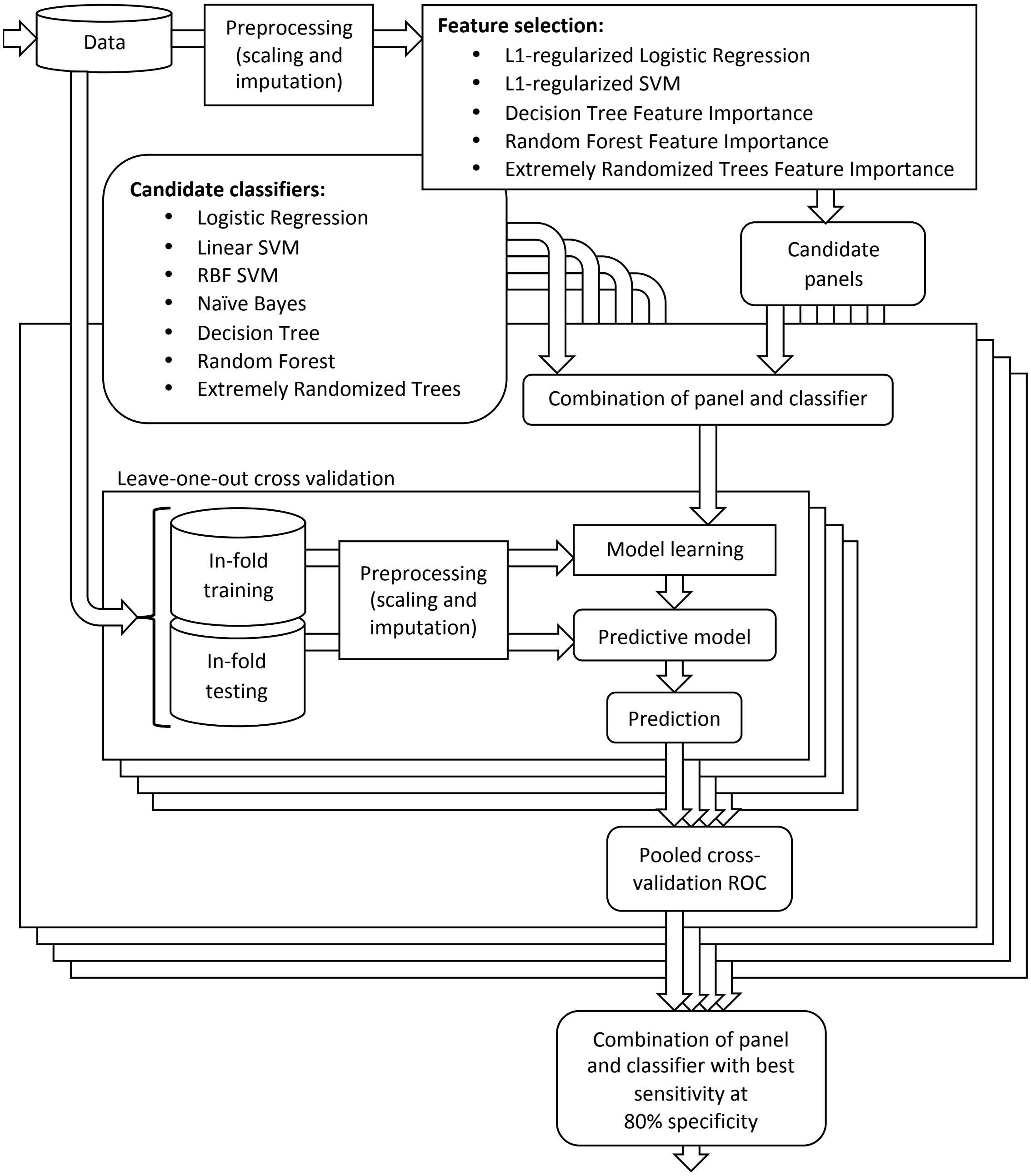
Appendix 1:

Peptide sequences and reference standards targeted for the 17 biomarkers surveyed in this study

Protein Name	Protein Symbol	Endogenous Peptide Sequence	Peptide Reference Standard Sequence*	Dominant Charge State	Peptide Ions Monitored (Qtrap 5500)	Peptide Ion Quantified
CD44 antigen	CD44	YGFIEGHVVIPR	YGFIEGHVVI**P***R	3	y4, y5, y6	y4
Complement Factor I	CFI	VFSLQWGEVK	VFSLQWGE**V***K	2	y5, y6, y7, y8	y5
Leucine-Rich alpha-2-glycoprotein	Lrg1	VAAGAFQGLR	VAAGAFQG**L***R (13C labeled only)	2	y5, y7, y8, y9	y8
Epidermal Growth Factor Receptor	EGFR	IPLENLQIIR	IPLENLQI**I***R (13C labeled only)	2	y5, y6, y7, y8	y7
Inter-alpha-trypsin inhibitor heavy chain H3	Itih3	EVSFDVELPK	EVSFDVE**L***PK (13C labeled only)	2	y5, y6, y7, y8	y5
Coagulation factor V	F5	NFFNPPIISR	NFFNPPI**I***SR (13C labeled only)	2	y6, y7, y8, b8	y6
Hemopexin	Hpx	LWWLDLK	LWWLD**L***K (13C labeled only)	2	y4, y5, y6, b5	y5
Vitamin D-binding protein	Gc	VLEPTLK	VLEPT**L***K (13C labeled only)	2	y4, y5, y6	y4
Inter-alpha-trypsin inhibitor, Heavy chain 4	ITIH4	FAHTVVTSR	FAHTVV**T***SR	3	y3, y4, b3, b4	y3
Serum Amyloid P	APCS	GYVIIKPLVWV	GYVIIKP**L***VWV (13C labeled only)	2	y6, y7, y8, b9	b9
Fetuin B	FetuB	IFFESVYGQCK	IFFESVYGQ**C***K	2	y6, y7, y8, y9	y9
C-reactive protein	Crp	ESDTSYVSLK	ESDTSYVS**L***K (13C labeled only)	2	y5, y6, y8	y6
Sulfhydryl Oxidase 1	QSOX1	LAGAPSEDPQFPK	LAGAPSEDPQF**P***K	2	y5, y7, y9, b4	y9
Peptidase inhibitor 16	PI16	WDEELAAFAK	WDEELAAF**A***K	2	y6, y8, y9	y8
Cadherin-2 (N-Cadherin)	Cdh2	GPFPQELVR	GPFPQE**L***VR (13C Labeled Only)	2	y6, y7, b3	y6
Dipeptidyl peptidase 4	DPP4	WEYYDSVYTER	WEYYDSV**Y***TER	2	y7, y8, y9, b5	y9
extracellular superoxide dismutase [Cu-Zn]	Sod3	VTGVVLFR	VTGVV**L***FR (13C Labeled Only)	2	y4, y5, y6, y7	y6

Bolded* indicates $^{13}C^{15}N$ amino acid unless otherwise indicated

Appendix 2: Detailed pipeline for the use of training cases to select a classification method and panel consisting of 2-5 biomarkers. The data is first used to produce a set of candidate panels, then each panel and each classifier from an array of methods is evaluated using leave-one-out cross validation on the dataset to produce a pooled ROC curve for that panel and classifier combination. The combination that yields the highest sensitivity at 80% specificity is selected.


