Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant Emory University and its agents the non-exclusive license to archive, make accessible and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter know, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use the future works (such as articles or books) all or part of this thesis or dissertation.

Signature

_____ _____

Pei-Yu Lin Date

Genome-wide Coexpression Dynamics in Lung Adenocarcinoma

By

Pei-Yu Lin
Master of science in Public Health

Department Biostatistics and Bioinformatics
Rollins School of Public Health

Tianwei Yu, Ph.D., Committee chair	Date
Hao Wu, Ph.D., Committee member	Date

Genome-wide Coexpression Dynamics in Lung Adenocarcinoma

By

Pei-Yu Lin

M.S.
National Taiwan University
2008

Thesis Committee Chair: Tianwei Yu, Ph.D.

An abstract of
A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
In Partial fulfillment of the requirement for the degree of
Master of Science in Public Health
In Department Biostatistics and Bioinformatics
Rollins School of Public Health
2011

Abstract

Genome-wide Coexpression Dynamics in Lung Adenocarcinoma

By Pei-Yu Lin

Lung cancer is one of the leading cause of cancer death in the United States and worldwide. Lung cancer in never smokers is almost universally non-small cell lung carcinoma (NSCLC), with a sizeable majority being adenocarcinoma. A microarray is a multiplex lab-on-a-chip consisting of an arrayed series of thousands of microscopic spots of DNA oligonucleotides that are used to hybridize a cDNA or cRNA sample. The high- throughput technology can accomplish many genetic tests in parallel and has dramatically accelerated many types of investigation. Liquid Association (LA) is a method to systematically study the co-expression pattern between functionally related genes as the cellular-state changes. In this study we used LA to analyze NSCLC DNA microarray data from 4 different lung cancer institutes. We found genes APOL3 and CEACAM1 to be a potential gene pair that exhibit dynamic correlations for blood glucose level maintenance when using CMKLR1 as the surrogate of the cellular-status. A study from Sergentanis et al. pointed out that the defect of glucose metabolism is a meaningful factor to elevate lung cancer risk. Our result on the potential dynamic correlation of genes APOL3 and CEACAM1 could provide biologists and doctors candidate genes for further study of the relationship between glucose metabolism and lung cancer formation.

Genome-wide Coexpression Dynamics in Lung Adenocarcinoma

By

Pei-Yu Lin

M.S.
National Taiwan University
2008

Thesis Committee Chair: Tianwei Yu, Ph.D.

A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
In Partial fulfillment of the requirement for the degree of
Master of Science in Public Health
In Department Biostatistics and Bioinformatics
Rollins School of Public Health
2011

Table of contents

Introduction

In this section, we provide an overview of the background of our study, the content of the data sets we used, the challenge we faced and our goal for this study by using the statistical method Liquid Association.

Lung cancer

Lung cancer is the leading cause of cancer death in the United States and worldwide. It claims more than 150,000 lives every year in the United States, thus exceeding the combined mortality from breast, prostate and colorectal cancers. (Doll and Peto 1981; Devesa, Silverman et al. 1987; Landis, Murray et al. 1999; Govindan, Page et al. 2006; Hu, Chen et al. 2010). The 5-year, overall survival rate is 15% and has not improved over many decades.

Approximately 98% of lung cancers are carcinoma, which are tumors composed of cells with epithelial characteristics, and there are 8 major groups of lung carcinomas recognized in WHO (Vardiman, Harris et al. 2002): squamous cell carcinoma, small cell carcinoma, adenocarcinoma, large cell carcinoma, adenosquamous carcinoma, sarcomatoid carcinoma, carcinoid tumor, and salivary gland-like carcinoma. Non-small cell lung carcinoma (NSCLC) is any type of epithelial lung cancer other than small cell lung carcinoma (SCLC). As a class, NSCLCs are relatively insensitive

to chemotherapy, compared to small cell carcinoma. When possible, they are primarily treated by surgical resection with curative intent, although chemotherapy is increasingly being used.

The most common types of NSCLC are squamous cell carcinoma, large cell carcinoma, and adenocarcinoma. Lung cancer in never smokers is almost universally NSCLC, with a sizeable majority being adenocarcinoma. Adenocarcinomas account for approximately 40% of lung cancers.

DNA Microarray

A microarray is a multiplex lab-on-a-chip (Sambrook and Russell 2001; Chomczynski and Sacchi 2006). It is a 2D array on a solid substrate that assays large amounts of biological material using high-throughput screening methods. DNA microarray consists of an arrayed series of thousands of microscopic spots of DNA oligonucleotides, each containing picomoles of a specific DNA sequence, known as probes. These can be a short section of a gene or other DNA element that are used to hybridize a cDNA or cRNA sample. The outcome of hybridization is usually detected and quantified by the intensity of fluorophore-, silver-, or chemiluminescence-signal. Since an array can contain tens of thousands of probes, a microarray experiment can accomplish many genetic tests in parallel. Therefore arrays have dramatically accelerated many types of investigation.

