RAI SERVICES COMPANY

A blood-based smoking-related gene expression signature using a machine learning approach

Gang (Michael) Liu and G.L. Prasad
RAI Services Company
401 N Main St, Winston-Salem, NC 27101

Introduction

RAI SERVICES COMPANY

- Cigarette smoking is a major risk factor for lung cancer, cardiovascular disease, and respiratory diseases such as chronic obstructive pulmonary disease.
- Chronic cigarette smoking induces oxidative stress, chronic inflammation, and negatively impacts cellular function and signaling pathways, which may eventually culminate in smoking-related diseases.
- Comparative analyses of blood transcriptomics (gene expression profiles) between generally healthy smokers (SMK) and non-tobacco consumers (NTC) enable a better understanding of pre-clinical molecular mechanisms affected by smoking that may lead to disease states.

Motivation

RAI SERVICES COMPANY

- Gene expression signatures (a small set of genes) offer a cost effective option as biomarkers of potential harm (BoPH) to characterize biological effects due to tobacco product exposure, compared to transcriptome profiling.
- The gene signature can be measured from relatively easily accessible tissues such as blood that are obtained in a minimally invasive manner.
- Machine learning methods have been successfully used to build disease-related gene signatures (e.g., FDA-cleared MammaPrint test, a 70-gene signature).

Previously Reported Gene Signatures

RAI SERVICES COMPANY

- Previous smoking-related gene signature studies[1-3] (5-, 11-, and 20-gene signatures) applied a specific classifier and/or feature selector in their analyses.

- Limited number of independent transcriptomic datasets were used for validation of these gene signatures.

1. Arimilli, S. et al., BMC Genomics 2017;18(1):156.
2. Martin, F. et al., Human & Experimental Toxicology 2015;34(12):1200.
3. Beineke, P. et al., BMC Medical Genomics 2012;5:58.

Objective

RAI SERVICES COMPANY

- To develop a robust gene signature with validated clinical performance using all eight publicly available transcriptomics datasets.

Gene Expression Data

RAI SERVICES COMPANY

	Dataset	# (NTC)	# (SMK)	Sample Type	Microarray Platform	References
Training Dataset	**GSE87072**	40	40	PBMC	Affymetrix U133 Plus 2.0	(Arimilli, et al., 2017)
Independent Validation Datasets	**EMTAB-5279**	29	30	Whole Blood	Affymetrix U133 Plus 2.0	(Martin, et al., 2015)
	EMTAB-5278	114	60	Whole Blood	Affymetrix U133 Plus 2.0	(Martin, et al., 2015)
	GSE20189	21	27	Whole Blood	Affymetrix U133 Plus 2.0	(Rotunno, et al., 2011)
	GSE23323	22	22	Whole Blood	Agilent	(Jennen, et al., 2015)
	GSE47415	24	24	Whole Blood	Agilent	(Paul and Amundson, 2014)
	GSE15289	211	74	Whole Blood	ABI Human Genome Survey Microarray Version 2	(Dumeaux, et al., 2010)
	GSE42057	0	13	PBMC	Affymetrix U133 Plus 2.0	(Bahr, et al., 2013)

1. Arimilli, S. et al., BMC Genomics 2017;18(1):156.
2. Martin, F. et al., Human & Experimental Toxicology 2015;34(12).
3. Bahr, T.M. et al., American Journal of Respiratory Cell and Molecular Biology 2013;49(2).
4. Rotunno, M. et al., Cancer Prevention Research 2011;4(10).
5. Jennen, D.G. et al., Chemical Research in Toxicology 2015;28(10).
6. Paul, S. and Amundson, S.A, Journal of Carcinogenesis & Mutagenesis 2014;5.
7. Dumeaux, V. et al., PLoS Genetics 2010;6(3).

