INTEGRATIVE SYSTEM BIOLOGY STUDIES ON HIGH THROUGHPUT GENOMICS AND PROTEOMICS DATASET

Madhankumar Sonachalam

Submitted to the faculty of the Bioinformatics Graduate Program

in partial fulfillment of the requirements for the degree

Master of Science in Bioinformatics

in the School of Informatics

Indiana University

May 2012

Accepted by the Faculty of Indiana University, in partial fulfillment of the requirements for the degree of Master of Science in Bioinformatics

Dr. Jake Yue Chen, Ph.D., Chairperson

Dr. Li Shen, Ph.D

Master's Thesis

Committee

Dr. Yaoqi Zhou, Ph.D

Dedicated to

Amma & Appa

Acknowledgements

This thesis would not exist if not for the support and guidance of many of my friends and family. To each of them I would like to express my gratitude. Furthermore, I would like to explicitly mention a select few, whose help has been invaluable to me.

I owe my most sincere gratitude to my research advisor, Dr. Jake Yue Chen, for his motivation, enthusiasm and immense knowledge. His energy, vision, and great efforts had been a constant source of inspiration. I am very grateful to him for the constant support and for playing a major role in honing my writing and presentation skills.

Besides my advisor, I would like to thank the rest of my thesis committee Dr.Shen Li and Dr. Yaoqi Zhou for their encouragement and insightful comments. I would also like to extend my thanks to our collaborators Dr. Tim Ratliff and Dr. Shen Li for providing us an excellent dataset to explore various bioinformatics techniques.

I am grateful to Dr.Xiaogang Wu for his friendly help, constructive criticism, and excellent advice during the preparation of this thesis.

It gives me great pleasure in acknowledging the members of my research group, Ragini Pandey, Dr. Fan, Huang Hui for providing insight and encouragement throughout my research work.

My thanks are due to Stephanie Burks and Teresa Hunter of Research and Technical Services and Kimberly Melluck of school of informatics for their timely help on computing resources.

Finally, I would like to thank my parents, brother, sister and uncle for supporting me always, for having faith in me. Without their encouragement and understanding it would have been impossible for me to finish this work.

Chapter 2 in part, quotes from the materials published in Madhankumar Sonachalam, Jeffrey Shen, Hui Huang, Xiaogang Wu, *Systems biology approach to identify gene network signatures for colorectal cancer*, Frontiers in System Biology, Accepted. The dissertation author was the first author on this work, responsible for design and data analysis. Xiaogang Wu, Hui Huang, Madhankumar Sonachalam, Sina Reinhard, Jeffrey Shen, Jake Y. Chen, *Reordering Based Integrative Expression Profiling for Microarray Classification*, BMC Bioinformatics, Accepted. The dissertation author was the second author on this work, responsible for Integrative Expression Profiling model building and microarray classification.

Chapter 3 in full, a re-editing of the materials submitted for the publication. Xiaogang Wu, Madhankumar Sonachalam, Sungeun Kim, Andrew J Saykin, Li Shen, Jake Y Chen, and Alzheimer's Disease Neuroimaging Initiative, *Identifying Plasma-Based Subnetwork Signatures for Alzheimer's disease using a Multiplex Proteomic Immunoassay Panel in Alzheimer's Disease Neuroimaging Initiative cohort*. The dissertation author was the first author on this work, responsible for design, data analysis and implementation.

ABSTRACT OF THE DISSERTION

Madhankumar Sonachalam

INTEGRATIVE SYSTEM BIOLOGY STUDIES ON HIGHTHROUGHPUT GENOMIC AND PROTEOMIC DATASET FOR BIOLOGICAL PATHWAY DISCOVERY

The post genomic era has propelled us to the view that the biological systems are complex network of interacting genes, proteins and small molecules that give rise to biological form and function. The past decade has seen the advent of number of new technologies designed to study the biological systems on a genome wide scale. These new technologies offers an insight in to the activity of thousands of genes and proteins in cell thereby changed the conventional reductionist view of the systems. However the deluge of data surpasses the analytical and critical abilities of the researches and thereby demands the development of new computational methods.

Gene expression microarrays can take a snapshot of all the transcriptional activity in a biological sample, while it also generates a huge amount of data with intrinsic noise (sample or instrument noise), which is still a quite challenging task to interpret it even by exploiting modern computational and statistical tools. The challenge no longer lies in the acquisition of gene expression profiles, but rather in the interpretation for the results to gain insights into biological mechanisms. Gene Set Enrichment Analysis (GSEA) is one of the widely used Gene Set Analysis (GSA) methods that aim to test the activity of gene clusters rather than individual genes.

In Chapter 2, we integrated prior knowledge from gene signatures (curated gene sets from MSigDB and/or GeneSigDB); gene set enrichment analysis (GSEA), and gene/protein network modeling to identify gene network signatures from microarray data. We demonstrated how to apply this approach in discovering gene network signatures for colorectal cancer (CRC) from

microarray datasets. We compared the network generated from two different gene set sources and showed that the integrated network generated from both MSigDB and GeneSigDB can be used to identify novel pathways involved in colorectal cancer.

In Chapter 3, we identified plasma based Subnetwork signatures for Alzheimer's disease (AD) using proteomics dataset. Current plasma based AD signatures uses feature selection methods that was originally designed for microarray analysis. We evaluated various feature selection and classification approaches to select the best set of features for a specific proteomics dataset. Our combination of feature selection and classification techniques showed better performance than the existing results and we further provided biological validation by identifying relevant Subnetwork signatures.

Finally, we applied the network based strategy to identify the role of MicroRNA in Myeloid Derived Suppressor Cells (MDSC) induced T-Cell Suppression. In Chapter 4, we used network based ranking algorithm to prioritize miRNA and genes by utilizing both network topology and differential information. We showed that the global and local topology characteristics of the miRNA-gene-gene/protein network along with the differential expression values obtained from microarray experiment can be used to identify biologically significant pathways.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. INTRODUCTION

1.1 Motivation

The term omics refers to the comprehensive analysis of biological systems. A variety of omics disciplines have begun to emerge after the Human Genome Project. Genomics deals with the systematic use of genomic information and it includes investigations about the function of the genes. Transcriptomics examines the expression level of mRNAs of the genes in a given cell population. Proteomics focus on the large scale study of proteins while Metabolomics addresses the metabolites involved in the cellular process.

System biology is the study of biology through the systematic perturbation, global read out of the multifaceted response through various omics studies and integration of those data to formulate predictive models. System biology investigates the behavior and relationships of all the elements in a biological system rather than focusing on a single gene or protein. As a result, inputs from various disciplines such as statistics, computer science and mathematics are necessary to a "systems" approach of analyzing data. We employed system biology techniques to solve complex biological problems using genomics, proteomics and integrated MicroRNA and gene expression dataset.

1.2 Messages

The sheer volume of data generated in the post genomic era surpasses the analytical and critical abilities of a single researcher and demands the new computational methods to assist in the analysis of these data. In these projects, we devised new methodologies, integrated existing machine learning and network biology techniques to solve various complex biological problems. We adapted different strategies depending on the data sets that we are handling; genomics data sets are filtered using gene set enrichment techniques and integrated it with network biology to explore biological mechanism of colorectal cancer. While for proteomics data, we evaluated various feature selection and classification algorithms to select best features which were then used to identify sub-network signatures. Finally, we combined the miRNA and gene transcriptomics data to generate an integrated network which explains the molecular mechanism of T-cell suppression in cancer.

1.3 Microarrays

Gene expression microarrays can take a snapshot of all the transcriptional activity in a biological sample, while it also generates a huge amount of data with intrinsic noise (sample or instrument noise), which is still a quite challenging task to interpret it even by exploiting modern computational and statistical tools. These high-throughput genomics technologies have tremendously changed biomedical research, which allow researchers to simultaneously monitor the expression of tens of thousands of genes [1]. Microarray data analysis has also become a common practice in many experimental laboratories. Numerous literatures describe the innovative insights within microarray data analysis [2, 3]. It has been widely applied into many medical areas, including distinguishing disease subtypes [4], identifying candidate biomarkers[5], and revealing the underlying molecular mechanisms of disease [6] or drug response [7]. There are several repositories such as Gene Expression Omnibus (GEO) and Array Express that hosts Microarray dataset from various experiments.

2. SYSTEMS BIOLOGY APPROACH TO IDENTIFY GENE NETWORK SIGNATURES FOR COLORECTAL CANCER

2.1 Background

In microarray analysis, crucial genes show relatively slight changes, and many genes selected are also poorly annotated [2]. From a biological perspective, functionally related genes often display a coordinated expression to accomplish their roles in the cell [8]. Hence, to translate such lists of differentially expressed genes into a functional profile will help us 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 (gene cluster) [2].

Accordingly, Gene Set Analysis (GSA) methods aim to test the activity of such gene clusters instead of testing the activity of individual genes - individual gene analysis (IGA) [9]. In recent years, GSA approach has received a great deal of attention, since it is free from the problems of the ‘cutoff-based’ methods. In this direction, GSA methods enable the understanding of cellular processes as an intricate network of functionally related components [8].Among these GSA methods, gene set enrichment analysis (GSEA) is one of the most widely used methods [10]. GSEA analyzes pre-defined gene sets based on prior biological knowledge to determine whether this gene set as a whole exhibits differential expression. GSEA has many advantages as it does not employ an arbitrary cutoff to select significant genes. Instead, it uses all the information about every gene involved in the experiment. However, GSEA does rely on pre-defined gene sets (without gene interaction information); making IGA more beneficial when not much is known about the biological function being considered. Furthermore, GSEA still assumes that more differentially expressed genes are more crucial to the biology, which is not always true [11]. In many cases, extensive upstream data processing, comprehensive gene selection statistics, and downstream pathway/network analysis cannot be replaced by GSEA. Therefore, gene expression signature analysis and pathway analysis (using tools such as DAVID [12]) remain two separate processes.

Network-based gene expression analysis is proposed for candidate biomarker discovery by integrating disease susceptibility genes, their gene expressions, and their gene/protein interaction network [13, 14]. In 2007, Marc Vidal's group at Harvard constructed a protein interaction network for breast cancer susceptibility using various 'omics' data sets, and identified HMMR as a new susceptibility locus for the disease[13]. Later, Trey Ideker's group at UCSD integrated protein network and gene expression data to improve the prediction of metastasis formation in patients with breast cancer [14]. The two studies marked the exciting beginning of a new paradigm which suggests networks and pathways, although drafty, error-prone, and incomplete, can serve as a molecular-level conceptual roadmap to guide future microarray analysis.

Recent advances in genomics, transcriptomics, proteomics, epigenomics, and metabolomics have begun to help discover DNA/RNA-based prognostic and predictive markers for early and advanced colorectal cancer (CRC) [15]. Systems biology results show that cancer genes and proteins do not function in isolation; instead, they work in interconnected pathways and molecular networks [16]. However, systematically building disease-specific network models, integrated at multiple Omics level - transcriptome (RNA-based markers from microarray data) and proteome (protein-based markers from network and pathway data), has not yet been done in CRC biomarker discovery.

In this work, we integrated prior knowledge from GWAS studies or gene signatures (curated gene sets from MSigDB and/or GeneSigDB), gene set enrichment analysis (GSEA), and gene/protein network modeling together to identify gene network signatures from microarray data. We demonstrated how to apply this approach into discovering gene network signatures for colorectal cancer (CRC) from microarray datasets at three levels - genome, transcriptome, and proteome. First, we use GSEA to analyze the microarray data through enriching differential genes in different CRC-related gene sets from two publicly-available up-to-date gene set databases - Molecular Signatures Database (MSigDB) and Gene Signatures Database (GeneSigDB). Second, we compare the enriched gene sets through enrichment score (ES), false-discovery rate (FDR) and nominal *p*-value. Third, we construct an integrated protein-protein interaction (PPI) network through connecting these enriched genes by using a human annotated and predicted protein interaction (HAPPI) database, with a confidence score (CS) labeled for each interaction. Finally, we map differential expression values onto the constructed network to build a comprehensive network model containing visualized genome, transcriptome, and proteome data. The results show that although MSigDB is more suitable for GSEA analysis than GeneSigDB, the integrated PPI network connecting the enriched genes from both MSigDB and GeneSigDB can provide more complete view for discovering gene signatures. We also find several important sub-network signatures for colorectal cancer, such as TP53 sub-network, PCNA sub-network and IL8 sub-network, corresponding to apoptosis, DNA repair, and immune response respectively.

2.2 Methods

2.2.1 Microarray data

From GEO (http://www.ncbi.nlm.nih.gov/geo/), we download a CRC-related microarray dataset - GSE8671, which compared the transcriptomes of 32 prospectively collected adenomas with those of the normal mucosa from the same individuals. Hence we have 32 CRC samples and 32 normal samples. We use maximal expression values for same proteins mapped from different Probe IDs. We use Affy package in BioConductor for quantile normalization. For background correction, we use the built-in MicroArray Suite (MAS5).

2.2.2 Gene sets

Gene sets are obtained from MSigDB and GeneSigDB. MsigDB has almost 6769 gene sets and are divided in to five major collections, of which “C2” are curated gene sets collected from various sources such as online pathway databases, publications in pubmed, and knowledge of domain experts. We searched in that collection with keyword “colon” and obtained 73 gene sets. GeneSigDB is a manually curated database of gene expression signatures. And it shares minimum overlap between MSigDB C2 Category of around 8%. It provides the standardized gene list for different search criteria. Searching as “Colon” had retrieved 36 gene sets.

2.2.3 Gene set enrichment analysis

Though there are many variations on the GSEA method, we describe the version of the algorithm developed by Subramanian and colleagues [10], which will be called the standard implementation of the method, since it is the most widely used form of the GSEA method.

Suppose that a microarray dataset is obtained from two different phenotypes, phenotype 1 and phenotype 2 (e.g. control vs. experimental). This microarray dataset has expression values for the genes across the samples and each row has been identified by unique probe identification. Consider also a given gene set *S*, usually derived from some common biological category. The objective of the GSEA method is to see if the gene set *S* shows differential expression between the two phenotypes.

The Broad Institute provides an easy to use standalone Java implementation of the GSEA method on their website (http://www.broadinstitute.org/gsea/). All gene sets with more than 500 genes or less than 15 genes were automatically excluded. The difference between signal-to-noise ratios was used as the association score. The number of phenotype permutations involved in the nominal p-value calculation was 1000. For each analysis, we report the number of gene sets with FDR<25%. Along with these gene sets with FDR<25%, we report the number of gene sets whose nominal *p*-values are <1% or 5%.

2.2.4 Network modeling

To optimize computation time and information generation, we use a combined network construction strategy, based on the enriched genes from both MSigDB and GeneSigDB.

