# SYSTEMS BIOLOGY ANALYSES OF GENE EXPRESSION AND GENOME WIDE ASSOCIATION STUDY DATA IN OBSTRUCTIVE SLEEP APNEA

YU LIU

*Center for Proteomics &*
*Bioinformatics, Case Western*
*Reserve University (CWRU),*
*Cleveland, Ohio, 44106, USA,*
*Email: yxl442@case.edu*

SANJAY PATEL

*Division of Pulmonary, Critical*
*Care and Sleep Medicine, CWRU,*
*Cleveland, Ohio, 44106, USA,*
*Email: srp20@case.edu*

ROD NIBBE
SEAN MAXWELL

*Center for Proteomics &*
*Bioinformatics, CWRU, Cleveland,*
*Ohio, 44106, USA, Email:*
*rkn6@case.edu; stm@case.edu*

SALIM A. CHOWDHURY

*Department of Electrical*
*Engineering & Computer Science,*
*CWRU, Cleveland, Ohio, 44106,*
*USA, Email: sxc426@case.edu*

MEHMET KOYUTURK

*Department of Electrical*
*Engineering & Computer Science,*
*CWRU, Cleveland, Ohio, 44106,*
*USA, Email: mxk331@case.edu*

XIAOFENG ZHU

*Department of Epidemiology and*
*Biostatistics, CWRU, Cleveland,*
*Ohio, 44106, USA,*
*Email:xiaofeng.zhu@case.edu*

EMMA K. LARKIN

*Division of Allergy, Pulmonary and*
*Critical Care, Vanderbilt University*
*Medical Center, 1215 21st Ave S.,*
*Nashville, Tennessee, 37232, USA,*
*Email: emma.larkin@vanderbilt.edu*

SARAH G BUXBAUM

*Jackson Heart Study, Jackson State*
*University, Jackson, MS 39213, USA*
*Email: sarah.g.buxbaum@jsums.edu*

NARESH M. PUNJABI

*Department of Epidemiology, Johns*
*Hopkins Bloomberg School of*
*Public Health, Baltimore,*
*MD 21205,USA,*
*Email: npunjabi@jhmi.edu*

SINA A. GHARIB

*Center for Lung Biology, Division of*
*Pulmonary and Critical Care*
*Medicine, University of Washington,*
*Seattle, WA 98109, USA. Email:*
*sagharib@u.washington.edu*

SUSAN REDLINE

*Department of Medicine, CWRU,*
*Cleveland, Ohio, 44106, and Depart*
*of Medicine, Brigham & Women's*
*Hospital and Beth Israel Deaconess*
*Medical School, Harvard Medical*
*School, Boston, MA, 02115 Email:*
*sredline@partners.org*

MARK R. CHANCE

*Center for Proteomics &*
*Bioinformatics, Department of*
*Genetics, Case Western Reserve*
*University , Cleveland, Ohio, 44106,*
*USA, Email: mark.chance@case.edu*

The precise molecular etiology of obstructive sleep apnea (OSA) is unknown; however recent research indicates that several interconnected aberrant pathways and molecular abnormalities are contributors to OSA. Identifying the genes and pathways associated with OSA can help to expand our understanding of the risk factors for the disease as well as provide new avenues for potential treatment. Towards these goals, we have integrated relevant high dimensional data from various sources, such as genome-wide expression data (microarray), protein-protein interaction (PPI) data and results from genome-wide association studies (GWAS) in order to define sub-network elements that connect some of the known pathways related to the disease as well as define novel regulatory modules related to OSA. Two distinct approaches are applied to identify sub-networks significantly associated with OSA. In the first case we used a biased approach based on sixty genes/proteins with known associations with sleep disorders and/or metabolic disease to seed a search using commercial software to discover networks associated with disease followed by information theoretic (mutual information) scoring of the sub-networks. In the second case we used an unbiased approach and generated an interactome constructed from publicly available gene expression profiles and PPI databases, followed by scoring of the network with p-values from GWAS data derived from OSA patients to uncover sub-networks significant for the disease phenotype. A comparison of the approaches reveals a number of proteins that have been previously known to be associated with OSA or sleep. In addition, our results indicate a novel association of Phosphoinositide 3-kinase, the STAT family of proteins and its related pathways with OSA.

## 1. Introduction

Although its precise functions are not entirely known, sleep is important for numerous physiological and cognitive functions. Sleep disorders can have a range of consequences, from minor to severe, such as untimely drowsiness, motor vehicle collisions, and workplace accidents as well as increase risk of hypertension, diabetes and mortality. Of the more than 70 known sleep disorders, obstructive sleep apnea (OSA) is one of the most common[1,2]. OSA is a complex disorder caused by a repetitive collapse of the upper airway during sleep, disrupting breathing and sleep. Repetitive episodes of obstruction cause intermittent drops in blood oxygen and increases in carbon dioxide levels, which can lead to frequent arousals from sleep. OSA is a major cause of chronic sleep deprivation and excessive daytime sleepiness. It is estimated that up to 5% of adults in Western countries are likely to have OSA syndrome[3]. Treatments for OSA include behavioral therapies (such as changing sleeping positions), use of mechanical devices, and surgery to increase the patency of the airway. However, after decades of research the molecular mechanisms underlying OSA remain unclear.