Data
Preprocessing (scaling and imputation)
Feature selection:
L1-regularized Logistic Regression
L1-regularized SVM
Decision Tree Feature Importance
Random Forest Feature Importance
Extremely Randomized Trees Feature Importance
Candidate classifiers:
Logistic Regression
Linear SVM
RBF SVM
Naïve Bayes
Decision Tree
Random Forest
Extremely Randomized Trees
Candidate panels
Combination of panel and classifier
Leave-one-out cross validation
In-fold training
In-fold testing
Preprocessing (scaling and imputation)
Model learning
Predictive model
Prediction
Pooled cross-validation ROC
Combination of panel and classifier with best sensitivity at 80% specificity

Appendix 3: SUPPLEMENTARY METHOD INFORMATION

Description of measurements

For all target biomarkers, a relative chromatographic area ratio of the biomarker compared to its reference standard was calculated in 3-5 technical replicates per patient. The geometric mean of the replicates' relative area ratios was taken to be the numeric measurement representative of the level of the biomarker for that patient. Relative ratios were generated using internal reference standards that were spiked into each of the patient protein samples in the same amount prior to trypsin digestion. The reference standards were made in lots of approximately 150 trypsin digests per lot. Two lots were used in this analysis and were compared to ensure correct scaling of the data by analyzing ten patient cases with both master mixes. A standard curve was generated comparing the patient pairs with both master mix lots. The linear equation generated from the two compared lots was used to scale the second set of patient data so that it had the same relative internal standard expression as the set of patient data generated from the first lot of material. Hemopexin only utilized the first five patient cases because the very high hydrophobicity of its reference peptide was such that one of the two trypsin digests used for the scaling was not representative of the first digest.

Protein Name	Protein Abbreviation	Equation for Old/New Conversion
Leucine-Rich Alpha-2-Glycoprotein	LRG1	y= 2.065x + 0.1801
Serum Amyloid P	APCS	y= 3.2542x + 2.3711
Fetuin B	FETUB	y = 4.9706x + 0.0248
Inter-Alpha-trypsin inhibitor heavy chain 4	ITIH4	y= 1.4203x - 0.1016
C-Reactive Protein	CRP	y = 1.7692x + 0.017
Epidermal Growth Factor Receptor	EGFR	y= 1.4633x + 0.0165
Inter-Alpha-trypsin inhibitor heavy chain H3	ITIH3	y= 1.3013x + 0.3541
Coagulation Factor V	F5	y= 2.0608x + 0.1706
Hemopexin	HPX*	y= 2.9936x - 1.1312
Vitamin D Binding Protein	VitD	y= 1.2839x - 0.1121
Complement Factor I	CFI	y = 3.076x + 1.5671
CD44	CD44	y= 0.8927x + 0.5184
Peptidyl Peptidase Inhibitor 16	PI16	y= 2.0276x + 0.0919
Superoxide Dismutase 3	SOD3	y = 2.4956x + 0.0408
Quiescin Q6 Sulfhydryl Oxidase 1	QSOX1	y= 1.8174x -0.004
Dipeptidyl Peptidase 4	DPP4	y= 1.1712x + 0.0355
Cadherin 2	CDH2	y= 1.6377x + 0.0034

Development of screening models based on blood protein data

We sought to identify small panels of 2-5 biomarkers that could identify cancer compared to low-risk cancer-free cases and localized cancer (TNM stages 1&2) compared to regional cancer (TNM stage 3). We cast the screening tasks as classification tasks, and evaluated an array of classification methods for each task. We cast the problem of selecting a panel of 2-5 biomarkers as a feature selection problem and used several feature selection methods for identifying candidate panels of biomarkers in a data-driven manner. We use an objective criterion for identifying the best-performing panel and classifier, and perform an evaluation on held-aside data. Figure 1 (main text) illustrates the evaluation process and Appendix 2 illustrates the panel selection process.

Classification methods

For both classification tasks described above we evaluated an array of classification methods: logistic regression, support vector machine classifiers with both linear and radial basis function (RBF) kernels, Gaussian naïve Bayes, the decision tree classifier, the random forest classifier, and extremely randomized trees classifier. We used the implementations from the scikit-learn toolkit (version 0.17.1)[1] respectively linear_model.LogisticRegression, svm.SVC, naive_bayes.GaussianNB, tree.DecisionTreeClassifier, ensemble.RandomForestClassifier, and ensemble.ExtraTreesClassifier. The decision tree, random forest, and extremely randomized trees classifiers used the Gini criterion for impurity[2]. Both the random forest and extremely randomized trees classifiers used ensembles of 100 trees.

Data preprocessing

In using the classification methods above, we use one subset of the data available to us, a training set, to build a predictive model, and use the model to make a prediction about a second subset of the data, a test set, which we use to evaluate the prediction by comparing the prediction about the patient's condition to the observed condition. Since we used leave-one-out cross validation in our analyses, different subsets of the data served as training and test sets, depending on the stage of the analysis. Each time we used a training set and a test set, we preprocessed the biomarker measurements in a manner tailored to the classification method used, and care was taken to never use information from a test set in preprocessing a training set (while one can use information from a training set in preprocessing a test set). For logistic regression and the support vector machine classifier, training data were shifted to zero mean and scaled to unit variance, and the same shift and scaling (estimated from the training data) were applied to test data. For the naïve Bayes classifier, data were log-transformed (natural logarithm). No scaling was used for the tree-based methods. For all methods, wherever particular biomarker measurements were missing, training data were mean-imputed and test data were imputed with the mean of the training data.

Data-driven panel and classifier selection

In order to select small panels of biomarkers (2-5) for screening based on a dataset D which contains measurements from all 17 markers and class labels (patient conditions) corresponding to each of the two classification tasks, we used several feature selection approaches, specifically: L1 regularized logistic regression, L1 regularized support vector machine classifiers with a linear kernel, and feature importance rankings for classification trees, random forests, and extremely randomized trees models. For each panel size of interest, each method was used to produce a candidate panel.

To find a panel of proteins of a given size using each of the L1-regularized methods, we employed a grid search over the values of the regularization constant. We preprocessed the data by mean-imputing the biomarker measurements and shifting and scaling them to zero mean and unit variance over the dataset D. We used the implementations in scikit-learn (linear_model.LogisticRegression and svm.LinearSVC for logistic regression and linear support vector machine respectively), which represent regularization strength using a parameter C that takes values between 0 and 1, with lower values corresponding to stronger regularization and 1 corresponding to no regularization. We varied the C parameter from 0.001 to 1.0 in steps of 0.001. For each setting of C and a given set of training samples, we learned a model and model coefficients. Strong L1 regularization leads to models where all but a few features (biomarkers) have zero coefficients associated with them. We then took the lowest value of C for which the number of nonzero coefficients was equal to the panel size we were interested in, and took the proteins associated with those nonzero coefficients to be our candidate panel.

To find a panel of proteins of a given size using each of the tree-based methods, we considered feature importances. For each method, we preprocessed the dataset D by mean-imputing the biomarker measurements and learned the tree-based model the data. We considered the Gini importance measures of the features in the resultant model. We took the biomarkers with the highest importance to form the candidate panels (e.g. for a panel of size 3, the three biomarkers with the highest Gini importance were used).