NTC: non-tobacco consumers
SMK: smokers

➢ **Eight blood-based smoking-related microarray datasets were used**

Machine Learning Algorithms

RAI SERVICES COMPANY

Classification

- ❖ SVM: support vector machine
- ❖ RF: random forest
- ❖ LDA: linear discriminant analysis
- ❖ NB: naïve Bayes

Feature (gene) Selection

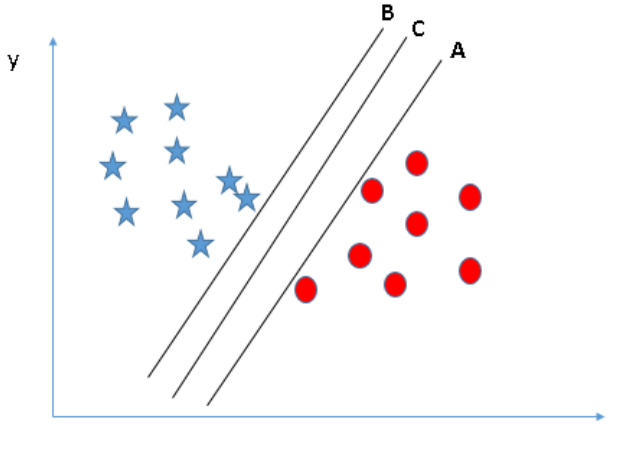
- ❖ RFE: recursive feature elimination
- ❖ CFS: correlation feature selection
- ❖ IGR: information gain ratio
- ❖ CS: Chi-squared method

- ➢ Classification models are used to predict smoking status.
- ➢ Feature selection methods are implemented to select a subset of genes from eight thousand genes contained within microarray data.

SVM+RFE

RAI SERVICES COMPANY

- SVM (support vector machine) defines an optimal hyperplane to separate the samples of different classes with maximization of separating margin.



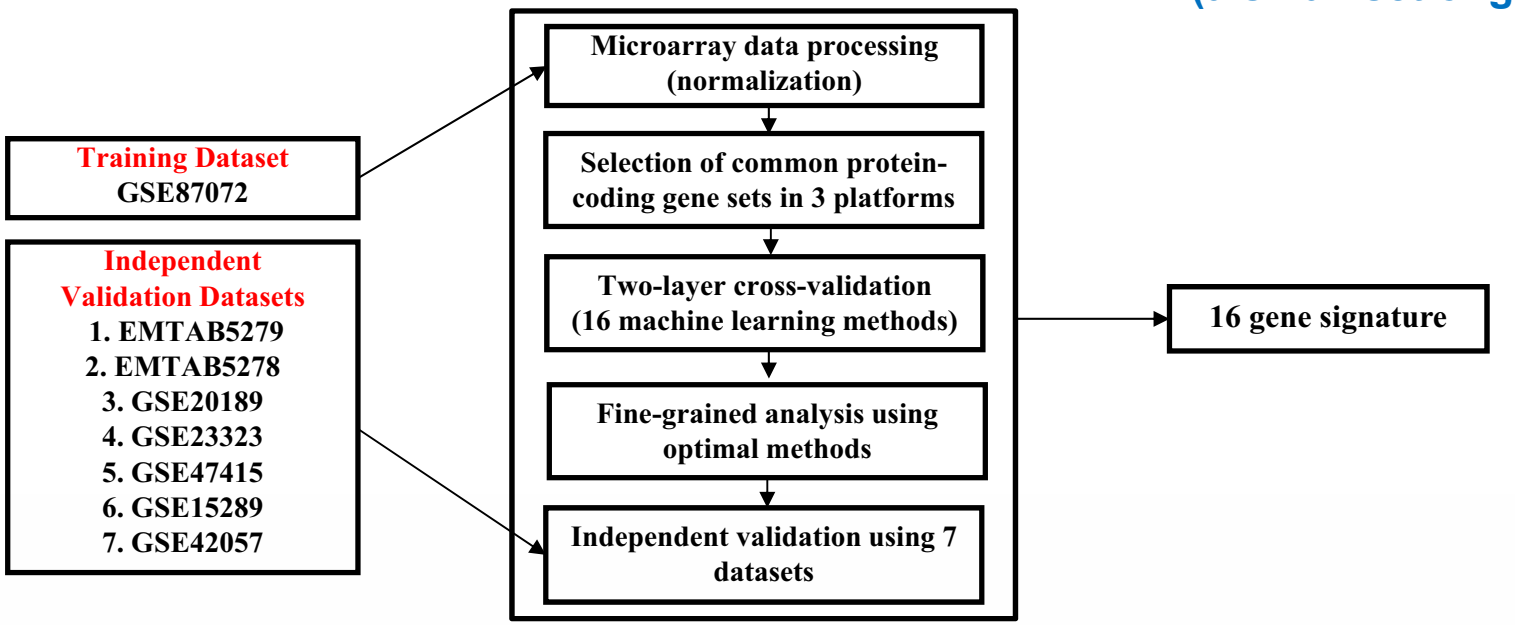
- ✓ Hyperplane: A, B and C to separate two classes (blue stars and red circles)
- ✓ Hyperplane C has maximum separating margin

