Identify Condition Specific Gene Co-expression Networks

THESIS

Presented in Partial Fulfillment of the Requirements for the Degree Master of Science in the Graduate School of The Ohio State University

By

Vikram Gajanan Kalluru

Graduate Program in Electrical and Computer Science

The Ohio State University

2012

Thesis Committee:

Dr. Kun Huang, Advisor

Dr. Raghu Machiraju

Abstract

Since co-expressed genes often are co-regulated by a group of transcription factors, different conditions (e.g., disease versus normal) may lead to different transcription factor activities and therefore different co-expression relationships. A method for identifying condition specific co-expression networks by combining the recently developed network quasi-clique mining algorithm and the Expected Conditional F-statistic has been proposed. This method has been applied to compare the transcriptional programs between the non-basal and basal types of breast cancers. This work is a translational bioinformatics study integrating network analysis which lifts the traditional gene list based disease biomarker discovery to the gene and protein interaction level.

This work presents a method for identifying condition specific gene co-expression networks. The method involves construction of a Weighted Graph Co-expression Network (WGCN) and mining the WGCNs to identify dense co-expression networks followed by a chi-square test based enrichment analysis for detecting condition specific co-expression relationship. The expression values in all the conditions for the genes constituting a condition specific co-expression network are visualized as heat maps which suggest that the genes are highly correlated in a specific condition but the correlations are disrupted in other conditions.

Dedication

This document is dedicated to my late grandparents without whom I would not be what I am today

Acknowledgments

First and foremost I would like to thank and acknowledge the support and guidance of my advisors Dr. Kun Huang and Dr. Raghu Machiraju. I express my gratitude for their invaluable support, guidance and advice. Apart from just research there are several other things more important to life that I have learnt from them. It has been an invaluable experience working with them.

At this point I take the opportunity to thank my lab mate Raghuram for his invaluable guidance and support throughout my two years at Ohio State. I would also like to thank my other lab mates Sundar Raman, Dr. Shantanu Singh, Chao Wang, Nan Meng and Hao Ding for all their help and support.

Finally I would like to thank all my friends and Ohio State University for these two challenging years and not so easy life. The problems, challenges, impediments, hurdles, unfavorable conditions made me understand and appreciate life and what it takes to succeed at the highest level.

Vita

April 2009 .....B.Eng. Electronics and Communication, Anna University

September 2010 ……Project Engineer, Central Electronic Engineering Research Institute

June 2012 .....M.S. Electrical & Computer Engineering, The Ohio State University

Publications

Kalluru V, Machiraju R, Huang K. Identify Condition Specific Gene Co-expression Networks, accepted to *ICIBM, 2012* (To appear in International Journal)

Fields of Study

Major Field: Electrical and Computer Engineering

Table of Contents

List of Tables

List of Figures

Chapter 1 - Introduction

Gene co-expression network analysis is a widely adopted bioinformatics method in biomedical research with many applications including discovering protein-protein interaction relationships [1-4], predicting new gene functions [5, 6] and pathways [7], and identifying disease biomarkers or genes [8-13]. In addition, many algorithms have been developed to identify gene modules or networks composed of highly co-expressed genes [3, 11, 14, 15]. These co-expression networks can be used as quantitative traits or disease biomarkers [11, 16-18].

However, a typical gene co-expression relationship is not static, it changes given different biological or disease conditions. For instance, in [19], it has been shown that the correlation between the *ARG2* and *CAR2* genes in yeast changes from positive to negative as the expression level of another gene (*CPA2*) changes. In addition, changes of co-expression relationships between gene pairs have been detected in cancers [20, 21]. In particular, condition specific co-expression relationship reveals condition specific potential biological mechanisms.

In this project, an expansive search for condition-specific gene co-expression between gene pairs is conducted to discover condition specific gene co-expression networks.

Since co-expressed genes often are co-regulated by a group of transcription factors, different conditions (e.g., disease versus normal) may lead to different transcription factor activities and therefore different co-expression networks. In this thesis a method has been proposed for identifying condition-specific co-expression networks by combining gene co-expression network mining and the Expected Conditional F-statistic (ECF) developed in [20] for evaluating changes in the co-expression relationship among different conditions.

Specifically, the recently developed network quasi-clique mining algorithm for weighted gene co-expression networks is used. In gene co-expression networks, Pearson correlation coefficient (PCC) is often used as the metric to measure co-expression between two genes in a microarray dataset [3, 5, 22]. A weighted graph can be established by setting genes as vertices and PCC values (or their absolute values) as weights of the edges. In some network mining algorithms, a threshold is imposed on the PCC values to derive an un-weighted network such that two genes are only connected by an edge if the PCC value between them is higher than a pre-defined threshold [3]**.** The un-weighted gene co-expression network approach has several drawbacks including the selection of the threshold which may be too rigid for weights around that threshold. Therefore weighted GCN (WGCN) method has been recently widely adopted and a series of tools have been developed to identify networks from WGCNs using hierarchical clustering based approach [8, 14, 22]. However**,** this approach often identifies large clusters (e.g., with hundreds or even thousands of genes), cannot directly control the intra-cluster connectivity and it does not allow shared genes between two clusters even

though many genes have multiple functions. Recently a WGCN dense network finding algorithm named eQCM [11] which is a derivation of the quasi-clique merging (QCM) algorithm [23] has been developed. This algorithm guarantees a lower bound on the densities of the identified networks and allows overlaps between networks. In this paper, the eQCM algorithm is used to detect co-expression networks in each condition and identify networks which are enriched with edges showing large weight changes between different conditions as measured by the ECF statistics, which is a metric for evaluating changes of correlation relationship in different conditions [20].

The developed method is applied to compare the transcriptional programs between two subtypes of breast cancers, namely the non-basal and basal types of breast cancers which are well known for their different molecular markers and prognosis in patients. This work is a translational bioinformatics study integrating network analysis which lifts the traditional gene list based disease biomarker discovery to the gene and protein interaction level.

Chapter 2 – The Expected Conditional F-Statistic

To understand cancer it is important to explore molecular changes in cellular processes from normal state to cancerous state. Differentially expressed genes are potential markers for clinical diagnoses and medical treatments. The F-statistic and its variants like t-test, signal-to-noise statistic and SAM method are commonly used to identify differentially expressed genes. A clustering algorithm may be used to find groups of genes that behave similarly across a dataset. However all these methods may miss groups of genes which form differential co-expression patterns under different experimental conditions.

Statistical tests such as the t-test or ANOVA, identify genes that are differentially expressed under one or more conditions. The output of such tests is a simple list of genes, with an associated test statistic and p-value. There is no indication of which genes may be interacting with one another. Alternatively clustering algorithms are often used to find groups of genes which display similar expression profiles across a dataset, and these clusters are subsequently analyzed visually for patterns of interest.

However, genes which show highly correlated patterns of expression in one biological state, but not in another, may not be highly correlated across the entire dataset, and therefore would not be associated with one another if a clustering algorithm is used.

Variation may exist in this and may lead to that gene being grouped incorrectly. Furthermore, clustering algorithms do not provide methods to identify groups that are behaving differently in different biological conditions.

Identification of differentially co-expressed gene clusters or gene pairs usually do not use a priori defined gene sets or pairs but try to find the best ones among all possible combinations without considering prior knowledge. Thus the biological interpretation of the clusters or pairs may also need the ontology and pathway based annotation analysis.