OSA is unlikely to be a simple condition associated with a few genes or proteins; instead, it is likely a manifestation of multiple interconnected aberrant pathways and numerous molecular abnormalities[4]. In addition, it is a risk factor for many other diseases and many other diseases increase the risk of OSA. For example, OSA is associated with inflammatory states[5-8] and oxidative stress[9,10]. While obesity is one of the strongest risk factors for OSA[11], other co-morbidities include insulin resistance, hypertension, and cardiovascular disease[12-14].

Multiple studies indicate an important genetic basis for OSA, and genetic factors alone can explain approximately 30-40% of the variance of the apnea hypopnea index (AHI), a quantitative measure of OSA, defined by the number of apneas and hypopneas per hour of sleep[15,16]. OSA is also mediated by environmental factors, most obviously through those that link it to related traits such as obesity[17], but which may include influences associated with irritant exposures, alcohol use and sleep deprivation. Efforts to identify genetic variants related to OSA, include family as well as genome-wide, case-control studies and are an important attempt to provide diagnostic and/or prognostic information related to the disease. In linkage analysis of families with an affected OSA member, Larkin et al. have identified several chromosomal regions linked to the AHI[18]. Some, but not all, of the genetic pathways were believed to be obesity dependent[18]. In another study, based on a pre-selected gene set and SNP data, the same group found five genes significantly associated with OSA and the AHI[19]. Many additional genes in these and related genome-scale studies are likely relevant to mediating disease, but due to the multiple hypotheses testing problem when thousands of genes are analyzed, only a few genes with very significant p-values are allowed to pass the relevant filters for significance. The problem of identifying as many biologically significant genes as possible in such an analysis remains very important.

Network modeling of protein-protein interactions provides a relatively new context to study disease and identify disease-related genes. The effectiveness of network-based approaches to the identification of multiple disease markers has been demonstrated in the context of various diseases, such as colon cancer[20]. The aim of this study is to uncover protein-protein sub-networks associated with OSA by integrating data from multiple high-dimensional studies, both to demonstrate the power of systems biology data integration in developing novel mediators of OSA and to use the novel data available in the field to explore and validate new computational approaches. To achieve these goals, we applied two approaches, 1: candidate gene approach integrated with adipose tissue microarray data; 2: genome-wide approach integrated with adipose microarray data. The first approach is based on the method proposed by Nibbe et al.[20]: we use 56 seed proteins to drive a search of a protein-protein interaction (PPI) to discover rank-ordered sub-networks associated with OSA. In this method, proteins known or suspected to be associated with OSA and related co-morbidities are used to seed a search of a well-annotated, human PPI for candidate sub-networks, which are subsequently scored with gene expression data to derive candidate sub-networks underlying the disease phenotype. We demonstrate the utility of this approach, using a biased seed set, to provide interesting candidate sub-networks for further exploration in the etiology of OSA.

In a second unbiased approach, we mapped p-values obtained from a case-control GWAS study based on OSA phenotypes to nodes of an adipose tissue-specific interactome constructed from gene expression data[21-23]. Subsequently, we used Cytoscape based tools to identify sub-networks significantly associated with OSA. A novel feature of the study is that nodes that were highly significant along with nodes that were not the most significant in the GWAS analysis both provided important contributions to discovering the sub-networks that are of potential biological significance for the phenotype. This indicates an approach for extending the value of GWAS data to other complex phenotypes. By incorporating data from both approaches, sub-networks were identified that included targets known to be associated with OSA or sleep in general and also indicated that PI3K, STAT family, and related pathways may have important functional roles in OSA.

## 2. Material and Methods

### 2.1 *Network construction*

Two methods to construct PPI networks were used in this study. First, 56 seed genes/proteins were selected based on knowledge of the underlying biology and prior genetics studies of OSA[19]. The list of genes is provided as supplementary material (Supplementary Table 1), and can be reached at website (http://proteomics.case.edu/news_events.aspx?newsid=38). Most of the genes are known to be in one or more pathways representing intermediate phenotypes for OSA: craniofacial morphology, obesity, inflammation, and ventilatory control pathways, or across multiple pathways, through biologic pleiotropy. A traditional association study has been conducted on this set of proteins and has been recently published[19]. Ingenuity Pathway Analysis (IPA) software (Ingenuity® Systems, www.ingenuity.com) was used to construct networks by the following steps[24]:a) seed proteins are combined into networks that maximize their specific connectivity, which is their interconnectedness with each other relative to all molecules they are connected to in the Ingenuity Knowledge Base; b) additional proteins from the Ingenuity Knowledge Base are added to specifically connect two or more smaller networks by merging them into a larger one. The networks were limited to 70 nodes each to permit ease of computational scoring of sub-networks using mutual information (see below). The overall network score is based on the number of seed proteins they contain. The two top scoring networks were used in the analysis (See Results and Discussion). Note that IPA will cluster a protein complex or protein family into a single node if a number of components or family members are present in the network. For scoring purposes (see below), the expression value of the family or complex is represented by the maximum expression value among its components.

Second, an interactome specific to adipose tissue, which has been previously constructed by combining gene expression data from adipose tissues and PPI information from public databases and published papers[22,23,25,26], was optimized following curation with recent next-generation sequencing data[23]. Briefly, mRNA expression levels from Su et al[25,26] and Wang et al[23], which are used to determine the significance of a gene to the network, were estimated by combining results from microarray (chip based) experiments along with next generation sequencing results from selected references[23,25,26]. Protein nodes with mRNA levels below a defined threshold were considered as absent (the threshold for data from next generation sequencing is 20 reads; in the case of data from microarray experiments, the threshold for normalized expression level is 200[25,26]). Interactions between two proteins supported by at least three databases and two experiments were added to the interactome[22]. The adipose specific interactome in SIF format is provided as supplementary data (Supplementary Table 2, http://proteomics.case.edu/ news_events.aspx?newsid=38). Network-Analyzer is used to compute the network properties, such as the average shortest path length and the node degree distribution[27].

