



## Potential of Genomic Data on PAHs to Inform Cumulative Assessment

*Lyle D. Burgoon, Ph.D.  
National Center for Environmental Assessment  
Office of Research and Development  
United States Environmental Protection Agency*



## **Disclaimer**

Does not necessarily reflect US Government Policy, US EPA Policy, or the views of the US EPA or the US Government.



## Overview

- PAH Background
- Brief Introduction to the NexGen Program
- Overview of PAH Data
- Analysis Overview
- Pilot Study
- Future Directions
- Summary

## Background on Polycyclic Aromatic Hydrocarbons (PAH)

- PAHs are produced through incomplete combustion of organic compounds
- Large body of evidence demonstrating the human carcinogenicity of several PAH-containing mixtures, such as soot, coal tars, coal-tar pitch, and household combustion of coal.
- Benzo[a]pyrene: the most well studied carcinogenic PAH
- U.S. EPA Superfund program routinely measure 15-17 PAHs in environmental media and use these data to estimate PAH population cancer risks

## Occupational Exposure to PAHs Associated with Induced Cancer

- Aluminum production
- Chimney sweeping
- Coal gasification
- Coal-tar distillation
- Coke production
- Iron and steel founding
- Paving and roofing with coal tar pitch
- Carbon black and diesel exhaust

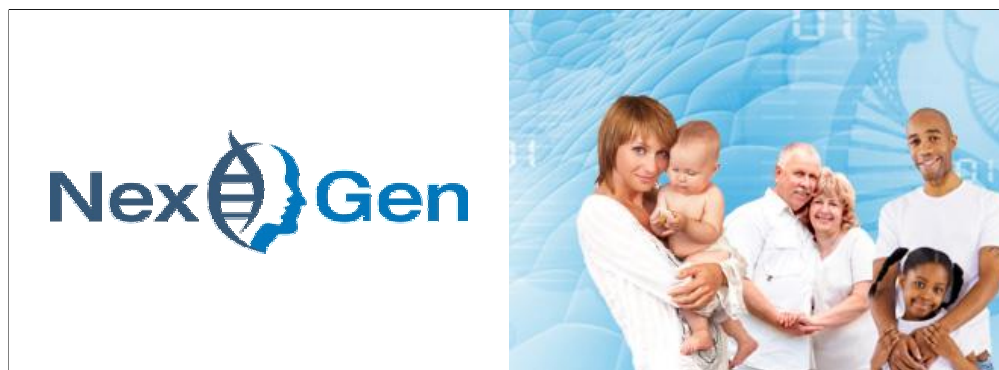


© Walt Disney, Mary Poppins 1964



## About this Pilot Project

- Advancing the Next Generation of Risk Assessment Program (NexGen)
  - PAH lung cancer risk assessment prototype
- *NexGen's Goal*
  - *To advance risk assessment science via incorporation of recent progress in molecular systems biology*
  - *Collaborative effort with several federal and state agencies*



## NexGen Partners

- US Environmental Protection Agency, Office of Research and Development
- National Institutes of Environmental Health Sciences & National Toxicology Program
- Centers for Disease Control & Agency for Toxic Substances and Disease Registry
- NIH Chemical Genomics Center
- California's Environmental Protection Agency, Office of Environmental Health Hazard Assessment
- FDA National Center for Toxicological Research
- Department of Defense
- Health Canada, Government of Canada



## Today's Focus

- Predicting lung cancer phenotype using toxicogenomic data
- *In vivo* human exposures to cigarette smoke
- Compared to small cell lung carcinoma

## Identifying the Data for Analysis: A Pilot

- Genes Expression Omnibus
  - NIH gene expression data repository
- ArrayExpress
  - European Bioinformatics Institute (EBI) gene expression data repository
- Search goal
  - Experimental data
    - Treatment levels are at least “smoker” and “non-smoker”
  - Disease marker data
    - Phenotype levels are “normal” and “lung cancer” or “small cell lung carcinoma”
- Searched:
  - “cigarette smoke lung cancer” (47 entries)
  - “cigarette smoke” (444 entries)
  - “lung cancer” (5,046)
  - “small cell lung carcinoma” (166)

## Identifying the Data for Analysis: A Pilot (cont'd)

### Overall Goal

- Predict known/accepted public health outcomes from PAH exposures using molecular systems biology and high throughput assay data

### Aims of this pilot

- Predict lung cancer phenotype using cigarette smoking data in humans
- Explore and develop network theoretic methods for data mining, pattern recognition to predict public health outcomes

## Decision Point: Data Used for Pilot Data Analysis

- Technology: Affymetrix microarray data
  - Near-term: All applicable microarray technologies
  - Middle-term: All expression technologies (next generation sequencing)
  - Longer-term: All omics technologies (e.g., proteomics, metabolomics)
- Published studies
  - All data must be from peer-reviewed published studies
- Rawest usable data
  - The rawest usable data must be available for the respective technology
  - Rawest usable data = data in non-binary, free-text, open source binary, or verified and validated reverse engineered binary format that has not been standardized or normalized

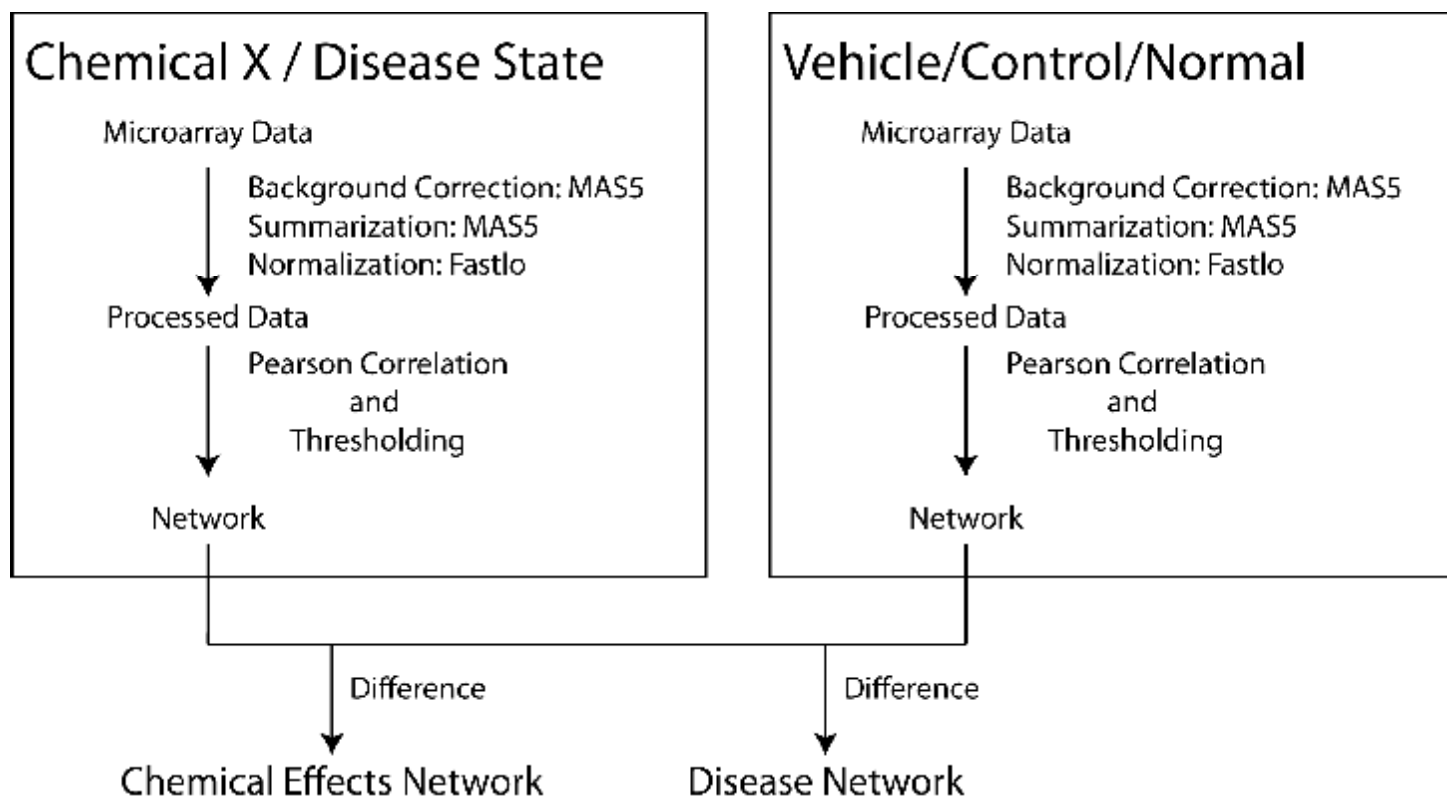
## Datasets for Pilot Analysis

| GSE      | Species | Mixture         | Model                   | Characteristics   |
|----------|---------|-----------------|-------------------------|---|
| GSE10072 | Human   | Cigarette smoke | Lung tissue             | Smokers, non-smokers; adenocarcinoma positive and negative patients |
| GSE5060  | Human   | Cigarette smoke | Large airway epithelium | Phenotypically normal smokers and non-smokers                       |



