Weighted Gene Co-expression Network Analysis in COPD patients: Finding a Molecular Signature for Smoking

Tova Fuller, MD, PhD
Columbia University Medical Center

August 8, 2012

A Study Purpose and Rationale

The effects of cigarette exposure on gene expression have been studied, showing that anywhere from tens to hundreds of genes are significantly altered by smoking. PBMCs stimulated with cigarette smoke extract were found to have 80 genes upregulated and 37 genes downregulated by 1.5-fold or greater after 8 hours; these genes were largely related to cell survival, including antioxidants, chaperones, folding proteins, and ubiquitin/proteosome pathway genes [1]. Charlesworth, et al. studied lymphocytes' response to smoking in 297 individuals in the San Antonio Family Heart study, identifying 323 unique genes and 342 transcripts that were significantly correlated with smoking behavior at a FDR of 5% [2]. Other studies have been performed on buccal mucosa [3], B cells in women [4], small airway epithelium, and alveolar macrophages in healthy individuals [5].

Gene coexpression network analysis, a means for looking at the relationship between different gene transcripts, has also been used to study smoking, but mostly in lung cancer as opposed to other types of pulmonary pathology. Recently, these networks have been constructed in smoking patients with lung adenocarcinoma [6], finding a 7-gene signature for diagnosis and prognosis. Likewise, after defining candidate genes related to lung cancer survival, a coexpression network based on prediction logic for smoker group and non-smoker group was used to find a 6-gene signature for prognosis in lung cancer that had an accurate diagnosis with an accuracy of 73% [7]. These studies postulate that understanding the molecular signature may help with patient selection for adjuvant chemotherapy, thus providing an option for intervention.

However, relatively less human expression studies have been published regarding COPD, another lung-related injury. Ezzie, et al., compared subjects with COPD and smokers without COPD, finding differential expression of 70 miRNAs and 2667 mRNAs [8]. These results solely represent differential expression. Gene network studies are even more rare. Acquaah-Mensah, et al. studied human airway cells restricting to posited ontological categories: apoptosis, response to oxidative stress, and inflammatory response; the genes involved in these categories were used to generate a transcriptional regulatory network. These genes and previously studied differential expressed genes in COPD were used to find COPD susceptibility loci. Not surprisingly, they found a number of senescence-related genes were related to COPD, considering they restrict their analysis to sensecence-related concepts [9]. Gene networking has been used to study COPD in muscle, with small sample sizes of 12 and 18 healthy and COPD subjects [10]. However, no well-powered studies to our knowledge have used data from solely COPD patients to construct a gene network.

In our study, we aim to use weighted gene coexpression network analysis to study smoking in COPD patients. By using PBMCs, we aim to identify a network based signature for smoking in those known to be critically affected. The rationale behind using PBMCs is that these cells can be easily harvested from future patients. Using network analysis provides two benefits: 1) numerically, this serves as a means for decreasing the inherent problem of multiple comparisons, as it is a data reduction strategy. 2) Using a systems biology perspective in identifying groups of coexpressed genes ideally would, in turn, identify pathways of genes that interact with one another, thereby being functionally linked.

B Study Design and Statistical Analysis

MESA, or the Multi-Ethnic Study of Atherosclerosis, is a community-based study that enrolled 6,814 asymptomatic men and women from 45 to 84 years of age across multiple field centers in 2000-2002; current participating centers are the University of Washington (Coordinating center), Columbia University, Johns Hopkins University, Northwestern University, University of Minnesota, University of California at Los Angeles, Wake Forest University, University of Vermont, New England Medical Center, the National Heart, Lung, and Blood Institute, and University of Virginia. As the name suggests, the study aims to find risk factors for subclinical or overt cardiovascular disease.

The MESA lung study, an ancillary study to MESA, is a four year epidemiological study to understand the endothelial hypotheses of COPD. It has recruited 3,959 participants, all of whom are MESA participants, with the requirements that they had consented to genetic analyses. Note that minority race/ethnic groups are oversampled with 35% Caucasian, 24% African American 23% Latino, and 18% Chinese-Americans. 97 of these individuals will be chosen who have COPD based on spirometry.

Weighted gene coexpression network analysis has been used to study multiple phenotypes [11–14]. First pairwise correlations will be identified in the gene expression microarray data, from which a topological overlap matrix can be defined based on these pairwise correlations and the connectivity, or row sum of the correlation matrix. Specifically,

$$a_{ij} = |\frac{1 + cor(x_i, x_j)}{2}|^{\beta}. \tag{1}$$

represents the adjacency (connection strength) between genes i an and j in a signed network. An important network concept is the connectivity measure k_i (also known as degree), which measures how connected the i-th gene is with other genes in the network. The (whole network) connectivity is defined as the sum of connection strengths (adjacencies) with the other network genes:

$$k_i = \sum_{u \neq i} a_{ui}. \tag{2}$$

A major step in our module centric analysis is to cluster genes into network modules using a network dissimilarity measure. Roughly speaking, a pair of genes has a small dissimilarity if it is closely interconnected.

