### Supplemental Figure S1. Gene promoter methylation amplification curves

Examples of gene promoter amplification curves for the each one of the three samples comparing the studied gene and B-actin. On the top plot B-actin is represented by the purple curves showing a sigmoidal curve with positive cycle threshold (Ct) of 24 and the green curve represents SOX17 with a Ct of 31 in all of the three replicates respectively. On the bottom plot B-actin (purple curve) had a Ct of 29 in the three samples and CDO1 (orange curve) had a late positive sigmoidal amplification Ct of 38 in two of the replicates and a negative amplification curve without sigmoidal shape in the other one with Ct of 43.

### Supplemental Figure S2 Receiver operator classification curves for lung cancer detection for the genes obtained from sputum

### Supplemental Figure S3 Receiver operator classification curves for lung cancer detection for the genes obtained from Plasma

### Supplemental Figure S4 Variable importance plot for random forest prediction

The plot details the relative importance of each of the variables to the model’s accuracy (including: methylation μ 2-ΔCT values, nodule size, age, pack-year, COPD status and FVC values). The x-axis is the mean decrease in the Gini co-efficient that results when that variable is included in the model. The Gini coefficient is a measure of inequality among the trees in the random forest, and in this case represents the performance of the random forest model with and without a variable included. Those variables that have the highest decrease in the Gini coefficient were most likely to create consensus among the individual decision trees used in the model (or reduce inequality) when included in the model. These variables are therefore most predictive of the outcome of the model overall. Those variables with a small decrease in the mean Gini coefficient are relatively less important to the prediction made by the random forest model.

### Supplementary Table S1 Primer and probe sequences.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Gene** | **Forward 5' - 3'** | **Reverse 5' - 3'** | **Probe 5'** | **Size (bp)** | **Temperature** |
| SOX17 | TTGGATTGGGACGTGGGATTCG | GAAAACGAACCGATCCCGCG | GTGGGTTTAACGACGCGGGATCGGTTCG | 138 | 65 |
| TAC1 | CGGTTAATTAAATATTGAGTAGAAAGTCGCG | CTCCGCACTCTCGATAACTACCG | GGAGAATGTTACGTGGGTTTGGAGGTTTAAGGAGG | 153 | 65 |
| HOXA7 | GTGGTTTCGTTTCGTAGGGTTCG | CCGCCAACTAAACGCTCCC | GCGGAAAAAGATTTGGAGGTTTCGCGGGAG | 98 | 65 |
| CDO1 | CGGAGGCGGGGAGATTTTGCG | CGAAACCCACGCGATCCCTAAAACG | TTTCGTTGTTTTCGGCGTTTTAGGGATCGCGTGGG | 116 | 65 |
| HOXA9 | AGTTATATAGGTTGGCGGGAAGTCG | TCGCCTCTCCCGAAAACC | AGGTTAGATTGTTGTGTTTGGTTGGCGAGTT | 92 | 65 |
| ZFP42 | TTCGGGTTGAGGGTGAGCG | CGACCCCGCCCTAAAACG | CGTCGTTTAGGTGTTAGGCGGTTTCG | 104 | 65 |
| B-ACTIN | TAG GGA GTA TAT AGG TTG GGG AAGTT | AACACACAATAACAA ACA CAA ATT CAC | CGA CTG CGT GTG GGG TGG TGA TGG AGG AGG TTT AGG CAG TCG | 103 | 65 |