Microarray data sets are commonly very large, and analytical precision is influenced by a number of variables. Statistical challenges include taking into account effects of background noise and appropriate normalization of the data. Algorithms that affect statistical analysis include image analysis, data processing, identification of statistically significant changes (Leung and Cavalieri 2003; Wei, Li et al. 2004; Ben-Gal, Shani et al. 2005; Priness, Maimon et al. 2007) and network-based methods (Emmert-Streib and Dehmer 2008). Microarray data may require further processing aimed at reducing the dimensionality of the data to aid comprehension and more focused analysis (Wouters, Gohlmann et al. 2003). Other methods permit analysis of data consisting of a low number of biological or technical replicates; for example, the Local Pooled Error (LPE) test pools standard deviations of genes with similar expression levels in an effort to compensate for insufficient replication (Wouters, Gohlmann et al. 2003).

Theoretical Concepts

Microarray technologies have made it straightforward to monitor simultaneously the expression pattern of thousands of genes. These new data promise to enhance fundamental understanding of life on the molecular level, from regulation of gene expression and gene function to cellular mechanisms, and many prove useful in

medical diagnosis, treatment, and drug design. The challenge now is to interpret such massive database. A natural basis for organizing gene expression data is to group together genes with similar patterns of expression (Eisen, Spellman et al. 1998; Marcotte, Pellegrini et al. 1999; Tamayo, Slonim et al. 1999; Alter, Brown et al. 2000; Brown, Grundy et al. 2000). Such profile-similarity analysis argued that those co-expressed genes are likely to be functionally associated, thus indicating their encoded proteins have higher probabilities of participating in a common structural complex metabolic pathway or biological process (Eisen, Spellman et al. 1998; Marcotte, Pellegrini et al. 1999).

Although many successful applications were reported, whether information about the underlying genetic architecture and regulatory interconnections can be derived from the analysis of coexpression patterns remains to be determined. There is an issue that is hard to be addressed by profile-similarity analysis. As is known, biological processes are interlocked and many proteins have multiple cellular roles. The role of each gene might be different along with the intrinsic cellular-state changes, which implies that the relationship between a pair of genes would not be consistent over time. In this condition, the profiles of most functionally associated genes would turn out to be uncorrelated under profile-similarity analysis.

Study Goal

Li (Li 2002) proposed another idea of Liquid Association (LA) to capture pairs of genes with low direct correlation but their correlation becomes high when conditioned on certain cellular states or the expression of a third gene. These methods help to capture potential co-regulation information that gene clustering may not reveal. In this study, we will follow this method to analyze the microarray data from NSCLC patients. By finding candidate gene pairs or potential regulatory genes, we could provide information to biologist or doctors for further investigation of possible cellular molecular mechanisms and identify susceptibility genes related to NSCLC.

Data Description

Data (Shedden, Taylor et al. 2008) were collected by investigators at four institutions, Moffitt Cancer Center (HLM), University of Michigan Cancer Center (UM), Dana-Farber Cancer Institute (CAN/ DF) and Memorial Sloan-Kettering Cancer Center (MSK). NSCLC samples were collected and studied using a common protocol. Subjects' samples along with all relevant clinical, pathological and outcome data were collected to ensure scientific validity of the results. The summary statistics of clinical variables in each data set are given as table 1. Gene expression data on subsets of lung adenocarcinomas were generated by each of four different laboratories using a

common platform and following a protocol previously demonstrated to be robust and reproducible (Dobbin, Beer et al. 2005). The data was collected, analyzed and evaluated by following strict protocol. It provides a rich dataset for many different kinds of analysis. The study is an example of how large data sets can be generated and tested by cooperation and pooling of resources among many investigators.

Methods

In this section, we introduce our analysis procedure and the main methods we used to construct our gene signature. We separated the analyzing procedure into two parts: gene filtering and gene selection. We first excluded the genes with low expression or small variation of expression. Then we extracted scouting genes and its high LA score gene pairs. Finally we performed functional study of the candidate genes that are highly correlated. Both the correlation and liquid association results were computed using the website: http://kiefeistat2.sinica.edu.tw/LAP3/index.php. The summary of the procedure was illustrated in figure 1.

Gene filtering

The major challenge in microarray data analysis is the large dimensionality. The number of genes is in the range ten to fifty thousand but the sample size is only about hundreds. There are 22,215 probe sets on Affymatrix U133A microarray (Langdon, da Silva Camargo et al. 2007), but the proportion of truly expressed gene might be not greater than half. To reduce the effect of microarray noise, we first filtered out non-informative genes by excluding the genes with low expressions. Only the genes with a mean expression level over the medium of the mean expressions from all genes were retained.

Another gene filtering criterion we used was the variation of each gene expression profile. Some housekeeping genes expressed constantly high or low for basic reactions. The variation of these gene expression profiles were small and may not relate to the question of our interest. However, selecting genes by normal quintile transformed correlation coefficients compared the correlations between gene expressions at the same scale of variation. Some non-informative genes might be selected. Therefore, we excluded the genes with small sample variation of expressions. Only the genes with variation expression over the medium of variance were retained.