### 2.2 *Gene expression data processing*

Experimentally derived mRNA expression data for subcutaneous and visceral fat tissues were measured by cDNA microarray using the Affymetrix Human Gene 1.0 ST Array on intra-operative samples from 10 OSA patients and 8 controls undergoing elective ventral hernia repair surgery. Adipose tissue was chosen for expression studies since it is accessible and because of the central role of obesity in the pathogenesis of OSA. The information about these samples, such as sample IDs, AHIs, are provided as supplementary data (Supplementary Table 3, http://proteomics.case.edu/news_events.aspx?newsid=38). Expression values were generated using the aroma package from bioconductor[28]. Robust multichip average (RMA) and quantile normalization methods were used for background correction and normalization. In an initial analysis, two subcutaneous and three visceral samples (i.e., five out of 36 samples) had much larger variances than other samples (GEGF ID 14, 15, 16, 21, 22, all of them are control samples, see Supplementary Table 3, http://proteomics.case.edu/news_events. aspx?newsid=38), these were treated as outliers, and removed.

### 2.3 *Subnetwork scoring and detecting using mutual information (MI)*

Once a network enriched in seed proteins is constructed, we identify dysregulated sub-networks within this network using mRNA expression data. The aim of this procedure is to find sets of genes that exhibit coordinate differential expression, in that they can discriminate case and control samples when their expression profiles are considered together. For this purpose, we use an information-theoretic measure of coordinate dysregulation that was developed by Chuang et al.[29] and was previously used to detect dysregulated subnetworks in breast cancer metastasis[29] and late stage colorectal cancer[20]. This measure of sub-network dysregulation is powerful in that it provides a multivariate assessment of the coordination between multiple genes in their differential expression.

Namely, for a given set of proteins $S=\{g_1, g_2, ..., g_k\}$, let $e_i$ denote the mRNA expression level of $g_i \in S$. Then the *subnetwork activity* of $S$ is defined as $e_S = \sum_{i=1}^{k} e_i/\sqrt{k}$, that is the aggregate mRNA-level expression of the proteins in the sub-network. Subsequently, mutual information is used to measure the dependence of two discrete random variables: in this case the health status *vs.* subnetwork activity of *S*. Denoting health status vector as $c$ (i.e., $c(j)$ denotes the health status of the $j^{th}$ sample) and quantized subnetwork activity of $S$ as $\hat{e}_S$, (i.e., $\hat{e}_S(j)$ denotes the aggregate expression of the gene products in $S$ in the $j^{th}$ sample), the dysregulation of $S$ is defined as $I(c,\hat{e}_S)=H(c)-H(c|\hat{e}_S)$. Here, $H(c)$ denotes the Shannon entropy of random variable $c$ (that is the uncertainty on the health status of a sample) and $H(c|\hat{e}_S)$ denotes the entropy of random variable $c$ after the observation of random variable $\hat{e}_S$ (that is the uncertainty on the health status of a sample given the subnetwork activity of $S$ in that sample). Consequently, the mutual information (MI) $I(c,\hat{e}_S)$ is a measure of the expression levels of all genes in the subnetwork in discriminating OSA patients from control. To this end, a high MI score for a sub-network is an indicator of the coordinate mRNA-level dysregulation of the proteins in the subnetwork, i.e., although the gene coding for each protein in the sub-network may not be significantly differential expressed in OSA, the total mRNA-level expression of these proteins exhibits significant difference between OSA patients and control. This information theoretic formulation of coordinate dysregulation has been shown to be effective in identification of subnetwork markers that were powerful in prediction of breast and colon cancer metastasis [29, 51].

While Chuang et al. originally used a greedy algorithm to identify subnetworks with high MI [29], we exhaustively searched for subnetworks of the IPA network to identify sets of genes with high MI. This is because the network obtained from IPA analysis is already filtered to obtain a concise network of proteins that are functionally associated with proteins that are already known to play a role in sleep apnea. Consequently, an exhaustive search for reasonably sized subnetworks (we search for subnetworks composed of up to 6 proteins in this study) is feasible on this network, which is guaranteed to find all subnetworks with a maximum MI, as opposed to a greedy algorithm[20].

### 2.4 *Analyzing adipocyte interactome using SNP association scores from GWAS*

The Candidate Gene Association Resource (CARe) project initiated by the National Heart, Lung, and Blood Institute, conducted analyses of genetic variation in cardiovascular, pulmonary, hematological, and sleep-related traits in nine community-based cohorts[21]. Polysomnography data, providing objective measurements of OSA, were only available for a subset of these cohorts, and of these, a genome-wide assay (Affymetrix 6.0) was only performed in the African American participants (n=647) in the Cleveland Family Study, which provided p-values for the associations between 867,496 SNPs with OSA (defined as an AHI > 15 for identifying cases).