First, we connect the enriched MSigDB genes from GSE8671 in HAPPI (http://bio.informatics.iupui.edu/HAPPI) with confidence score (*CI* >=0.75, i.e. 4-star rating) for interactions, to obtain a protein-protein interaction (PPI) network. The local topological property (e.g. node degree, cluster coefficient, betweenness centrality, neighborhood connectivity etc. [17]) for each node is calculated based on this network. Then genes with absolute fold change $|FC|>=1.5$, equals to Log $_2(FC)>=0.585$, are kept.

Second, we connect the enriched GeneSigDB genes from GSE8671 in HAPPI (http://bio.informatics.iupui.edu/HAPPI) with confidence score (*CI* >=0.75, i.e. 4-star rating) for interactions, to obtain another protein-protein interaction (PPI) network. In the same way, the local topological property for each node is calculated based on this network. Then genes with absolute fold change $|FC|>=1.5$, equals to Log $_2(FC)>=0.585$, are kept.

Finally, we combine these two networks to build a node-weighted edge-scored CRC-specific PPI network model by using Cytoscape [18], with node color representing the fold change for each gene, node size representing the local topological property for each gene/protein, edge color and edge width representing confidence score for each protein interaction.

2.3 Results

2.3.1 Enriched gene sets

We run the GSEA analysis for the gene expression microarray data - GSE8671 with gene sets obtained from MSigDB and GeneSigDB separately. We use the default values in GSEA which filtered out 22 gene sets from MSigDB as the number of genes in those sets falls below the threshold value of 15 in GSEA. So we run the GSEA analysis on remaining 51 gene sets. Of those 51, 22 gene sets are up-regulated in normal and remaining 29 are up-regulated in cancer samples. Summary of the results are shown in Table 1.

There are 22 gene sets that are significantly enriched in normal and 29 in Colorectal cancer, of which the gene set - GRADE_COLON_CANCER_DN tops the list with enrichment score of 0.79 in Normal vs. Cancer, and the gene set - SANA_RESPONSE_TO_IFNG_DN tops the list in Cancer vs. Normal with the enrichment score of -0.67.

Table 1: Summary of CRC Gene Set Results enriched in MSigDB

Enrichment	**Normal vs. Cancer**	**Cancer vs. Normal**
Up-regulated	22 gene sets	29 gene sets
Significant at FDR < 25%	8 gene sets	14 gene sets
Nominal p-value for S from ES_{NULL} < 5%	7 gene sets	12 gene sets
Nominal p-value for S from ES_{NULL} < 1%	5 gene sets	6 gene set

As with the case of MSigDB, GSEA had filtered only 22 gene sets out of 34 based on the default filter criteria. Of these 22, 11 gene sets are enriched in normal and remaining 11 on cancer. Summary of the results are shown in table 2. Of the enriched gene sets - 16091735-TABLE1 tops the list in Normal vs. Cancer with the enrichment score of 0.52 and the gene set - 11906190-TABLE2B-2 tops the list with enrichment score of -0.53 in Cancer vs. Normal.

Table 2: Summary of CRC Gene Set Results enriched in GeneSigDB

Enrichment	**Normal vs. Cancer**	**Cancer vs. Normal**
Up-regulated	11 gene sets	11 gene sets
Significant at FDR < 25%	7 gene sets	8 gene sets
Nominal p-value for S from ES_{NULL} < 5%	4 gene sets	5 gene sets
Nominal p-value for S from ES_{NULL} < 1%	1 gene sets	2 gene set

2.3.2 A PPI network based on enriched genes from MSigDB

We construct a PPI network (325 genes and 686 interactions) with *CI* >=0.75 based on the 694 enriched genes (mapped to 678 proteins) from MSigDB, and visualize the network layout by using spring embedded network layout in Cytoscape 2.8.1. After filtering out genes with |*FC*|<1.5, there are 244 genes and 422 interactions. We also map the differential expression values onto the genes in the network by representing them as node colors. Since we also simply represent node degree as node size, we can easily access the relationship between differential expression value and topological property for each gene in the network. As shown in Figure 1, the gene sets from MSigDB connected very well. Most important cancer genes, such as TP53 and PCNA, related to apoptosis and DNA repair are included. It indicates that MSigDB is suitable for GSEA analysis, unsurprisingly, since MSigDB is generated by the same group which developed GSEA.

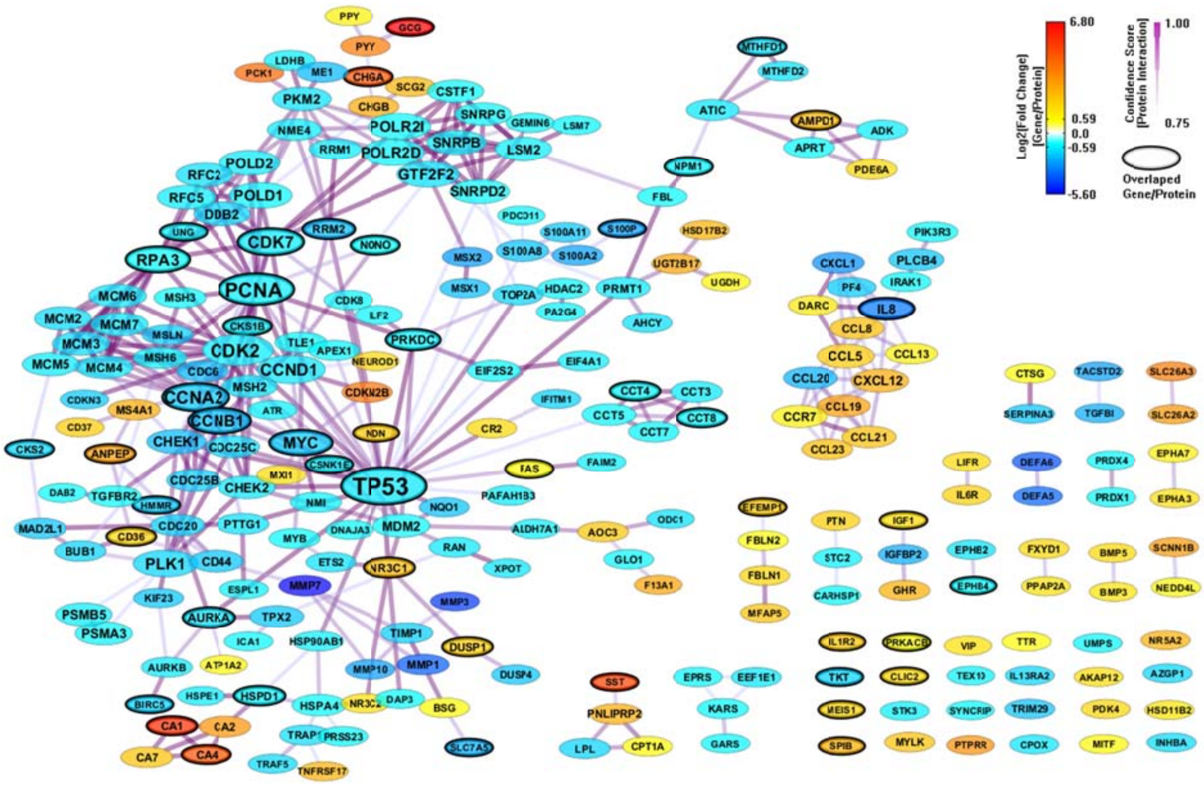


Figure 1: PPI Network based on the Enriched Genes from MSigDB *by through analyzing GSE8671 by GSEA. Black-circled genes represent the enriched genes from both MSigDB and GeneSigDB*

2.3. 3 A PPI network based on enriched genes from GeneSigDB

We also construct a PPI network (112 genes and 169 interactions) with *CI* >=0.75 based on the 303 enriched genes (mapped to 301 proteins) from GeneSigDB, and visualize the network layout by using spring embedded network layout in Cytoscape 2.8.1. After filtering out genes with |*FC*|<1.5, there are only 68 genes and 62 interactions (shown in Figure 2). Although the gene sets from GeneSigDB are directly from gene expression profile (most of them are microarray data) analysis, the scale of the PPI network built on the enriched genes from GeneSigDB is really poor. It implies that GeneSigDB may not be applicable for GSEA analysis, at least, cannot be used singly.

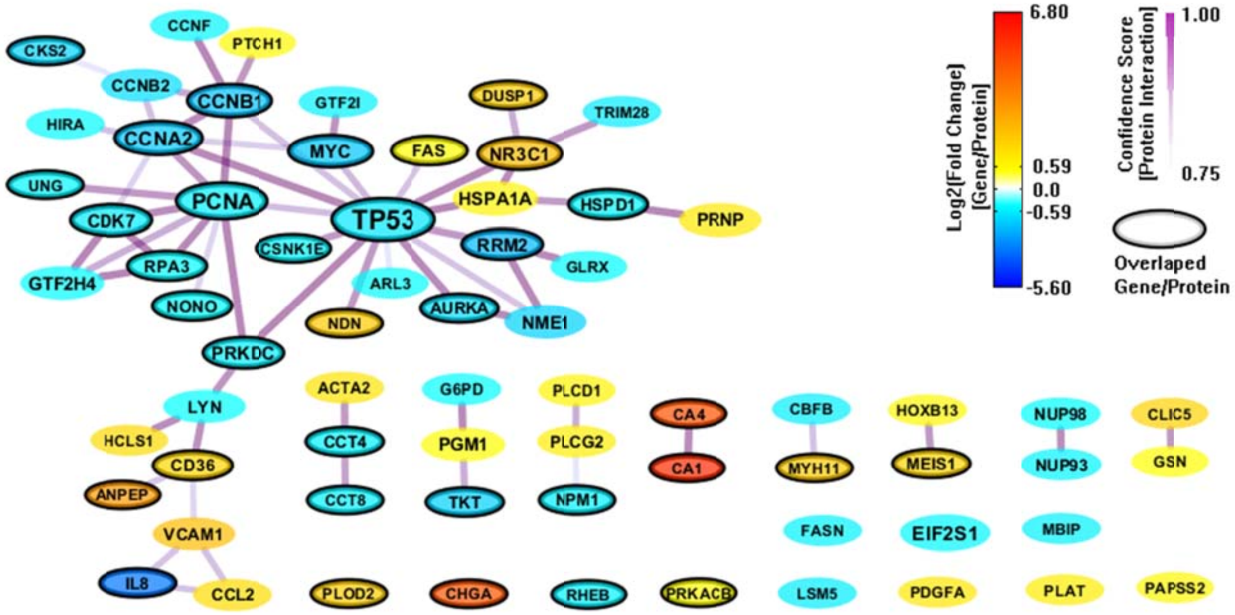


Figure 2: PPI Network based on the Enriched Genes from GeneSigDB *through analyzing GSE8671 by GSEA. Black-circled genes represent the enriched genes from both MSigDB and GeneSigDB*

2.3.4 Integrated CRC specific network signature

There are only 85 genes (mapped to 84 proteins) overlapped between the 694 enriched genes from MSigDB and the 303 enriched genes from GeneSigDB. So we combine the two PPI network together to build an integrated network signature specific for colorectal cancer. We construct a PPI network (443 genes and 1070 interactions) with CI >=0.75 based on the 895 enriched genes from both MSigDB and GeneSigDB. After filtering out genes with |FC|<1.5, there are 311 genes and 541 interactions (shown in Figure 3). As we can see, the integrated network has more genes/proteins connected, especial for the gene sub-network surrounding IL8, which relates to inflammation and immune response. This gene has been recognized playing an important role in regulates various aspects of immune response, cell death, and differentiation as well as cancer [19].

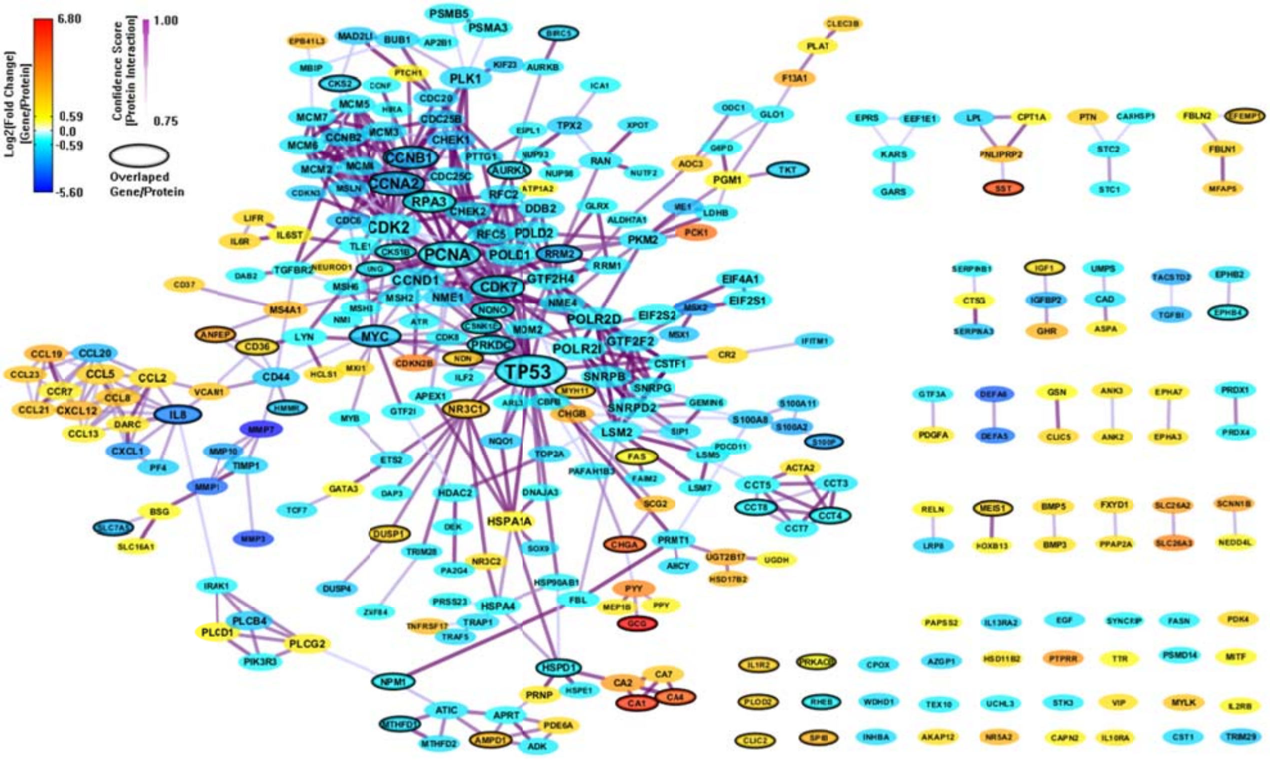


Figure 3: An integrated CRC-specific Network Signature *based on the Enriched Genes from both MSigDB and GeneSigDB through analyzing GSE8671 by GSEA. Black-circled genes represent the enriched genes from both MSigDB and GeneSigDB.*

2.4 Discussion

In this work, we integrated gene signatures; gene set enrichment analysis (GSEA), and protein-protein interaction network to identify gene network signatures from microarray data. We demonstrated how to apply this approach in discovering gene network signatures for colorectal cancer from microarray datasets at three levels - genome, transcriptome, and proteome. The results show that MSigDB is more suitable for GSEA analysis than GeneSigDB.