Each of the L1-regularized feature selection methods and each of the tree-based feature selection methods above produced a candidate biomarker panel for a given panel size, together producing a set of at most 5 candidate biomarker panels for each panel size. Given these panel candidates and the seven classifiers listed above, we applied leave-one-out cross validation within the dataset D and pooled the predictions of the classifiers across cross-validation folds to produce a receiver operator characteristic (ROC) curve evaluating the performance of each combination of classifier and panel candidate. We took the best combination to be the one that reached the highest sensitivity at 80% specificity or above.

By the process above, we determine a best combination of panel and classifier by our criterion. We then used that panel and classifier combination to learn a classification model from the entire dataset D.

Evaluation pipeline

We evaluated the process described above, of finding a small panel of proteins and associated classifier in terms of how well it generalizes to new data that was not used for training the model. We set aside almost 50% of the data, stratified based on patient class, as validation data, leaving the rest of the data to comprise the training set. For each panel size, within the training set, we ran the panel and method selection process within leave-one-out cross validation folds. This produced a panel, method, and patient class prediction for each fold. The resultant ROC curve represents the performance of our panel and method selection process. We also examined how consistent the panel and method selections were across folds.

For the best performing (in terms of highest sensitivity reached at 80% specificity or above) panel size, we used the panel and method selection process to select a panel and method using the entire training set, and show ROC curves produced when evaluating the resultant model on the evaluation set.

1. Fabian Pedregosa GV, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, Édouard Duchesnay: Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research 12:5, 2011

2. Breiman L, Friedman, J., Olshen, R., Stone, C.: Classification and regression trees. Monterey, Calif., USA, Wadsworth, 1984

Appendix 4: Detailed summary of enrolled patients in the study

Overall Summary:

Group:	# of Cases	Sex (M/F)	Median Age, (range)	Number of Polyps (0/1-2/3+)	Advanced Adenoma Cases	Cancer Stage (1/2/3/Other)	Median Pre-Operative CEA Level, (range)*
Colonoscopy (Non-Cancers)	212	104/108	60, (30-80)	54/107/51	72	N/A	N/A
Colectomy (Cancers)	47	26/21	62, (37-88)	N/A	N/A	22/13/12/2	1.95 (0.7-119)

*Does not include 3 cases with CEA levels <0.5, 6 cases with unknown/not measured levels, and one post-operative only measurement

Colonoscopy/Cancer Free Cases:

Specimen ID	Sex	Age	Race/Eth.	POLYP COUNT	Polyp Count BIN	Pathology Rating	Polyp Location (Colon vs. Rectum)	Polyp Location effect (Ascending, Transverse, Descending, Rectal)	Pathology Summary
21154	M	76	White	1	1-2	Low	Rectal	Rectal	Tubular Adenoma
21155	M	51	White	0	0	Normal	Normal	Normal	Screening normal
21214	F	52	White	1	1-2	Low	Colon	Descending	Tubular Adenoma
21216	F	65	White	1	1-2	Low	Colon	Descending	Hyperplastic polyp
21344	F	73	White	3	3+	Advanced Adenoma	Colon	Ascending & Descending	Non-neoplastic mucosa
21474	M	66	White	3	3+	Low	Colon	Ascending & Descending	Non-dysplastic polyp, tublar adenoma
21478	M	50	Hispanic	1	1-2	Low	Rectal	Rectal	Hyperplastic polyp
21488	M	68	White	2	1-2	Low	Colon and Rectal	Transverse & Rectum	Hyperplastic polyp, tubular adenoma
21489	F	60	White	1	1-2	Advanced Adenoma	Rectal	Rectal	Tubulovillous Adenoma
21527	F	72	White	1	1-2	Low	Colon	Ascending & Descending	Tubular Adenoma
21528	M	60	White	2	1-2	Advanced Adenoma	Colon	Ascending	Sessile serrated adenoma, Tubular adenoma
21529	M	80	Asian	1	1-2	Low	Colon	Descending	Tubular Adenoma
21530	F	65	White	0	0	Normal	Normal	Normal	Screening normal
21684	M	64	White	0	0	Normal	Normal	Normal	Screening normal
21685	F	50	White	0	0	Normal	Normal	Normal	Screening normal
21755	F	50	White	0	0	Normal	Normal	Normal	Screening normal
21756	M	51	White	2	1-2	Advanced Adenoma	Colon and Rectal	Descending & Rectal	Tubular adenoma, tubulovillous adenoma
21836	M	52	White	2	1-2	Low	Colon	Descending	Tubular Adenoma
21837	F	57	Black	1	1-2	Low	Colon	Descending	Tubular Adenoma
21958	F	69	White	0	0	Normal	Normal	Normal	Screening normal
21968	M	50	White	2	1-2	Low	Colon	Ascending & Descending	Hyperplastic polyp
22797	M	52	White	7	3+	Advanced Adenoma	Colon and Rectal	Ascending, Transverse, Descending, Rectum	Hyperplastic polyp, sessile serrated adenoma
22798	M	58	Black	5	3+	Advanced Adenoma	Colon and Rectal	Ascending, Transverse, Rectum	Tubular adenoma, tubulovillous adenoma
22800	M	63	White	1	1-2	Low	Colon	Ascending	Sessile serrated adenoma
22847	M	50	White	0	0	Normal	Normal	Normal	Screening normal
22848	M	67	White	0	0	Low	Colon	Normal	Screening normal
22849	M	61	White	0	0	Normal	Normal	Normal	Screening normal
22869	M	51	White	1	1-2	Low	Colon	Ascending	Tubular Adenoma

Specimen ID	Sex	Age	Race/Eth.	POLYP COUNT	Polyp Count BIN	Pathology Rating	Polyp Location (Colon vs. Rectum)	Polyp Location effect (Ascending, Transverse, Descending, Rectal)	Pathology Summary
22872	F	60	White	1	1-2	Low	Colon and Rectal	Ascending & Rectal	Screening normal
22887	M	55	White	9	3+	Advanced Adenoma	Colon and Rectal	Transverse, Descending, Rectal	Hyperplastic polyp, tubular adenoma
22888	F	68	White	0	0	Normal	Normal	Normal	Screening normal
22891	F	58	White	3	3+	Low	Colon	Ascending, Transverse, Descending	Hyperplastic polyp, tubular adenoma
22892	F	51	White	3	3+	Low	Colon and Rectal	Transverse, Descending, Rectal	Hyperplastic polyp, tubular adenoma
22909	F	55	White	0	0	Normal	Normal	Normal	Screening normal
22910	F	55	White	1	1-2	Advanced Adenoma	Colon	Ascending	Tubular Adenoma
22911	F	63	White	2	1-2	Advanced Adenoma	Colon	Transverse, Descending	mucosal polyp, tubular adenoma
22927	F	64	White	1	1-2	Low	Colon and Rectal	Ascending & Rectal	Hyperplastic polyp
22928	F	63	White	2	1-2	Low	Colon	Ascending	Screening normal
22929	M	63	White	1	1-2	Low	Colon	Ascending	Tubular Adenoma
22987	F	66	White	0	0	Normal	Normal	Normal	Screening normal
22988	F	64	White	0	0	Normal	Normal	Normal	Screening normal
22989	M	50	Asian	1	1-2	Low	Colon	Transverse	Tubular Adenoma
22996	M	68	White	0	0	Normal	Normal	Normal	Screening normal
22997	M	63	White	3	3+	Low	Colon	Ascending & Descending	Tubular Adenoma
23031	F	51	White	0	0	Normal	Normal	Normal	Screening normal
23032	M	51	White	0	0	Normal	Normal	Normal	Screening normal
23051	F	50	White	7	3+	Advanced Adenoma	Rectal	Rectal	Hyperplastic polyp
23059	M	52	White	0	0	Normal	Normal	Normal	Screening normal
23060	F	62	White	2	1-2	Low	Colon	Ascending	Tubular Adenoma
23061	M	56	White	3	3+	Advanced Adenoma	Colon and Rectal	Ascending, Transverse, Rectum	Lymphoid Aggregate
23062	M	60	White	1	1-2	Low	Colon	Ascending	Tubular Adenoma
23076	M	59	White	0	0	Normal	Normal	Normal	Screening normal
23247	M	56	White	2	1-2	Low	Colon and Rectal	Transverse & Rectum	Hyperplastic polyp, tubular adenoma
23248	M	67	White	1	1-2	Low	Colon	Descending	Screening normal
23250	M	56	White	4	3+	Low	Colon	Ascending, Transverse, Descending	Lymphoid nodule, sessile serrated adenoma
23251	M	51	White	3	3+	Low	Colon	Ascending & Descending	Lymphoid nodule
23252	F	58	White	2	1-2	Advanced Adenoma	Colon and Rectal	Descending & Rectal	Tubular Adenoma
23253	M	69	White	0	0	Normal	Normal	Normal	Screening normal
23254	M	66	White	2	1-2	Low	Colon	Ascending & Descending	Lymphoid nodule

