A SYSTEMATIC STUDY REVEALS NEW INSIGHTS OF CANCER

by

KUN XU

(Under the Direction of Dr.Ying Xu)

ABSTRACT

The decreasing trend of cancer mortality has been mostly due to the improved diagnostic techniques for detecting the early stage of cancer as well as development of therapeutic strategy which heavily depends on the understanding of the fundamental biology of tumor cell. We applied a systematic study by using bioinformatics and computational biology methods on gene expression data to address problems that related to those two issues. (1).A comparative study of public gene-expression data of seven types of cancers was conducted with the aim of deriving serum marker genes for early detection, The analysis results indicate that (1a) each cancer type can be distinguished from its corresponding control tissue based on the expression patterns of a small number of genes; (1b) the expression patterns of some genes can distinguish multiple cancer types from their corresponding control tissues, potentially serving as general markers for all or some groups of cancers; (1c) the proteins encoded by some of these genes are predicted to be blood secretory providing potential cancer markers in blood. (2). A comparative analysis of two types of skin cancers, melanoma and basal cell carcinoma in comparison with other cancer types, was conducted with the aim of improving the understanding and identifying key regulatory factors that either cause or contribute to the

aggressiveness of melanoma. Our findings include the following. (2a) Advanced melanoma shows substantial up-regulation of key genes involved complimentary metabolism process, providing a source of the energetics necessary to support the rapid growth. (3) A comparative analysis of six solid cancer types in micro-environmental study with the aim of proposing a model of how cancer cells utilize a few mechanisms to keep the protons outside of the cells. (3a)The model consists of a number of previously studied, well or partially, mechanisms for transporting out the excess protons and a new mechanism that neutralizes protons. (3b)We hypothesize that these processes are regulated by cancer related conditions making these encoded processes not available to normal cells under acidic conditions. We believe this systematic study will bring important insight regarding to both topics to the cancer research field.

INDEX WORDS: Cancer, System Biology, Bioinformatics, Biomarkers, Microarray gene expression data, Pathway analysis, Secretory protein, Oncogenic Metabolism, Warburg Effect, Melanoma, Glycolysis, Tumor Micro-Environment, Acidosis, Statistics.

A SYSTEMATIC STUDY REVEALS NEW INSIGHTS OF CANCER

by

KUN XU

B.S., Jilin University, China, 2005

M.S., The University of Georgia, 2008

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2012

A SYSTEMATIC STUDY REVEALS NEW INSIGHTS OF CANCER

by

KUN XU

Major Professor: Ying Xu

Committee: Jaxk Reeves
Lily Wang
Xiangrong Yin
Shaying Zhao

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
December 2012

DEDICATION

Dedicate to my father who has been fighting against cancer for years. Also dedicate to my mother who makes tremendous effort to keep this family together. Even I did research at fundamental level, still hope my research can eventually help patients and their family out from suffering.

ACKNOWLEDGEMENTS

First of all, I would like to show my deepest gratitude to my supervisor Dr. Ying Xu for providing me valuable guidance and continues support during the last six years. I thank him for sharing with me his enthusiasm in science which truly moved me and his tremendous efforts on guiding me to go through the most import transformation in my life from a student knowing little about science into a mature researcher who is capable of doing independent research.

I wish to express my sincere gratitude to all faculty members in my advisory committee for their guidance and encouragement for helping me to turning my research ideas into complete studies. My particular thanks will go to Dr Jaxk Reeves, Dr. Lily Wang and Dr. Xiangrong Yin for their expert guidance and constant help with each of my research. They all set good examples to me as excellent statisticians. I also thank Dr Shaying Zhao for her inspiring suggestion on the biological problems involved in my research. Their passion and dedication in science has influenced me all the way along the path.

I also would like to thank Professor J Dave Puett, Dr Victor Olman, Dr Juan Cui, Dr Xizeng Mao for their valuable help on my projects. Special thanks to the Dr Minesh Mehta for the cooperative effort and inspiration on the projects.

Many thanks to Chi Zhang for providing me a fun living place for the last period of my stay at Athens. Thanks to Leon Li for offering me a job when I was in a difficult financial situation. Thanks to Lai Xu for cooking for me for the first 2 years and it's my pleasure to work with Dr Jianfeng Zhou and Dr Bingqiang Liu in the kitchen for a while. Thanks to my basketball teammates Dr Qin Ma, Dr Jianing Xu, Deli Liu, Dong Zhang, Dr Zhibin Huang, Dr Yanchun Yu, Dr Yajun Yan, Bo Feng and Chi Zhang for fighting with me and giving me the best basketball memory so far in my life. Thanks to my dragon boat teammates for winning me the first and the last champion as a captain for UGA. Thanks for Xianghao Wu and Xin li for watching football games with me. Thanks for Di Long for drawing some fabulous pictures for me as gift. Last but not the least, thanks to Jiannan Peng for sitting on the backseat of my motorcycle and risking his life to finish the great cross america tour as the first Chinese biker and giving me the best memory of my life.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

INTRODUCTION AND LITERATURE REVIEW

Purpose of the Study

Cancer is a key threat to people's health and life, accounting for ~13% of all disease-causing deaths in the world (WHO 2006). In 2007, 7.6 million people died of cancer world-wide (Dunham 2007). In the U.S, over 1.4 million new cancer cases were reported each year in the past few years, and cancer is the second leading cause of death following heart disease. Statistics from the SEER reports indicate that the mortality rate across all cancer types in the U.S. went from 195.4 per 100,000 cases in 1950, continued an upward trend till 1978 reaching 204.4, and then steadily decreased to 184.0 in 2005 (Ries LAG 2008). This decreasing trend has been mostly due to the improved diagnostic techniques for detecting the early stage of cancer as well as development of therapeutic strategy. Regarding to these two major issues, we applied a systematic study with bioinformatics and computational biology method on gene expression data of cancer.

Since the most patients are asymptomatic in the early stages of cancer, and only a few effective cancer-screening tests are clinically available. While some tests have proved to be effective in detecting cancer at its early stage, they are often too invasive, such as colonoscopy, to be routinely used during regular physicals and are currently limited to only a small number of cancer types. Often a cancer is already in an advanced stage when diagnosed; clearly, more effective techniques for early cancer detection are urgently

needed. Since traditional method of identifying novel tumor markers is labor intensive and time consuming, it has been very difficult to find markers with high sensitivity and specificity. In recent years, gene expression profiling has been a popular method for biomarker discovery (Simon 2003; Sommer and Haendler 2003; Yanagisawa, Xu et al. 2003; Rai and Chan 2004). Studies of this nature have been fruitful in identifying novel genes that are altered in expression in disease states, as the method can assess the levels of thousands of genes simultaneously.

Other than the early detection problem, another challenge is the development of new therapeutic strategy which heavily depends on the understanding the fundamental biology of tumor cell. Since the finding that cancerous cells divide fast make the milestone contribution to the development of the most popular chemotherapeutic drugs that target fast-dividing cancer cells (Li 2006). More than 50% of people diagnosed with cancer are treated with chemotherapy. The understanding of the cancer biology helped people to identify the specific feature of the tumor behavior and eventually lead the development of the cancer treatment and saves peoples life. And during the last decade the conceptual progress have been made in understanding cancer mechanistic underpinnings. This progress provides a big leap for the entire cancer research industry. The limitation of current popular treatment is obvious as a great number of patients are not successfully cured. As the entire field improving the understanding on the biology of cancer, new drugs and therapies invented to help saving patients' life and reduce their suffering.

To address the issues in these two aspects, we applied the bioinformatics and computational biology methods on gene expression data to address problems that related to those two issues. In this comprehensive study, we present a work on serum marker for cancer detection. To meet the need of deeper understanding of tumor-genesis at the molecular level, new experimental technologies were developed. To analyze and explain the experiment results, bioinformatics technics are widely used. Here we present a comparative by using the bioinformatics methods to solve the aforementioned two major problems.

Gene expression data by microarray technology and application to cancer research

Our work mainly based on the data analysis and data mining of the transcriptomic data, which is cancer gene-expression data generated from the microarray experiment. Microarray technology is introduced to the scientific community for decades. A DNA microarray is a collection of microscopic DNA spots attached to a small chip. DNA microarrays are used to measure the expression levels of large numbers of genes simultaneously. Each DNA spot contains picomoles (10−12 moles) 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 (called target) under high-stringency conditions. Probe-target hybridization is usually detected and quantified by detection of fluorophore-, silver-, or chemiluminescence-labeled targets to determine relative abundance of nucleic acid sequences in the target. Since an array can contain tens of thousands of probes, a microarray experiment can accomplish many genetic tests in parallel.

Microarray technology is proven to be among the most useful techniques for molecular biology and widely used in the field of cancer research (Guo 2003). Figure 1.1 shows a typical cDNA microarray experiment. Gene expression arrays are used to detect RNA expression levels in the cell. By comparing RNA expression levels among the samples of interest (tumor sample in certain stage or subtypes, the normal tissue as control sample), gene expression changes can be profiled on a genome-wide scale to reflect possible biologic or clinical relevance.

Current situation of the data analysis and data mining on cancer microarray data

DNA microarray technology appears to be the most comprehensive and productive approach to characterize human malignancies molecularly. Gene expression profiling of cancers expanded exponentially in the past several years and represents the largest category of research based on this technology. Gene expression profiling using DNA microarray can offer a global view of networking events in multiple genes and pathways and generate exciting new hypotheses. The power of this approach has been demonstrated in the studies of a wide variety of malignancies, including cancers of prostate, breast, liver, pancreas, ovary, stomach, lung, and head and neck (Bhattacharjee, Richards et al. 2001; Dhanasekaran, Barrette et al. 2001; Garber, Troyanskaya et al. 2001; Tonin, Hudson et al. 2001; Al Moustafa, Alaoui-Jamali et al. 2002; Belbin, Singh et al. 2002; Chen, Cheung et al. 2002; Han, Bearss et al. 2002; Hedenfalk, Ringner et al. 2002; Hippo, Taniguchi et al. 2002; Luo, Dunn et al. 2002). These microarray studies have revealed a large set of genes differentially expressed between cancerous and normal

cells, including those genes known to be important for neoplastic transformation. Although the information obtained through these studies have contributed to a better understanding of tumor genesis, the potential of gene expression profiling has not been fully realized because of the lack of knowledge about the functions of many genes and the lack of more adequate bioinformatics or statistical tools.

Feature of the microarray data and analysis methods

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. Normalization methods may be suited to specific platforms and, in the case of commercial platforms, the analysis may be proprietary. Algorithms that affect statistical analysis include: (A) Image analysis: gridding, spot recognition of the scanned image (segmentation algorithm), removal or marking of poor-quality and low-intensity features (called *flagging*). (B) Data processing: background subtraction (based on global or local background), determination of spot intensities and intensity ratios, visualization of data, and log-transformation of ratios, global or local normalization of intensity ratios, and segmentation into different copy number regions using step detection algorithms. (C) Identification of statistically significant changes: T-test, ANOVA, Bayesian method (Ben-Gal, Shani et al. 2005) Mann–Whitney test methods tailored to microarray data sets, which take into account multiple comparisons (Leung and Cavalieri 2003) or cluster analysis(Priness, Maimon et al. 2007). These methods assess statistical power based on the variation present in the data and the number of experimental replicates, and can help minimize Type I and type II

errors in the analyses (Wei, Li et al. 2004). (D) Network-based methods: Statistical methods that take the underlying structure of gene networks into account, representing either associative or causative interactions or dependencies among gene products.

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 standar deviations of genes with similar expression levels in an effort to compensate for insufficient replication (Jain, Thatte et al. 2003).

The systematic study of cancer gives new insight

To make improvement on the two aforementioned fields in the cancer research, we present a comprehensive computational study, based on public microarray gene-expression data, on solving the following related issue: 1) a prediction of serum markers for seven major internal cancer types as for early detection purpose and a systematic analysis on the cancer related cancer hallmarks. For understanding fundamental cancer biology in purpose of development of new therapy that target on the tumor specific feature, 2) a metabolic study on skin cancers to reveal a novel boosted oncogenic metabolism for the melanoma and 3) a micro-environmental study on six cancer types to reveal a complicated up-regulation of cell de-acidification mechanism in tumor. The results suggest promising detection and drug target for cancer and bring new insights of the cancer biology.

Figures

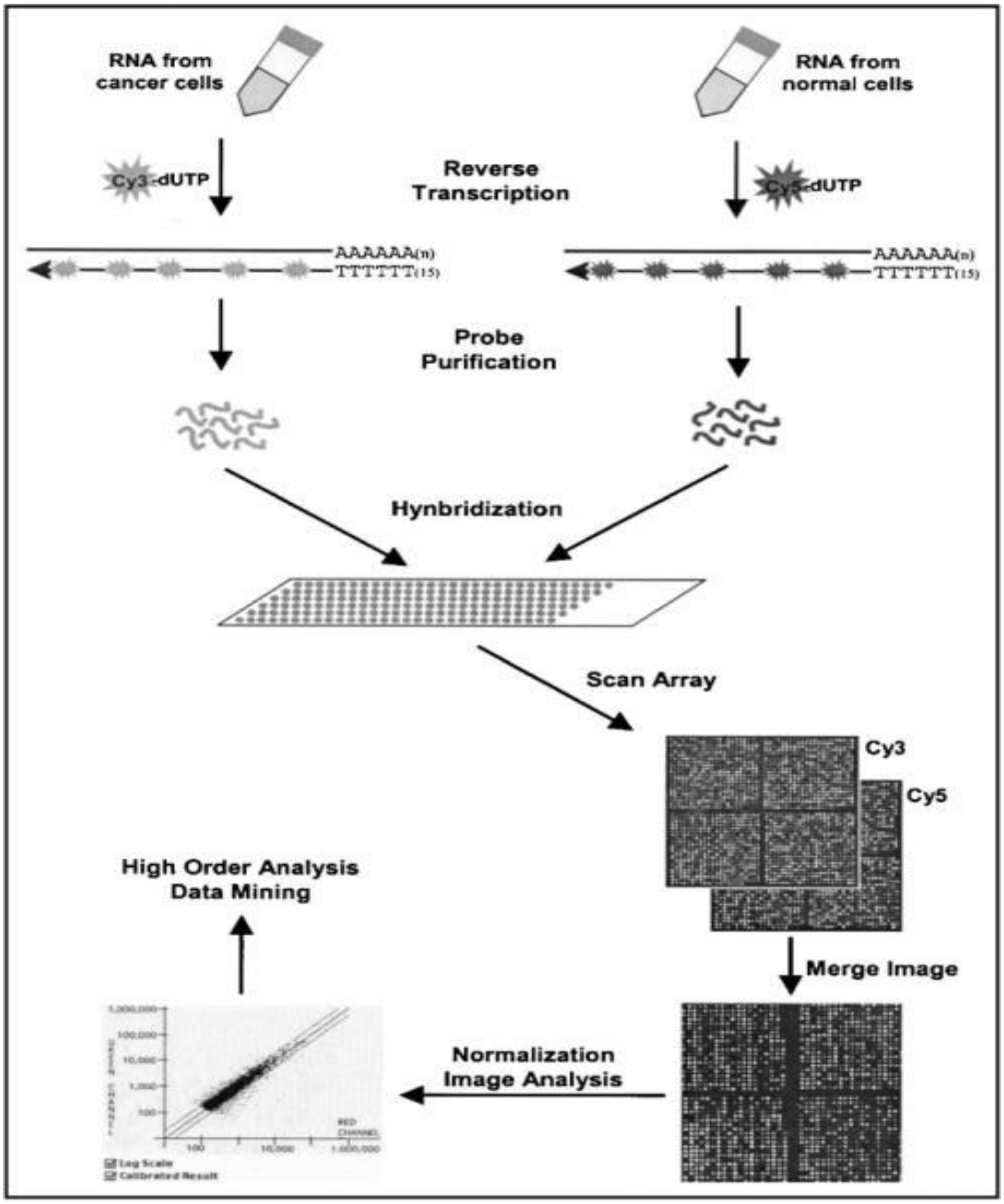


Figure 1: Schematic of a cDNA microarray experiment. (Guo 2003)

CHAPTER 2

A COMPARATIVE ANALYSIS OF GENE-EXPRESSION DATA OF MULTIPLE CANCER TYPES[1]

[1]Xu K and Cui J. et al. 2010. PLoS ONE 5(10): e13696. doi:10.1371/journal.pone.0013696

Abstract

A comparative study of public gene-expression data of seven types of cancers (breast, colon, kidney, lung, pancreatic, prostate and stomach cancers) was conducted with the aim of deriving marker genes, along with associated pathways, that are either common to multiple types of cancers or specific to individual cancers. The analysis results indicate that (a) each of the seven cancer types can be distinguished from its corresponding control tissue based on the expression patterns of a small number of genes, e.g., 2, 3 or 4; (b) the expression patterns of some genes can distinguish multiple cancer types from their corresponding control tissues, potentially serving as general markers for all or some groups of cancers; (c) the proteins encoded by some of these genes are predicted to be blood secretory, thus providing potential cancer markers in blood; (d) the numbers of differentially expressed genes across different cancer types in comparison with their control tissues correlate well with the five-year survival rates associated with the individual cancers; and (e) some metabolic and signaling pathways are abnormally activated or deactivated across all cancer types, while other pathways are more specific to certain cancers or groups of cancers. The novel findings of this study offer considerable insight into these seven cancer types and have the potential to provide exciting new directions for diagnostic and therapeutic development.

Introduction

Cancer is a key threat to people's health and life, accounting for ~13% of all disease-causing deaths in the world (WHO 2006). In 2007, 7.6 million people died of cancer world-wide (Dunham 2007). In the U.S, over 1.4 million new cancer cases were

reported each year in the past few years, and cancer is the second leading cause of death following heart disease. Statistics from the SEER reports indicate that the mortality rate across all cancer types in the U.S. went from 195.4 per 100,000 cases in 1950, continued an upward trend till 1978 reaching 204.4, and then steadily decreased to 184.0 in 2005 (Ries LAG 2008). This decreasing trend has been mostly due to the improved diagnostic techniques for detecting the early stage of cancer. General survival statistics of cancer indicate that early detection and treatment are the key to longer survival across all cancer types (Ries LAG 2008).

Challenges in early cancer detection arise mainly from the reality that most patients are asymptomatic in the early stages of cancer, and only a few effective cancer-screening tests are clinically available. While some tests have proved to be effective in detecting cancer at its early stage, they are often too invasive, such as colonoscopy, to be routinely used during regular physicals and are currently limited to only a small number of cancer types. Frequently a cancer is already in an advanced stage when diagnosed; clearly, more effective techniques for early cancer detection are needed.

A number of genetic markers have been proposed for various cancers, such as BRCA1 and BRCA2 for familial breast cancer (Ford, Easton et al. 1998) and CDH1 (CD324 or E-cadherin) for gastric cancer (Guilford, Hopkins et al. 1998). Recent studies have identified a number of promising serum markers for cancer that are being used clinically (Diamandis 2004). Among them, PSA (prostate-specific antigen) is probably the most well known and has been widely used for diagnosing prostate cancer through blood tests

(Catalona, Smith et al. 1995). However, the effectiveness of PSA in prostate cancer detection is far from adequate, widely considered as having a false positive rate that is too high to be a reliable indicator (Dhanasekaran, Barrette et al. 2001; Lilja, Ulmert et al. 2008). Similar observations have been made about other serum markers such as CA125 for ovarian cancer (Matei, Graeber et al. 2002).

Herein we present a computational study on prediction of both genetic and serum markers for seven cancer types, as well as for groups of these cancers, based on public microarray gene-expression data and a computer program for prediction of blood-secretory proteins, which we previously developed (Cui, Liu et al. 2008). Compared to earlier studies on cancer marker identification, the present study has the following key unique features: (i) a focus on identification of multi-gene markers that was achieved through exhaustive analysis of all possible combinations of genes, taking full advantage of available high-level computing power, rather than using heuristic approaches that may not necessarily find the optimal markers; (ii) an attempt to find markers for groups of cancers in addition to those for individual cancers; and (iii) an attempt to link the information derived from transcriptomic data of tissues to marker prediction in serum using a novel prediction program (Cui, Liu et al. 2008). In addition, our pathway enrichment analysis is also focused on identification of pathways that are abnormally activated or deactivated across multiple types, with the aim of identifying commonalities and uniqueness among different groups of cancers. It is anticipated that these novel data will prove highly valuable in elucidating the genetic alterations in various cancers, as well as offering potential directions for new approaches in diagnostics and therapeutics.

Results

This study is focused on seven cancer types, namely breast, colon, kidney, lung, pancreas, prostate and stomach, which are chosen because there are large sets of microarray gene-expression data in the public domain, collected on a genome scale from tissues for each of these cancer types as well as from their corresponding control tissues. By working on multiple cancer types simultaneously, our goal is to derive potential markers either specific to individual cancer types or general to all or groups of cancers, as well as to identify abnormally activated or deactivated pathways across all cancer types or some groups of cancers.

1. Predicted marker genes for individual cancer types

We have searched for individual genes and gene combinations whose expression patterns can best distinguish between cancer and associated reference tissues for each of the seven cancer types considered in this study. Specifically, all 1-, 2-, 3- and 4-gene combinations encoded in the human genome were ranked in terms of their discerning power in distinguishing the cancer samples from the corresponding reference samples for each cancer type. In addition, we have also ranked *k*-gene combinations, based on their discerning power between early cancer samples and control samples if the relevant data are available and sufficiently large. Throughout the remainder of this paper, *k*-gene groups refer to combinations of *k*-genes for k = 1, 2, 3, 4 unless stated otherwise.

A: Breast cancer: The analysis was done on a gene-expression dataset consisting of 43 paired breast cancer and cancer-adjacent reference tissues from the same patients (Pau

Ni, Zakaria et al. 2010). Of the 43 samples, 32 were early-stage cancers (stages I and II). 314 genes were found to be consistently and abnormally expressed with at least a 2-fold change in their expression across the cancer and the reference tissues in our training data, 88 of which were up-regulated and 226 down-regulated in the cancer tissues. For the differentially expressed genes, our prediction program (Cui, Liu et al. 2008) indicates that 76 of the encoded proteins are secreted and could thus serve as potential serum biomarkers(Appendix Table A2.1).

An analysis of the microarray data was then conducted, with the goal of identifying *k*-gene combinations whose expression patterns can accurately distinguish between the cancer and the reference samples. For this, a linear classifier for each *k* was trained on the microarray data. Figure 2.1 (a) and (c) show the classification accuracies of the best 100 *k*-gene combinations on the whole training set and on the training set containing only of early stage samples, respectively. An independent evaluation set is used to assess the performance of the trained classifier, which consists of 31 breast cancer and 27 cancer-adjacent reference samples from the same patients (some cancer-adjacent samples of the patients are missing) (Pedraza, Gomez-Capilla et al. 2010), of which 12 are early stage samples paired with corresponding control samples. Figure 2.1 (b) and (d) show the classification performance by the trained classifiers on the evaluation set. The detailed list of these 100 *k*-gene combinations is given in Appendix Table A2.1.

As can be seen in Figure 2.1, the majority of the top *k*-gene combinations, particularly for $k > 1$, perform well on the independent test sets, although their ranking orders, derived

based on the training data, may not be well preserved on the test sets. We believe that the fluctuations in their classification accuracies on the test sets by the trained classifiers are caused by the limited training data so some *k*-gene combinations did not generalize well to the test set.