There have been several methods proposed to compute differential co-expression between a pair of genes. The differential CoxS algorithm for differential co-expression analysis of paired gene sets between conditions has the benefits of both differential co-expression and gene set-wise analyses [29]. Kostka and Spang [27] described a method to investigate differentially co-expressed groups of genes, using an additive model for scoring gene-gene co-expression and then a stochastic search algorithm to find groups of genes showing differential co-expression patterns. Jen *et al* [28] developed ACT, the Arabidopsis Co-expression Tool, which allows users to calculate co-expression across user-defined data sets and uses a correlation cut-off to define groups of genes.

We use Expected Conditional F-Statistic (ECF) an extension of the F-Statistic to identify differential gene co-expression in this work. While the F-statistic is a widely used method to test whether a gene is differentially expressed, ECF operates on a pair of genes and evaluates the difference in correlation of the two genes across different conditions. It is

essentially a method to find gene pairs that are in principle positively correlated in one condition and not correlated in other. Higher the difference in correlation, higher is the ECF value.

As in [20] the ECF is given as:

$$\boldsymbol{E_Y}(\lambda_{(X|Y=y)}) = \left[\sum_i p_i\,(1 - \rho_i^2)\right]^{-1} \sum_k \sum_{i<j} p_k\, p_i\, p_j \left\{\left[\frac{\mu_{Xi} - \mu_{Xj}}{\sigma_X} - \frac{\rho_i(\mu_{Yi} - \mu_{Yk})}{\sigma_Y} + \frac{\rho_j(\mu_{Yj} - \mu_{Yk})}{\sigma_Y}\right]^2 + (\rho_i - \rho_j)^2\right\}$$

where $\lambda_{(X|Y=y)}$ is the conditional F-statistic and taking its expectation we get the ECF

ρ_i is the Pearson Correlation Coefficient and is given as

$$PCC_i = \frac{\sum_{j=1}^{n_i}(x_{ij} - \widehat{\mu_{Xi}})(y_{ij} - \widehat{\mu_{Yi}})}{\sqrt{\sum_{j=1}^{n_i}(x_{ij} - \widehat{\mu_{Xi}})^2} * \sqrt{\sum_{j=1}^{n_i}(y_{ij} - \widehat{\mu_{Yi}})^2}}$$

in which

the estimated mean is given as $\widehat{\mu_{Xi}} = \frac{1}{n_i} * \sum_{j=1}^{n_i} x_{ij}$,

and the estimated variance is given as $\widehat{\sigma_{Xi}^2} = \frac{1}{n_i - 1} * \sum_{j=1}^{n_i}(x_{ij} - \widehat{\mu_{Xi}})^2$

Chapter 3- Materials and Methods

The method includes the following major steps as outlined in the following workflow:

1. Pre-processing of the microarray data.

For a gene expression (microarray) dataset with multiple samples, it is normalized using standard microarray data normalization algorithms. For Affymetrix GeneChip data, they are normalized using the Robust Multi-array Analysis (RMA) algorithm for normalization [30]. For any gene with multiple probesets in the microarray, the values from the probeset with the highest mean expression value is used to represent that gene as suggested by [26].

2. Construct WGCN for each condition.

First compute the Pearson correlation coefficients (PCC) between every pair of genes in the specific condition and then apply our recently developed weighted graph quasi-clique mining algorithm eQCM to identify tightly co-expressed gene networks [11]. For a pair of genes (X, Y) the Pearson Correlation Coefficient in the i^{th} group is given as:

$$PCC_i = \frac{\sum_{j=1}^{n_i}(x_{ij} - \widehat{\mu_{Xi}})(y_{ij} - \widehat{\mu_{Yi}})}{\sqrt{\sum_{j=1}^{n_i}(x_{ij} - \widehat{\mu_{Xi}})^2} * \sqrt{\sum_{j=1}^{n_i}(y_{ij} - \widehat{\mu_{Yi}})^2}}$$

where x_{ij} is the j^{th} observation of X in the i^{th} group, the estimated mean $\widehat{\mu_{Xi}} = \frac{1}{n_i} * \sum_{j=1}^{n_i} x_{ij}$, and the estimated variance $\widehat{\sigma_{Xi}^2} = \frac{1}{n_i - 1} * \sum_{j=1}^{n_i}(x_{ij} - \widehat{\mu_{Xi}})^2$. The WGCN is constructed as a weighted graph G (V, E, W) in which the PCC values constitute the edge weights W between the nodes (genes).

3. Identify dense co-expression networks from the WGCNs.

The eQCM algorithm is applied to the G (V, E, W) to identify dense networks. For a network of k nodes, the density of the network is defined as $d = \frac{2\sum_{i=1}^{k}\sum_{j=1, j\neq i}^{k} W_{ij}}{k(k-1)}$, where W_{ij} is the weight between the i-th and j-th nodes in the network. In the eQCM algorithm there are two parameters $\boldsymbol{\gamma}$ and $\boldsymbol{t}$, which all contributes to density of the detected co-expression networks [11]. In this study, we set $\boldsymbol{\gamma}$ = 0.99 and $\boldsymbol{t}$ = 1.

4. Compute the combined ECF score between every gene pair across multiple conditions. For every pair of genes, the ECF-statistics is computed as described in [21], which is essentially a metric that evaluates a change of PCC between different conditions. Specifically, ECF-statistic is given in as:

$$\boldsymbol{E_Y}(\lambda_{(X|Y=y)}) = \left[\sum_i p_i\,(1 - \rho_i^2)\right]^{-1} \sum_k \sum_{i<j} p_k\, p_i\, p_j \left\{\left[\frac{\mu_{Xi} - \mu_{Xj}}{\sigma_X} - \frac{\rho_i(\mu_{Yi} - \mu_{Yk})}{\sigma_Y} + \frac{\rho_j(\mu_{Yj} - \mu_{Yk})}{\sigma_Y}\right]^2 + (\rho_i - \rho_j)^2\right\}$$

and the combined ECF-statistic score ECF(X,Y) = $E_{\mathbf{Y}}(\lambda_{X/Y=y})$ + $E_{\mathbf{X}}(\lambda_{Y/X=x})$. A relatively high value of ECF(X,Y) signifies that the correlation between gene X and gene Y changed significantly between different conditions [20].

5. Identify high ECF gene pairs.

The graph constructed in step 1 is examined. For N genes in a graph there exists N*(N-1)/2 edges and corresponding ECF values. The ECF values are ranked in the descending order and the gene pairs with top 5% of the ECF values are selected. All such gene pairs are referred to as *high ECF gene pairs*.

6. Identify co-expression networks enriched with high ECF gene pairs.

Since the threshold is the top five percentile of the all ECF values, it is expected that on an average 5% of the edges in a co-expression network will have ECF scores above the threshold. Therefore chi-square tests can be applied to determine if a co-expression network is significantly enriched with edges with high ECF scores. For a network with ***k*** nodes, a contingency table can be derived as in Table 1. The Bonferroni method is applied to compensate for multiple tests in determining the threshold on the chi-square test p-values. Specifically, if ***M*** networks are identified in Step 3, then the p-value

threshold is set at ***0.05/M***

Table 1. An example of contingency table for the chi-square test.

