Supplemental Information:

Supplemental Methods

Supplemental Tables:

Supplemental Figures:

Supplemental Datasets: see separate excel files

Supplemental Methods

Software versions referenced

Data Processing

Illumina CASAVA v1.8.2
TopHat v2.0.4
RSeQC v2.3.3
HTSeq-count v0.5.4
R v3.0.0
edgeR v3.4.2
RSEM v1.2.1
Bowtie v1.0.0

Data Analysis

Limma v3.18.13
edgeR v3.4.2
sva v3.6.0
GSVA v1.10.3

Gene expression-based prediction of smoking status

Microarray data from Beane *et al*. [1] Gene Expression Omnibus [GEO] [2] Accession Number GSE7895) was re-analyzed using Robust Multi-array Average (RMA) [3] and the Ensembl CDF file v16.0.0 file (http://brainarray.mbni.med.umich.edu/Brainarray/Database/CustomCDF/16.0.0/ensg.asp). The R package was used to identify genes differentially expressed between current (n=52) and never (n=21) smokers, using the linear model presented in the paper additionally correcting for quality covariates (NUSE and RLE). Ninety-four genes (FDR<0.001) were differentially expressed between current and never smokers. The weighted voting algorithm [4] was trained on z-score normalized microarray data (n=73) across the 94 genes and used to predict smoking status in z-scored log2-transformed counts per million (cpm) from the 82 mRNA-Seq samples.

Processing of Publically Available Datasets

Cancer Cell Line Compendium (CCLE). The Entrez ID gene expression file labeled 10/18/2012 and the sample information file were downloaded from CCLE website (http://www.broadinstitute.org/ccle/home). After matching the sample annotation to the expression file, we used ComBat [5] to adjust the data for batch effects (n=14 batches across 1019 samples). After batch correction, the lung cell lines (n=186) were selected and GSVA was used to calculate a pathway enrichment score for each lung cell line for the following pathways: KEGG oxidative phosphorylation, KEGG glycolysis gluconeogenesis, BioCarta glycolysis, and Reactome glycolysis. The GSVA scores for the glycolysis pathways were averaged per sample.

The Cancer Genome Atlas (TCGA). RSEM gene-level (Entrez IDs) counts derived from RNA-Seq data were downloaded from the TCGA data portal on August 27, 2013 for lung squamous cell carcinomas and adjacent matched control tissue (n=100 samples from n=50 subjects). After

applying the mixture model referenced in the paper, 14,178 out of 20,531 genes were expressed as signal in at least 15% of samples (n=15). Differential gene expression between tumor and adjacent normal tissue was determined using limma and voom-transformed data [6, 7] via a linear model with cancer status as the main effect and a random patient effect modeled using the duplicateCorrelation function. Gene sets containing the top 200 up- and down-regulated differentially expressed genes associated with cancer status were used as input for GSEA.

Microarray Data. CEL files for GSE19188 and GSE18842 were downloaded from GEO and processed using Robust Multi-array Average (RMA) [3] and the Ensembl Gene CDF v16.0.0 file (http://brainarray.mbni.med.umich.edu/Brainarray/Database/CustomCDF/16.0.0/ensg.asp).
Samples with a median RLE greater than 0.1 or a median NUSE greater than 1.05 were excluded, yielding n=146 samples for GSE19188 and n=82 samples for GSE18842. For GSE19188, differential gene expression between squamous cell tumors (n=23) and normal lung tissue (n=64) was conducted using limma and a linear model that included RLE and NUSE covariates. For GSE18842, paired normal and tumor tissue from the same subjects (n=37 subjects, n=74 samples) were selected, and differential gene expression was conducted in an analogous manner as described above for TCGA, additionally correcting for RLE and NUSE metrics.

CEL files for GSE4115 were processed using RMA and the CDF file above. The n=164 samples described in Spira *et al*. [8], were used to determine genes differentially expressed in airway brushings from subjects with and without lung cancer, using limma and a linear model with terms for cancer status, RLE, NUSE, smoking status, and pack-years. Gene sets containing the top 200 up- and down-regulated differentially expressed genes associated with cancer status were used as input for GSEA.

Biomarker Development
Upstream gene filtering. In order to provide cross-platform compatibility, we ran the biomarker discovery and validation pipelines using 11,926 genes commonly present on the RNA-Seq platform (Illumina HiSeq 2500 used with Ensembl v64 GTF) and two microarray platforms (Affymetrix GeneChip Human Gene 1.0 ST Array used with custom ENSG *Homo sapiens* CDF from Brainarray v11 and Affymetrix Human Genome U133A Array used with custom ENSG *Homo sapiens* CDF from Brainarray v16).

Data generation and summarization. Samples (n=75) were run across 4 flow cells (4 batches), and samples run in batches 1, 2, and 3 (n=58) were assigned to a discovery set, while the remaining samples (n=17) were used as an independent validation set and not included in the biomarker development. Alignments and gene level summarization were conducted as described in the paper methods. Alignment and quality metrics were calculated using RSeQC (v2.3.3) [9]. Using the gene body measure computed by RSeQC, a ratio between the average read coverage at 80% of the gene length and the average coverage at 20% of the gene length was derived as an additional quality metric (gb-ratio) to assess 3' bias per sample. The metric was highly correlated

with a surrogate variable applied in the identification of differentially expressed genes, and was used as a quality control metric in the biomarker pipeline.

Biomarker discovery pipeline. The biomarker discovery pipeline has been outlined generally in the main text. A graphical representation of data flow as well as processing and analysis steps is provided in Supplementary Figure 3. Each computational step outlined is detailed in the following sections.

Balancing signature. We tested gene signatures consisting either of an equal or unequal number of genes up- and down-regulated in subjects with dysplastic lesions.

Input data preprocessing. We tested 3 input data types. HTSeq-count (v0.5.4) [10] was used to derive gene count estimates (raw counts). In addition, Cufflinks (v2.0.2) [11] was used to derive reads per kilobase per million mapped reads (RPKM) using BAM files containing only properly paired reads. We also calculated log2-transformed counts per million (CPM) by applying edgeR (v3.8.6) [12] to raw counts using the "TMM" method (weighted trimmed mean of M-values [13]).

Gene filtering. Signal-based gene filtering was conducted as described in detail in the Methods. In short, a gene was included in downstream analyses if the mixture model classified it as "on" in at least 1%, 5%, 10% or 15% of the samples. For CPM input data type, we recalculated CPM values using raw counts after filtering out genes.

Feature selection. To identify genes differentially expressed (DE) between samples with and without premalignant lesions (PMLs), we applied several algorithms to our filtered dataset. The algorithms used were as follows:

1. edgeR: We applied the edgeR package (v3.8.6) [12] to raw counts only. After calculating normalization factors (calcNormFactors) and estimating common (estimateGLMCommonDisp) and tagwise (estimateGLMTagwiseDisp) dispersion factors, we identified DE genes associated with the presence of PMLs using a generalized linear model, correcting for sex, COPD status, and smoking status covariates. For balanced signatures, the sign of the log2-fold change of expression between conditions determined gene directionality. For all models regardless of balancing, gene importance was defined by FDR-adjusted p-value from likelihood ratio tests (glmLRT).
2. edgeRgb: We used the edgeR package as described in #1, additionally correcting for gb-ratio (described in *Data generation and summarization* section).
3. lm: We applied the limma package (v3.22.7) [6] to CPMs, RPKMs, or voom-transformed raw counts [7]. Voom transformation was applied using a linear model, adjusting for sex, COPD status, and smoking status covariates, after calculating normalization factors. We used the same model to identify DE genes associated with the presence of PMLs. For balanced signatures, the sign of the moderated t-statistic obtained via eBayes and topTable

determined gene directionality. For all models regardless of balancing, gene importance was defined by the magnitude of the t-statistic.