After the two-step gene filtering, there were around 10,000 genes retained in our data from each of the 4 institutes. We then used Liquid Association to analyze candidate genes.

Gene selection

After gene filtering, we reduced the total number of genes for further statistical analysis. For analyzing the microarray data by using Liquid Association, we next used a normal score transformation to preprocess those remaining gene expression data so that the expression of each gene is normally distributed with mean 0 and variance 1.

Liquid Association is a method to systematically study the co-expression pattern between functionally related genes as the cellular-state changes (figure 2). In contrast

to direct approach, which would be to specify a number of profiling according to different intercellular conditions, LA approaches the problem in a reverse manner. For studying the correlation between any pair of two given genes (X, Y), the third gene Z is introduced. The assumption is that the cellular-state variable is correlated with the expression of a third gene Z. If this is the case, we may use Z as a scouting gene to detect such a 'liquid' (as opposed to 'solid') pattern of statistical association between a pair of genes (X and Y).

Liquid Association has been used to analyze the dynamic metabolic networks in yeast (Li 2002), anticancer drug screen, and candidate gene research for Alzheimer disease (Li, Liu et al. 2004) and Multiple sclerosis(Li, Palotie et al. 2007).

A liquid association score LA(X, Y|Z) can be computed using a simple statistical formula given below:

I. Standardize each gene-expression profile with a normal score transformation. Specifically, we rank the m observations in the profile $R_1, \ldots, R_m$. The ranks are then used to obtain the transformed profile, $\Phi^{-1}(R_1/(m+1)), \Phi^{-1}(R_2/(m+1)), \ldots, \Phi^{-1}(R_m/(m+1))$, where $\Phi(.)$ is the cumulative normal distribution.

II. Compute the average product of the three transformed profiles,

$$(X_1Y_1Z_1 + \ldots\ldots\ldots + X_mY_mZ_m) / m$$

which is the LA score that we need.

Specifically, if an increase in Z is associated with a decrease in the correlation of (X, Y), then gene Z is a negative LA-scouting gene for (X, Y), and a negative score is assigned. Likewise, a positive LA score implies that the change of the gene expression of Z is simultaneously increased or decreased with the correlation of (X,Y) being positive or negative.

There are two ways of applying LA. For a given pair of X and Y, one can look for genes that may mediate X,Y co-expression by computing the LA score LA(X, Y|Z) for each gene Z in the genome and obtaining a genome-wide ranking. Alternatively, given one gene Z, we may ask which pairs of genes Z may mediate. In this study, we used the later method to analyze the data.

In this study, we treated each gene in the dataset as a scouting gene. First we calculated LA scores of each scouting gene with all possible pair of gene combinations, and selected the scouting genes with the highest and lowest 1000 LA scores from each of the four institutes. After comparing the scouting genes we selected among the four institutes and finding the overlapping scouting genes, we then analyzed the common highly correlated gene pairs among those overlapping scouting genes. Finally, by biological functional study, we investigated those high LA score gene pairs and found their potential metabolic relationships on the

biological level. According to the functional studies, we then evaluated the possible roles of those gene pairs that may contribute to NSCLC.

Results

In this section, we present our results by Liquid Association. The gene functional studies provided clues of the relationships among the genes we selected by Liquid Association and Non-small-cell lung carcinoma.

Two gene pairs were selected by Liquid Association

After filtering out genes with low expression or low variation expression, we had approximately 10,000 genes retained in each dataset from the four institutes. Next we considered each of the remaining genes as a scouting gene and calculated its LA scores with all the possible gene pairs. For each of the dataset we used from the four different institutes, we retained the first 2000 scouting genes that had the highest positive LA scores and the last 2000 scouting genes that had the lowest negative LA scores. We called the scouting genes we retained in this step top master genes or bottom master genes respectively since these genes helped finding gene pairs that were highly correlated. By comparing the master genes from each of the four institutes, we found there were 10 top master genes and 14 bottom master genes overlapped in all those four institutes.

Next, we used the 24 overlapped master genes to find candidate gene pairs that were functionally highly correlated. We again calculated the LA score of the 14 master

genes with all the possible combination of two genes by Liquid Association. The gene pairs that created the top 500 highest LA scores with each of the top master genes were retained from each institute. Also the gene pairs that created the top 500 lowest LA scores with each of the bottom master genes were retained. Finally, we compared the gene pairs from each of the institutes and searched the gene pairs appeared among all those four institutes by cross-matching. We analyzed two gene pairs from the data sets: APOL3/CEACAM1/CMKLR1 (scouting) and PCNA/SEC31A/VBP1 (scouting).