# of expected high ECF gene pairs $\left(0.05 \times \frac{k(k-1)}{2}\right)$	$0.95 \times \frac{k(k-1)}{2}$
# of observed h/igh ECF gene pairs	$\frac{k(k-1)}{2}$ - # of observed high ECF gene pairs

7. Mapping genes to transcription factors.

Transcription factors are proteins that bind to specific DNA sequences, thereby controlling the expression of genes. Having identified the sub-networks enriched with high ECF edges it is further important to investigate the transcription factors that co-regulate the sub-networks enriched with high ECF genes. The gene motifs dataset from the molecular signatures database of Broad Institute was used (http://www.broadinstitute.org/gsea/msigdb/collections.jsp). It lists 614 gene motifs and their respective targets. We determine the gene motifs which are enriched with targets (genes) contained in the sub-networks enriched with high ECF edges. To determine the enriched gene motifs hyper-geometric test is used. It describes the probability of K successes in n draws from a finite population of size N containing M successes without replacement.

$$P(X=k) = [{}_{K}C_{x} * {}_{(M-K)}C_{(N-x)}]/{}_{M}C_{N}$$

In the figure below M represents the genes in the microarray dataset, K represents the genes which are the transcription factor targets (success states), N represents the genes in the condition specific sub-networks (random selection) and X represents the number of successes.

Figure 1

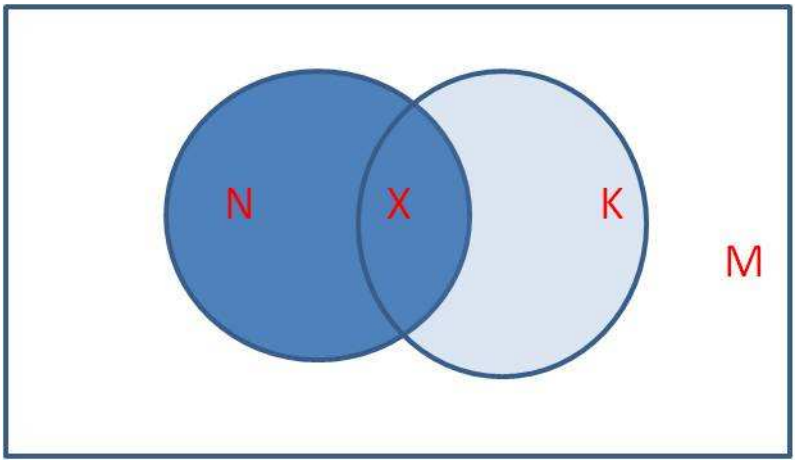


Illustration of hyper-geometric test

Once the enriched gene motifs are determined the transcription factor is extracted from it (example: ‘ER’ is extracted from ‘V$ER_07’). The extracted transcription factor is identified in the microarray dataset and its fold change over different conditions and the respective p-values (t-test) are computed and reported

Chapter 4 - Results

This method has been applied to a breast cancer gene expression dataset from the NCBI Gene Expression Omnibus with accession numbers GDS2250 which contains human samples for control, non-basal like and basal like human breast cancer tissues. For non-basal like subtype, 171 networks were identified using eQCM algorithm and out of which 88 show enriched high ECF gene pairs ($p < 0.05/171$) suggesting extensive disruption of transcriptional programs in the other (i.e., basal) type of breast cancer. For basal-like subtype, 23 out of 99 total networks were identified. These are the subtype specific networks for further analysis including gene ontology/pathway enrichment analysis and correlation with survival times. Table 2 shows examples of the networks identified for the non-basal type breast cancer samples. In addition, the heatmap of the gene expression values for the network 10 is shown in Figure 2. It is clear that the gene expression profiles show high correlations in the non-basal like breast cancers but not in the basal like breast cancers.

An important aspect of this exercise is to learn the causal mechanisms of such disruption of co-expression relationships. Here this question is investigated using the first network

in Table 2 as an example. There are 46 genes in this network. Using the gene enrichment analysis GeneCoDis2 (http://genecodis.dacya.ucm.es/), it is determined that this set of genes are highly enriched with the transcription factor LEF1 targets (hyper-geometric test p-value 3.34317×10^{-10}) as shown in Figure 3. Interestingly, in checking the original microarray data, it is found that LEF1 shows higher expression levels among the non-basal like breast cancer samples compared to the basal like breast cancer samples with log mean fold change being 1.31 (fold change 2.48) and one-tail p-value for Student t-test being measured as 0.022. This observation suggests that LEF1 plays an important role in maintaining the co-expression relationship in *network 1* in non-basal like breast cancer samples while its absence contributes to the disruption of the co-expression network in the basal like breast cancer samples. In fact, LEF1 has been shown to be involved in many different cancers including breast cancer as an important component of the cancer related **Wnt** pathway [24, 25].

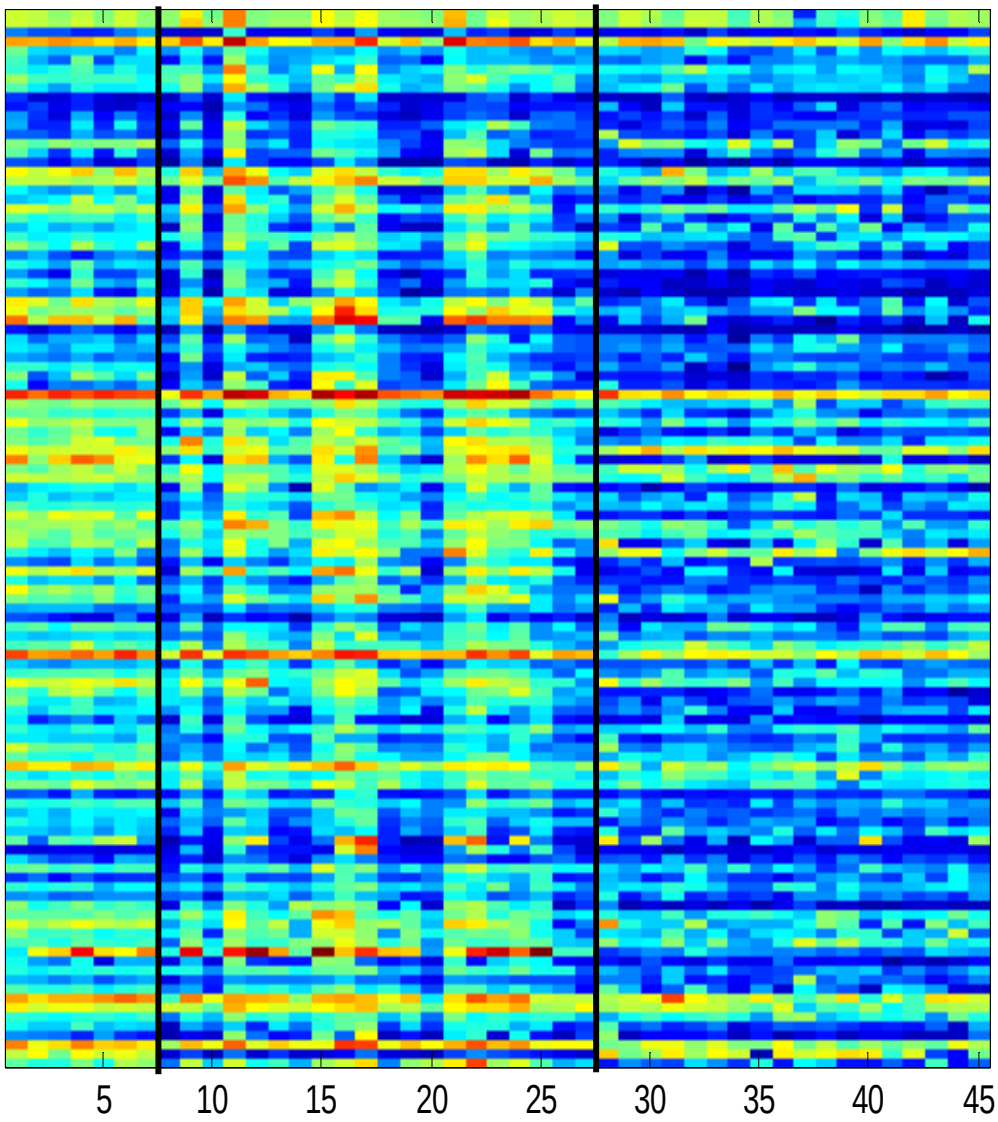
We then expand the transcription factor enrichment analysis to all the identified networks as described in the previous section. Results obtained by using this method are tabulated as shown in Table 3 and Table 4. Amongst all the transcription factors, AR which is enriched in both sub-groups, basal and non-basal, seems to be the most prominent (fold change = -1.81 and p-value = $3.66*10^{-10}$). Besides AR, GATA2 and CEBPB which are enriched only in non-basal sub-type show significant fold-change.