The best three single gene discriminators are PCOLCE2, ANGPTL4 and LEP, having 88.4%, 88.4% and 87.2% classification accuracy on the training set and 94.8% , 84.5% and 96.6% on the test set, respectively. The top three 2-, 3- and 4-gene combinations are (Al Moustafa, Alaoui-Jamali et al. 2002), {RRM2+COL1A1+PCOLCE2, RRM2+COL1A1+PPARG, RRM2 + STBD1 + MAOA}, and {RRM2+COL1A1+GPR109B+IGJ, RRM2+COL1A1+GPR109B+IGJ, RRM2+COL1A1+GPR109B + SPINT2}, respectively. Similarly, for early breast cancer, the best three *k*-gene discriminators are {GPR109B, PCOLCE2, PCSK5}, {PCSK5+COL10A1, FERMT2+SPINT2, MAOA+IGJ}, {COL1A1+PCSK5+TF, GPX3+COL1A1+SPINT2, GPX3+FAP+TMEM97}, and {RRM2+COL1A1+GPR109B+IGJ, RRM2+COL1A1+GPR109B+IGJ, RRM2+ COL1A1+ GPR109B+SPINT2}, respectively.

Among these top discriminators, some have been considered as possible breast cancer marker genes by previous studies. For example, ADIPOQ (adiponectin) is found to be closely associated with a breast-cancer risk (Miyoshi, Funahashi et al. 2003). The SPINT2, an inhibitor of HGF activator, was reported to have higher expression levels in early stage breast cancer and associated with a poor prognosis (Parr, Watkins et al. 2004),

consistent with our findings. Some others are involved in the activities of cancer cells in general (Hanahan and Weinberg 2000). For example, CAV1, down-regulated in the cancer samples, was found to inhibit breast cancer growth and metastasis (Sloan, Stanley et al. 2004); the down-regulation of PPARG is associated with local recurrence and metastasis in breast cancer (Jiang, Douglas-Jones et al. 2003); and ANGPTL4 may act as a regulator of angiogenesis.(Le Jan, Amy et al. 2003). Other top discriminators represent new discoveries. For example, MAOA, ACSM5 and GPR109B have not been reported related to cancer. To the best our knowledge, all the 2-, 3- and 4-gene discriminators are novel predictions.

Similar analyses have been carried out on six other cancer types. For each cancer type, a training dataset was collected, and a linear classifier was trained on the dataset for each *k.* The key findings on each of these six cancer types are highlighted below, with the summary being given in Appendix Table A2.1 and other details in Appendix Tables A2.2 – A2.7.

B. Colon cancer: Our analysis was done on a microarray dataset consisting of 53 colon cancer and 28 cancer-adjacent reference tissues from the same patients (some of the cancer samples have no reference samples) (Ki, Jeung et al. 2007). 248 genes were found to be consistently and abnormally expressed with at least a 2-fold change in their expression across the cancer and the reference tissues in our training data, 56 of which are up-regulated and 192 are down-regulated in colon cancer tissues. An independent set, consisting of 24 colon cancer and 24 cancer-adjacent reference samples from the same

patients (Jiang, Tan et al. 2008) was used to test the performance of the trained classifier. Figure 2.2 shows the classification accuracies by the best 100 *k*-gene discriminators on both the training and the testing sets.

The best three single-gene discriminators are MMP7, DPT and MMP1 having 97.5%, 96.3% and 95.1% classification accuracy on the training set and 97.9% , 97.9% and 91.7% on the test set, respectively. The top three 2-gene discriminators are MMP7+FAM107A, FRZB+MMP7, and SLIT3+MMP7 (we did not try k > 2 since the best 2-gene combinations already gives 100% classification accuracy). Some of our top discriminators have been previously studied in the context of colorectal cancer. For example, MMP1 is an invasion-promoting factor, and its up-regulation, as observed in our data, is associated with the invasiveness of the cancer (Behrens, Mathiak et al. 2003). MMP7 is known to play an important role in cancer growth, and its up-regulation could be a key mechanism for cancer cells' escape from the immune surveillance (Wang, Chen et al. 2006). FRZB, a down-regulated gene, is annotated to function in the negative regulation of the Wnt signaling pathway, which promotes tumor growth. Other top discriminators represent new findings. For example, ADAMDEC1, down-regulated in tumor tissue, is a gene moderately expressed in the colon; and it may play a role in the immune response but has never been reported as being related to colon cancer.

C. Kidney cancer: The analysis was carried on a microarray gene-expression dataset consisting of 49 kidney cancer and 23 cancer-adjacent reference tissue samples from the same patients (Jones, Otu et al. 2005). 232 genes were found to be consistently and abnormally expressed with at least a 2-fold change in their expression across the cancer

and reference tissues in our training data, 130 of which are up-regulated and 102 are down-regulated in cancer. An independent evaluation set, consisting of 35 kidney cancer samples and 12 cancer-adjacent reference samples from the same patients was applied (Someya, Yamasoba et al. 2008). Figure 2.3 shows the classification accuracies by the top *k*-gene discriminators on both the training and the testing sets.

The best three single gene discriminators are CCL18, ACPP and UMOD, having the same classification accuracy, 98.6% on the training set and 89.4%, 95.7% and 100% on the test set, respectively. The top three 2-gene combinations are EGF+ALB, ACPP+UMOD, and UMOD+ALB. Among the top discriminators, UMOD has been reported to be related to kidney disease (Hart, Gorry et al. 2002). SERPINA5, down-regulated in the cancer, regulates the invasive potential of renal cancer growth and invasion. Other top discriminators represent new discoveries. For example, AFM has not been reported to be related to cancer, and C6orf155 does not have a characterized function.

D. Lung cancer: The analysis was done on a microarray dataset consisting of 58 lung cancer tissue and 49 cancer-adjacent reference tissue samples from the same patients (Landi, Dracheva et al. 2008). 700 genes were found to be consistently and abnormally expressed with at least a 2-fold change in their expression across the cancer and reference tissues in our training data, 259 of which are up-regulated and 441 are down-regulated in lung cancer tissues. An independent set, consisting of 27 lung cancer and 27 cancer-adjacent reference samples from the same patients (Su, Chang et al. 2007), was used to

assess the performance of our trained classifier. Figure 2.4 shows the classification accuracies by the top 100 *k*-gene discriminators on both the training and the testing sets. The best three single gene discriminators are CAV1, SFTPC and TNXB, having the same classification accuracy, 99.1% on the training set and 98.2%, 96.3% and 94.4% on the test set, respectively. The top three 2-gene combinations are FERMT2+GREM1, TEK+NFASC, CAV1+MMP1. Among the top discriminators, CAV1 has been found to be down-regulated in breast cancer (Park, Kim et al. 2005), and has been reported to be associated with metastasis in lung cancer (Ho, Huang et al. 2002). SFTPC has been reported to be associated with interstitial lung disease (Bridges, Wert et al. 2003). FAM107A, which suppresses cell growth, may play a role in cancer development (Kholodnyuk, Kozireva et al. 2006). CD93 is believed to be involved in intercellular adhesion and in the clearance of apoptotic cells (Ikewaki, Tamauchi et al. 2007). NMU, up-regulated in lung cancer, is a promoter for cancer formation and a promoter for lung cancer metastasis and cancer cachexia (Wu, McRoberts et al. 2007). Other top discriminators represent new observations. For examples, SPP1 and EMCN have not previously been reported as cancer-related.

E. Pancreatic cancer: The analysis was done on a microarray dataset consisting of 39 paired pancreatic cancer and cancer-adjacent reference tissue samples from the same patients (Badea, Herlea et al. 2008). 969 genes were found to be consistently and abnormally expressed with at least a 2-fold change in their expression across the cancer and reference tissues in the training data, 690 of which are up-regulated and 279 are down-regulated in pancreatic cancer. An independent set, consisting of 36 lung cancer

samples and 16 cancer-adjacent reference samples from the same patients (Pei, Li et al. 2009), was used to assess the classification performance of our trained classifier. Figure 2.5 shows the classification accuracies by the top 100 *k*-gene discriminators on both the training and the testing sets.

The best three single-gene discriminators are KRT17, COL10A1 and FAM19A5, having the same classification accuracy, 93.6% on the training set and 88.5% , 84.6% and 84.6% on the test set, respectively. The top three 2- and 3-gene discriminators are {MMP7+AZGP1; MMP7+ELA3B; MMP7+FGL1} and {COL8A2+SGPP2+CCL18; COL8A2+PMEPA1+TMEM45B; LCN2+COL8A2+PMEPA1}, respectively. Among the top discriminators, KRT17 is known to be involved in tissue repair, as is CTHRC1 (Tang, Dai et al. 2006). AZGP1 has been reported to cause extensive loss of fat, often associated with advanced cancers (Groundwater, Beck et al. 1990; Bing, Bao et al. 2004). ELA3B has been proposed as a pancreatic cancer marker (Shimada, Yamaguchi et al. 2002). PLA2G1B regulates the inhibition of pancreatic phospholipase a2, which is involved in the uptake of dietary fat (Pan and Bahnson 2007). Other top discriminators represent new findings. For examples, RSAD2, involved in antiviral defense, has not been reported as being related to cancer, as well as SGPP2, known to be involved in pro-inflammatory signaling (Mechtcheriakova, Wlachos et al. 2007), and CST4.

F. Prostate cancer: The analysis was done on a microarray dataset consisting of 65 prostate cancer and 63 cancer-adjacent reference tissue samples from the same patients (Chandran, Dhir et al. 2005). 139 genes were found to be consistently and abnormally

expressed with at least a 2-fold change in their expression across the cancer and reference tissues in our training data, of which 44 are up-regulated and 95 are down-regulated in lung cancer tissues. We then used an independent set, consisting of 62 prostate cancer samples and 47 cancer-adjacent reference samples from the same patients (Lapointe, Li et al. 2004). Figure 2.6 shows the classification accuracies by the top 100 *k*-gene discriminators on both the training and testing sets.

The best three single gene discriminators are CRISP3, MYLK and PALLD, having 75.8%, 73.4% and 71.9% classification accuracy on the training set and 72.5% , 83.5% and 69.6% on the test set, respectively. The top three 2- and 3-gene discriminators are {LTF+IGF1; LTF+SPARCL1; SMTN+CCK}, {SMTN+CCK+CCL2; SMTN+CCK+COMP; SMTN+CCK+PLA2G7}, respectively. Among the top discriminators, CRISP3 has been reported to be a potential prostate cancer marker and is up-regulated in the prostate cancer tissues (Kosari, Asmann et al. 2002). LTF is known to inhibit the growth of tumors (Varadhachary, Wolf et al. 2004). IGF1, a growth factor, plays a role in the development of prostate cancer (Soulitzis, Karyotis et al. 2006) and has been reported as an indicator of advanced prostate cancer (Chan, Stampfer et al. 2002). EDNRA is known be relevant to the progression of prostate cancer (Akhavan, McHugh et al. 2006). Other top discriminators represent new discoveries. For example, CHRDL1 may play a role in regulating angiogenesis (Kane, Godson et al. 2008) but has not been reported to be cancer-related to cancer. The same is with SMTN.

G. Stomach cancer: The analysis was done on a microarray dataset consisting of 89 stomach cancer and 23 cancer-adjacent reference tissues from the same patients (Chen, Leung et al. 2003). Out of the 89 cancer tissue samples, 31 are early-stage cancers. 336 genes were found to be consistently and abnormally expressed with at least a 2-fold change in their expression across the cancer and reference tissues in our training data, 156 of which are up-regulated and 180 are down-regulated in lung cancer tissues. An independent set, consisting of 38 stomach cancer samples and 31 cancer-adjacent reference samples from the same patients (D'Errico, de Rinaldis et al. 2009) was used to assess the performance of our trained classifier, of which 12 are early stage samples partially paired with 10 reference samples. Figure 2.7 shows the classification accuracies by the top 100 *k*-gene discriminators on both training and testing sets.

The best three single-gene discriminators are SERPINH1, BGN and COL12A1, having 99.1%, 98.2% and 98.2% classification accuracy on the training set and 94.2% , 88.4% and 84.1% on the test set, respectively. The top three 2-gene combinations are CHGA+SERPINH1, PGC+SERPINH1 and TGFBI+CHGA, respectively. For early stomach cancer, the best three *1*-gene discriminators are also SERPINH1, BGN and COL12A1, respectively. Among the top discriminators, BGN is known to have a role in controlling cell growth in cancer (Chen, Lenschow et al. 2002). The abnormal expression of CTHRC1, a regulator of matrix deposition, has been widely found across different solid cancers and is considered to be associated with cancer invasion and metastasis (Tang, Dai et al. 2006). NID2, which inhibits nidogen expression, has a potential pathogenic role in gastrointestinal cancer (Ulazzi, Sabbioni et al. 2007). SPARC, a

regulator of cell growth, is known to be associated with the development of gastric cancer (Wang, Lin et al. 2004). Of particular interest is that PGC has been proposed as an indicator of gastric cancer (Ning, Sun et al. 2004), and the serum level of PGC has been used as a biomarker for precancerous lesions of the stomach (Broutet, Plebani et al. 2003). Other top discriminators represent new discoveries. For example, ABCA5, ADAMTS12 and CLEC3B have not been reported to be cancer related.

Interestingly, the number of differentially expressed genes across different cancer types has a wide spread[2], ranging from 139 (prostate), 232 (kidney), 248 (colon), 249 (breast), 336 (stomach) to 554 (lung) and 733 (pancreatic). One possible explanation is that these numbers may reflect the aggressiveness of the corresponding cancers. We did notice that there is strong correlation between the number of differentially expressed genes in a given cancer type and the five-year survival rate of patients with that cancer (CancerFact 2006) (detailed statistics in Table 2.1), as shown in Figure 2.8 Another interesting observation is that, while the majority of the differentially expressed genes with at least a 2-fold change in five cancer types (breast, colon, lung, prostate, stomach) are down-regulated, in kidney and pancreatic cancers, the majority of such genes are up-regulated, possibly suggesting unique characteristics of these two cancer types.

[2] While the measured expression levels of most genes, including 135 house-keeping genes, vary substantially across different types of cancers, the relative gene expression changes for each cancer *versus* its reference tissues from different datasets were found to be consistent. Hence, the 2-fold change cutoff in gene-expression level changes has the same meaning across the seven cancer datasets, i.e., comparing the numbers of the genes with at least 2-fold expression-level changes is meaningful.

2. Markers for multiple cancer types

We have also sought to identify genes that could be used as indicators for cancer in general or for a group of cancers. It is possible to find common gene "markers" across different cancer types because of the observation that the majority of the cancers, if not all, undergo a common set of alterations (Hanahan and Weinberg 2000) during oncogenesis, namely (a) self-sufficiency in growth signals, (b) insensitivity to antigrowth signals, (c) evasion of apoptosis, (d) limitless replication potential, (e) sustained angiogenesis and (f) tissue invasion and metastasis. Some of these biological processes may be executed by the same groups of proteins during the formation and progression of different cancers, hence possibly giving rise to common markers for different cancer types.

A. Identification of genes differentially expressed across multiple cancer types: We have examined differentially expressed genes with at least 2-fold changes between cancer and reference tissues across all seven cancer types and attempted to find those genes common to multiple cancer types. The key findings are summarized in Table 2.2.

As can been seen from Supplementary Table 2.3, 92 genes are differentially expressed (same direction of regulation) across at least three cancer types, among which 20 genes are across at least four cancer types, four genes (ABCA8, DPT, FHL1 and TOP2A) across five cancer types and one gene, CDC2, across six cancer types. The differences in the gene expression across different cancer types may indicate either a general or a specific relevance of a gene to these types of cancers, which has been partially confirmed

by the functional analysis and an extensive literature search. The detailed molecular function of these genes is summarized in Table 2.3. 67 out of the 92 genes have been reported to be cancer associated in previous studies. For example, CDC2, up-regulated in six of the seven cancers studied, has been reported to be related to colon (Nozoe, Honda et al. 2003), prostate (Chen, Xu et al. 2006) and stomach cancer (Masuda, Inoue et al. 2003), which is not surprising in view of its role in regulating the cell cycle, e.g. entry from G_1 to S; TOP2A, again up-regulated in six of seven cancers, has been reported to be associated with gastric (Varis, Zaika et al. 2004), breast (Koren, Rath-Wolfson et al. 2004) and ovarian cancer (Chekerov, Klaman et al. 2006), consistent with a function in DNA strand regulation; RRM2, up-regulated in four of the seven cancers, has been suggested to be related to esophageal and gastric cancers and prostate cancer (Kolesar, Huang et al. 2009), consistent with its critical role in DNA synthesis which must be maintained in rapidly dividing cells; And LCN2, up-regulated in four of the seven cancers, has been reported in breast (Bauer, Eickhoff et al. 2008), colon (Lee, Lee et al. 2006) and pancreatic cancer (Tong, Kunnumakkara et al. 2008). The function of LCN2 is believed to be involved in transport into cells, consistent with maintaining sufficient substrates for metabolically active cancer cells. Of the 92 genes, 49 have been reported to be relevant to immune diseases, such as CXCL12, COL1A1, MMP9, CD36 and ALOX5 (Aota, Sumi et al. 2004; Lee, Kim et al. 2005; Piovan, Tosello et al. 2005; Herb, Thye et al. 2008), likely reflecting an inflammatory–type response often associated with cancer. In addition, MMP9, important in extracellular matrix degradation, is up-regulated in three of the seven cancers, and CD36, which may function in cell adhesion, is down-regulated

in three of the seven cancers; both of these changes are consistent with a role of the gene products in metastasis.

B. Pathway enrichment analysis of differentially expressed genes: We have carried out a pathway-enrichment analysis on genes that are differentially expressed in any of the seven cancer types. Overall, a number of signaling pathways are consistently and highly enriched across all seven types of cancers, such as Wnt, p53 and integrin signaling pathways, as well as a few other processes like phospho-APC/C-mediated degradation of cyclin A and inflammation determined by chemokine and cytokine signaling pathways (in addition to the general cellular processes such as cell cycle, DNA replication and repair, apoptosis and various metabolic pathways). Notably, these pathways are mostly enriched with up-regulated genes in cancer, indicating a possible activation of these processes. In addition, a few metabolic pathways such as tyrosine, histidine, phenylalanine, butanoate and 5-hydroxytryptamine pathways are enriched only with down-regulated genes across all cancers. This may indicate a possible deficiency of the relevant metabolic enzymes in cancer, which could for example arise from loss-of-function mutations in their genes. These observations may suggest the essential roles played by these processes in cancer formation and progression. Note that the increased enzymatic activity of histidine decarboxylase (HDC) has been observed in colorectal cancer (Garcia-Caballero, Neugebauer et al. 1988). This is opposite to our observation, so additional information is clearly needed regarding the levels of activities of these processes across different subtypes/stages of a cancer.

Other than the above processes common to all cancers, a few pathways are enriched only in specific cancers. For example, arginine, proline, glutamate and riboflavin (vitamin B2) metabolism are enriched with up-regulated genes only in lung cancer; folate biosynthesis and nitrogen metabolism pathways are enriched in breast cancer; formyltetrahydroformate biosynthesis in stomach cancer; and NF-kB activation and Csk activation by cAMP-dependent protein kinase inhibits signaling through T-cell receptor in kidney cancer. Of particular interest is the finding that, compared to other cancer types, pancreatic cancer has the greatest number of differentially expressed genes involved in a complex network consisting of the EGF signaling pathway, purine and aminosugar metabolism, PKC-catalyzed phosphorylation of inhibitory phosphoprotein of myosin phosphatase, metabotropic glutamate receptor group II pathway, Fc epsilon receptor I signaling and the BCR and IL 4 signaling pathways. This suggests a highly active state of the underlying cells in terms of cell growth, differentiation, invasion and metastasis, consistent with the aggressiveness of the cancer. Seeking the genes and their products that are responsible for the more aggressive behaviors of pancreatic cancer may provide new targets for treating the cancer or preventing the cancer from progression.

A number of pathways specific to a group of cancers have also been identified, which may suggest common characteristics of the underlying neoplasms. For example, the glutathione metabolic pathway is enriched across five cancer types, excluding breast and prostate cancer; *E. coli* infection-related pathways are activated in kidney, lung, pancreatic and stomach cancers but not in other cancers; the thyrotropin-releasing hormone receptor signaling pathway is activated in pancreatic and kidney cancer, but not

in the other five cancers; and steroid biosynthesis is activated in breast, lung and pancreatic cancer but not in the other four cancers. Cancer-specific pathway activations have been previously reported. For example, the thyrotropin-releasing hormone receptor signaling pathway was reported to promote programmed cell death in pancreatic cancer (Mulla, Geras-Raaka et al. 2009); steroid biosynthesis in pancreatic cancer was found based on analyses of several steroidogenic enzymes, such as the cytochrome P-450scc enzymatic complex (P450scc) that is responsible for the conversion of cholesterol into pregnenolone (Morales, Cuellar et al. 1999). These diverse findings indicate that comparative analyses of cancer microarray data can reveal interesting and undetected relationships across different cancer types/subtypes, thus providing useful guiding information for further investigation. The detailed pathway-enrichment information across different cancer types is summarized in Table 2.4.

C. Top k-gene markers for multiple cancer types: We have examined the *k*-gene combinations among genes that are differentially expressed in each cancer type to find gene combinations that are common to multiple cancer types. The idea is to identify commonalities of gene combinations with differential expression patterns between cancer and reference tissue across multiple cancer types, which could provide useful information about common underlying mechanisms of carcinogenesis of different cancers. Supplementary Table 2.5 – 2.7 gives the detailed list of all the *k*-gene combinations with classification accuracies at least 75% across at least three cancer types.

As shown in Table 2.5, the top two 2-gene combinations, CDC2 + DPT and CDC2 + TOP2A, are found to be good markers for five types of cancers, namely breast, colon, lung, prostate and stomach cancers. Similarly, ABCA8+ALDH1A1+DPT and ABCA8+AURKA+DPT are good 3-gene markers for four types of cancers with higher classification accuracies than the top 2-gene markers, as shown in Table 2.6.

As noted, CDC2 and DPT appear in all of the top 2-, 3- and 4-gene discriminators, and, consequently, we have examined the functions of these genes. CDC has been reported to play a key role in cell proliferation (Wang, Hasham et al. 2003) and apoptosis (Ababneh, Gotz et al. 2001), and DPT is suggested to have a possible role in carcinogenesis through its interaction with a known oncogene, TGFB1. Moreover, some of the top discriminator genes have been reported to be cancer relevant. For example, ECT2 is reported to be involved in cancer development, influencing processes such as the cell cycle, apoptosis and cell division (Eguchi, Takaki et al. 2007); FABP4 is involved in the activation of the immune response and is reported to be related to breast cancer (Li, Lu et al. 2007) and bladder cancer (Ohlsson, Moreira et al. 2005); and TOP2A is involved in stomach cancer (Varis, Zaika et al. 2004). These independent observations confirm that the findings herein are meaningful.

D. Top k-gene markers that are blood secretory: Using our prediction program in conjunction with the above top gene discriminators, it is possible to identify proteins that may be secreted into circulation, thus possibly providing candidate serum marker proteins for cancer detection. Table 2.8 summarizes the top *k*-gene markers that are predicted to

have their proteins secreted into blood. Some genes involved in these top candidate markers have been previously reported to be cancer related, e.g. KLF4 and MMP7 (Wei, Gong et al. 2005; Zhang, Jin et al. 2005), and MMP7 that has been reported relevant to five out of seven cancer types in this study. Other predicted blood-secretory marker proteins such as DPT, PAICS, CHRDL1, KLF2, COL10A1 and MYL9 have not heretofore been reported to be cancer related.

While Table 2.8 gives a detailed list of all the gene combinations whose proteins are predicted to be blood secretory, with discerning power between cancer and corresponding reference tissues higher than 70%, a few top candidates for these seven cancer types are highlighted. One 2-gene combination, DPT+KLF4, covers four cancer types, namely breast, colon, lung and stomach cancer, with 70% classification accuracy. Note that DPT has not been previously found to be cancer related. Three types of cancers are covered by 22 2-gene combinations, with MMP11+RRM2 and MMP7+MMP9 representing the top 2-gene markers with at least 75% classification accuracy. The best 4-gene combination, MMP7+MMP9+MMP11+RRM2, gives at least 86% classification accuracy for lung, pancreatic and stomach cancers, and all of these four genes are up-regulated by at least 2-fold in the cancer tissues, suggesting the potential of this combination as a good blood marker for these cancer types. CCL18+TGFBI represents a good discriminator for kidney, pancreatic and stomach cancer, which are up-regulated by at least 2-fold in cancer tissues. Similarly, CN2+THBS2 are both up-regulated by 2-fold in kidney, lung and pancreatic cancer. MMP11+RRM2 are up-regulated in lung cancer, pancreatic cancer and stomach cancer tissues, and hence may also make a good marker for these

three cancer types. The summary of the top k-gene for each of the seven cancer types is in the Table 2.9.