# ANALYSIS OVERVIEW

## Conceptual Overview of the Analysis Method

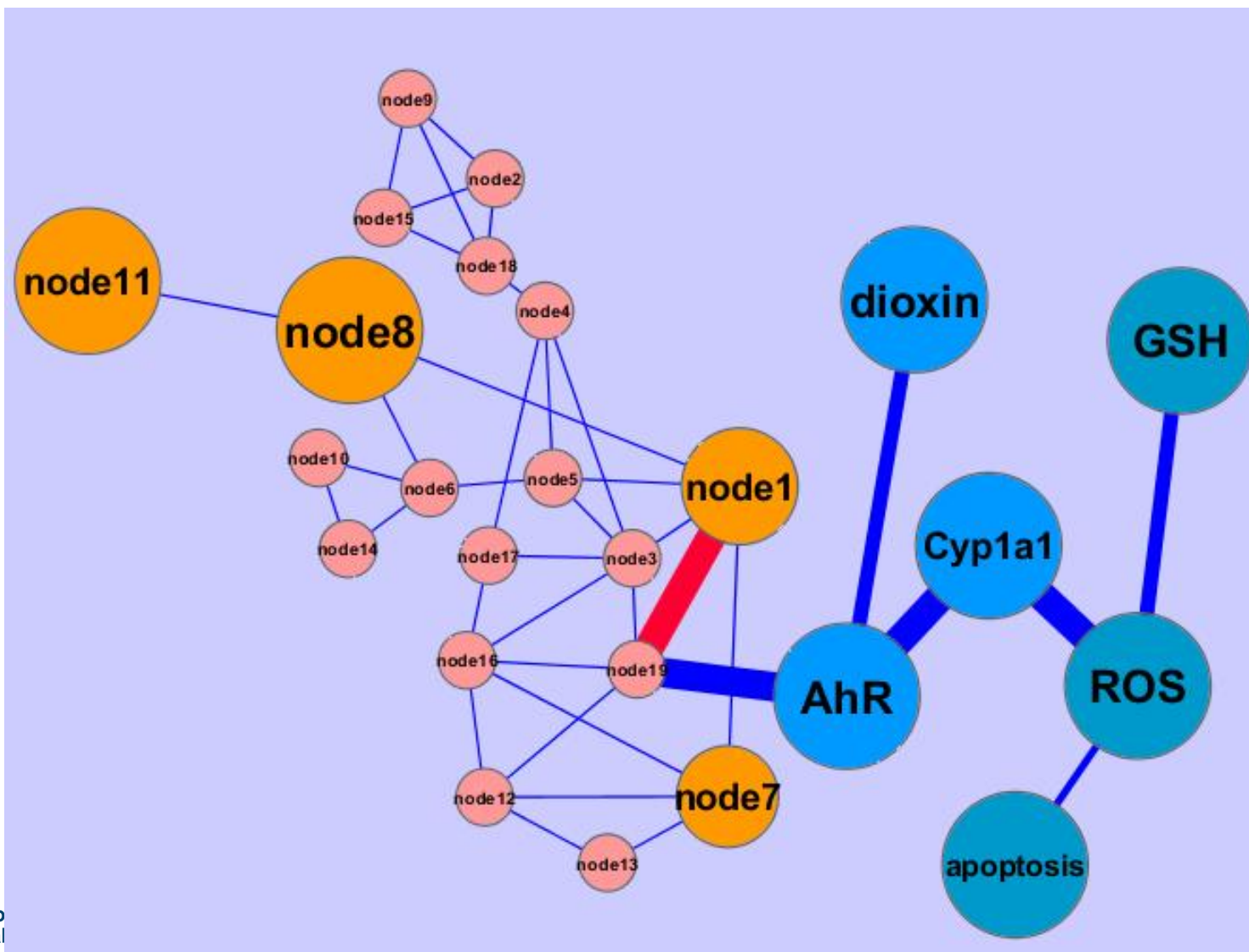
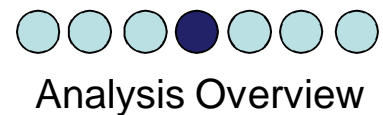


## Laying the Theoretical Foundations

- Diseases, no matter their etiology, result from the deviation of a biological system away from homeostasis
- Cellular pathways are a means of communicating information
- Exposures to stressors do not create new cellular pathways, *per se*
- Stressors elicit disease outcomes by “encouraging” or “forcing” information/signals to travel through pathways that are not normally activated/used in a particular context



# Contrived Graphical Example



## The Biological Problem

- Assumptions:
  - We have complete knowledge of what “normal” is
  - We know all of the molecular and signaling “connectivities”
  - We know how a stressor will impact information flow through the system
- In reality we know...
  - Very little, actually
- What we don’t know
  - The connectivity of the “normal” state
  - How information flows through the cell
  - All of the functions of all genes
  - The complete interactome

