Network Expansion and Pathway Enrichment Analysis towards Biologically Significant Findings from Microarrays

Xiaogang Wu[1,2,*], Hui Huang[1,*], Tao Wei[3], Ragini Pandey[2], Christoph Reinhard[3], Shuyu D. Li[3,†], Jake Y. Chen[1,2,†]

[1]School of Informatics, Indiana University, Indianapolis, IN 46202, USA

[2]MedeoLinx, LLC, Indianapolis, IN 46280, USA

[3]Eli Lilly and Company, Indianapolis, IN 46285, USA

Summary

In many cases, crucial genes show relatively slight changes between groups of samples (e.g. normal vs. disease), and many genes selected from microarray differential analysis by measuring the expression level statistically are also poorly annotated and lack of biological significance. In this paper, we present an innovative approach - network expansion and pathway enrichment analysis (NEPEA) for integrative microarray analysis. We assume that organized knowledge will help microarray data analysis in significant ways, and the organized knowledge could be represented as molecular interaction networks or biological pathways. Based on this hypothesis, we develop the NEPEA framework based on network expansion from the human annotated and predicted protein interaction (HAPPI) database, and pathway enrichment from the human pathway database (HPD). We use a recently-published microarray dataset (GSE24215) related to insulin resistance and type 2 diabetes (T2D) as case study, since this study provided a thorough experimental validation for both genes and pathways identified computationally from classical microarray analysis and pathway analysis. We perform our NEPEA analysis for this dataset based on the results from the classical microarray analysis to identify biologically significant genes and pathways. Our findings are not only consistent with the original findings mostly, but also obtained more supports from other literatures.

1 Background

Microarrays make possible the discovery of new functions and pathways of known genes, as they measure all the transcriptional activity in a biological sample [1]. This high-throughput procedure can be used in medical diagnostics, in biomarker discovery, and in investigating the ways a drug, disease, polymorphism or environmental condition affects gene expression and function [2, 3]. However, one challenge has arisen because microarray technology generates a large amount of transcriptional data, which is hard to interpret for the results to gain insights into biological mechanisms [4]. As a result, researchers have sought to analyze microarray data through the use of modern computational tools and statistical methods.

In many cases, crucial genes show relatively slight changes, and many genes selected from differential analysis between groups of samples (e.g. normal vs. disease) by measuring the expression level statistically are also poorly annotated [2]. From a biological perspective, functionally related genes often display a coordinated expression to accomplish their roles in

* These authors contributed equally to this work.

† To whom correspondence should be addressed. Email: jakechen@iupui.edu, li_shuyu_dan@lilly.com

doi:10.2390/biecoll-jib-2012-213

the cell [5]. Hence, to translate such lists of differentially expressed genes into a functional profile able to understand the underlying biological phenomena, one approach to aid interpretation is to look for changes in a group of genes with a common function [2].

Gene set enrichment analysis (GSEA) is one of the most widely used methods for identifying both statistically and biologically significant genes from high-throughput data such as gene-expression assays [4]. GSEA relies on pre-defined gene sets, while neglect gene/protein interaction, pathway upstream or downstream information. Furthermore, GSEA still assumes that more differentially expressed genes are more crucial to the biology, which is not always true [6]. Currently, gene expression signature analysis and pathway analysis remain two separate processes.

From a view of network biology [7], cancer genes and proteins do not function in isolation; instead, they work in interconnected pathways and molecular networks at multiple levels [8], one study re-characterized them in a molecular interaction network for BRCA, and identified HMMR as a new susceptibility locus [9]. Another study integrated protein interaction network and gene expression data to improve the prediction of BRCA metastasis [10]. These works suggest that protein interaction networks and pathways, although noisy, incomplete and static, can serve as a molecular-level conceptual roadmap to guide future microarray analysis [11].

In this paper, we present an innovative approach - network expansion and pathway enrichment analysis (NEPEA) for integrative microarray analysis. We assume that organized knowledge will help microarray data analysis in significant ways, and the organized knowledge could be represented as molecular interaction networks or biological pathways. Based on this hypothesis, we develop the NEPEA framework based on network expansion [12] from the human annotated and predicted protein interaction (HAPPI) database [13], and pathway enrichment from the human pathway database (HPD) [14].

We use a recently-published microarray dataset (GSE24215) related to insulin resistance and type 2 diabetes (T2D) as case study, since this study provided a thorough experimental validation for both genes and pathways identified computationally from classical microarray analysis and pathway analysis [15]. In this study, skeletal muscle samples were collected in all participants (n = 20) in both the basal and insulin-stimulated state before and after bed rest. We perform our NEPEA analysis for this dataset based on the results from the classical microarray analysis to identify biologically significant genes and pathways. Our findings are not only consistent with the original findings mostly, but also obtained more supports from other literatures.