Table 2: Examples of the networks enriched with high ECF gene pairs for the non-basal subtype of breast cancers.

Network Index	Chi-square p-value	# of expected high ECF edges	# of observed high ECF edges
1	3.81E-08	89	167
2	0	83	342
3	6.04E-07	77	143
4	0	83	234
5	0	770	1787
6	0	92	374
7	0	322	1105
8	1.56E-05	620	742
9	5.57E-06	179	260
10	1.64E-06	158	241
11	3.52E-08	107	192
12	0	289	1120
13	2.99E-08	80	156
14	0	334	1117
15	3.34E-06	107	176
16	0	150	492
17	0	413	1271
18	0	95	359
19	0	121	584

Figure 2: Heat map representing gene expression in a co-expression network showing high correlation among the non-basal like breast cancer samples but not in the basal-like breast cancer samples. This finding reveals potentially specific transcriptional programs associated with non-basal like breast cancer which is disrupted in the basal like breast cancer

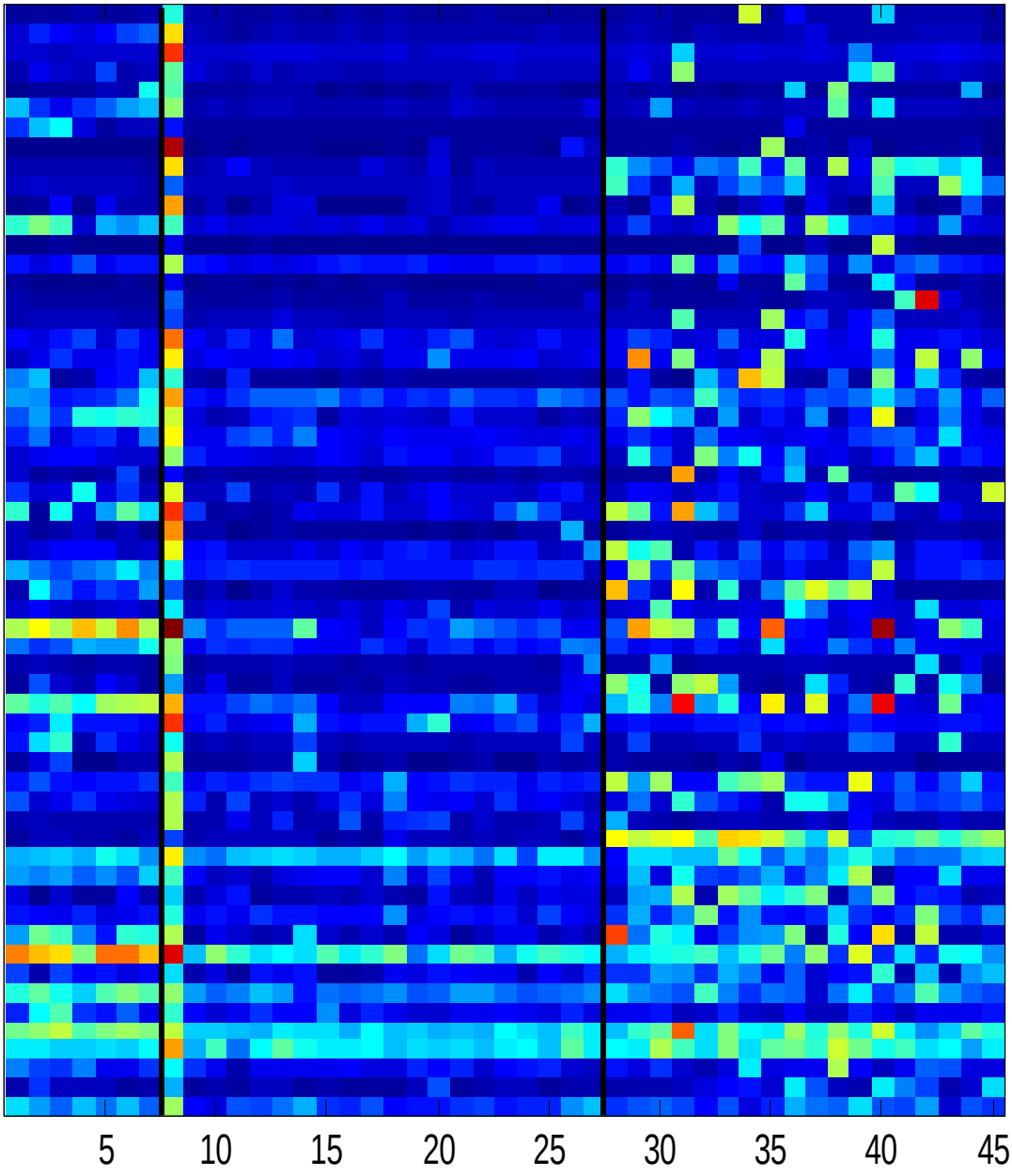
a)