4. lmgb: We used the limma package as described in #3, additionally correcting for gb-ratio (described in *Data generation and summarization* section).
5. glmnet: We applied the glmnet package (v1.9-8) [14] to CPMs, RPKMs, or voom-transformed raw counts (as in #3) to identify DE genes associated with the presence of PMLs. For balanced signatures, gene directionality was determined by the sign of the t-statistic obtained via limma by running a linear model described in #3. We carried out the following series of steps using all genes for unbalanced signatures and separately using up- and down-regulated genes for balanced signatures: First, RPKMs and CPMs were z-score normalized, while raw counts were voom-transformed. Then, due to the binary character of our response variable (dysplasia status), a logistic regression model was fit using the binomial distribution family and elastic net mixing parameter $\alpha = 0.5$ (indicating a tradeoff between ridge and lasso regressions). The standardize option was set to FALSE, causing the coefficients to be returned on the original scale, thus allowing their magnitude to be interpreted as gene importance. Next, a range of regularization parameters λ was generated via leave-one-out cross-validation (nfolds = 46), and the λ giving the minimum mean cross-validated error (lambda.min) was chosen to estimate the coefficients. Finally, DE genes were defined as having non-zero coefficients and then sorted by importance based on the coefficients' magnitude.
6. randomForest: We applied the randomForest package (v4.6-12) [15] to CPMs, RPKMs, and voom-transformed raw counts (as in #3), setting the number of trees (ntree) to 100 and importance to TRUE. For balanced signatures, the sign of the t-statistic as described in #5 determined gene directionality. For all models regardless of balancing, gene importance was determined by the magnitude of the importance variable, defined as the mean decrease in accuracy over both conditions.
7. DESeq: We applied the DESeq package (v1.18.0) [16] to unmodified raw counts only. DE analysis to find genes associated with the presence of PMLs included data normalization (estimation of the effective library size), variance estimation, and inference for two experimental conditions, as outlined in the DESeq package vignette (https://www.bioconductor.org/packages/3.3/bioc/vignettes/DESeq/inst/doc/DESeq.pdf).
   For balanced signatures, the sign of the log2-fold change of expression between the two conditions determined gene directionality. For all models regardless of balancing, gene importance was defined by FDR.
8. SVA: We applied the sva package (v3.12.0) [17] to CPMs, RPKMs, or voom-transformed raw counts. Raw counts were voom-transformed using a linear model including only dysplasia status as the predictor variable. The number of surrogate variables (SVs) not associated with dysplasia status was estimated using the default approach of Buja and Eyuboglu [18] ("be" method). SVs were then identified using the empirical estimation of control probes ("irw" method), and up to 5 were added as covariates in the linear model (limma package). The adjusted model was then used to once again voom-transform raw

counts, and subsequently fitted to identify DE genes associated with the presence of PMLs. For balanced signatures, the sign of the moderated t-statistic obtained via topTable determined gene directionality. For all models regardless of balancing, gene importance was defined by the magnitude of the t-statistic.

9. pAUC (partial AUC) [19] : We applied the rowpAUCs function in the genefilter package (v1.48.1) [20] to CPMs, RPKMs, or voom-transformed raw counts (as in #3). We used 10 class label permutations and a sensitivity cutoff of 0.1 for a specificity range of 0.9-1. For balanced signatures, the sign of the moderated t-statistic obtained via limma's topTable determined gene directionality. For all models regardless of balancing, gene importance was defined by the magnitude of the t-statistic.

Gene signature size. After the feature selection step, we selected the top scoring 10, 20, 40, 60, 80, 100, or 200 genes, making sure that for balanced signatures, half originated from an ordered list of up-regulated genes, and half from an ordered list of down-regulated genes.

Prediction method. For each set of genes, we applied multiple prediction methods to predict dysplasia status (presence of PMLs) in a training set of 46 samples and a test set of 12 samples. These training and test set samples differed in each iteration, which resulted from randomly splitting the 58 discovery set samples (Supplementary Figure 3). The following prediction methods were used:

1. glmnet: We used glmnet (v1.9-8) [14] to first estimate a range of penalty parameters λ in 10-fold cross validation using the binomial distribution family parameter and $\alpha = 0$ to ensure all feature-selected genes were included in predictions. Dysplasia status was then predicted as a binary class, using lambda.min penalty.
2. wv (weighted voting) [4]: We used the weighted voting algorithm to predict dysplasia status.
3. svm (Support Vector Machine) [21]: We used the svm function in the e1071 package (v1.6-7) [21] with linear kernel and 5-fold cross validation for class prediction.
4. rf (random forest): We used the randomForest package (v4.6-12) [15] with 1000 trees, requesting a matrix of class probabilities as output.
5. nb (Naïve Bayes): We used the naiveBayes function in the e1071 package (v1.6-7) with default parameters.

Each of the prediction algorithms generated a vector of predicted scores and a vector of predicted labels for all samples in the training and test sets.

Performance metrics. We considered 6,160 statistically and computationally viable combinations of the above parameters. The predicted class labels calculated for each model (i.e., a combination

of parameters), coupled with true class labels were then used to calculate performance metrics for the biomarker as follows:

Accuracy $\frac{TP + TN}{TP + TN + FP + FN}$

Sensitivity $\frac{TP}{TP + FN}$

Specificity $\frac{TN}{FP + TN}$

Positive Predictive Value $\frac{TP}{TP + FP}$

Negative Predictive Value $\frac{TN}{TN + FN}$

Matthew's Correlation Coefficient (MCC) $\frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$

AUC for ROC (Receiver Operating Characteristic)

MAQCII metric $0.5 \times AUC + 0.25 \times (MCC + 1),$

where TP = true positives; FP = false positives; TN = true negatives; FN = false negatives; MCC = Matthews's Correlation Coefficient; and AUC = Area Under the Curve.

For each model, we calculated these metrics for each of the 500 iterations (different training and test sets assembled from the discovery set samples) and then averaged over all iterations. In addition to the standard performance metrics, we calculated model overfitting and gene selection consistency. The overfitting metric was calculated as the difference between the train set AUC and the test set AUC. Specifically, a model performing well on the training set but poorly on the test set would achieve a high overfitting score. For each model, the gene selection consistency metric was calculated as the average ("normalized" to biomarker size in a given model) percentage of genes passing the gene filter, that were selected into the final gene committee in all 500 iterations:

$$consistency = 1 - \frac{\#\ unique\ genes\ in\ all\ iterations - biomarker\ size}{(biomarker\ size\ \times\ \#\ iterations) - biomarker\ size}$$

For example, a model requiring a 10-gene biomarker would have the highest consistency (1) if it selected the same 10 genes in all 500 iterations (10 unique genes selected altogether). The same model would have the lowest consistency (0) if it selected a different set of 10 genes in all iterations (10 genes x 500 iterations = 5000 unique genes altogether).