Although MSigDB may be more suitable for GSEA analysis than GeneSigDB (the two databases are both updated very quickly), the integrated PPI network connecting the enriched genes from both MSigDB and GeneSigDB can provide more complete view for discovering gene signatures. We also find several important sub-network signatures for colorectal cancer, such as

TP53 sub-network, PCNA sub-network and IL8 sub-network, corresponding to apoptosis, DNA repair, and immune response respectively [19-21].

However, gene-to-gene or gene-to-protein interaction may be even more accurately represented by a network. One limitation of our restrictive approach and of the GSEA method in general, is that it is not able to generate new hypotheses for unsuspected gene sets. This has proved to be a major limitation of the GSEA method in general, especially since one of the main goals of gene expression microarray analysis is to find new sets of relevant genes. Another disadvantage of the GSEA method is that genes that are more differentially expressed are assumed to be more crucial. However, this assumption has not been thoroughly tested.

Currently, it is important to realize that no single method of gene expression microarray analysis works best, but rather information generated by the different analyses should be integrated together with the knowledge from biological research. In future work, we aim to combine GSEA, gene ontology (GO) enrichment, network expanding/enriching methods together to identify biologically significant genes/proteins. We will use more gene expression microarray datasets to validate this integrated strategy. We will also use newly generated gene expression profiles by using RNA-sequencing (RNA-seq) technique to test our new hypothesis.

3. IENTIFYING PLASMA-BASED SUBNETWORK SIGNATURES FOR ALZHEIMER'S DISEASE USING A MULTIPLEX PROTEOMIC IMMUNOASSAY PANEL

3.1 Background

Currently, the only way to confirm Alzheimer's disease (AD) definitively comes from autopsy, with the presence of characteristic lesions in the brain caused by extracellular plaques of Amyloid β (Aβ) peptide and intracellular neurofibrillary tangles (NFTs) formed by hyperphosphorylated Tau protein [22]. Intensive research has been conducted for discovering reliable AD biomarkers in peripheral blood. Although there are many publications on potential plasma-based AD biomarkers, follow-up studies by other research groups have often failed to show accurate, efficient and consistent diagnostic values [23]. There is an urgent need for benchmarks to be able to evaluate the performance of these biomarker panels/signatures. Moreover, it also remains unknown what relationships between proteins within each signature, and what relationships between these signatures are involved.

In 2007, Ray et al. [24] screened 120 proteins involved in cell communication, and found a 18-protein signature that can be used to classify blinded samples from Alzheimer's and control subjects with close to 90% accuracy. Their analysis was based on a shrunken centroid algorithm called predictive analysis of microarrays (PAM), with 83 archived plasma samples as training set and 92 separate samples as testing set (AD against control). Biological interpretation based on these 18 signaling proteins indicates systemic dysregulation of hematopoiesis, immune responses, apoptosis and neuronal support in presymptomatic AD. This pilot study has made a significant contribution for discovering diagnostically useful plasma-based AD biomarkers.

In 2008, using the same proteomic data, Gomez and Moscato [25] reported a 5-protein signature (which is a subset of the 18-protein signature) that achieves, on average, a 96% total accuracy in predicting clinical AD (80 for training and 92 for testing). This 5-protein signature (the abundance of IL-1α, IL-3, EGF, TNF-α and G-CSF) was chosen by using their spectacular feature selection approach based on Fayyad and Irani's entropy minimization algorithm, which was originally designed for microarray analysis [26]. The 5-protein signature demonstrated the same performance with the original 18-protein signature when using the same classifiers, such as Simple Logistic or Logistic Model Trees. The performance was verified by using over 20 different classifiers available in the widely-used Weka software package [27]. In 2011, by using methods from combinatorial optimization and information theory, Moscato's team reanalyzed the same proteomic data, and uncovered novel biomarkers, which confirms ANG-2, IL-11, PDGF-BB, CCL15/MIP-1δ; and supports the joint measurement of other signaling proteins not previously discussed: GM-CSF, NT-3, IGFBP-2 and VEGF-B [28].

Although the accuracy reported for these two plasma-based AD signatures are high enough, the consistency for their stable clinical application still remains unknown, especially when evaluating these signatures in new cohorts. In 2009, Soares et al. tested the reproducibility of another subset of the 18-protein signature by using quantitative multiplex proteomic immunoassay, which suggest diagnostic accuracy using this subset can only achieved 61%. By using multivariate analysis for feature selection and linear discriminant and random forest analysis for classification, an 89-protein signature was found, which can yield a diagnostic accuracy of 70%. This result suggests that the current plasma-based AD signatures may be useful as AD screening tools, but are still far from AD diagnostic purpose.

Another major concern on the view of bioinformatics is that the current plasma-based AD signatures are all selected by the feature selection approaches originally designed for microarray analysis. The performance of these approaches is doubtful when applying to proteomic studies (typically 120-250 protein analytes) with much less feature than in high-throughput transcriptomic studies (generally 20,000 genes or 50,000 probes). Feature selection is used as a preprocessing step before building models for classification. It aims to limit the amount and dimensionality of the data and thereby selecting significant features that correlate well with the target class [29-31]. Feature selection methods are often categorized as filters, wrappers, or embedded methods depending on how they combine the feature selection search with the construction of the classification model [32].

Filter methods evaluates each feature by looking only at the intrinsic properties of the data. Most methods works by assigning a score value for each feature and set a threshold as a criterion to evaluate performance. If the score values of a feature is greater than the threshold, then the feature will be selected otherwise it will be removed. Filter methods can rely on univariate and multivariate statistics [33]. Univariate methods such as chi square and Pearson correlation assumes each attribute as independent and assess the relevance of individual attributes for a specific class at a time [34]. In this kind of analysis, attributes that are not individually relevant but become significant in the context of other attributes will be missed out. Since univariate features selection methods are not able to capture feature interactions, it can result in redundant features which have high score with the class. Multivariate methods such as ReliefF overcome this constraint by considering feature interactions [33].

The wrapper approach uses search algorithms to select various subsets of features in the space of possible subsets, and evaluates the specific subset of features using a specific classification model [29, 30]. Wrapper methods consider the feature dependencies and since there is an interaction between the feature subset and the model selection, it helps to select best subset. Embedded methods also involve classification models, but unlike wrapper methods, the search for an optimal subset of features is built in to the classifier construction. Since it perform feature selection as a part of the classifier training process it takes computationally less time than wrapper[33]. SVM Attribute Evaluator and Elastic Net are some well-known embedded methods [35]. These methods are very effective in dealing with genomics and proteomics dataset in various bioinformatics approach. The main drawback with these methods is they do not integrate any external knowledge source to explain the biology behind the interactions.

In this work, we developed five criteria – accuracy, efficiency, consistency, significance and connectivity, to evaluate different feature selection approaches for identifying protein biomarkers under varied conditions. We compared 15 feature selection approaches and 20 classification methods in Weka [27] on a new dataset generated from the samples (AD vs. health control) collected by Alzheimer's Disease Neuroimaging Initiative (ADNI). We chose five feature selection approaches – ReliefF, SVM, OneR, Elastic Net and lasso for thorough evaluation on the presented five criteria, according to their overall performance on accuracy. Based on the top 32 protein analytes selected by each of these five feature selection approaches, we finally built a 64-protein network consisting of four subnetworks, which are defined as subnetwork signatures here. We also compared the protein analytes in the network with those analytes in the existing plasma-based AD signatures, and found these four plasma-based AD subnetwork signatures corresponding to G-Protein coupled receptor (GPCR) ligand pathway activation, complement, immune response, and apoptosis respectively. In these functions, GPCR ligand pathway activation is newly reported for plasma-based AD signatures, especially the important role of an analyte –follicle stimulating hormone (FSH), which bridges the functional subnetworks of hemostasis and apoptosis.

3.2 Methods

3.2.1 ADNI multiplex proteomic immunoassay data collection

We used the ADNI dataset providing expression values for 146 analytes in 108 Alzheimer disease patients and 53 healthy controls which we call as master dataset. We randomly permuted the master dataset to create 20 partitions with same number of Alzheimer and healthy samples. Then we approximately halved each partition in to training and testing set thereby maintaining equal samples for both class in training and testing set say 53-54 samples of Alzheimer and 26-27 samples of Healthy control.

3.2.2 Feature selection and Classification

Waikato Environment for knowledge analysis suite (Weka, version 3.6) was used for applying classification and feature selection methods to our datasets [27]. Weka is a java based tool that provides implementations for various machine learning algorithms. Weka has been used for various genomics and proteomics studies in Bioinformatics [36]. The default parameters set within Weka has been used for all the attribute selection and classification method. We are aware that the results can be optimized by tuning the parameters for classification algorithm, since we are interested in selecting possibly best features for downstream analysis rather than the classification accuracy we used only default parameters and moreover it ensures the reproducibility of results.

Weka provides implementation for various feature selection belongs to different categories such as Filter methods based on univariate statistics (CFSSubsetEval, ChiSquaredAttributeEval), based on multivariate statistics (ReliefF), meta-evaluators (CostSensitiveAttributeEval, CostSensitiveSubsetEval, FilteredAttribute and SubsetEval), Embedded methods (SVMAttributeEval) and other filter methods ConsistencySubsetEval, GainRatioAttributeEval, OneRAttributeEval and SymmetricalUncertAttributeEval. We used all the implementations in Weka version 3.6 except principal component analysis and latent semantic analysis which transforms the set of attributes. Hence, these methods are not widely used for constructing classification model. Elastic Net and Lasso (Least Absolute Shrinkage and Selection Operator) are two well-known embedded methods that use penalty functions to select the best features.

Both these methods have been used for various bioinformatics studies [35]. Glmnet package, MATLAB implementation of Elastic Net and Lasso was used for feature selection [37].

ReliefF

ReliefF is one of the widely used instance based attribute ranking scheme [38]. The main idea of ReliefF is to iteratively estimate feature weights according to their ability to discriminate between neighboring patterns. Each time, random samples are drawn from the dataset and for this instance the neighbors of the same class and the opposite class are determined. Based on these neighboring cases the weights of the attributes are adjusted [39, 40]. ReliefF doesn't remove statistically dependent attributes but relies on a multivariate relevance criterion that ranks the attributes in context of other attributes.

SVM Attribute Evaluator

SVM Attribute Evaluator uses recursive feature elimination (RFE) method in combination with linear support vector machine (SVM). The algorithm builds a model using linear support vector machines and ranks the attributes based on the size of the coefficients. During iteration, it computes the attribute ranking criterion for each attribute and removes the attribute with the smallest ranking criterion. Finally we will have ranked attributes as output.

Logistic Model Tree (LMT)

A Logistic Model Tree (LMT) is an algorithm for supervised learning tasks which combines both linear logistic regression and tree induction. Linear logistic regression tries to fit a simple stable model to the data with low variance and high bias while the tree induction searches a less restricted space of models and capture nonlinear patterns in the data with high variance and low bias. LMT combines the best features from both the methods. It creates a model tree with a standard decision tree structure with logistic regression functions at leaf nodes. In LMT, leaves have an associated logic regression functions instead of just class labels [41]. Tree induction method has been used to identify potential biomarkers from proteomics data in cancer classification studies [42].

3.2.4 Performance Evaluation

Accuracy is usually a good measure for binary class problem consisting positive and negative samples. It is defined as

$$\text{Accuracy} = (\text{TP}+\text{TN})/\ (\text{TP}+\text{FP}+\text{TN}+\text{FN})$$

Here, TP and TN represent the number of correct predictions in the positive and negative class while FP and FN represents the misclassifications.

Determining the ROC Area under the curve (AUC) and Mathew's correlation coefficient (MCC) are possible other metric to assess classification quality especially when there is obvious disparity in the number of samples between two classes [33]. The ROC curve is a plot of the true-positive rate (sensitivity) versus the false-positive rate (1-specificity), and the AUC is equivalent to the non-parametric Wilcoxon–Mann–Whitney test of ranks. For a binary classification problem, an AUC value of 0.5 suggests that the model built by the classifier will perform no better than random guessing while a value of 1.0 shows that it is a best classifier.

MCC is defined as,

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FN})(\text{TP} + \text{FP})(\text{TN} + \text{FN})(\text{TN} + \text{FP})}},$$

It combines both the sensitivity and specificity in to one measure and the values lie in the range of -1 to +1, while 1 means complete prediction accuracy and 0 means every prediction was random [43].

3.3 Results

3.3.1 Accuracy based on selected analytes

Accuracy for evaluating a feature selection approach is defined as the overall classification performance here, including not only percent agreement with clinical diagnosis (accuracy rate), but also area under ROC curve (AUC) and Mathew's correlation coefficient (MCC), by using only the features selected or top-ranked by that feature selection approach. Although this definition has been widely used, there're three key points need to be considered in classification processes. First, which classifier is used? Second, how many features are using? Third, highest accuracy rate with poor AUC or MCC will NOT be considered as the best accuracy.

We used training set from partition coverage 1 to select the best features from various methods listed above. For fair comparison we choose only the top 12 features which is the least number of features selected by most of the methods on partition coverage 1. We know that it is impossible to find a classification algorithm that performs well with all feature selection algorithm as every classification algorithm has its own learning bias. So in order to find the best combination of feature selection and classification algorithm for this kind of proteomics data we choose 20 different classifiers available in Weka.

Logistic Model Tree (LMT) performs consistently well for the features selected by different set of feature selection algorithm with highest accuracy of 84% and AUC 0,94 for the features selected using ReliefF. LMT has been proven to be performing well with the proteomics data with small set of features earlier [25]. So we decided to use LMT as a classifier for further analysis.

Table 3: Results of best classifiers with Top 12 analytes selected using ReliefF and SVM

Classifier	**ReliefF**			**SVMAttributeEval**		
	Accuracy	MCC	AUC	Accuracy	MCC	AUC
BaysianLogistic Reg	0.80	0.55	0.77	0.84	0.67	0.85
Naive Bayesian	0.84	0.63	0.90	0.85	0.66	0.90
LibSVM	0.83	0.60	0.78	0.85	0.66	0.82
Logistic	0.80	0.55	0.85	0.81	0.62	0.92
SMO	0.80	0.54	0.75	0.86	0.71	0.87
ClassViaRegress	0.83	0.60	0.90	0.78	0.47	0.90
JRIP	0.80	0.54	0.75	0.77	0.46	0.73
LMT	0.84	0.63	0.94	0.85	0.67	0.93

Once the classifier was fixed, we tried to identify the best feature selection method by repeating the above analysis with all the partitions. The number of attributes selected for classification varies between the partitions, depending on the minimum number of features selected by the feature selection algorithm for that particular partition. And the results showed that ReliefF and SVMAttributeEval classify with better accuracy and AUC for different partitions followed by Elastic Net and CostSensitiveAttributeEval. ReliefF achieves highest accuracy of 80% with AUC 0.90 for partition coverage 1 and SVM achieves 80% accuracy with AUC 0.86 for partition 4.