- RFE (recursive feature elimination) iteratively fits the model and discards the features ranked as least important to classification performance until a specified number of features are met.

Machine Learning Workflow

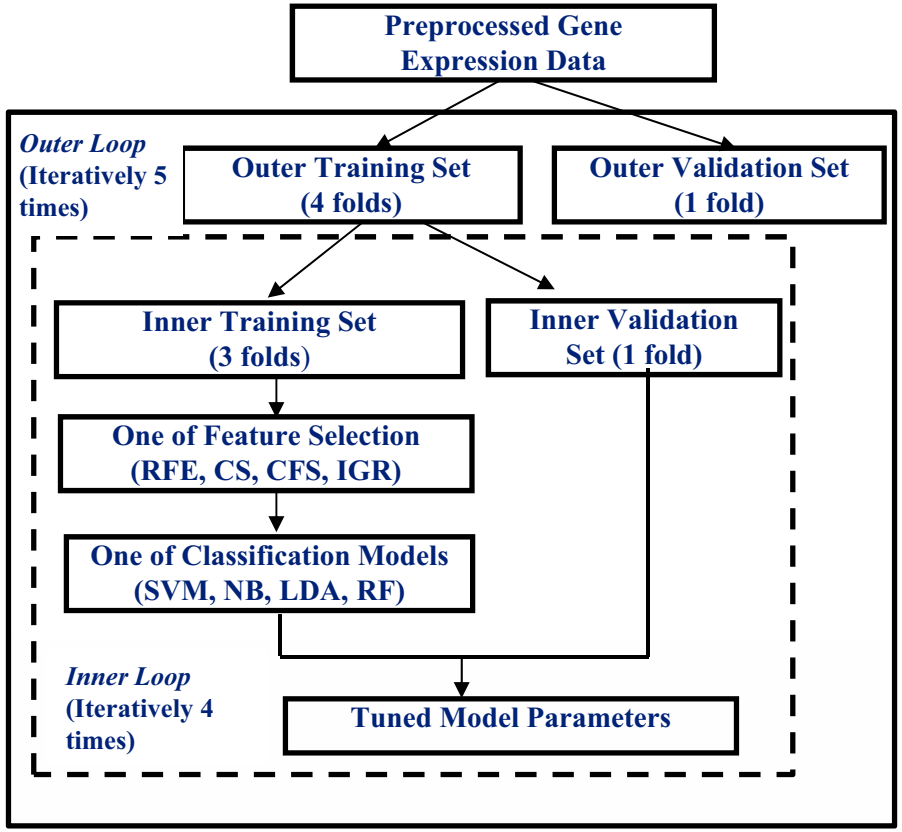
RAI SERVICES COMPANY

Gene Expression Data → Machine Learning → Gene Signature (a small set of genes)