We then map these p-values to proteins/nodes in the adipose tissue-specific interactome map as follows. For each protein $g_i$ in the network, the most significant p-value that is associated with a SNP located in the coding region of $g_i$ is designated as the p-value of the association of $g_i$ with OSA. In other words, letting *p(s)* denote the p-value of the association of SNP $s$ with OSA, we define $p_i = \min_{s \in R_i} p(s)$ where $R_i$ denotes the set of SNPs that reside within the coding region of $g_{i.}$[30,31]. Subsequently, we apply a Cytoscape tool, jactivemodule, to extract sub-networks of the adipose tissue-specific interactome map that are enriched in proteins with high total significance of association with OSA[31].

jactivemodule is a subnetwork search algorithm that was originally developed  to identify *active subnetworks* in a network of interactions, where an active subnetwork refers to a connected subgraph of the interactome that has high total significance of differential mRNA-level expression with respect to a particular perturbation[32]. It takes as input p-values associated with each protein in the network, converts these p-values to $z$-scores (so that a higher $z$-score indicates more significant differential expression), and greedily identifies subnetworks with high aggregate *z-score.* More precisely, the score of a subnetwork $S$={$g_1$, $g_{2,}$ …, $g_k$} is defined as $A(S) = \sum_{i=1}^{k} z_i/\sqrt{k},$ where $z_i$ denotes the z-score corresponding to p-value $p_{i.}$

Although this method was originally developed to identify differentially expressed subnetworks, it can as well be used to identify disease-associated subnetworks since the p-values of differential expression can be replaced by p-values of association with the disease. Motivated by this insight, we use this algorithm to identify subnetworks that are implicated in OSA by GWAS. Observe also that, a high-scoring sub-network is not necessarily one that is enriched in proteins with very significant p-values, but it can also be comprised of many proteins with moderately significant p-values. Consequently, this method has the potential of uncovering groups of proteins that exhibit seemingly insignificant association with OSA when considered individually, but exhibit strong association when considered together. Since such subnetworks are connected by a network of interactions by the construction of the algorithm, they are likely to be functionally associated and therefore might be underlying a potential genetic interaction that underlies the manifestation of  the disease. The details of procedure can be found at the documentation of cytoscape (www.cytoscape.org) and in the literature[33].

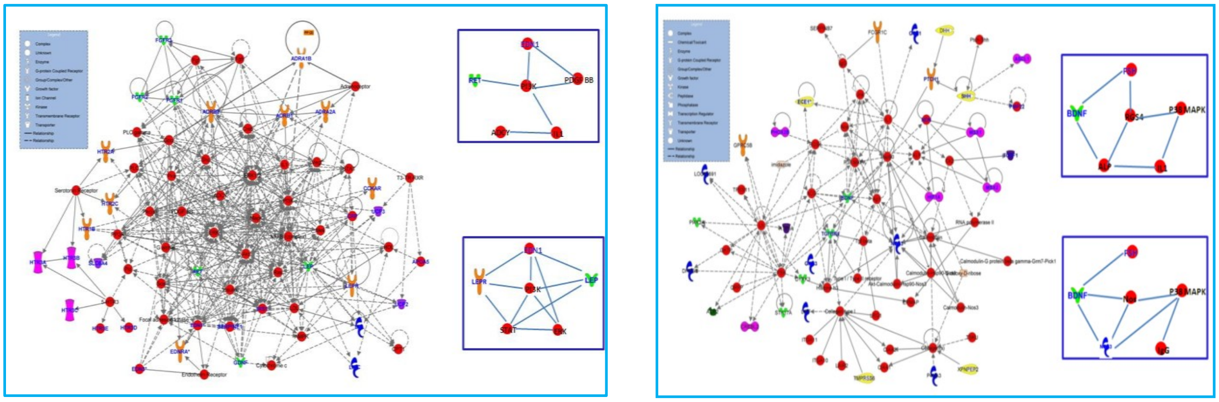
Finally, MCODE and BiNGO were applied to analyze the sub-networks detected, e.g., detecting the functional modules and identifying the enrichment of GO category[34,35].

## 3. Results and Discussion

### 3.1 *Generating and analyzing networks from seed genes/proteins*

We used IPA to generate networks using the 56 seed proteins related to sleep disorders. The top two scoring networks were used for further analysis. Among 70 proteins in each network, network 1 (Left figure in Figure 1) contains 32 seed proteins. The enriched functions for this network as identified by IPA include neurological disease, organismal injury and abnormalities, and genetic disorders. Network 2 (Right figure of Fig. 1) includes 16 seeds, and the associated functions are genetic disorder, neurological disease, and respiratory disease.

A quantitative method to detect and score sub-networks within the networks was applied to identify sub-networks that are highly discriminative for the OSA phenotype based on transcriptional dysregulation using mRNA expression data from subcutaneous and visceral fat tissues[20]. To limit the computational overhead of the calculation while using exhaustive search, we constrained the search where sub-networks were limited to six nodes. This analysis of network 1 provided 108 sub-networks of 6 nodes using expression data from subcutaneous fat tissue, and 97 sub-networks of 6 nodes from visceral tissue that had the maximum possible values of MI. In case of network 2, 9 sub-networks are detected for subcutaneous tissue, and 8 for visceral tissue. Further analyses focus on these sub-networks.



Network 1 Network 2

Fig. 1 Networks generated using IPA with highest score (proteins name in blue indicates seed proteins), subnetworks with 6 nodes are identified by MI scores for subcutaneous and visceral fat tissues. Larger and high resolution picture can be found at http://proteomics.case.edu/news_events. aspx?newsid=38

In order to analyze the sub-networks, we calculated the frequency of occurrence of proteins in these sub-networks. We assume that the proteins that appear most frequently will likely be significant in terms of defining differences between the OSA phenotype and control. To reduce the incidence of false positives, we focused on the proteins that are in the top 6 in frequency for both tissues, which are listed in Table 1.