Specimen ID	Sex	Age	Race/Eth.	POLYP COUNT	Polyp Count BIN	Pathology Rating	Polyp Location (Colon vs. Rectum)	Polyp Location effect (Ascending, Transverse, Descending, Rectal)	Pathology Summary
30628	M	52	Black	2	1-2	Low	Colon and Rectal	Transverse & Rectum	Tubular Adenoma
30653	F	63	White	4	3+	Advanced Adenoma	Colon	Ascending	Sessile serrated adenoma
30431	F	50	White	0	0	Normal	Normal	Normal	Screening normal
30432	M	63	Black	0	0	Normal	Normal	Normal	Screening normal
30497	F	50	White	1	1-2	Low	Colon	Descending	Hyperplastic polyp
30517	M	50	White	0	0	Normal	Normal	Normal	Screening normal
30629	F	52	White	1	1-2	Low	Colon	Descending	Tubular Adenoma
30651	M	60	Hispanic, American IndiaUnspecifiedlaska Native	2	1-2	Advanced Adenoma	Colon	Transverse, Descending	Hyperplastic polyp, tubular adenoma
30771	M	65	White	6	3+	Advanced Adenoma	Colon	Ascending, Transverse, Descending	Tubular Adenoma
30772	F	70	White	2	1-2	Low	Colon	Ascending	Sessile serrated adenoma, Tubular adenoma
30788	M	60	White	0	0	Undetermined	Undetermined	Undetermined	Screening normal
30807	F	61	White	1	1-2	Advanced Adenoma	Colon	Transverse	Sessile serrated adenoma
30808	F	52	White	0	0	Low	Colon	Ascending	Tubular Adenoma
30918	F	62	White	4	3+	Advanced Adenoma	Colon	Ascending & Descending	Sessile serrated adenoma
30831	M	55	White	3	3+	Low	Colon	Ascending & Transverse	Lymphoid Aggregate, tubular adenoma
30919	F	61	White	3	3+	Low	Colon and Rectal	Ascending & Rectal	Tubular Adenoma
30945	F	60	White	0	0	Normal	Normal	Normal	Screening normal
30967	F	59	White	0	0	Normal	Normal	Normal	Screening normal
30968	M	60	White	0	0	Normal	Normal	Normal	Screening normal
31047	F	50	White	0	0	Normal	Normal	Normal	Screening normal
31089	F	56	White	0	0	Normal	Normal	Normal	Screening normal
31090	F	50	White	0	0	Normal	Normal	Normal	Screening normal
31130	M	68	White	0	0	Normal	Normal	Normal	Screening normal
31131	F	70	White	1	1-2	Advanced Adenoma	Colon	Ascending	Tubulovillous Adenoma
31151	M	64	White	3	3+	Low	Colon	Ascending & Descending	Lymphoid Aggregate, tubular adenoma
31167	M	50	White	2	1-2	Low	Colon	Transverse	Tubular Adenoma
31227	F	42	White	1	1-2	Advanced Adenoma	Colon	Ascending	Sessile serrated adenoma
31248	F	55	Hispanic, American IndiaUnspecifiedlaska Native	0	0	Normal	Normal	Normal	Screening normal

Specimen ID	Sex	Age	Race/Eth.	POLYP COUNT	Polyp Count BIN	Pathology Rating	Polyp Location (Colon vs. Rectum)	Polyp Location effect (Ascending, Transverse, Descending, Rectal)	Pathology Summary
31273	F	62	White	4	3+	Advanced Adenoma	Rectal	Rectal	Hyperplastic polyp
31471	F	51	White	2	1-2	Low	Colon	Ascending	Sessile serrated adenoma
31472	M	30	White	1	1-2	Advanced Adenoma	Colon	Descending	Hyperplastic polyp
31488	F	62	White	2	1-2	Low	Colon	Transverse	Hyperplastic polyp
31489	F	56	White	2	1-2	Advanced Adenoma	Colon	Ascending	Sessile serrated adenoma, Tubular adenoma
31527	F	60	White	9	3+	Advanced Adenoma	Colon	Ascending, Transverse, Descending	Tubular Adenoma
31528	M	72	White	0	0	Normal	Normal	Normal	Screening normal
31532	F	66	White	2	1-2	Advanced Adenoma	Colon	Transverse, Descending	Hyperplastic polyp, tubular adenoma
31530	M	68	White	4	3+	Advanced Adenoma	Colon	Ascending, Transverse, Descending	Sessile serrated adenoma
31571	M	50	White	0	0	Normal	Normal	Normal	Screening normal
31611	M	55	White	1	1-2	Advanced Adenoma	Colon	Transverse	Tubular Adenoma
31612	F	51	White	1	1-2	Low	Colon	Ascending	Tubular Adenoma
31634	F	51	White	0	0	Low	Colon	Normal	Screening normal
31633	F	50	White	1	1-2	Advanced Adenoma	Colon	Ascending & Transverse	Sessile serrated adenoma
31757	F	40	White	0	0	Normal	Normal	Normal	Screening normal
31947	F	50	White	1	1-2	Low	Colon	Ascending	Tubular Adenoma
31948	F	50	White	1	1-2	Low	Colon	Transverse	Tubular Adenoma
31949	M	64	White	0	0	Normal	Normal	Normal	Screening normal
32312	M	54	White	0	0	Normal	Normal	Normal	Screening normal
32391	M	69	White	2	1-2	Low	Colon	Ascending & Transverse	Sessile serrated adenoma, Tubular adenoma
32392	F	50	White	5	3+	Advanced Adenoma	Colon and Rectal	Ascending, Descending, Rectum	Lymphoid Aggregate, tubular adenoma
32393	F	50	White	3	3+	Low	Colon	Ascending & Descending	Hyperplastic polyp, tubular adenoma
32488	M	55	White	0	0	Normal	Normal	Normal	Screening normal
32491	F	53	White	0	0	Normal	Normal	Normal	Screening normal
32535	F	53	White	2	1-2	Advanced Adenoma	Colon	Ascending	Sessile serrated adenoma
32669	F	77	White	1	1-2	Low	Colon	Ascending	Tubular Adenoma
32670	M	69	White	3	3+	Advanced Adenoma	Colon and Rectal	Ascending, Descending, Rectum	Tubular Adenoma
32688	M	51	White	4	3+	Advanced Adenoma	Colon	Transverse, Descending	Tubular Adenoma
32687	F	65	White	3	3+	Advanced Adenoma	Colon	Ascending & Transverse	Tubular Adenoma
30627	F	69	White	3	3+	Advanced Adenoma	Colon	Ascending, Transverse, Descending	Tubular Adenoma
32767	F	56	White	1	1-2	Low	Colon	Descending	Hyperplastic polyp
32951	F	57	White	3	3+	Low	Colon	Ascending & Descending	Tubular Adenoma
32987	M	52	White	5	3+	Low	Colon	Descending	Hyperplastic polyp
32990	M	50	White	1	1-2	Low	Colon	Transverse	Sessile serrated adenoma
32991	M	51	White	2	1-2	Low	Colon	Descending	Hyperplastic polyp, tubular adenoma
33010	M	74	White	3	3+	Low	Colon and Rectal	Ascending & Rectal	Hyperplastic polyp, tubular adenoma