Methods

1. Microarray gene expression data for human cancers

Microarray gene expression data were downloaded for seven cancer types, namely, breast, colon, kidney, lung, pancreatic, prostate and stomach cancer from the GEO database of NCBI (Edgar, Domrachev et al. 2002). To ensure that our prediction results can be generalized to larger datasets (i.e., not over-trained), a training data and an independent testing data were downloaded for each cancer type (Table 2.10). For each dataset, we have included the following data items: (a) (normalized) gene expression levels for each gene in the cancer tissue of each patient, (b) (normalized) gene expression levels for each gene in the control tissue of each patient and (c) stage information for the majority of the cancer samples (this information is not available for some data).

We have chosen microarray datasets normalized by RMA, which has been reported to reflect more accurately gene-expression changes due to biology compared to other normalization methods. Expression levels of 135 house-keeping genes were examined, and large variations were observed for individual genes in the reference tissues across the seven types of cancers. The distributions of the fold-changes (FC) of individual genes between cancer and corresponding reference tissues across the seven types of cancers were also checked, and it was found that the distributions are highly similar. Figure 2.9 shows one such comparison of FC distributions between breast cancer and lung cancer.

Hence we conclude that comparisons of fold-changes across different cancer datasets in our study are meaningful.

2. Identification of differentially expressed genes

For each gene in a dataset, we consider the two distributions of gene expression values across all cancer samples and across all the control samples, respectively. The Mann-Whitney test (Wilcoxin 1947) was applied first to identify those genes differentially expressed in cancer *versus* the control samples, using a p-value cutoff = 0.05. In addition, the fold-change for each gene in the cancer *versus* the control samples was calculated using the following formula:

$$FC = \overline{X_{tumor}} / \overline{X_{normal}}$$

where $\overline{X_{tumor}}$ and $\overline{X_{normal}}$ represent the mean expression level of each gene among all cancer and reference samples for each cancer type, respectively. A positive FC indicates up-regulation in cancer *versus* reference tissues, while a negative FC indicates down-regulation. Overall, we consider a gene being *differentially expressed* if the p-value of the Mann-Whitney test is < 0.05 and its fold-change is at least 2 or at most 0.5.

3. Prediction of blood secreted proteins

All genes that are predicted to be differentially expressed between cancer and corresponding reference samples were analyzed by a computer program developed to predict blood-secretory proteins (Cui, Liu et al. 2008). The basic idea of the algorithm is

that through an extensive literature search, a large number of human proteins were identified that are documented in the literature to be blood secretory; we then trained a support vector machine (SVM)-based classifier using various sequence-based features to distinguish between the blood-secretory proteins and proteins that are not secreted, using features such as signal peptides, transmembrane domains, glycosylation sites, disordered regions, secondary structural content, hydrophobicity and polarity measures. On a large independent test set containing 105 secretory proteins and 7,258 non-secretory proteins of humans, the classifier achieved ~94% prediction sensitivity and ~98% prediction specificity.

This program was applied to all annotated human proteins in Swissprot (Boeckmann, Bairoch et al. 2003), and 2,842 were predicted to be blood-secretory. To ensure that potential serum proteins are not overlooked, we have also included proteins reported to be extracellularl, which adds additional 1,277 proteins to the above predicted protein list, giving rise to a total of 3,882 proteins as potential blood secretory proteins.

4. Prediction of marker genes for each cancer type

Based on the identified differentially expressed genes, the following approach was employed to assess the discerning power of each k-gene combination in terms of classification accuracy of cancer tissue samples *versus* control samples. For each *k*-gene combination out of the differentially expressed gene list for each cancer type, an SVM-based classifier was trained to maximize the classification accuracy defined as

$$Overall\ accuracy = (TP+TN)/N,$$

where TP and NP are true positives and negatives, respectively, and N is the total number of samples. The linear kernel function was used to find the optimum linear separation plane for the SMV classifier, and the training and testing were conducted using the LIBSVM (Chang and Lin 2001) software package.

For each cancer type, a 5-fold cross-validation was done on the training data and the genes ranked according to their classification performance. In order to find markers that are generalized well to other datasets, we applied the identified gene markers on an independent testing dataset. Because a total independent testing dataset was used for each cancer type, the intent was to show that the markers have consistent performance in the testing dataset. The LIBSVM, with 5-fold cross validation, was applied to the test dataset to determine the classification accuracy. The markers based on the classification accuracy in the training dataset was ranked, and if more than 2 markers have the same training accuracy the markers with the greatest testing accuracy are ranked higher.

5. Prediction of markers for multiple cancer types

The following procedure to identify *k*-gene discriminators for multiple cancer types was employed. All of the genes that consistently exhibited differential expression in at least 2 types of cancers were collected. For each *k*-gene combination in this gene list, its classification accuracy between each cancer type and the corresponding reference tissues for k = 1, 2, 3, 4 was calculated. This was done for every cancer type. Then, the *k*-gene combinations exhibiting discerning power across multiple cancer types were

determined. By applying a fixed cut-off on classification accuracies, the top discriminators for multi-cancer types were identified.

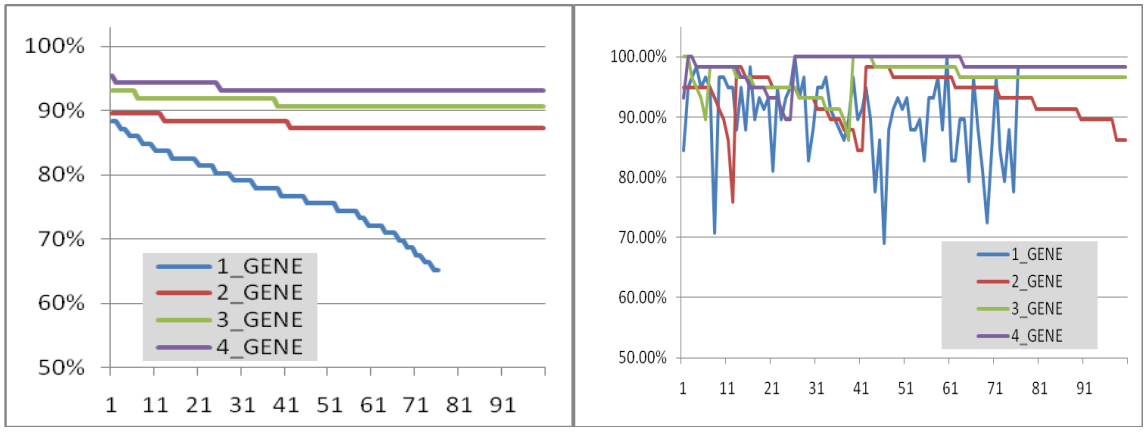
6. Pathway enrichment analysis of differentially expressed genes

Functional analysis and pathway enrichment analysis were conducted using DAVID (Dennis, Sherman et al. 2003), where the pathway information is based on the annotation from KEGG (Kanehisa and Goto 2000), BBID (Becker, White et al. 2000) and BIOCARTA (www.biocarta.com). A p-value < 0.05 was used to guarantee the significance level of the enriched pathway.

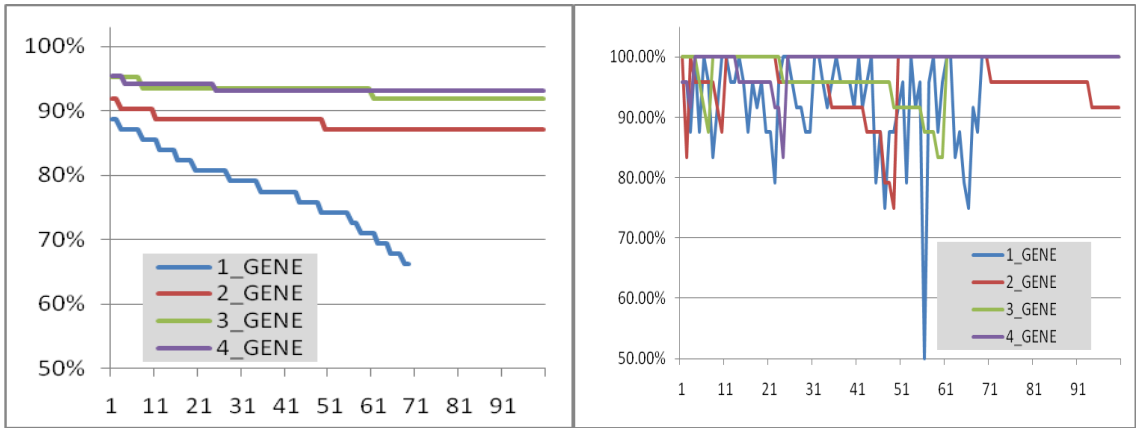
Concluding remarks

A computational protocol for predicting gene markers in cancer tissues and protein markers in serum was developed for seven cancer types. In addition to individual gene markers, we have focused on gene combinations that can be used to distinguish multiple cancer types and their corresponding reference tissues. The pathway enrichment analysis among the differentially expressed genes across multiple cancer types, as well as those specific to individual cancer types, has identified a number of abnormally activated or deactivated pathways across multiple cancers and for specific cancers. The information provided on individual genes and pathways, along with potential serum biomarkers, should provide highly useful information for elucidating pathways in cancer, as well as expediting the search for potential serum biomarkers of specific cancers.

Figures

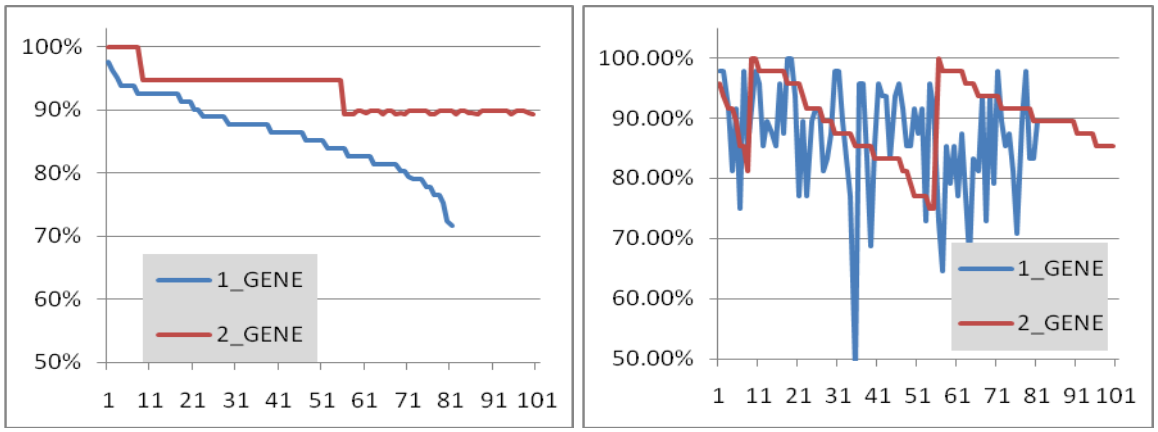


(a) (b)



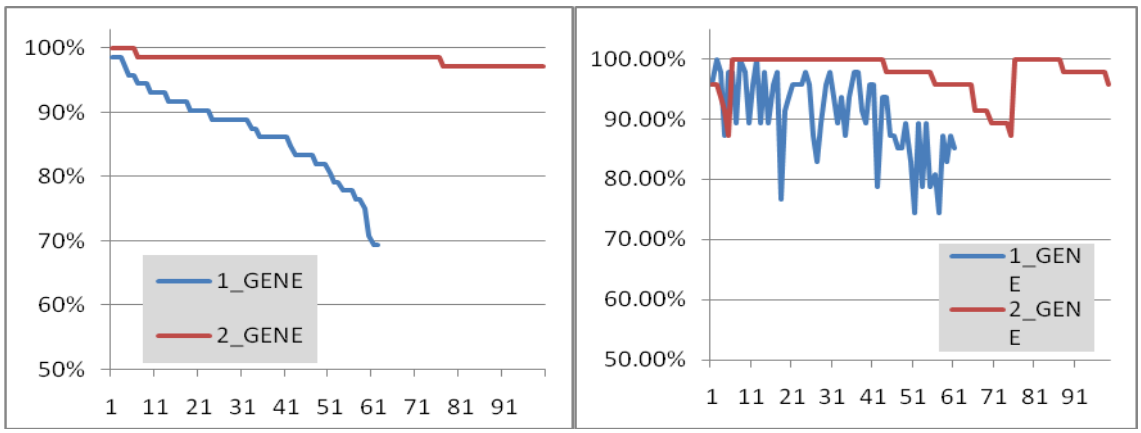
(c) (d)

Figure 2.1: Classification performance by top k-gene groups of breast cancer. Classification accuracies by the top 100 *k*-gene markers on the training and the test sets. For each panel, the x-axis is the list of 100 *k*-gene markers ordered by their classification performance on the training datasets, and the y-axis represents the classification accuracy. (a) classification accuracies by the top 100 *k*-gene combinations between breast cancer and reference samples in the training set, and (b) on the test set; (c) classification accuracies by top 100 *k*-gene combinations between early breast cancer and corresponding reference samples in the training set and (d) on the test set.



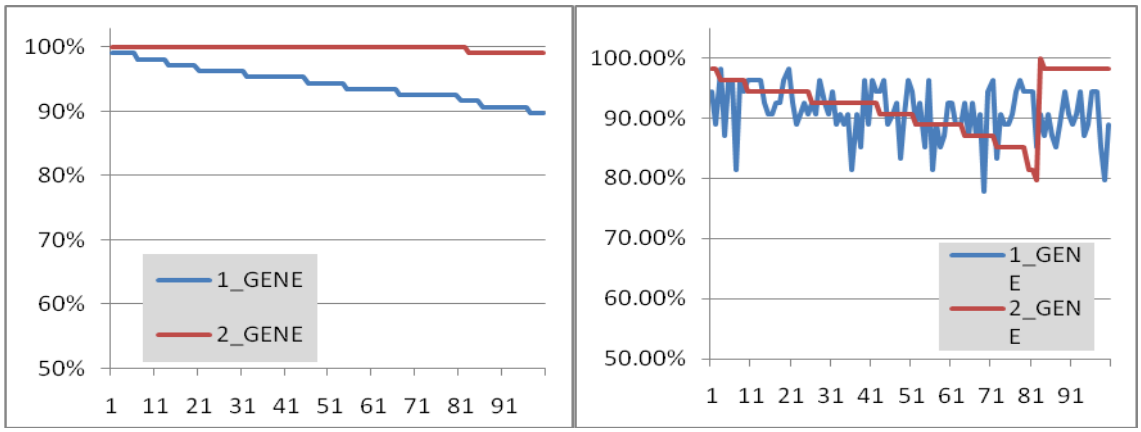
(a) (b)

Figure 2.2: Classification performance by top k-gene groups of colon cancer. (a) classification accuracies by the top 100 *k*-gene combinations between colon cancer and reference samples in the training set. (b) classification accuracies by the top 100 *k*-gene combinations on the test set.



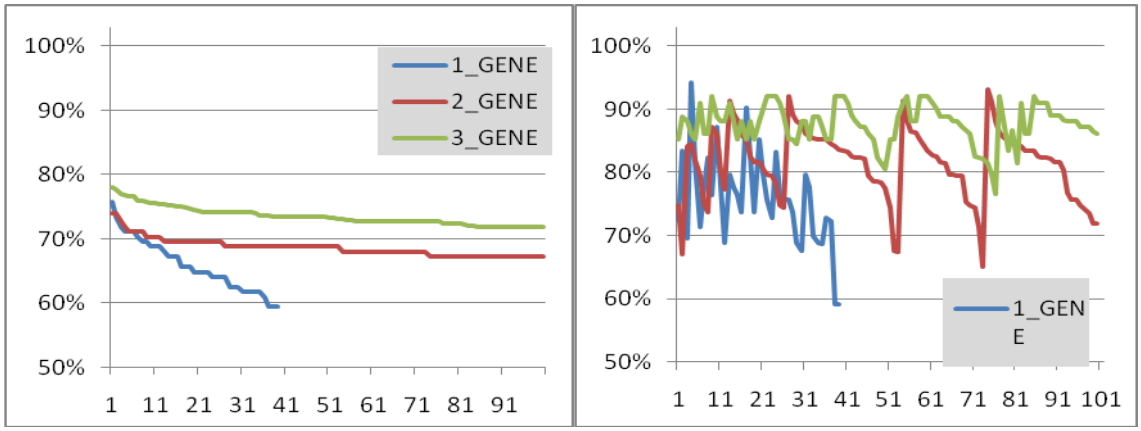
(a) (b)

Figure 2.3: Classification performance by top k-gene groups of kidney cancer. (a) classification accuracies by the top 100 *k*-gene combinations between kidney cancer and reference samples in the training set. (b) classification accuracies by the top 100 *k*-gene combinations on the test set.



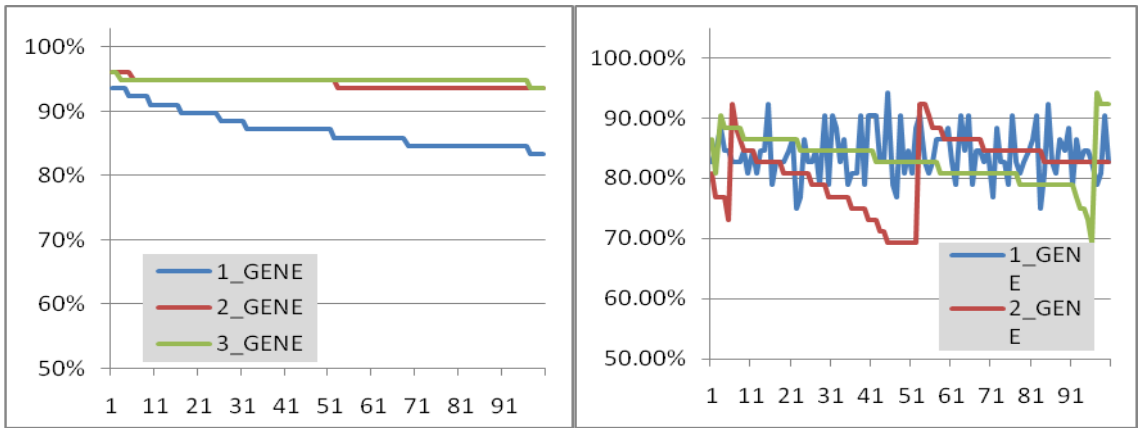
(a) (b)

Figure 2.4: Classification performance by top k-gene groups of lung cancer. (a) classification accuracies by the top 100 *k*-gene combinations between lung cancer and reference samples in the training set. (b) classification accuracies by the top 100 *k*-gene combinations on the test set.



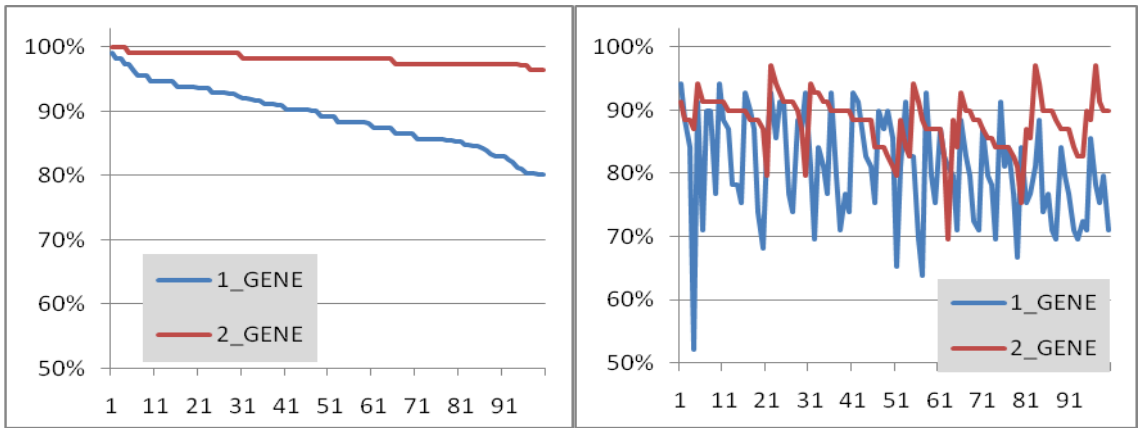
(a) (b)

Figure 2.5: Classification performance by top k-gene groups of pancreatic cancer. (a) Classification accuracies by the top 100 *k*-gene combinations between pancreatic cancer and reference samples in the training set. (b) classification accuracies by the top 100 *k*-gene combinations on the test set.

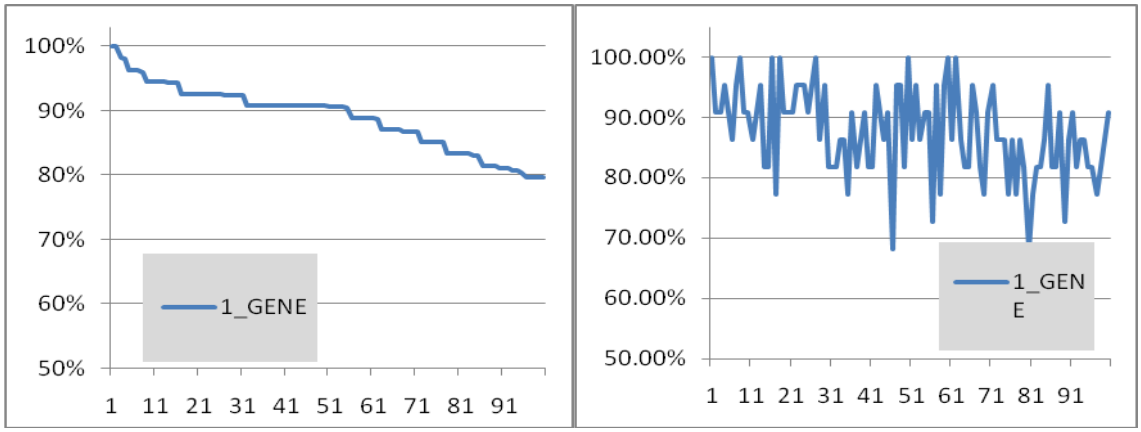


(a) (b)

Figure 2.6: Classification performance by top k-gene groups of prostate cancer. (a) classification accuracies by the top 100 *k*-gene combinations between prostate cancer and reference samples in the training set. (b) classification accuracies by the top 100 *k*-gene combinations on the test set.