continued on next page

continued from page 15

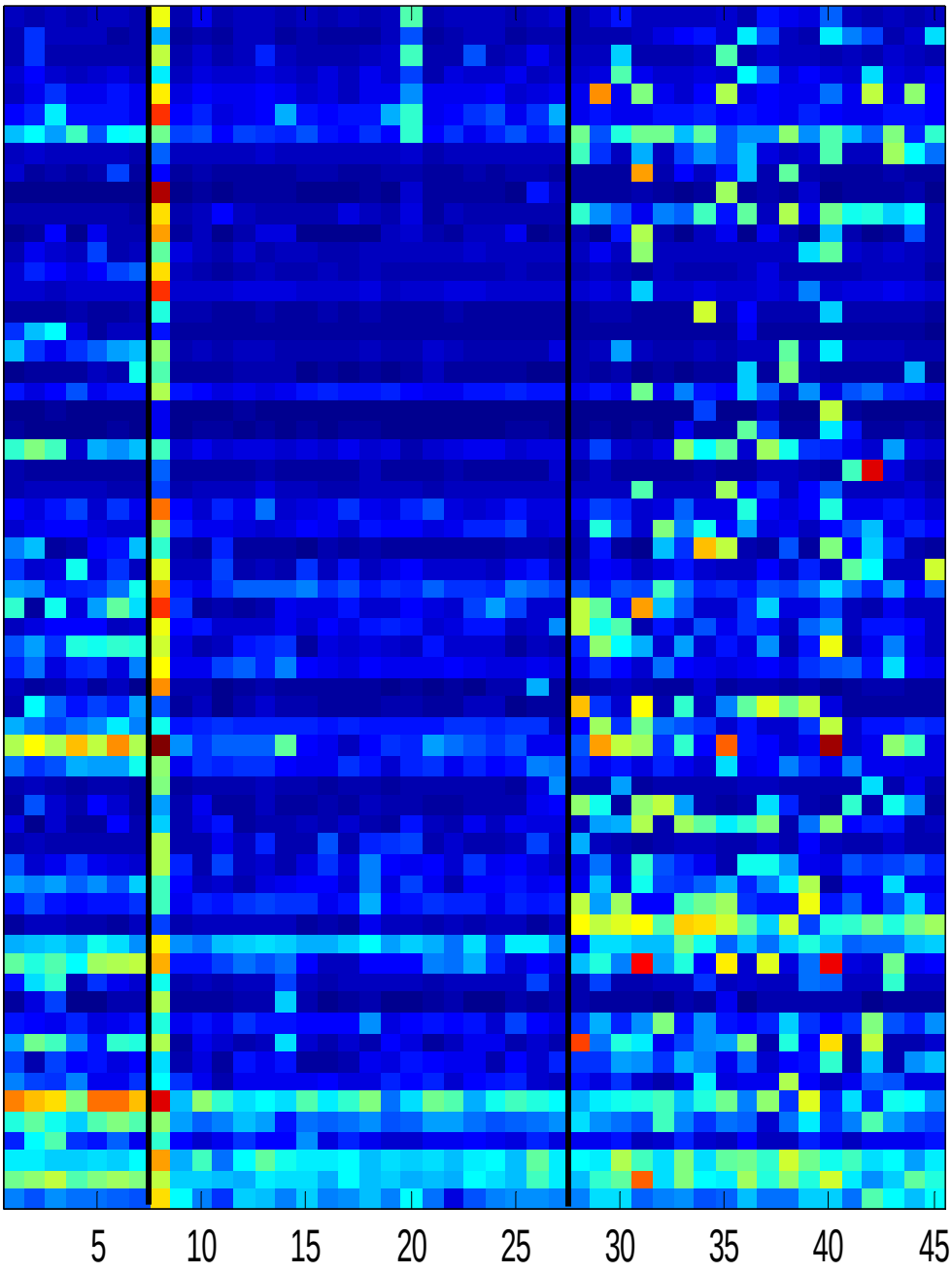
b)



continued on next page

continued from page 16

c)



continued on next page

continued from page 17

d)

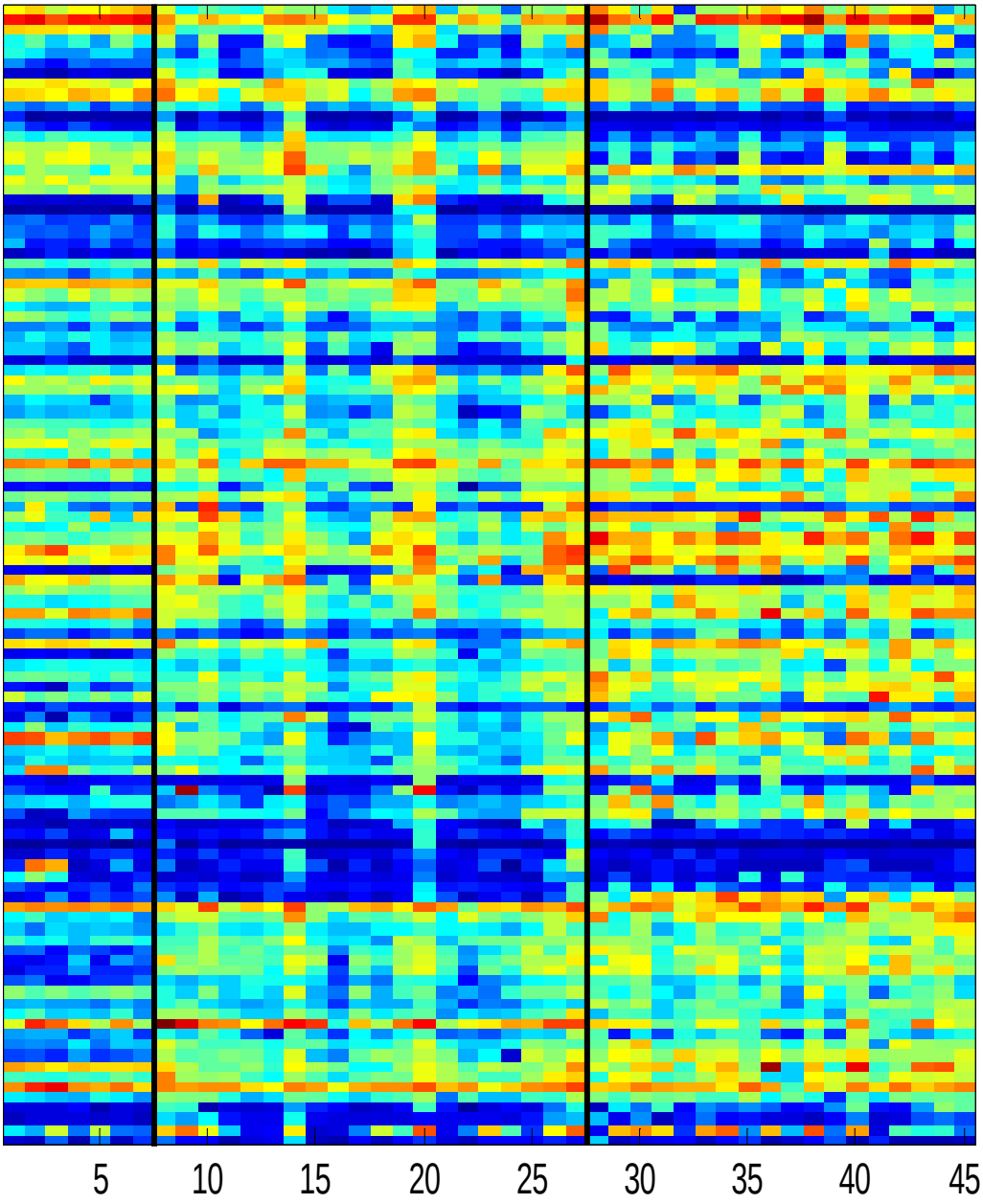


Figure 3 Top ten enriched biological terms. Top 10 enriched biological terms (including gene ontology, transcription factor, microRNA and KEGG pathways) in the 46 genes in the network 1 in Table 2. The x-axis enumerates the number of genes associated with the terms listed on the left. The transcription factor LEF1 is on the top with 18 genes associated with it.