Specimen ID	Sex	Age	Race/Eth.	POLYP COUNT	Polyp Count BIN	Pathology Rating	Polyp Location (Colon vs. Rectum)	Polyp Location effect (Ascending, Transverse, Descending, Rectal)	Pathology Summary
33012	F	56	White	3	3+	Advanced Adenoma	Colon and Rectal	Ascending, Transverse, Rectum	Tubulovillous Adenoma
33011	F	61	White	2	1-2	Low	Colon and Rectal	Ascending & Rectal	Hyperplastic polyp, tubular adenoma
33028	F	71	White	4	3+	Advanced Adenoma	Colon	Transverse	Tubular Adenoma, Traditional serrated adenoma
33029	F	62	White	0	0	Normal	Normal	Normal	Screening normal
33072	F	56	White	0	0	Normal	Normal	Normal	Screening normal
33287	M	55	White	0	0	Normal	Normal	Normal	Screening normal
33291	M	66	White	1	1-2	Low	Colon	Ascending	Tubular Adenoma
33467	F	52	Asian	0	0	Normal	Normal	Normal	Screening normal
33635	F	54	White	2	1-2	Advanced Adenoma	Colon	Ascending & Descending	Hyperplastic polyp, sessile serrated adenoma
33636	M	50	White	2	1-2	Low	Colon	Rectal	Hyperplastic polyp
33637	M	52	White	3	3+	Advanced Adenoma	Colon	Ascending & Descending	Tubular Adenoma
33738	M	56	White	2	1-2	Low	Colon	Descending	Hyperplastic polyp, tubular adenoma
33740	M	64	White	1	1-2	Low	Colon	Descending	Tubular Adenoma
33742	M	52	White	1	1-2	Low	Colon	Descending	Hyperplastic polyp
33797	M	51	White	0	0	Normal	Normal	Normal	Screening normal
33798	F	62	White	0	0	Normal	Normal	Normal	Screening normal
33800	M	66	White	2	1-2	Advanced Adenoma	Colon	Transverse, Descending	Sessile serrated adenoma
33802	M	66	White	0	0	Normal	Normal	Normal	Screening normal
34032 & 30433	F	73	White	3	3+	Advanced Adenoma	Colon	Ascending	Tubular Adenoma
34027	F	52	White	0	0	Normal	Normal	Normal	Screening normal
34073	F	65	White	2	1-2	Low	Colon	Descending	Hyperplastic polyp, sessile serrated adenoma
34075	F	67	White	1	1-2	Low	Colon	Ascending	Tubular Adenoma
34151	M	52	White	2	1-2	Advanced Adenoma	Colon	Ascending & Descending	Hyperplastic polyp, sessile serrated adenoma
34156	F	66	White	1	1-2	Advanced Adenoma	Colon	Descending	Hyperplastic polyp
34175	F	51	White	3	3+	Advanced Adenoma	Colon	Ascending & Descending	Tubular adenoma, tubulovillous adenoma

Specimen ID	Sex	Age	Race/Eth.	POLYP COUNT	Polyp Count BIN	Pathology Rating	Polyp Location (Colon vs. Rectum)	Polyp Location effect (Ascending, Transverse, Descending, Rectal)	Pathology Summary
34188	M	50	White	0	0	Normal	Normal	Normal	Screening normal
34189	M	67	White	3	3+	Low	Colon	Descending & Rectal	Hyperplastic polyp, tubular adenoma
34207	M	67	White	6	3+	Low	Colon	Ascending, Transverse, Descending	Hyperplastic polyp, tubular adenoma
34209	F	55	White	1	1-2	Low	Colon	Descending	Tubular Adenoma
34208	F	68	White	1	1-2	Low	Colon	Ascending	Tubular Adenoma
34239	F	57	White	2	1-2	Advanced Adenoma	Colon	Ascending & Descending	Hyperplastic polyp, sessile serrated adenoma
34244	M	74	White	1	1-2	Low	Colon	Rectal	Hyperplastic polyp
34263	F	52	White	1	1-2	Low	Colon	Rectal	Tubular Adenoma
34262	M	66	White	4	3+	Advanced Adenoma	Colon	Ascending & Descending	Sessile serrated adenoma, Tubular adenoma
34313	F	66	White	2	1-2	Low	Colon	Ascending & Descending	Hyperplastic polyp, sessile serrated adenoma
34311	M	50	White	2	1-2	Advanced Adenoma	Colon	Descending	Tubular Adenoma
34315	F	60	White	0	0	Normal	Normal	Normal	Screening normal
34312	F	56	White	0	0	Normal	Normal	Normal	Screening normal
34314	F	71	White	6	3+	Advanced Adenoma	Colon	Ascending, Transverse, Descending	Lymphoid Aggregate, non-neoplastic mucosa, tubular adenoma
34347	F	62	White	1	1-2	Advanced Adenoma	Colon	Transverse	Sessile serrated adenoma
34367	M	50	White	1	1-2	Low	Colon	Rectal	Hyperplastic polyp
34368	F	62	White	1	1-2	Low	Colon	Transverse	Screening normal
34369	M	60	White	3	3+	Low	Colon	Ascending & Descending	Lymphoid Aggregate, tubular adenoma
34370	F	50	White	1	1-2	Low	Colon	Rectal	Hyperplastic polyp
34375	M	59	White	1	1-2	Low	Colon	Ascending	Tubular Adenoma
34392	F	51	White	4	3+	Advanced Adenoma	Colon	Transverse, Descending, Rectal	Lymphoid aggregate, tubulovillous adenoma, tubular adenoma, hyperplastic polyp
34394	F	70	White	1	1-2	Low	Colon	Descending	Hyperplastic polyp, lymphoid aggregate
34395	F	64	White	2	1-2	Low	Colon	Ascending & Descending	Hyperplastic polyp, tubular adenoma
34408	F	63	White	0	0	Low	Colon	Normal	Screening normal
34409	M	63	White	1	1-2	Low	Colon	Descending	Non-neoplastic mucosa
34410	F	66	White	2	1-2	Low	Colon	Ascending	Tubular Adenoma
34411	F	57	White	0	0	Normal	Normal	Normal	Screening normal
34412	M	51	White	0	0	Normal	Normal	Normal	Screening normal
34413	F	73	White	2	1-2	Low	Colon	Ascending & Transverse	Sessile serrated adenoma, Tubular adenoma