Interpretation of significant finding

The LA scores of APOL3/CEACAM1 were -0.5336 in Dana-Farber Cancer Institute(CAN/ DF), -0.4063 in University of Michigan Cancer Center (UM), -0.4558 in Memorial Slean-Kettering Cancer Center(MSK), -0.0183 in Moffitt Cancer Center (HLM) with CMKLR1 being the scouting gene. The negative LA scores indicated that the APOL3/CEACAM1 was a negative LA pair (LAP) of CMKLR1. As a result, the increase of CMKLR1 is associated with a decrease in the correlation of APOL3 and CEACAM1. As a negative LA-scouting gene of the gene pair APOL3/CEACAM1, the expression of CMKLR1 can be seen as a monitor to detect the change of the relationship between the gene pair APOL3/CEACAM1, from co-expression to contra-expression.

Carcinoembryonic antigen-related cell adhesion molecule 1 (biliary glycoprotein) (CEACAM1) is also known as CD66a (Thompson, Zimmermann et al. 1992). Multiple cellular activities have been attributed to the encoded protein, including roles in the differentiation and arrangement of tissue three-dimensional structure, angiogenesis, apoptosis, tumor suppression, metastasis, and the modulation of innate and adaptive immune responses. Recent studies revealed that insulin stimulates phosphorylation of CEACAM1 which in turn leads to up-regulation of receptor-mediated insulin endocytosis and degradation in the hepatocyte (Formisano, Najjar et al. 1995; Najjar, Philippe et al. 1995). The gene expression of ECACAM1 has a role of down-regulation on insulin resistance. APOL3 is a member of the apolipoprotein L gene family, and it is present in a cluster with other family members on chromosome 6 (Dunham, Hunt et al. 1999; Horrevoets, Fontijn et al. 1999). The encoded protein is found in the cytoplasm, where it may affect the movement of lipids, including cholesterol, and/or allow the binding of lipids to organelles. In addition, expression of this gene is up-regulated by tumor necrosis factor-alpha (TNF-alpha) in endothelial cells lining the normal and atherosclerotic iliac artery and aorta (Shah, Lu et al. 2009). TNF-alpha has been reported positively correlated to insulin resistance. This indicated that TNF-alpha-induced APOL3 expression has a role of up-regulation on insulin resistance. Chemokine receptor-like 1 also known as ChemR23 (Chemerin

Receptor 23) is a protein that in humans is encoded by the CMKLR1 gene (Gantz, Konda et al. 1996). Chemokine receptor-like 1 is a G protein-coupled receptor for the chemoattractant adipokine chemerin (Wittamer, Franssen et al. 2003) and the omega-3 fatty acid derived molecule resolvin E1. CMKLR1 shows wide RNA expression profile but is notably high in plasmacytoid dendritic cells, macrophages, cardiomyocytes, adipocytes and endothelial cells. Activating CMKLR1 by an agonist mobilizes intracellular calcium and causes the activation of several other signaling cascades like the ERK1 and NF-κB. Initial studies of CMKLR1 suggested that it might have a role in the inflammatory pathways. Its cognate ligand, chemerin, also known as retinoic acid receptor responder protein 2 (RARRES2), tazarotene-induced gene 2 protein (TIG2), or RAR-responsive protein. Studies in mice have shown neither chemerin nor CMKLR1 are highly expressed in brown adipose tissue, indicating that chemerin plays a role in energy storage rather than thermogenesis. Interestingly, it was found incubation of 3T3-L1 cells with recombinant human chemerin protein facilitated insulin-stimulated glucose uptake (Cash, Hart et al. 2008). This suggests chemerin plays a role in insulin sensitivity and may be a potential therapeutic target for treating type II diabetes. The role of chemerin played in insulin-stimulated glucose up take indicated that its receptor- CMKLR1 also has functions correlated to insulin resistance.

According to the functional studies of APOL3, CEACAM1 and CMKLR1, the three genes have functions related to insulin-based glucose metabolism in human. Based on those studies, APOL3 plays a role in facilitating insulin resistance while CEACAM1 inhibit insulin resistance. In addition, the expression of CMKLR1 is as marker of insulin-stimulated glucose uptake. As a result, CMKLR1 could be a sensor or a surrogate of the blood glucose level changes in human body. When the blood glucose level was high, CEACAM1 was down regulated as APOL3 was up-regulated, to maintained the concentration of insulin in the blood high enough and facilitate the uptake of glucose in the blood. Therefore CEACAM1 and APOL3 are negatively correlated when the human body is in a high blood glucose cellular state.

The other triplet we found by Liquid Association was PCNA/SEC31A/VBP1. The LA score of PCNA/SEC31A were 0.0026 in CAN/ DF, -0.3771 in UM -0.4558 in MSK, -0.0777 in HLM with VBP1 being the scouting gene. The LA scores were close to zero in CAN/ DF and HLM, which indicated that the correlation between PCNA and SEC31A was not significant under the supervision of VBP1 according to the data from the two institutes.