Selection of best model. In selecting the best model from among the 6,160 we tested, we considered the degree of model overfitting, model gene selection consistency and test set AUC. First, we identified top 10% (n=616) least overfitting models. Simultaneously, we identified top 10% (n=616) most consistent models. Finally, the model with the highest test set AUC among models fulfilling both criteria (n=121) was chosen as the final model.

Selection of final gene signature. The biomarker genes selected may differ between iterations due to changes in the training set. Therefore, to generate a final gene signature, we trained the biomarker using all 58 discovery set samples and best model parameters.

Positive and negative controls. The biomarker discovery pipeline was also used to develop control biomarkers. As positive controls, we used smoking status and sex phenotypes to identify biomarkers that could successfully distinguish former from current smokers (AUC=0.99), and females from males (AUC=0.96). As negative controls, we used randomly shuffled labels for dysplasia status (AUC=0.48), smoking status (AUC=0.52), and sex (AUC=0.51). Label shuffling was conducted preserving the association between gene expression profiles and remaining phenotypes; i.e., in the case of shuffled dysplasia status, only dysplasia status was shuffled while other phenotypes and the corresponding gene expression profile remained unchanged and linked to the same sample ID.

Validations. We tested the performance of the final biomarker using the biomarker discovery pipeline in validation mode. In this mode, the pipeline takes in the entire discovery set (n = 58) as the training set, and an external validation set as the test set. The test set is first corrected for gb-ratio (RNA-Seq quality metric) using limma, and the residual data is used as input. Both training and test sets are then z-score normalized. The pipeline is run using only the final model to generate prediction labels and prediction scores for the test set samples. Finally, pROC package (v1.8) [(22)] is used to visualize and quantify biomarker performance by plotting a ROC curve using prediction scores as the response and the dichotomous phenotype as the predictor, and extracting the AUC value from the resulting ROC object.

Detecting PML presence in validation set samples

In order to validate the biomarker's ability to detect the presence of PMLs, we tested the performance of the biomarker in smokers with and without PMLs (n=17 samples) left out of the biomarker discovery process. To assess the robustness of the results, we randomly permuted dysplasia status labels 100 times, obtaining biomarker scores for all 17 samples in each of the iterations. We then concatenated the 100 newly generated biomarker score sets for randomized labels, creating a predictor vector consisting of 1700 scores. Similarly, we concatenated 100 identical copies of biomarker score sets for true labels, creating a response vector of the same length. This allowed us to visualize the performance of the biomarker on true and randomized labels in a single ROC curve (Figure 5).

Predicting PML progression in longitudinally-collected samples

In order to validate the biomarker's ability to predict sample progression/regression, we first used the biomarker to score the longitudinally collected RPCI samples (n=51). Next, we calculated the difference in scores between two consecutive time points for each patient (later time point biomarker score - earlier time point biomarker score). For example, a subject with 3 samples from 3 different time points would have 3 scores, and thus two score differences could be calculated; a subject with 2 samples from 2 time points would have 2 scores, and thus 1 score difference.

Each pair of samples was assigned a "progressing/stable" or "regressing" phenotype. A "progressing/stable" phenotype indicated that the worst histological grade of PMLs sampled during the baseline procedure increased in severity or remained unchanged at follow-up; while a "regressing" phenotype indicated that the worst histological grade of PMLs sampled at baseline decreased in severity at follow-up.

We quantified the ability of the score difference to predict the "progression/regression" phenotype by plotting a ROC curve, using the vector of score differences as the predictor variable, and the progression/regression phenotype as the response variable.

Implementation of the method. The framework and structure of this pipeline are based on principles outlined by Joshua Campbell, PhD for microarray data applications. The pipeline outlined in this paper was substantially modified to accommodate RNA-Seq data as well as RNA-Seq-specific methods.

Subject inclusion/exclusion criteria for samples from the British Columbia Cancer Agency (BCCA)

The samples with normal/hyperplasia histology are part of the Pan-Canadian Study and included subjects between 50 and 75 years old, current or former smokers who have smoked cigarettes for 20 years or more, and that had an estimated 3-year lung cancer risk of greater than or equal to 2%. Exclusion criteria included medical conditions, such as severe heart disease, that would jeopardize the subject's safety during participation in the study, previously diagnosed lung cancer, ex-smokers of greater than or equal to 15 years, anti-coagulant treatment, and pregnancy. The subjects with airway dysplasia were participants in three different chemoprevention studies for green tea extract (n=27 samples), sulindac (n=4 samples), and myo-inositol (n=13 samples) or from the Pan-Canadian Study described above (n=6). All samples were collected at the BCCA at baseline prior to administration of therapeutic interventions. Inclusion criteria for these chemoprevention trials can be summarized as subjects between 40 and 79 years of age, current or former smokers with at least 30 pack-years, no lung cancer history or stage 0/I curatively treated NSCLC either at least 1 year or 6 months prior to the trial (depending on trial). Exclusion criteria varied by trial but included medical conditions that would jeopardize the subject's safety during participation of the study and pregnancy. See details below:

Green Tea:

Inclusion Criteria

- Women or men age 45 to 74 years of age
- Current or former smokers who have smoked at least 30 pack-years, e.g. 1 pack per day for 30 years or more (a former smoker is defined as one who has stopped smoking for one or more years)
- ECOG performance status 0 or 1
- C-Reactive Protein >1.2 mg/L
- One or more areas of dysplasia with a surface diameter larger than 1.2 mm on autofluorescence bronchoscopy
- Willing to take Polyphenon E/placebo twice a day regularly
- Since it is unknown if Polyphenon E or EGCG will cause fetal harm when administered during pregnancy, women subjects must be postmenopausal (no menstrual periods > 1year or elevated FSH > 40 mIU/ml), surgically sterile, or using birth control pill. Women of childbearing age must have normal β-HCG within 14 days to exclude pregnancy.
- Normal renal and liver function defined as serum creatinine bilirubin, AST, ALT or alkaline phosphatase levels below the upper limit of normal
- Agreeing to sign, on initial interview, informed consent forms for screening procedures (sputum cytometry analysis, fluorescence bronchoscopy, and low dose spiral thoracic CT scan). Once eligibility has been determined for the chemoprevention trial participation, agreeing to sign a study- specific treatment informed consent form.