Specimen ID	Sex	Age	Race/Eth.	POLYP COUNT	Polyp Count BIN	Pathology Rating	Polyp Location (Colon vs. Rectum)	Polyp Location effect (Ascending, Transverse, Descending, Rectal)	Pathology Summary
34431	F	61	White	4	3+	Advanced Adenoma	Colon	Ascending & Descending	Tubular Adenoma
34433	M	51	White	2	1-2	Low	Colon	Descending	Hyperplastic polyp, tubular adenoma
30939	M	71	White	1	1-2	Low	Colon	Rectal	Tubular Adenoma
1	M	74	Unspecified	1	1-2	Low	Rectal	Rectal	Tubular Adenoma
2	M	57	Unspecified	1	1-2	Low	Colon	Descending	Tubular Adenoma
3	F	67	Unspecified	1	1-2	Low	Colon	Transverse	Tubular Adenoma
4	M	42	Unspecified	1	1-2	Low	Rectal	Rectal	Tubular Adenoma
5	F	56	Unspecified	2	1-2	Advanced Adenoma	Colon and Rectal	Ascending & Rectal	Hyperplastic polyp, tubular adenoma
6	M	56	Unspecified	2	1-2	Advanced Adenoma	Colon and Rectal	Descending & Rectal	Hyperplastic polyp, tubular adenoma
7	M	60	Unspecified	1	1-2	Advanced Adenoma	Colon	Ascending	Tubulovillous Adenoma
8	F	58	Unspecified	3	3+	Advanced Adenoma	Colon	Descending	Tubular Adenoma
9	M	64	Unspecified	1	1-2	Low	Colon	Ascending	Tubular Adenoma
10	F	57	Unspecified	1	1-2	Advanced Adenoma	Colon	Ascending	Inflammatory polyp
11	M	65	Unspecified	2	1-2	Low	Colon	Transverse, Descending	Tubular Adenoma
12	F	66	Unspecified	2	1-2	Advanced Adenoma	Colon	Ascending	Tubular Adenoma
13	M	66	Unspecified	2	1-2	Advanced Adenoma	Colon	Ascending	Sessile serrated adenoma
14	Unspecified	Unspecified	Unspecified	0	0	Low	Unspecified	Unspecified	Inflammatory polyp
15	F	60	Unspecified	3	3+	Advanced Adenoma	Colon	Ascending & Transverse	Tubular Adenoma
16	M	54	Unspecified	3	3+	Advanced Adenoma	Colon and Rectal	Transverse, Descending, Rectal	Tubular Adenoma
17	M	63	Unspecified	2	1-2	Low	Colon	Ascending & Descending	Hyperplastic polyp, tubular adenoma
18	M	54	Unspecified	2	1-2	Low	Colon and Rectal	Transverse & Rectum	Hyperplastic polyp, tubular adenoma
19	M	63	Unspecified	3	3+	Advanced Adenoma	Colon	Descending	Tubular Adenoma
20	M	59	Unspecified	1	1-2	Advanced Adenoma	Colon	Descending	Tubulovillous Adenoma
21	M	61	Unspecified	2	1-2	Advanced Adenoma	Colon	Ascending & Descending	Tubular adenoma, tubulovillous adenoma
22	M	53	Unspecified	3	3+	Advanced Adenoma	Colon	Ascending	Tubular Adenoma
23	F	76	Unspecified	3	3+	Advanced Adenoma	Colon and Rectal	Ascending, Descending, Rectum	Tubulovillous Adenoma
24	M	58	Unspecified	1	1-2	Advanced Adenoma	Rectal	Rectal	Tubulovillous Adenoma
25	M	64	Unspecified	5	3+	Advanced Adenoma	Colon and Rectal	Ascending, Transverse, Descending, Rectum	Hyperplastic polyp, tubular adenoma
26	M	61	Unspecified	1	1-2	Low	Colon and Rectal	Descending & Rectal	Hyperplastic polyp
27	Unspecified	Unspecified	Unspecified	0	0	Low	Normal	Normal	Screening normal
28	F	63	Unspecified	5	3+	Advanced Adenoma	Colon and Rectal	Ascending, Transverse, Descending, Rectum	Hyperplastic polyp, tubular adenoma
29	F	55	Unspecified	2	1-2	Advanced Adenoma	Colon	Ascending & Transverse	Sessile serrated adenoma
30	F	55	Unspecified	3	3+	Advanced Adenoma	Colon	Ascending	Tubular adenoma, tubulovillous adenoma
32	M	66	Unspecified	1	1-2	Low	Colon	Descending	Tubular Adenoma
33	F	67	Unspecified	2	1-2	Advanced Adenoma	Colon	Ascending & Transverse	Sessile serrated adenoma, Tubular adenoma
34	M	76	Unspecified	2	1-2	Advanced Adenoma	Colon	Ascending & Transverse	Tubular Adenoma
35	M	79	Unspecified	2	1-2	Advanced Adenoma	Colon	Ascending & Transverse	Tubular Adenoma