- A machine learning workflow was developed to search for optimal algorithm

Two-layer Cross Validation



- **Two-layer (nested) cross validation method minimizes the bias introduced through the overuse of the training data, and provides a better approach to select the best-performing model.**

Model Performance: Accuracy

RAI SERVICES COMPANY

	Predicted Positive	Predicted Negative
Actual P	True Positive (TP)	False Negative (FN)
Actual N	False Positive (FP)	True Negative (TN)

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

- **Accuracy is the ratio of total number of correct predictions to the total number of samples**

Model Performance: AUCROC

RAI SERVICES COMPANY

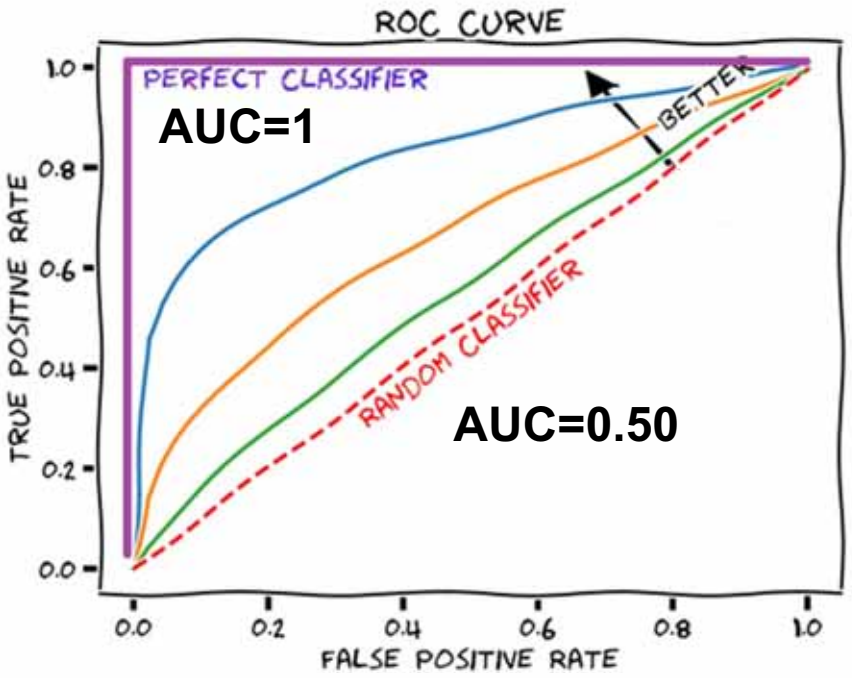
	Predicted Positive	Predicted Negative
Actual P	True Positive (TP)	False Negative (FN)
Actual N	False Positive (FP)	True Negative (TN)

$$\text{True Positive Rate} = \frac{TP}{TP+FN}$$

$$\text{False Positive Rate} = 1 - \frac{TN}{FP+TN}$$

- **AUCROC (area under curve of ROC) provides another metric of how the model performs**

