Identification of Candidate Biomarkers of Tobacco Smoking-Related Diseases through Gene/Disease Associations

Jeffery Edmiston[1], Walter Jessen[2], Sinnathamby Gomathinayagam[3], William Rees[1], Mohamadi Sarkar[1]

[1]Altria Client Services LLC, Center for Research and Technology, 601 East Jackson Street, Richmond, VA 23219, USA
[2]Laboratory Corporation of America & [2]Covance Greenfield Laboratories, 671 S. Meridian Road, Greenfield, IN

CORESTA SSPT 2017, October 8-12, 2017, Kitzbühel, Austria

This poster may be accessed at www.altria.com/ALCS-Science

Introduction

Biomarkers can be useful tools in measuring the biological effects of tobacco product use. Although there is a small group of biomarkers that are typically used to assess the biological effect of smoking tobacco, there are few publications summarizing recent developments in this area. The purpose of this project was to investigate a data-mining approach to analyze recent (past 5 years) publications to identify potential biomarkers associated with tobacco smoking-related disease mechanisms.

Methods

- Candidate biomarkers were identified by investigating gene/protein associations in the published literature for three indications and a term: chronic obstructive pulmonary disease (COPD), cardiovascular disease (CVD), lung cancer (LC), and tobacco smoke (TS).
- Approach used query terms, association words, and database terms using the PolySearch web server and PolySearch Relevancy Index (PRI)[1] {pattern recognition relevancy ranking} to identify gene-disease/term associations in the published literature (MEDLINE) over the past 5 years (title and abstract only, up to 5000 abstracts for each search). The search was limited to the ~50 highest ranked gene/proteins for each indication and term (Table 1).
  - COPD, synonym keywords: chronic obstructive pulmonary disease; COAD; COLD – chronic obstructive lung disease; COPD; COPD – chronic obstructive pulmonary disease; chronic obstructive airways disease; chronic obstructive lung disease; chronic airflow limitation; chronic airway disease; chronic airway obstruction; chronic irreversible airway obstruction; chronic obstructive airway disease; pulmonary disease, chronic obstructive
  - Cardiovascular disease, synonym keywords: cardiovascular disease; circulatory system disorder; cardiovascular system diseases; circulatory disorders; circulatory disease; circulatory system diseases; diseases of the circulatory system; disorder of the circulatory system; circulatory disorder
  - Lung Cancer, synonym keywords: lung cancer; cancer of lung; cancer of the lung; cancer, lung; cancer, pulmonary; lung cancers; malignant lung neoplasm; malignant lung tumor; malignant neoplasm of the lung; malignant tumor of the lung; malignant neoplasm of lung; malignant tumor of lung; pulmonary cancer; pulmonary cancers
- Identified gene/proteins for each indication were then compared with the TS-associated genes/proteins (Figures 1 & 2).
- Disease models (i.e. protein-protein interaction networks) based on published peer-reviewed research curated by Thomson Reuters were constructed to simulate disease biology using MetaCore.[2]
- Dijkstra's shortest path algorithm[3,4] was used to create a network for each gene set. Each indication model is tissue specific: COPD, LC, and TS models were constructed using genes expressed in the lung; the CVD model was constructed using genes expressed in the cardiovascular system. (Figures 3 & 4).
- Disease models were then compared to the TS model to identify overlapping gene/proteins (Figure 5).
- A functional analysis was performed to identify enriched pathways (Biocarta, KEGG, Panther or Reactome) for the overlapping gene/proteins using the Database for Annotation, Visualization, and Integrated Discovery (DAVID).[5,6] Most enriched pathways were associated with immune and inflammatory response with the highest rankings for the JAK-STAT and Cytokine-cytokine receptor interactions (Table 2).

REFERENCES

1. Cheng D et al. (2008) Nucleic Acids Res. 36:W399-405.
2. http://thomsonreuters.com/en/products-services/pharma-life-sciences/pharmaceutical-research/metacore.html
3. Dijkstra EW (1959) Numerische Mathematik. 1: 269–271.
4. Ekins S et al. (2006) Xenobiotica. 36:877-901.
5. Huang DW et al. (2009) Nature Protoc. 4:44-57.
6. Huang DW et al. (2009) Nucleic Acids Res. 37:1-13.

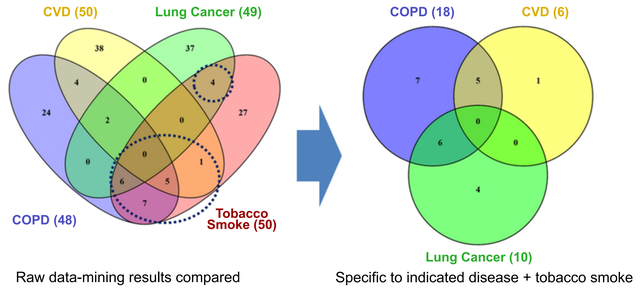
Results

Table 1. PolySearch Identified ~50 Gene/Proteins for Each Condition or Term

Association	PolySearch Relevancy Index score threshold	Gene number
Chronic obstructive pulmonary disease (COPD)	400 (top score: 2939)	48
Cardiovascular disease (CVD)	116 (top score: 890)	50
Lung cancer (LC)	347 (top score: 11805)	49
Tobacco smoke (TS)	45 (top score: 290)	50