Colectomy/Cancer Cases

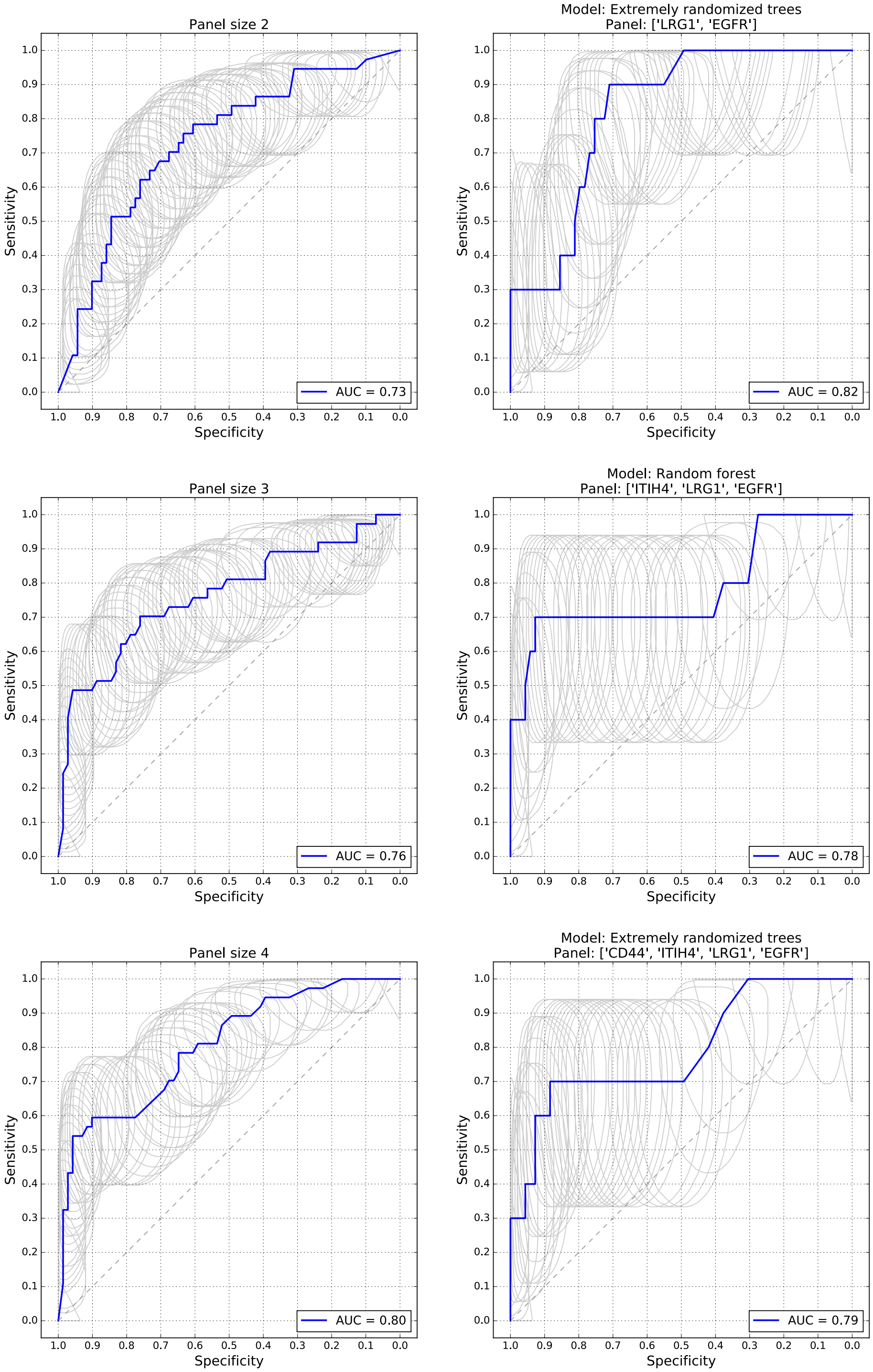
Sample Code (Specimen ID)	Sex	Age	Race/Eth.	T-Stage	N-Stage	Generic stage	Age at diagnosis	Lymphovascular invasion	Neural invasion	Pre-Op CEA Level	Known Familial Cancer Syndrome
21651	M	82	White	T2	N0	Stage 1	82	Present	Absent	1.3	No
22287	F	75	White	T2	N0	Stage 1	75	Absent	Absent	1.9	No
22529	M	56	White	T3	N1C	Stage 3	56	Absent	Absent	5.5	Unspecified
22567	F	78	White	T3	N2A	Stage 3	78	Absent	Absent	3.6	Unspecified
22611	M	55	White	T1	N0	Stage 1	55	Absent	Absent	2.2	Yes-Mother
22707	M	75	White	T3	N0	Stage 2	75	Absent	Absent	2.3	No
22741	F	49	White	T3	N0	Stage 2	49	Not marked	Absent	less than 0.5	No
23311	F	86	White	T3	N0	Stage 2	86	Absent	Absent	2.7	No
23548	F	59	White	T2	N0	Stage 1	59	Absent	Absent	1.7	No
23651	F	72	Black	T3	N0	Stage 2	72	Absent	Absent	6.3	Yes
23891	M	49	American Indian/Alaskan Native	T2	N0	Stage 1	49	Absent	Absent	1.8	No
23996	M	78	White	T3	N0	Stage 2	78	Absent	Absent	Unspecified	No
24447	F	41	White	T2	N1A	Stage 3	41	Absent	Absent	1.29	No
25253	F	86	White	T2	N0	Stage 1	86	Absent	Absent	1.7	Unspecified
25274	F	60	White	T1	N0	Stage 1	60	Absent	Absent	less than 0.5	Yes
25307	M	73	White	T1	N0	Stage 1	73	Absent	Absent	1.3	No
25449	F	72	White	T3	N0	Stage 2	72	Absent	Absent	1.5	No
25527	M	50	White	T1	N0	Stage 1	50	Absent	Absent	1.7	Yes
25628	M	74	White	T4a	N1A	Stage 3	74	Absent	Absent	1.9	No
25873	M	74	White	T3	N0	Stage 2	74	Present	Present	2	Unspecified
27876	F	88	White	T3	N0	Stage 2	88	Absent	Absent	43	Unspecified
27871	F	53	White	T3	N1C	Stage 3	53	Absent	Absent	119	Unspecified
27711	M	47	White	T2	N2A	Stage 3	47	Present	Absent	1.3	No
27671	M	64	Black	T1	N0	Stage 1	64	Absent	Absent	6.4	Yes
27616	F	79	White	T3	N0	Stage 2	79	Absent	Absent	Unspecified	Yes
27431	M	74	White	T3	N1A	Stage 3	74	Absent	Absent	1.2	No
27213	M	65	White	T3	N0	Stage 2	65	Absent	Absent	2	No
27067	M	57	White	T3	N0	Stage 2	57	Present	Absent	4.2	Yes

Sample Code (Specimen ID)	Sex	Age	Race/Eth.	T-Stage	N-Stage	Generic stage	Age at diagnosis	Lymphovascular invasion	Neural invasion	Pre-Op CEA Level	Known Familial Cancer Syndrome
26747	F	68	White	N/A	N/A	Neuroendocrine Tumor	68	Unspecified	Unspecified	Unspecified	Unspecified
26727	M	69	White	T2	N0	Stage 1	69	Absent	Absent	0.9	Yes
26689	F	71	White	T1	N0	Stage 1	71	Unspecified	Unspecified	1.8	Yes
26447	F	61	White	T3	N2B	Stage 3	61	Present	Absent	1.9	No
26131	F	75	White	T2	N0	Stage 1	75	Absent	Absent	0.7	No
26108	M	87	White	T2	N0	Stage 1	86	Absent	Absent	3.5 (post op)	Unspecified
26067	M	56	White	T2	N0	Stage 1	56	Absent	Absent	Not uploaded	Yes
28112	M	87	White	T2	N0	Stage 1	86	Absent	Absent	Unspecified	No
28753	M	37	White	T2	N1b	Stage 3	37	Absent	Absent	Unspecified	No
28281	F	62	White	T2	N0	Stage 1	62	Present	Absent	1.1	No
28007	M	61	White	T3	N0	Stage 2	61	Absent	Absent	3.1	No
28907	M	38	White	T2	N0	Stage 1	38	Present	Absent	1	No
29527	F	44	White	T2	N0	Stage 1	44	Absent	Absent	<0.5	No
29992	M	55	White	T3	N1A	Stage 3	55	Absent	Absent	3.2	No
29588	M	67	White	T3	N2B	Stage 3	67	Present	Absent	2.5	No
30707	M	59	White	T1	N0	Stage 1	59	Absent	Absent	2	No
30920	F	58	White	T2	N0	Stage 1	58	Absent	Absent	1	First Cousin
30871	F	50	White	T2	N0	Stage 1	50	Absent	Absent	27.5	Yes
30887	M	39	White	pT4	pN1	Stage 3	35	Unspecified	Unspecified	2.6	No
31207	F	61	White	ypT0	ypN0	Post Chemoradiation	60	Absent	Absent	2.2	No
31971	F	47	White	T3	N0	Stage 2	47	Absent	Absent	23.7	Unspecified