Proliferating Cell Nuclear Antigen (PCNA) is a protein that acts as a processivity factor for DNA polymerase in eukaryotic cells (Leonardi, Girlando et al. 1992). The protein acts as a homotrimer and helps increase the processivity of leading strand

synthesis during DNA replication. In response to DNA damage, this protein is ubiquitinated and is involved in the RAD6-dependent DNA repair pathway (Hoege, Pfander et al. 2002). The protein encoded by SEC31A is similar to yeast Sec31 protein. Yeast Sec31 protein is known to be a component of the COPII protein complex which is responsible for vesicle budding from endoplasmic reticulum (ER). The von Hippel-Lindau binding protein 1 (VBP1), also known as "prefoldin 3", is a chaperone protein that binds to von Hippel-Lindau protein and transports it from perinuclear granules to the nucleus or cytoplasm inside the cell (Tsuchiya, Iseda et al. 1996). It is also involved in transporting nascent polypeptides to cytosolic chaperonins for post-translational folding. According to the functional studies of gene PCNA and SEC31A, we did not find those two genes share a common function related to each other. Thus, based on the biological studies, together with the statistical result from Liquid Association, we did not find a solid relationship between PCNA and SEC31A when we treated VBP1 as a scouting gene.

Discussion

Statistical similarity analysis has been instrumental in analysis of microarray data. Gene with correlated expression profiles tend to be functionally associated. However, many functionally associated genes turn out to be uncorrelated. Despite of the high noise level of microarray data and the complexity of the mechanism of gene regulation, the co-expression pattern of a pair of genes can be sensitively dependent on cellular state. This possibility can be described in terms of LA. The idea of LA is that it uses a third gene as a whistle blower, which may not have a direct physical contact with its LAPs, to describe subtle co-expression pattern between a pair of genes. Traditionally, we could investigate the subtle co-expression by inspecting the scatter plot. However, for the genome-wide study, there is a computational hurdle to overcome, because there are too many combinations for choosing three genes. Human, for example, has >20,000 genes and the total number of triplets would be >1000, 000, 000, 000. We surely cannot afford to inspect every scatter plot to find all triplets with LA patterns. Consequently, an easy-to-compute index of how likely one is to find a LA pattern is desirable. The formula of LA that was given by Li (Li 2002), Which turns out to be simple enough to serve the purpose.

In this study we were able to demonstrate some applications of LA using the

Non-small cell lung carcinoma (NSCLC) data. We found the potential pair of genes APOL3/CEACAM1 that both have the functions related to insulin resistance according to recent biological studies. According to the biological studies, CMKLR1 is also has a function of maintaining the glucose level in the blood. Thus, CMKLR1 serves as an indicator of the cellular state change that could change the correlation between APOL3/CEACAM1.

Insulin resistance is a condition with increased prevalence and significance in the contest of obesity, metabolic syndrome, nonalcoholic fatty liver disease, and type 2 diabetes mellitus (Tritos and Mantzoros 1998; Samuel, Petersen et al. 2010), and malignancy (Kaaks 1996; Kaaks, Lukanova et al. 2002). The large cohort studies including the National Health and Nutrition Examination Survey III in the United States (Parekh, Lin et al. 2010) and the Heart Disease and Diabetes in a Screened Cohort in the United Kingdom (Loh, North et al. 2010) have independently pointed to insulin resistance as a risk factor for overall cancer mortality. In addition, insulin resistance, quantified by means of *HOMA*-IR (i.e. *HOMA*-IR index= insulin（μU/mL）× glucose（mmol/L）/22.5), emerged as a meaningful factor pointing out to elevated lung cancer risk in recent study (Petridou, Sergentanis et al. 2011).The same results appeared when the analysis only considered the NSCLC cases. Combined with those studies, our investigation of the dynamic correlation of genes APOL3 and

CEACAM1 could provide biologists and doctors with candidate genes for further study of the relationship between insulin resistance and lung cancer formation.