Receiver Operating Characteristics



https://glassboxmedicine.com/2019/02/23/measuring-performance-auc-auroc/

Two-layer Cross Validation Results

RAI SERVICES COMPANY

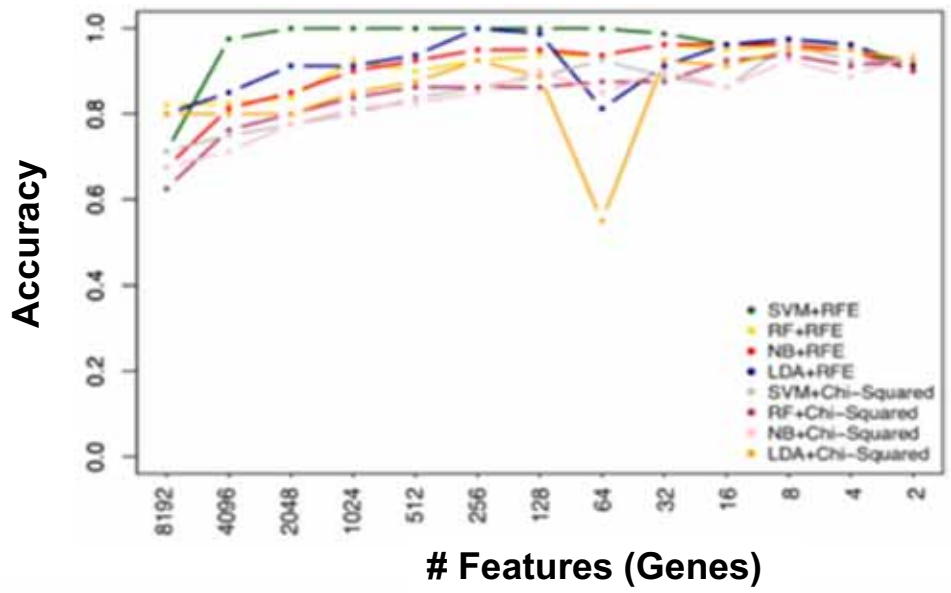
Classifier	Feature Selection	Accuracy	# Features (Genes)
SVM	RFE	1	64
	Chi-Squared	0.96	8
	CFS	0.94	1
	IGR	0.93	2
RF	RFE	0.96	8
	Chi-Squared	0.94	8
	CFS	0.88	1
	IGR	0.95	32
NB	RFE	0.96	8
	Chi-Squared	0.94	2
	CFS	0.93	1
	IGR	0.94	2
LDA	RFE	1	256
	Chi-Squared	0.95	4
	CFS	0.91	1
	IGR	0.96	4

- **SVM+RFE and LDA+RFE perform better than the others**
- **SVM+RFE identifies fewer genes in the signature**

SVM: support vector machine
RF: random forest
LDA: linear discriminant analysis
NB: naïve Bayes
RFE: recursive feature elimination
CFS: correlation feature selection
IGR: information gain ratio
CS: Chi-squared method