## Network Hypothesis

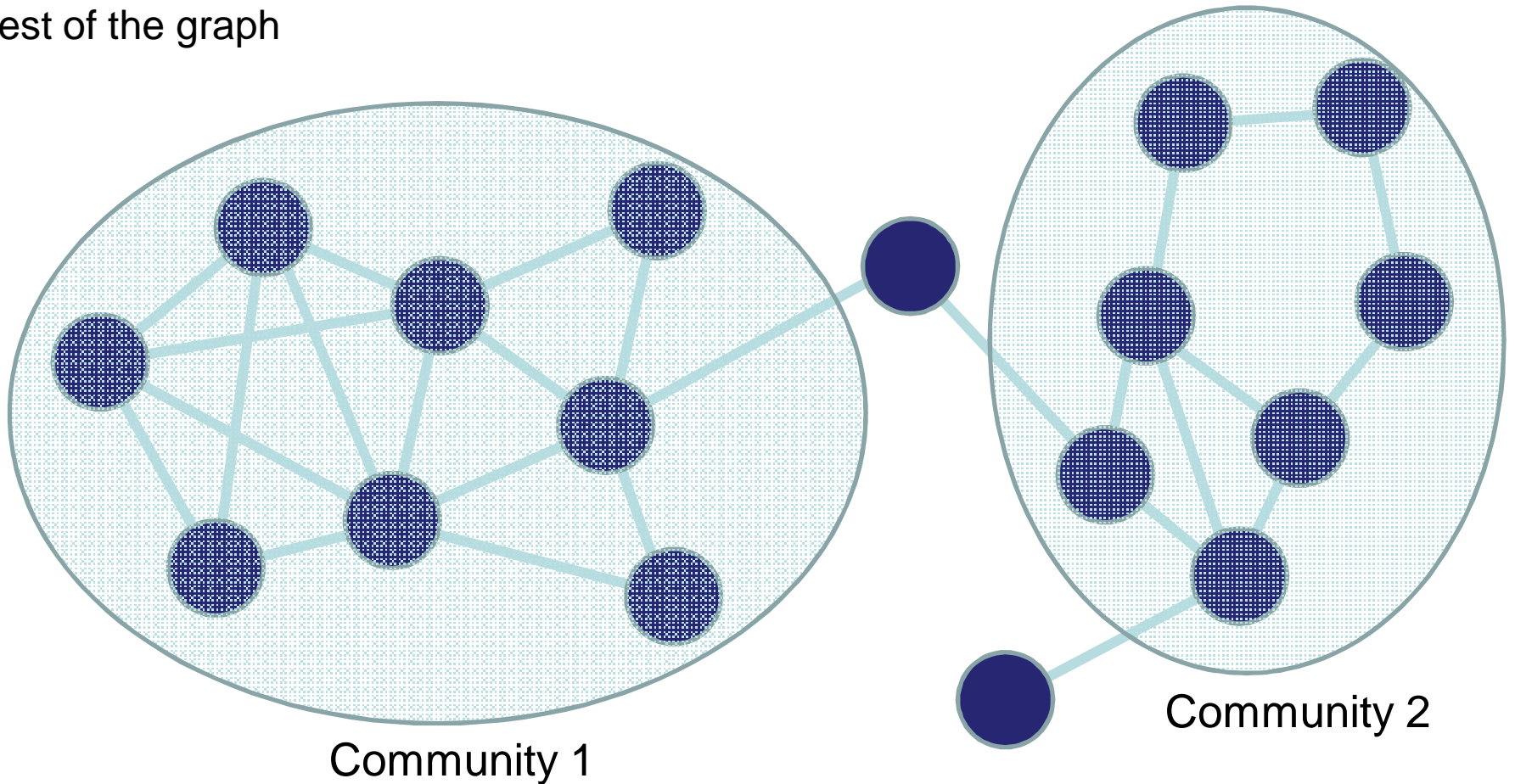
- Correlation does not identify all of the cellular connections
  - Correlation does not result in a complete cellular information network
- Correlation networks of “normal” state can be compared to “disease” and “exposed” states
  - Changes in connectivity may
    - Predictive of hazard
    - Adaptive responses
    - Part of the mode of action or pathogenesis
- Comparing disease and “exposure-response” networks may enable us to predict disease outcomes

## What About Biomarkers?

- Agglomerative biomarkers (ABs)
  - Biomarker sets that are mechanistically tied to a disease, no matter what the stressor
  - May be defined from omics datasets
  - Prefer to define based on network community analysis
- An AB will likely be associated with multiple diseases
  - Define probabilities of AB association to specific diseases
    - By using the wealth of NIH-funded disease research
    - Analysis using frequent itemset mining (e.g., Apriori algorithm)
- Associate a chemical exposure (at specific doses) to disease probabilities
  - Based on ABs

# Community

- A subgraph where nodes are more connected to each other than the rest of the graph

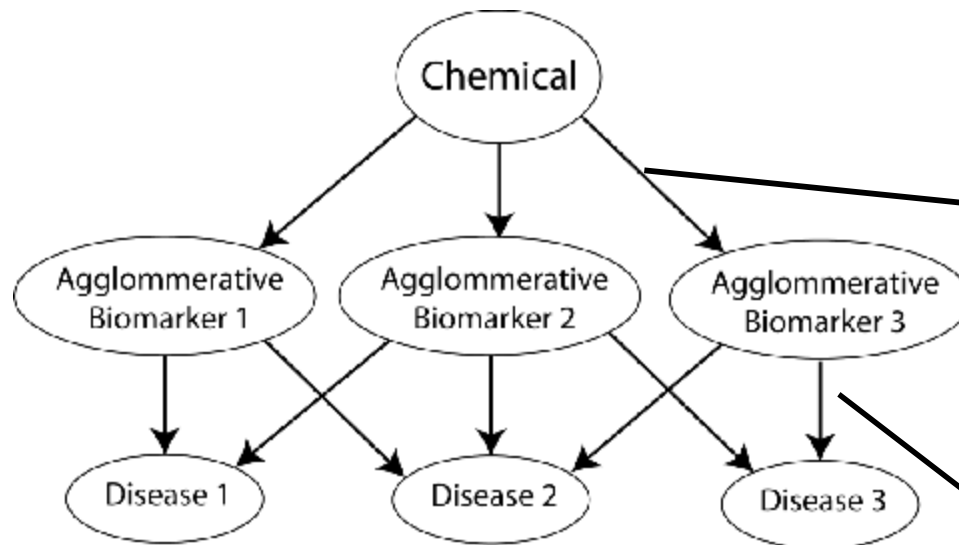


## Agglomerative Biomarkers

- Agglomerative Biomarkers = network communities
- Can easily identify “similar” agglomerative biomarkers from multiple chemicals, diseases or other non-chemical stressors
- Predict public health outcomes using Bayesian Belief Networks

# Bayesian Networks for Probabilistic Chemical-Disease Prediction

## Contrived Model



Probabilistic Community  
Similarity  
and  
Probability of AB  
associated with chemical

Probability AB  
associated with disease

$P(D_d, C | AB_i)$  the probability of interest



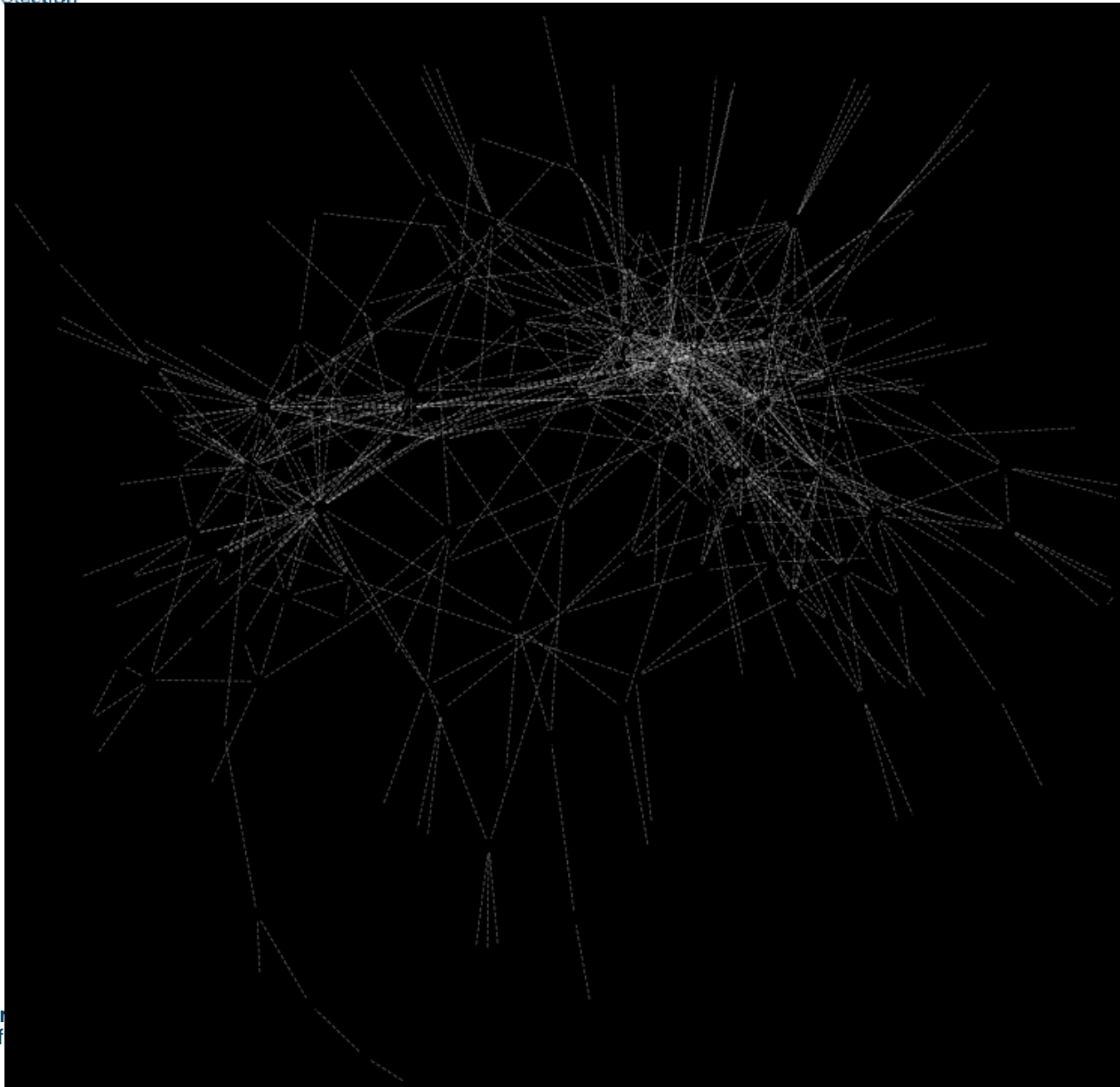
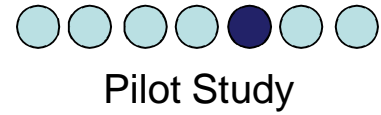
# PILOT RESULTS



