of total RNA was subjected to polyadenylation and reverse transcription reactions. RNA was incubated in the presence of poly (A) polymerase, MnCl2, ATP, oligodT primer and SuperScript II RT (Invitrogen) for one hour at 37°C. A subset of the training samples was used to select the final set of miRNAs and these were assayed for the expression of 96 probes. The C_T (Cycle Threshold) values were calculated based on output from the ABI 7500 software, namely the FAM™ and ROX™ vectors for each qRT-PCR curve. The calculation was performed using in-house software, in a manner similar to the way the ABI 7500 software calculates C_TS. The baseline function was calculated and subtracted from the Rn (FAM™ divided by ROX™) vector. The C_T was defined as the cycle in which the delta Rn vector crosses the pre-defined, miRNA-specific, threshold. References to "expression" throughout the article refer to inverted C_TS (C_T values subtracted from 50), so that high values represent high expression. Training samples were profiled at Rosetta Genomics Israel laboratory (RG-IL) and/or at the Rosetta Genomics US laboratory (RGL-US), a CAP-accredited, Clinical Laboratory Improvement Amendments (CLIA) certified laboratory. All the samples in the validation set were profiled at RGL-US.

Test protocol and classifier: In the final assay, the C_TS of 24 miRNAs are measured. Total RNA is extracted and 20ng are used for creating cDNA. The samples are processed together with two negative controls and a positive control. The negative controls serve to

detect potential contaminations, and the expression of the miRNAs in the negative controls should not be above pre-determined thresholds in the PCR reaction. One is generated for each batch by performing the cDNA reaction in DDW (double distilled water) and the other is generated by performing the RNA extraction process in DDW. The positive control is a specific mix of RNA derived from resected thyroid tumors and synthetic RNA, such that all miRNA in the assay are detected. If either the positive or negative controls fail, the process is repeated. Each sample is profiled in duplicate. Quality control (QC) of each well is based on the fluorescence amplification curve, using thresholds on the maximal fluorescence and the linear slope as a function of the measured C_T. For each miRNA, the C_T value used is the average C_T of the two repeats. If one or both repeats fail, or if the signal difference of the probe duplicate wells exceeds a threshold, data for the sample may be discarded and the sample rerun. Additional QC steps (sample QC) include: 1) sufficient thyroid epithelial marker, and 2) minimal expression thresholds for miRNA.

The classifier combines several Linear Discriminant Analysis (LDA) steps and a K-Nearest Neighbor (KNN) classifier step (K=9), which uses a Pearson correlation distance metric over the C_T values of six of the measured miRNAs, and uses a training set of 314 samples. Samples with at least four benign neighbors are classified as benign by the KNN classifier. Only a subset of the training samples were used to determine the final classifier. For example, samples with low expression for the miRNA in the KNN step were not included in the KNN training set.

Statistical Analyses: 95% confidence intervals were calculated using the Clopper-Pearson method. Sensitivity and specificity estimates for the training cohort were calculated as the mean of 10 10-fold cross-validation runs. For these cross-validation runs, medullary

carcinoma samples and samples with low expression of the assay miRNA were excluded. Medullary carcinoma samples were classified separately. For 2x2 contingency tables, either the χ^2 test or Fisher's exact test (in the case where the conditions for performing a χ^2 test were not met) was used.

Supplementary Results

FNA Cell blocks: In order to assess the performance of the classifier on cell blocks, the classifier was also assessed on a set of 48 FNA cell blocks, which were run in the final assay format. These samples were part of the set used in the discovery stages. The performance of the classifier on these samples (72% sensitivity and 79% specificity) can be seen in Supplementary Table S1, with a slightly lower sensitivity and a slightly higher specificity relative to the FNA smears, however these differences are not statistically significant (p=0.2 and p=0.46, respectively, χ^2 test). The reference diagnosis of these samples, as for other training samples, was not confirmed by additional pathologists. The six cell block Bethesda II and Bethesda VI samples performed well, as all were correctly classified.