### Supplemental Table S2. Gene Methylation Sensitivity, Specificity, at optimal cutoffs with AUC using Sputum and Plasma.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Sputum** | **Sensitivity** | **Specificity** | **PPV** | **NPV** | **AUC** | **95% CI** |  |
|  | SOX17 | 84% | 88% | 96% | 59% | 0.84 | (0.75 -0.94) |  |
|  | TAC1 | 84% | 79% | 94% | 57% | 0.84 | (0.74 - 0.94) |  |
|  | HOXA7 | 63% | 92% | 97% | 40% | 0.77 | (0.67 - 0.86) |  |
|  | CDO1 | 78% | 67% | 90% | 45% | 0.70 | (0.57 - 0.84) |  |
|  | HOXA9 | 77% | 42% | 83% | 32% | 0.56 | (0.41 - 0.69) |  |
|  | ZFP42 | 88% | 62% | 90% | 58% | 0.73 | (0.60 - 0.87) |  |
|  | TAC1, HOXA7, SOX17 | 93% | 79% | 94% | 75% | 0.89 | (0.80 - 0.98) |  |
|  |  |  |  |  |  |  |  |  |
|  | **Plasma** | **Sensitivity** | **Specificity** | **PPV** | **NPV** | **AUC** | **95% CI** |  |
|  | SOX17 | 71% | 86% | 93% | 54% | 0.78 | (0.70 - 0.86) |  |
|  | TAC1 | 76% | 78% | 90% | 57% | 0.78 | (0.70 - 0.86) |  |
|  | HOXA7 | 33% | 94% | 93% | 36% | 0.60 | (0.51 - 0.69) |  |
|  | CDO1 | 65% | 74% | 86% | 46% | 0.68 | (0.58 - 0.77) |  |
|  | HOXA9 | 81% | 52% | 81% | 52% | 0.62 | (0.52 - 0.73) |  |
|  | ZFP42 | 81% | 58% | 83% | 55% | 0.66 | (0.56 - 0.75) |  |
|  | CD01, TAC1, SOX17 | 91% | 64% | 86% | 74% | 0.77 | (0.68 - 0.86) |  |
|  |  |  |  |  |  |  |  |  |

*Abbreviations*: area under the curve (in the ROC curves): AUC, 95 % confidence interval: 95% CI.

### Supplemental Table S3. Performance for lung cancer diagnosis of the independent blinded random forest prediction models on the testing subset

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | **Sensitivity** | **Specificity** | **PPV** | **NPV** | **AUC** | **95% CI** |  |
|  | Prediction from Sputum | 93% | 86% | 96% | 75% | 0.85 | 0.59-1 |  |
|  | Prediction from Plasma | 93% | 67% | 87% | 80% | 0.89 | 0.79-0.99 |  |
|  | Clinical Predictors alone | 84% | 26% | 75% | 38% | 0.64 | 0.50-0.78 |  |
|  |  |  |  |  |  |  |  |  |

*Abbreviations*: area under the curve (in the ROC curves): AUC, 95 % confidence interval: 95% CI.

### Supplemental Table S4. Baseline Characteristics of the 155 Smokers Subjects (Former and current).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Patient Characteristics** | **Cancer**  **(N=114)** | **Control**  **(N=41)** | ***p* Value** |  |
|  | Age at surgery (years) (IQR) | 67 (62-75) | 63 (55-73.25) | 0.007 |  |
|  | Gender |  |  |  |  |
|  | Male (%) | 48 (42%) | 26 (63%) | 0.03 |  |
|  | Female (%) | 66 (58%) | 15 (37%) |  |
|  | Race |  |  |  |  |
|  | White (%) | 94 (83%) | 36 (88%) | 0.06 |  |
|  | Black (%) | 15 (13%) | 2 (5%) |  |
|  | Other (%) | 5 (4%) | 3 (7%) |  |
|  | Stage |  |  |  |  |
|  | IA-IB (%) | 103 (90%) | NA | NA |  |
|  | IIA (%) | 11 (10%) | NA |  |
|  | Histology |  |  |  |  |
|  | Adenocarcinoma (%) | 87 (76%) | NA | NA |  |
|  | Squamous-cell (%) | 24 (21%) | NA |  |
|  | Adenosquamous (%) | 3 (3%) | NA |  |
|  | Smoking status |  |  |  |  |
|  | Current (%) | 27 (24%) | 7 (17%) | 0.5 |  |
|  | Former (%) | 87 (76%) | 34 (83%) |  |
|  | Never (%) | 0 (0%) | 0 (0%) |  |
|  | Pack-year (IQR) | 30 (10-50) | 20 (0-35) | 0.01 |  |
|  | COPD (%) | 34 (32%) | 10 (28%) | 0.83 |  |
|  | FEV1 % Predicted (IQR) | 84 (70-99) | 85 (70-100) | 0.86 |  |
|  | FVC % Predicted (IQR) | 91 (80-103) | 87 (80-110) | 0.68 |  |
|  | FEV1/FVC % Ratio (IQR) | 73 (68-77) | 77 (70-80) | 0.07 |  |
|  | Nodule size (cm) | 2 (1.5-7.5) | 1.5 (1-3) | 0.01 |  |
|  | < 1cm | 4 (3%) | 12 (30%) | <0.001 |  |
|  | 1-2 cm | 37 (33%) | 13 (32%) |  |
|  | > 2 cm | 73 (64%) | 16 (39%) |  |
|  | Nodule volume (cm3) | 17.9 (1.8-14.1) | 1.6 (0.5-7.9) | <0.001 |  |
|  |  |  |  |  |  |

*Abbreviations*: Chronic obstructive pulmonary disease: COPD, Forced Expiratory Volume in one second: FEV1, Forced vital capacity: FVC, Interquartile range: IQR.