2 Methods

The NEPEA method has three main components: 1) classical microarray analysis for data preprocessing consisting of quality control, normalization and differential analysis, 2) network expansion analysis for significant gene identification consisting of disease gene curation, network construction and significance score calculation, and 3) pathway enrichment analysis consisting of pathway search, pathway differential analysis and ranking. Using the microarray dataset - GSE24215 as an example, we introduce the detailed steps below:

2.1 Microarray data preprocessing

2.1.1 Quality Control

We use AffyQCReport (applicable for Affymetrix platform) and ArrayQualityMetrics (applicable for Agilent platform) packages in Bioconductor to generate three plots to detect bad chips for each microarray dataset as: 1) examine a heat map that shows array-array

Spearman rank correlation coefficients. The map enabled us to plot outliers, failed hybridizations, and mis-tracked samples; 2) make a box plot of all perfect match intensities. The plot enabled us to detect outliers in terms of average intensity; and 3) make a distribution plot of kernel density estimates for perfect match intensities, which enables us to detect outliers in terms of shaped density. After applying ArrayQualityMetrics packages into quality control for microarray dataset - GSE24215, total 3 suspects out of 48 samples are flagged, which are kicked off as bad chips.

2.1.2 Normalization

We use Quantile normalization to normalize all the four qualified microarray datasets; MAS5 for Affymetric platform and normexp for Agilent platform on background correction. We also perform the steps background correction, normalization, probe specific correction, and summary value computation as following: 1) bgcorrect.method: mas; 2) normalize.method:quantiles; 3) pmcorrect.method:pmonly; and 4) summary.method:mas.

2.1.3 Differential analysis

We use Limma (Linear Models for Microarray Data) package [16] in Bioconductor to identify differentially-expressed genes for each clinical group comparison from the qualified and normalized microarray datasets as 1) The package Limma uses an approach called linear models to analyze designed microarray experiments; 2) For statistical analysis and assessing deferential expression, Limma uses an empirical Bayes method for more stable inference and improved power, especially for experiments with small numbers of arrays; and 3) Differential genes are obtained by using the filters with p-Value <= 0.05, Fold Change (FC) >= 1.3, and Average Expression Level (AEL) >= 40% after applying Limma package in Bioconductor. Average expression levels (AEL>=40%) have been checked to ensure the presences of the differential genes in the tissue - muscle. Duplicated genes with lower fold changes are eliminated, which implies that only the highest fold change for one gene will be kept. For microarray dataset - GSE24215, we get 495 differential genes from insulin before-bed (IBB) group, and 930 differential genes from insulin after-bed (IAB) group

2.2 Network expansion analysis

2.2.1 Disease gene curation

The network expansion analysis is knowledge-guided approach, which relies on the disease-associated genes. Here we use T2D as an example to demonstrate how to curate disease-associated genes, but our method can be applied to any other disease phenotypes.

We curate T2D-associated genes from OMIM (http://www.ncbi.nlm.nih.gov/omim) manually, evaluates them semi-automatically through searching in PubMed (http://www.ncbi.nlm.nih.gov/pubmed/) as following: 1) Query: ("Type II Diabetes"[All Fields] OR "Type 2 Diabetes"[All Fields]) AND (prefix star[prop] OR prefix plus[prop]); 2) Results: Records (Entries) -> Genes (Gene Symbol) -> Proteins (Uniprot ID); 3) GENE: Gene name, linked to GeneCards.org; 4) UNIPROT: Uniprot ID, linked to UniProt.org; 5) PUBMED: Count number of references where both term ("Type II Diabetes" OR "Type 2 Diabetes") AND "GENE" appeared in PubMed, linked to PubMed; and 6) Obtain interactions from HAPPI 1.31 for these T2D-associated genes (seed genes) curated from OMIM.

2.2.2 Network reconstruction

We construct a T2D-specific protein-protein interaction (PPI) network by using Oracle SQL Developer with high-quality interaction data in HAPPI version 1.31 and map differentially-expressed genes onto the T2D-specific PPI network by using Cytoscape as following: 1) Expand 39 seed genes (PUBMED >=50) in HAPPI 1.31 (4-Star, h-Score >=0.75), and obtain 702 genes (including 32 seed genes); 2) The left 7 seed genes are also added into the network in order to show their expressions; and 3) Construct a T2D-specific PPI network with 709 nodes and 944 edges, by using Nearest Neighbor Expansion (NNE) approach [12].

2.2.3 Significant gene identification

We measure and rank all the differential genes in a T2D-specific protein-protein interaction (PPI) network by considering both differential expressions and network properties. Differential genes are obtained by applying filters with p-Value <= 0.05, Average Expression Level (AEL) >= 40%, and Absolute Fold Change (ABS_FC) >= 1.3. Duplicated genes with lower fold changes have been eliminated, which implies that only the highest fold change for one gene will be kept. The T2D-specific PPI network is reconstructed by expanding all the seed genes curated from OMIM (PUBMED >=0), in HAPPI_1.31 (3-Star) (Confidence: h-Score >=0.45)

We define Gene Significance Score (integrating both gene expression fold change - FC and network connectivity - NC) here as:

Sig_Score = $\left(\alpha_1 + log_2^{|FC|}\right) \times log_2(\alpha_2 + NC)$, $|FC|$ = ABS_FC, absolute fold change.