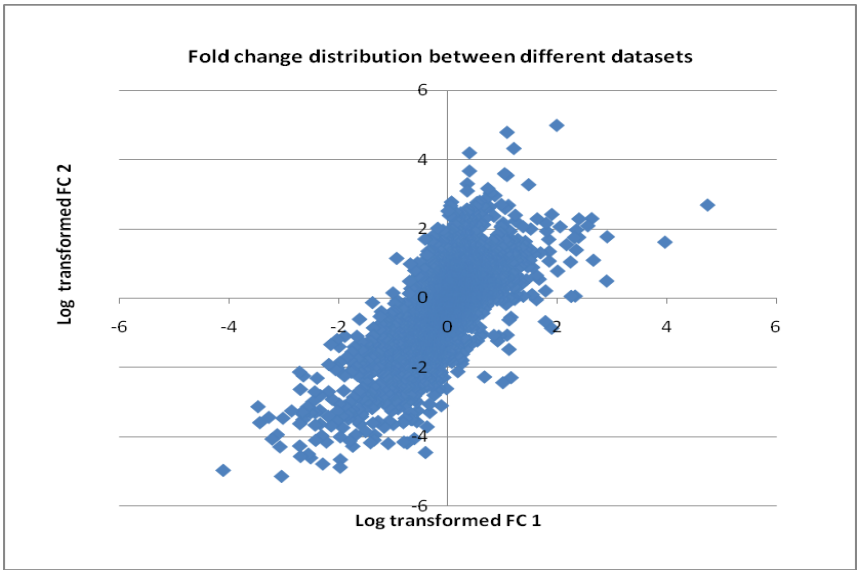


(a) (b)

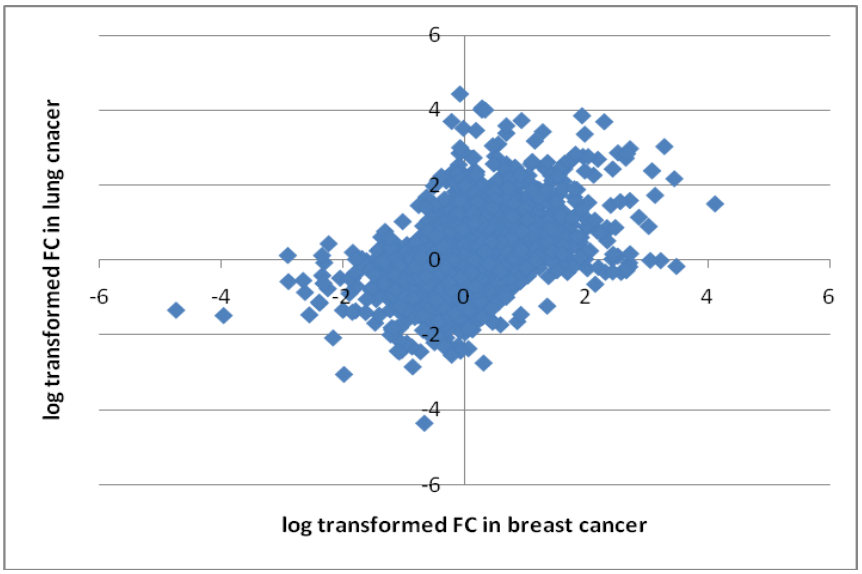


(c) (d)

Figure 2.7: Classification performance by top k-gene groups of stomach cancer. (a) classification accuracies by the top 100 *k*-gene combinations between stomach cancer and reference samples in the training set. (b) classification accuracies by the top 100 *k*-gene combinations on the test set. (c) classification accuracies by top 100 *k*-gene combinations between early stomach cancer and corresponding reference samples in the training set and (d) on the test set.



(a)



(b)

Figure 2.8: Comparison of the gene expression fold changes (a) between breast cancer training and testing datasets (b) between breast cancer and lung cancer

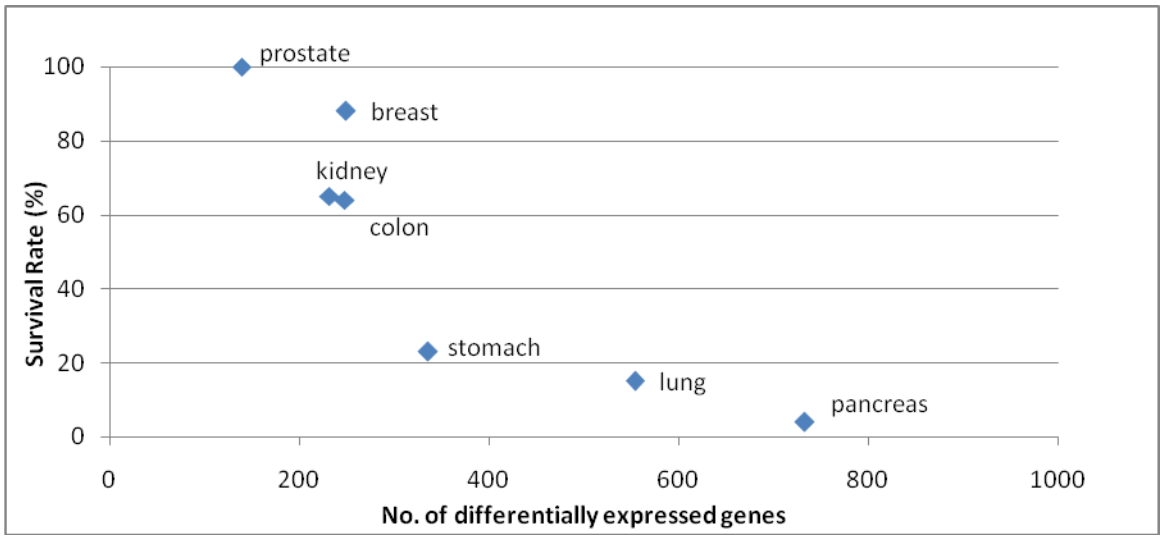


Figure 2.9: Correlation between 5-year survival rate and the number of differentially genes in each cancer type.

Tables

Table 2.1: Statistics of 5-year relative survival rates by race and year of diagnosis, US. 1974-2001 (all numbers are in percentage)

Cancer Site	Relative 5-Year Survival Rate (%)									No. of differential genes
	White			African American			All Race			
	1974-76	1983-85	1995-2001	1974-76	1983-85	1995-2001	1974-76	1983-85	1995-2001	
Prostate	68	76	100	58	64	97	67	75	100	139
Breast	75	79	90	63	64	76	75	78	88	249
Kidney	52	56	65	49	55	64	52	56	65	232
Colon	51	58	65	46	49	55	50	58	64	248
Stomach	15	16	21	16	19	23	15	17	23	336
Lung & bronchus	13	14	16	11	11	13	12	14	15	554
Pancreas	3	3	4	3	5	4	3	3	4	733

Table 2.2: List of genes that are differentially expressed in more than 4 cancer types and their relevance to different cancer types. "↑" indicates up-regulated gene expression in the corresponding cancer type while "↓"is down-regulation. "*****" indicates that a gene has been reported as relevant to the corresponding cancer type. "B." for breast cancer; "C." for colon cancer; "K." for kidney cancer"; "L." for lung cancer"; "Pa." for pancreatic cancer"; "Pr." for prostate cancer" and "S." for stomach cancer".

	Direction of regulation							**Reported to be related to cancers**							
Gene ID	**Breast**	**Colon**	**Kidney**	**Lung**	**Pancreas**	**Prostate**	**Stomach**	**B.**	**C.**	**K.**	**L.**	**Pa.**	**Pr.**	**S.**	**Other cancer types**
CDC2	↑	↑		↑	↑	↑	↑	*	*		*		*	*	liver cancer; squamous cell carcinoma;nasopharyngeal carcinoma
AURKA	↑	↑		↑		↑	↑	*	*		*	*	*	*	ovarian cancer;esophageal squamous cancer;uterine cancer;bladder cancer
ABCA8	↓	↓	↓	↓			↓								
DPT	↓	↓		↓		↓	↓								
TOP2A	↑	↑		↑	↑		↑	*	*					*	bladder cancer;ovarian cancer; squamous cell carcinoma
MMP7		↑		↑	↑		↑	*	*		*	*		*	ovarian cancer; oral cancer; rectal cancers;

															bladder cancer; liver cancer
MAD2L1		↑		↑	↑		↑		*					*	thyroid carcinomas; oesophageal squamous cancer
KLF4	↓	↓		↓			↓	*	*					*	esophageal cancer;bladder cancer
MELK	↑			↑	↑		↑	*							brain cancer;endometrial cancer
C7		↓	↓	↓		↓		*					*		uterine cervical cancers
ECT2		↑		↑	↑		↑					*			
PRC1	↑			↑	↑		↑	*							
RRM2	↑			↑	↑		↑				*	*	*		
ALDH1A1	↓	↓		↓			↓					*			non-small cell bronchopulmonary cancer; liver cancer;T-cell leukemia
PMAIP1	↑	↑		↑	↑				*		*	*			
FABP4	↓	↓		↓			↓	*							Bladder cancer;
LCN2		↑	↑	↑	↑			*	*			*			ovarian cancer; leukemia
COL11A1	↑	↑		↑	↑										adenomas;
TTK		↑		↑		↑	↑								
CENPF	↑			↑	↑		↑	*							

Table 2.3: The list of genes that differentially expressed in more than 3 cancer type. "↑" indicates that a gene is up-regulated in the corresponding cancer type while "↓" indicates that a gene is down-regulated

Gene ID	Direction of regulation							Function
	Bre.	Con.	Kid.	Lun.	Pan.	Pro.	Sto.	
ABCA8	↓	↓	↓	↓			↓	ATP-dependent lipophilic drug transporter
ACADL	↓			↓	↓			NA
ADH1B	↓		↓	↓				NA
ADH1C	↓	↓					↓	NA
AGR2	↑			↑	↑			NA
ALDH1A1	↓	↓		↓			↓	Binds free retinal and cellular retinol-binding protein-bound retinal. Can convert/oxidize retinaldehyde to retinoic acid
ANLN		↑			↑		↑	Required for cytokinesis. Essential for the structural integrity of the cleavage furrow and for completion of cleavage furrow ingression.
AOC3	↓	↓		↓				Cell adhesion protein that participates in lymphocyte recirculation by mediating the binding of lymphocytes to peripheral lymph node vascular endothelial cells in an L-selectin- independent fashion. Has a monoamine oxidase activity.
ASPM	↑			↑	↑			Probable role in mitotic spindle regulation and coordination of mitotic processes. May have a preferential role in regulating neurogenesis.
AURKA	↑	↑		↑		↑	↑	May play a role in cell cycle regulation during anaphase and/or telophase, in relation to the function of the centrosome/spindle pole region during chromosome segregation. May be involved in microtubule formation and/or stabilization. May play a key role during tumor development and progression. Phosphorylates ARHGEF2 and BORA.
BUB1		↑		↑			↑	Involved in cell cycle checkpoint enforcement. Can interact and phosphorylate BUB3.
C7		↓	↓	↓		↓		C7 is a constituent of the membrane attack complex. C7 binds to C5b forming the C5b-7 complex, where it serves as a membrane anchor.
CAV1	↓			↓		↓		May act as a scaffolding protein within caveolar membranes. Interacts directly with G-protein alpha subunits and can functionally regulate their activity
CCL18			↑		↑		↑	Chemotactic factor that attracts lymphocytes but not

								monocytes or granulocytes. May be involved in B-cell migration into B-cell follicles in lymph nodes. Attracts naive T-lymphocytes toward dendritic cells and activated macrophages in lymph nodes, has chemotactic activity for naive T-cells, CD4+ and CD8+ T-cells and thus may play a role in both humoral and cell-mediated immunity responses.
CD36	↓			↓			↓	Seems to have numerous potential physiological functions. Binds to collagen, thrombospondin, anionic phospholipids and oxidized LDL. May function as a cell adhesion molecule. Directly mediates cytoadherence of Plasmodium falciparum parasitized erythrocytes. Binds long chain fatty acids and may function in the transport and/or as a regulator of fatty acid transport.
CDC2	↑	↑		↑	↑	↑	↑	Plays a key role in the control of the eukaryotic cell cycle. It is required in higher cells for entry into S-phase and mitosis. p34 is a component of the kinase complex that phosphorylates the repetitive C-terminus of RNA polymerase II.
CDH3		↑		↑	↑			Cadherins are calcium dependent cell adhesion proteins. They preferentially interact with themselves in a homophilic manner in connecting cells; cadherins may thus contribute to the sorting of heterogeneous cell types.
CENPF	↑			↑	↑		↑	Probably required for kinetochore function, involved in chromosome segregation during mitosis. Interacts with retinoblastoma protein (RB), CENP-E and BUBR1.
CHRDL1	↓			↓		↓		Antagonizes the function of BMP4 by binding to it and preventing its interaction with receptors. Alters the fate commitment of neural stem cells from gliogenesis to neurogenesis. Contributes to neuronal differentiation of neural stem cells in the brain by preventing the adoption of a glial fate. May play a crucial role in dorsoventral axis formation. May play a role in embyonic bone formation (By similarity). May also play an important role in regulating retinal angiogenesis trough modulation of BMP4 actions in endothelial cells.
CKS2	↑				↑		↑	Binds to the catalytic subunit of the cyclin dependent kinases and is essential for their biological function.
CLDN4				↑	↑		↑	Plays a major role in tight junction-specific obliteration of the intercellular space.
CLEC3B	↓			↓			↓	Tetranectin binds to plasminogen and to isolated kringle 4. May be involved in the packaging of molecules destined for exocytosis.
CNN1		↓		↓		↓		Thin filament-associated protein that is implicated in the regulation and modulation of smooth muscle contraction. It is capable of binding to actin, calmodulin, troponin C and tropomyosin. The interaction of calponin with actin inhibits the actomyosin Mg-ATPase activity (By similarity).
COL10A1	↑			↑	↑			Type X collagen is a product of hyperthrophic chondrotocytes and has been localized to presumptive

								mineralization zones of hyaline cartilage.
COL11A1	↑	↑		↑	↑			May play an important role in fibrillogenesis by controlling lateral growth of collagen II fibrils.
COL1A1	↑			↑	↑			Type I collagen is a member of group I collagen (fibrillar forming collagen).
COX7A1	↓			↓		↓		This protein is one of the nuclear-coded polypeptide chains of cytochrome c oxidase, the terminal oxidase in mitochondrial electron transport.
CXCL12		↓		↓		↓		Chemoattractant active on T-lymphocytes, monocytes, but not neutrophils. SDF-1-beta(3-72) and SDF-1-alpha(3-67) show a reduced chemotactic activity. Binding to cell surface proteoglycans seems to inhibit formation of SDF-1-alpha(3-67) and thus to preserve activity on local sites.
DMD		↓			↓	↓		May play a role in anchoring the cytoskeleton to the plasma membrane.
DPT	↓	↓		↓		↓	↓	Seems to mediate adhesion by cell surface integrin binding. May serve as a communication link between the dermal fibroblast cell surface and its extracellular matrix environment. Enhances TGFB1 activity. Inhibits cell proliferation. Accelerates collagen fibril formation, and stabilizes collagen fibrils against low-temperature dissociation.
ECT2		↑		↑	↑		↑	Binds highly specifically to RhoA, RhoC and Rac proteins, but does not appear to catalyze guanine nucleotide exchange.
ESM1		↑			↑		↑	May have potent implications in lung endothelial cell-leukocyte interactions.
FABP4	↓	↓		↓			↓	Lipid transport protein in adipocytes. Binds both long chain fatty acids and retinoic acid. Delivers long-chain fatty acids and retinoic acid to their cognate receptors in the nucleus.
FAM107A		↓		↓			↓	When transfected into cell lines in which it is not expressed, suppresses cell growth. May play a role in tumor development.
GPX3	↓			↓			↓	Protects cells and enzymes from oxidative damage, by catalyzing the reduction of hydrogen peroxide, lipid peroxides and organic hydroperoxide, by glutathione.
GREM1				↑	↑		↑	Cytokine that may play an important role during carcinogenesis and metanephric kidney organogenesis, as a BMP antagonist required for early limb outgrowth and patterning in maintaining the FGF4-SHH feedback loop. Down-regulates the BMP4 signaling in a dose-dependent manner. Acts as inhibitor of monocyte chemotaxis.
HBB	↓	↓		↓				LVV-hemorphin-7 potentiates the activity of bradykinin, causing a decrease in blood pressure.
HPGD			↓	↓			↓	Inactivation of prostaglandins.
HS3ST1			↑	↑	↑			Rate limiting enzyme for synthesis of HSact. Performs the crucial step modification in the biosynthesis of anticoagulant heparan sulfate (HSact) that is to

								complete the structure of the antithrombin pentasaccharide binding site.
INHBA	↑				↑		↑	Inhibins and activins inhibit and activate, respectively, the secretion of follitropin by the pituitary gland. Inhibins/activins are involved in regulating a number of diverse functions such as hypothalamic and pituitary hormone secretion, gonadal hormone secretion, germ cell development and maturation, erythroid differentiation, insulin secretion, nerve cell survival, embryonic axial development or bone growth, depending on their subunit composition. Inhibins appear to oppose the functions of activins.
KLF4	↓	↓		↓			↓	Transcription factor which acts as both an activator and repressor. Binds the CACCC core sequence. Binds to multiple sites in the 5'-flanking region of its own gene and can activate its own transcription. Required for establishing the barrier function of the skin and for postnatal maturation and maintenance of the ocular surface. Involved in the differentiation of epithelial cells and may also function in skeletal and kidney development.
KRT8	↑			↑	↑			Together with KRT19, helps to link the contractile apparatus to dystrophin at the costameres of striated muscle.
LCN2		↑	↑	↑	↑			Transport of small lipophilic substances (Potential).
MAD2L1		↑		↑	↑		↑	Required for the execution of the mitotic checkpoint which monitors the process of kinetochore-spindle attachment and delays the onset of anaphase when this process is not complete. It inhibits the activity of the anaphase promoting complex by sequestering CDC20 until all chromosomes are aligned at the metaphase plate.
MCM4	↑			↑			↑	Involved in the control of DNA replication.
MDK	↑			↑	↑			Has heparin binding activity, and growth promoting activity. Involved in neointima formation after arterial injury, possibly by mediating leukocyte recruitment. Also involved in early fetal adrenal gland development (By similarity).
MELK	↑			↑	↑		↑	Phosphorylates ZNF622 and may contribute to its redirection to the nucleus. May be involved in the inhibition of spliceosome assembly during mitosis.
MME	↓		↓	↓				Thermolysin-like specificity, but is almost confined on acting on polypeptides of up to 30 amino acids. Biologically important in the destruction of opioid peptides such as Met- and Leu-enkephalins by cleavage of a Gly-Phe bond. Involved in the degradation of atrial natriuretic factor (ANF).
MMP1		↑		↑	↑			Cleaves collagens of types I, II, and III at one site in the helical domain. Also cleaves collagens of types VII and X. In case of HIV infection, interacts and cleaves the secreted viral Tat protein, leading to a decrease in

								neuronal Tat's mediated neurotoxicity.
MMP11				↑	↑		↑	May play an important role in the progression of epithelial malignancies.
MMP12				↑	↑		↑	May be involved in tissue injury and remodeling. Has significant elastolytic activity. Can accept large and small amino acids at the P1' site, but has a preference for leucine. Aromatic or hydrophobic residues are preferred at the P1 site, with small hydrophobic residues (preferably alanine) occupying P3.
MMP7		↑		↑	↑		↑	Degrades casein, gelatins of types I, III, IV, and V, and fibronectin. Activates procollagenase.
MMP9				↑	↑		↑	May play an essential role in local proteolysis of the extracellular matrix and in leukocyte migration. Could play a role in bone osteoclastic resorption. Cleaves KiSS1 at a Gly-|-Leu bond.
MT1M				↓	↓		↓	Metallothioneins have a high content of cysteine residues that bind various heavy metals; these proteins are transcriptionally regulated by both heavy metals and glucocorticoids.
MT1X		↓				↓	↓	Metallothioneins have a high content of cysteine residues that bind various heavy metals; these proteins are transcriptionally regulated by both heavy metals and glucocorticoids.
MXRA5				↑	↑		↑	NA
MYB	↑					↑	↑	Transcriptional activator; DNA-binding protein that specifically recognize the sequence 5'-YAAC[GT]G-3'. Plays an important role in the control of proliferation and differentiation of hematopoietic progenitor cells.
MYH11		↓		↓		↓		Muscle contraction.
MYL9	↓	↓		↓				Myosin regulatory subunit that plays an important role in regulation of both smooth muscle and nonmuscle cell contractile activity via its phosphorylation. Implicated in cytokinesis, receptor capping, and cell locomotion.
MYLK		↓		↓		↓		Calcium/calmodulin-dependent enzyme implicated in smooth muscle contraction via phosphorylation of myosin light chains (MLC). Implicated in the regulation of endothelial as well as vascular permeability. In the nervous system it has been shown to control the growth initiation of astrocytic processes in culture and to participate in transmitter release at synapses formed between cultured sympathetic ganglion cells. Critical participant in signaling sequences that result in fibroblast apoptosis.
NDC80	↑			↑	↑			Acts as a component of the essential kinetochore-associated NDC80 complex, which is required for chromosome segregation and spindle checkpoint activity. Required for kinetochore integrity and the organization of stable microtubule binding sites in the outer plate of the kinetochore.
NUSAP1				↑	↑		↑	Microtubule-associated protein with the capacity to bundle and stabilize microtubules (By similarity). May

								associate with chromosomes and promote the organization of mitotic spindle microtubules around them.
PAICS		↑		↑			↑	NA
PCK1	↓	↓	↓					NA
PDK4				↓	↓		↓	Inhibits the mitochondrial pyruvate dehydrogenase complex by phosphorylation of the E1 alpha subunit, thus contributing to the regulation of glucose metabolism.
PHLDA2	↑			↑	↑			May play a role in regulating placenta growth.
PLA2G7			↑			↑	↑	Modulates the action of platelet-activating factor (PAF) by hydrolyzing the sn-2 ester bond to yield the biologically inactive lyso-PAF. Has a specificity for substrates with a short residue at the sn-2 position. It is inactive against long-chain phospholipids.
PMAIP1	↑	↑		↑	↑			Promotes activation of caspases and apoptosis. Promotes mitochondrial membrane changes and efflux of apoptogenic proteins from the mitochondria. Contributes to p53-dependent apoptosis after radiation exposure. Promotes proteasomal degradation of MCL1. Competes with BAK1 for binding to MCL1 and can displace BAK1 from its binding site on MCL1 (By similarity). Competes with BIM/BCL2L11 for binding to MCL1 and can displace BIM/BCL2L11 from its binding site on MCL1.
PRC1	↑			↑	↑		↑	KIF4A translocates PRC1 to the plus ends of interdigitating spindle microtubules during the metaphase to anaphase transition, an essential step for the formation of an organized central spindle midzone and midbody and for successful cytokinesis. Required for KIF14 localization to the central spindle and midbody. Acts as a microtubule-binding and bundling protein both in vivo and vitro. May function as an in vivo cyclin- CDK substrate.
PTGER4		↓		↓	↓			Receptor for prostaglandin E2 (PGE2). The activity of this receptor is mediated by G(s) proteins that stimulate adenylate cyclase. Has a relaxing effect on smooth muscle. May play an important role in regulating renal hemodynamics, intestinal epithelial transport, adrenal aldosterone secretion, and uterine function.
PTRF	↓			↓		↓		Termination of transcription by RNA polymerase I involves pausing of transcription by TTF1, and the dissociation of the transcription complex, releasing pre-rRNA and RNA polymerase I from the template. PTRF is required for dissociation of the ternary transcription complex (By similarity).
PTTG1		↑			↑		↑	Regulatory protein, which plays a central role in chromosome stability, in the p53/TP53 pathway, and DNA repair. Probably acts by blocking the action of key proteins. During the mitosis, it blocks Separase/ESPL1 function, preventing the proteolysis of the cohesin

								complex and the subsequent segregation of the chromosomes. At the onset of anaphase, it is ubiquitinated, conducting to its destruction and to the liberation of ESPL1. Its function is however not limited to a blocking activity, since it is required to activate ESPL1. Negatively regulates the transcriptional activity and related apoptosis activity of TP53. The negative regulation of TP53 may explain the strong transforming capability of the protein when it is overexpressed. May also play a role in DNA repair via its interaction with Ku, possibly by connecting DNA damage-response pathways with sister chromatid separation.
RNASE1		↓			↓		↓	Endonuclease that catalyzes the cleavage of RNA on the 3' side of pyrimidine nucleotides. Acts on single stranded and double stranded RNA.
RRM2	↑			↑	↑		↑	Provides the precursors necessary for DNA synthesis. Catalyzes the biosynthesis of deoxyribonucleotides from the corresponding ribonucleotides. Inhibits Wnt signaling.
S100P	↑			↑	↑			NA
SDPR	↓	↓		↓				May play a role in targeting PRKCA to caveolae (By similarity).
SFN			↑	↑	↑			p53-regulated inhibitor of G2/M progression.
SLC16A3			↑	↑	↑			Proton-linked monocarboxylate transporter. Catalyzes the rapid transport across the plasma membrane of many monocarboxylates such as lactate, pyruvate, branched-chain oxo acids derived from leucine, valine and isoleucine, and the ketone bodies acetoacetate, beta-hydroxybutyrate and acetate (By similarity).
SOX4				↑	↑		↑	Transcriptional activator that binds with high affinity to the T-cell enhancer motif 5'-AACAAAG-3' motif.
SPARCL1		↓		↓		↓		NA
SULF1				↑	↑		↑	Exhibits arylsulfatase activity and highly specific endoglucosamine-6-sulfatase activity. It can remove sulfate from the C-6 position of glucosamine within specific subregions of intact heparin. Diminishes HSPG (heparan sulfate proteoglycans) sulfation, inhibits signaling by heparin-dependent growth factors, diminishes proliferation, and facilitates apoptosis in response to exogenous stimulation.
TFAP2A				↑	↑		↑	Sequence-specific DNA-binding protein that interacts with inducible viral and cellular enhancer elements to regulate transcription of selected genes. AP-2 factors bind to the consensus sequence 5'-GCCNNNGGC-3' and activate genes involved in a large spectrum of important biological functions including proper eye, face, body wall, limb and neural tube development. They also suppress a number of genes including MCAM/MUC18, C/EBP alpha and MYC. AP-2 alpha is the only AP-2 protein required for early morphogenesis of the lens vesicle (By similarity).