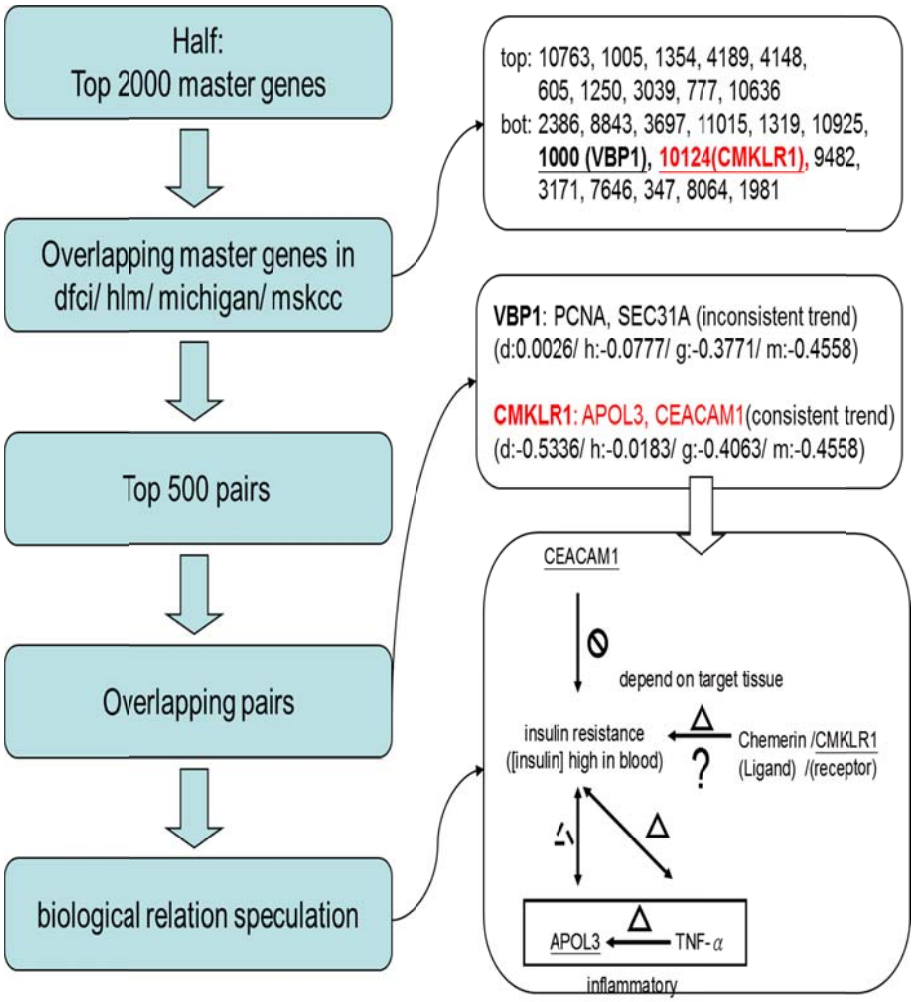
Liquid Association, as a statistical method of analyzing the correlation of a pair of genes, points to a new source of information hidden in the microarray data. For n genes, the algorithm returns a huge amount of message, in the order of n^3, that can be stored and used in a variety of ways to meet different researchers' needs. In general, the better the LA score is, the more likely we can detect the LA pattern when visually examining the profile plots. Therefore, in our illustrations, we only analyzed the high-scoring LAPs. But how we use the information to infer the functional relationship of the genes depends on the available biological knowledge. We can use the existing bioinformatics resources, such as Kyoto Encyclopedia of Genes and Genomes (KEGG), Online Mendelian Inheritance in Man (OMIM), Information Hyperlinked over Proteins (iHOP), transcription factor database (TRANSFAC), protein data bank (PDB) and GeneCards, to investigate the plausible biological hypotheses and construct the potential metabolic map of the genes. LA can be extended in several directions. In addition to P value and visual inspection, one may bring in methods from multiple comparison / false discovery, an area with renewed interest fueled by microarray analysis. Similarity-based methods such as principle component analysis can be applied to high LA-scoring genes (Li, Liu et al.

2004).According to the information provided from the data we used in this study, we can also use the survival time, age, cancer stage as the scouting factors (Z) to select the important genes related to those different forms of cellular-state variables.

	HLM	UM	CAN/DF	MSK
sample size	79	177	82	104
age (mean)	67	64	61	65
cell type (% adenocarcinomas)	100	100	100	100
sex (%male)	51	56	56	36
stage IA	11	39	13	26
stage IB	43	27	55	35
stage II	27	16	32	19
stage III	19	18	0	20
stage IV	0	0	0	0
median follow up (months)	39	54	51	43
number of deathes	60	102	35	39