Table 4 Results of top six feature selection methods for partition coverage 4

Feature Selection method	Accuracy	MCC	AUC
ReliefFAttributeEval	0.80	0.54	0.90
SVMAttributeEval	0.75	0.44	0.81
CostSensitiveAttributeEval	0.70	0.29	0.73
Elastic Net	0.73	0.36	0.80
Lasso	0.73	0.36	0.80
OneRAttributeEval	0.68	0.26	0.69

3.3.2 Efficiency of selected analytes

Efficiency for evaluating a feature selection approach is defined as the least number of selected or top-ranked features used for classification that can achieve the best accuracy.

Apart from the methods that give the minimum number of features for each partition, there are five methods viz... CostSensitiveAttributeEval, ReliefF, SVMAttributeEval, ElasticNet and Lasso provide either weights or ranks for all the 146 features. There is a possibility that information contained in the top features produced by this algorithm may be lost when we compare all the feature selection with the minimum number of features in each partition. So we compared the 10 fold classification accuracy and AUC of the top 80 features selected by the five methods in the master data set and the results are shown below.

Figure 4: Comparison of Accuracy of five feature selection methods

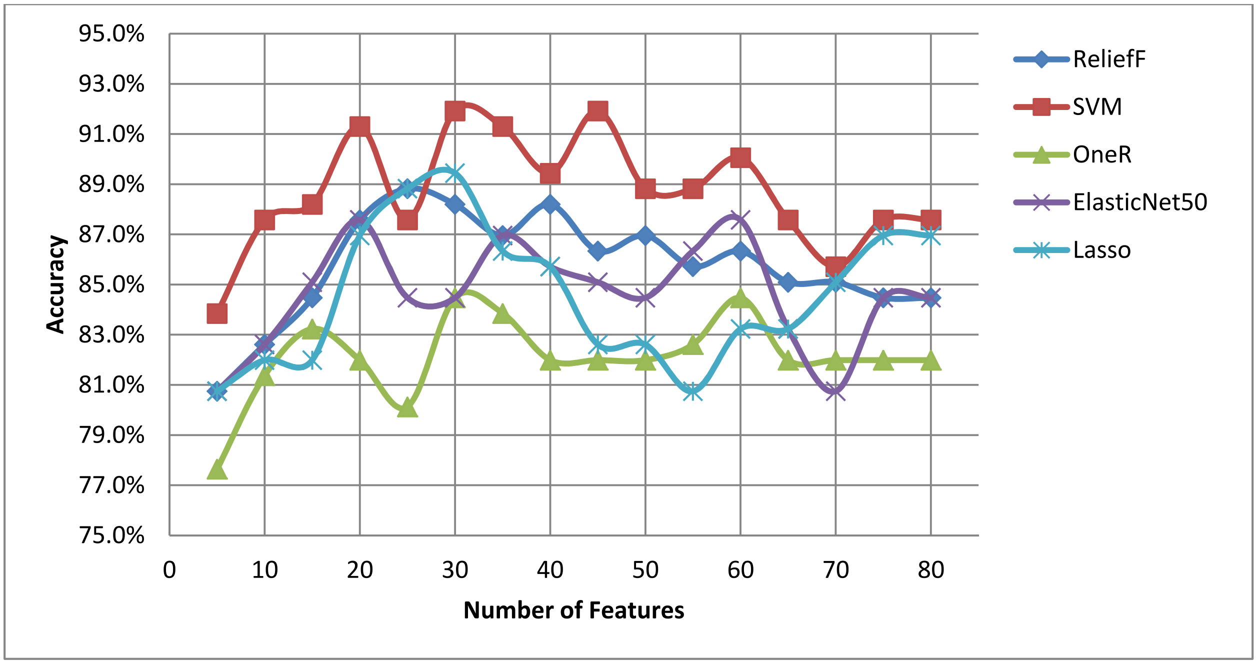
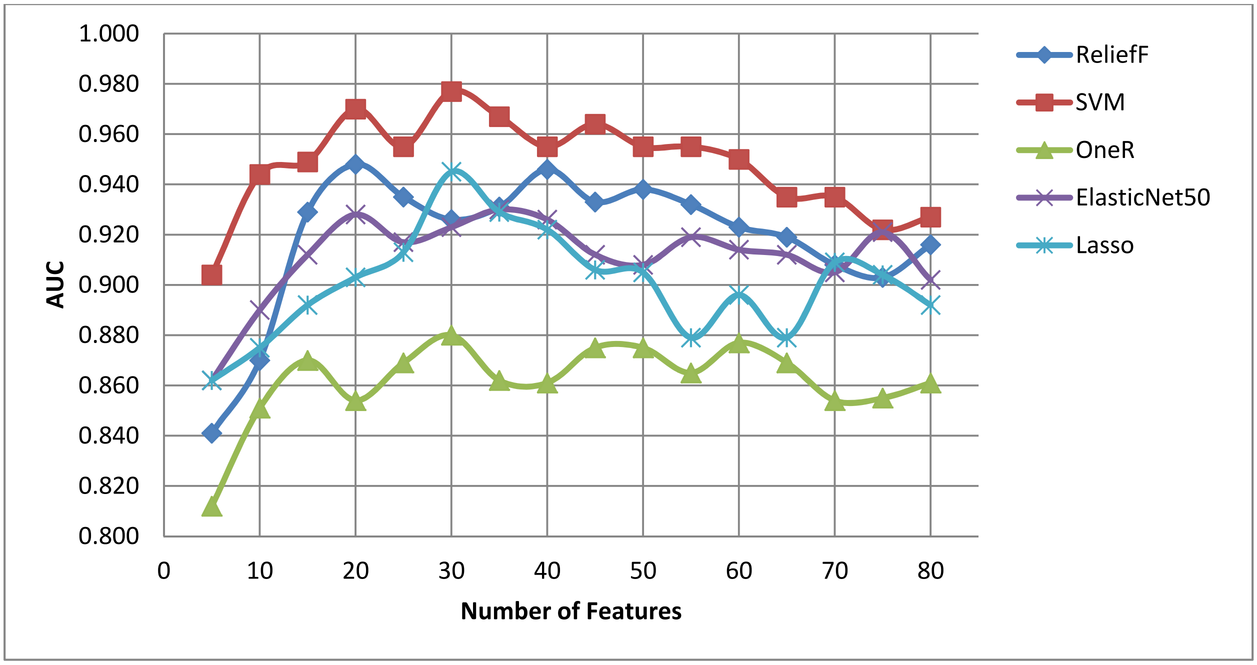


Figure 5: Comparison of AUC of five feature selection methods



As we infer from the figure, all the five methods achieves the highest possible accuracy with features not more than 32. Top 32 features selected using SVMAttributeEval classifies the samples with accuracy of 92.5% and AUC 0.98. Both ReliefF and Lasso shows accuracy of about 89% and AUC 0.94. OneR is the least performing method in case of both accuracy and AUC. As we used 10 fold cross validation accuracy to compare the results, it is no surprise that SVMAttributeEval perform better in this case. Since it is an embedded feature selection method which uses linear SVM classifier weights to rank the attribute, it may over fit the model. To overcome this problem, we run our analysis in partition coverage using separate training and testing set. We compared the predictive power of top features from both ReliefF and SVMAttributeEval by reducing the number of features in the multiples of 3. Results shows that top 15 features performs better in most of the partitions with ReliefF achieving a maximum accuracy of 89% with AUC 0.93 in the case of partition coverage1.

Table 5: Classification accuracy of Top 30 analytes using LMT

Classifier	**ReliefF**			**SVMAttributeEval**		
	Accuracy	MCC	AUC	Accuracy	MCC	AUC
Top 30 features	0.80	0.59	0.89	0.77	0.49	0.83
Top 27 features	0.84	0.65	0.92	0.80	0.56	0.86
Top 24 features	0.80	0.56	0.88	0.74	0.41	0.80
Top 21 features	0.86	0.70	0.92	0.75	0.46	0.81
Top 18 features	0.86	0.70	0.94	0.74	0.43	0.83
Top 15 features	0.89	0.75	0.93	0.80	0.57	0.87
Top 12 features	0.83	0.61	0.90	0.72	0.37	0.81
Top 09 features	0.80	0.54	0.90	0.75	0.44	0.81

3.3.3 Consistency of selected analytes

We earlier assessed the consistency of the five feature selection algorithm by their prediction accuracy with LMT classifier in all the partition. In the five methods, SVMAttributeEval ranks the features while all the other four methods assign weight to them. We decided to compare the consistency of these methods in selecting analytes across the partition. Since the weight values produced by the algorithms are heterogeneous with weights ranging from -1 to 1, 0-100, a unified weighting schema was devised.

First we ranked the features based on their weight values for all the feature selection method except SVMAttributeEval for which we already have the ranks available from Weka for all the partition. When more than one analyte has same weight, they were alphabetically sorted.

Once we have the standardized weight for each analyte, the average standard deviation is calculated which we call as consistency score. Lower the score, higher the consistency. Consistency score for the five methods were shown in table 6. As we infer from the table 6, ReliefF has low score of 0.13 with higher consistency in selecting the features across 20 partitions. It was followed by Lasso and ElasticNet with scores 0.15 and 0.15 respectively.

3.3.4 Significance of selected analytes

We identified that top 32 features selected by the five feature selection methods contributes more for the classification. To explore the association between these features and Alzheimer's disease we performed a literature analysis for all the top 32 analytes. We searched the pubmed with search criteria ("Analyte" OR "Full Name" OR "Gene Symbol") AND "Alzheimer" and get the number of records. A significant score was calculated for each method by taking average of log transformed count. Higher the score, more significant are the features selected by the approach. The significance here means the agreement between the analytes selected by the feature selection methods with the evidence found in the current literatures. Features selected by Elastic Net shows higher significancy with significant score of 1.22 followed by ReliefF and Lasso.

Table 6: Consistency and significance scores of Top 32 analytes

	ReliefF	SVM	OneR	Elastic Net	Lasso	Integrated
Consistency score	0.13	0.18	0.21	0.15	0.15	N/A
Significance score	1.21	0.94	1.07	1.22	1.13	1.19
Few Supported	4	8	8	6	6	10
Nature 2007 (18)	0	1	0	3	2	3
PloS one 2008 (5)	0	1	0	3	2	3
ANYAS 2008 (89)	13	15	10	16	15	28

3.3.5 Protein Interaction network

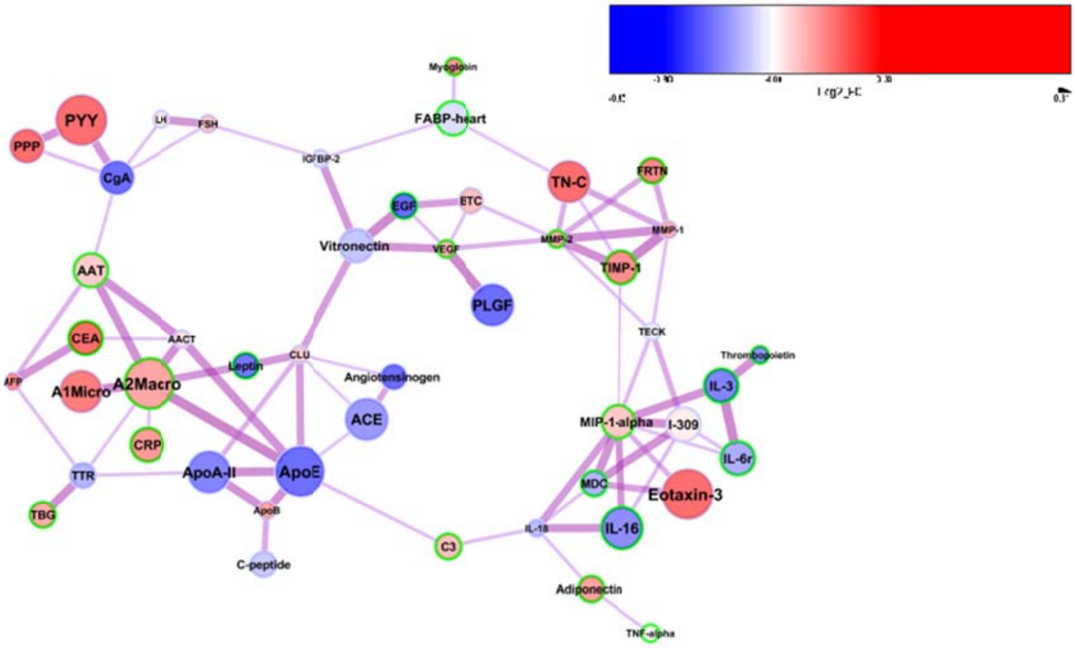


Figure 6: Integrated CRC network showing Subnetwork signatures

We identified that top 32 features selected by the five feature selection methods contributes more for the classification. We expanded the top features in HAPPI with confidence score (*CI* >=0.75, i.e. 4-star rating) for interactions, to obtain a protein-protein interaction (PPI) network.

3.4 Discussion

Sub network 1 shows the GPCR Signaling pathway induced by Follicle Stimulating Hormone (FSH), Luteinizing Hormone (LH) and pancreatic polypeptide (PPY). Both FSH and LH has been linked to Neuroactive ligand receptor interaction pathway too. G Protein Coupled Receptors (GPCRs) are involved in the process of cleavage of amyloid precursor proteins and also in various key neurotransmitters system. There are various studies which supports the notion that GPCRs and activation of their downstream signal cascades increases the non-amyloidogenic processing of APP [44, 45]. GPCRs are also involved in neuroinflammation and plays role in Amyloid β mediated toxicity. Class A receptors of GPCRs seen in the hippocampus and cortex of the brain are abundantly expressed in the microglial cells of AD patients. Several attempts had been made to use this adenosinergic system as a potential therapeutic target for managing cognitive dysfunction in AD [46]. Though the involvement of these hormones with AD has been reported earlier, role of hormone induced GPCR signaling in AD is quite a fascinating one. Surprisingly both FSH and LH induce GPCR signaling through Class A receptors. This provides an interesting insight in to the pathways involved in AD which could be a potential therapeutic target and complement the current treatment approaches that focus mainly on secretase inhibitors and amyloid immunotherapy.