TGFBI			↑		↑		↑	Binds to type I, II, and IV collagens. This adhesion protein may play an important role in cell-collagen interactions. In cartilage, may be involved in endochondral bone formation.
TGFBR3	↓	↓		↓				Binds to TGF-beta. Could be involved in capturing and retaining TGF-beta for presentation to the signaling receptors.
THBS2			↑	↑	↑			Adhesive glycoprotein that mediates cell-to-cell and cell-to-matrix interactions. Can bind to fibrinogen, fibronectin, laminin and type V collagen.
TOP2A	↑	↑		↑	↑		↑	Control of topological states of DNA by transient breakage and subsequent rejoining of DNA strands. Topoisomerase II makes double-strand breaks.
TOX3	↑			↑	↑			NA
TPX2				↑		↑	↑	NA
TRIM29		↑			↑		↑	It is able to complement the radiosensitivity defect of an ataxia telangiectasia (AT) fibroblast cell line.
TTK		↑		↑		↑	↑	Phosphorylates proteins on serine, threonine, and tyrosine. Probably associated with cell proliferation.
TWIST1					↑	↑	↑	Probable transcription factor, which seems to be involved in the negative regulation of cellular determination and in the differentiation of several lineages including myogenesis, osteogenesis, and neurogenesis. Inhibits myogenesis by sequestrating E proteins, inhibiting trans-activation by MEF2, and inhibiting DNA-binding by MYOD1 through physical interaction. This interaction probably involves the basic domains of both proteins (By similarity). Also represses expression of proinflammatory cytokines such as TNFA and IL1B.
UBE2C		↑		↑			↑	Catalyzes the covalent attachment of ubiquitin to other proteins. Required for the destruction of mitotic cyclins.

Table 2.4: Enriched pathways by differentially expressed genes in different cancer types (enrichment P-value cutoff = 0.05)

Pathways	COUNT	BREAST	COLON	KIDNEY	LUNG	PANCREASE	PROSTATE	STOMACH
Complement and coagulation cascades	4		X	X	X	X		
ECM-receptor interaction	4	X			X	X		X
Focal adhesion	4	X			X	X	X	
Cell Communication	4	X			X	X	X	
Cell adhesion molecules (CAMs)	3				X	X		X
PPAR signaling pathway	3	X	X	X				
Glycine, serine and threonine metabolism	3		X	X		X		
p53 signaling pathway	2				X	X		
Cell cycle	2				X			X
Glycolysis / Gluconeogenesis	2			X		X		
Platelet Amyloid Precursor Protein Pathway	2			X		X		
PKC-catalyzed phosphorylation of inhibitory phosphoprotein of myosin phosphatase	2		X				X	
Fibrinolysis Pathway	2			X		X		
Eicosanoid Metabolism	2		X			X		
RBphosphoE2F	2	X			X			
Adipocytokine signaling pathway	1	X						
Small cell lung cancer	1					X		
Hematopoietic cell lineage	1					X		
TGF-beta signaling pathway	1				X			
Cytokine-cytokine receptor interaction	1					X		
Calcium signaling pathway	1						X	
ABC transporters – General	1		X					
Metabolism of xenobiotics by cytochrome P450	1							X
Nitrogen metabolism	1							X
3-Chloroacrylic acid degradation	1	X						
Propanoate metabolism	1	X						
Pyruvate metabolism	1	X						
Linoleic acid metabolism	1							X
Arachidonic acid metabolism	1							X
Glycerophospholipid metabolism	1	X						
Glycerolipid metabolism	1	X						
O-Glycan biosynthesis	1					X		
Tryptophan metabolism	1	X						

Tyrosine metabolism	1			X				
Histidine metabolism	1	X						
Arginine and proline metabolism	1			X				
Urea cycle and metabolism of amino groups	1	X						
Bile acid biosynthesis	1	X						
Visceral Fat Deposits and the Metabolic Syndrome	1	X						
Vitamin C in the Brain	1					X		
Inhibition of Matrix Metalloproteinases	1				X			
IGF-1 Receptor and Longevity	1	X						
Low-density lipoprotein (LDL) pathway during atherogenesis	1				X			
Classical Complement Pathway	1			X				
Pertussis toxin-insensitive CCR5 Signaling in Macrophage	1		X					
Chemokine_families	1		X					

Table 2.5: The top 2-gene markers for multiple cancer types. Each numerical value shows the classification accuracy between a cancer and its corresponding reference. Each entry represents the classification accuracy between a cancer set and its corresponding reference set on the training (train) and the testing (test) datasets, respectively.

Count	Markers	Breast		Colon		Kidney		Lung		Pancreas		Prostate		Stomach	
		train	test	train	test	train	test	train	test	train	test	train	test	train	test
5	CDC2+DPT	70.7 %	94.8 %	91.7 %	97.9%	_	_	88.9 %	92.6 %	_	_	60.2 %	81.6 %	66.7 %	85.5 %
	CDC2+TOP2A	72.4 %	94.8 %	75.0 %	100.0 %	_	_	85.2 %	85.2 %	71.2 %	71.2 %	_	_	78.3 %	85.5 %
4	CDC2+ECT2	_	_	85.4 %	97.9%	_	_	83.3 %	77.8 %	78.8 %	86.5 %	_	_	75.4 %	78.3 %
	ABCA8+AURKA	81.0 %	96.6 %	91.7 %	100.0 %	_	_	94.4 %	94.4 %	_	_	_	_	75.4 %	92.8 %
	ABCA8+FABP4	79.3 %	96.6 %	89.6 %	97.9%	_	_	96.3 %	98.1 %	_	_	_	_	79.7 %	84.1 %
	DPT+FABP4	79.3 %	87.9 %	95.8 %	89.6%	_	_	94.4 %	96.3 %	_	_	_	_	82.6 %	75.4 %
	FABP4+TOP2A	77.6 %	94.8 %	85.4 %	100.0 %	_	_	96.3 %	94.4 %	_	_	_	_	78.3 %	85.5 %
3	CDC2+SULF1	_	_	_	_	_	_	90.7 %	88.9 %	96.2 %	90.4 %	_	_	95.7 %	88.4 %

Table 2.6: The top 3-gene discriminators for multiple cancer type types

Count	Markers	Breast		Colon		Kidney		Lung		Pancreas		Prostate		Stomach	
		train	test	train	test	train	test	train	test	train	test	train	test	train	test
4	ABCA8+ALDH1A1+DPT	75.9%	96.6%	93.8%	97.9%	_	_	92.6%	96.3%	_	_	_	_	75.4%	91.3%
4	ABCA8+AURKA+DPT	77.6%	96.6%	95.8%	100.0%	_	_	92.6%	94.4%	_	_	_	_	76.8%	91.3%
4	ALDH1A1+FABP4+TOP2A	79.3%	93.1%	87.5%	85.4%	_	_	96.3%	77.8%	_	_	_	_	76.8%	76.8%

Table 2.7: The top 4-gene discriminators for multiple cancer types

Count	Markers	Breast		Colon		Kidney		Lung		Pancreas		Prostate		Stomach	
		train	test	train	test	train	test	train	test	train	test	train	test	train	test
4	ABCA8+CDC2+KLF4+TOP2A	79.3%	94.8%	93.8%	100.0%	_	_	94.4%	98.1%	_	_	_	_	88.4%	94.2%
4	ABCA8+FABP4+KLF4+TOP2A	79.3%	96.6%	89.6%	100.0%	_	_	96.3%	98.1%	_	_	_	_	87.0%	89.9%
4	ALDH1A1+FABP4+KLF4+TOP2A	79.3%	94.8%	89.6%	97.9%	_	_	96.3%	96.3%	_	_	_	_	85.5%	91.3%
4	CDC2+COL11A1+PMAIP1+TOP2A	79.3%	100.0%	93.8%	100.0%	_	_	90.7%	90.7%	90.4%	86.5%	_	_	_	_
4	DPT+FABP4+KLF4+TOP2A	79.3%	94.8%	91.7%	100.0%	_	_	98.1%	96.3%	_	_	_	_	85.5%	91.3%

Table 2.8: Top *k*-gene discriminators with their proteins to be blood secretory. Each numerical value represents the classification accuracy between cancer tissues and their corresponding reference tissues.

		Breast		Colon		Kidney		Lung		Pancreas		Prostate		Stomach	
Count	Markers	train	test	train	test	train	test	train	test	train	test	train	test	train	test
4	DPT+KLF4	70.7%	84.5%	89.9%	97.9%	_	_	94.4%	94.4%	_	_	_	_	84.1%	91.3%
3	GREM1+MMP7	_	_	_	_	_	_	88.9%	79.6%	92.3%	73.1%	_	_	89.9%	75.4%
3	MMP7+MMP9	_	_	_	_	_	_	75.9%	79.6%	96.2%	78.9%	_	_	77.9%	76.8%
3	MMP11+MMP7+MMP9+RRM2	_	_	_	_	_	_	85.2%	96.3%	96.2%	88.5%	_	_	88.1%	88.4%
3	CCL18+TGFBI	_	_	_	_	74.5%	80.9%	_	_	82.7%	82.7%	_	_	71.0%	75.4%
3	LCN2+THBS2	_	_	_	_	74.5%	87.2%	88.9%	85.2%	96.2%	82.8%	_	_	_	_
3	DPT+MMP7	_	_	97.9%	89.6%	_	_	85.2%	88.9%	_	_	_	_	84.2%	81.2%
3	FAM107A+KLF4	_	_	87.5%	100.0%	_	_	94.4%	92.6%	_	_	_	_	91.3%	92.8%
3	FAM107A+KLF4+MMP7+PAICS	_	_	100.0%	100.0%	_	_	94.4%	94.4%	_	_	_	_	91.3%	91.3%
3	INHBA+RRM2	74.1%	100.0%	_	_	_	_	_	_	94.2%	88.5%	_	_	78.3%	81.2%
3	GPX3+RRM2	81.0%	96.6%	_	_	_	_	88.9%	94.4%	_	_	_	_	85.5%	81.2%
3	COL11A1+DPT	72.4%	96.6%	97.9%	89.6%	_	_	92.6%	94.4%	_	_	_	_	_	_
3	MMP11+RRM2	_	_	_	_	_	_	88.9%	90.7%	86.5%	88.5%	_	_	75.0%	82.6%

Table 2.9: A summary of the top three *k*-gene discriminators for each of the seven cancer types along with discriminators for early stage breast and stomach cancer. "C" for cancer and "R" for reference tissues; and "---" indicates that the corresponding *k*-gene combinations were not assessed since (*k-1*)-gene combinations already give 100% classification accuracy.

Cancer type	No. of samples (C/R) in training and testing set	Top three k-gene discriminators			
		1-gene	2-gene	3-gene	4-gene
Breast cancer	43/43 31/27	PCOLCE2 ANGPTL4 LEP	ADIPOQ+TMEM97 PPARG+TMEM97 TACSTD2+CAV1	RRM2+COL1A1+PCOLCE2 RRM2+COL1A1+PPARG RRM2 + STBD1 + MAOA	RRM2+COL1A1 + GPR109B+IGJ RRM2+COL1A1 + GPR109B+IGJ RRM2+COL1A1+GPR109B + SPINT2
Breast cancer Early stage	31/31 12/12	GPR109B PCOLCE2 ADIPOQ	PCSK5+COL10A1 FERMT2+SPINT2 MAOA+IGJ	COL1A1+PCSK5+TF GPX3+COL1A1+SPINT2 STBD1+TMEM97+COL10A1	RRM2+COL1A1+GPR109B+IGJ RRM2+ COL1A1+ GPR109B+SPINT2 COL1A1+MAOA+SPINT2+COL11A1
Colon cancer	53/28 24/24	MMP7 DPT MMP1	MMP7+FAM107A FRZB+MMP7 SLIT3+MMP7	---	---
Kidney cancer	49/23 35/12	CCL18 ACPP UMOD	EGF+ALB ACPP+UMOD UMOD+ALB	---	---
Lung cancer	58/49 27/27	CAV1 SFTPC TNXB	FERMT2+GREM1 TEK+NFASC CAV1+MMP1	---	---
Pancreatic cancer	39/39 36/16	KRT17 COL10A1 FAM19A5	MMP7+AZGP1 MMP7+ELA3B MMP7+FGL1	COL8A2+SGPP2+CCL18 COL8A2+PMEPA1+TMEM45B LCN2 + COL8A2 + PMEPA1	---
Prostate cancer	65/63 62/47	CRISP3 MYLK PALLD	LTF+IGF1 LTF+SPARCL1 SMTN+CCK	SMTN+CCK+CCL2 SMTN+CCK+COMP SMTN+CCK+PLA2G7	---

Stomach cancer	89/23 38/31	SERPINH1 BGN COL12A1	CHGA+SERPINH1 PGC+SERPINH1 TGFBI+CHGA	---	---
Stomach cancer Early stage	31/23 12/10	SERPINH1 BGN COL12A1	---	---	---

Table 2.10: A summary of the training and the testing set used in our analysis

Cancer	GEO dataset ID training / testing data	# reference/ #cancer samples
breast cancer	GSE15852 / GSE10810	43/43 (27/31)
colon cancer	GSE6988 / GSE10950	28/53 (24/24)
kidney cancer	GSE15641 / GSE4866	23/49 (12/35)
lung cancer	GSE10072 / GSE7670	49/58 (27/27)
pancreatic cancer	GSE15471 / GSE16515	39/39 (16/36)
prostate cancer	GSE6606 / GSE3933	63/65 (47/62)
stomach cancer	GSE2701 / GSE13911	23/89 (31/38)

CHAPTER 3

A COMPARATIVE STUDY OF GENE-EXPRESSION DATA OF BASAL CELL CARCINOMA AND MELANOMA REVEALS NEW INSIGHTS ABOUT THE TWO CANCERS[3]

[3]Xu K and Mao X. et al. 2012. PLoS ONE 7(1): e30750. doi:10.1371/journal.pone.0030750

Abstract

A comparative analysis of genome-scale transcriptomic data of two types of skin cancers, melanoma and basal cell carcinoma in comparison with other cancer types, was conducted with the aim of identifying key regulatory factors that either cause or contribute to the aggressiveness of melanoma, while basal cell carcinoma generally remains a mild disease. Multiple cancer-related pathways such as cell proliferation, apoptosis, angiogenesis, cell invasion and metastasis, are considered, but our focus is on energy metabolism, cell invasion and metastasis pathways. Our findings include the following. (a) Both types of skin cancers use both glycolysis and increased oxidative phosphorylation (electron transfer chain) for their energy supply. (b) Advanced melanoma shows substantial up-regulation of key genes involved in fatty acid metabolism (β-oxidation) and oxidative phosphorylation, with aerobic metabolism being far more efficient than anaerobic glycolysis, providing a source of the energetics necessary to support the rapid growth of this cancer. (c) While advanced melanoma is similar to pancreatic cancer in terms of the activity level of genes involved in promoting cell invasion and metastasis, the main metastatic form of basal cell carcinoma is substantially reduced in this activity, partially explaining why this cancer type has been considered as far less aggressive. Our method of using comparative analyses of transcriptomic data of multiple cancer types focused on specific pathways provides a novel and highly effective approach to cancer studies in general.

Introduction

The improvement of the cancer treatment and therapy heavily relies on the understanding of the fundamental tumor biology. The rapidly increasing pool (Sherlock, Hernandez-Boussard et al. 2001; Barrett, Troup et al. 2007) of large-scale transcriptomic data for various cancer types has provided unprecedented opportunities for computational cancer biologists to study common characteristics across multiple cancer types as well as distinct properties of individual cancer types, which could provide novel insights about different cancer phenotypes at the molecular level. To target the unique feature of the most aggressive cancer types melanoma, here we present a comparative analysis of gene expression data collected on cancer and control tissue samples of two skin cancer types, melanoma and basal cell carcinoma, which have very distinct characteristics.

Skin cancer is one of the most common cancer types in the USA. Currently over 3.5 million cases of skin cancers are diagnosed and reported annually (Rogers, Weinstock et al. 2010). It has been estimated that three out of ten Caucasians will develop skin cancer during their lifetime (Polsky and Wang 2011). The most common skin cancer is basal cell carcinoma (BCC), which develops in the basal cell layer of the skin, and primarily occurs in fair-skinned individuals. Sunlight is known to be a major factor for causing the disease. BCC is rarely deadly since it generally does not metastasize (Jemal, Siegel et al. 2010). In contrast, melanoma is a rare type of skin cancer but is among the deadliest forms of cancers (Jerant, Johnson et al. 2000). The tumor is derived from melanocytes, cells that produce the dark pigment. While melanoma is not limited to skin, it generally starts from the skin. A number of genes or their mutations have been found to be associated with the

development of melanoma such as MC1R (Box, Duffy et al. 2001), CDK4 (Zuo, Weger et al. 1996) and CDKN2A (Hughes-Davies 1998). The early stage of the disease is referred to as the *radial growth phase* when the tumor grows mostly horizontally. The behavior of the tumor drastically changes as soon as it starts to grow vertically, i.e., entering the *vertical growth phase*. It generally starts invading neighboring tissues when its thickness goes beyond 1mm (Balch, Buzaid et al. 2001).

While some information is known about the potential causes of the two skin cancer types, such as excessive exposure to sunlight and development of the basal-cell nevus syndrome being the main causes of basal cell carcinoma and a few rare mutations in the aforementioned genes being the main reason for the development of melanoma, a detailed understanding about why the two skin cancer types behave so differently remains to be very limited.

Through comparative analyses of genome-scale transcriptomic data on the two cancer types, we have gained a number of new insights which could shed new lights on our efforts to understand the detailed mechanisms of these two rather different skin cancer types. To put our analysis in a larger context, seven other cancer types have also been included, which range from relatively slow growing cancer to the fastest growing cancers, i.e, prostate, breast, kidney, colon, stomach, lung and pancreatic. By using transcriptomic data collected on cancer *versus* control tissues and comparing expression changes of the genes involved in different pathways associated with energy metabolism, we found that: (i) multiple genes involved in oxidative phosphorylation are up-regulated

in both melanoma and BCC, which is unique to only skin cancers among the nine types of cancer we examined and is inconsistent with Warburg's thesis (Warburg 1956); (ii) interestingly, the key enzyme in ATP generation in the oxidative phosphorylation pathway is up-regulated only in advanced melanoma but not in any form of BCC; (iii) the level and scale of up-regulated genes involved in cell invasion and metastasis in advanced melanoma are comparable to those of pancreatic cancer, while the corresponding values in BCC are essentially at the lower end among all the nine cancer types we examined. We believe that our comparative transcriptomic data analyses of multiple cancer types focused on specific cancer related pathways provide a novel and highly effective approach to cancer studies, which could lead to substantial new insights about cancer formation (when the relevant data are available) and progression.

Results

Our analysis was done on two gene-expression datasets. One set consisted of 52 tissue samples for the study of melanoma. Of the 52 tissue samples, 18 were common nevi (moles) (CMN), 11 were dysplastic nevi (pre-cancerous) (DN), 8 in the radial growth phase (RGP) (early stage) and 15 in the vertical growth phase (VGP) (advanced stage) (Scatolini, Grand et al. 2010). For this particular dataset, the common nevi tissues were used as the control set since the original study did not include normal skin tissues (Scatolini, Grand et al. 2010). The second set consisted of 31 tissues for the study of BCC. Of the 31 tissue samples, 8 were in the superficial form (early stage), 7 in the morphea form (intermediate stage) and 8 in the nodular form (advanced stage), along with 8 normal skin epithelial tissues as the control (Lo, Yu et al. 2010).

1. Differentially expressed genes in the two skin cancer types

In this study, a gene is considered *differentially* expressed at a specific stage of a cancer if the distribution of its expression levels among cancer tissues at that stage is deemed to be statistically different[4] from the distribution of its expression levels among the control tissues (see Material and Methods). For BCC, 158, 406 and 494 genes were found to be differentially expressed in superficial, morphea and nodular forms, respectively, compared to the controls, which is consistent with our previous observation that the number of differentially expressed genes increases as a cancer advances (Xu, Cui et al. 2010; Cui, Chen et al. 2011). Using the same cutoff, 123, 326 and 1,647 genes were deemed to be differentially expressed in DN, RGP and VGP melanoma. In our previous study, we found that there is a strong correlation between the number of differentially expressed genes and the five-year survival rate associated with a particular cancer (Xu, Cui et al. 2010). Thus, the high number of differentially expressed genes in VGP melanoma is consistent with the clinical statistics regarding the mortality rate of this cancer. There is a possibility that this number could be potentially under-estimated since the controls (moles) for the melanoma analysis are not normal skin tissues and moles are probably the first step moving towards melanoma. The detailed lists of differentially expressed genes are given in Appendix Table A3.1. Overall it was found that the numbers of differentially expressed genes in the two cancer types are comparable in their early stages; and a substantial rise in the number of differentially expressed genes in the advanced stage, VGP, of melanoma was observed.

[4] Among all the genes deemed to be differentially expressed in our study, their expression values are either consistently up-regulated or consistently down-regulated.

A careful analysis of the pathways in the KEGG database that are enriched by the differentially expressed genes among the cancer tissues of BCC and melanoma was conducted with the DAVID program (see Material and Methods). For BCC, no pathways were found to be clearly enriched in the early stage, while 19 and 18 pathways were enriched in the intermediate and advanced stages, respectively. For melanoma, 1, 8 and 61 pathways were enriched in the DN, RGP and VGP forms, respectively. The names of these enriched pathways are given in Table 3.1.

It was noted that 11 enriched pathways are specific to VGP melanoma, the deadliest form among all the skin cancer types, including pathways associated with amino sugar and nucleotide sugar metabolism, linoleic acid metabolism and the citratric acid cycle (TCA cycle). In addition, some pathways are significantly enriched in only melanoma among the two skin cancer types considered, such as fatty acid metabolism, cell cycle, apoptosis and the ErbB signaling pathway. For BCC, it was noted that its advanced stage cancer has three pathways uniquely enriched among all the cancer types under consideration, including the spliceosome, GnRH signaling and long-term potentiation pathways.

In addition, we have checked if some of the annotated proto-oncogene and tumor-suppressor genes (http://www.uniprot.org/keywords/) show differential expressions in BCC and melanoma. Overall it was found that 2, 5 and 4 oncogenes are over-expressed in early, intermediate and advanced stage of BCC samples, respectively, and 0, 1 and 0 tumor suppressor genes are respectively under-expressed in the three stages. Similarly, for melanoma, 9, 1 and 32 oncogenes are over-expressed in precancerous, early stage and

advanced melanoma, respectively, and 0, 4 and 4 tumor suppressor genes are respectively under-expressed in the three stages.