Two-layer Cross Validation Results

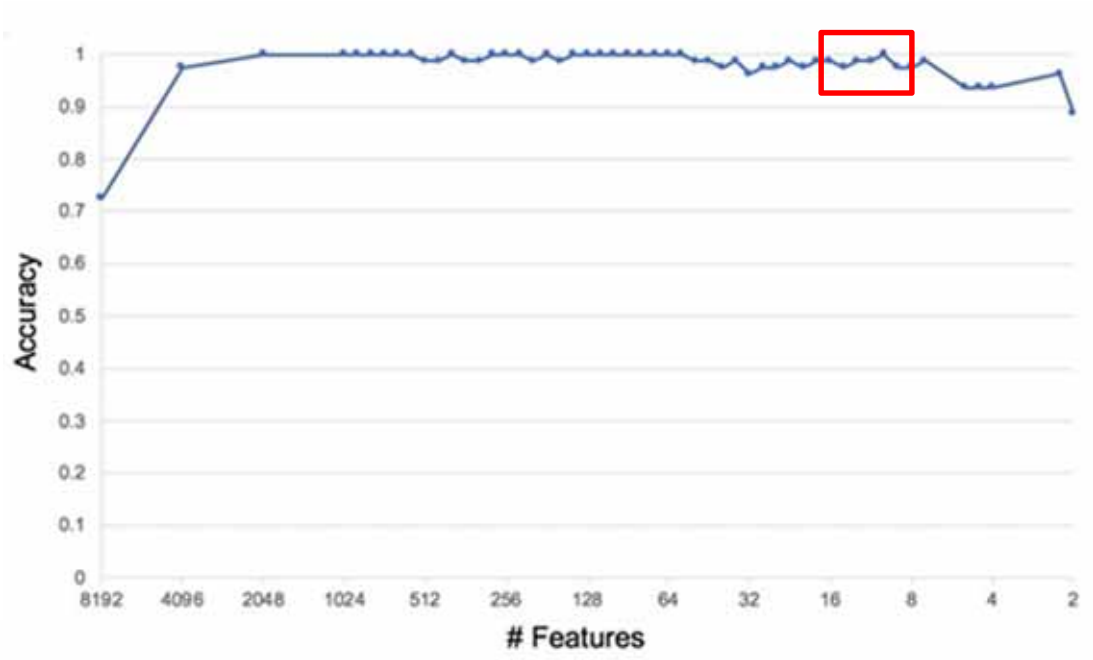
RAI SERVICES COMPANY



- **Model performance changes with the number of features**
- **SVM+RFE outperforms the others**

Fine-grained Analysis (SVM+RFE)

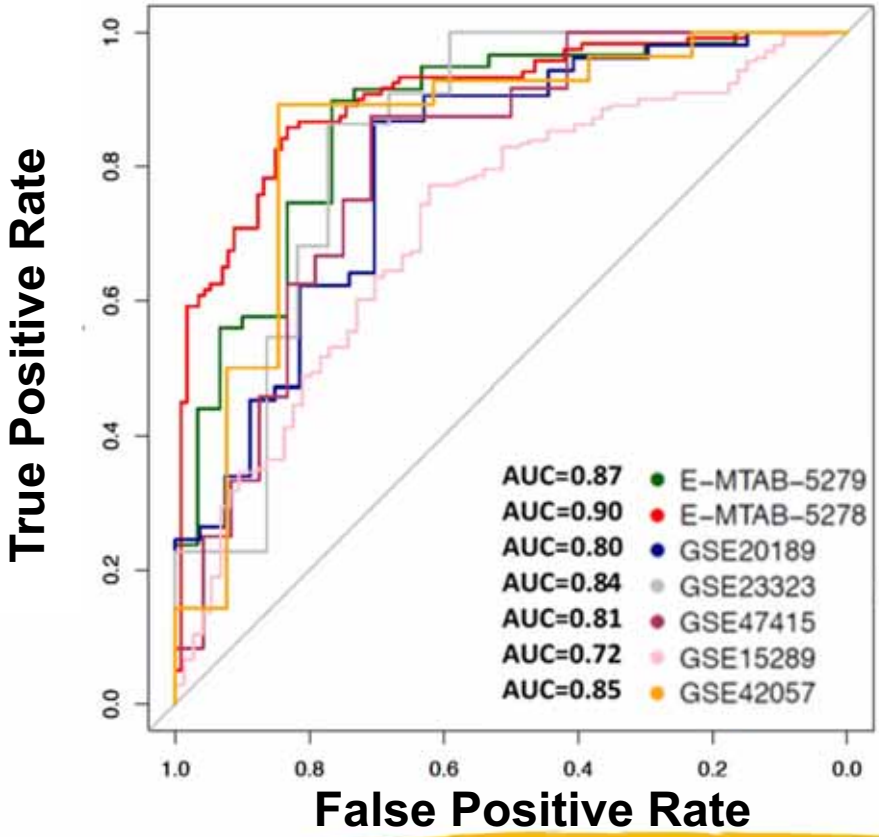
RAI SERVICES COMPANY



- The accuracy of 8-18 gene signatures were greater than 0.95
- 16 gene signature performs best in independent validation among 8-18 gene signatures

Independent Validation of 16 Gene Signature

RAI SERVICES COMPANY



- All AUC of independent validation datasets are greater than 0.80 except GSE15289 (AUC=0.72)