Analytes from sub network 2 involved in four pathways viz. lipid metabolism, Complement activation, Renin Angiotensin System and Hemostasis. Role of Lipid metabolism in AD has been well known with central obesity is related to a high risk of Late Onset Alzheimer's disease

(LOAD) [47]. Apart from this, the majority of the analytes from this sub network involves in complement and coagulation pathways. The interactions between the components in complement activation pathways and hemostasis is well established [48]. Both the complementation cascade and the blood clotting were activated by same kind of stimuli. Multiple regulatory loops between these two systems provide an effective host response against infection. Complement system activation due to accumulation of Amyloid β and the involvement of other key analytes in hemostasis was captured in sub network 2.

Sub network 3 connected to network 2 through C3 has around 8 analytes that involves in Cytokine – cytokine receptor interaction and Chemokine signaling. Cytokines plays crucial role in innate and adaptive inflammatory responses, cell growth, differentiation, angiogenesis and homeostasis. There are considerable evidences to suggest that an inflammatory response is involved in the AD neurodegenerative cascade. A detailed review on the cytokine AD association highlighted the elevated levels of several key analytes that was shown in sub network 2 such as TNF-α, IL-6r, IL-16 and IL-1847.

Sub network 4 shows analytes that involves in various pathways such as Cytokine – cytokine receptor interaction, various cancer pathways (Pancreas and Bladder cancer), Hemostasis and mTOR Signalling pathway. Analytes that participates in cancer related pathways are involved in two key processes, while Epidermal Growth Factor (EGF) involves in evading apoptosis, Vascular Endothelial growth factor (VEGF) and Matrix metallopeptidase 1 (MMP1) involves in VEGF Signaling which eventually helps in Angiogenesis. Moreover, mTOR signaling pathway plays a central role in various neuronal functions and maintains hemostasis, it also regulates different forms of learning and memory.

Another promising feature in the integrated network is the coherence between the sub networks achieved through bridge analytes. Sub network 1 connects to sub network 4 through Insulin-like growth factor binding protein 2 (IGFB2). Insulin signaling plays a role in learning and memory and deregulated insulin signaling occurred in the brains of patients with AD[46]. Hence Type 2 diabetes has been identified as a major risk factor for AD, and the onset of diabetes worsens cognitive disorders even in the absence of amyloid plaques. Cognitive decline associated with neuronal cell death (apoptosis) has been targeted in AD treatment using anti diabetic medicine. IGFB2 involves in insulin signaling pathway which controls vital brain functions such as cell survival, energy metabolism and neuroregeneration [49]. Similarly, sub network 2 connects with GPCR Signaling sub network 4 through AAT and with Sub network 2 through analytes that in complement system activation.

We also confirmed the results from sub network by Gene Ontology (GO) analysis. Results from GO performed with DAVID and pathway analysis with Kyoto Encyclopedia of Genes and Genomes (http://www.genome.jp/kegg/) and Reactome (http://www.reactome.org/ReactomeGWT/entrypoint.html) also confirmed the involvement of the selected analytes in Chemokine Signalling pathway, Hematopoetic Cell Lineage, Complement Activation pathway and Focal Adhesion.

4. INTEGRATIVE NETWORK ANALYSIS OF MICRO RNA AND MRNA

4.1 Background

4.1.1 MDSC and T-cell suppression

Myeloid derived suppressor cells (MDSC) constitute a unique component of immune system that expand during cancer, inflammation, and infection, and capable of suppressing T-cell responses. In addition to T-cell suppression, MDSCs have also been linked to innate immune response regulation through cytokines. MDSCs were described more than 20 years ago in patients with cancer and found to play significant roles in tumor angiogenesis and metastasis [50]. They are heterogeneous group of cells that consists of myeloid progenitor cells and immature myeloid cells (IMCs). In normal conditions, these IMCs matures in to granulocytes, macrophages and dendritic cells, while in pathological conditions such as cancer, auto immune disorders, sepsis and in some infectious disease, a partial block in the differentiation of IMCs in to mature myeloid cells results in the expansion of this population. It results in upregulation of immune suppressive Arginase 1 (ARG1), inducible Nitric oxide synthase (iNOS), Nitric oxide (NO) and Reactive oxygen Species.

In Mouse, MDSCs are characterized by the co-expression of myeloid cell differentiation antigen (GR-1) and CD11b (α integrin). Subtypes of MDSC have been defined in the mouse based on the antibody specificity of GR1's two epitopes LY6G and LY6C. Granulocytic MDSCs have a CD11b+Gr1+ phenotype, whereas MDSCs with monocytic morphology are CD11b+Gr1-. These two subsets have different functions in cancer, infectious and auto immune diseases and employs different mechanism to suppress T cell function. Mouse bone marrow has 20-30% of these cells, while spleen has 2-3% and absent completely in Lymph nodes. Previous studies had observed significant functional activity in freshly isolated cells at the site of infection (functional MDSC) while it is completely absent in peripheral cells (MDSC precursors) [51].

4.1.2 MicroRNA and Immune system

Micro RNAs are small, single stranded non-coding RNAs that are involved in the regulation of protein expression in many biological systems. They are about 22 nucleotides long and they predominantly bind to the 3' untranslated region (3'UTR) of messenger RNAs (mRNAs) to inhibit translation or to induce cleavage. So far more than 700 miRNAs have been identified in human genome and each have the potential to suppress the expression of thousands of genes. More than 100 different miRNAs are expressed by cells of the immune system; they have the potential to broadly influence the molecular pathways that control the development and function of innate and adaptive immune responses. Depending on the nature of the target, miRNAs have tumor suppressive or tumor promoting effect on various cancers of immunological origin [52].

In this study, we identified and validated crucial miRNA-gene associations that can be used to study the difference in the molecular mechanism between functional MDSCs and MDSC precursor. We have compiled all the existing Micro RNA resources and used a knowledge guided approach to build an integrated miRNA-gene network to identify the significant genes. miRNAs and pathways through which functional MDSCs differ from their precursors.

4.2 Approach

4.2.1 Differential expression analysis

We isolated both the granulocytic (G^{high}) and monocytic ($G^{\ low}$) subtypes of MDSC cells from spleen and peritoneal cavity (PC) that has a peritoneal tumor. Spleen cells are G^{high} or $G^{\ low}$. Peritoneal cavity cells are G^{low}, G^{mid} and G^{high}. Totally, there are six contrast groups: PC G^{high} vs.

Sp $G^{high,}$ PC G^{mid} vs. Sp G^{high}, PC G^{low} vs. Sp G^{high}, PC G^{high} vs. Sp G^{low}, PC G^{mid} vs. Sp G^{low}, PC G^{low} vs. Sp G^{low}. We have mRNA arrays (GeneChip® Mouse GENE 1.0 ST) and micro-RNA arrays (GeneChip® miRNA array) from the same RNA samples.

Data preprocessing, including quality control and normalization, will be implemented by using standard packages in Bioconductor. Filters based on fold changes, p-values, and detection numbers will be applied to obtain differential miRNAs for each contrast group. Criteria of miRNA array filters will be determined based on existing literatures [53].

For the same six contrast groups, data preprocessing, including quality control and normalization, will also be implemented by using standard packages in Bioconductor. Filters based on fold changes, p-values, and presence/absence calls of mRNA probe IDs will be applied to obtain differential genes for each contrast group. In many cases, crucial genes show relatively slight changes, and many genes selected are also poorly annotated [2]. So criteria of mRNA array filters will be determined according to how many genes are finally obtained.

4.2.2 Data integration for miRNA-gene associations

Table 7 shows some of the primary databases that provide a comprehensive view of microRNAs.

Table 7: Primary microRNA databases and feature comparisons

Database **Features**	**miRBase**	**HMDD**	**miRecords**	**TarBase**	**miR2Disease**	**miRGator**	**miRo**
Target Gene Information	Only Predicted genes	NA	Predicted and Validated genes	Only Validated genes	Yes	Yes	Yes
Disease	NA	Yes	NA	NA	Yes	Yes	NA

Association							
Gene Ontology	NA	NA	NA	NA	NA	NA	Yes
Download	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Statistics for Human	1048 records	450 miRNA genes, 258 diseases,	548 miRNAs, 1579 target genes	1300 validated targets	349 miRNA, 134 Disease	Expression profile	NA
Comments	Central Repository of miRNA	Tissue & Gene wise disease association	Largest source of validated target	Second largest of validated targets	HMDD with Additional features	Expression Profile	GO info

There are around 12 prediction tools (Table 8 shows main 8 tools of them) are available which can predict the miRNA-targeted genes. These algorithms uses various structural features such as hairpin length, hairpin loop length, thermodynamic stability of miRNA-mRNA duplex, base pairing, and distance of microRNA from the loop of its hairpin precursor; and sequence features such as nucleotide content and location, 3`UTR sequence complementarity, and nucleotide repeats [54]. However, for the most part, all these target prediction methods generate a large number of false positives. Several algorithms addressed this problem by considering conservation of sequences across the species which eliminates poorly conserved sites [55].

Table 8: Tools for microRNA-targeted gene prediction and feature comparisons

Database Features	miRDB	TargetScan	picTar	microRNA	RNAhybrid	Diana MicroT	PITA
Search Features	miRNA, Gene Name and batch	Gene Name	miRNA and Gene Name	miRNA name	miRNA sequence(s)	miRNA, Gene Name	miRNA sequence
Download	Yes	Yes	No	Yes	Yes	Yes	Yes
Statistics	2295 microRNA	17821 Human Genes	NA	1100 Human miRNA	NA	NA	NA
Update	April 2009	Release 5.2, June 2011	March 26, 2007	August 2010	2006	April 2009	NA
Comments	SVM based target prediction method.	Sequence similarity and conservation	Mouse based, looks for conservati on in Human	Official source of expression profile	Free energy	Has miRPath and miRExtra tools	Seq. similiarit y

A recent review paper [56] shows that three computational algorithms - TargetScan, DianaMicroT and miRanda/mirSVR can provide miRNA target gene prediction with higher precision. TargetScan 5.2 [57] is one of the most widely used microRNA prediction algorithm. It predicts the microRNA binding sites through the identification of 7 nucleotide seed matches on the 3`UTR of mRNAs and the assessment of their evolutionary conservation across several species. It uses RNA Fold to calculate thermodynamic free energy of the binding, and scores both the single and multiple binding sites. DianaMicroT [58] is an algorithm based on several parameters calculated individually for each microRNA and it combines conserved and non-conserved microRNA recognition elements (MRE) in to a final prediction score. The total predicted score of a miRNA-gene association is the weighted sum of conserved and unconserved MREs of a gene. Both DianaMicroT and TargetScan estimated to achieve a precision level of 66% and 60% respectively in a recent study, outperforming most other prediction algorithms. Algorithm miRanda [59] uses a position weighted matrix to emphasize binding on microRNA 5`end segment, uses RNA Fold for free energy calculation and relies on evolutionary conservation of binding sites. Algorithm mirSVR [55] is a most recent machine learning method that ranks microRNA target sites by a down regulation score. The algorithm trains a regression model using the sequence and contextual features extracted from the target sites predicted by miRanda, thereby combining the efficiency of two methods. Algorithm miRanda/mirSVR is competitive with other target prediction methods and in addition it has a unique ability to predict the extent of downregulation by specific miRNA at mRNA or protein level. Importantly, this method identifies a significant number of experimentally determined non-canonical and non-conserved sites. All these three algorithms will be used to predict the genes targeted by specific miRNAs in our project.

4.2.3 Correlation analysis between miRNA and mRNA arrays

Both Pearson's correlation and Spearman's correlation (non-parametric) will be calculated to correlate the expressions of all miRNAs with all mRNAs through all samples. According to the distributions of miRNA arrays and mRNA arrays, filters based on correlation coefficients and p-values will be applied to obtain statistically significant miRNA-gene correlations. Then the filters based on fold changes, p-values, and detection numbers of miRNAs will also be applied to obtain statistically significant miRNA-gene correlations for differential miRNAs. These differential miRNA-gene correlations will be used as a supplementary data for miRNA-gene associations retrieved from databases and computational prediction after a functional enrichment validation.

4.2.4 Validation for miRNA-gene associations at pathway-level

Since miRNA-gene association data retrieved from databases is far from complete, while miRNA-gene association data predicted by computational algorithms is very noisy (high FPR - false positive rate), it is not easy to validate miRNA-gene correlations from miRNA and mRNA array correlation analysis at molecule-level. Hence, functional enrichment for miRNA target genes will be processed for miRNA-gene association validation first. For functional enrichment, both pathway analysis and Gene Ontology (GO) analysis will be applied.

As shown in Table 9, there are six online tools which can be used for miRNA target gene functional enrichment and pathway analysis. Of these tools, Diana miRPath uses only list of miRNAs to predict target genes and enrich these predicted genes in KEGG pathways, while other tools requires array datasets. We will mainly use miRPath for microRNA-targeted gene functional enrichment analysis at pathway-level.

Table 9: Tools for miRNA target gene functional enrichment and annotation

Database **Features**	**miRPath**	**miTALOS**	**Magia**	**miRGen**	**mirAct**
Target Gene Information	Uses DianaMicroT to predict gene	Uses TargetScan	miRanda, PITA & TargetScan	From miRanda, picTar, TargetScan & DianaMicroT	miRanda, picTar, TargetScan & PITA
Functional	KEGG	KEGG, tissue	GO, Network	Maps to UCSC	Clustering

Annotation		specific expression	enrichment	genome browser	
Expression Profiling	No	Yes	Based on user input	No	User Input data
Comments	Both single and batch processing	Tissue specific enrichment method	Analyze user input exp. Data	Positional relationships & Cluster info	Using expression data

First, miRNAs will be input into miRPath (and other tools will be also tested) for target gene prediction and functional enrichment, which will generate a list of pathways ranked by -log(p-value) [60]. Second, experimentally-validated miRNA target genes will be enriched in a comprehensive human pathway database (HPD) [61], which has integrated heterogeneous pathways from five data sources - NCI-Nature curated Pathway Interaction Database (PID), Reactome, BioCarta, KEGG and ProteinLounge. An online pathway analysis tool based on HPD will also generate a list of pathways ranked by similarity scores [62]. Third, two ranked pathway lists will be compared to assess whether predicted miRNA target genes have same enriched functions with experimentally-validated miRNA target genes at pathway-level. Finally, differential miRNA-gene correlations will also be validated by the same way of pathway enrichment analysis, but using both experimentally-validated and computationally-predicted miRNA target genes.

4.2.5 Network of differential miRNAs and genes

An integrated network will be constructed to connect miRNAs and genes differentially-expressed in miRNA arrays and mRNA arrays respectively. First, differential miRNAs will be connected to target genes from the above integrated miRNA database and also computational predictions. Validated miRNA-gene correlations will be also added into connections. Second, differential genes will be connected to the genes targeted by differential miRNAs by using high-quality interactions from the human annotated and predicted protein interaction (HAPPI) database [63]. One or two intermediate proteins will be used to bridge the connections. Third, we will integrate miRNA-gene associations and gene/protein-gene/protein interactions together, to build the network connecting differential miRNAs and differential genes. Finally, the comprehensive network will not only provide a systems-level view for the study on functional activities of MDSCs, but will also serve as a molecular interaction network model to identify significant miRNAs and genes, which could be used as biomarkers to distinguish functional MDSCs from MDSC precursors.