Exclusion Criteria

- Consumption of more than 7 cups of tea a week
- Use of other natural health products containing green tea compounds
- Chronic active hepatitis/liver cirrhosis
- Severe heart disease, e.g. unstable angina, chronic congestive heart failure, use of antiarrhythmic agents
- Ongoing gastric ulcer
- Have on-going rectal bleeding
- Have a history of chronic diverticulitis and/or colitis
- Experiencing symptoms of gastritis or hemorrhoids in which medical treatment is required
- Experiencing any symptomatic gastrointestinal condition that may predispose the individual to gastrointestinal bleeding
- Acute bronchitis or pneumonia within one month
- Carcinoma in-situ or invasive cancer on bronchoscopy or abnormal spiral chest CT suspicious of lung cancer
- Known reaction to Xylocaine salbutamol, midazolam, and alfentanil
- Known allergy to green tea and/or corn starch, gelatin, or other nonmedicinal ingredients
- Any medical condition, such as acute or chronic respiratory failure, or bleeding disorder, that in the opinion of the investigator could jeopardize the subject's safety during participation in the study
- On anti-coagulant treatment such as warfarin or heparin
- Breastfeeding
- Pregnancy

- Unwilling to have a bronchoscopy
- Unwilling to have a spiral chest CT
- Unwilling to sign a consent

Sulindac:

Inclusion Criteria

- Men and women 40 through 79 years of age
- Current or former smokers with a ≥ 30 pack-year smoking history and (a) no prior lung cancer, (b) stage I NSCLC resected at least one year prior to Registration/Randomization, or (c) stage I Non-Small Cell Lung Cancer (NSCLC) with a > 1 year interval since adjuvant chemotherapy conclusion
- Women of childbearing potential and men must agree to use adequate contraception (hormonal or barrier method of birth control; abstinence) prior to study entry and for the duration of study participation. Should a woman become pregnant or suspect she is pregnant while participating in this study, she should inform her treating physician immediately.
- A negative (serum or urine) pregnancy test done ≤ 7 days prior to
- Registration/Randomization, for women of childbearing potential only
- Willingness to provide tissue blocks and sputum samples for research purposes
- Participants must have normal organ and marrow function as defined below and obtained ≤ 45 days prior to Registration/Randomization:
  - Hemoglobin ≥ lower limit of institutional normal (LLN)
  - Leukocytes ≥ 3,000/μL
  - Absolute neutrophil count ≥ 1,500/μL
  - Platelets ≥ 100,000/μL
  - Direct bilirubin ≤ 1.5 x institutional upper limit of normal (ULN)
  - ALT (SGPT) ≤ 1.5 x institutional ULN
  - Creatinine ≤ 1.5 x institutional ULN or calculated creatinine clearance ≥ 30 ml/min
- ≥ 1 site of histologically-confirmed bronchial dysplasia
- ECOG performance status ≤ 1
- Negative chest x-ray
- Negative electrocardiogram

Exclusion Criteria

- Prior history of cancer (within the previous 3-years). Exception: Stage I NSCLC as outlined above, nonmelanomatous skin cancer, localized prostate cancer, carcinoma in situ (CIS) of cervix, or superficial bladder cancer with conclusion of treatment > 6 months prior to Registration/Randomization.
- Prior pneumonectomy
- Solid organ transplant recipients
- History of GI ulceration, bleeding or perforation
- Uncontrolled intercurrent illness including, but not limited to: ongoing or active infection, symptomatic congestive heart failure, unstable angina pectoris, cardiac arrhythmia, recent (≤ 6 months) history of MI, chronic renal disease, chronic liver disease, difficult to control hypertension or psychiatric illness/social situations that would limit compliance with study requirements.
- Recent (≤ 6 months) participation in another chemoprevention trial

- Participant currently receiving any other investigational agents
- Any supplemental oxygen use (continuous or intermittent use) or documented
- Room Air (RA) SaO2 < 90%
- Pregnant women. Note: because there are no adequate, well-controlled studies in pregnant women and sulindac is absolutely contraindicated in the 3rd trimester.
- Breastfeeding women. Note: because there is an unknown but potential risk for adverse events in nursing infants secondary to treatment of the mother with sulindac, women who are breast-feeding will be excluded.
- Individuals who are known to be HIV positive. Note: HIV positive individuals are excluded for the following two reasons:
  - First, HIV positive individuals are known to have altered immune function. Since one of the potential mechanisms of action of sulindac is proposed to be enhancement of immune function in preventing lung cancer progression, it is not known how the presence of HIV infection would alter this enhancement of immune function as compared to non- HIV infected individuals.
  - Second, individuals with HIV are also known to be at higher risk for lung cancer then non-HIV infected individuals which would alter the risk/incidence of lung cancer in our study population.
- Regular NSAID or corticosteroid use during the 6-month period prior to intervention (may be eligible after washout period of 12 weeks for NSAIDs and
  - 6 weeks for corticosteroids)
- Regular aspirin use. Exception: Aspirin can be used if prescribed by a physician for prevention. Maximum of one aspirin (81mg) per day allowed.
- History of allergic reactions or hypersensitivity to sulindac or other NSAIDS, including aspirin-sensitive asthma
- Women of childbearing potential who are unwilling to employ adequate contraception (hormonal or barrier method of birth control; abstinence) prior to study entry and for the duration of study participation. Note: Effects of sulindac on the developing human fetus at the recommended therapeutic dose are fetal harm early in pregnancy. However, there are known harmful adverse events in the third trimester of pregnancy. Should a woman become pregnant or suspect she is pregnant while participating in this study, she should inform her treating physician immediately.
- Current use of methotrexate, corticosteroids, (anti-platelet agents) warfarin, ticlopidine, clopidogrel, aspirin, abciximab, dipyridamole, eptifibatide, tirofiban, lithium, cyclosporine, hydralazine, ACE inhibitors

Myo-inositol:

Inclusion Criteria

- Ability to understand and willingness to sign a written informed consent document
- Age ≥45 to ≤79
- ECOG performance status (PS) 0 or 1 (see Appendix A)
- One or both of the following:
  - Stage 0/I curatively treated non-small cell lung cancer (NSCLC) with a ≥30 pack-year smoking history (surgery, adjuvant chemotherapy or radiotherapy must be completed ≥ 6 months prior to screening); OR
  - Current or former smokers with a ≥ 30 pack-year smoking history without a history of lung cancer. Pack-years is determined by multiplying the number of

packs smoked per day by the number of years smoked.

- Women of childbearing capacity who agree to use an acceptable form of birth control for the duration of the study (e.g. condom, oral contraceptives, etc.)

Exclusion Criteria

- Prior history of cancer, with the following exceptions:
  - ≥ 3-year disease free interval (with the exception of stage I NSCLC as described above)
  - Non-melanomatous skin cancer
  - Localized prostate cancer with conclusion of treatment > 6 months prior to screening
  - Carcinoma in situ (CIS) of cervix with conclusion of treatment > 6 months prior to screening
  - Superficial bladder cancer with conclusion of treatment > 6 months prior to screening
- Prior pneumonectomy
- Solid organ transplant recipients
- Uncontrolled intercurrent illness including, but not limited to: ongoing or active infection, symptomatic congestive heart failure, unstable angina pectoris, cardiac arrhythmia, severe chronic obstructive pulmonary disease requiring supplemental oxygen, difficult to control hypertension, or psychiatric illness/social situations that would limit compliance with study requirements.
- Schizophrenia
- Bipolar disorder
- Lithium treatment
- Carbamazepine treatment
- Valproate treatment
- Diabetes
- Currently using other natural health products containing inositol
- Anticoagulant use such as Coumadin or heparin. Exception: participant is off those drugs for ≥ 7 days prior to pre-registration.
- Recent (≤ 6 months) participation in another chemoprevention trial
- Participant currently receiving any other investigational agents
- Any supplemental oxygen use (continuous or intermittent use) or documented Room Air (RA) SaO2 < 90%
- Pregnant women. (Excluded because the effects of high doses of myo-inositol on the fetus or newborn are not known.)
- Breastfeeding women. (Excluded because the risk for adverse events in nursing infants secondary to treatment of the mother with high doses of myo-inositol are not known.)
- History of allergic reactions attributed to myo-inositol
- History of allergies to any ingredient in the study product or placebo

Early Detection of Lung Cancer – A Pan-Canadian Study:

Inclusion Criteria

- Women or men age 50 to 75 years
- Current or former smokers who have smoked cigarettes for 20 years or more (a former smoker is defined as one who has stopped smoking for one or more years)
- An estimated 3-year lung cancer risk of ≥2% based on the risk prediction model.