Some of these up-regulated oncogenes have been reported as key regulatory genes for some of the skin cancer types. Among the up-regulated oncogenes in melanoma, ABL2, NRAS, PDGFC and FGF1 have been reported to be melanoma-associated oncogenes (Polsky and Cordon-Cardo 2003). Thus, RAB6B, REL and WHSC1L1, as identified by our analysis, may represent additional oncogenes for melanoma. For BCC, HRAS, RRAS and RUNX1 have been reported as BCC-associated oncogenes (Iwasaki, Srivastava et al. 2010). Hence, ECT2, PLAG1, RAB6C and SSPN, which were identified by our analysis, may represent additional oncogenes for BCC. It is interesting that nine oncogenes exhibit up-regulation in the pre-cancer stage of melanoma, which may be the initial switch of the tumorgenesis leading to melanoma. Moreover, the large increase in the number of up-regulated oncogenes in VGP melanoma may suggest the aggressiveness of the cancer. A detailed list of all the identified oncogene and tumor-suppressor genes is given in Figure 3.1.

2. Differentially expressed gene involved in energy metabolism

We have examined expression fold-changes of genes involved in four energy metabolic pathways: glycolysis, fatty acid metabolism, the TCA cycle and oxidative phosphorylation (also called the *electron transfer chain*) in the two skin cancer types and compared them with the other seven non-skin cancer types. Figure 3.2 shows expression level changes of genes involved in the four energy pathways of the two skin cancer types

and the other seven cancer types. By examining the figure, the following observations can be made.

1. Substantial increases in expression levels of multiple genes involved in glycolysis are observed in both the advanced forms of the two skin cancer types and five of the seven reference cancer types, namely kidney, colon, stomach, lung and pancreatic, consistent with Warburg's thesis (Hanahan and Weinberg 2011).

2. Two enzyme-encoding genes involved in fatty acid metabolism show substantial up-regulation in both BCC and melanoma, in contrast to the seven non-skin cancer, indicating a unique way that skin cancer obtains energy not only from glucose metabolism like other cancers but also from fatty acid metabolism, which is a more efficient energy generation pathway.

3. Moderately increased expression levels of genes involved in the TCA cycle are also observed in advanced melanoma, along with breast, colon, stomach and lung cancers.

4. Most interesting is the finding that multiple genes involved in oxidative phosphorylation are up-regulated in the advanced form of both skin cancer types, which strongly suggests that both skin cancer types obtain much of their energy through oxidative phosphorylation, which produces an order of magnitude more ATPs than each of the other three energy pathways (per glucose). This is very surprising as this indicates

the two skin cancer types, even in their advanced forms, are not under hypoxic condition, and hence do not show the Warburg effect.

5. ATP synthase is up-regulated only in advanced melanoma, in addition to the protein complex responsible for electron transfer but not in any form of BCC, indicating that ATP synthesis is faster in advanced melanoma than BCC. Further analysis of the differentially expressed genes in advanced melanoma suggests that the increased levels of acetyl-CoA, NADH and $FADH_2$ derived from fatty acid oxidation can enter mitochondria and undergo oxidative phosphorylation. This process would be highly advantageous for a tumor as β-oxidation of fatty acids yields a larger number of acetyl-CoA molecules compared to glycolysis, and thus it has a larger number of substrates for the TCA cycle and subsequent oxidative phosphorylation. In order to utilize the increased number of acetyl-CoAs, the cells may need to have an increased rate of oxidative phosphorylation. Acetyl-CoA is an allosteric inhibitor of the PDH enzymes. As we observed, PDHA2 is substantially down-regulated in VGP melanoma, which can be attributed to the increased level of acetyl-CoA. As the PDHA2 activity decreases, pyruvate will naturally be diverted to lactic acid formation, which is shown by the up-regulation of LDHA and lactate transporters (MCT proteins SLC16As 3&6). Figure 3.3 gives an energy model for advanced melanoma based on the results of our data analysis.

3. Differentially expressed gene involved in tumor invasion and metastasis

We have studied expression changes of genes involved in the metastatic process, with the goal of identifying possible reasons why the two skin cancer types have

substantial differences in their ability to metastasize. For this, focus was placed on the pro-metastasis gene families, namely the positive regulation of the epithelial-mesenchymal transition (EMT), the negative regulation of cell adhesion, the chemokine and MMP families, both of which promote degradation of extracellular matrices. Figure 3.4 shows the observed expression changes of genes involved in these processes of the two skin cancer types, along with the other seven cancer types. We have made the following observations from the data in Figure 3.4.

1. The VGP melanoma has the greatest number of up-regulated genes involved in the pro-metastasis gene family, having even more such genes than pancreatic cancer; While simply counting the number of up-regulated genes may be a rather crude way to assess the ability of a cancer to metastasize, Figure 3.5 shows that there is a strong (negative) correlation between this number and the five-year survival rate of a cancer.

2. VGP melanoma is the only skin cancer type with up-regulated genes involved in positive regulation of the epithelial-mesenchymal transition, which is considered to be the crucial developmental and regulatory program for cell invasion and metastasis (Klymkowsky and Savagner 2009; Polyak and Weinberg 2009); in addition, the significant up-regulation of genes involved in degrading the cell-cell adhesion molecules and the increased negative regulation of the cell adhesion, the matrix metalloproteinases and chemokines all suggest that VGP is highly metastatic (Hanahan and Weinberg 2000).

3. The lymphatic spread represents a major way for metastasis. We noted that the number of differentially expressed genes in lymphocyte proliferation is much higher in VGP melanoma compared to all the other skin cancer subtypes under consideration (see Figure 3.6).

4. It has been reported that a chaotic circadian rhythm may lead to faster tumor growth (Fu and Lee 2003). From the above figure, we noted that key genes of the circadian rhythm pathway are differentially expressed in VGP melanoma but only to a very limited extent in other skin cancer subtypes under consideration. CRY2, the most important gene controlling cell circadian rhythm (van der Horst, Muijtjens et al. 1999), shows down-regulation only in VGP melanoma among all the cancer types. Items (2) – (4) above suggest that VGP melanoma has the greatest activities for cell-invasion and metastasis.

5. In contrast, only a few genes of BCC are up-regulated in the aforementioned processes. Among the two different BCC subtypes, it is the morphea form, not the nodular form, that has the most up-regulated genes, which is consistent with previous reports that this form represents the BCC form with the greatest number of metastasis cases (Bozikov and Taggart 2006). We observed that in the morphea BCC, a number of genes encoding the collagens and proteins involved in anti-metastasis are highly up-regulated, suggesting that the cancer is under the control of the immune system in inhibiting its invasion and metastasis throughout its development.

Overall our analysis suggests that VGP melanoma has the highest potential and unique ability to metastasize among the cancer types studied herein, while the BCC's ability to metastasize is weakest among the nine types of cancer under consideration.

3. Differentially expressed genes involved in other cancer-related processes

3a. *Differentially expressed genes involved in cell proliferation*: Self-sufficiency in growth signals is a major acquired capability of any cancer. Our analysis of differentially expressed genes involved in positive regulation of cell proliferation indicates that VGP melanoma has substantially more genes up-regulated in this category than all the other skin cancer forms. Specifically, we noted that the number of such up-regulated genes is comparable to that of pancreatic cancer with the detailed data given in Figure 3.7. Particularly worth noting is that VGP melanoma is the only cancer type among the nine cancer types showing substantial up-regulation of genes of the Jak-STAT signaling pathway, which is a crucial pathway that promotes cell growth.

3b. *Differentially expressed genes involved in apoptosis*: The morphea BCC, nodular BCC and VGP melanoma all have substantial numbers of differentially expressed genes involved in negative regulation of cell death as shown in Figure 3.8. From this figure we can see an increasing trend in the number of up-regulated genes as the cancer type becomes more aggressive. Specifically it was noted that this number for VGP is higher than all the other cancer types under consideration except for pancreatic cancer.

3c. *Differentially expressed genes involved in angiogenesis*: None of the skin cancer types show increased activities of angiogenesis, unlike the advanced form of the other

cancers as shown in Figure 3.9. This observation is consistent with our earlier finding that skin cancers are generally not under hypoxic stress and hence probably have no pressure to activate the angiogenesis pathway.

4. Signature genes for the two skin cancer types

We have predicted signature genes and gene groups for melanoma and BCC, respectively, where a *signature* (or *marker*) gene or gene group refers to genes whose expression pattern is unique to a specific cancer type. We then tested the performance of the identified signature genes for the two cancer types, respectively, on two datasets, independent of the training datasets used, i.e., GSE3189 for melanoma and GSE12542 for BCC. For melanoma, the top five 1-gene signatures, OAS3 (82.9% on training and 71.4% on testing set), RALBP1 (82.9% on training and 71.4% on testing set), GLA (80.4% on training and 71.4% on testing set), LLGL1 (80.4% on training and 71.4% on testing set) and SERPINA6 (80.4% on training and 71.4% on testing set), all have better than 80% classification accuracy between melanoma cases and control cases in the test set. Among the top markers, OAS3 and SERPINA6 are predicted to be blood secretory by our prediction program (Cui, Liu et al. 2008), hence suggesting the potential feasibility in identifying diagnostic markers for melanoma through blood tests. Similarly, four 1-gene signatures with classification accuracy better that 71% are predicted to be urine excretory, CCL18 (73.2% on training and 71.4% on testing set), HEXB (73.2% on training and 71.4% on testing set), IFI30 (73.2% on training and 71.4% on testing set) and STC1 (73.2% on training and 71.4% on testing), by using our prediction program (Hong, Cui et al. 2011), suggesting the potential feasibility in identifying diagnostic

markers for melanoma through urine tests. Among the top 2-gene signatures, three pairs reach classification accuracy better than 85% on both the training and the testing datasets. Three pairs, with training classification accuracies better than 90% and testing classification accuracy better than 70%, are predicted to be blood secretory and only 1 pair CTSK_RNASE6 (87.8% on training and 73.0% on testing sets), with training classification accuracy better than 87% and testing classification accuracy better than 70%, are predicted to be urine excretory.

For BCC, the top two 1-gene markers, CS (90.3% on training and 87% on testing set) and TACSTD1 (87.1% on training and 87% on testing dataset), both have at least 87% classification accuracy on both the training and testing datasets. Among the top markers, EGR1 (87.1% on training and 81.3% on testing set) is predicted to be blood secretory. Similarly, RAB3D (80.6% on training and 80% on testing dataset) is predicted to be urine excretory, with a classification accuracy better that 80% in both sets, providing potential diagnostic markers for melanoma through urine tests. The top two 2-gene signatures all have classification accuracies better than 90% on both the training and testing sets. Also, four pair signatures, with classification accuracy better than 80% on both the training and testing sets, are predicted to encode blood secretory proteins. The detailed list of all these marker genes is given in Appendix Table A3.2.

Materials and Method

1. Microarray gene expression data for human cancers

Microarray gene expression data for both skin cancer types were downloaded from the GEO database of NCBI (Edgar, Domrachev et al. 2002). The melanoma data is the dataset GSE12391 and the BCC data is the dataset GSE6520. The gene-expression data for the seven cancer types used in this study: breast, colon, kidney, lung, pancreatic, prostate and stomach, are also downloaded from the GEO database of NCBI. For each dataset used for each cancer type, we have made sure that the dataset was generated using the same platform by the same research group. For each of the classification problem solved in this study, we have used a training dataset and a separate testing set for each cancer type. The details of the data are given in Table 3.2.

Considering that different microarray datasets used in this study cover different gene sets, we have considered the genes that belong to all microarray datasets used in this study for all the comparative analyses throughout this paper, which consists of 4,401 genes. The detailed list of these genes is given in Appendix Table A3.3. When mapping genes across different datasets, we rely on the NCBI gene IDs of the genes, i.e., two genes in different datasets are considered as the same genes if their IDs are identical.

The cancer gene list is downloaded from the Cancer Gene Census website, which contains 457 confirmed cancer genes (http://www.sanger.ac.uk/genetics/CGP/Census/).

2. Identification of differentially expressed genes

For each dataset used in this study, we have used the normalized expression data from the original study. We fully understand that gene-expression data across different datasets may not necessarily directly comparable; so we have compared the fold-changes between the diseased and the control tissues for each cancer type with fold-changes of expression data of another cancer type. Supplementary Figure 3.10 shows that the global fold-changes between two cancer types are generally comparable across the nine cancer types under consideration. It should be noted that when calculating the fold-change of an individual gene for a specific cancer type, we did not use the information of paired diseased-control tissues, instead we estimated the fold change based on the distributions of gene-expressions of the gene across all the cancer tissues versus control tissues, for each cancer type. We did so because some of the datasets have paired information wile other datasets do not.

For each dataset, the Mann-Whitney test was applied to identify genes that are differentially expressed in cancer *versus* control samples as follows: Given the null hypothesis H_0 that a gene is not differentially expressed between the cancer *versus* the control groups, rejection of this hypothesis means that the gene is differentially expressed in cancer. We consider a gene as *up-regulated* if the statistical significance, *p*-value, is less than 0.01 and its fold-increase is at least 1.5. A *down-regulated* gene is defined similarly.

For the non-skin cancer data, we only consider those genes with consistent up/down-regulation in both the training and the testing data sets as differentially expressed genes.

3. Pathway enrichment analysis of differentially expressed genes

Functional analysis and pathway enrichment analysis were conducted using DAVID (Dennis, Sherman et al. 2003), where the pathway information is based on the annotation from KEGG *(http://www.genome.ad.jp/kegg/)*. A *p*-value < 0.05 was used as the threshold to determine if a pathway is enriched or not by the identified differentially expressed genes. Note that the observations made throughout this paper are generally stable with respect to the p-value cutoff. Also note all the pathway enrichment analysis is based on the 4,401 genes that are shared by all the datasets used in this study.

4. Prediction of signature genes

To derive signature genes or gene groups, we conducted an exhaustive search for all the k-gene (k=1, 2) combinations among the differentially expressed genes, using a linear SVM-based classifier. We have used 5-fold cross validation to validate each identified signature. We refer the reader to (Cui, Li et al. 2011) for the detailed procedure used.

Concluding remarks

Our gene-expression data analysis revealed that both skin cancer types utilize oxidative phosphorylation as the key energy metabolism in addition to glycolysis. All evidence revealed by our study strongly indicates that the two skin cancer types are not

under hypoxic stress, hence explaining why the two cancer types do not show the Warburg effect. The high energy metabolism in melanoma, powered by the substantially more efficient energy pathway, the up-regulated oxidative phosphorylation compared to the alternatives, along with its high ability to metastasize, explained why the cancer is so aggressive. In contrast, BCC, while using energy from oxidative phosphorylation, seems to have an obvious block from the immune system, which appears to prevent cell invasion and metastasis, hence making the cancer one of the least deadly cancers.

The new insights derived in this study are global in nature and lack of detailed mechanism information due to the low-resolution nature of the non-paired datasets used in this study. We anticipate that higher-resolution insights could be derived using the same computational approach but on paired cancer-control datasets.

We believe that our study provides a new and highly effective way to gain new insights and understanding about certain unique characteristics of different cancers in their formation and progression through mining large-scale gene-expression data across multiple cancer types, but focused on key relevant cancer pathways.

Figures

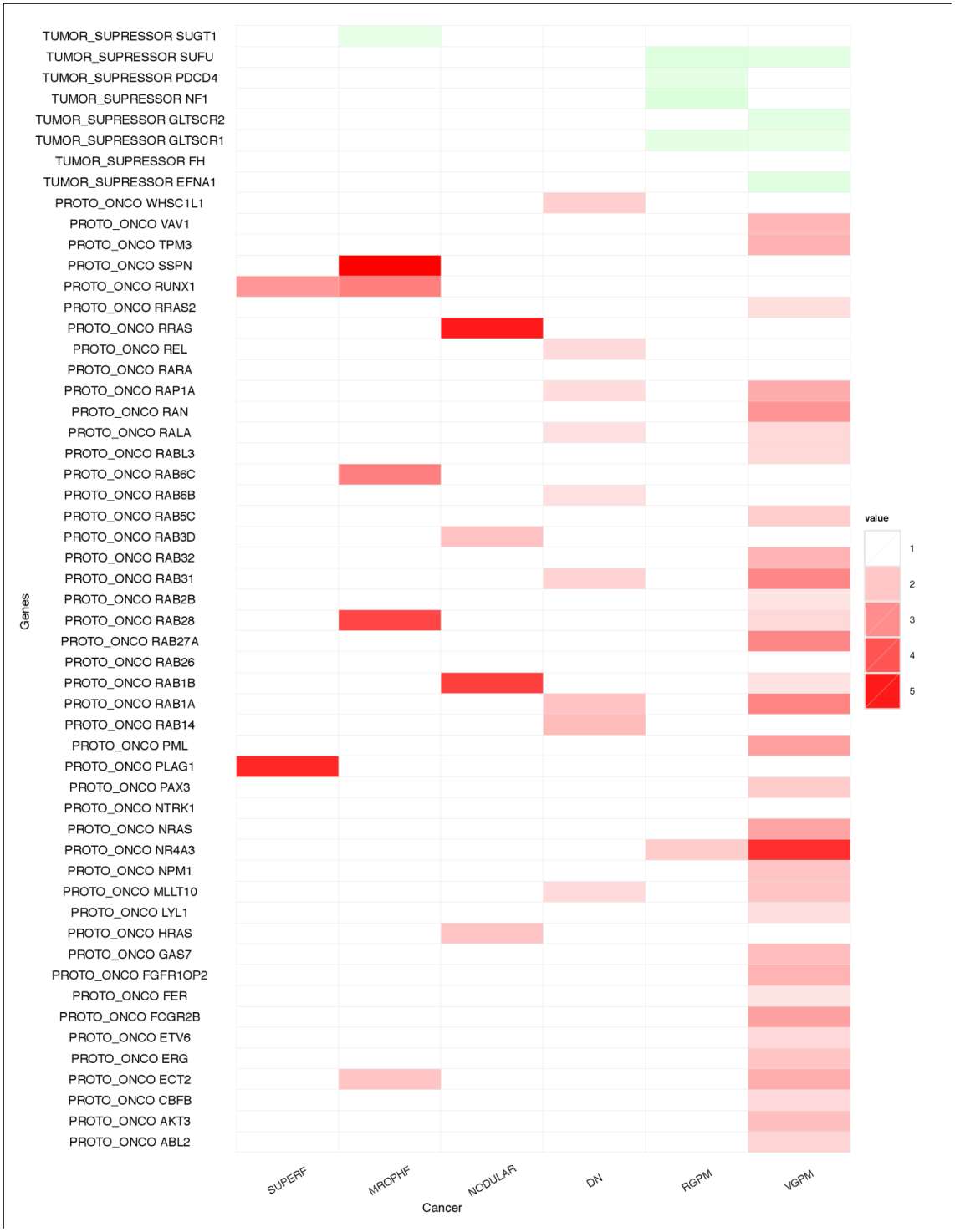


Figure 3.1: Expression level changes of proto-oncogene and tumor-suppressor genes for two skin cancer types. We only consider the up-regulated proto-oncogene and down-

regulated tumor-suppressor genes. Each row represents expression changes of a gene across all the cancer types under study. Each column represents one cancer type. The fold change of gene expression is color-coded with red, white and green for up-, no and down-regulation.

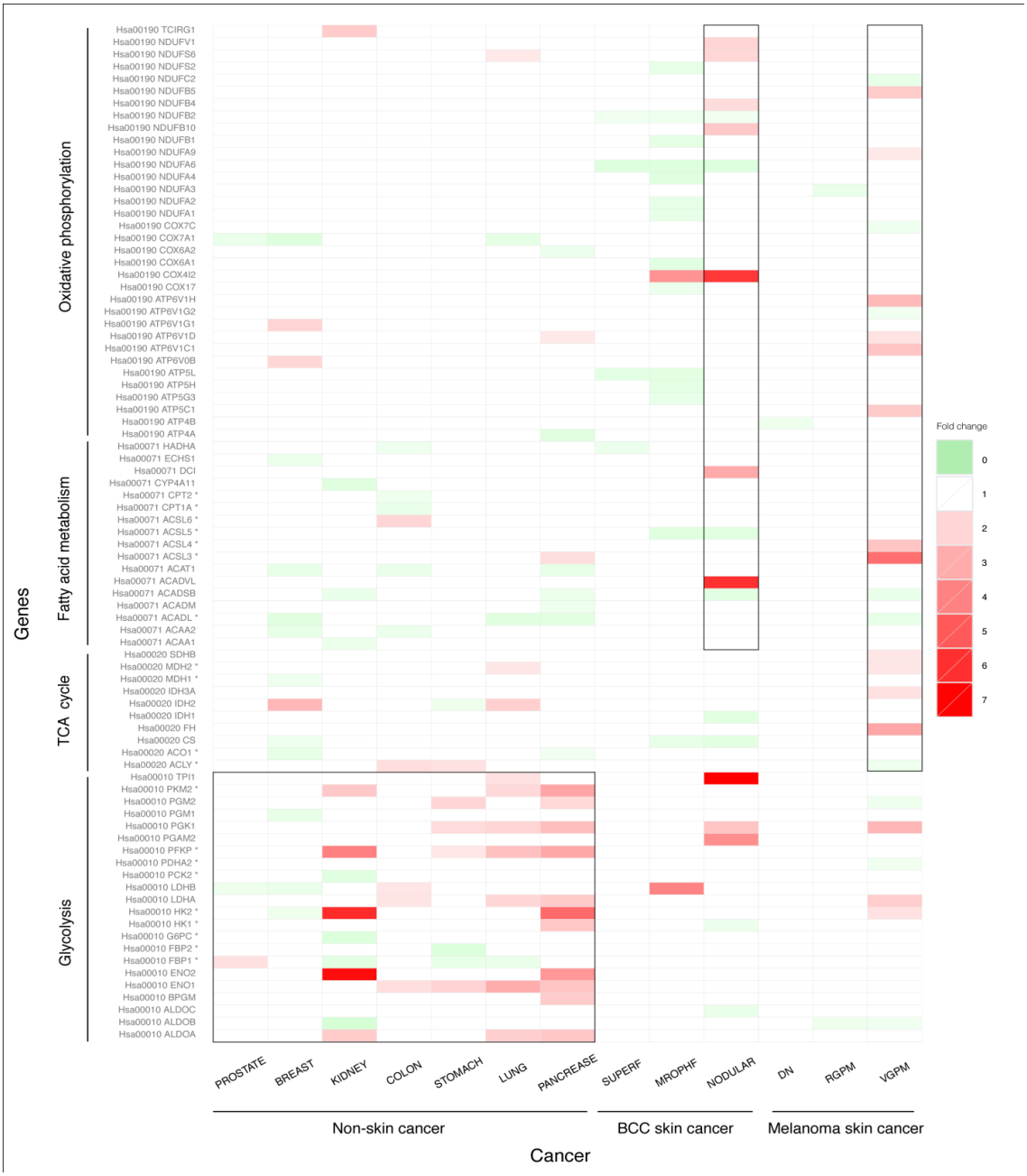


Figure 3.2: Expression level changes of genes involved in the four types of energy metabolism for two skin cancer types and seven non-skin cancer types. Each row represents expression changes of a gene across all the cancer types under study. Rate limiting enzymes in each pathway are indicated using *****. Each column represents one

cancer type. The fold change of gene expression is color-coded with red, white and green for up-, no and down-regulation.