Supplementary Table S1. Performance of the assay on FNA cell blocks

	#Malignant	#Benign	Sensitivity[a]	Specificity[a]
All	23	25	78 [56-93]	80 [59-93]
Indeterminate	18	24	72 [47-90]	79 [58-93]
Bethesda II and VI	5	1	100 [48-100]	100 [3-100]

[a]95% Confidence Intervals are in square brackets

Bethesda II and Bethesda VI: The performance of the classifier on benign and malignant samples in the training set (Bethesda II and VI) was estimated to be higher than for indeterminate samples, with a sensitivity of 96% and a specificity of 82% for these samples. In a separate study, performed subsequent to the evaluation of the validation set, the performance of the assay was assessed on a set of 41 benign and malignant FNA smears. This set of samples consisted of nine malignant (Bethesda VI) and 32 benign (Bethesda II) samples. All but one of the malignant samples were correctly classified (89%). The medullary carcinoma sample was correctly classified and demonstrated high expression of hsa-miR-375. Twenty of the benign samples (63%) were correctly classified. As for the training samples and the cell block samples, the diagnosis of these samples was not confirmed by additional pathologists, so these performance values likely underestimate the true performance.

Oncocytic lesions: Since oncocytic FCs and FAs present a diagnostic challenge, we examined the samples for which the resection was marked as such by the referring pathologist. In the training set, there were 41 indeterminate FA and FC samples with oncocytic features (9 Bethesda III, 30 Bethesda IV, two Bethesda V). Six were FC, with an estimated accuracy of 67% and 35 were FA, with an estimated accuracy of 69% (performance from the 10-fold cross-validation results). In the validation set, there were 21 oncocytic FA samples. Of these, 14 (67%) were correctly classified . 74 FAs were not marked as oncocytic lesions and, of these, 58 (78%) were correctly classified. The oncocytic FAs therefore demonstrate a lower accuracy, however this difference was not found to be statistically significant ($p=0.27$, χ^2 test). None of the follicular carcinoma samples in the validation set were marked as having oncocytic features.

References for Supplementary Data

1. Nass D, Rosenwald S, Meiri E, et al. MiR-92b and miR-9/9* are specifically expressed in brain primary tumors and can be used to differentiate primary from metastatic brain tumors. *Brain Pathol* 2009;**19**(3):375-83 doi: 10.1111/j.1750-3639.2008.00184.x [published Online First: 2008/07/16].
2. Rosenfeld N, Aharonov R, Meiri E, et al. MicroRNAs accurately identify cancer tissue origin. *Nat Biotechnol* 2008;**26**(4):462-9 doi: 10.1038/nbt1392 [published Online First: 2008/03/26].
3. Griffiths-Jones S. The microRNA Registry. *Nucleic Acids Res* 2004;**32**(Database issue):D109-11 doi: 10.1093/nar/gkh023 [published Online First: 2003/12/19].
4. Griffiths-Jones S, Grocock RJ, van Dongen S, Bateman A, Enright AJ. miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res* 2006;**34**(Database issue):D140-4 doi: 10.1093/nar/gkj112 [published Online First: 2005/12/31].
5. Meiri E, Mueller WC, Rosenwald S, et al. A second-generation microRNA-based assay for diagnosing tumor tissue origin. *Oncologist* 2012;**17**(6):801-12 doi: 10.1634/theoncologist.2011-0466 [published Online First: 2012/05/24].
6. Friedlander MR, Mackowiak SD, Li N, Chen W, Rajewsky N. miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. *Nucleic Acids Res* 2012;**40**(1):37-52 doi: 10.1093/nar/gkr688.
7. Lebanony D, Benjamin H, Gilad S, et al. Diagnostic assay based on hsa-miR-205 expression distinguishes squamous from nonsquamous non-small-cell lung carcinoma. *J Clin Oncol* 2009;**27**(12):2030-7 doi: 10.1200/jco.2008.19.4134 [published Online First: 2009/03/11].
8. Gilad S, Lithwick-Yanai G, Barshack I, et al. Classification of the four main types of lung cancer using a microRNA-based diagnostic assay. *J Mol Diagn* 2012;**14**(5):510-7 doi: 10.1016/j.jmoldx.2012.03.004 [published Online First: 2012/07/04].