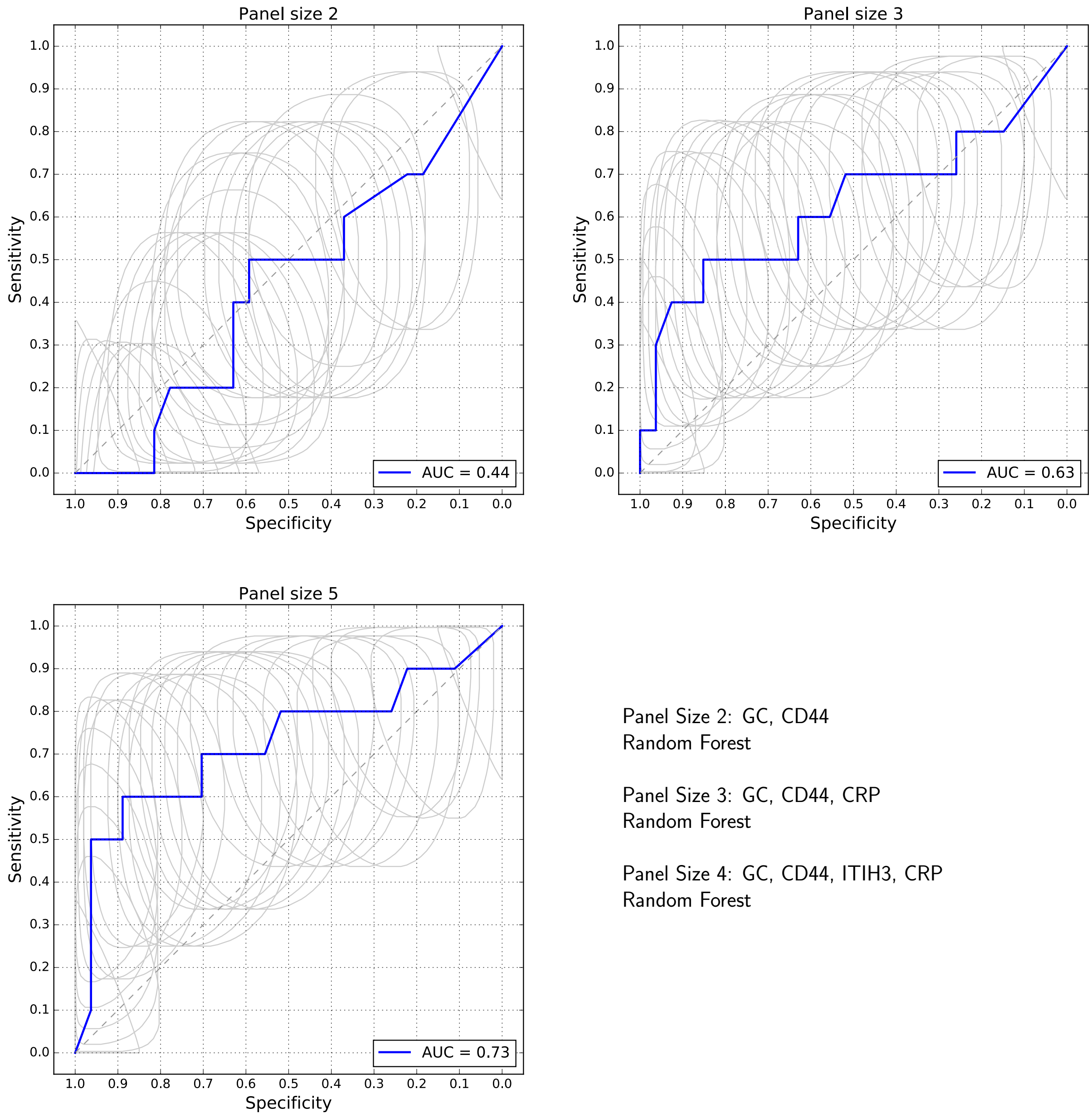
Appendix 5: Relative area ratio-to-reference standard for patient groups compared in the study.

	Relative Ratio-To-Reference Standard									
Protein	Cancer Median Expression	Non-Cancer Median Expression	Advanced Adenoma Expression	Normal and Non-advanced Adenoma Expression	Median Local Cancer Expression (Stages 1&2)	Median Regional Cancer Expression (Stage 3)	Median Female Non-cancer expression	Median Male Non-cancer Expression	Median Female CANCER expression	Median Male CANCER Expression
FETUB	3.85	4.77	4.565	4.83	3.78	3.855	4.87	4.57	3.74	3.855
EGFR	0.0451	0.050319	0.04969	0.05091	0.0451	0.0447	0.0496	0.0518	0.0451	0.0451
ITIH3	2.46	1.31	1.2481	1.34	2.54	1.8	1.29	1.345	2.86	2.07
LRG1	4.28	2.85	2.6974	2.96	4.28	4.26	2.97	2.72440482	4.46	4.11
ITIH4	2.21	1.66	1.65	1.68	2.21	2.17	1.72	1.639732	2.21	2.17
CD44	0.6	0.68	0.6725	0.69	0.65	0.47	0.67	0.688042	0.62	0.6
CRP	0.88	0.495	0.61	0.42	0.77	1.11	0.42	0.53	0.88	0.92
VITD	1.44	1.39	1.3783	1.3959	1.41	1.53	1.38	1.39842815	1.43	1.44
HPX	10.36	6.99	6.9022	7.0377	10.49	10.15	6.92	7.03237614	9.79	10.73
CFI	4.51	4.64	4.66289	4.62218	4.49	5.11	4.64	4.655955	4.11	5.23
APCS	6.49	6.73	6.9335	6.6372	6.45	7.27	6.18	7.54893619	5.47	9.22
F5	0.95	0.85	0.8654	0.84	1.02	0.87	0.83	0.87	0.95	0.95
SOD3	0.15	0.159	0.1593	0.1595	0.15	0.14	0.17	0.15	0.18	0.14
PI16	0.35	0.45	0.4519	0.4549	0.33	0.385	0.47	0.43673113	0.33	0.4
QSOX1	0.22	0.177	0.1715	0.1773	0.231	0.204	0.16536987	0.1925	0.23	0.22
DPP4	0.059	0.07705	0.0751	0.07759	0.0546	0.0599	0.07613524	0.07766597	0.056	0.06
CDH2	0.024	0.026	0.025	0.02613	0.025	0.0225	0.024	0.02746199	0.025	0.024

Appendix 6: ROC curves for Cancer vs. Cancer-free for 2-4 marker panels. Curves on the left represent the pooled ROC generated from the training data using leave-one-out cross validation. Curves on the right represent validation ROCs for the specified model and listed biomarkers.



Appendix 7: ROC curves for Localized Cancer vs. Regional for 2, 3, and 5 marker panels. Curves represent the pooled ROC generated from the training data using leave-one-out cross validation. The optimal method and panel selected for validation is listed.



Panel Size 2: GC, CD44
Random Forest

Panel Size 3: GC, CD44, CRP
Random Forest

Panel Size 4: GC, CD44, ITIH3, CRP
Random Forest