Table 1a Protein detected in subnetworks from network 1(Figure 1) and its frequency in the exhaustive search

| Protein (subcutaneous fat) | Frequency | Probability in detected subnetwork* | Protein (visceral fat) | Frequency | Probability in detected subnetwork* |
|---|---|---|---|---|---|
| PDGF BB | 55 | 50.9% | ERK | 53 | 54.6% |
| EDN1 | 43 | 39.8% | EDN1 | 34 | 35.0% |
| IL1 | 43 | 39.8% | STAT | 31 | 31.9% |
| PI3K | 38 | 35.1% | PI3K | 26 | 26.8% |
| RET | 27 | 25.0% | LEP | 26 | 26.8% |
| ADCY | 25 | 23.1% | LEPR | 24 | 24.7% |

Table 1b Protein detected in subnetworks from network 2 (Figure 1) and its frequency in the exhaustive search

| Protein (subcutaneous fat) | Frequency | Probability in detected subnetwork* | Protein (visceral fat) | Frequency | Probability in detected subnetwork* |
|---|---|---|---|---|---|
| P38 MAPK | 5 | 55.6% | BDNF | 4 | 50.0% |
| RGS4 | 4 | 44.4% | P38 MAPK | 4 | 50.0% |
| FSH | 4 | 44.4% | NOS3 | 3 | 37.5% |
| BDNF | 4 | 44.4% | FSH | 3 | 37.5% |
| IL1 | 3 | 33.3% | Nos | 3 | 37.5% |
| ALP | 3 | 33.3% | IgG | 3 | 37.5% |

* Calculated by Frequency/(total number of sub-networks with maximum MI)

Notably, 14 out of 24 proteins in table 1 are not seed proteins, and potentially indicate novel findings discovered by our approach. A number of proteins listed in table 1 are associated with OSA or other sleep phenotypes based on previous studies. For example, Endothelin 1 (EDN1), a potent vasoconstrictor implicated in hypertension, is both a seed protein and is ranked as second most frequent node for both tissues in the network 1 analysis (Table 1). Studies using knockout mice show that EDN1 is associated with respiratory distress[36], and more recently, association studies suggests that a missense coding SNP in EDN1 is linked with OSA in a European American sample[19]. The phosphorylation of ERK (Extracellular Signal-Regulated Kinase), the most frequently identified protein in visceral fat from network 1, is correlated with sleep patterns in flies[37]. PDGF BB (subunit of platelet-derived growth factor) the most frequently identified protein in subcutaneous fat from network 1, is a growth factor that regulates cell growth and division. There is evidence for the role of PDGF BB in disordered breathing from the responses of rats to hypoxia[38-40]. Follicle-stimulating hormone (FSH), seen in both fat analyses of network 2, is a hormone found in humans and other animals. Recent studies show that the concentration of FSH has a significant correlation with the obstructive apnea index in cerebrospinal fluid.[41]

Aside from many proteins that are directly related to OSA or sleep, the sub-networks also contain proteins that are known to be involved in processes related to sleep, but have not been reported to have specific associations with OSA. P38 MAPK (a frequently observed sub-network member from the analysis of network 2) is a member of the mitogen-activated protein kinases (MAPK) that play crucial roles in signaling the inflammatory response and are involved in pathways that respond to oxidative stress[42,43]. As indicated above, both processes are known to be related to OSA[4]. Another protein, Phosphatidylinositol 3-kinases (PI3K) is ranked in the top four in both tissues (Table 1, network 1). PI3Ks are a group of lipid kinases that catalyze the phosphorylation of phosphatidylinositols and phosphoinositides. They are composed of one 85 kDa regulatory subunit and one 110 kDa catalytic subunit. PIK3R genes (such as PIK3R1, PIK3R2, PIK3R3, PIK3R5, etc), encode the p85 regulatory subunit, while PIK3C genes (such as PIK3C3, PIK3CA, PIK3CB, PIK3CD, etc), code for the p110 catalytic subunit. It has been reported that PI3K is associated with fatty acid-induced insulin resistance[44], and although OSA and insulin resistance may be causally related, the exact mechanism linking them has not been fully elucidated [4]. Another top gene, STAT, encodes a family of transcription factors. In response to cytokines and growth factors, STAT family members are phosphorylated by the receptor associated kinases, and then translocated to the cell nucleus where they act as transcription activators. In a recent report, STAT4 was found to be involved in metabolic processes, especially in insulin resistance and inflammation in adipose tissue.[45]

### 3.2 *Analyzing interactome in adipose*

An interactome relevant to adipose tissue was generated from a combination of public interaction databases and gene expression profiles and contains 2909 proteins and 8323 interactions (Supplementary Table 2). Analyses of the topological parameters of the network show that it possesses typical properties of realistic networks,[46,47] such as small-world properties (the average shortest path length is 4.5). The node degree distribution fits a power law distribution.