## Lung Cancer Dataset

- Landi, et al. PLoS One 2008 3(2): e1651.
- Adenocarcinoma of the lung
  - Smokers and former smokers: 24 samples
- Phenotypically normal lung
  - Non-smokers: 15 samples

# Lung Cancer Network

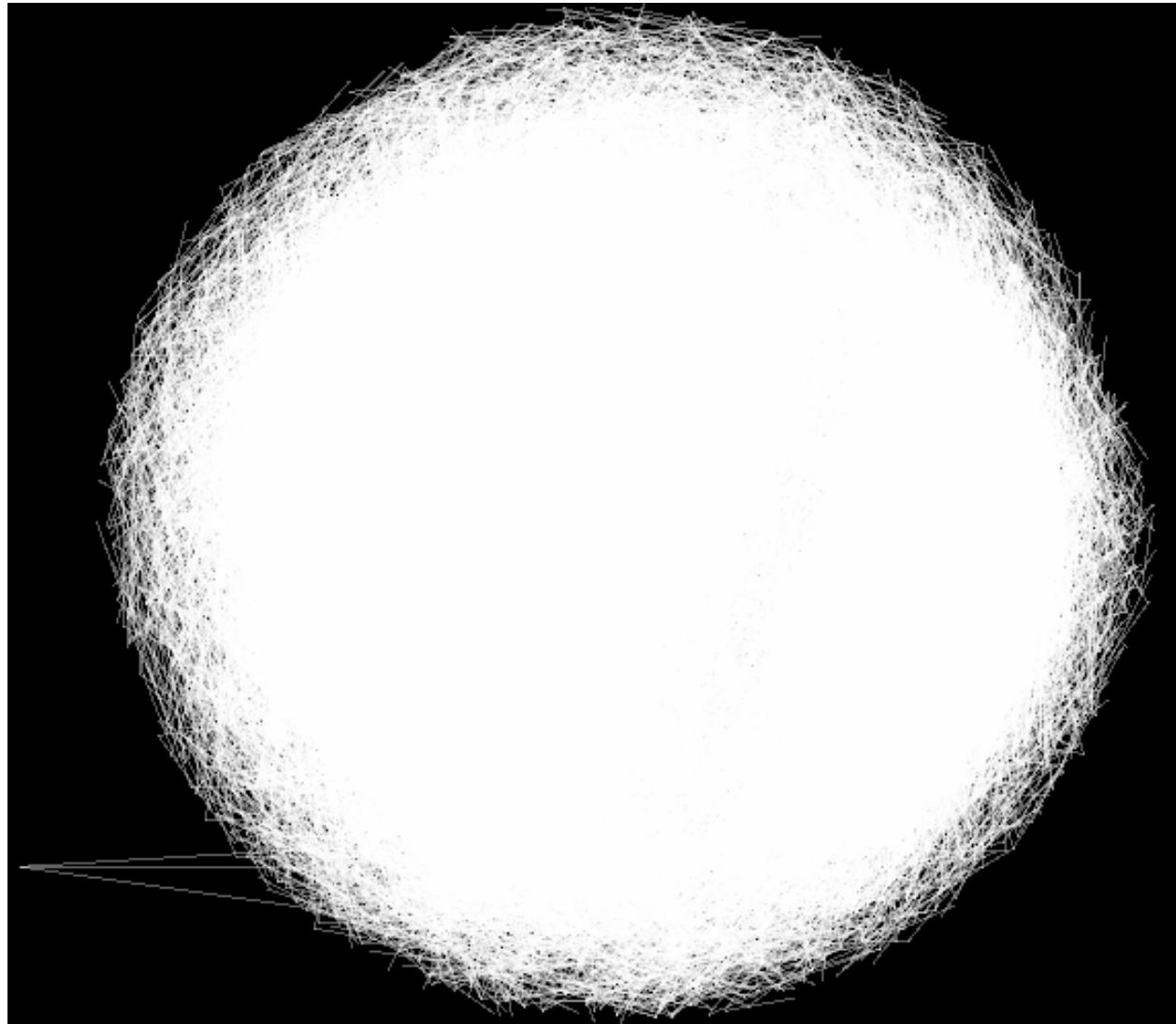
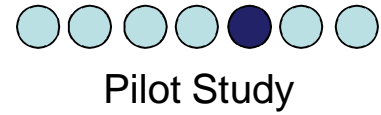