- ECOG performance status 0 or 1
- Capable of providing, informed consent for screening procedures (low dose spiral CT, AFB, spirometry, blood biomarkers)

Exclusion Criteria

- Any medical condition, such as severe heart disease (e.g. unstable angina, chronic congestive heart failure), acute or chronic respiratory failure, bleeding disorder, that in the opinion of the investigator could jeopardize the subject's safety during participation in the study or unlikely to benefit from screening due to shortened life-expectancy from the co-morbidities
- Have been previously diagnosed with lung cancer
- Have had other cancer with the exception of the following cancers which can be included in the study: non-melanomatous skin cancer, localized prostate cancer, carcinoma in situ (CIS) of the cervix, or superficial bladder cancer. Treatment of the exceptions must have ended >6 months before registration into this study.
- Ex-smoker for ≥ 15 years
- On anti-coagulant treatment such as warfarin or heparin
- Known reaction to Xylocaine, salbutamol, midazolam, and alfentanil
- Pregnancy
- Unwilling to have a spiral chest CT
- Chest CT within 2 years
- Unwilling to sign a consent

Subject inclusion/exclusion criteria for samples from RPCI

Subjects met the following high-risk lung screening criteria: 1) Personal cancer history of the lung, bronchus, head/neck, and/or esophagus and no evidence of disease at the time of enrollment, or 2) No personal history of upper aerodigestive cancer, age 50+, and a current smoker or a former smoker with 20+ pack years. In addition, subjects in the second group had to have one or more risk factors including chronic lung disease such as emphysema, chronic bronchitis, or chronic obstructive pulmonary disease, occupationally related asbestos disease, or a family history of lung cancer in a first degree relative.

Supplemental Table 1. ANOVA derived p-values for the association between the surrogate variables and demographic/phenotypic variables

Variable	SV1	SV2	SV3	SV4	SV5	SV6	SV7	SV8	SV9
Presence of premalignant lesion (2-level)	0.549	0.376	0.964	0.500	0.118	0.481	0.046	0.166	0.652
Smoking status	0.000	0.655	0.191	0.084	0.689	0.804	0.308	0.719	0.761
Smoking status by Gene Expression	0.000	0.363	0.801	0.045	0.819	0.780	0.130	0.827	0.663
Sex	0.961	0.058	0.000	0.032	0.492	0.801	0.433	0.884	0.991
COPD status	0.612	0.866	0.047	0.161	0.973	0.129	0.083	0.007	0.592
Pack-years	0.398	0.293	0.523	0.576	0.845	0.399	0.875	0.428	0.178
Age	0.300	0.153	0.562	0.845	0.166	0.618	0.037	0.050	0.528
FEV1	0.050	0.391	0.046	0.009	0.123	0.150	0.171	0.028	0.691
FEV1/FVC ratio	0.023	0.670	0.172	0.056	0.491	0.107	0.028	0.011	0.708
Barcode	0.870	0.605	0.006	0.500	0.745	0.444	0.695	0.119	0.187
Lane	0.335	0.748	0.682	0.351	0.037	0.792	0.402	0.996	0.549
Batch	0.676	0.730	0.474	0.426	0.861	0.037	0.145	0.688	0.261
GC content	0.599	0.886	0.057	0.902	0.257	0.157	0.001	0.416	0.210
Genebody 80/20 ratio (gb-ratio)	0.000	0.245	0.633	0.271	0.000	0.736	0.015	0.319	0.048
Number of Uniquely Aligning Reads	0.302	0.154	0.726	0.948	0.055	0.120	0.036	0.163	0.586
Number of Reads Aligning to Splice Junctions	0.545	0.605	0.498	0.442	0.000	0.383	0.170	0.745	0.942
Z-score (sample mean of z-score normalized data by gene)	0.514	0.371	0.238	0.595	0.024	0.031	0.005	0.353	0.021
Relative Expression (sample median of ratios computed for each gene by dividing the expression by the median expression)	0.814	0.615	0.996	0.740	0.918	0.887	0.214	0.274	0.111

Supplemental Table 2. Phenotypic information about the human biopsy cell cultures used in the bioenergetics and mitochondrial enumeration (MitoTraker Green FM) experiments.

Histology	Gender	Smoking Status	Bioenergetics	MitoTrackerFM
Normal	F	Current	X	
Normal	M	Current	X	
Normal	F	Former	X	
Normal	M	Former	X	
Normal	F	Current	X	X
Normal	F	Current	X	X
Moderate Dysplasia	M	Current	X	
Severe Dysplasia	M	Former	X	
Severe Dysplasia	M	Current	X	
Low grade dysplasia	M	Former	X	
Severe Dysplasia	M	Current	X	X
Low grade dysplasia	M	Former	X	X

Supplemental Table 3. Phenotypic information about the human biopsies used in the IHC experiments.

(*CS refers to current smoker and FS to former smoker)

Stain	PtID	Smoking Status	Worst Histology Description
TOMM22	Pt 3	FS	0 Normal, Negative, Benign Mucosa
COX4I1	Pt 3	FS	0 Normal, Negative, Benign Mucosa
TOMM22	Pt 4	FS	23 Squamous Metaplasia (non-specific), Mature Metaplasia, Squamous Hyperplasia
COX4I1	Pt 4	FS	23 Squamous Metaplasia (non-specific), Mature Metaplasia, Squamous Hyperplasia
TOMM22	Pt 3	FS	25 Moderate Dysplasia, Squamous Pre-invasive
COX4I1	Pt 3	FS	25 Moderate Dysplasia, Squamous Pre-invasive
TOMM22	Pt 1	CS	27 CIS Squamous Carcinoma In-Situ
Cox-IV	Pt 1	CS	27 CIS Squamous Carcinoma In-Situ

Supplemental Table 4. Demographic and clinical characteristics of the British Columbia Lung Health Study stratified by premalignant lesions status