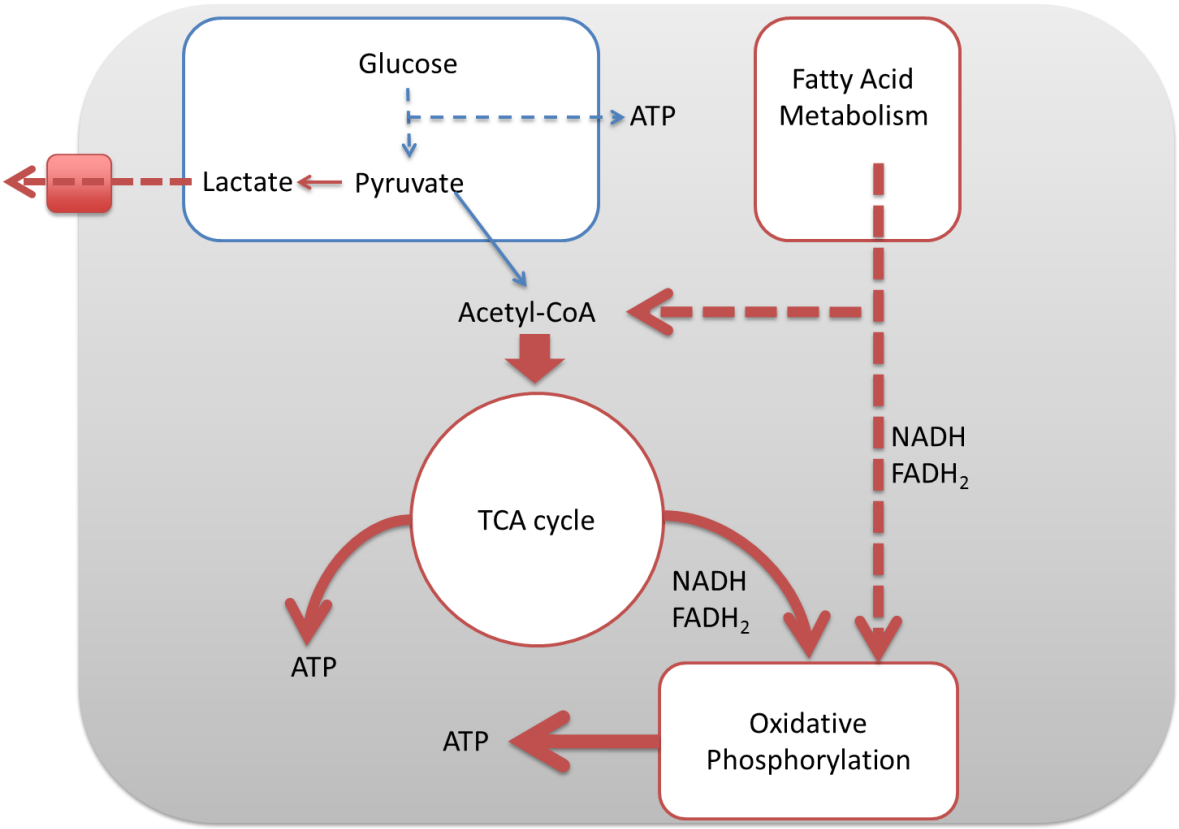


Figure 3.3: A model for energy metabolism for VGP melanoma.

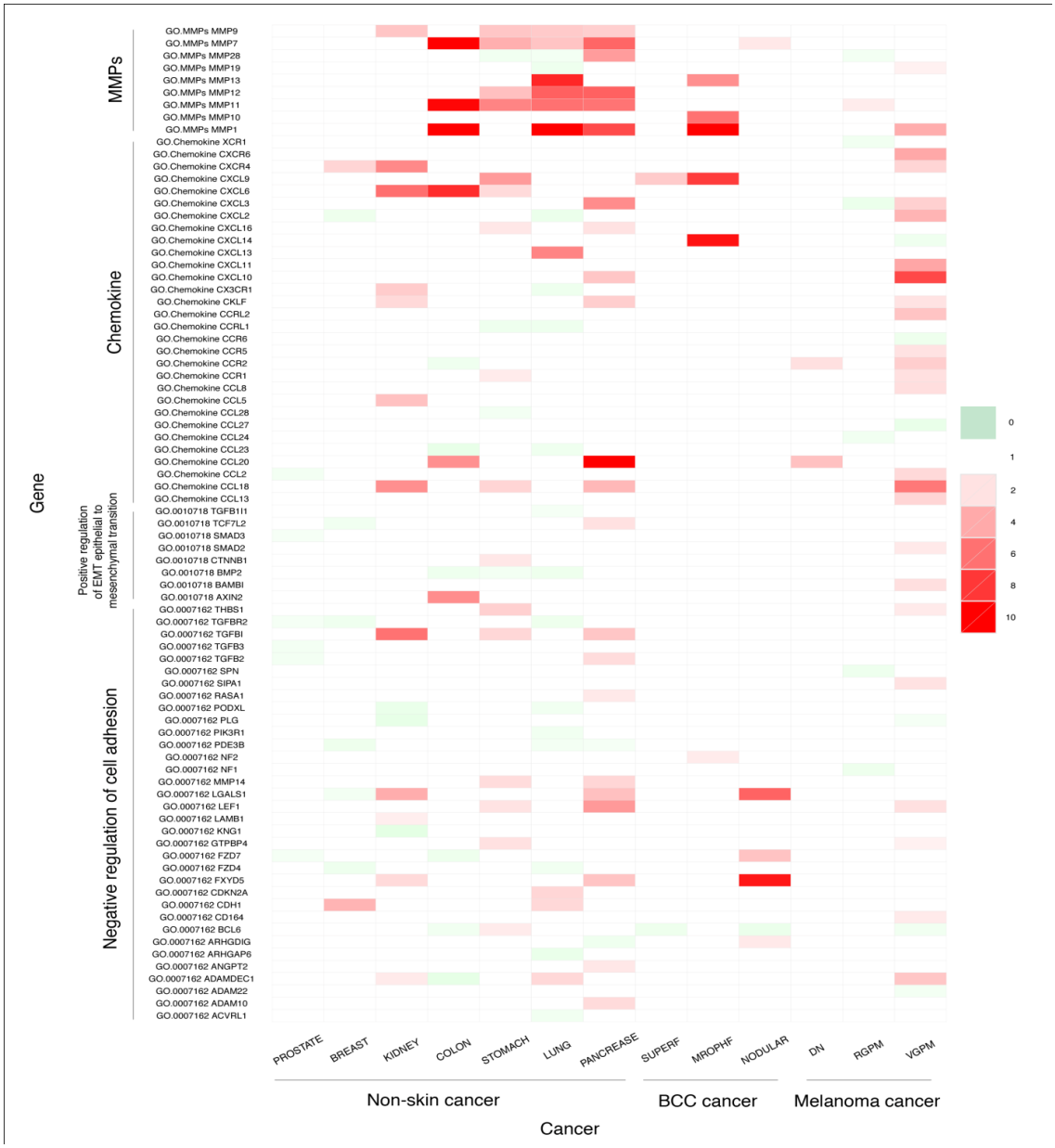


Figure 3.4: Expression changes of genes involved in the pro-metastasis of the two skin cancer types and other seven non-skin cancer types.

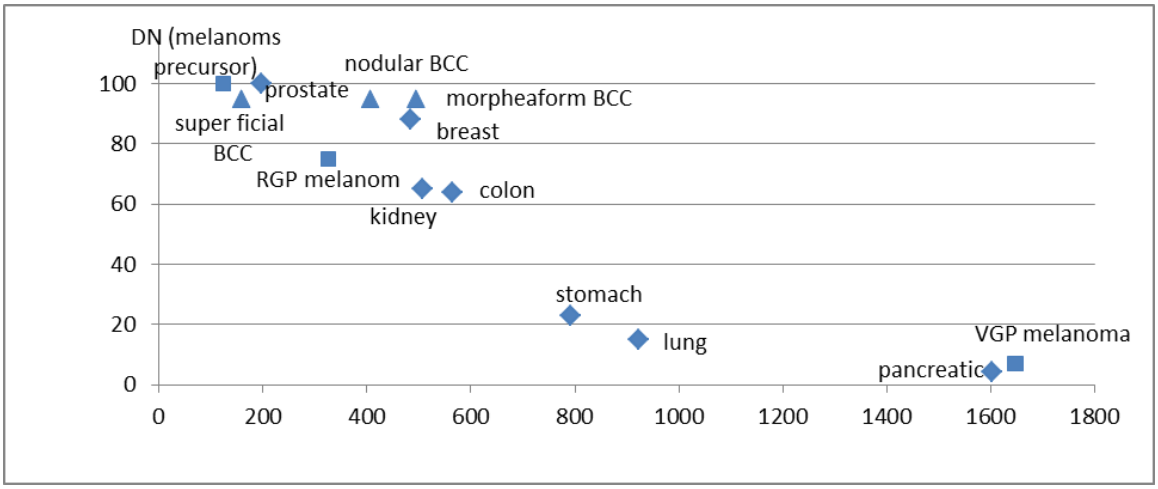


Figure 3.5: Correlation between 5-year survival rate and the number of differentially genes in each cancer type using the same statistical significance cutoff (see Material and Methods). The x-axis is the five-year survival rate ranging from 0 to 100% (data used from www.cancer.org), and the y-axis is the number of differentially expressed genes for each cancer type.

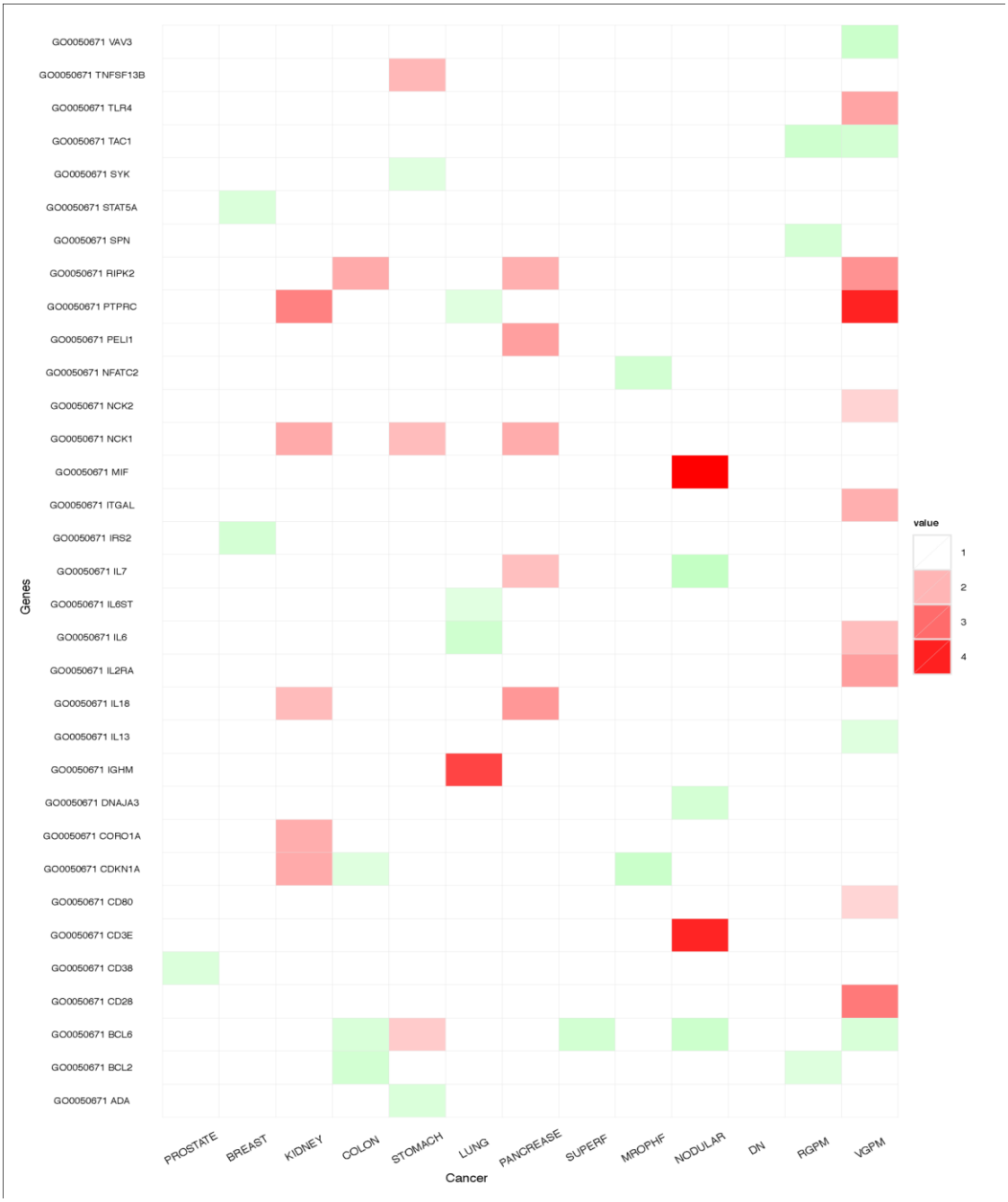


Figure 3.6: Expression level changes of genes involved in the positive regulation of lymphocyte proliferation for two skin cancer types and seven non-skin cancer types. Each row represents expression changes of a gene across all the cancer types under study. Each column represents one cancer type. The fold change of gene expression is color-coded with red, white and green for up-, no and down-regulation.

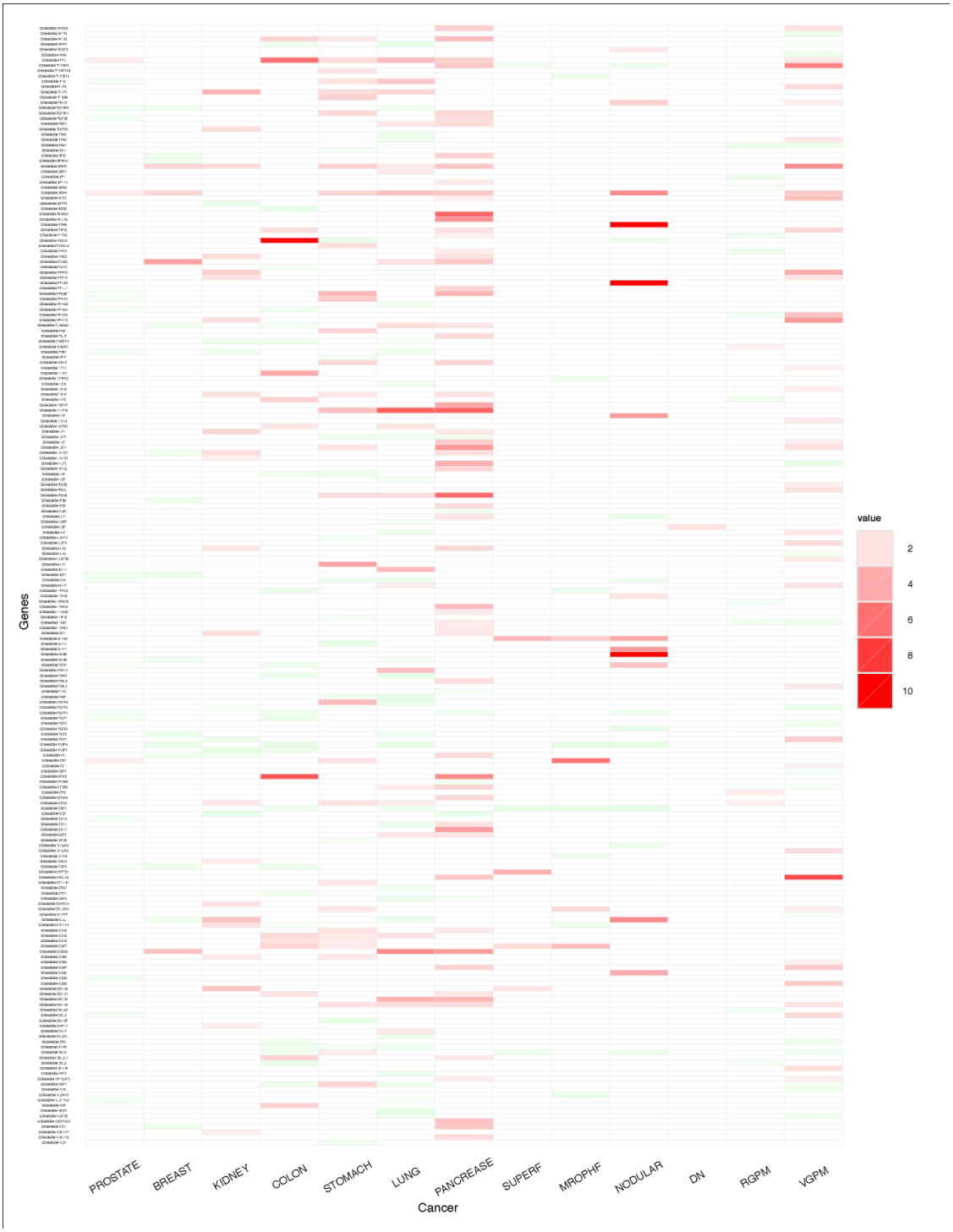


Figure 3.7: Expression level changes of genes involved in the positive regulation of cell proliferation for two skin cancer types and seven non-skin cancer types. Each row represents expression changes of a gene across all the cancer types under study. Each

column represents one cancer type. The fold change of gene expression is color-coded with red, white and green for up-, no and down-regulation.

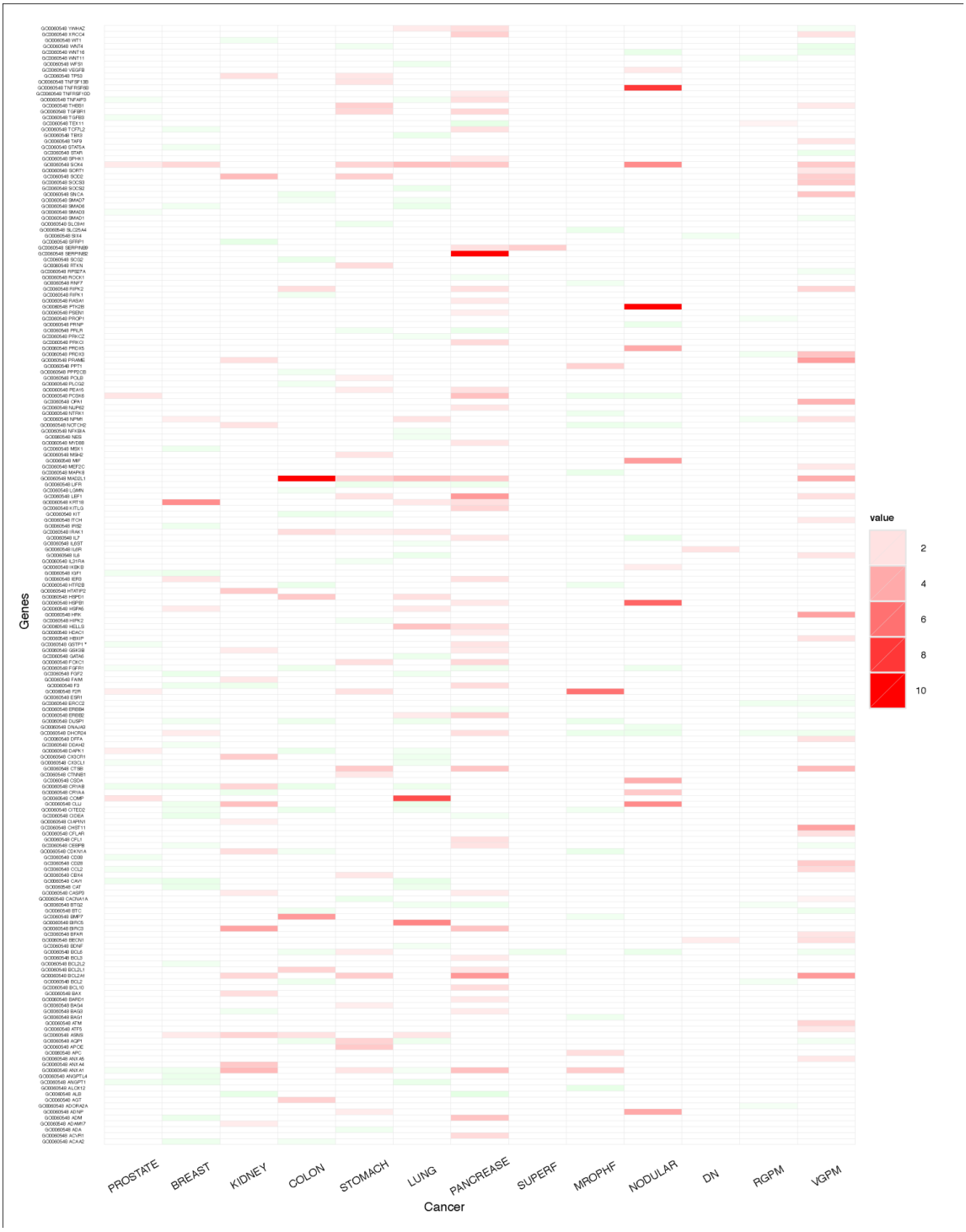


Figure 3.8: Expression level changes of genes involved in the negative regulation of cell death for two skin cancer types and seven non-skin cancer types. Each row represents expression changes of a gene across all the cancer types under study. Each column represents one cancer type. The fold change of gene expression is color-coded with red, white and green for up-, no and down-regulation.