Constant parameters α_1 and α_2 here are for the balance between differential expressions and network properties. In the implementation, α_1=3 and α_2=1 have best performance to rank known significant genes in the front. Network connectivity (for un-weighted networks) NC: Weight_1 = Number of direct neighbors for each node. Network connectivity (for weighted networks) NC: Weight_2 = Sum of connection strength values on all neighbored edges. In the implementation, we use Weight_2 here. Connection strength here is the confidence for an interaction: h-Score.

2.3 Pathway enrichment analysis

2.3.1 Pathway search

We search curated T2D-associated genes by using Oracle SQL Developer with comprehensive integrated pathway data in HPD version 2.1 (including pathway data from NCI-Nature curated, KEGG, BioCarta, and Protein Lounge), and map differentially-expressed genes onto the pathways obtained. We obtain 92 pathways with (HITS/Pathway Scale) >= 3.5% AND HITS >=2 by querying 39 seed genes (PUBMED >=50) in HPD 2.1.

2.3.2 Pathway differential analysis

We provide average differential gene expressions in a pathway as:

AVG_ABS_FC: The average of ABS_FC of all the available differential gene expressions in a pathway.

We define pathway differential expressions here as:

NORM_ABS_FC: The *p**-norm of ABS_FC of all the available differential gene expressions in a pathway

Usually, p-norm = $(\sum_{i=1}^{n}(x_i)^p)^{\frac{1}{p}} = (SUM((x_i)^p)^{1/p}$

For unification, we modify it as p*-norm = $(\frac{1}{n}\sum_{i=1}^{n}(x_i)^p)^{\frac{1}{p}} = (AVG((x_i)^p)^{1/p}$

In the implementation, p = 6 have best performance to emphasize highly differential expressions in a pathway.

We also provide maximal differential gene expressions in a pathway as:

MAX_ABS_FC: The maximum value of ABS_FC of all the available differential gene expressions in a pathway;

and count number of differentially expressed genes as:

CNT_DIFF: Count number of differentially expressed genes (FC >= 1.3 AND p-Value <= 0.05) in a pathway.

We rank all the pathways by their pathway differential expressions - NORM_ABS_FC defined above.

3 Results

3.1 Findings on insulin before-bed (IBB) group

3.1.1 Top-20 differential genes

Totally 495 differential genes are obtained, which are differentially-expressed in Insulin Before Bed (IBB) Group from the microarray dataset – GSE24215. Differential genes are obtained by using filters with p-Value <= 0.05, Fold Change (FC) >= 1.3, and Average Expression Level (AEL) >= 40% after applying Limma package in Bioconductor. Average expression levels (AEL>=40%) have been checked to ensure the presences of the differential genes in the tissue - muscle. Duplicated genes with lower fold changes are eliminated, which implies that only the highest fold change for one gene will be kept. Top-20 differential genes in IBB from GSE24215, ordered by absolute fold change (ABS_FC), are listed in Table 1.

3.1.2 Top-20 significant genes

Totally 130 significant genes in IBB from GSE24215 are obtained from all the differential genes in a T2D-specific protein-protein interaction (PPI) network, measured by using significant score (considering both differential expressions and network properties). The T2D-specific PPI network is reconstructed by expanding all the seed genes curated from OMIM (PubMed >=0), in HAPPI_1.31 (3-Star) (Confidence: h-Score >=0.45). Top-20 significant genes in IBB from GSE24215, ordered by significant score (Sig_Score), are listed in Table 2.

A T2D-significant protein-protein interaction (PPI) network (See Figure 1) is reconstructed a by connecting Top-20 significant genes in IBB from GSE24215, with and within the T2D-associated genes (seed genes) curated from OMIM (PubMed >= 50), in HAPPI_1.31 (3-Star) (Confidence: h-Score >=0.75).

Table 1: Top-20 differential genes in IBB from GSE24215, ordered by FC, (FC>=1.3, p-value <=0.05 and AEL >=40% after applying Limma package in Bioconductor). Note: Gene names are linked to GeneCards.org, UniProt IDs are linked to UniProt.org, and Evidences are linked to PubMed.