	Discovery Set				Validation Set			
Factor	**Overall (n=58)**	**No Lesions (n=20)**	**Lesions (n=38)**	**P***	**Overall (n=17)**	**No Lesions (n=5)**	**Lesions (n=12)**	**P***
Age	62.7 (7.1)	64.1 (5.8)	61.9 (7.6)	0.24	63.9 (8.6)	66 (5.8)	63 (9.7)	0.45
Male	37/58 (63.8)	12/20 (60)	25/38 (65.8)	0.78	14/17 (82.4)	4/5 (80)	10/12 (83.3)	1
Current smoker	28/58 (48.3)	9/20 (45)	19/38 (50)	0.79	8/17 (47.1)	2/5 (40)	6/12 (50)	1
Pack-years	48.2 (16.9)	49.4 (18.9)	47.5 (15.9)	0.71	44.6 (12.9)	40.5 (11.6)	46.3 (13.5)	0.39
FEV1% Predicted	86.5 (17.7)	87.8 (16.7)	85.7 (18.5)	0.66	69.5 (16.2)	71 (17.7)	68.9 (16.3)	0.83
FEV1/FVC Ratio	72.1 (7.7)	75.1 (6.3)	70.4 (8)	0.02	67 (8.1)	66.8 (8.5)	67.1 (8.3)	0.95
COPD (FEV1%<80 & FEV1/FVC<70)	11/58 (19)	2/20 (10)	9/38 (23.7)	0.3	11/17 (64.7)	3/5 (60)	8/12 (66.7)	1
Histology				<0.001				<0.001
Normal	11/58 (19)	11/20 (55)			1/17 (5.9)	1/5 (20)		
Hyperplasia	9/58 (15.5)	9/20 (45)			4/17 (23.5)	4/5 (80)		
Metaplasia	0/58 (0)				0/17 (0)			
Mild Dysplasia	29/58 (50)		29/38 (76.3)		6/17 (35.3)		6/12 (50)	
Moderate Dysplasia	6/58 (10.3)		6/38 (15.8)		6/17 (35.3)		6/12 (50)	
Severe Dysplasia	3/58 (5.2)		3/38 (7.9)				0/12 (0)	

Data are means (SD) for continuous variables and proportions (%) dichotomous variables. Reads are expressed in millions denoted by M. P* values are for the comparison of subjects with and without premalignant lesions. Two sample t-tests were used for continuous variables; Fisher's exact test was used for factors.

Supplemental Table 5. Alignment statistics of the British Columbia Lung Health Study Discovery and the Roswell Park Cancer Institute cohort

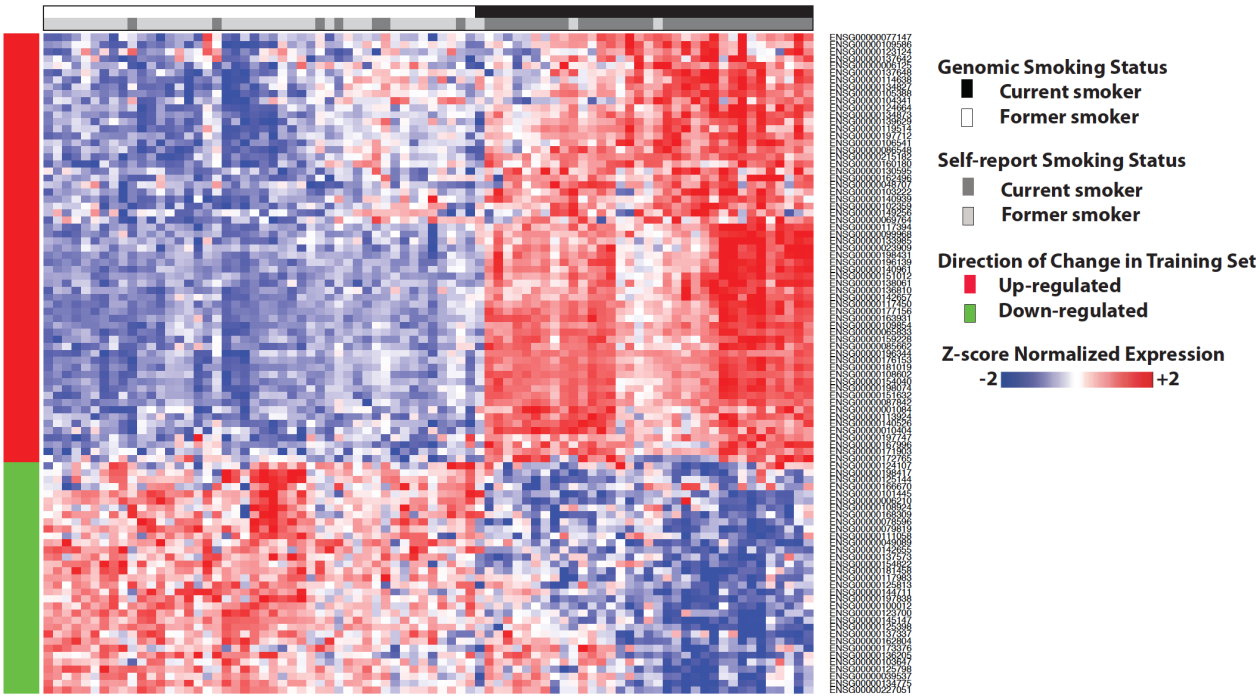
	BC-LHS Discovery Set				BC-LHS Validation Set				RPCI
Factor	**Overall (n=58)**	**No Lesions (n=20)**	**Lesions (n=38)**	**P***	**Overall (n=17)**	**No Lesions (n=5)**	**Lesions (n=12)**	**P***	**Overall (n=51)**
Total Alignments	90M (16M)	98M (15M)	91M (17M)	0.67	93M (22M)	94M (18M)	92M (24M)	0.86	95M (15M)
Unique Alignments	83M (15M)	82M (13M)	83M (16M)	0.65	85M (20M)	86M (16M)	84M (22M)	0.85	
Properly Paired Alignments	66M (1.2M)	65M (11M)	67M (12M)	0.63	68M (16M)	69M (13M)	67M (17M)	0.86	65M (9.6M)
Genebody 80/20 Ratio	1.3 (0.2)	1.3 (0.1)	1.3 (0.2)	0.39	1.3 (0.3)	1.2 (0.1)	1.4 (0.3)	0.15	1.8 (0.2)
Mean GC Content	48.1 (3.4)	47.5 (2.7)	48.4 (3.6)	0.33	47.4 (3.8)	46.9 (3.8)	47.6 (3.9)	0.74	49.2 (1.4)

Data are means (SD). Reads are expressed in millions denoted by M. P* values are for two sample t-tests for comparison of subjects with and without premalignant lesions.

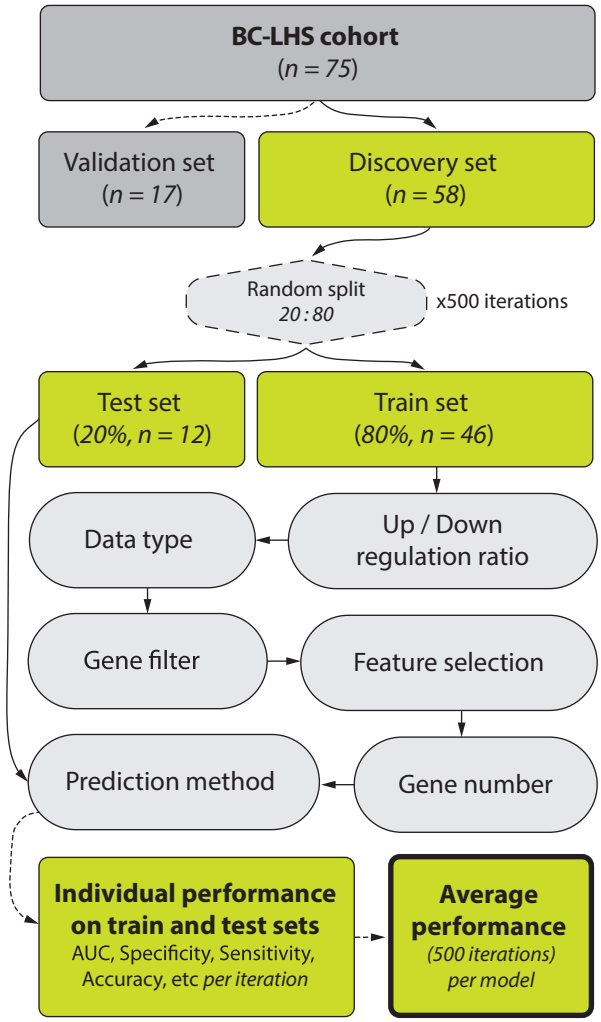
Supplemental Table 6. Demographic and clinical characteristics of the Roswell Park Cancer Institute Cohort (n=51 samples from n=23 subjects)