4.3 Results

Of the six contrast groups in this preliminary study on mRNA arrays for MDSC, we focused on a result of 2-way ANOVA analysis on PC Glow vs. SP Glow contrast in mRNA arrays. From a recent review [50], we selected 11 MDSC-related genes - ARG1, NOS2, IL1RL2, VDR, SLC7A2, TLR4, FOLR2, HIF1A, S100A9, CEBPB, and S100A8, which are all differentially expressed in PC Glow vs. SP Glow contrast from mRNA arrays (each group has 4 samples).

4.3.1 Experimentally validated miRNA-gene association network

We focused on a result of 2-way ANOVA analysis on PC Glow vs. SP Glow contrast in miRNA arrays (each group has 4 samples). From differential analysis, 153 miRNAs, of which 13 are duplicates (Same miRNA from different species), have been selected by using the filter (p-value <=0.05 and |Fold Change| >= 1.5). We combined 153 experimentally-validated miRNA-gene associations from two databases (miRecords and Tarbase) and 42 protein-protein interactions with 4-5 star quality retrieved from HAPPI to build a differential miRNA targeted gene network, shown in Figure 5. The network contains 35 differential miRNAs having target genes (validated by experiments from two databases – miRecords and Tarbase) out of totally 140 differential miRNAs. Although some interesting differential genes are found, such as VEGFA and MYC, the network that built from only experimentally-validated data shows limited information. In the network, many MDSC-related genes are not involved, which implies that high-quality miRNA-gene association data is quite incomplete.

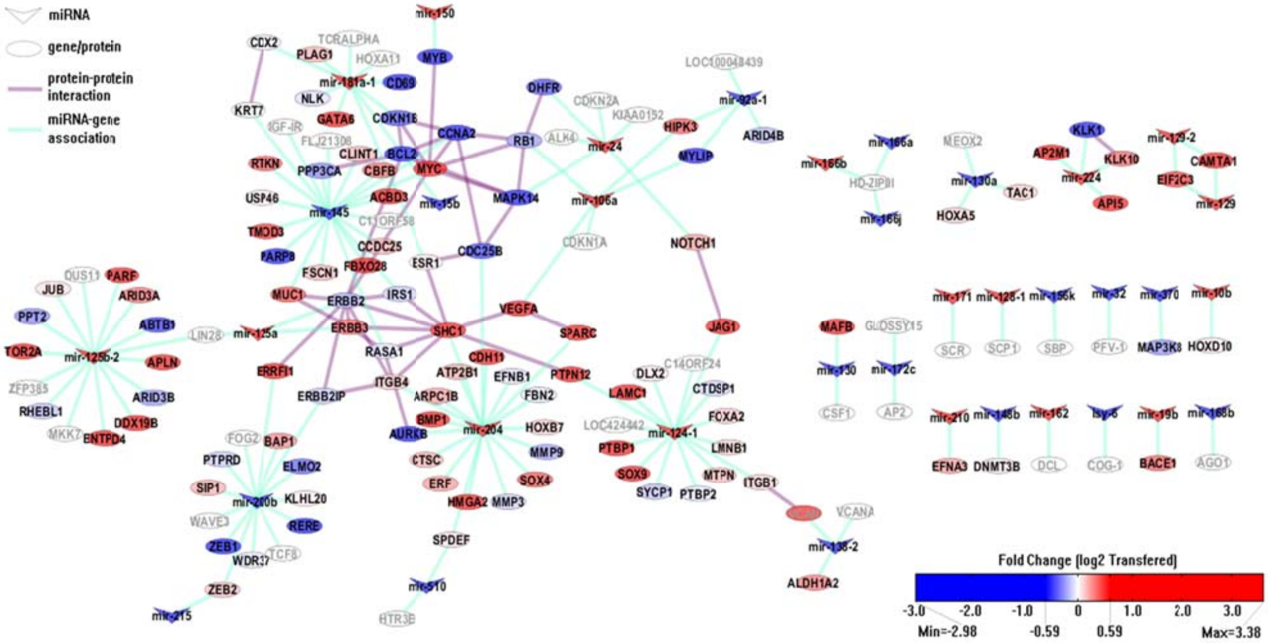


Figure 7: Differential miRNA target gene network from experimentally-validated databases. *The network contains 35 differential miRNAs having targeted genes (validated by experiments from two databases – miRecords and Tarbase) out of totally 140 differential miRNAs (p-value <=0.05 and |Fold Change| >= 1.5) from the 2-way ANOVA analysis on PC Glow vs. SP Glow contrast in miRNA array, and 107 targeted genes with expression values (not filtered) in mRNA array out of 140 targeted genes. Fold changes on PC Glow vs. SP Glow contrast in mRNA array for these107 targeted genes are also represented as node color, by using the same color map as for miRNA. Node color here represents Log2-transferd fold change. Genes without expression values are labeled with gray font. Vee node shape represents miRNA, while ellipse node represents gene/protein.*

4.3.2 Computationally predicted miRNA-gene association network

We connected the 11 significant genes and the 5 differential miRNAs to the network (shown in Figure 1), by using computationally-predicted miRNA-gene associations and 5-star interactions from HAPPI. If there is no direct interaction, one intermediate protein will be used to bridge the connections. Another network integrated with computationally-predicted miRNA-gene associations is shown in Figure 8, from which, we can see that the integrated network contains more interesting information than the one in Figure 7.

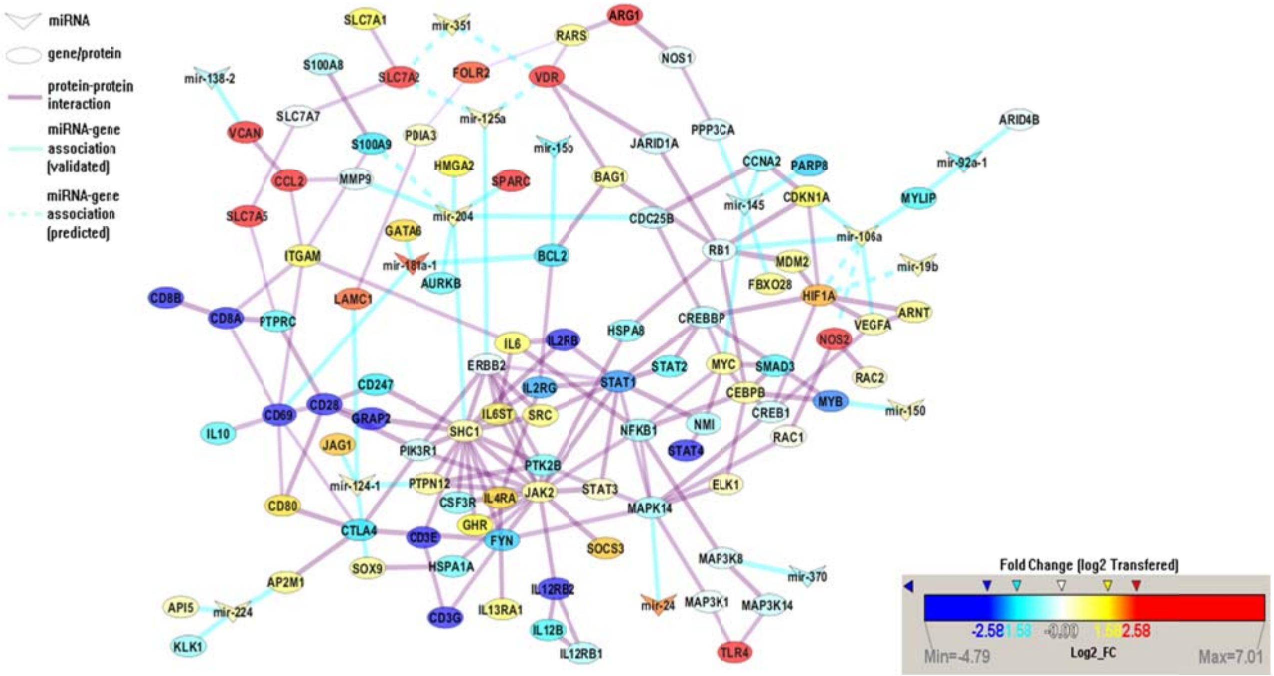


Figure 8: Differential miRNA targeted gene and significant gene network. *The network contains 15 differential miRNAs and 97 genes. Fold changes for miRNAs and genes on PC Glow vs. SP Glow contrast in miRNA and mRNA arrays are all represented as node color. Node color here represents Log2-transferd fold change. Vee node shape represents miRNA, while ellipse node represents gene/protein.*

4.3.3 Validation of computationally predicted association

Based on the three computational miRNA-gene association prediction algorithms - TargetScan, DianaMicroT and miRanda/mirSVR, 153 miRNAs (after removing duplicates, 115 unique miRNAs) have potentials to target these 11 significant MDSC-related genes. By using the filter (p-value <=0.05, |Fold Change| >= 1.5, and Max detection number > 0), 5 differential miRNAs (4 of which are in the network shown in Figure 1) are selected:

- mmu-miR-106a → HIF1A, NOS2
- mmu-miR-125a → VDR, SLC7A2
- mmu-miR-19b → HIF1A
- mmu-miR-204 → S100A9
- mmu-miR-351 → SLC7A2, VDR

By using miRPath, top-10 KEGG pathways associated with these 5 miRNAs are listed below:

(a) Axon guidance (b) MAPK signaling pathway (c) Long-term potentiation (d) Insulin signaling pathway **(e) mTOR signaling pathway (f) Renal cell carcinoma (g) Melanogenesis** (h) Glioma (i) Chronic myeloid leukemia **(j) Focal adhesion**

Most of the pathways (highlighted with fold font) are overlapped with the pathways enriched in HPD from the 5 differential miRNA-targeted genes, which implies that computationally-predicted miRNA-gene associations can be trusted at pathway-level.

We used an online gene set analysis toolkit Web Gestalt [64] for Gene Ontology (GO) analysis. Significant genes overrepresentation was found by hypergeometric statistical. Genes were enriched in the following three sub categories which are strongly associated with signal transduction pathways in immune response.

- Biological Process: Regulation of Lymphocyte Activation
- Molecular Function: Cytokine Receptor Activity
- Cellular Location: External Side of Plasma membrane

4.3.4 MDSC specific Molecular Interaction Network

We expanded (similar to enrichment process) the curated genes in HAPPI to construct a MDSC-specific molecular interaction network (shown in Figure 9). Many expanded genes are also highly differentially-expressed, such as ARG2, SMAD3, MYB, EGLN3, TAP1, STAT1, and MS4A1, CD40 etc., which implies that the network expansion/enrichment strategy is extremely useful to identify significant genes, even if those genes are neglected at first beginning curation.

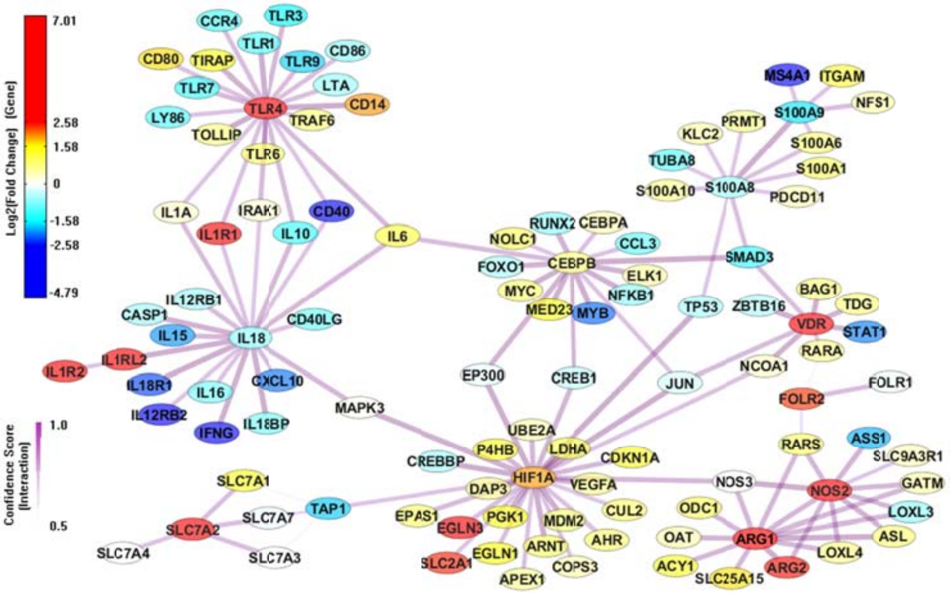


Figure 9: A MDSC-specific molecular interaction network, *constructed based on 11 MDSC-related genes - ARG1, NOS2, IL1RL2, VDR, SLC7A2, TLR4, FOLR2, HIF1A, S100A9, CEBPB, and S100A8. Other genes are expanded from HAPPI. There are totally 112 genes and 134 interactions in the network.*

4.3.5 MDSC specific Pathway Association Network

We also enriched the genes/proteins from the MDSC-specific molecular interaction in HPD to construct a MDSC-specific pathway association network (shown in Figure 10). The pathway results further confirmed the role of MDSC in T Cell Suppressor function through various signaling pathways.