Table 1. Summary statistics of the clinical and survival data

Figure 1. The flow chart of analysis procedure



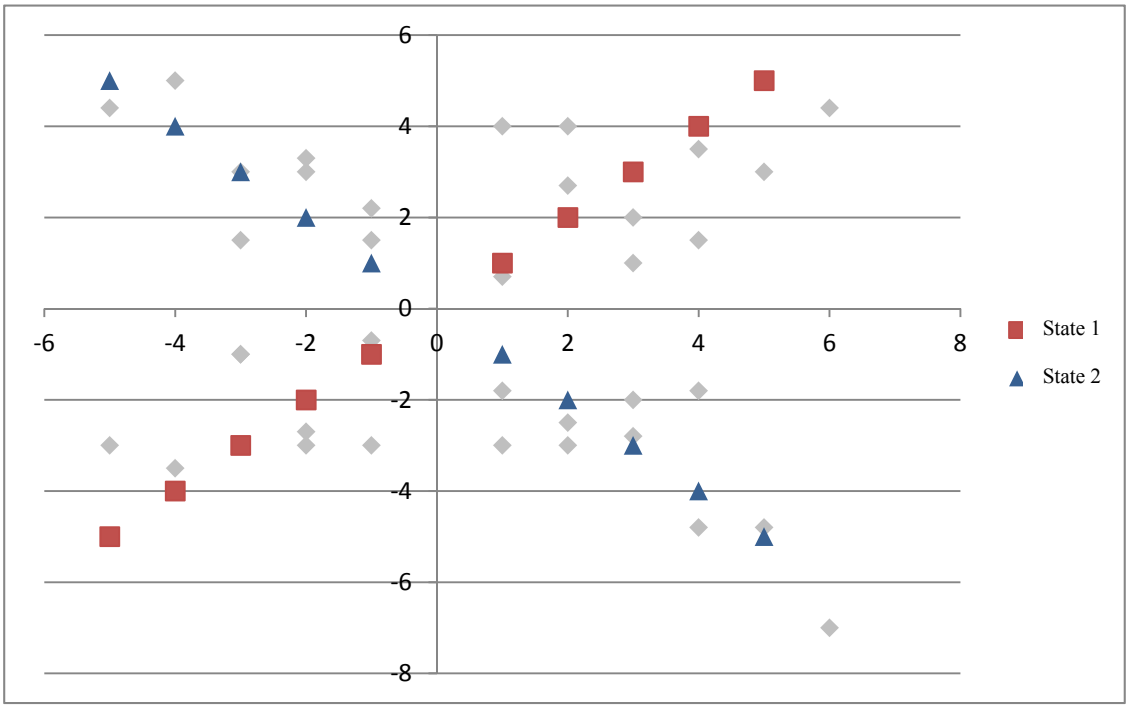


Figure 2. Co-expression dynamics

The correlation between two genes (X, Y) is displayed in a scatter plot. The red points represent the correlation between the two genes for cellular state 1 where are X and Y are co-expressed. Likewise, the blue points represent the correlation between the two genes for cellular state 2 where are X and Y are contra-expressed. To depict this kind of internal evolution in the association pattern, we say (X, Y) forms a Liquid Association Pair (LAP). Liquid Association is a statistical method using a third gene (Z) as a surrogate of relevant cellular state to detect genes (X, Y) about their LA activity.

Reference

Alter, O., P. O. Brown, et al. (2000). "Singular value decomposition for genome-wide expression data processing and modeling." Proceedings of the National Academy of Sciences of the United States of America **97**(18): 10101.

Ben-Gal, I., A. Shani, et al. (2005). "Identification of transcription factor binding sites with variable-order Bayesian networks." Bioinformatics **21**(11): 2657.

Brown, M. P. S., W. N. Grundy, et al. (2000). "Knowledge-based analysis of microarray gene expression data by using support vector machines." Proceedings of the National Academy of Sciences of the United States of America **97**(1): 262.

Cash, J. L., R. Hart, et al. (2008). "Synthetic chemerin-derived peptides suppress inflammation through ChemR23." The Journal of experimental medicine **205**(4): 767.

Chomczynski, P. and N. Sacchi (2006). "The single-step method of RNA isolation by acid guanidinium thiocyanate¡Vphenol¡Vchloroform extraction: twenty-something years on." Nature Protocols **1**(2): 581-585.

Devesa, S. S., D. Silverman, et al. (1987). "Cancer incidence and mortality trends among whites in the United States, 1947-84." Journal of the National Cancer Institute **79**(4): 701.

Dobbin, K. K., D. G. Beer, et al. (2005). "Interlaboratory comparability study of cancer gene expression analysis using oligonucleotide microarrays." Clinical cancer research **11**(2): 565.

Doll, R. and R. Peto (1981). "The causes of cancer: quantitative estimates of avoidable risks of cancer in the United States today." Journal of the National Cancer Institute **66**(6): 1191.

Dunham, I., A. Hunt, et al. (1999). "The DNA sequence of human chromosome 22." Nature **402**(6761): 489-495.

Eisen, M. B., P. T. Spellman, et al. (1998). "Cluster analysis and display of genome-wide expression patterns." Proceedings of the National Academy of Sciences

of the United States of America **95**(25): 14863.

Emmert-Streib, F. and M. Dehmer (2008). Analysis of microarray data: A network-based approach, Wiley-VCH.

Formisano, P., S. M. Najjar, et al. (1995). "Receptor-mediated internalization of insulin. Potential role of pp120/HA4, a substrate of the insulin receptor kinase." The Journal of biological chemistry **270**(41): 24073.

Gantz, I., Y. Konda, et al. (1996). "Molecular cloning of a novel receptor (CMKLR1) with homology to the chemotactic factor receptors." Cytogenetic and Genome Research **74**(4): 286-290.

Govindan, R., N. Page, et al. (2006). "Changing epidemiology of small-cell lung cancer in the United States over the last 30 years: analysis of the surveillance, epidemiologic, and end results database." Journal of Clinical Oncology **24**(28): 4539.

Hoege, C., B. Pfander, et al. (2002). "RAD6-dependent DNA repair is linked to modification of PCNA by ubiquitin and SUMO." Nature **419**(6903): 135-141.

Horrevoets, A. J. G., R. D. Fontijn, et al. (1999). "Vascular endothelial genes that are responsive to tumor necrosis factor-{alpha} in vitro are expressed in atherosclerotic lesions, including inhibitor of apoptosis protein-1, stannin, and two novel genes." Blood **93**(10): 3418.