Factor	Overall	Regressing	Progressing Stable	P*
No. Samples	51	34	22	
No. Sample Pairs	28	17	11	
No. Patients**	23	16	10	
Time between Procedures (Days)	343.8 (171.9)	350.9 (199.6)	332.8 (125.9)	0.77
Histological Grade Change	-0.9 (1.7)	-1.9 (1.0)	0.7 (1.3)	<0.001
Worst Histological Lesion Observed				
Normal	5/51 (9.8)	4/34 (11.8)	2/22 (9.1)	0.038
Hyperplasia	6/51 (11.8)	5/34 (14.7)	1/22 (4.5)	
Metaplasia	9/51 (17.6)	8/34 (23.5)	1/22 (4.5)	
Mild Dysplasia	3/51 (5.9)	3/34 (8.8)	0 (0)	
Moderate Dysplasia	20/51 (39.2)	9/34 (26.5)	15/22(68.2)	
Severe Dysplasia	8/51 (15.7)	5/34 (14.7)	3/22 (13.6)	
Age at Baseline	58.1 (6.5)	58.4 (6.9)	57.6 (6.1)	1
Male	13/28 (46.4)	7/17 (41.2)	6/11 (54.5)	0.7
Ever smoker at Baseline	27/28 (96.4)	17/17 (100)	10/11 (90.9)	0.39
Pack-years at Baseline	48.1 (22)	49.8 (24.8)	45.4 (17.6)	1

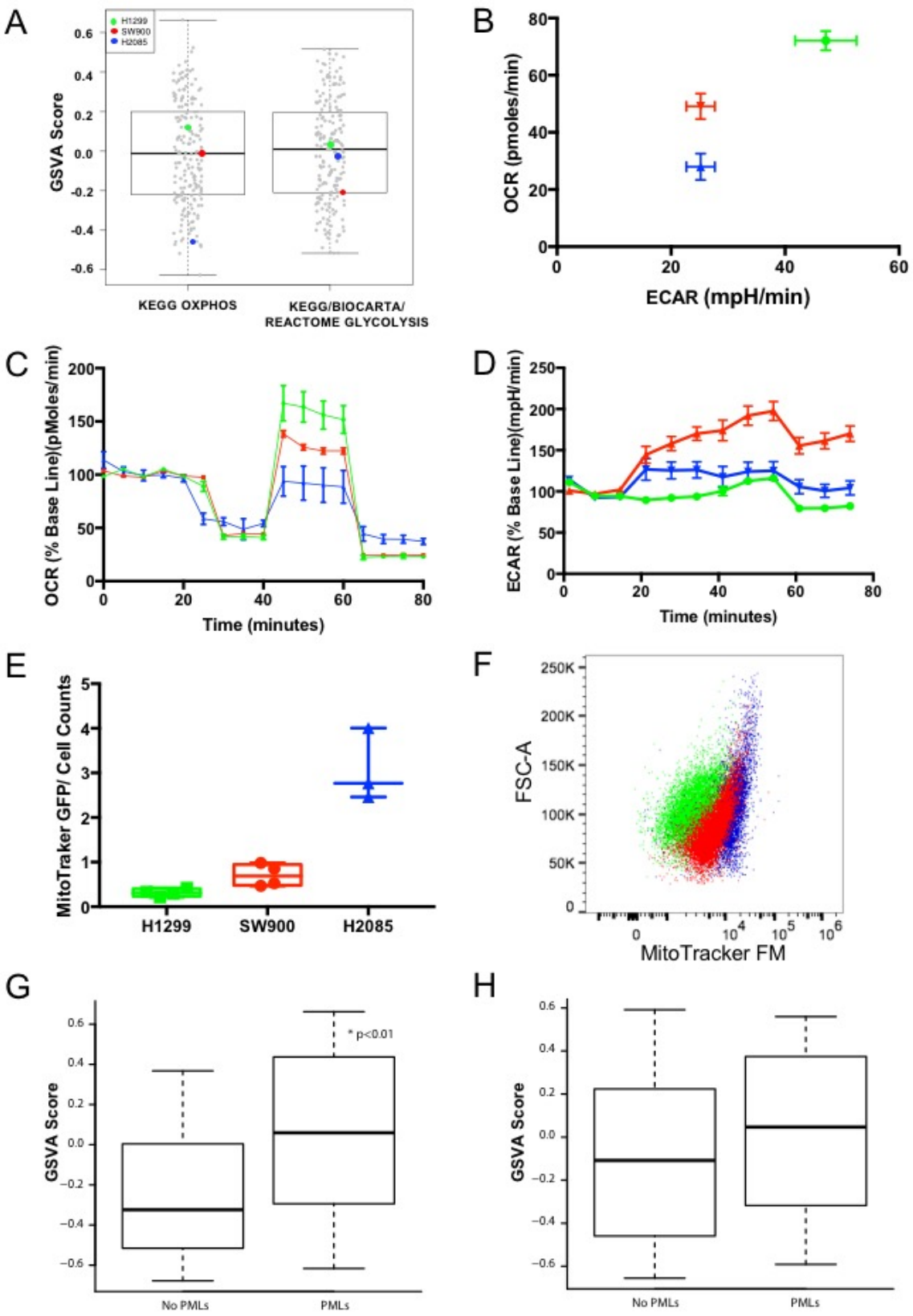
Data are means (SD) for continuous variables and proportions (%) for dichotomous variables. P* values are for the comparison of samples, sample pairs, or patients classified as having regressing or progressing/stable PMLs. Two sample t-tests were used for continuous variables; Fisher's exact test was used for factors. **Among the 23 patients, 3 patients had 2 sample pairs where one pair was classified as regressing and the other as progressing/stable. These patients are counted in both the regressing and progressing/stable columns.



Supplemental Figure 1. Unsupervised hierarchal clustering of genes associated with smoking status. The weighted voting algorithm was trained on z-score normalized microarray data (GSE7895) across 94 genes differentially expressed between current and never smokers and used to predict smoking status in log2-transformed counts per million (cpm) that were z-score normalized from the 82 mRNA-Seq samples. The heatmap shows the results of unsupervised Ward hierarchal clustering across the 82 mRNA-Seq samples and the 94 genes. The row color label indicates if genes were up-regulated (red) or down-regulated (green) in current smokers compared to never smokers in GSE7895. The lower column color labels indicate the smoking status in the clinical annotation (self-report) with light gray indicating former smokers and dark gray indicating current smokers. The upper column color labels indicate the predicted class of the samples based on the 94 genes with white indicating former smokers and black indicating current smokers. Log2-cpm mRNA-Seq data was z-score normalized prior to clustering.