Gene Symbol	*p*-Value	FDR	Log2_FC	ABS_FC	Evidences
SOCS3	0.00193	0.08858	2.54455	**5.83426**	28
PDK4	0.00000	0.00074	-2.34193	**5.06980**	16
THBD	0.00001	0.00243	2.25714	**4.78043**	0
CISH	0.00013	0.01380	2.19425	**4.57651**	0
G0S2	0.00000	0.00003	2.05403	**4.15264**	0
MYC	0.00064	0.04234	1.97513	**3.93164**	23
PDE4B	0.00000	0.00042	1.82895	**3.55280**	0
ADAMTS4	0.00061	0.04111	1.76371	**3.39569**	1
GADD45A	0.00002	0.00373	1.76132	**3.39008**	0
RGS16	0.00217	0.09630	1.72508	**3.30598**	1
EGR1	0.01342	0.29638	1.71863	**3.29125**	3
HES1	0.00000	0.00012	1.71837	**3.29065**	1
CCL2	0.00019	0.01796	1.71466	**3.28219**	111
KLF15	0.00000	0.00000	-1.66849	**3.17882**	3
PYCR1	0.00000	0.00000	1.66356	**3.16797**	0
CITED2	0.00000	0.00000	-1.65106	**3.14064**	0
OTUD1	0.00006	0.00778	-1.56650	**2.96186**	0
ARRDC4	0.00000	0.00000	1.51143	**2.85092**	0
NR1D1	0.00000	0.00003	-1.50730	**2.84277**	1
PIK3R1	0.00000	0.00000	1.50274	**2.83379**	9

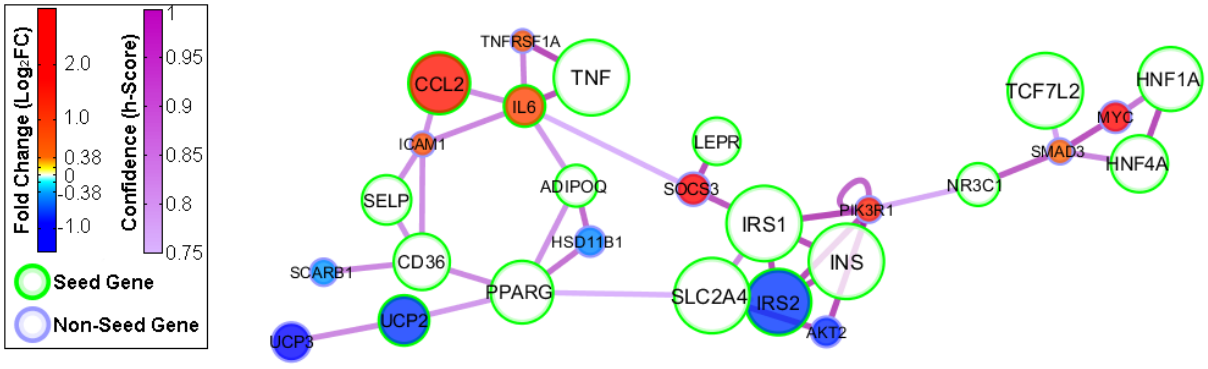

Figure 1: Top-20 significant genes in IBB from GSE24215, interacted with T2D-associated genes. Node size represents Evidence for each gene, node color represents Log2_FC, red color implies over-expressed and blue color implies under-expressed. Green circled nodes are seed genes (T2D-associated genes curated from OMIM). Edge color represents Confidence (h-Score) for each interaction.

Table 2: Top-20 significant genes in IBB from GSE24215, ordered by Sig_Score, which is measured in the T2D-specific PPI network (PubMed >=0, h-Score >=0.45) for all the differential genes (FC>=1.3, p-value <=0.05 and AEL >=40% after applying Limma package in Bioconductor) in IBB from GSE24215.

Gene Symbol	*p*-Value	FDR	Log2_FC	ABS_FC	Weight_1	Weight_2	Sig_Score	Evidences
CCL2	0.00019	0.01796	1.71466	3.28219	98	75.2645	**29.48048**	111
IL6	0.00164	0.08069	0.96338	1.94987	140	112.808	**27.07169**	52
AKT2	0.00133	0.06955	-0.83609	1.7852	104	66.6378	**23.32247**	21
IRS2	0.00011	0.01276	-0.78851	1.72729	60	49.3156	**21.4162**	124
VEGFA	0.01432	0.30888	0.52022	1.43417	57	44.6818	**19.40888**	28
PIK3R1	0	0	1.50274	2.83379	13	11.8228	**16.57294**	9
MYC	0.00064	0.04234	1.97513	3.93164	10	8.8236	**16.39929**	23
UCP3	0.00001	0.00147	-1.04039	2.05679	22	15.2626	**16.25647**	45
SOCS3	0.00193	0.08858	2.54455	5.83426	7	6.3574	**15.96385**	28
UCP2	0	0.00093	-0.80665	1.74915	22	15.7096	**15.46493**	78
SCARB1	0.00115	0.06287	-0.38842	1.30896	30	19.9446	**14.87011**	11
HSD11B1	0.01673	0.33975	-0.46192	1.37737	24	17.478	**14.56683**	20
SORBS1	0	0	-1.13404	2.19472	13	10.08	**14.34464**	2
KLF11	0	0.00001	-0.5958	1.51131	20	14.1964	**14.11588**	8
AQP7	0.00023	0.02066	-0.71786	1.64474	9	7.305	**11.35427**	8
RRAD	0.00018	0.01748	1.45449	2.7406	7	4.8134	**11.31166**	8
LPIN1	0.00898	0.23354	-0.49106	1.40548	13	7.849	**10.98119**	9
SMAD3	0.01644	0.33728	0.4402	1.35679	9	8.1113	**10.96617**	7
ICAM1	0.00219	0.09689	0.74517	1.67617	8	6.539	**10.91482**	4
TNFRSF1A	0.00039	0.02999	0.5936	1.509	8	6.6686	**10.56144**	0

3.1.3 Top-20 significant pathways

Totally 51 significant pathways (*p**-norm >= 1.2) in IBB from GSE24215 are obtained from all the differential pathways, measured by using pathway differential expressions (*p**-norm). Top-20 significant pathways in IBB from GSE24215, ordered by pathway differential expressions (*p**-norm), are listed in Table 3.