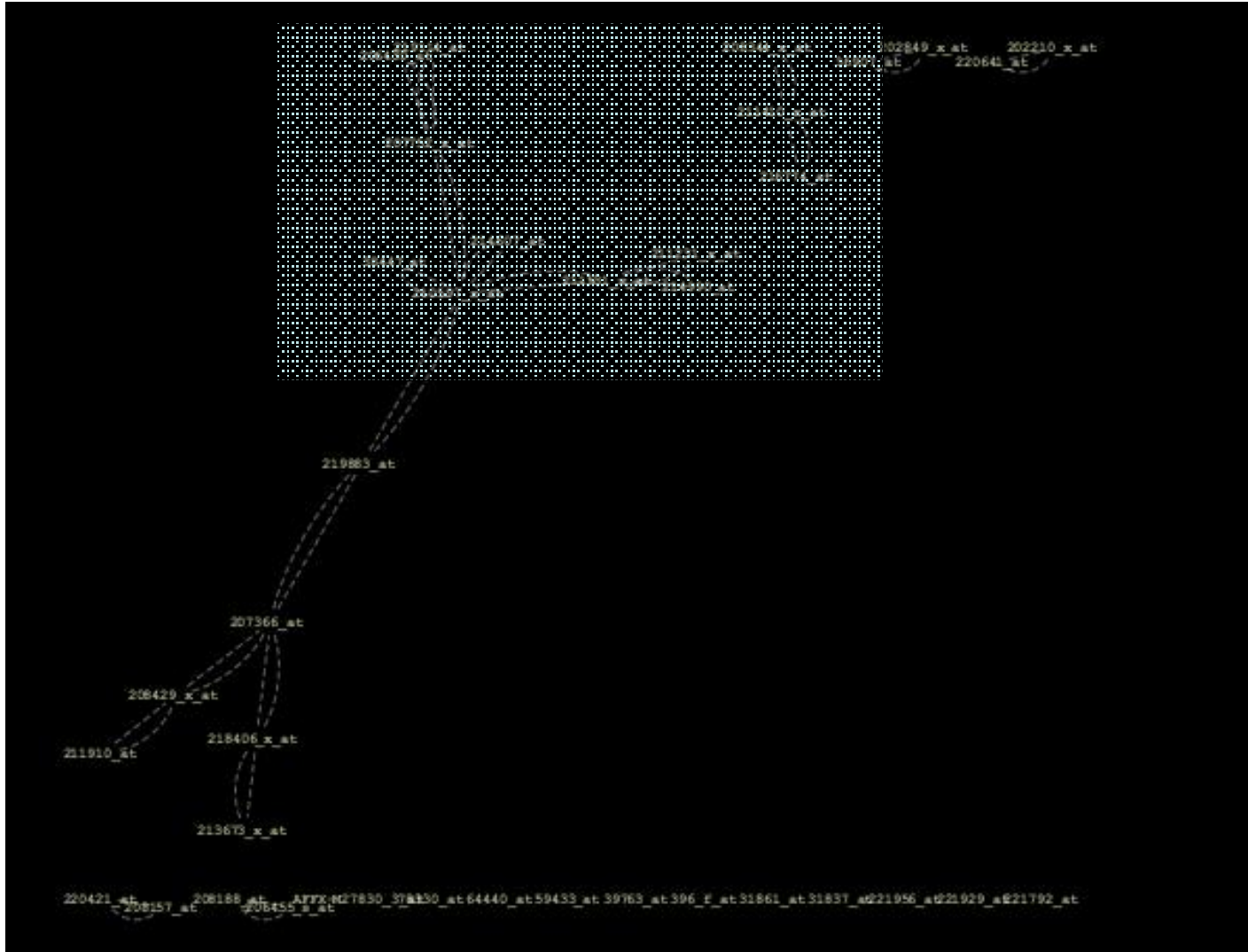
## Cigarette Smoking Dataset

- Smokers
  - 10 samples
  - 15-45 pack-years
- Non-smokers
  - 12 samples

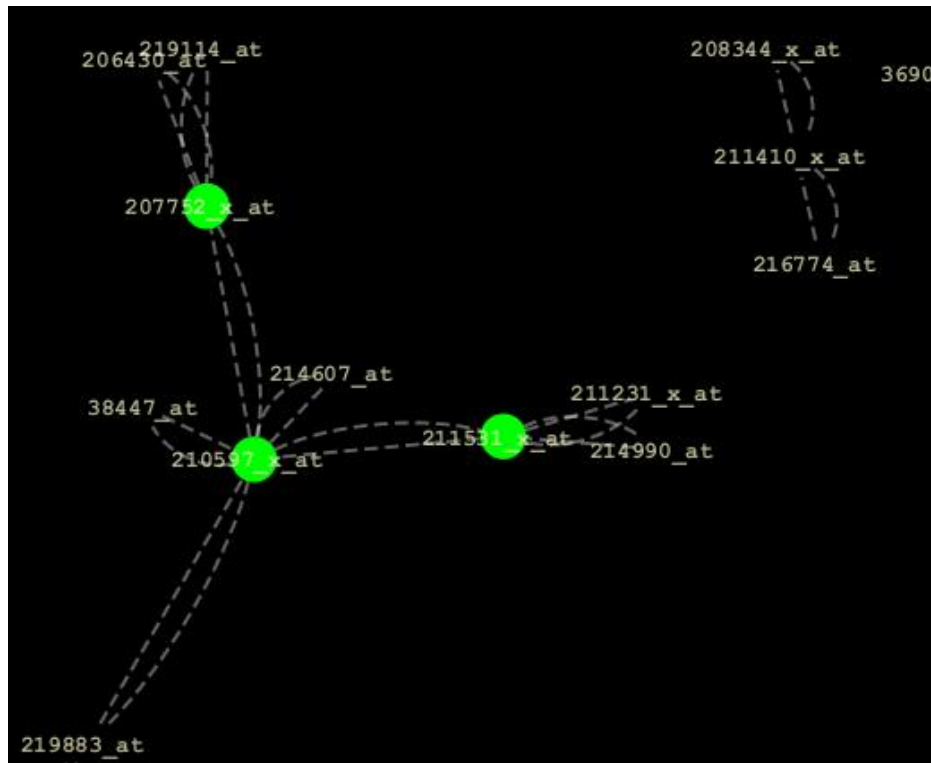
# Cigarette Smoking Network



# Intersection of Lung Cancer and Smoking Networks

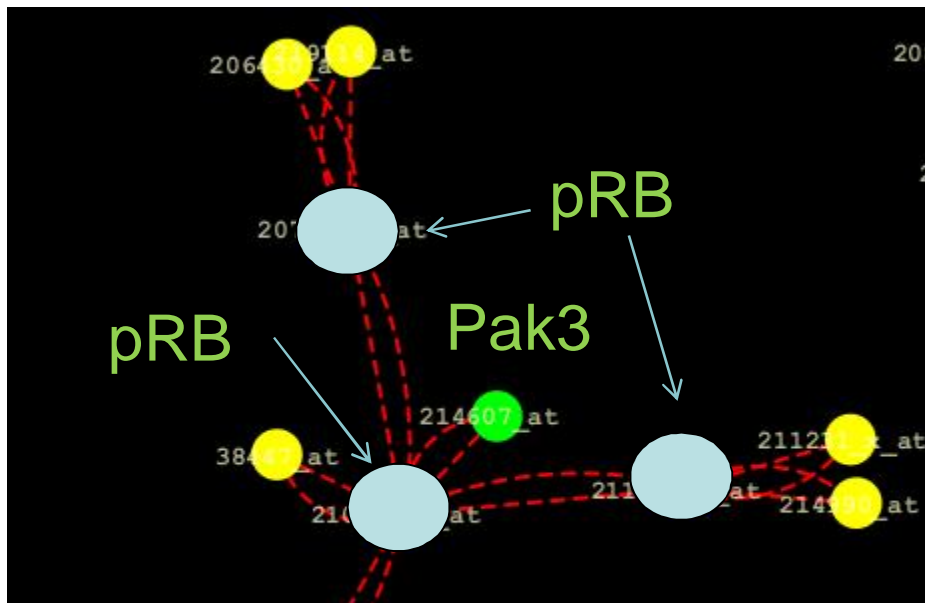


## Feature Analysis



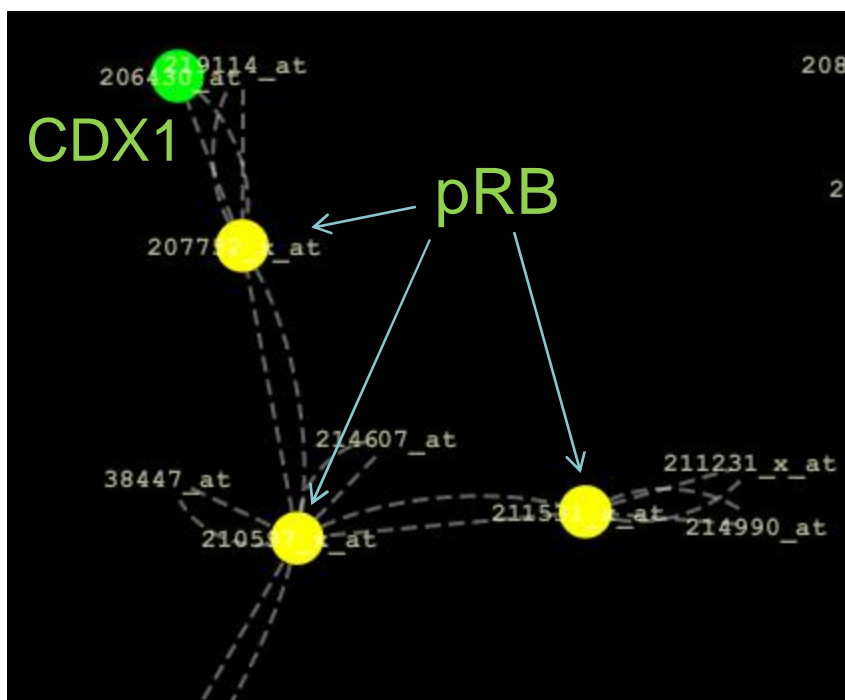
- pRB (retinoblastoma) protein
- Checkpoint that regulates G1 to S phase cell cycle transition
- Loss of RB function allows for unchecked progression to S phase