Results

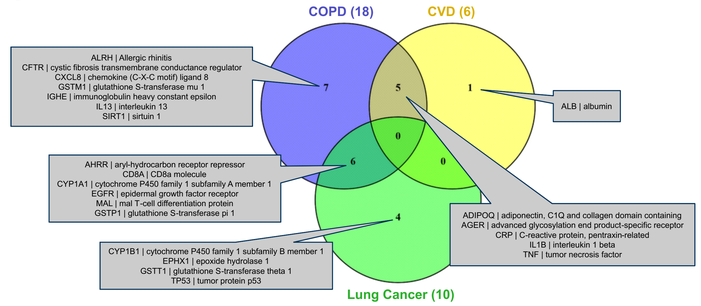
Figure 1. Venn Diagram Showing the Overlap Between the Identified Disease and Tobacco Smoke Gene/Proteins



Raw data-mining results compared

Specific to indicated disease + tobacco smoke

Figure 2. Identified Disease and Tobacco Smoke Related Gene/Proteins



Specific to indicated disease + tobacco smoke

Figure 3: Simple Modeling of Disease Biology

Data mine MEDLINE to identify high-scoring gene-disease associations (Figure 1)

candidate biomarkers

Build models by adding pathway "steps" between disease-associated genes

Example: Simple COPD model

Figure 4. Complex Modeling of Disease Biology

Complex candidate biomarker models include additional interacting gene/protein steps

All simple model genes are contained in the complex model

Example: Complex COPD model

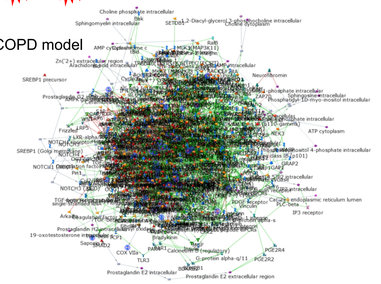
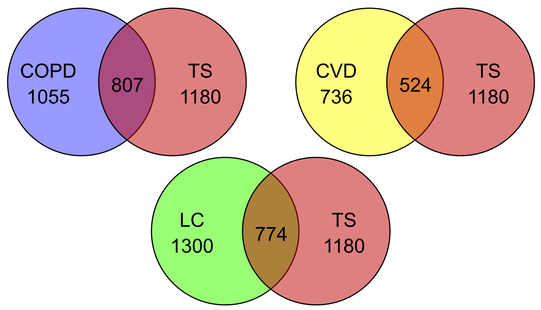


Figure 5: Complex Models Overlapping with Tobacco Smoke



Numbers indicate the number of Entrez gene IDs for each complex model.

Table 2. Top-Scoring Enriched Pathways for TS and All Three Complex Disease Models: JAK-STAT Signaling and Cytokine-Cytokine Receptor Interaction (A&B)

A: JAK-STAT signaling pathway genes in each condition evaluated

Condition	Genes	FDR (B&H)*
COPD+TS	AKT1, AKT2, AKT3,CD127, CD25, CSF2, IFNG, IL10, IL12A, IL12B, IL13, IL2, IL4, IL6, IL6R, LEP, PRL, STAT1, STAT3, STAT5	1.3E-15
CVD+TS	AKT1, AKT2, AKT3, CD25, IFNG, IL10, IL2, IL6, IL6R, LEP, PRL, STAT1, STAT3, STAT5	5.70E-10
LC+TS	AKT1, AKT2, AKT3, CD25, CSF2, IFNG, IL10, IL13, IL2, IL4, IL6, IL6R, LEP, STAT1, STAT3, STAT5	3.10E-12

B: Cytokine-cytokine receptor interaction genes in each condition evaluated

Condition	Genes	FDR (B&H)*
COPD+TS	CD127, CD25, CSF2, EGF, IFNG, IL10, IL12A, IL12B, IL13, IL1B, IL1R1, IL2, IL4, IL6, IL6R, IL8, KDR, LEP, PRL, TNF, TNFRSF1A, VEGFA	6.10E-14
CVD+TS	CD25, CD95, HGF, IFNG, IL10, IL18, IL1B, IL1R1, IL2, IL6, IL6R, IL8, KDR, LEP, PRL, TNF, TNFRSF1A, VEGFA	2.10E-11
LC+TS	CD95, CCL5, CSF2, EGF, HGF, IFNG, IL1R1, IL1B, IL10, IL13, IL2, CD25, IL4, IL6, IL6R, IL8, KDR, LEP, TNF, TNFRSF1A, VEGFA	8.80E-15

*Benjamini and Hochberg False Discovery Rates (FDR (B&H)) (Benjamini Y & Hochberg Y (1995) J Royal Stat Soc B. 57:289–300)

Limitations

- Associations were extracted from abstracts only and the search was limited to the past 5 years.
- Identified targets were based on a computer search algorithm using word and sentence structure with no assessment of directional changes or reproducibility of changes.
- Identified targets will require additional research to confirm their utility in tobacco product assessments.

Conclusions

- We identified 18 COPD + Tobacco Smoke targets, 6 CVD + Tobacco Smoke targets, and 10 LC + Tobacco Smoke targets.
- We identified many overlapping gene/proteins between the three diseases and tobacco smoke.
- The top-scoring enriched pathways (DAVID 6.8) for all three disease conditions were JAK-STAT signaling and cytokine-cytokine receptor interactions.
- This literature mining and data analysis approach is a potential tool for the identification of emerging biomarkers of smoking tobacco-related disease mechanisms.