Hu, Z., X. Chen, et al. (2010). "Serum MicroRNA Signatures Identified in a Genome-Wide Serum MicroRNA Expression Profiling Predict Survival of Non¡VSmall-Cell Lung Cancer." Journal of Clinical Oncology **28**(10): 1721.

Kaaks, R. (1996). "Nutrition, hormones, and breast cancer: is insulin the missing link?" Cancer causes and control **7**(6): 605-625.

Kaaks, R., A. Lukanova, et al. (2002). "Obesity, Endogenous Hormones, and Endometrial Cancer Risk." Cancer Epidemiology Biomarkers & Prevention **11**(12): 1531.

Landis, S. H., T. Murray, et al. (1999). "Cancer statistics, 1999." CA: A Cancer Journal for Clinicians **49**(1): 8-31.

Langdon, W., R. da Silva Camargo, et al. (2007). "Spatial defects in 5896 HG-U133A GeneChips." Critical Assesment of Microarray Data, Valencia: 13-14.

Leonardi, E., S. Girlando, et al. (1992). "PCNA and Ki67 expression in breast carcinoma: correlations with clinical and biological variables." Journal of clinical pathology **45**(5): 416.

Leung, Y. F. and D. Cavalieri (2003). "Fundamentals of cDNA microarray data analysis." TRENDS in Genetics **19**(11): 649-659.

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

Li, K. C., C. T. Liu, et al. (2004). "A system for enhancing genome-wide coexpression dynamics study." Proceedings of the National Academy of Sciences of the United States of America **101**(44): 15561.

Li, K. C., A. Palotie, et al. (2007). "Finding disease candidate genes by liquid association." Genome biology **8**(10): R205.

Loh, W. J., B. V. North, et al. (2010). "Insulin resistance-related biomarker clustering and subclinical inflammation as predictors of cancer mortality during 21.5 years of follow-up." Cancer causes and control **21**(5): 709-718.

Marcotte, E. M., M. Pellegrini, et al. (1999). "A combined algorithm for genome-wide prediction of protein function." Nature **402**(6757): 83-86.

Najjar, S. M., N. Philippe, et al. (1995). "Insulin-stimulated phosphorylation of recombinant pp120/HA4, an endogenous substrate of the insulin receptor tyrosine kinase." Biochemistry **34**(29): 9341-9349.

Parekh, N., Y. Lin, et al. (2010). "Longitudinal associations of blood markers of insulin and glucose metabolism and cancer mortality in the third National Health and Nutrition Examination Survey." Cancer causes and control **21**(4): 631-642.

Petridou, E. T., T. N. Sergentanis, et al. (2011). "Insulin resistance: an independent

risk factor for lung cancer?" Metabolism.

Priness, I., O. Maimon, et al. (2007). "Evaluation of gene-expression clustering via mutual information distance measure." BMC bioinformatics **8**(1): 111.

Sambrook, J. and D. W. Russell (2001). Molecular cloning: a laboratory manual, Cold spring harbor laboratory press.

Samuel, V. T., K. F. Petersen, et al. (2010). "Lipid-induced insulin resistance: unravelling the mechanism." The Lancet **375**(9733): 2267-2277.

Shah, R., Y. Lu, et al. (2009). "Gene profiling of human adipose tissue during evoked inflammation in vivo." Diabetes **58**(10): 2211.

Shedden, K., J. M. G. Taylor, et al. (2008). "Gene expression¡Vbased survival prediction in lung adenocarcinoma: a multi-site, blinded validation study." Nature medicine **14**(8): 822-827.

Tamayo, P., D. Slonim, et al. (1999). "Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation." Proceedings of the National Academy of Sciences of the United States of America **96**(6): 2907.

Thompson, J., W. Zimmermann, et al. (1992). "Long-range chromosomal mapping of the carcinoembryonic antigen (CEA) gene family cluster." Genomics **12**(4): 761-772.

Tritos, N. and C. Mantzoros (1998). "Clinical review 97: Syndromes of severe insulin resistance." The Journal of clinical endocrinology and metabolism **83**(9): 3025.

Tsuchiya, H., T. Iseda, et al. (1996). "Identification of a novel protein (VBP-1) binding to the von Hippel-Lindau (VHL) tumor suppressor gene product." Cancer research **56**(13): 2881.

Vardiman, J. W., N. L. Harris, et al. (2002). "The World Health Organization (WHO) classification of the myeloid neoplasms." Blood **100**(7): 2292.

Wei, C., J. Li, et al. (2004). "Sample size for detecting differentially expressed genes in microarray experiments." BMC genomics **5**(1): 87.

Wittamer, V., J. D. Franssen, et al. (2003). "Specific recruitment of antigen-presenting cells by chemerin, a novel processed ligand from human inflammatory fluids." The Journal of experimental medicine **198**(7): 977.

Wouters, L., H. W. Gohlmann, et al. (2003). "Graphical exploration of gene expression data: a comparative study of three multivariate methods." Biometrics **59**(4): 1131-1139.