## Pak3



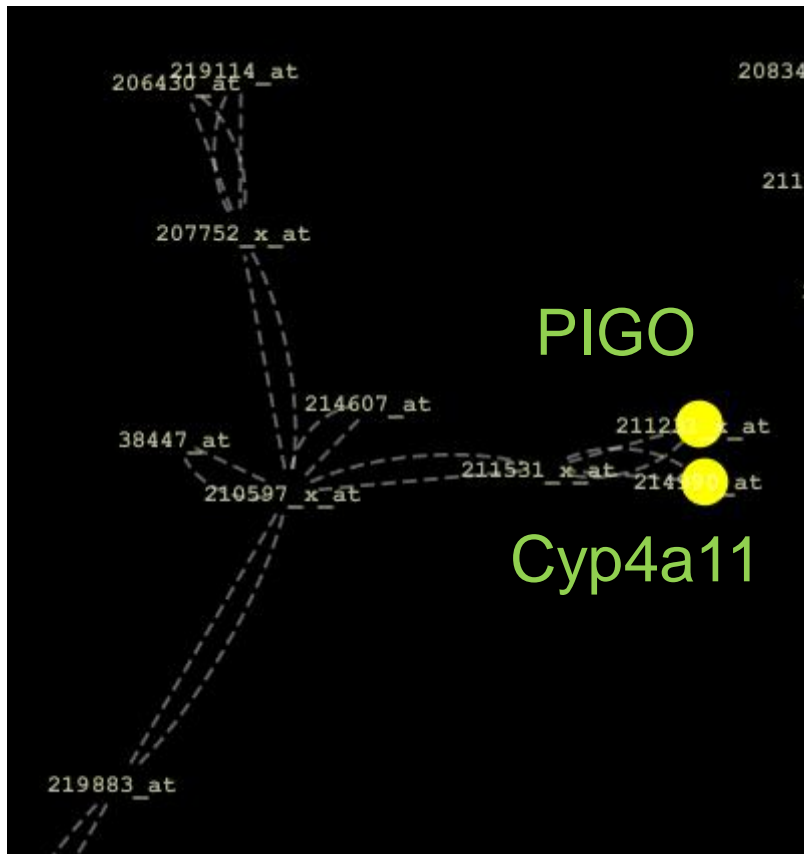
- Pak3 activated by p21, Cdc42, and RAC1
- Involved in cell motility and proliferation (cell cycle)
- We see Pak3 expression correlated with pRB expression

## CDX1 (caudal type homeobox 1)

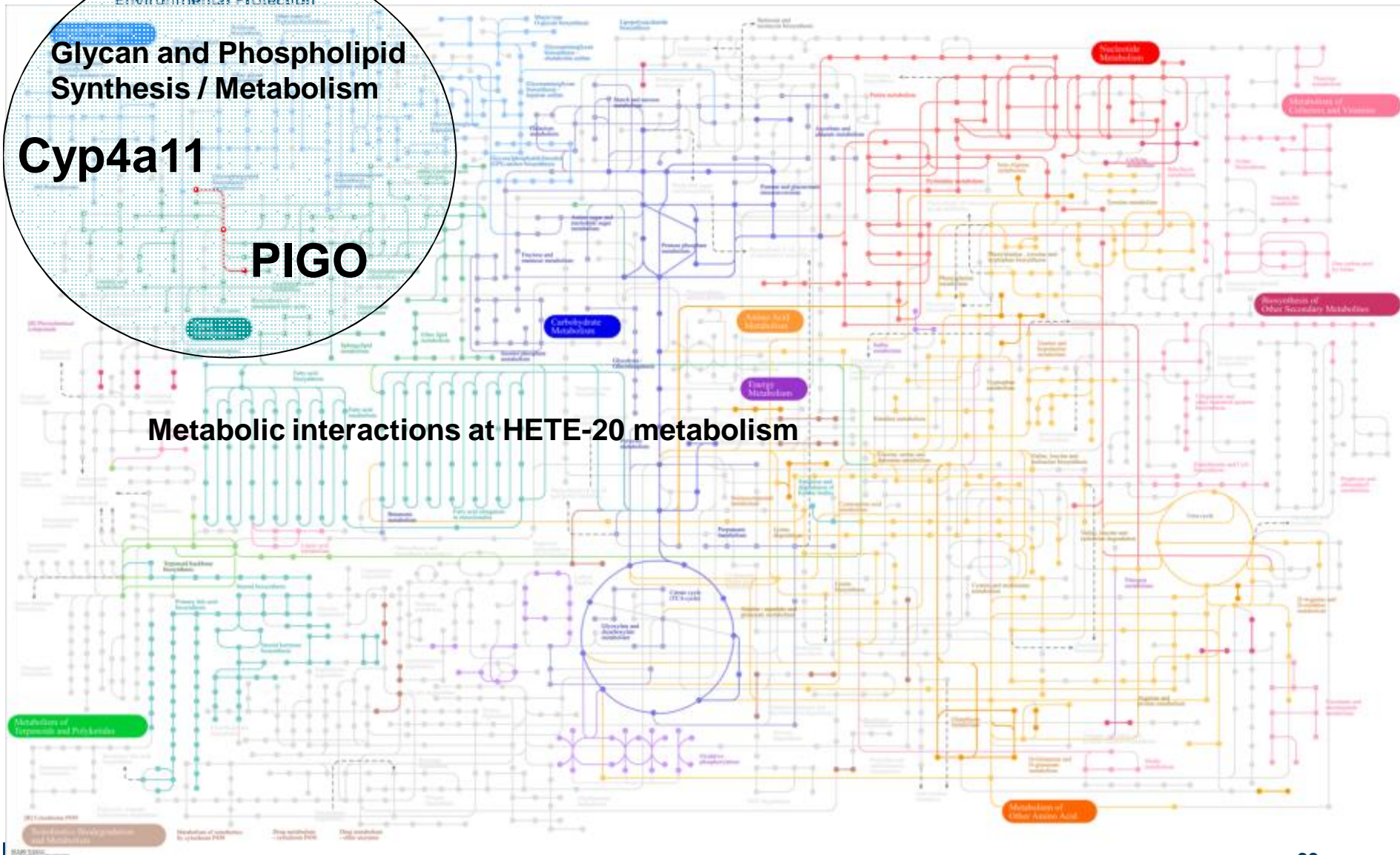


- CDX1 normally expressed in intestine
- Correlated in expression with MUC6 (mucin) in many lung cancers
- CDX hypothesized to play a role in aberrant MUC6 expression
- Expression of MUC2 and MUC6 in small adenocarcinomas associated with poor prognosis

## PIGO and Cyp4a11



- Phosphatidylinositol glycan anchor biosynthesis, class O
- Cytochrome P450 4A11
- Why would these be connected??





## Brief Review

- Identified several network features that may associate phenotypically normal lung expression from smokers with lung cancer disease phenotype
- Intersecting networks from the smoker and lung cancer datasets