We searched this interactome for OSA related sub-networks by mapping p-values from the GWAS study to proteins of the interactome [21]. Then, cytoscape and its plugin jactivemodule are applied to detect sub-networks that are significant. The jactivemodule combines the network structure and associated p-value of each protein to extract potential meaningful sub-networks. A subnetwork with 203 proteins and 324 interactions is identified with a significant score (7.09, Figure 2, subnetworks with score > 3.0 are considered as significant[32]). Similar to the whole interactome, this sub-network shows some typical properties, such as small-world and power-law distribution of node degree. Note that many of the nodes have modest p-values (low z-scores), and would not be seen as significant in a conventional GWAS analysis. For example, the p-value of hepatocyte growth factor-regulated tyrosine kinase substrate (HGS) is 0.62, but its interacting partners (neurofibromin 2 (NF2), signal transducing adaptor molecule (STAM and, STAM2)) have p-value less than 0.006, thus, it is included in the subnetwork. Other similar examples are minichromosome maintenance complex component 7 (MCM7, p-value: 0.97) and SHC (Src homology 2 domain containing) transforming protein 1 (SHC1, p-value: 0.35).

Another cytoscape plugin MCODE was applied to explore the protein complexes or other modules present in the sub-network identified by jactivemodule. MCODE detects densely connected regions in a network that may represent functional modules. It is based on vertex weighting by local neighborhood density and outward traversal from a locally dense seed protein to isolate the dense regions. The top two clusters identified by MCODE are listed in Figure 3. The cluster with best score has ten proteins that are densely connected. Nine out of ten components are proteasome subunits. The proteasome is a large, multimeric protein complex with regulatory and catalytic functions. It is responsible for degrading damaged, misfolded, nonfunctional and potentially toxic proteins. Notably, it has been reported that the proteasome pathway and proteasomal activity are associated with OSA and hypoxia, a central feature of OSA [48,49].

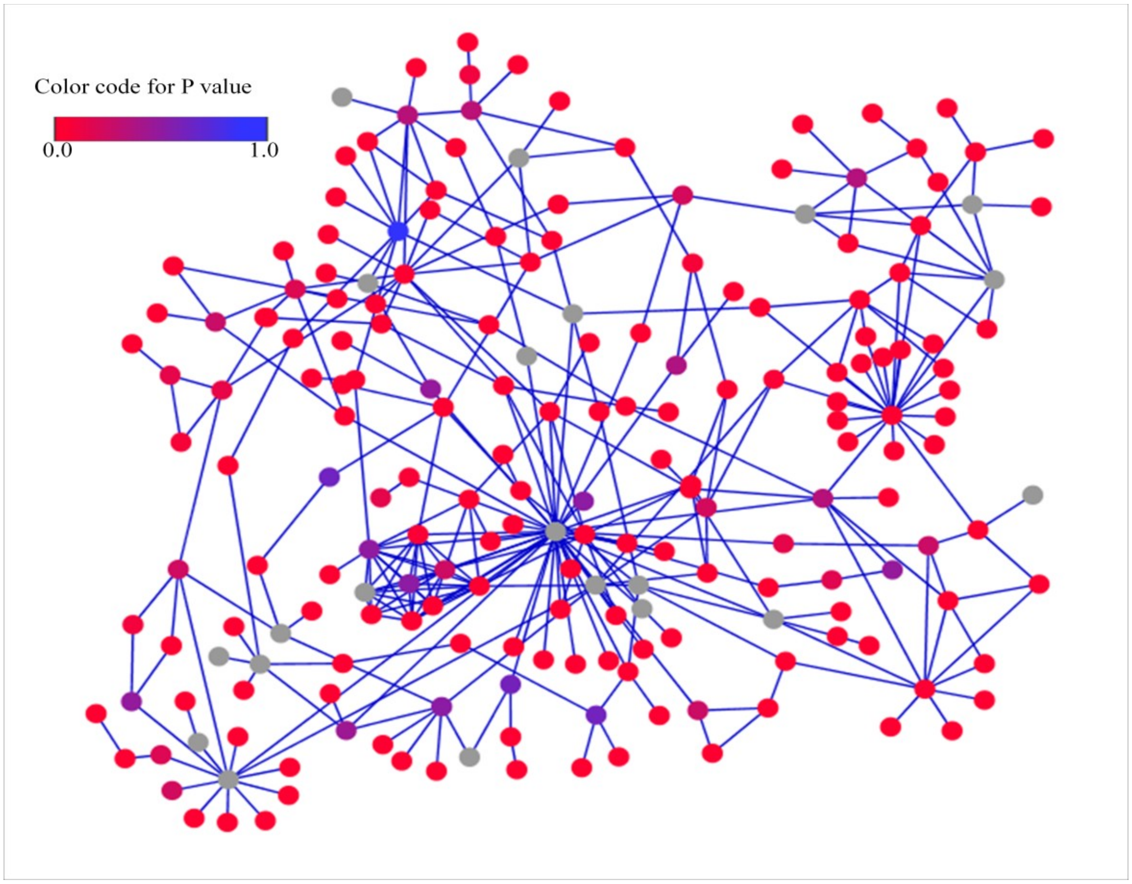


Figure 2 Network identified by jactivemodule using p-values from GWAS study, color represents the p-values and nodes with grey color indicate that the p-values are missing from GWAS. High resolution picture with the node lable can be found at http://proteomics.case.edu/news_events. aspx?newsid=38

To determine which Gene Ontology (GO) functional categories are statistically overrepresented in the sub-network, we further applied the BiNGO program to the sub-network of Figure 2. The detected functions include axon extension, spliceosome assembly, protein catabolic process, insulin receptor signaling pathway, and negative regulation of tyrosine phosphorylation of STAT3 proteins. Recent studies suggest that STAT3 tyrosine phosphorylation is critical for interleukin protein production in the inflammatory response [45]. Also, STAT family members are implicated in several processes relevant to tumor growth, providing an additional link aside from PI3K between OSA and cancer.