Supplemental Figure 2. **Biomarker discovery flowchart**. Samples (n=75) were split into a discovery set (n=58) and a validation set (n=17). The pipeline was run 500 times, and each time the discovery set was randomly split into training (80% of samples, n=46) and test (20% of samples, n=12) sets. The training set samples were used to train the biomarker using all combinations of pipeline parameters, including: 1. Up- / down-regulation ratio: TRUE or FALSE (see *Balancing signature*). 2. Data type: raw counts, RPKM or CPM (see *Input data preprocessing*). 3. Gene filter: genes with signal in at least 1%, 5%, 10%, or 15% of samples (see *Gene filter*). 4. Feature selection: edgeR, edgeR correcting for gb-ratio, limma, limma correcting for gb-ratio, glmnet, random forest, DESeq, SVA, or partial AUC (see *Feature selection*). 5. Gene number: 10, 20, 40, 60, 80, 100, or 200 genes (see *Biomarker size*). 6. Prediction method: weighted voting, random forest, SVM, naïve bayes, or glmnet (see *Prediction method*).


A
H1299
SW900
H2085
GSVA Score
KEGG OXPHOS
KEGG/BIOCARTA/
REACTOME GLYCOLYSIS
B
OCR (pmoles/min)
ECAR (mpH/min)
C
OCR (% Base Line)(pMoles/min)
Time (minutes)
D
ECAR (% Base Line)(mpH/min)
Time (minutes)
E
MitoTraker GFP/ Cell Counts
H1299
SW900
H2085
F
FSC-A
MitoTracker FM
G
GSVA Score
* p<0.01
No PMLs
PMLs
H
GSVA Score
No PMLs
PMLs

Supplemental Figure 3. **Cellular metabolism in cancer cell lines and in the airway field associated with premalignant lesions** (**A**) GSVA scores were calculated based on genes in KEGG OXPHOS pathway and KEGG, Biocarta, and Reactome Glycolysis pathways in the CCLE cell lines highlighting the H1229 (green) (high OXPHOS and moderate glycolysis), SW900 (red) (moderate OXPHOS and low glycolysis) and H2805 (blue) ((low OXPHOS and moderate glycolysis). (**B**) Baseline OCR/ECAR ratio values for the cancer cells lines demonstrating the relationship between elevated OXPHOS GSVA scores and oxygen consumption. (**C**) Elevation of respiratory capacity associated with high OXPHOS gene score in response to mitochondrial perturbation. (**D**) Elevated ECAR response in the H1299 and H205 is associated with the moderate glycolysis GSVA score, however, although the SW900 glycolysis GSVA scores agree with baseline ECAR, in the state of repressed OXPHOS, glycolysis is activated. **(E)** Enumeration of mitochondria within each cancer cell suggests that increased GSVA scores for OXPHOS or glycolysis did not correlate with mitochondrial number. H2085 cells had the lowest OXPHOS GSVA score, the lowest basal OCR, and the lowest respiratory capacity, but their mitochondrial content was significantly greater that H1299 and SW900 ($p=0.03$). **(F)** Cell area (FSC-A) is correlated with mitochondrial number (fluorescence of MitoTracker Green FM). **(G)** GSVA scores were calculated based on genes in KEGG OXPHOS pathway. The GSVA scores for OXPHOS activity were significantly elevated in the airway field of subjects with PMLs compared to subjects without PMLs ($p<0.01$). (**H**) GSVA scores were calculated based on genes in the KEGG, Biocarta, and Reactome Glycolysis pathways. The mean GSVA scores were moderately elevated in the airway field of subjects with PMLs compared to subjects without PMLs.

References

1. Beane J, Sebastiani P, Whitfield TH, Steiling K, Dumas YM, Lenburg ME, Spira A: A prediction model for lung cancer diagnosis that integrates genomic and clinical features, Cancer Prev Res (Phila) 2008, 1:56-64

2. Edgar R, Domrachev M, Lash AE: Gene Expression Omnibus: NCBI gene expression and hybridization array data repository., Nucleic Acids Research 2002, 30:207-210

3. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP: Exploration, normalization, and summaries of high density oligonucleotide array probe level data., Biostatistics (Oxford, England) 2003, 4:249-264

4. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES: Molecular classification of cancer: class discovery and class prediction by gene expression monitoring., Science 1999, 286:531-537

5. Johnson WE, Li C, Rabinovic A: Adjusting batch effects in microarray expression data using empirical Bayes methods., Biostatistics (Oxford, England) 2006, 8:118-127

6. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK: limma powers differential expression analyses for RNA-sequencing and microarray studies, Nucleic Acids Research 2015, 43:gkv007-e047

7. Law CW, Chen Y, Shi W, Smyth GK: Voom: precision weights unlock linear model analysis tools for RNA-seq read counts, Genome biology 2014,

8. Spira A, Beane J, Shah V, Liu G, Schembri F, Yang X, Palma J, Brody JS: Effects of cigarette smoke on the human airway epithelial cell transcriptome, Proceedings of the National Academy of Sciences of the United States of America 2004, 101:10143-10148

9. Wang L, Wang S, Li W: RSeQC: quality control of RNA-seq experiments., Bioinformatics (Oxford, England) 2012, 28:2184-2185

10. Anders S, Pyl PT, Huber W: HTSeq–A Python framework to work with high-throughput sequencing data, bioRxiv 2014,

11. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L: Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation., Nature biotechnology 2010, 28:511-515

12. Robinson MD, McCarthy DJ, Smyth GK: edgeR: a Bioconductor package for differential expression analysis of digital gene expression data., Journal of Gerontology 2010, 26:139-140

13. Robinson MD, Oshlack A: A scaling normalization method for differential expression analysis of RNA-seq data., Genome biology 2010, 11:R25

14. Friedman J, Hastie T, Tibshirani R: Regularization Paths for Generalized Linear Models via Coordinate Descent, Journal of statistical software 2010, 33:1-968

15. Liaw A, Wiener M: Classification and regression by randomForest, R news 2002,

16. Anders S, Huber W: Differential expression analysis for sequence count data, Genome biology 2010, 11:1

17. Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD: The sva package for removing batch effects and other unwanted variation in high-throughput experiments., Bioinformatics (Oxford, England) 2012, 28:882-883

18. Buja A, Eyuboglu N: Remarks on Parallel Analysis, Multivariate behavioral research 1992, 27:509-540

19. McClish DK: Analyzing a portion of the ROC curve., Medical decision making : an international journal of the Society for Medical Decision Making 1989, 9:190-195

20. Gentleman R, Carey V, Huber W, Hahne F: Genefilter: methods for filtering genes from high-throughput experiments. Edited by R package version, 2015,

21. Meyer D, Dimitriadou E, Hornik K, Weingessel A, Leisch F: e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien [R package e1071 version 1.6-7]. Edited by Comprehensive R Archive Network (CRAN), 2015, p.

22. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez J-C, Müller M: pROC: an open-source package for R and S+ to analyze and compare ROC curves., BMC bioinformatics 2011, 12:77