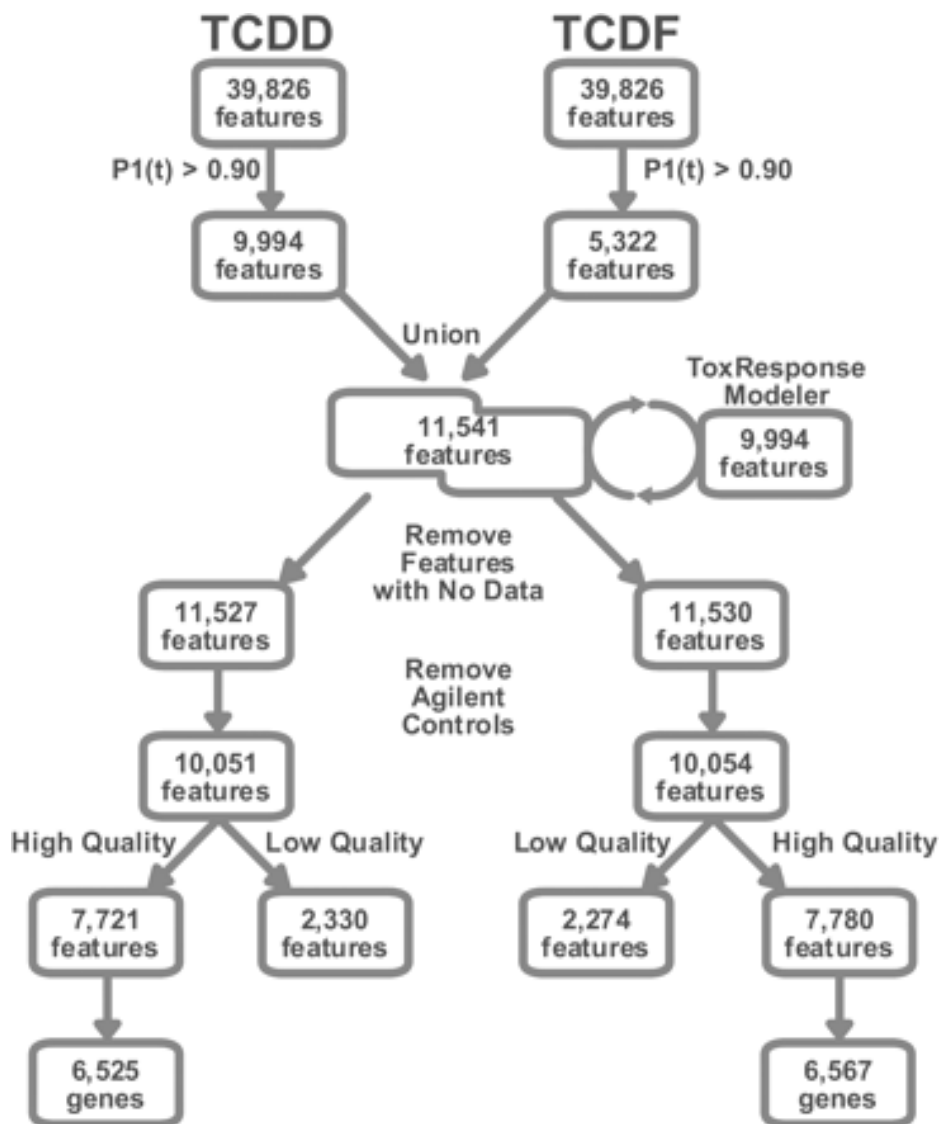
## Next Steps

- Community analysis to identify agglomerative biomarkers (ABs)
- Disease: Analyze additional datasets to create probabilistic ABs
- Mixture: Analyze additional datasets to create probabilistic ABs
- Single Chemical (BaP): Analyze additional datasets to create probabilistic ABs

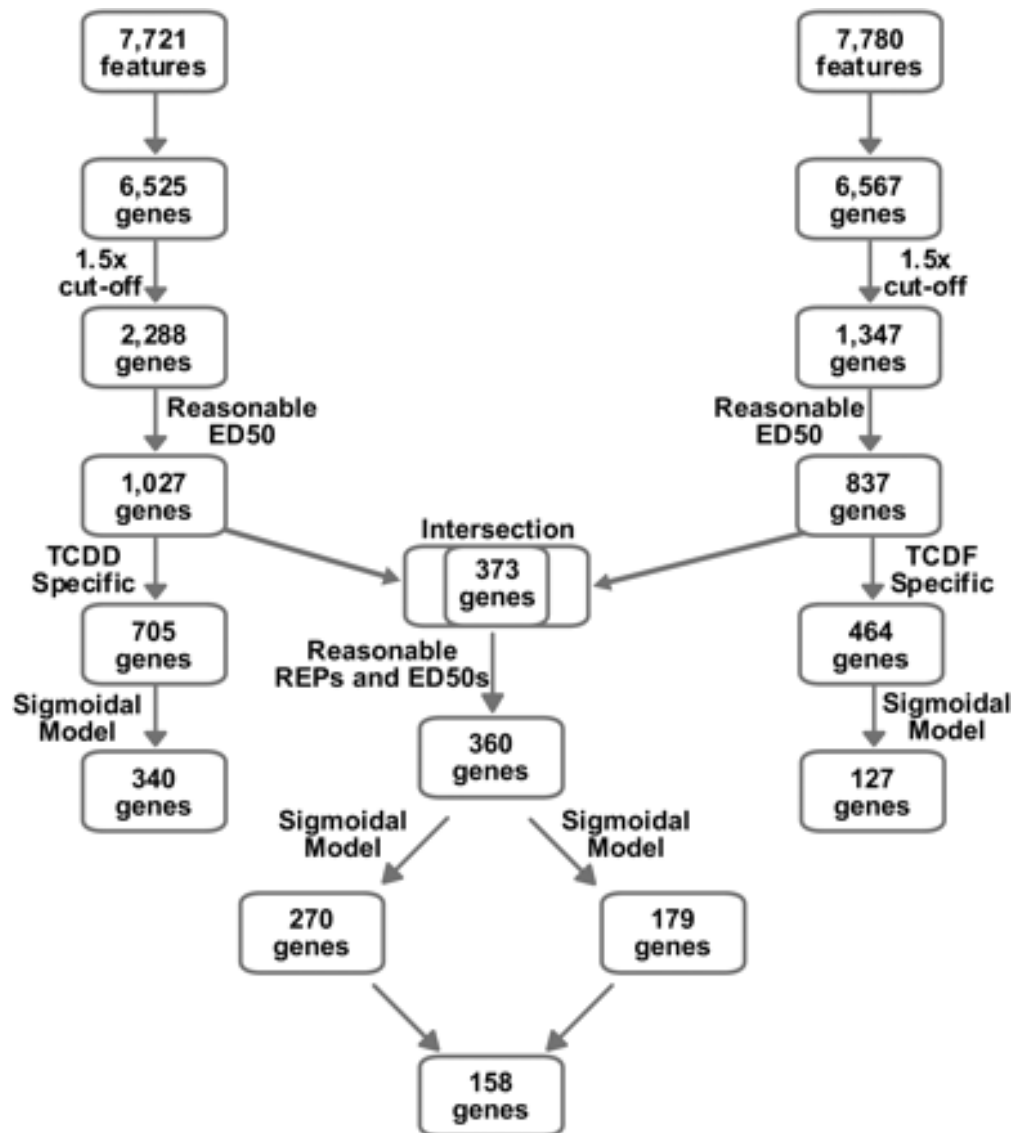
## Translating This to Risk Assessment

- Hazard identification
  - With additional datasets, and probabilistic ABs, we can quantitate certainty around hazard/end-point predictions
- Dose-response
  - Need additional dose-response datasets in this case (not many in the published literature)
  - I developed a toxicogenomic method for calculating relative potency in the past (Burgoon and Zacharewski, ToxSci 2009)
    - Working on improving the current method using ABs
- Improve mode of action understanding