3.2 Findings on insulin after-bed (IAB) group

3.2.1 Top-20 differential genes

Totally 930 differential genes are obtained, which are differentially-expressed After Bed (IAB) Group from the microarray dataset – GSE24215. Differential genes are obtained by using filters with *p*-Value <= 0.05, Fold Change (FC) >= 1.3, and Average Expression Level (AEL) >= 40% after applying Limma package in Bioconductor. Average expression levels (AEL>=40%) have been checked to ensure the presences of the differential genes in the tissue - muscle. Duplicated genes with lower fold changes are eliminated, which implies that only the highest fold change for one gene will be kept.

Top-20 differential genes in IAB from GSE24215, ordered by absolute fold change (ABS_FC), are listed in Table 4.

Table 3: Top-20 significant pathways in IBB from GSE24215, ordered by pathway differential expressions (p*-norm), which is measured with all the available differential gene expressions in IBB from GSE24215.

PATHWAY_NAME	DB_SOURCE_ID	NORM_ABS_FC	MAX_ABS_FC
IL-9 Pathway	KEGG	**2.56782**	3.97011
IL-10 Pathway	KEGG	**2.35238**	3.97011
IL23-mediated signaling events	NCI-Nature Curated	**2.29305**	3.97011
EPO signaling pathway	NCI-Nature Curated	**2.24041**	3.97011
Murine MSP-STK Signaling	KEGG	**2.20505**	3.28219
IL6-mediated signaling events	NCI-Nature Curated	**2.19750**	3.97011
Type II diabetes mellitus	KEGG	**2.16107**	3.97011
Signaling events mediated by PTP1B	NCI-Nature Curated	**2.13761**	3.97011
growth hormone signaling pathway	BioCarta	**2.09285**	3.31174
IL-4 Pathway	KEGG	**2.08410**	3.97011
Growth Hormone Signaling	KEGG	**2.06889**	3.97011
LDL Oxidation in Atherogenesis	KEGG	**2.05923**	3.28219
IL4-mediated signaling events	NCI-Nature Curated	**2.02714**	3.97011
Adipocytokine signaling pathway	KEGG	**1.97685**	3.97011
FoxO family signaling	NCI-Nature Curated	**1.97503**	3.39008
C. pneumoniae Infection in Atherosclerosis	KEGG	**1.89211**	3.28219
Calcineurin-regulated NFAT-dependent transcription in lymphocytes	NCI-Nature Curated	**1.87618**	3.29125
Jak-STAT signaling pathway	KEGG	**1.85834**	3.97011
MSP-RON Signaling	KEGG	**1.84756**	3.28219
Insulin signaling pathway	KEGG	**1.80963**	3.97011

3.2.2 Top-20 significant genes

Totally 237 significant genes in IAB from GSE24215 are obtained from all the differential genes in a T2D-specific protein-protein interaction (PPI) network, measured by using significant score (considering both differential expressions and network properties). The T2D-specific PPI network is reconstructed by expanding all the seed genes curated from OMIM (PubMed >=0), in HAPPI_1.31 (3-Star) (Confidence: h-Score >=0.45). Top-20 significant genes in IAB from GSE24215, ordered by significant score (Sig_Score), are listed in Table 5.

A T2D-significant protein-protein interaction (PPI) network (See Figure 2) is reconstructed a by connecting Top-20 significant genes in IAB from GSE24215, with and within the T2D-associated genes (seed genes) curated from OMIM (PubMed >= 50), in HAPPI_1.31 (3-Star) (Confidence: h-Score >=0.75)

Table 4: Top-20 differential genes in IAB from GSE24215, ordered by FC, (FC>=1.3, p-value <=0.05 and AEL >=40% after applying Limma package in Bioconductor). Note: Gene names are linked to GeneCards.org, UniProt IDs are linked to UniProt.org, and Evidences are linked to PubMed.