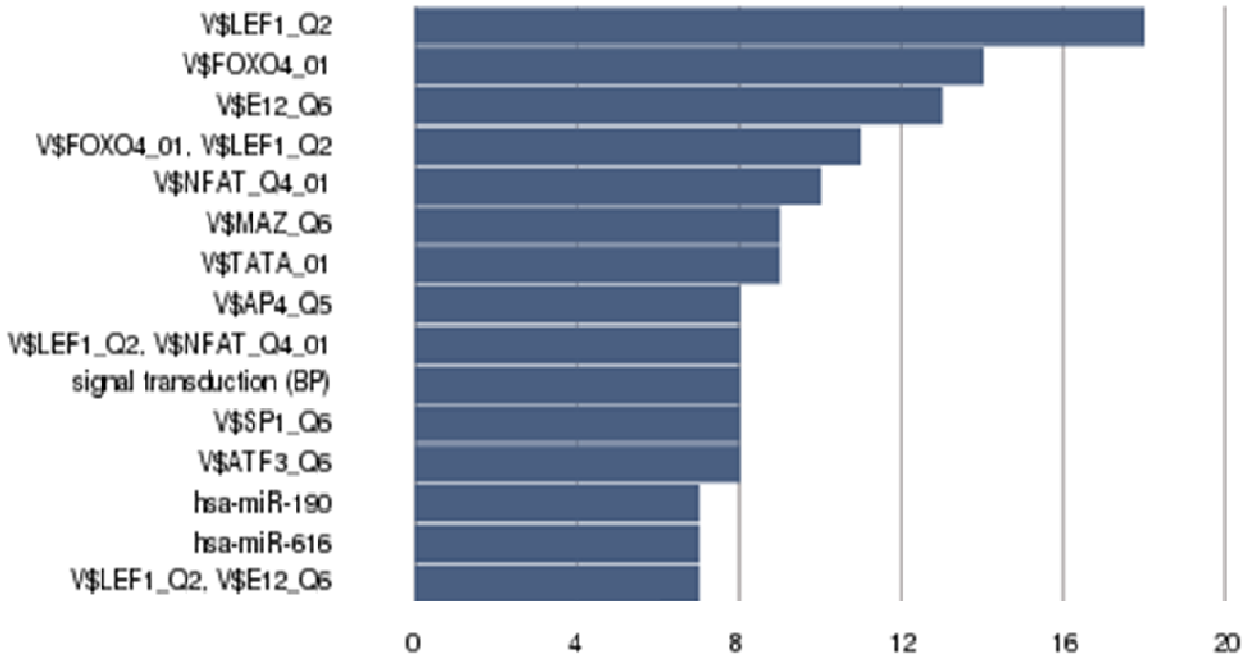


Table 3: Enriched Transcription Factors for the basal cancer subgroup

Transcription Factor	Gene Motif	Network Id	p-value	fold change	t-test p-value
STAT5B	TTCYNRGAA_V$STAT5B_01	4	0.02574	1.06	0.158
SRF	V$SRF_Q4	22	0.000004	1.08	0.0496
SRF	V$SRF_Q6	22	0.000008	1.08	0.0496
SRF	CCAWWNAAGG_V$SRF_Q4	22	0.000008	1.08	0.0496
SRF	V$SRF_Q5_01	22	0.000031	1.08	0.0496
SRF	V$SRF_C	22	0.000134	1.08	0.0496
NF1	V$NF1_Q6_01	22	0.037703	-1.02	0.584
NF1	V$NF1_Q6	22	0.040709	-1.02	0.584
STAT5B	TTCYNRGAA_V$STAT5B_01	24	0.037311	1.06	0.158
PAX2	V$PAX2_01	29	0.017916	-1.04	0.297
AR	V$AR_Q2	29	0.047996	-1.81	3.67E-10
PAX8	V$PAX8_01	54	0.034517	1.02	0.149
PAX4	V$PAX4_04	91	0.015003	-1.03	0.446
FOXD3	V$FOXD3_01	91	0.015287	1.21	0.066
VDR	V$VDR_Q6	91	0.024445	-1.09	0.00218
GATA6	V$GATA6_01	91	0.025562	1.31	0.0102

Table 4: Enriched Transcription Factors for the non-basal cancer subgroup

Transcript Factor	Gene Motif	Network ID	p-value	Fold change	p-value (t-test)
ATF3	TGAYRTCA_V$ATF3_Q6	1	0.0120	1.10383	0.06546528
ATF1	V$ATF1_Q6	1	0.0144	-1.0211	0.431335
HSF1	V$HSF1_01	1	0.01761	1.08890	0.04290051
HSF1	RGAANNTTC_V$HSF1_01	1	0.03473	1.08890	0.15733549
E2F1	V$E2F1_Q4_01	1	0.0358	1.08213	0.14605310
PAX3	CGTSACG_V$PAX3_B	1	0.04004	1.20141	0.13477072
ATF3	TGACGTCA_V$ATF3_Q6	1	0.04177	1.10383	0.12348833
MSX1	V$MSX1_01	13	0.02351	-1.09740	0.11220595
ATF1	V$ATF1_Q6	14	0.00462	-1.02113	0.10092356
ATF3	TGAYRTCA_V$ATF3_Q6	14	0.02746	1.10383	0.08964118
ATF3	TGACGTCA_V$ATF3_Q6	14	0.03759	1.10383	0.07835879
ATF1	V$ATF1_Q6	16	0.00079	-1.0211	0.06707641
ATF3	TGAYRTCA_V$ATF3_Q6	16	0.00854	1.10383	0.05579402
ATF3	TGACGTCA_V$ATF3_Q6	16	0.00943	1.10383	0.04451164
ATF3	V$ATF3_Q6	16	0.01461	1.10383	0.06546528
HSF1	V$HSF1_01	16	0.01552	1.08890	0.04290051
FOXO4	TTGTTT_V$FOXO4_01	16	0.03677	1.03702	0.07875606
E4F1	GTGACGY_V$E4F1_Q6	16	0.03996	-1.1088	0.00116162
FOXO1	V$FOXO1_01	16	0.04053	-1.06697	0.02810615
HSF1	RGAANNTTC_V$HSF1_01	16	0.04137	1.08890	0.04290051
HOXA4	V$HOXA4_Q2	16	0.04834	1.04547	0.4742465

continued on next page

continued from page 21

NF1	V$NF1_Q6_01	16	0.04834	-1.01564	0.5844932
ATF3	TGAYRTCA_V$ATF3_Q6	17	0.01926	1.10383	0.06546528
ATF1	V$ATF1_Q6	17	0.02025	-1.02113	0.431335
HSF1	V$HSF1_01	17	0.02448	1.08890	0.04290051
E2F1	V$E2F1_Q4_01	17	0.04617	1.08213	0.04129277
HSF1	RGAANNTTC_V$HSF1_01	17	0.04789	1.08890	0.0429005
PAX3	CGTSACG_V$PAX3_B	17	0.04936	1.20141	0.0188914
PAX2	V$PAX2_02	21	0.04999	-1.04258	0.2972777
SOX5	V$SOX5_01	24	0.03990	1.02546	0.3815135
NF1	V$NF1_Q6_01	24	0.04489	-1.01562	0.5844932
PBX1	V$PBX1_02	31	0.02393	-1.15393	0.0002210
PAX8	V$PAX8_01	32	0.04651	1.02327	0.148986
LEF1	V$LEF1_Q2	44	0.04749	-1.23831	0.0402204
NF1	V$NF1_Q6	51	0.02772	-1.01562	0.5844932
MSX1	V$MSX1_01	51	0.02984	-1.09740	0.1747607
ATF3	TGAYRTCA_V$ATF3_Q6	60	0.01541	1.10383	0.0654652
HSF1	V$HSF1_01	60	0.02089	1.08890	0.0429005
E2F1	V$E2F1_Q4_01	60	0.04085	1.08213	0.0412927
HSF1	RGAANNTTC_V$HSF1_01	60	0.04109	1.08890	0.0429005
PAX3	CGTSACG_V$PAX3_B	60	0.04461	1.20141	0.0188914
ATF3	TGACGTCA_V$ATF3_Q6	60	0.04747	1.10383	0.0654652
SRF	V$SRF_Q6	62	0.03264	1.08099	0.0495677
LHX3	YTAATTAA_V$LHX3_01	62	0.03657	1.02490	0.5307498