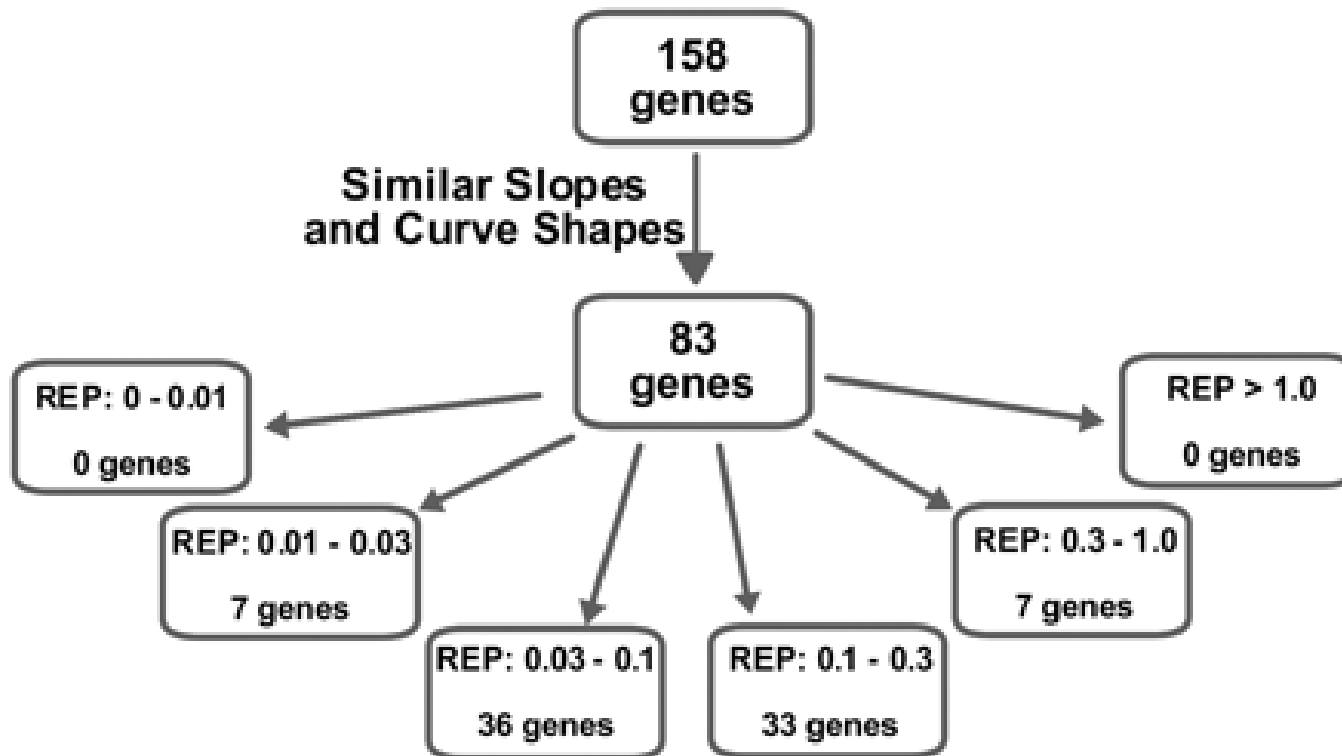
# Toxicogenomic Relative Potency Calculation



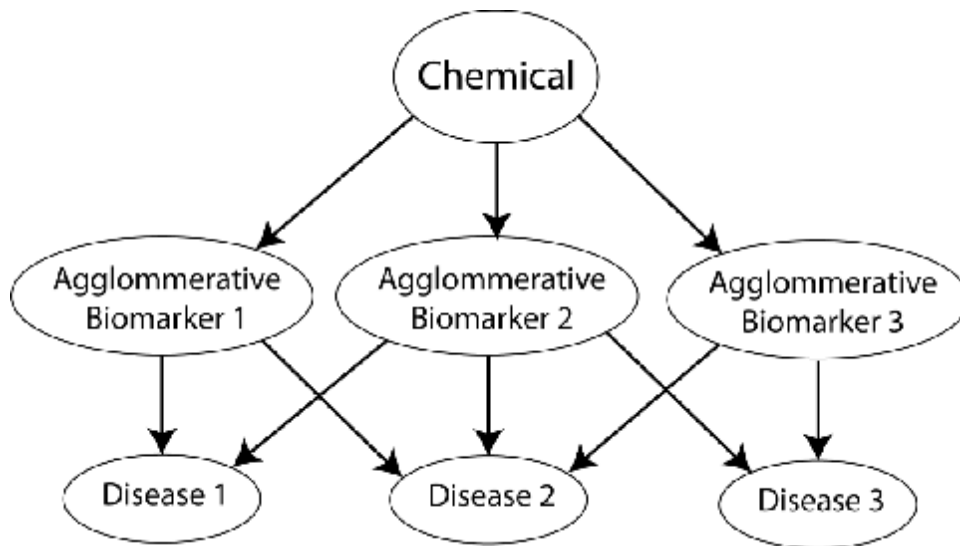
# Toxicogenomic Relative Potency Calculation



## Toxicogenomic Relative Potency Calculation



## Translating this to Risk Assessment (cont'd)



- Model is flexible
- Can incorporate probabilistic ABs for non-chemical stressors
- Model can also incorporate dose-response

## Future Work (and needs)

- Additional omics data about other diseases to see how specific predictions from this data will be for lung cancer
- *In vitro* datasets to see how well they predict *in vivo* ABs, and if they can be used to specifically predict lung cancer
- Need: more single chemical data
  - Dose-responses at relevant doses
- Improving community analysis methods (identifying ABs)
- Improving the computational speed and throughput
  - Development of automated analysis workflows

## Summary

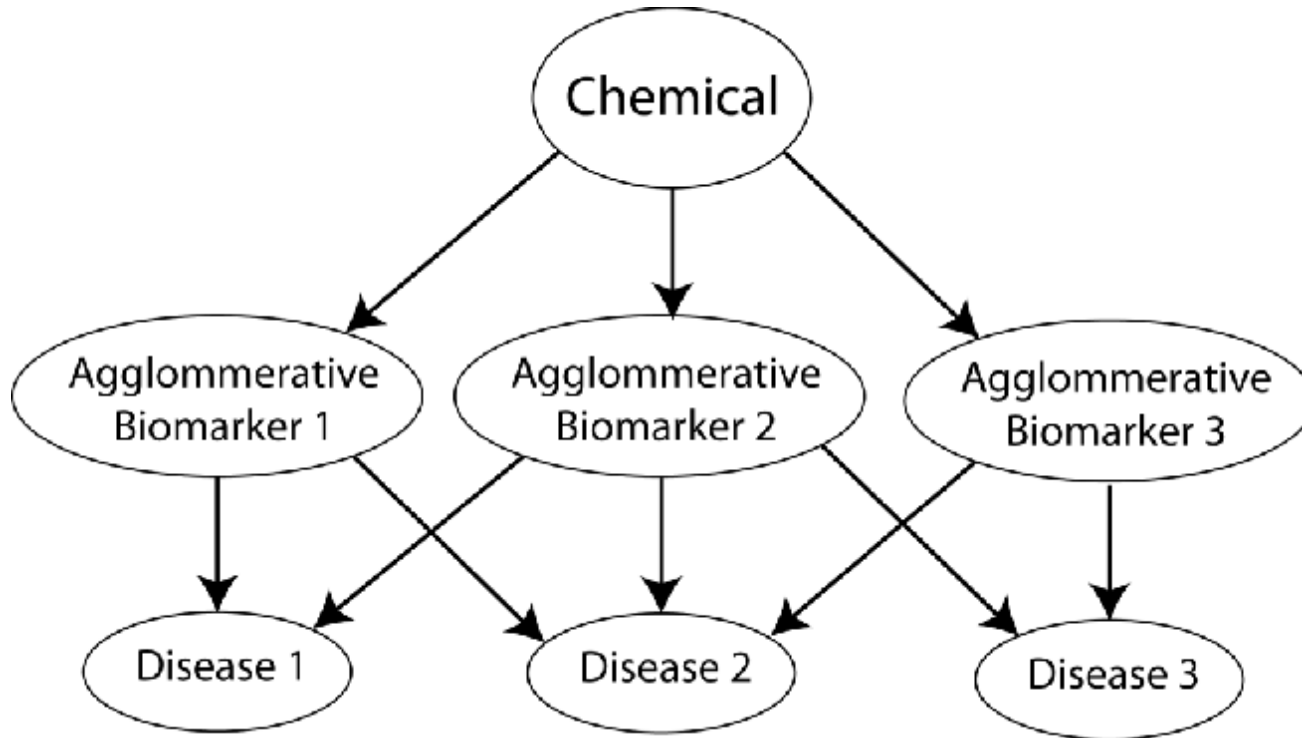
- We have one example of toxicogenomics being able to predict a health outcome
- Additional work is needed to make these predictions probabilistic
- We have built a prototype computational analysis system for combining data from all relevant stressors (chemical and non-chemical)
- We need more data to evaluate the system and our predictions



## Acknowledgements (only listing people who helped on this pilot)

- NexGen Team
  - Ila Cote
  
- NexGen PAH Prototype Team
  - Peter McClure
  - Heather Carlson-Lynch
  - Julie Stickney
  
- Health Canada
  - Carole Yauk
  - Ivy Moffat
  
- US EPA
  - Stephen Edwards
  
- Michigan State University
  - Shannon Bell

# Bayesian Networks for Probabilistic Chemical-Disease Prediction



Non-chemical stressors, too!