Gene Symbol	*p*-Value	FDR	Log2_FC	ABS_FC	Evidences
NR4A3	0.00000	0.00000	4.18431	**18.18032**	2
SOCS3	0.00005	0.00780	4.12982	**17.50651**	28
GADD45B	0.00054	0.03440	3.56927	**11.87016**	1
THBD	0.00000	0.00100	3.48248	**11.17714**	0
ADAMTS4	0.00019	0.01838	3.40269	**10.57580**	1
PDE4B	0.00000	0.00005	3.33522	**10.09258**	0
FOS	0.00031	0.02436	3.31416	**9.94630**	18
EGR1	0.00002	0.00362	3.11271	**8.65008**	3
JUNB	0.00004	0.00743	3.08829	**8.50488**	2
RGS16	0.00044	0.03038	2.96393	**7.80246**	1
ZFP36	0.00012	0.01425	2.92543	**7.59700**	2
MYC	0.00026	0.02179	2.86543	**7.28754**	23
CISH	0.00000	0.00049	2.78449	**6.88992**	0
CCL2	0.00013	0.01462	2.59339	**6.03513**	111
CXCL2	0.00006	0.00858	2.35828	**5.12758**	2
ATF3	0.00202	0.07254	2.31257	**4.96766**	9
SERPINA3	0.00732	0.14867	2.16589	**4.48742**	0
NFIL3	0.00002	0.00434	2.15060	**4.44013**	0
GADD45A	0.00382	0.10196	2.14953	**4.43682**	0
IL6	2.0786	0.00044	0.03051	**4.22398**	52

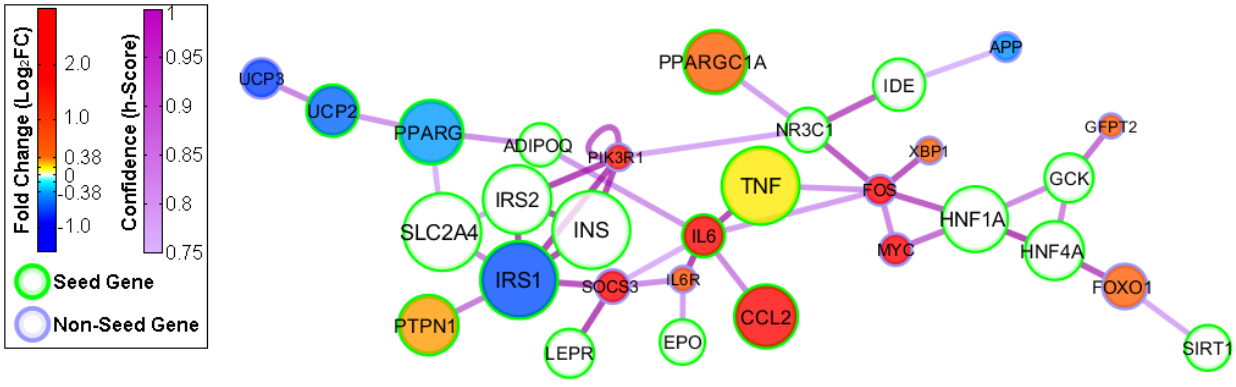

Figure 2: Top-20 significant genes in IAB from GSE24215, interacted with T2D-associated genes. Node size represents Evidence for each gene, node color represents Log2_FC, red color implies over-expressed and blue color implies under-expressed. Green circled nodes are seed genes (T2D-associated genes curated from OMIM). Edge color represents Confidence (h-Score) for each interaction.

Table 5: Top-20 significant genes in IAB from GSE24215, ordered by Sig_Score, which is measured in the T2D-specific PPI network (PubMed >=0, h-Score >=0.45) for all the differential genes (FC>=1.3, p-value <=0.05 and AEL >=40% after applying Limma package in Bioconductor) in IAB from GSE24215. Note: Gene names are linked to GeneCards.org, UniProt IDs are linked to UniProt.org, and Evidences are linked to PubMed.

Gene Symbol	*p*-Value	FDR	Log2_FC	ABS_FC	Weight_1	Weight_2	Sig_Score	Evidences
CCL2	0.00013	0.01462	2.59339	6.03513	98	75.2645	**34.9751**	111
IL6	0.00044	0.03051	2.0786	4.22398	140	112.808	**34.68919**	52
IRS1	0.00902	0.16494	-0.68912	1.6123	103	83.1045	**23.58864**	280
IL6R	0.00002	0.00404	0.79678	1.73722	70	55.9728	**22.14359**	7
VEGFA	0.00017	0.01724	0.92831	1.90304	57	44.6818	**21.6589**	28
APP	0.03329	0.31458	-0.43503	1.35194	80	68.3971	**21.01141**	15
SOCS3	0.00005	0.0078	4.12982	17.50651	7	6.3574	**20.52815**	28
ADRB2	0.00016	0.0164	1.0105	2.01461	37	31.2254	**20.09311**	10
FOXO1	0.00008	0.01056	0.52622	1.44015	65	44.529	**19.42493**	59
MYC	0.00026	0.02179	2.86543	7.28754	10	8.8236	**19.33394**	23
SOD2	0.00164	0.06478	0.82246	1.76841	50	29.5486	**18.85629**	4
DGKD	0.02661	0.28238	0.5992	1.51487	42	34.932	**18.59775**	2
FOS	0.00031	0.02436	3.31416	9.9463	7	6.3552	**18.17698**	18
PIK3R1	0	0.00008	1.67218	3.18695	13	11.8228	**17.1966**	9
XBP1	0.00338	0.09606	0.40352	1.32273	42	26.9454	**16.35234**	9
AGT	0.00295	0.08852	0.59516	1.51064	24	18.032	**15.28071**	42
UCP3	0.00163	0.06449	-0.75312	1.68543	22	15.2626	**15.10061**	45
UCP2	0.00186	0.06897	-0.60181	1.51762	22	15.7096	**14.63273**	78
PPARGC1A	0.00479	0.117	0.53018	1.44411	17	13.6	**13.65437**	111
GFPT2	0.04473	0.36274	0.54879	1.46286	14	12.2852	**13.2432**	2