As there is an association between OSA and diabetes[50], the functional enrichment for the insulin receptor-signaling pathway deserves closer investigation. Three proteins in the sub-network are responsible for the enrichment of this function: PIK3R1, IRS2, and IGF1R. PIK3R1 (phosphoinositide-3-kinase, regulatory subunit 1) phosphorylates the inositol ring of phosphatidylinositol at the 3-prime position and plays an important role in the metabolic actions of insulin; IRS2 (insulin receptor substrate 2) mediates effects of insulin by acting as a molecular adaptor between diverse receptor tyrosine kinases and downstream effectors; IGF1R (insulin-like growth factor 1 receptor) binds insulin-like growth factor with a high affinity and modulates insulin's actions.. Notably, these three proteins plus YWHAG (tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein, gamma polypeptide) densely connect, forming a cluster in the subnetwork that is also detected by MCODE (Figure 3).

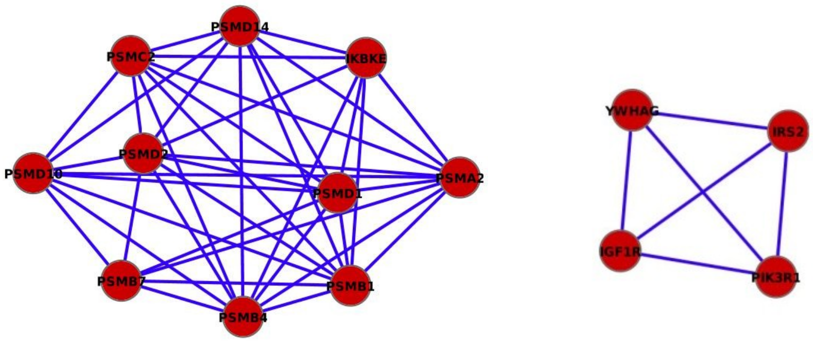


Figure 3 Densely connected subnetworks identified using MCODE, those represent potentially functional module or protein complex

### 3.3 *Comparison and limitation of approaches*

In this study, we took two systems biology approaches to detect subnetworks which are likely associated with OSA. Because of the nature of two approaches (the first one is biased and based on prior knowledge of OSA; the second one is unbiased), it is hard to compare them, and it is not surprising that the results are different. These two approaches use SNP data from GWAS and gene expression data from microarray experiments respectively, and treat them independently. Also the data are from two different sources (SNP data derived from CARe project[21], and gene expression data from other sources (Patel, S, et al, unpublished data).

One limitation of our approach is that the method for detection of subnetworks using MI is computationally extensive, and can only be applied on small networks. Further efforts are necessary to improve its efficacy. Another limitation is the method to derive the significance level of proteins based on the SNP data. Usually, there are multiple SNPs located within the regions for each gene. Although several methods have been proposed to condense this informaiton[30,31,52], we applied the simple and most commonly used one: consider the most significant p-value among SNPs as p-value of proteins as other methods may provide conflicting results.

## 4. Conclusion

Our integrated analysis of mRNA expression from adipose tissues, PPI networks, and SNP data from genome-wide association studies provides a novel approach for combining data from disparate sources to identify candidate pathways for potential validation studies. Some of the associations identified may reflect pathways that predispose to OSA, while others may indicate pathways that are perturbed by OSA-related stresses which contribute to co-morbidities such as diabetes. The results of this initial study suggest that the PI3K, the STAT protein family, and insulin signaling may be associated with OSA. Further investigation is needed to elucidate the exact role of these genes and their gene products in OSA. In addition, our approach outlines a novel application of SNP data in sub-network discovery relevant to disease that is consistent with other well-accepted methodologies. Thus, we suggest this approach could be generally applied to the analysis of GWAS data that is available for over 100 other diseases.

## 5. Acknowledgments

This research is supported by the Case Center for Transdisciplinary Research on Energetics and Cancer, U54-CA-116867 (PI: Berger, Nate) and Case Comprehensive Cancer Center, P30-CA-043703 (PI: Gerson, Stanton), from the National Cancer Institute, UL1-RR024989 (PI: Davis, Pamela) from the National Center for Research Resources (Clinical and Translational Science Awards), Case Proteomics Center for HIV/AIDS and Drug Abuse, P20-DA26113 (PI: Chance, Mark) from the National Institute for Drug Abuse, CFS HL 463680 (PI: Redline, Susan) and by NSF CAREER Award CCF-0953195 (PI: Koyuturk, Mehmet). NHLBI's Candidate Gene Association Resource has supported extensive genotyping and has created a genotype-phenotype database using data and samples from nine cohorts, including the Atherosclerosis Risk in Communities (ARIC) study, the Cardiovascular Health Study (CHS), the Cleveland Family Study, the Cooperative Study of Sickle Cell Disease (CSSCD), the Coronary Artery Risk Development in Young Adults (CARDIA) study, the Framingham Heart Study (FHS), the Jackson Heart Study (JHS), the Multi-Ethnic Study of Atherosclerosis (MESA), and the Sleep Heart Health Study (SHHS). Detailed information on the grants and contracts that have supported CARe and its parent cohorts can be found at http://public.nhlbi.nih.gov/GeneticsGenomics/home/care.aspx.