continued on next page

continued from page 22

SOX9	V$SOX9_B1	66	0.01722	1.21783	0.0010167
SRY	V$SRY_02	66	0.02056	1.05458	0.0154536
GATA1	V$GATA1_01	66	0.04071	1.00752	0.4984056
MSX1	V$MSX1_01	66	0.04917	-1.09740	0.1747607
GATA2	V$GATA2_01	89	0.04766	-1.30677	2.7500E-06
GATA2	V$GATA2_01	92	0.03616	-1.30677	2.7500E-06
HSF1	V$HSF1_01	98	0.00596	1.08890	0.0429005
E2F1	V$E2F1_Q4_01	98	0.01173	1.08213	0.0412927
ATF3	TGAYRTCA_V$ATF3_Q6	98	0.01663	1.10383	0.0654652
ATF1	V$ATF1_Q6	98	0.01819	-1.02113	0.431335
HSF1	RGAANNTTC_V$HSF1_01	98	0.04331	1.08890	0.0429005
LEF1	CTTTGA_V$LEF1_Q2	98	0.04519	-1.23831	0.0402204
PAX3	CGTSACG_V$PAX3_B	98	0.04618	1.20141	0.0188914
ATF3	TGACGTCA_V$ATF3_Q6	98	0.04943	1.10383	0.0654652
ATF1	V$ATF1_Q6	99	0.00151	-1.02113	0.431335
ATF6	V$ATF6_01	99	0.00158	1.04160	0.1132485
PAX3	CGTSACG_V$PAX3_B	99	0.00214	1.20141	0.0188914
E4F1	GTGACGY_V$E4F1_Q6	99	0.00306	-1.10887	0.0011616
HSF1	V$HSF1_01	99	0.00833	1.08890	0.0429005
SRF	V$SRF_01	99	0.00842	1.08099	0.0495677
ATF3	TGAYRTCA_V$ATF3_Q6	99	0.009393	1.103836	0.06546528
E4F1	V$E4F1_Q6	99	0.009645	-1.10887	0.00116162
E2F1	V$E2F1_Q4_01	99	0.015448	1.082136	0.04129277

continued on next page

continued from page 23

ATF3	TGACGTCA_V$ATF3_Q6	99	0.018907	1.103836	0.06546528
HSF1	RGAANNTTC_V$HSF1_01	99	0.023231	1.088903	0.04290051
SRF	V$SRF_Q4	99	0.023591	1.080993	0.04956771
ATF3	V$ATF3_Q6	99	0.027031	1.103836	0.06546528
STAT5A	V$STAT5A_04	99	0.038608	1.062279	0.1018519
FOXO4	TTGTTT_V$FOXO4_01	99	0.045005	1.037023	0.07875606
SOX5	V$SOX5_01	115	0.019848	1.025464	0.3815135
ATF3	TGAYRTCA_V$ATF3_Q6	122	0.015411	1.103836	0.06546528
ATF1	V$ATF1_Q6	122	0.017207	-1.02113	0.431335
HSF1	V$HSF1_01	122	0.020894	1.088903	0.04290051
E2F1	V$E2F1_Q4_01	122	0.040854	1.082136	0.04129277
HSF1	RGAANNTTC_V$HSF1_01	122	0.04109	1.088903	0.04290051
PAX3	CGTSACG_V$PAX3_B	122	0.044616	1.201415	0.0188914
ATF3	TGACGTCA_V$ATF3_Q6	122	0.047475	1.103836	0.06546528
GATA2	V$GATA2_01	123	0.025985	-1.30677	2.75002E-06
AR	V$AR_01	123	0.046482	-1.81346	3.66744E-10
ATF1	V$ATF1_Q6	128	0.000732	-1.02113	0.431335
HSF1	V$HSF1_01	128	0.001037	1.088903	0.04290051
SRF	V$SRF_01	128	0.001371	1.080993	0.04956771
ATF6	V$ATF6_01	128	0.002816	1.041603	0.1132485
ATF3	TGAYRTCA_V$ATF3_Q6	128	0.0029	1.103836	0.06546528
E4F1	GTGACGY_V$E4F1_Q6	128	0.003041	-1.10887	0.00116162
PAX3	CGTSACG_V$PAX3_B	128	0.003771	1.201415	0.0188914

continued on next page

continued from page 24

E4F1	V$E4F1_Q6	128	0.005012	-1.10887	0.00116162
ATF3	TGACGTCA_V$ATF3_Q6	128	0.008934	1.103836	0.06546528
SRF	V$SRF_Q4	128	0.011738	1.080993	0.04956771
FOXO1	V$FOXO1_01	128	0.012784	-1.06697	0.02810615
ATF3	V$ATF3_Q6	128	0.013887	1.103836	0.06546528
HSF1	RGAANNTTC_V$HSF1_01	128	0.01664	1.088903	0.04290051
E2F1	V$E2F1_Q4_01	128	0.024716	1.082136	0.04129277
SRF	V$SRF_C	128	0.027726	1.080993	0.04956771
FOXO4	TTGTTT_V$FOXO4_01	128	0.033669	1.037023	0.07875606
HSF2	V$HSF2_01	128	0.034274	1.047504	0.1739826
ATF4	V$ATF4_Q2	128	0.037798	1.019135	0.2626214
CEBPB	V$CEBPB_02	128	0.047262	1.101568	1.16468E-05
SRF	V$SRF_Q6	128	0.049257	1.080993	0.04956771
SOX5	V$SOX5_01	137	0.038361	1.025464	0.3815135
NF1	V$NF1_Q6_01	137	0.043201	-1.01562	0.5844932
IRF2	V$IRF2_01	148	0.023861	1.043192	0.3332626
ATF3	TGAYRTCA_V$ATF3_Q6	171	0.014242	1.103836	0.06546528
ATF1	V$ATF1_Q6	171	0.016252	-1.02113	0.431335
HSF1	V$HSF1_01	171	0.019765	1.088903	0.04290051
HSF1	RGAANNTTC_V$HSF1_01	171	0.038917	1.088903	0.04290051
E2F1	V$E2F1_Q4_01	171	0.039138	1.082136	0.04129277
PAX3	CGTSACG_V$PAX3_B	171	0.043071	1.201415	0.0188914
ATF3	TGACGTCA_V$ATF3_Q6	171	0.045544	1.103836	0.06546528

Chapter 5 - Discussion

Network-based representation and analysis of data from high-throughput experimental technologies is increasingly being used to both visualize and identify the components and their interactions involved in a given cellular system. In particular, construction of co-expression networks from gene expression microarray datasets has recently become a popular alternative to the conventional analytic approaches, such as the detection of differential expression using statistical testing or the co-expression analysis using unsupervised clustering. Representing dependencies in the dataset as interaction networks allows the researcher to explore the whole spectrum of pair wise relationships among the genes as opposed to flat lists of genes from statistical tests or distinct groups of genes from clustering tools.

While network methods are increasingly used in biology, the network vocabulary of computational biologists tends to be far more limited than that of, say, social network theorists. The relationship between network theory and the field of microarray data analysis helps to clarify the meaning of and the relationship among network concepts in gene co-expression networks. Network theory offers a wealth of intuitive concepts for describing the pair wise relationships among genes, which are depicted in cluster trees and heat maps. Conversely, high throughput microarray data analysis technologies

(singular value decomposition, tests of differential expressions) can also be used to promote interesting problems in network theory.

This work presents a method for identifying condition-specific gene co-expression networks. The method combines the weighted co-expression network mining and condition- specific co-expression relationship detection using a chi-square test based enrichment analysis. By applying this method to a breast cancer microarray data set replete with different subtypes, we were able to identify a large number of condition-specific co-expression networks in non-basal like breast cancers suggesting that the underlying co-expression relationship has been disrupted in the basal like breast cancers. These results provide a new perspective for studying gene interaction dynamics in cancers and assessing the effects of perturbation on key genes such as transcription factors. Specifically, using the first network as an example, we suggest that the decreased gene expression level of the LEF1 transcription factor in basal like breast cancers may be associated with the disruption of this co-expression network, thus linking this network to potential cancer development. Our work provides a way for dynamically characterizing the gene interaction networks.