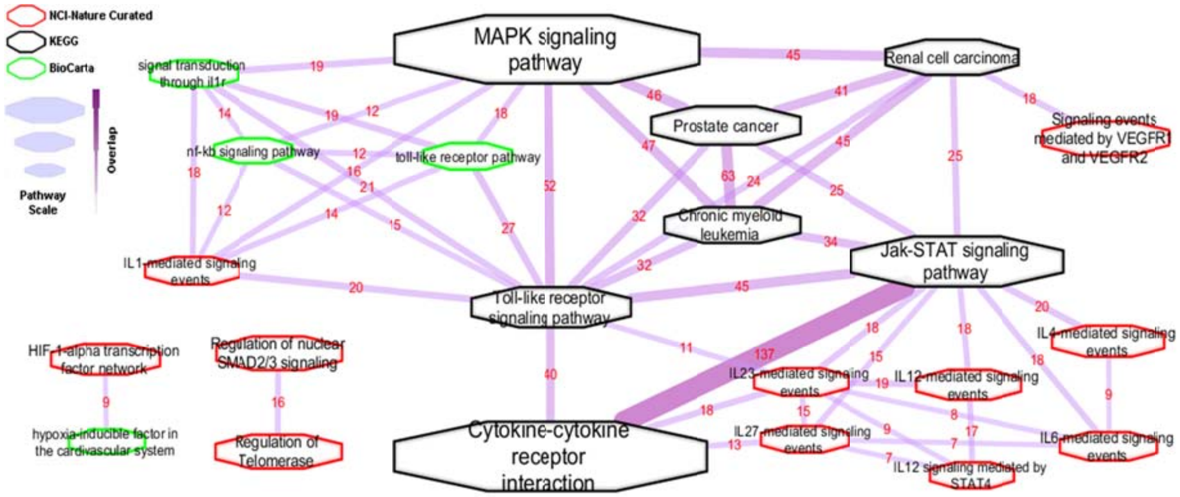

NCI-Nature Curated
KEGG
BioCarta
Pathway Scale
Overlap
MAPK signaling pathway
Renal cell carcinoma
signal transduction through il1r
Prostate cancer
Signaling events mediated by VEGFR1 and VEGFR2
nf-kb signaling pathway
toll-like receptor pathway
Chronic myeloid leukemia
Jak-STAT signaling pathway
IL1-mediated signaling events
Toll-like receptor signaling pathway
HIF-1-alpha transcription factor network
Regulation of nuclear SMAD2/3 signaling
IL4-mediated signaling events
IL23-mediated signaling events
IL12-mediated signaling events
hypoxia-inducible factor in the cardiovascular system
Regulation of Telomerase
Cytokine-cytokine receptor interaction
IL27-mediated signaling events
IL6-mediated signaling events
IL12 signaling mediated by STAT4
19
45
41
18
14
19
12
18
46
47
45
25
18
12
16
21
52
63
24
25
12
14
15
27
32
34
32
20
45
20
18
18
11
137
15
19
18
9
16
40
9
18
15
9
8
17
13
7
7

Figure 10: A MDSC-specific pathway association network, *enriched from the MDSC-specific molecular interaction network shown in Figure 1. All the pathways are enriched from HPD. The number of overlapped genes between two pathways is also labeled on each edge.*

We further combined the interrelated ontological process and pathways together through GARNET [65] (Gene Annotation Relationship NEtwork Tools), an integrative platform for gene set analysis. The result is shown in Figure 11, which provides a systems-level view for the study on functional activities of MDSCs.

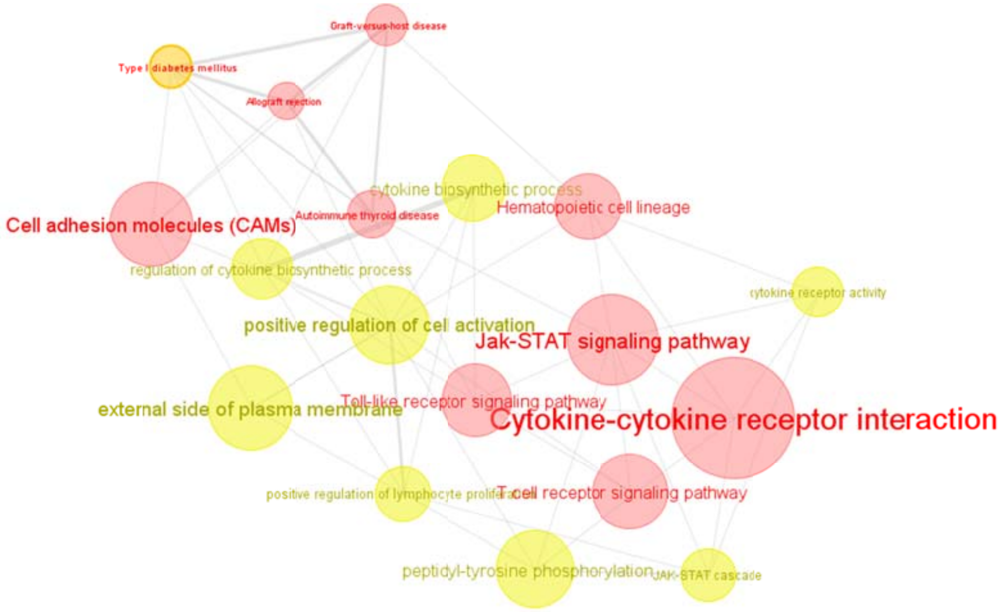

Graft-versus-host disease
Type I diabetes mellitus
Allograft rejection
cytokine biosynthetic process
Hematopoietic cell lineage
Cell adhesion molecules (CAMs)
Autoimmune thyroid disease
regulation of cytokine biosynthetic process
cytokine receptor activity
positive regulation of cell activation
Jak-STAT signaling pathway
external side of plasma membrane
Toll-like receptor signaling pathway
Cytokine-cytokine receptor interaction
positive regulation of lymphocyte proliferation
T cell receptor signaling pathway
peptidyl-tyrosine phosphorylation
JAK-STAT cascade

Figure 11: Significantly enriched Ontology and Pathway network

4.4 Discussion

We used integrated expression profiling, protein interaction information and existing miRNA related databases to identify significant miRNA's that regulates T-cell suppression through Monocytic subset of MDSC. We also verified our analysis on pathway level and identified the role of MAPK signaling pathway, mTOR signaling pathway and Insulin signaling pathways through which T-cell suppression happens in Tumor microenvironment. Rather than pure data driven approach, we employed knowledge guided approach to fill the gaps in the existing information about miRNA. We performed analysis both from mRNA and miRNA and finally integrated them to identify potential miRNAs and pathways. As our analysis focused more on the monocytic subset, we would be expanding our studies to granulocytic subset in the future.

5. Lesson's learned and future directions

We have used various system biology techniques to identify sub network signatures from genomics, proteomics and integrated MicroRNA data set. We used existing gene signature and pathway databases to identify significant molecular expression patterns from high throughput data. Currently, gene expression signature analysis and pathway analysis remains two separate processes, since in many cases, extensive data preprocessing, comprehensive gene selection statistics, and downstream pathway/network analysis cannot be replaced by GSEA. Having a single repository for comprehensive disease associated gene and network/pathway enrichment analysis will be of great use to the scientific community. As a future study, we decided to build an integrated online database - Pathway and Gene Enrichment Database (PAGED), to enable comprehensive search for phenotype-associated gene sets, network modules, and pathways, by integrating gene set based molecular patterns at three dimensions – DNA/genome, RNA/transcriptome, and Protein/proteome. First, disease-gene association data are curated and integrated from Online Mendelian Inheritance in Man (OMIM) database and Genetic Association Database (GAD). Second, functionally-grouped gene sets are evaluated and integrated by using gene signatures in Molecular Signatures Database (MSigDB) and Gene Signatures Database (GeneSigDB). Third, signaling pathways/protein interaction networks and transcription factors/gene regulatory networks are retrieved from Human Pathway Database (HPD) and Human Annotated and Predicted Protein Interaction (HAPPI) database. This integrated database will be of great use to the system biology studies on high throughput data sets.

Bibliographies

1. Allison DB, Cui X, Page GP, Sabripour M: **Microarray data analysis: from disarray to consolidation and consensus**. *Nature reviews Genetics* 2006, **7**(1):55-65.
2. Reimers M: **Making informed choices about microarray data analysis**. *PLoS computational biology* 2010, **6**(5):e1000786.
3. Slonim DK, Yanai I: **Getting started in gene expression microarray analysis**. *PLoS computational biology* 2009, **5**(10):e1000543.
4. Sørlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, Hastie T, Eisen MB, Van De Rijn M, Jeffrey SS: **Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications**. *Proceedings of the National Academy of Sciences of the United States of America* 2001, **98**(19):10869–10874.

5. Giltnane JM, Rimm DL: **Technology insight: Identification of biomarkers with tissue microarray technology**. *Nature clinical practice Oncology* 2004, **1**(2):104-111.
6. Segal E, Friedman N, Kaminski N, Regev A, Koller D: **From signatures to models: understanding cancer using microarrays**. *Nature genetics* 2005, **37 Suppl**:S38-45.
7. Potti A, Dressman HK, Bild A, Riedel RF, Chan G, Sayer R, Cragun J, Cottrill H, Kelley MJ, Petersen R: **Genomic signatures to guide the use of chemotherapeutics**. *Nature medicine* 2006, **12**(11):1294-1300.
8. Glez-Pena D, Gomez-Lopez G, Pisano DG, Fdez-Riverola F: **WhichGenes: a web-based tool for gathering, building, storing and exporting gene sets with application in gene set enrichment analysis**. *Nucleic acids research* 2009, **37**(Web Server issue):W329-334.
9. Medina I, Montaner D, Bonifaci N, Pujana MA, Carbonell J, Tarraga J, Al-Shahrour F, Dopazo J: **Gene set-based analysis of polymorphisms: finding pathways or biological processes associated to traits in genome-wide association studies**. *Nucleic acids research* 2009, **37**(Web Server issue):W340-344.
10. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES *et al*: **Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles**. *Proceedings of the National Academy of Sciences of the United States of America* 2005, **102**(43):15545-15550.
11. Huang DW, Sherman BT, Lempicki RA: **Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists**. *Nucleic Acids Research* 2009, **37**(1):1-13.
12. Dennis Jr G, Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, Lempicki RA: **DAVID: database for annotation, visualization, and integrated discovery**. *Genome Biol* 2003, **4**(9):R60.
13. Pujana MA, Han JD, Starita LM, Stevens KN, Tewari M, Ahn JS, Rennert G, Moreno V, Kirchhoff T, Gold B *et al*: **Network modeling links breast cancer susceptibility and centrosome dysfunction**. *Nature genetics* 2007, **39**(11):1338-1349.
14. Chuang HY, Lee E, Liu YT, Lee D, Ideker T: **Network-based classification of breast cancer metastasis**. *Molecular systems biology* 2007, **3**:140.
15. Walther A, Johnstone E, Swanton C, Midgley R, Tomlinson I, Kerr D: **Genetic prognostic and predictive markers in colorectal cancer**. *Nature reviews Cancer* 2009, **9**(7):489-499.
16. Goymer P: **Cancer genetics - Networks uncover new cancer susceptibility suspect**. *Nature Reviews Genetics* 2007, **8**(11):823-823.
17. Wu X, Chen JY: **Molecular Interaction Networks: Topological and Functional Characterizations**. *Automation in Proteomics and Genomics: An Engineering Case-Based Approach* 2009:145.
18. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T: **Cytoscape: a software environment for integrated models of biomolecular interaction networks**. *Genome research* 2003, **13**(11):2498-2504.
19. JA R, JL M, JW P, F Y, P R, PB D: **Modulation of NF-κB-dependent gene transcription using programmable DNA minor groove binders**. *Proc Natl Acad Sci U S A* 2011, **Dec 27**.

20. Poole EM, Curtin K, Hsu L, Kulmacz RJ, Duggan DJ, Makar KW, Xiao L, Carlson CS, Slattery ML, Caan BJ *et al*: **Genetic variability in EGFR, Src and HER2 and risk of colorectal adenoma and cancer**. *International journal of molecular epidemiology and genetics* 2011, **2**(4):300-315.
21. Tang FY, Pai MH, Chiang EP: **Consumption of high-fat diet induces tumor progression and epithelial-mesenchymal transition of colorectal cancer in a mouse xenograft model**. *The Journal of nutritional biochemistry* 2012.
22. Holtzman DM, Morris JC, Goate AM: **Alzheimer's Disease: The Challenge of the Second Century**. *Science Translational Medicine* 2011, **3**(77):77sr71.
23. Hampel H, Frank R, Broich K, Teipel SJ, Katz RG, Hardy J, Herholz K, Bokde ALW, Jessen F, Hoessler YC: **Biomarkers for Alzheimer's disease: academic, industry and regulatory perspectives**. *Nature Reviews Drug Discovery* 2010, **9**(7):560-574.
24. Ray S, Britschgi M, Herbert C, Takeda-Uchimura Y, Boxer A, Blennow K, Friedman LF, Galasko DR, Jutel M, Karydas A *et al*: **Classification and prediction of clinical Alzheimer's diagnosis based on plasma signaling proteins**. *Nature medicine* 2007, **13**(11):1359-1362.
25. Gomez Ravetti M, Moscato P: **Identification of a 5-protein biomarker molecular signature for predicting Alzheimer's disease**. *PloS one* 2008, **3**(9):e3111.
26. Cotta C, Sloper C, Moscato P: **Evolutionary search of thresholds for robust feature set selection: application to the analysis of microarray data**. *Applications of Evolutionary Computing* 2004:21-30.
27. Witten IH, Frank E: **Data Mining: Practical machine learning tools and techniques**: Morgan Kaufmann Pub; 2005.
28. de Paula MR, Ravetti MG, Berretta R, Moscato P: **Differences in Abundances of Cell-Signalling Proteins in Blood Reveal Novel Biomarkers for Early Detection Of Clinical Alzheimer's Disease**. *PloS one* 2011, **6**(3):e17481.
29. Dash M, Liu H: **Feature selection for classification**. *Intelligent data analysis* 1997, **1**(3):131-156.
30. Guyon I, Elisseeff A: **An introduction to variable and feature selection**. *The Journal of Machine Learning Research* 2003, **3**:1157-1182.
31. Hall MA, Holmes G: **Benchmarking attribute selection techniques for discrete class data mining**. *IEEE Transactions on Knowledge and Data engineering* 2003:1437-1447.
32. Saeys Y, Inza I, Larrañaga P: **A review of feature selection techniques in bioinformatics**. *Bioinformatics* 2007, **23**(19):2507.
33. Kastenmuller G, Schenk ME, Gasteiger J, Mewes HW: **Uncovering metabolic pathways relevant to phenotypic traits of microbial genomes**. *Genome biology* 2009, **10**(3):R28.
34. Shin H, Markey MK: **A machine learning perspective on the development of clinical decision support systems utilizing mass spectra of blood samples**. *Journal of biomedical informatics* 2006, **39**(2):227-248.
35. Ma S, Huang J: **Penalized feature selection and classification in bioinformatics**. *Briefings in bioinformatics* 2008, **9**(5):392-403.
36. Witten IH, Frank E, Hall MA: **Data mining practical machine learning tools and techniques**. In: *[Morgan Kaufmann series in data management systems]*. 3rd edn. Burlington, MA: Morgan Kaufmann; 2011: xxxiii, 629 p.

37. Kim Y, Koo I, Jung BH, Chung BC, Lee D: **Multivariate classification of urine metabolome profiles for breast cancer diagnosis**. *BMC bioinformatics* 2010, **11 Suppl 2**:S4.

38. Hall MA, Holmes G: **Benchmarking attribute selection techniques for discrete class data mining**. *IEEE transactions on knowledge and data engineering* 2003, **15**(6):1437-1447.

39. Wang Y, Makedon F, Pearlman J: **Tumor classification based on DNA copy number aberrations determined using SNP arrays**. *Oncology reports* 2006, **15 Spec no.**:1057-1059.

40. Robnik-Sikonja M, Kononenko I: **Theoretical and empirical analysis of ReliefF and RReliefF**. *Mach Learn* 2003, **53**(1-2):23-69.