## References

1 Reite, M., Ruddy, J. & Nagel, K. *Concise guide to evaluation and management of sleep disorders*. (American Psychiatric Publishing, Inc., 2002).
2 Vgontzas, A. N. & Kales, A. *Annu Rev Med* **50**, 387-400, (1999).
3 Young, T., Peppard, P. E. & Gottlieb, D. J. *Am J Respir Crit Care Med* **165**, 1217-1239 (2002).
4 Arnardottir, E. S., et al. *Sleep* **32**, 447-470 (2009).
5 Waradekar, et al. *Am J Respir Crit Care Med* **153**, 1333-1338 (1996).
6 Donadio, V. *et al. J Sleep Res* **16**, 327-332, (2007).

7 Bravo Mde, L. *et al. Sleep Breath* **11**, 177-185 (2007).

8 Minoguchi, K. *et al. Am J Respir Crit Care Med* **172**, 625-630, (2005).

9 Schulz, R. *et al. Am J Respir Crit Care Med* **162**, 566-570 (2000).

10 Dyugovskaya, L., Lavie, P. & Lavie, L. *Am J Respir Crit Care Med* **165**, 934-939 (2002).

11 Young, T. *et al. N Engl J Med* **328**, 1230-1235 (1993).

12 Ip, M. S. *et al. Am J Respir Crit Care Med* **165**, 670-676 (2002).

13 McNicholas, W. T. & Bonsigore, M. R. *Eur Respir J* **29**, 156-178, (2007).

14 Logan, A. G. *et al. Eur Respir J* **21**, 241-247 (2003).

15 Strohl, K. P., Saunders, N. A., Feldman, N. T. & Hallett, M. *N Engl J Med* **299**, 969-973 (1978).

16 Redline, S. *et al. Am J Respir Crit Care Med* **151**, 682-687 (1995).

17 Riha, R. L. *Respiration* **78**, 5-17, (2009).

18 Larkin, E. K. *et al. Ann Hum Genet* **72**, 762-773,(2008).

19 Larkin, E. K. *et al. Am. J. Respir. Crit. Care Med.* in press (2010).

20 Nibbe, R. K. *et al. Mol Cell Proteomics* **8**, 827-845, (2009).

21 Musunuru, K. *et al. Circ Cardiovasc Genet*, in press (2010).

22 Bossi, A. & Lehner, B. *Mol Syst Biol* **5**, 260, (2009).

23 Wang, E. T. *et al. Nature* **456**, 470-476, (2008).

24 Szabo, P. M. *et al. Oncogene* **29**, 3163-3172, (2010).

25 Su, A. I. *et al. Proc Natl Acad Sci U S A* **99**, 4465-4470, (2002).

26 Su, A. I. *et al. Proc Natl Acad Sci U S A* **101**, 6062-6067, (2004).

27 Assenov, Y. et al. *Bioinformatics* **24**, 282-284, (2008).

28 Bengtsson, H., et al. (Dep. of Statistics, Univ. of California, Berkeley, February 2008).

29 Chuang, H. Y., Lee, E., Liu, Y. T., Lee, D. & Ideker, T. *Mol Syst Biol* **3**, 140, (2007).

30 Wang, K., Li, M. & Bucan, M. *Am J Hum Genet* **81**, (2007).

31 Baranzini, S. E. *et al. Hum Mol Genet* **18**, 2078-2090, (2009).

32 Ideker, T., et al. *Bioinformatics* **18 Suppl 1**, S233-240 (2002).

33 Cline, M. S. *et al. Nat Protoc* **2**, 2366-2382, (2007).

34 Bader, G. D. & Hogue, C. W. *BMC Bioinformatics* **4**, 2 (2003).

35 Maere, S., Heymans, K. & Kuiper, M. *Bioinformatics* **21**, 3448-3449, (2005).

36 Kuwaki, T. *et al. Am J Physiol* **270**, R1279-1286 (1996).

37 Foltenyi, K., Greenspan, R. J. & Newport, J. W. *Nat Neurosci* **10**, 1160-1167, (2007).

38 Alea, O. A. *et al. Am J Physiol Regul Integr Comp Physiol* **279**, R1625-1633 (2000).

39 Vlasic, V., Simakajornboon, N., Gozal, E. & Gozal, D. *Pediatr Res* **50**, 236-241 (2001).

40 Gozal, D. *et al. J Neurochem* **74**, 310-319 (2000).

41 Capatina, C., et al. *Endocrine Abstracts* **22**, P631 (2010).

42 Ryan, S., et al. *Biochem Biophys Res Commun* **355**, 728-733, (2007).

43 Hu, J. Y. *et al. Cell Mol Life Sci* **67**, 321-333, (2010).

44 Kruszynska, Y. T. *et al. J Clin Endocrinol Metab* **87**, 226-234 (2002).

45 Samavati, L. *et al. Mol Immunol* **46**, 1867-1877, (2009).

46 Uetz, P. *et al. Nature* **403**, 623-627, (2000).

47 Ito, T. *et al. Proc Natl Acad Sci U S A* **97**, 1143-1147 (2000).

48 Gozal, D. *et al. J Neurochem* **86**, 1545-1552, (2003).

49 Taylor, C. T., et al. *Proc Natl Acad Sci U S A* **97**, 12091-12096, (2000).

50 Shaw, J. E., et al. *Diabetes Res Clin Pract* **81**, 2-12, (2008).

51 Chowdhury S.A., Nibbe, R.K., Chance, M.R., Koyuturk, M. *Proc. RECOMB*, LNCS 6044, 80-95, (2010)

52 Yu, K, et al. *Genet Epidemiol*. 33(8):700-9, (2009)