3.2.3 Top-20 significant pathways

Totally 64 significant pathways (*p**-norm >= 1.2) in IAB from GSE24215 are obtained from all the differential pathways, measured by using pathway differential expressions (*p**-norm). Top-20 significant pathways in IAB from GSE24215, ordered by pathway differential expressions (*p**-norm), are listed in Table 6.

4 Discussion

The key finding of the present study on GSE24215 was that bed rest was associated with a paradoxically increased response to insulin of genes involved in acute-phase response and inflammation, including IL-6 signaling, IL-10 signaling, and the ER stress pathway, contrasting the development of severe peripheral insulin resistance of glucose metabolism in young healthy men. The present study demonstrated that 9 days of bed rest induces severe transcriptional changes of genes potentially involved in the pathogenesis of insulin resistance and T2D in skeletal muscle, which might to some extent explain the harmful effect of a sedentary lifestyle on human metabolism. Impaired expression of HK2, VEGFA, NDUFB6, PPARGC1A, and OXPHOS genes in general, as well as a markedly increased expression of

RRAD, are among the prime candidates contributing to the development of insulin resistance during bed rest.

Table 6: Top-20 significant pathways in IAB from GSE24215, ordered by pathway differential expressions (p*-norm), which is measured with all the available differential gene expressions in IAB from GSE24215.

PATHWAY_NAME	DB_SOURCE_ID	NORM_ABS_FC	MAX_ABS_FC
IL-9 Pathway	KEGG	6.70137	10.92515
IL-10 Pathway	KEGG	6.43693	10.92515
IL6-mediated signaling events	NCI-Nature Curated	6.34297	10.92515
EPO signaling pathway	NCI-Nature Curated	6.13163	10.92515
IL23-mediated signaling events	NCI-Nature Curated	6.01717	10.92515
igf-1 signaling pathway	BioCarta	5.98831	9.94630
Type II diabetes mellitus	KEGG	5.90782	10.92515
Signaling events mediated by PTP1B	NCI-Nature Curated	5.83706	10.92515
IL-4 Pathway	KEGG	5.71129	10.92515
signal transduction through il1r	BioCarta	5.68010	9.94630
Growth Hormone Signaling	KEGG	5.67342	10.92515
Calcineurin-regulated NFAT-dependent transcription in lymphocytes	NCI-Nature Curated	5.66004	9.94630
IL4-mediated signaling events	NCI-Nature Curated	5.53802	10.92515
FOXM1 transcription factor network	NCI-Nature Curated	5.44978	9.94630
Adipocytokine signaling pathway	KEGG	5.40776	10.92515
GDNF-Family Ligands and Receptor Interactions	KEGG	5.18210	9.94630
HIF-1-alpha transcription factor network	NCI-Nature Curated	4.99991	9.94630
Regulation of nuclear SMAD2/3 signaling	NCI-Nature Curated	4.91493	9.94630
Insulin signaling pathway	KEGG	4.89924	10.92515
Jak-STAT signaling pathway	KEGG	4.80603	10.92515

Our analysis on this microarray dataset also shows that Insulin-stimulation After Bed-rest (IAB) is associated with the same significant genes: VEGFA (Rank: 5), PPARGC1A (Rank: 19), HK2 (Rank: 23), and RRAD (Rank: 29). We also found IAB is associated the same/similar pathways: IL-10 Pathway (Rank: 2) from KEGG database, IL6-mediated signaling events (Rank: 3) from NCI-Nature Curated pathway database, igf-1 signaling pathway (Rank: 6)from BioCarta database, Type II diabetes mellitus (Rank: 7) from KEGG database, Growth Hormone Signaling (Rank: 11) from KEGG database, Insulin signaling pathway (Rank: 19) from KEGG database, Jak-STAT signaling pathway (Rank: 20) from KEGG database, il 6 signaling pathway (Rank: 27) from BioCarta database, and role of erbb2 in signal transduction and oncology (Rank 31) from BioCarta database.

5 Conclusions

In this paper, we apply both classical microarray analysis (such as differential analysis in Bioconductor) and our knowledge-guided analysis (network expansion analysis and pathway

enrichment analysis). From the evidence from literature (PubMed), Top 20 significant genes from our analysis have more supports than Top 20 differential genes from simple differential analysis, in the case study on the microarray dataset - GSE24215. This implies the vitality of our hypothesis on which organized knowledge will help microarray data analysis in significant ways.