To identify modules, we define a network dissimilarity given by

$$dissTOM_{ij} = 1 - \frac{\sum_{u \neq i} a_{iu} a_{uj} + a_{ij}}{min(k_i, k_j) + 1 - a_{ij}} \tag{3}$$

with k_i defined above. This measure combines the adjacency of two genes and the connection strengths these two genes share with other 'third party' genes.

We define the module eigengene E as the first principal component of the standardized expression profiles of a given module, which may be used to summarize a module's expressions. It can be considered a weighted average gene expression profile or the representative of the gene expression profiles in a module. When a gene expression sample trait y is available–here pack-years–one can correlate the module eigengenes with this outcome. The correlation coefficient or corresponding p-value is referred to as eigengene significance.

Intramodular connectivity or $k_{IM,i}$ measures how connected, or co-expressed, the i-th gene is with respect to the genes of a particular module. The intramodular connectivity may be interpreted as a measure of module membership. Intramodular connectivity is calculated as the sum of the adjacencies within the module of interest. The loosely defined term 'hub gene' is used as an abbreviation of 'highly connected gene.' By definition, genes inside coexpression modules tend to have high connectivity.

In our analysis, we will relate all module eigengenes (n=20 or so) and hub genes to pack-years. Linear regression models will be tested for prediction of pack-years based on the significantly correlated modules and hubs based an alpha of 0.001, thus using the Bonferroni correction to correct for 50 tests at the significance level of 0.05.

Power analyses predict that, with 47 individuals in each group, and a standard deviation of 2.2, which preliminary analyses support, we would be able to show an effect size of 1.9 with a power of 0.8 and an alpha of 0.001. This alpha is calculated using the Bonferroni correction for 50 or less module and hub representations, combined. In other words, with the current enrollment numbers for this study, we are powered to find a difference of 0.86 standard deviations.

C Study Drugs

No drugs will be used.

D Medical Device

No medical devices will be used.

E Study Questionnaires

The patients will fill out questionnaires regarding their age, gender, smoking status, and number of pack-years (number of years smoking times packs smoked per day).

F Study Subjects

The patient population is racially mixed, with 38% white, 28% African-American, 22% Latino, and 12% Asian. Patients were considered eligible if they were between the ages of 45 and 84, were African-American, Chinese-American, Caucasian, or Hispanic, and if they did not meet any of the exclusion criteria.

Exclusion criteria, based on the initial MESA IRB protocol, were:

- Age younger than 45 or older than 84 years
- Physician-diagnosed heart attack
- Physician-diagnosed angina or taking nitroglycerin
- Physician-diagnosed stroke or TIA
- Physician-diagnosed heart failure
- Current atrial fibrillation
- Having undergone procedures related to cardiovascular disease (CABG, angioplasty, valve replacement, pacemaker or defibrillator implantation, any surgery on the heart or arteries)
- Active treatment for cancer
- Pregnancy
- Any serious medical condition which would prevent long-term participation
- Weight >300 pounds
- Cognitive inability as judged by the interviewer
- Living in a nursing home or on the waiting list for a nursing home
- Plans to leave the community within five years
- Language barrier (speaks other than English, Spanish, Cantonese or Mandarin)
- Chest CT scan in the past year

G Recruitment of Subjects

Sites have recruited equal numbers of participants with the race proportions as given above. Wake Forest, Johns Hopkins, Minnesota, and Northwestern worked to create community awareness of the study and enlisted the support and endorsement of community-based organizations and leadership. Columbia worked with the 1199 National Benefit Fund during recruitment, and UCLA used random-digit dialing. All sites employed staff fluent in Spanish, Cantonese, and Mandarin when applicable.

H Confidentiality of Data

Patient confidentiality will be strictly maintained, and information will not be shared with any insurance company or employer. All patient samples and records will be maintained in a locked and secure location. Anonymous identifier keys will be assigned to samples in both expression and phenotype data.

I Potential Conflict of Interest

There are no conflicts of interest in this analysis.

J Location of the Study

The study, as mentioned previously, is currently taking place at the University of Washington (Coordinating center), Columbia University, Johns Hopkins University, Northwestern University, University of Minnesota, University of California at Los Angeles, Wake Forest University, University of Vermont, New England Medical Center, the National Heart, Lung, and Blood Institute, and University of Virginia.

K Potential Risks

The risks involved with this study are entirely related to the phlebotomy required for PBMC retrieval: infection and bleeding at the puncture site.

L Potential Benefits

There are no anticipated benefits to participants.

M Alternative Therapies

There will be no alternative therapies.

N Compensation to Subjects

There is no compensation to subjects.

O Costs to Subjects

There are no costs to the subjects.

P Minors as Research Subjects

No minors will be studied.

Q Radiation or Radioactive Substances

There will be no radiation or radioactive substances used in this study.