Nodule size % <1cm, 1-2, >2cm

### Supplemental Table S5. Gene Methylation Sensitivity, Specificity, AUC and Association with Cancer Diagnosis for genes obtained from Sputum and Plasma in Only Smokers.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Sputum** | **Sensitivity** | **Specificity** | **PPV** | **NPV** | **AUC** | **95% CI** |  |
|  | SOX17 | 79% | 85% | 95% | 55% | 0.81 | (0.70 – 0.92) |  |
|  | TAC1 | 81% | 85% | 95% | 57% | 0.84 | (0.73 – 0.95) |  |
|  | HOXA7 | 67% | 85% | 94% | 44% | 0.76 | (0.65 – 0.88) |  |
|  | CDO1 | 91% | 45% | 85% | 60% | 0.66 | (0.50 – 0.81) |  |
|  | HOXA9 | 72% | 45% | 81% | 32% | 0.55 | (0.41 – 0.70) |  |
|  | ZFP42 | 81% | 70% | 90% | 52% | 0.76 | (0.62 – 0.90) |  |
|  | TAC1, HOXA7, SOX17 | 94% | 80% | 94% | 80% | 0.89 | (0.79 – 0.99) |  |
|  |  |  |  |  |  |  |  |  |
|  | **Plasma** | **Sensitivity** | **Specificity** | **PPV** | **NPV** | **AUC** | **95% CI** |  |
|  | SOX17 | 71% | 84% | 93% | 49% | 0.78 | (0.68 – 0.88) |  |
|  | TAC1 | 78% | 78% | 91% | 54% | 0.80 | (0.70 – 0.90) |  |
|  | HOXA7 | 34% | 91% | 92% | 32% | 0.58 | (0.47 – 0.68) |  |
|  | CDO1 | 67% | 84% | 93% | 46% | 0.76 | (0.66 – 0.86) |  |
|  | HOXA9 | 89% | 53% | 85% | 61% | 0.66 | (0.53 – 0.79) |  |
|  | ZFP42 | 82% | 60% | 86% | 53% | 0.69 | (0.58 – 0.80) |  |
|  | CDO1, TAC1, SOX17 | 93% | 72% | 91% | 77% | 0.85 | (0.76 – 0.94) |  |
|  |  |  |  |  |  |  |  |  |

*Abbreviations*: Positive predictive value: PPV, Negative Predictive Value: NPV, area under the curve (in the ROC curves): AUC, 95 % confidence interval: 95% CI.

### Supplemental Table S6. Greatest Transverse Diameter (cm) in Methylated Vs Unmethylated patients.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Sputum** | **Methylated** | **Number** | **Unmethylated** | **Number** | ***p* Value** |
| SOX17 | 2.63 | 76 | 2.30 | 14 | 0.57 |
| TAC1 | 2.47 | 77 | 3.04 | 13 | 0.13 |
| HOXA7 | 2.57 | 57 | 2.57 | 33 | 0.76 |
| CDO1 | 2.52 | 70 | 2.74 | 20 | 0.79 |
| HOXA9 | 2.57 | 84 | 2.53 | 6 | 0.91 |
| ZFP42 | 2.62 | 78 | 2.23 | 12 | 0.57 |
|  |  |  |  |  |  |
| **Plasma** | **Methylated** | **Number** | **Unmethylated** | **Number** | ***p* Value** |
| SOX17 | 2.74 | 91 | 2.38 | 34 | 0.54 |
| TAC1 | 2.49 | 95 | 3.01 | 30 | 0.53 |
| HOXA7 | 2.70 | 42 | 2.60 | 83 | 0.32 |
| CDO1 | 2.54 | 81 | 2.79 | 44 | 0.40 |
| HOXA9 | 2.69 | 108 | 2.30 | 17 | 0.34 |
| ZFP42 | 2.60 | 105 | 2.80 | 20 | 0.57 |