For GSE24215 dataset, both of the two networks (shown in Figure 1 and Figure 2) consist of two subnetworks. The bigger one includes genes that are highly related to diabetic type 2. Some of genes are shared between before bed network and after bed network, like insulin receptor, peroxisome proliferator-activated receptor gamma and so on. For those shared genes, their expressions are different between these two conditions. We can see from the figure that IRS1, PPARG are under-expressed in after bed condition while IRS2 are under-expressed in before bed condition. Beside those shared molecules, some only show in after bed condition, like PPARGC1A, IDE, IL6R, APP, and PTPN1 while others only show in before bed condition, like CD36, SCARB1, SELP, ICAM1, TNFRSF1A, HSD11B1, and AKT2. The smaller one is relatively small sub-network. Commonly shared gene between before bed and after bed are HNF1A, HNF4A and MYC with similar expression level. Some genes like TCF7L2 only show up in before bed network while GCK, GFPT2, FOXO1, and SIRT1 only show in after bed network.

Another interesting finding is that the molecules which connect the red sub-network and blue-subnetwork are different. In before bed network, SMAD3 play this important role while in after bed network it is FOS that connect these two subnetworks. In fact, FOS and SMAD3 are physically interacting with each other and together Smad3 cooperates with c-Jun/c-Fos to mediate TGF-beta-induced transcription. Finally though IGF1 doesn't show up in the network, yet the IGF1 pathway is highly ranked (refer to pathway analysis part) in the after bed condition.

Acknowledgements

We would like to thank Eli Lilly and Company, and MedeoLinx, LLC for their great financial support.

References

[1] D. B. Allison, X. Cui, G. P. Page, and M. Sabripour. Microarray data analysis: from disarray to consolidation and consensus. *Nature Reviews Genetics*, 7:55–65, 2006.

[2] M. Reimers. Making Informed Choices about Microarray Data Analysis. *PLoS Computational Biology*, 6:e1000786, 2010.

[3] D. K. Slonim and I. Yanai. Getting started in gene expression microarray analysis. *PLoS Computational Biology*, 5:e1000543, 2009.

[4] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, and E. S. Lander. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102:15545–15550, 2005.

[5] D. Glez-Pena, G. Gomez-Lopez, D. G. Pisano, and F. Fdez-Riverola. WhichGenes: a web-based tool for gathering, building, storing and exporting gene sets with application in gene set enrichment analysis. *Nucleic Acids Research*, 37:W329–W334, 2009.

[6] D. W. Huang, B. T. Sherman, and R. A. Lempicki. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Research*, 37:1–13, 2009.

[7] A. L. Barabasi and Z. N. Oltvai. Network biology: understanding the cell's functional organization. *Nature Reviews Genetics*, 5:101–113, 2004.

[8] P. Goymer. Cancer genetics: Networks uncover new cancer susceptibility suspect. *Nature Reviews Genetics*, 8:823–823, 2007.

[9] M. A. Pujana, J. D. J. Han, L. M. Starita, K. N. Stevens, M. Tewari, J. S. Ahn, G. Rennert, V. Moreno, T. Kirchhoff, and B. Gold. Network modeling links breast cancer susceptibility and centrosome dysfunction. *Nature Genetics*, 39:1338–1349, 2007.

[10] H. Y. Chuang, E. Lee, Y. T. Liu, D. Lee, and T. Ideker. Network-based classification of breast cancer metastasis. *Molecular Systems Biology*, 3:140–149, 2007.

[11] N. McCarthy. Tumour profiling: Networking, protein style. *Nature Reviews Cancer*, 7:892–893, 2007.

[12] J. Y. Chen, C. Shen, and A. Y. Sivachenko. Mining Alzheimer disease relevant proteins from integrated protein interactome data. *In* Pacific Symposium on Biocomputing, pages 367–378, 2006.

[13] J. Y. Chen, S. Mamidipalli, and T. Huan. HAPPI: an online database of comprehensive human annotated and predicted protein interactions. *BMC Genomics*, 10(Suppl 1):S16, 2009.

[14] S. R. Chowbina, X. Wu, F. Zhang, P. M. Li, R. Pandey, H. N. Kasamsetty, and J. Y. Chen. HPD: an online integrated human pathway database enabling systems biology studies. *BMC Bioinformatics*, 10(Suppl 11):S5, 2009.

[15] A. C. Alibegovic, M. P. Sonne, L. Højbjerre, J. Bork-Jensen, S. Jacobsen, E. Nilsson, K. Færch, N. Hiscock, B. Mortensen, and M. Friedrichsen. Insulin resistance induced by physical inactivity is associated with multiple transcriptional changes in skeletal muscle in young men. *American Journal of Physiology-Endocrinology And Metabolism*, 299:E752–E763, 2010.

[16] G. Smyth. Limma: linear models for microarray data. *In* Bioinformatics and computational biology solutions using R and Bioconductor, pages 397–420, 2005.