Independent Validation of 16 Gene Signature

RAI SERVICES COMPANY

Dataset	Actual	Predicted SMK	Predicted NTC	Accuracy
E-MTAB-5279	29 NTC	3	26	0.81
	30 SMK	23	7	
E-MTAB-5278	60 NTC	11	49	0.82
	114 SMK	98	16	
GSE20189	21 NTC	0	21	0.79
	27 SMK	15	12	
GSE23323	22 NTC	3	19	0.82
	22 SMK	17	5	
GSE47415	24 NTC	8	16	0.71
	24 SMK	18	6	
GSE15289	211 NTC	41	170	0.73
	74 SMK	38	36	
GSE42057	13 SMK	11	2	0.73

- **Accuracy of independent validation datasets are all greater than 0.70**

Independent Validation of 16 Gene Signature using Former Smokers' Data

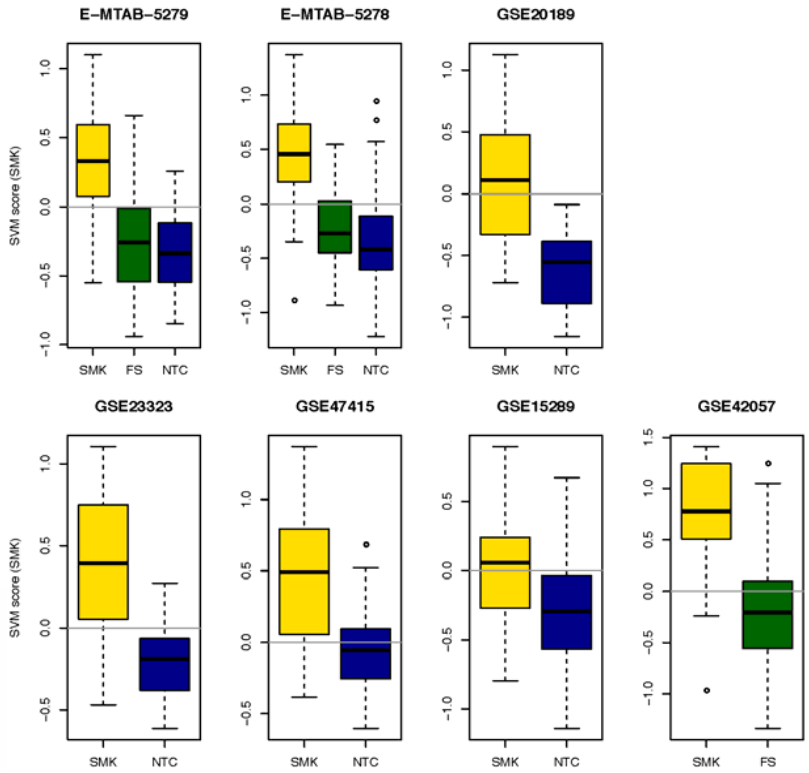
RAI SERVICES COMPANY

Dataset	Former Smokers (as proxy to NTC)	True rate
E-MTAB-5278	15	0.75
E-MTAB-5279	7	0.77
GSE42057	9	0.68

- **All true rates are >0.65**

Independent Validation of 16 Gene Signature

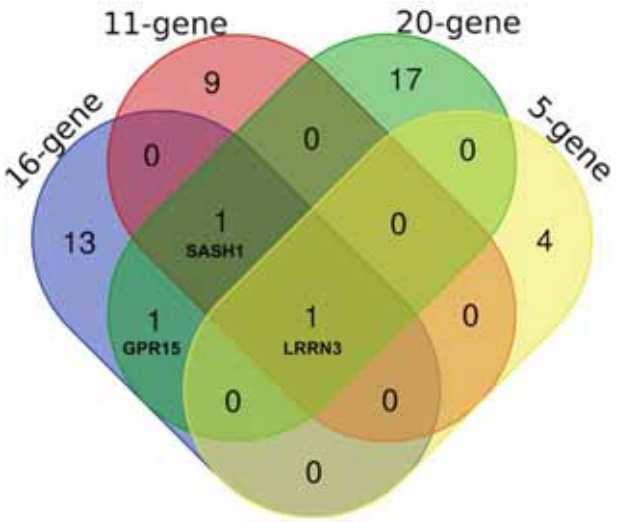
RAI SERVICES COMPANY



- 16 gene signature scores (SVM scores) can distinguish SMK from NTC in all 8 datasets
- All p values comparing gene signature scores of SMK with NTC are less than 0.05
- SVM scores of FS (former smokers) are similar to NTC