By mapping genes to transcription factors it is found that AR (Androgen Receptor) is mostly significantly enriched in sub-networks of both sub groups shows a high fold change. There is sufficient proof in the literature that AR is related to breast cancer. Further our method can tell us the genes which are regulated by AR. In other words our method finds condition specific gene co-expression networks and the transcription factors

regulating the condition specific co-expressed genes. Besides AR, transcription factors such as GATA2 and CEBPB are found to be significantly enriched.

References

1. Bhardwaj N, Lu H: **Co-expression among constituents of a motif in the protein-protein interaction network**. *Journal of bioinformatics and computational biology* 2009, **7**(1):1-17.

2. Langfelder P, Horvath S: **Eigengene networks for studying the relationships between co-expression modules**. *BMC systems biology* 2007, **1**:54.

3. Hu H, Yan X, Huang Y, Han J, Zhou XJ: **Mining coherent dense subgraphs across massive biological networks for functional discovery**. *Bioinformatics (Oxford, England)* 2005, **21 Suppl 1**:i213-221.

4. Sun Y, Li H, Liu Y, Mattson MP, Rao MS, Zhan M: **Evolutionarily conserved transcriptional co-expression guiding embryonic stem cell differentiation**. *PloS one* 2008, **3**(10):e3406.

5. 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.

6. Kais Z, Barsky SH, Mathsyaraja H, Zha A, Ransburgh DJ, He G, Pilarski RT, Shapiro CL, Huang K, Parvin JD: **KIAA0101 interacts with BRCA1 and regulates centrosome number**. *Mol Cancer Res*, **9**(8):1091-1099.

7. Li H, Sun Y, Zhan M: **Exploring pathways from gene co-expression to network dynamics**. *Methods in molecular biology (Clifton, NJ* 2009, **541**:249-267.

8. Horvath S, Zhang B, Carlson M, Lu KV, Zhu S, Felciano RM, Laurance MF, Zhao W, Qi S, Chen Z *et al*: **Analysis of oncogenic signaling networks in glioblastoma identifies ASPM as a molecular target**. *Proceedings of the National Academy of Sciences of the United States of America* 2006, **103**(46):17402-17407.

9. Zhang J, Xiang Y, Jin R, Huang K: **Using Frequent Co-expression Network to Identify Gene Clusters for Breast Cancer Prognosis**. In: *International Joint Conference on Bioinformatics, Systems Biology and Intelligent Computing (IJCBS).* Shanghai: IEEE Computer Society; 2009.

10. Zhang J, Ding L, Keen-Circle K, Borlawsky T, Xiang Y, Ozer H, Jin R, Payne P, Huang K: **Predicting biomarkers for chronic lymphocytic leukemia using gene co-expression network analyses for ZAP70**. In: *AMIA Summit of Translational Bioinformatics: 2010; San Francisco, CA*; 2010.

11. Xiang Y, Zhang C-Q, Huang K: **Predicting Glioblastoma Prognosis Networks Using Weighted Gene Co-expression Network Analysis on TCGA Data**. *BMC*

bioinformatics 2011, **accepted to Specifial Issue for Proceedings of the Great Lake Bioinformatics Conference**.

12. Wenzke K, Cantemir C, Zhang J, Marsh C, Huang K: **Identifying Common Genes and Networks in Multi-Organ Fibrosis**. In: *AMIA Summit on Translational Bioinformatics: March 2012; San Francisco*; 2012.

13. Presson AP, Sobel EM, Papp JC, Suarez CJ, Whistler T, Rajeevan MS, Vernon SD, Horvath S: **Integrated weighted gene co-expression network analysis with an application to chronic fatigue syndrome**. *BMC systems biology* 2008, **2**:95.

14. Langfelder P, Horvath S: **WGCNA: an R package for weighted correlation network analysis**. *BMC bioinformatics* 2008, **9**:559.

15. Li A, Horvath S: **Network module detection: Affinity search technique with the multi-node topological overlap measure**. *BMC research notes* 2009, **2**:142.

16. Yip AM, Horvath S: **Gene network interconnectedness and the generalized topological overlap measure**. *BMC bioinformatics* 2007, **8**:22.

17. Zhang J, Xiang Y, Ding L, Keen-Circle K, Borlawsky TB, Ozer HG, Jin R, Payne P, Huang K: **Using gene co-expression network analysis to predict biomarkers for chronic lymphocytic leukemia**. *BMC bioinformatics*, **11 Suppl 9**:S5.

18. Voineagu I, Wang X, Johnston P, Lowe JK, Tian Y, Horvath S, Mill J, Cantor RM, Blencowe BJ, Geschwind DH: **Transcriptomic analysis of autistic brain reveals convergent molecular pathology**. *Nature*, **474**(7351):380-384.

19. Li KC: **Genome-wide coexpression dynamics: theory and application**. *Proceedings of the National Academy of Sciences of the United States of America* 2002, **99**(26):16875-16880.

20. Lai Y, Wu B, Chen L, Zhao H: **A statistical method for identifying differential gene-gene co-expression patterns**. *Bioinformatics (Oxford, England)* 2004, **20**(17):3146-3155.

21. Ma H, Schadt EE, Kaplan LM, Zhao H: **COSINE: COndition-SpecIfic sub-NEtwork identification using a global optimization method**. *Bioinformatics (Oxford, England)*, **27**(9):1290-1298.

22. Zhang B, Horvath S: **A general framework for weighted gene co-expression network analysis**. *Statistical applications in genetics and molecular biology* 2005, **4**:Article17.

23. Ou Y, Zhang C-Q: **A new multimembership clustering method**. *Jounral of Industrial and Management Optimization* 2007, **3**(4):619-624.

24. Ayyanan A, Civenni G, Ciarloni L, Morel C, Mueller N, Lefort K, Mandinova A, Raffoul W, Fiche M, Dotto GP *et al*: **Increased Wnt signaling triggers oncogenic conversion of human breast epithelial cells by a Notch-dependent mechanism**. *Proceedings of the National Academy of Sciences of the United States of America* 2006, **103**(10):3799-3804.

25. Holmes KA, Song JS, Liu XS, Brown M, Carroll JS: **Nkx3-1 and LEF-1 function as transcriptional inhibitors of estrogen receptor activity**. *Cancer research* 2008, **68**(18):7380-7385.

26. Dziuda D: **Data mining for genomics and proteomics: analysis of gene and protein expression data**: Wiley-Interscience; 2010.

27. Kostka D, Spang R: **Finding disease specific alterations in the co-expression of genes**: Bioinformatics, 20(1): 194-199, 2004.

28. Jen CH, Manfield IW, Michalopoulos I, Pinney JW, Willats WG, Gilmartin PM, Westhead DR: **The Arabidopsis co-expression tool (ACT): a WWW-based tool and database for microarray-based gene expression analysis**: Plant Journal, 46(2):336-48, 2006

29. Sung B Cho, Jihun Kim, Ju H Kim: **Identifying set-wise differential co-expression in gene expression microarray data**: BMC Bioinformatics, 10:109, 2009

30. Rafael A. Irizarry, Bridget Hobbs, Francois Collin, Terence P. Speed: **Exploration, Normalization and summaries of high density oligonucleotide array probe level data**: Biostatistics, **4**, 2 249:264, 2003