References

[1] William R Wright, Katarzyna Parzych, Damian Crawford, Charles Mein, Jane A Mitchell, and Mark J Paul-Clark. Inflammatory transcriptome profiling of human monocytes exposed acutely to cigarette smoke. *PLoS One*, 7(2):e30120, 2012.

[2] Jac C Charlesworth, Joanne E Curran, Matthew P Johnson, Harald Hh Göring, Thomas D Dyer, Vincent P Diego, Jack W Kent, Jr, Michael C Mahaney, Laura Almasy, Jean W MacCluer, Eric K Moses, and John Blangero. Transcriptomic epidemiology of smoking: the effect of smoking on gene expression in lymphocytes. *BMC Med Genomics*, 3:29, 2010.

[3] Doris M Kupfer, Vicky L White, Marita C Jenkins, and Dennis Burian. Examining smoking-induced differential gene expression changes in buccal mucosa. *BMC Med Genomics*, 3:24, 2010.

[4] Feng Pan, Tie-Lin Yang, Xiang-Ding Chen, Yuan Chen, Ge Gao, Yao-Zhong Liu, Yu-Fang Pei, Bao-Yong Sha, Yan Jiang, Chao Xu, Robert R Recker, and Hong-Wen Deng. Impact of female cigarette smoking on circulating b cells in vivo: the suppressed icoslg, tcf3, and vcam1 gene functional network may inhibit normal cell function. *Immunogenetics*, 62(4):237–51, Apr 2010.

[5] Brendan J Carolan, Ben-Gary Harvey, Neil R Hackett, Timothy P O'Connor, Patricia A Cassano, and Ronald G Crystal. Disparate oxidant gene expression of airway epithelium compared to alveolar macrophages in smokers. *Respir Res*, 10:111, 2009.

[6] Ying-Wooi Wan, Rebecca A Raese, James E Fortney, Changchang Xiao, Dajie Luo, John Cavendish, Laura F Gibson, Vincent Castranova, Yong Qian, and Nancy Lan Guo. A smoking-associated 7-gene signature for lung cancer diagnosis and prognosis. *Int J Oncol*, Jul 2012.

[7] Nancy Lan Guo and Ying-Wooi Wan. Pathway-based identification of a smoking associated 6-gene signature predictive of lung cancer risk and survival. *Artif Intell Med*, 55(2):97–105, Jun 2012.

[8] Michael E Ezzie, Melissa Crawford, Ji-Hoon Cho, Robert Orellana, Shile Zhang, Richard Gelinas, Kara Batte, Lianbo Yu, Gerard Nuovo, David Galas, Philip Diaz, Kai Wang, and S Patrick Nana-Sinkam. Gene expression networks in copd: microrna and mrna regulation. *Thorax*, 67(2):122–31, Feb 2012.

[9] George K Acquaah-Mensah, Deepti Malhotra, Madhulika Vulimiri, Jason E McDermott, and Shyam Biswal. Suppressed expression of t-box transcription factors is involved in senescence in chronic obstructive pulmonary disease. *PLoS Comput Biol*, 8(7):e1002597, Jul 2012.

[10] Nil Turan, Susana Kalko, Anna Stincone, Kim Clarke, Ayesha Sabah, Katherine Howlett, S John Curnow, Diego A Rodriguez, Marta Cascante, Laura O'Neill, Stuart Egginton, Josep Roca, and Francesco Falciani. A systems biology approach identifies molecular networks defining skeletal muscle abnormalities in chronic obstructive pulmonary disease. *PLoS Comput Biol*, 7(9):e1002129, Sep 2011.

[11] Tova F Fuller, Anatole Ghazalpour, Jason E Aten, Thomas A Drake, Aldons J Lusis, and Steve Horvath. Weighted gene coexpression network analysis strategies applied to mouse weight. *Mamm Genome*, 18(6-7):463–72, Jul 2007.

[12] Christiaan G J Saris, Steve Horvath, Paul W J van Vught, Michael A van Es, Hylke M Blauw, Tova F Fuller, Peter Langfelder, Joseph DeYoung, John H J Wokke, Jan H Veldink, Leonard H van den Berg, and Roel A Ophoff. Weighted gene co-expression network analysis of the peripheral blood from amyotrophic lateral sclerosis patients. *BMC Genomics*, 10:405, 2009.

[13] Simone de Jong, Tova F Fuller, Esther Janson, Eric Strengman, Steve Horvath, Martien J H Kas, and Roel A Ophoff. Gene expression profiling in c57bl/6j and a/j mouse inbred strains reveals gene networks specific for brain regions independent of genetic background. *BMC Genomics*, 11:20, 2010.

[14] Wei Zhao, Peter Langfelder, Tova Fuller, Jun Dong, Ai Li, and Steve Hovarth. Weighted gene coexpression network analysis: state of the art. *J Biopharm Stat*, 20(2):281–300, Mar 2010.