41. Landwehr N, Hall M, Frank E: **Logistic model trees**. *Mach Learn* 2005, **59**(1-2):161-205.

42. Adam BL, Qu Y, Davis JW, Ward MD, Clements MA, Cazares LH, Semmes OJ, Schellhammer PF, Yasui Y, Feng Z *et al*: **Serum protein fingerprinting coupled with a pattern-matching algorithm distinguishes prostate cancer from benign prostate hyperplasia and healthy men**. *Cancer research* 2002, **62**(13):3609-3614.

43. Tian J, Wu N, Guo X, Guo J, Zhang J, Fan Y: **Predicting the phenotypic effects of non-synonymous single nucleotide polymorphisms based on support vector machines**. *BMC bioinformatics* 2007, **8**:450.

44. Buxbaum JD, Koo EH, Greengard P: **Protein phosphorylation inhibits production of Alzheimer amyloid beta/A4 peptide**. *Proceedings of the National Academy of Sciences of the United States of America* 1993, **90**(19):9195-9198.

45. Hung AY, Haass C, Nitsch RM, Qiu WQ, Citron M, Wurtman RJ, Growdon JH, Selkoe DJ: **Activation of protein kinase C inhibits cellular production of the amyloid beta-protein**. *The Journal of biological chemistry* 1993, **268**(31):22959-22962.

46. Thathiah A, De Strooper B: **The role of G protein-coupled receptors in the pathology of Alzheimer's disease**. *Nature reviews Neuroscience* 2011, **12**(2):73-87.

47. Luchsinger JA, Cheng D, Tang MX, Schupf N, Mayeux R: **Central Obesity in the Elderly is Related to Late-onset Alzheimer Disease**. *Alzheimer disease and associated disorders* 2011.

48. Markiewski MM, Nilsson B, Ekdahl KN, Mollnes TE, Lambris JD: **Complement and coagulation: strangers or partners in crime?** *Trends in immunology* 2007, **28**(4):184-192.

49. Moloney AM, Griffin RJ, Timmons S, O'Connor R, Ravid R, O'Neill C: **Defects in IGF-1 receptor, insulin receptor and IRS-1/2 in Alzheimer's disease indicate possible resistance to IGF-1 and insulin signalling**. *Neurobiology of aging* 2010, **31**(2):224-243.

50. Gabrilovich DI, Nagaraj S: **Myeloid-derived suppressor cells as regulators of the immune system**. *Nature reviews Immunology* 2009, **9**(3):162-174.

51. Haverkamp JM, Crist SA, Elzey BD, Cimen C, Ratliff TL: **In vivo suppressive function of myeloid-derived suppressor cells is limited to the inflammatory site**. *European journal of immunology* 2011, **41**(3):749-759.

52. O'Connell RM, Rao DS, Chaudhuri AA, Baltimore D: **Physiological and pathological roles for microRNAs in the immune system**. *Nature reviews Immunology* 2010, **10**(2):111-122.

53. Thomson JM, Parker J, Perou CM, Hammond SM: **A custom microarray platform for analysis of microRNA gene expression**. *Nature Methods* 2004, **1**(1):47-53.

54. Bentwich I: **Prediction and validation of microRNAs and their targets**. *FEBS letters* 2005, **579**(26):5904-5910.

55. Betel D, Koppal A, Agius P, Sander C, Leslie C: **Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites**. *Genome biology* 2010, **11**(8):R90.

56. Selbach M, Schwanhausser B, Thierfelder N, Fang Z, Khanin R, Rajewsky N: **Widespread changes in protein synthesis induced by microRNAs**. *Nature* 2008, **455**(7209):58-63.

57. Friedman RC, Farh KKH, Burge CB, Bartel DP: **Most mammalian mRNAs are conserved targets of microRNAs**. *Genome research* 2009, **19**(1):92.

58. Maragkakis M, Reczko M, Simossis V, Alexiou P, Papadopoulos G, Dalamagas T, Giannopoulos G, Goumas G, Koukis E, Kourtis K: **DIANA-microT web server: elucidating microRNA functions through target prediction**. *Nucleic acids research* 2009, **37**(suppl 2):W273.

59. John B, Enright AJ, Aravin A, Tuschl T, Sander C, Marks DS: **Human microRNA targets**. *PLoS biology* 2004, **2**(11):e363.

60. Papadopoulos G, Alexiou P, Maragkakis M, Reczko M, Hatzigeorgiou A: **DIANA-mirPath: Integrating human and mouse microRNAs in pathways**. *Bioinformatics* 2009, **25**(15):1991.

61. Chowbina SR, Wu X, Zhang F, Li PM, Pandey R, Kasamsetty HN, Chen JY: **HPD: an online integrated human pathway database enabling systems biology studies**. *BMC bioinformatics* 2009, **10 Suppl 11**:S5.

62. Chowbina S, Deng Y, Ai J, Wu X, Guan X, Wilbanks M, Escalon B, Meyer S, Perkins E, Chen J: **A new approach to construct pathway connected networks and its application in dose responsive gene expression profiles of rat liver regulated by 2, 4DNT**. *BMC Genomics* 2010, **11**(Suppl 3):S4.

63. Chen JY, Mamidipalli S, Huan T: **HAPPI: an online database of comprehensive human annotated and predicted protein interactions**. *BMC Genomics* 2009, **10 Suppl 1**:S16.

64. Zhang B, Kirov S, Snoddy J: **WebGestalt: an integrated system for exploring gene sets in various biological contexts**. *Nucleic acids research* 2005, **33**(suppl 2):W741.

65. Rho K, Kim B, Jang Y, Lee S, Bae T, Seo J, Seo C, Lee J, Kang H, Yu U: **GARNET–gene set analysis with exploration of annotation relations**. *BMC bioinformatics* 2011, **12**(Suppl 1):S25.

CURRICULUM VITAE

MADHANKUMAR SONACHALAM
714 Blake Street, Apt L, Indianapolis, IN 46202, (631)741-6353, madhanprema@gmail.com

CAREER OBJECTIVE

To obtain a responsible and challenging position with a Life sciences industry where my technical expertise in both software development and bioinformatics can be applied to solve interdisciplinary problems in biologyx

SUMMARY OF QUALIFICATIONS

- Over 5 years of focused IT experience with a proven track record in all the phases of Software development life cycle (SDLC) and Configuration of distributed environments for Fortune 500 companies
- Strong programming skills in Java, J2EE, PERL, Python, R, MATLAB, Hibernate, JDBC, JavaScript, Jakarta Struts, Spring, JSP, Servlets, HTML, XML, Web logic, JBoss, Tomcat, Oracle PL/SQL and experience in developing multi-tier Client-Server and Object Oriented Technologies in Internet/intranet environments
- Proficiency in processing documentation; used for Use cases, Unit & Integration Test cases, Functional Requirement Documents (FRD), System Requirement Specifications (SRS) and Requirement Traceability matrix (RTM)
- Adroit at High throughput Genomics and Proteomics analysis, Next generation sequencing technologies, Network generation and visualization, R, Bioconductor and various Statistical, artificial intelligence techniques
- Sound practical and theoretical knowledge in various domains of Biology viz, Genetics, Microbiology, Cell and Molecular biology

EDUCATION

Master of Science, Bioinformatics, Dean's List	DEC 2011
Indiana University, School of Informatics, Indianapolis, IN	GPA: 3.9/4.0
Bachelor of Engineering, Biotechnology	MAY 2006
Anna University, India	GPA:8.12/10.0
Sun Certified Java Programmer (SCJP)	DEC 2007
Sun Microsystems, USA	
Cognizant Certified Professional in Java and J2EE	JAN 2007
Teaneck, NJ, USA	

COMPUTATIONAL PROFICIENCY

Languages:	Java, J2EE, Java scripts, Perl, Python, PL SQL, PHP
Packages:	Matlab, R and Bioconductor
RDBMS:	Oracle 10g, 11g, My SQL
Development tools:	Eclipse, IBM Rational Application Developer (RAD), RSM
Web Frameworks:	Jakarta Struts MVC, Spring, Hibernate, Apache FOP
Web/App Servers:	Apache, Jakarta Tomcat, JBoss, Bea Weblogic, IBM Websphere
Version control:	Rational Clear Case, CVS, PVCS
Java Skills:	JSP, Servlets, HTML, XML, CSS, JNDI, JUnit, Javabeans, JavaMail
Database tools:	Aqua Data Studio, SQL Navigator, SQL Developer, Oracle APEX
Operating System:	Windows, Windows server 2003, UNIX

BIOINFORMATICS PROFICIENCY

Microarray analysis:	MeV, SAM, Array track, GSEA, Bioconductor, GenePattern, Babelomics
Database search:	SRS, Entrez, BLAST and PSI-BLAST
Statistical packages:	R, SPSS, SAS
Visualization tools:	Cytoscape, MetaCore, GeneGo
Data Mining:	WEKA, LibSVM, Adminer, Oracle Data Miner (ODM), CD-Hit
MiRNA Resources:	mirBase, miRecords, Tarbase, Pictar, Diana MicroT, miRpath, miRExtra
Enrichment tools:	David, GOmapper, WebGestalt, GenMAPP, GARNET
Other Analysis:	NGS Techniques, Network Generation, Downstream pathway analysis

PROFESSIONAL EXPERIENCE

GRADUATE RESEARCH ASSISTANT (SEP '10 - PRESENT)

INDIANA CENTER FOR SYSTEM BIOLOGY AND PERSONALIZED MEDICINE, IN

- Worked in various computational system biology projects that requires strong programming and bioinformatics skills
- Analyzed mRNA and miRNA microarray data to build a miRNA and gene interaction network, correlated the miRNA with their target genes at pathway level to identify the significant pathways that involves in T cell suppression.
- Reviewed the existing miRNA specific databases, worked with miRNA expression profiling data, used functional enrichment and pathway analysis tools and build the network
- Evaluated various data mining tools and compared the efficiency of various feature selection algorithms (ReliefF, SVM, OneR, Elastic net, Lasso) with wide range of classifiers (SVM, Bayesian, Logistic Model Tree, Decision Table, Random Forest) using

the high throughput Proteomics data from Multiplex Proteomic Immunoassay Panel for Alzheimer's disease

- Performed exploratory analysis, Hypothesis testing and various statistical tests between the control and Type II Diabetes clinical data set using SAS
- Built a 3D network terrain model for cancer early detection, sub-classification, diagnosis, and prognosis using breast cancer and prostate cancer as case studies
- Served as database administrator and supported database projects developed using Oracle Application Express
- The projects involved expertise in Oracle 11g, Oracle SQL Developer, Oracle APEX, PERL, R, MATLAB, UNIX Shell scripting, High throughput Genomic and Proteomic Analysis, Network Generation and visualization, Data Mining techniques, WEKA, Oracle Data Miner, Cancer genomics and Immunology

PCR DATABASE DEVELOPER INTERN (MAY '10 - AUG '10)
LIFE TECHNOLOGIES FOSTER CITY, CA

- Part of the Algorithms and software development group which plays critical role in developing algorithms to automate detection, quantify and to improve precision and accuracy in gene expression and genotyping
- Worked as a primary developer of database, input/output converter tool to store Real Time (RT) PCR data produced from diverse instruments viz StepOne, StepOne Plus, 7500, 7900 and Paragon
- Interacted closely with scientists and Engineers to gather requirements and integrated the file readers for all the instruments in to a single application developed using Jakarta Struts Model View Controller (MVC) pattern
- Designed and developed database using Oracle PL/SQL and Graphical User Interface (GUI) using HTML, CSS, JSP and JavaScript that provide the advance search capabilities to the user

PROJECT LEAD/ SENIOR SOFTWARE DEVELOPER (AUG '06 - DEC'09)

COGNIZANT TECHNOLOGY SOLUTIONS CHENNAI,INDIA

- Exposure to all the areas of SDLC starting from requirement analysis to project delivery including post production support and maintenance in methodologies like Waterfall, V, Agile and Spiral
- Interacted with clients in gathering requirements and lead the team to develop prototypes using MS Visio and HTML
- Successfully implemented the open source modules developed using Apache POI and Apache FOP to the applications
- Designed the functional specifications and architecture of the web-based module and developed applications based on Jakarta Struts and Spring MVC framework
- Extensively worked in database design and development using Oracle PL/SQL and its integration with the front end

- Involved in code refactoring, Migration of Java 1.4 to 1.5, Ant to maven, created web sphere configuration files & project deployment scripts
- Lead a five member team in requirement, design and development phase of the project developed using Struts MVC, JavaScript, CSS and MySQL
- Conducted training and presentations to UAT and Operation groups for the web applications and their interfaces with the core product, Java and J2EE technologies to the incoming team members
- As an active member of Cognizant Training Academy, conducted classroom training and tests for incoming freshers on Java, J2EE technologies, UML and Design patterns
- Part of the Cognizant R&D team to develop Aetna Enterprise Framework service, a Java based framework

HONORS

- Secured Achiever of the Month award at Cognizant Technology Solutions, February, 2007 for outstanding contribution in design and development of a project developed for Aetna, a leading Health insurer in US
- Four international publication, a book chapter to the credit and received the 2010 Most Cited Paper Award from Pattern Recognition letters Journal for my very first publication

POSTERS/PUBLICATIONS

- Piyush Mundra, Madhan Kumar, K.Krishna Kumar, V.K Jayaraman, B.D.Kurkarni. (2007)."Using Pseudo Amino Acid Composition to Predict Protein Sub nuclear Localization: Approached with PSSM". Pattern Recognition Letters, 28 1610-1615
- Xiaogang Wu, Hui Huang, Madhankumar Sonachalam, Sina Reinhard, Jeffrey Shen, Jake Y. Chen "Reordering Based Integrative Expression Profiling for Microarray Classification", Bioinformatics, Accepted
- Madhankumar Sonachalam, Jeffrey Shen, Hui Huang, Xiaogang Wu. Systems biology approach to identify gene network signatures for colorectal cancer, Frontiers in System Biology, Accepted
- Madhankumar Sonachalam, Xiaogang Wu, Sungeun Kim, Andrew J Saykin, Li Shen, Jake Y Chen, and Alzheimer's Disease Neuroimaging Initiative "Identifying Plasma-Based Subnetwork Signatures for Alzheimer's disease using a Multiplex Proteomic Immunoassay Panel in Alzheimer's Disease Neuroimaging Initiative cohort", Bioinformatics, Submitted for review
- Madhankumar Sonachalam , Hui Huang, Xiaogang Wu, Ragini Pandey, Jake Y. Chen, PAGED: an integrated pathway and gene enrichment database enabling molecular phenotype discoveries, Manuscript in preparation
- Hui Huang, Madhankumar Sonachalam, Xiaogang Wu, Fan Zhang, Jake Y. Chen "Computational Biomarker Discovery in Cancer: From systems biology to predictive and personalized medicine applications", Cancer Research Day 2010, IU Simon Cancer Center