16 Gene Signature

RAI SERVICES COMPANY



Gene Signatures	Gene Symbols
16-gene	**LRRN3**, **SASH1**, **GPR15**, STAB1, NDST2, COCH, PHACTR1, MKRN3, EPB41L3, PTGDR, PAFAH2, CDK8, TPSG1, TBX21, GZMM, NCBP1
11-gene	**LRRN3**, **SASH1**, PALLD, RGL1, TNFRSF17, CDKN1C, JCHAIN, RRM2, ID3, SERPING1, FUCA1
20-gene	**LRRN3**, **SASH1**, **GPR15**, GPM6B, RIPK2, ASGR2, PTGDS, ADGRG1, ERAP1, PID1, MS4A4A, CLEC1B, CENPK, ITGB8, S1PR3, TOB1, PCGF3, FCRL5, AP5M1, HLA-DPB2
5-gene	**LRRN3**, MUC1, GOPC, LEF1, CLDND1

- **LRRN3 and SASH1 were shared among 16, 11, and 20 gene signatures.**

LRRN3: leucine rich repeat neuronal 3
SASH1: SAM and SH3 domain containing 1
GPR15: G-protein coupled receptor 15

Cancer-related Genes in the Signature

RAI SERVICES COMPANY

Gene name	Disease Type	Full Name and Biological Functions
GZMM	Cancer	Granzyme M, a serine protease playing important role in innate immunity
LRRN3		Leucine-rich repeat neuronal protein 3, a membrane protein involved in cognitive and immune functions
SASH1		SAM and SH3 Domain Containing 1, a scaffold protein involved in immune signaling
PTGDR		Prostaglandin D2 receptor, a membrane receptor protein involved in inflammation signaling
EPB41L3		Erythrocyte member protein band 4.1-like 3, a membrane protein with tumor suppressive properties
TBX21		T-box transcription factor 21, a transcription factor regulating immune function
CDK8		Cyclin-dependent protein kinase 8, a colorectal cancer oncogene and putative tumor suppressor gene in other cancers
STAB1		Stabilin-1, a transmembrane receptor protein with a role in regulating angiogenesis

Lung and Cardiovascular Diseases-related Genes in the Signature

RAI SERVICES COMPANY

Gene name	Disease Type	Full Name and Biological Functions
TPSG1	Lung Disease	Tryptase Gamma 1, a trypsin-like serine protease implicated as mediators in the pathogenesis of inflammatory disorders
PAFAH2		Platelet-activating factor acetylhydrolase isoform 2, an intracellular enzyme involved in platelet homeostasis
NCBP1	Cardiovascular Disease	Nuclear cap binding protein subunit 1, a component of the nuclear cap-binding complex involved in various processes including translation regulation and pre-mRNA splicing
PHACTR1		Phosphatase and actin regulator 1, an intracellular protein associated with coronary artery disease
MKRN3		Makorin Ring Finger Protein 3, a ubiquitin ligase associated with coronary artery disease and cancer risk

Summary

RAI SERVICES COMPANY

- An optimal machine learning algorithm was identified for deriving a gene signature from blood-based gene expression data sets.
- A 16-gene signature was developed to characterize biological responses to chronic cigarette smoking.
- It demonstrates consistent and robust performance across seven independent validation datasets.
- This gene signature can serve as a BoPH to differentiate biological responses in consumers of different types of tobacco products.

Acknowledgements

- AccuraScience
  - Justin Li

- RAIS
  - Wei Chao

RAI SERVICES COMPANY

Thank you!
Questions?