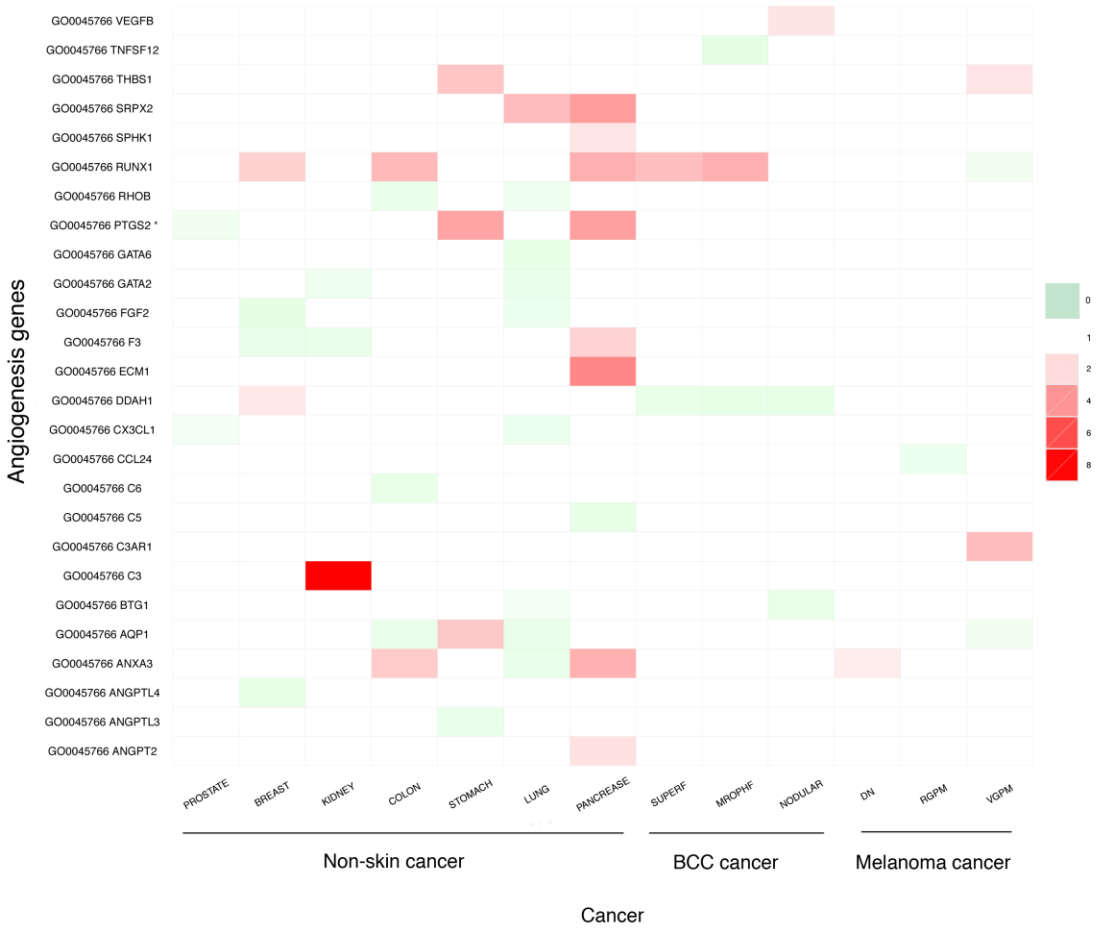
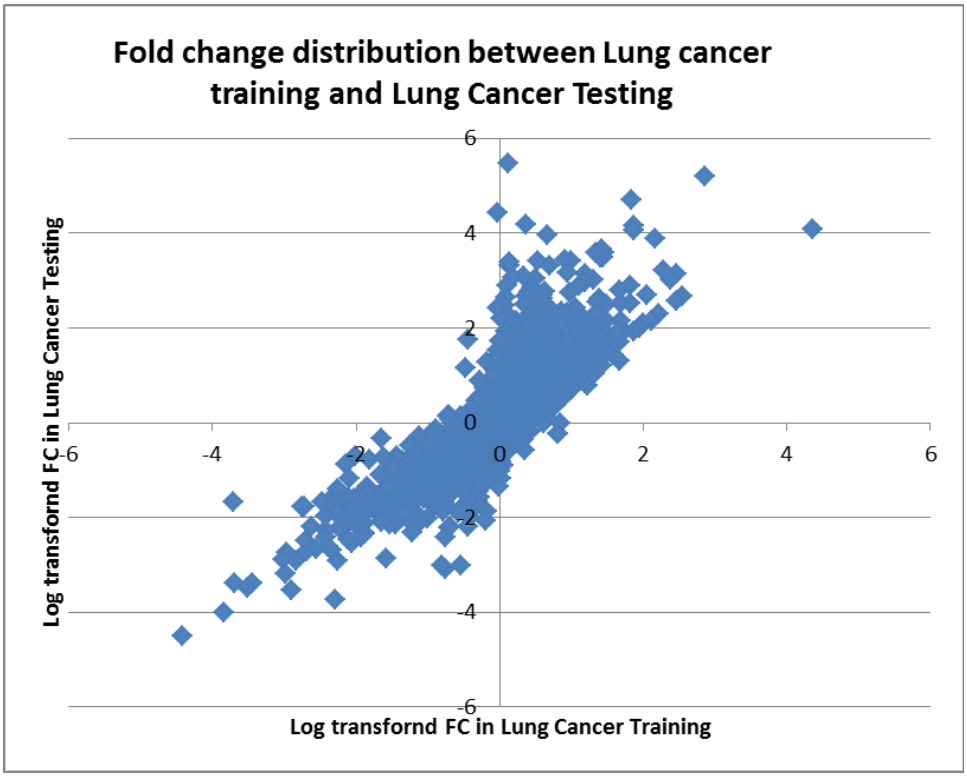
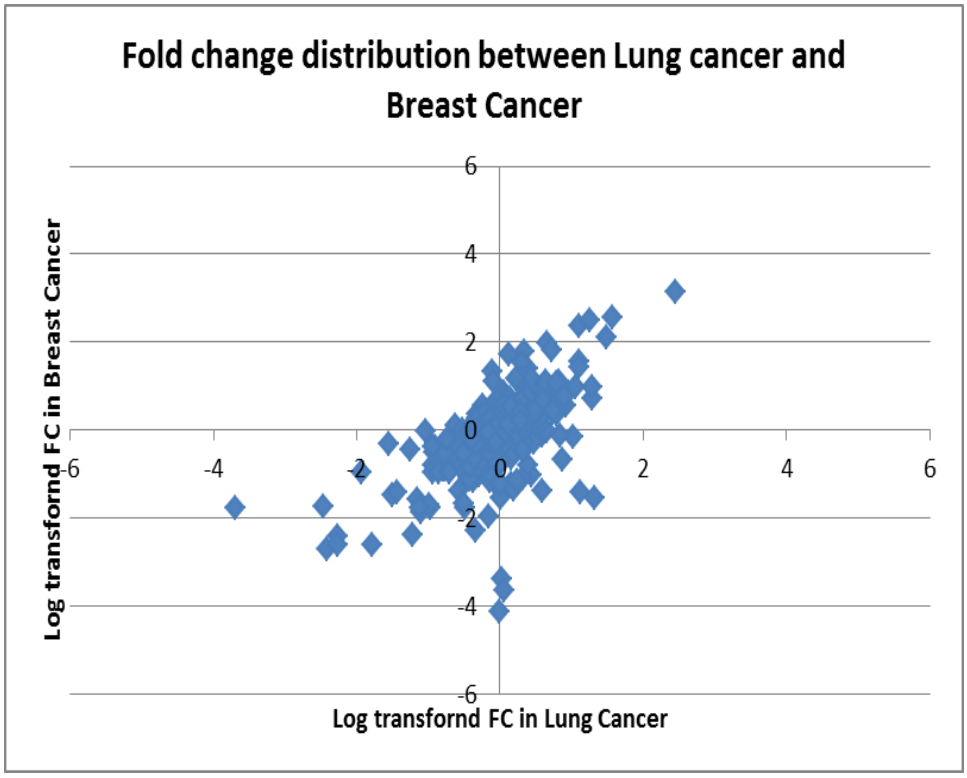


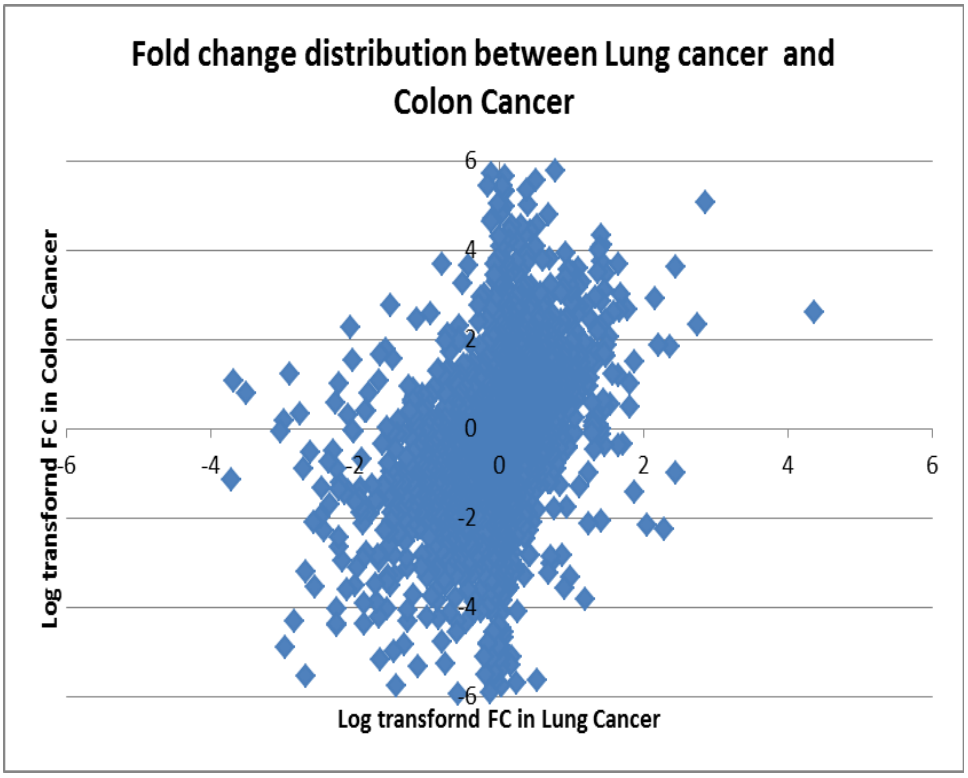
Figure 3.9: Expression changes of genes involved in the pro-angiogenesis of the two skin cancer types and other seven non-skin cancer types.


Fold change distribution between Lung cancer training and Lung Cancer Testing
Log transfornd FC in Lung Cancer Testing
Log transfornd FC in Lung Cancer Training
6
4
2
0
-2
-4
-6
-6
-4
-2
0
2
4
6

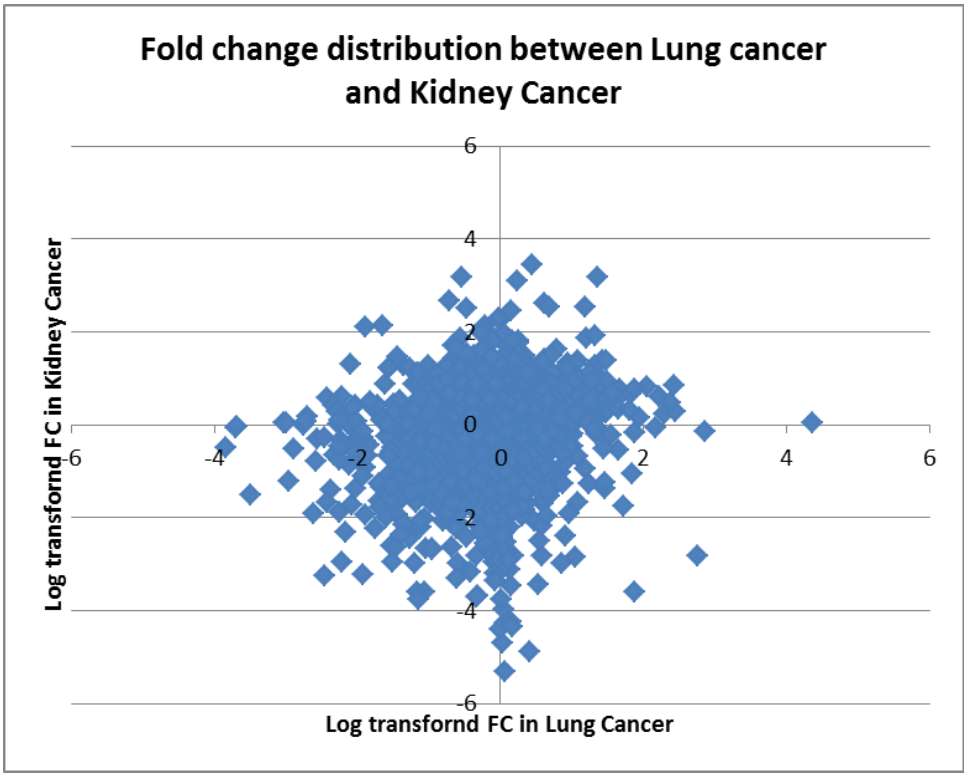
(A)


Fold change distribution between Lung cancer and Breast Cancer
Log transfornd FC in Breast Cancer
Log transfornd FC in Lung Cancer
6
4
2
0
-2
-4
-6
-6
-4
-2
0
2
4
6

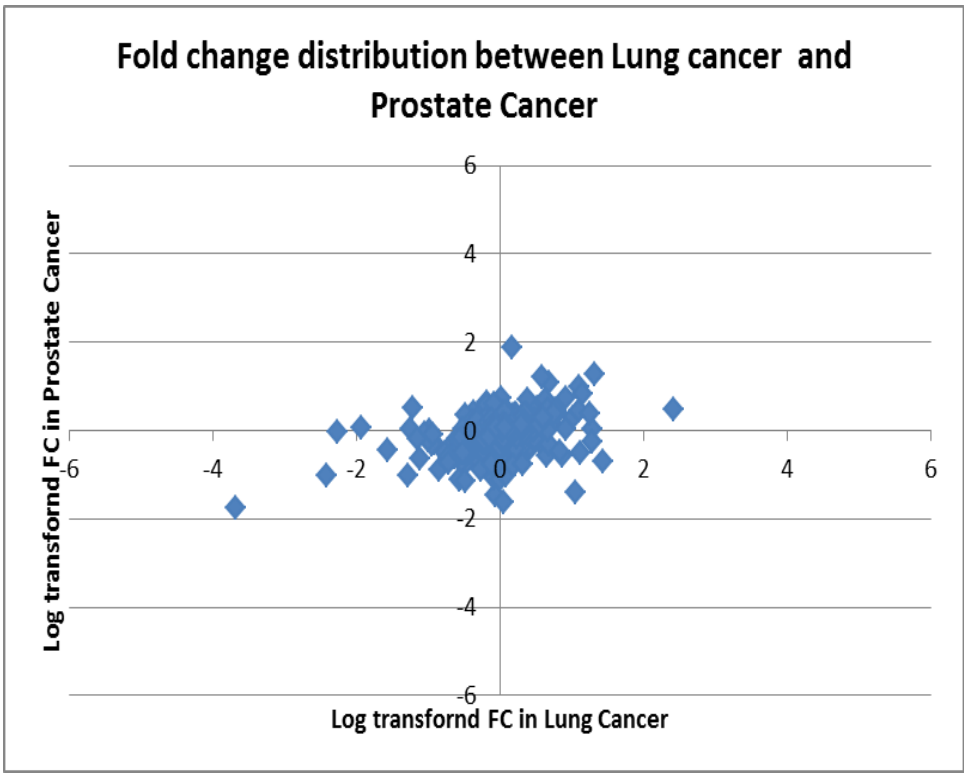
(B)


Fold change distribution between Lung cancer and Colon Cancer
Log transfornd FC in Colon Cancer
Log transfornd FC in Lung Cancer
6
4
2
0
-2
-4
-6
-6
-4
-2
0
2
4
6

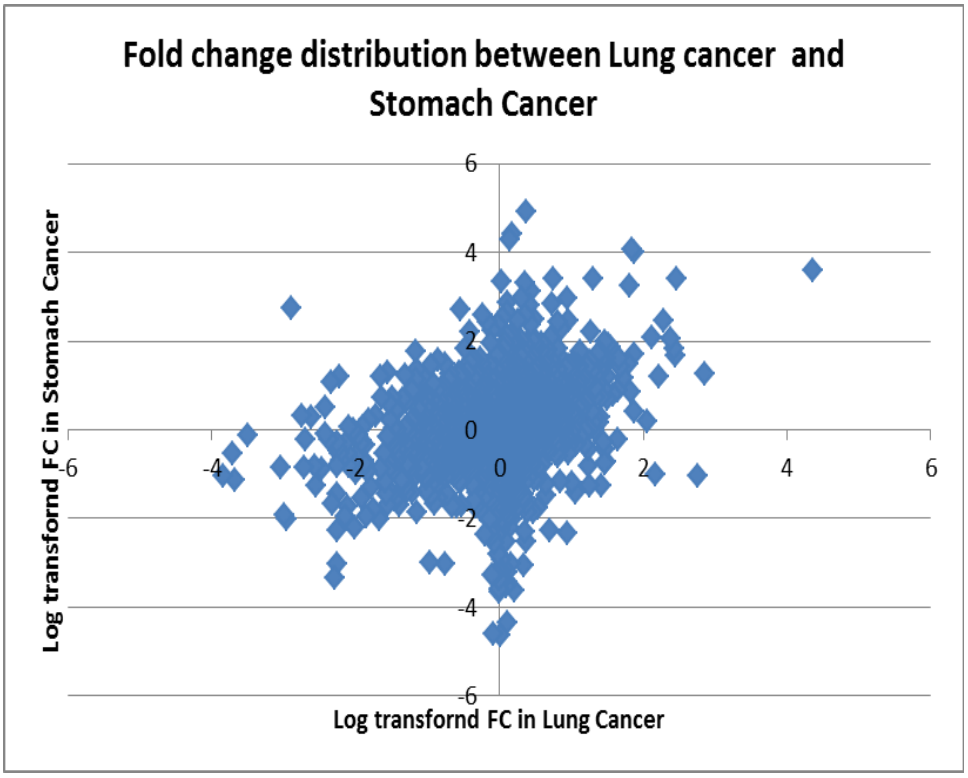
(C)


Fold change distribution between Lung cancer and Kidney Cancer
Log transfornd FC in Kidney Cancer
Log transfornd FC in Lung Cancer
6
4
2
0
-2
-4
-6
-6
-4
-2
0
2
4
6

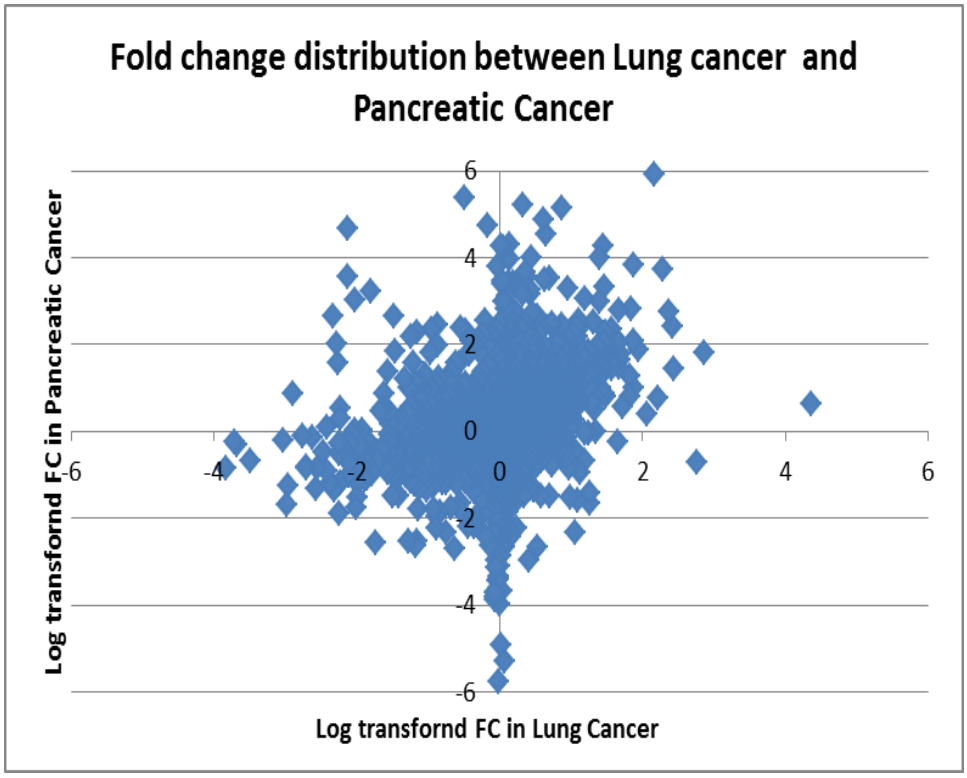
(D)


Fold change distribution between Lung cancer and Prostate Cancer
Log transfornd FC in Prostate Cancer
Log transfornd FC in Lung Cancer
6
4
2
0
-2
-4
-6
-6
-4
-2
0
2
4
6

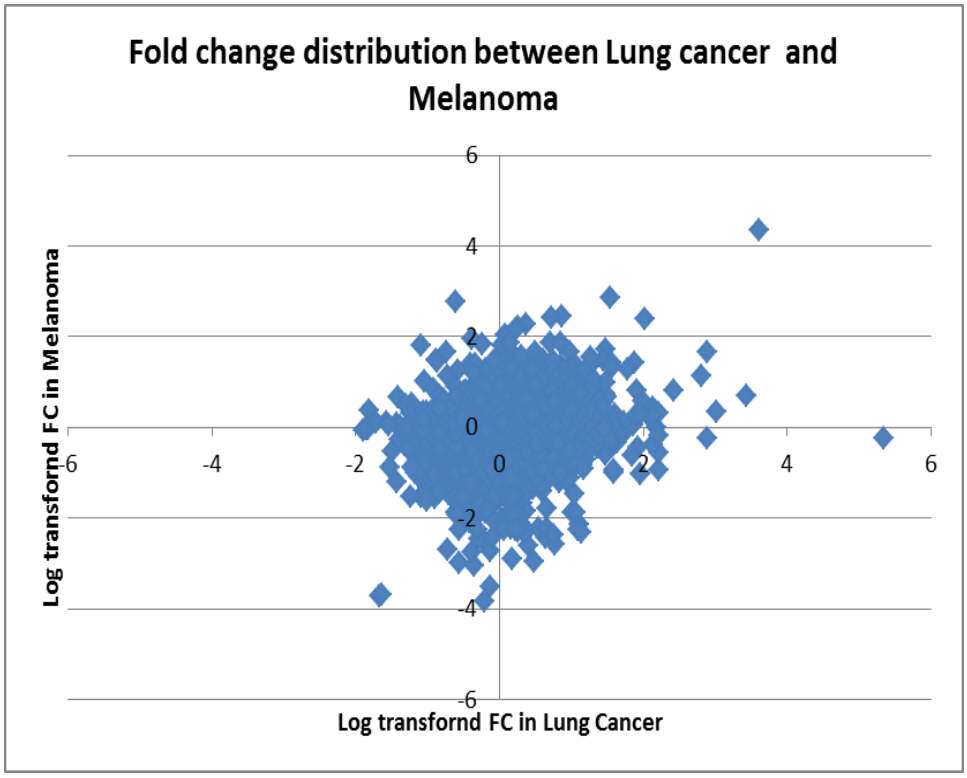
(E)


Fold change distribution between Lung cancer and Stomach Cancer
Log transfornd FC in Stomach Cancer
Log transfornd FC in Lung Cancer
6
4
2
0
-2
-4
-6
-6
-4
-2
0
2
4
6

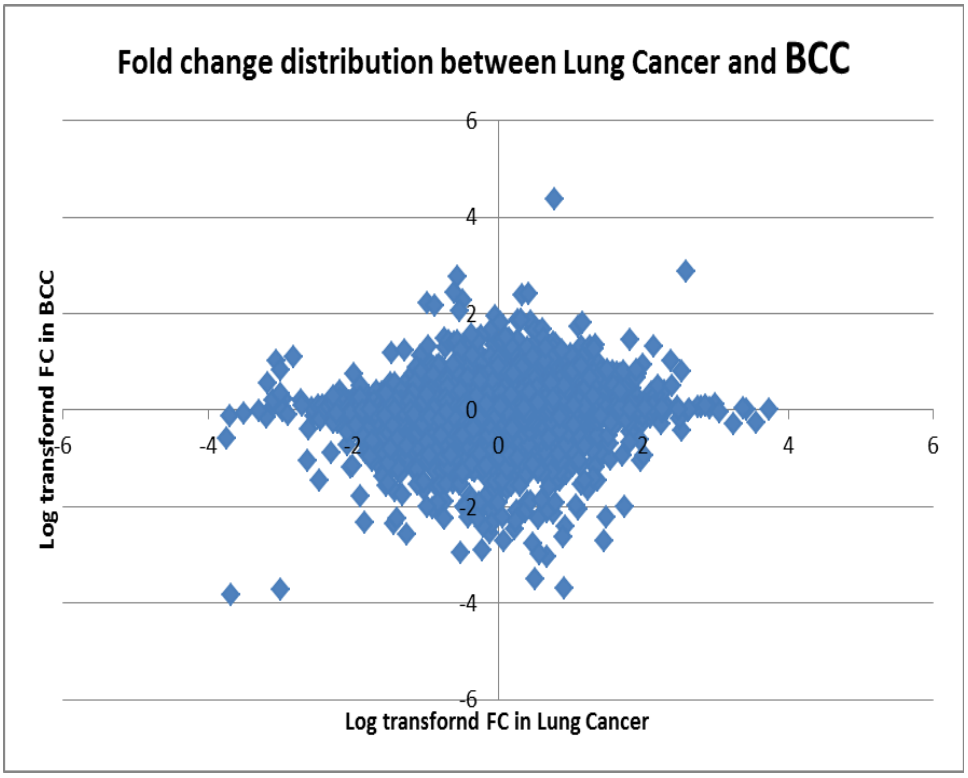
(F)


Fold change distribution between Lung cancer and Pancreatic Cancer
Log transfornd FC in Pancreatic Cancer
Log transfornd FC in Lung Cancer
6
4
2
0
-2
-4
-6
-6
-4
-2
0
2
4
6

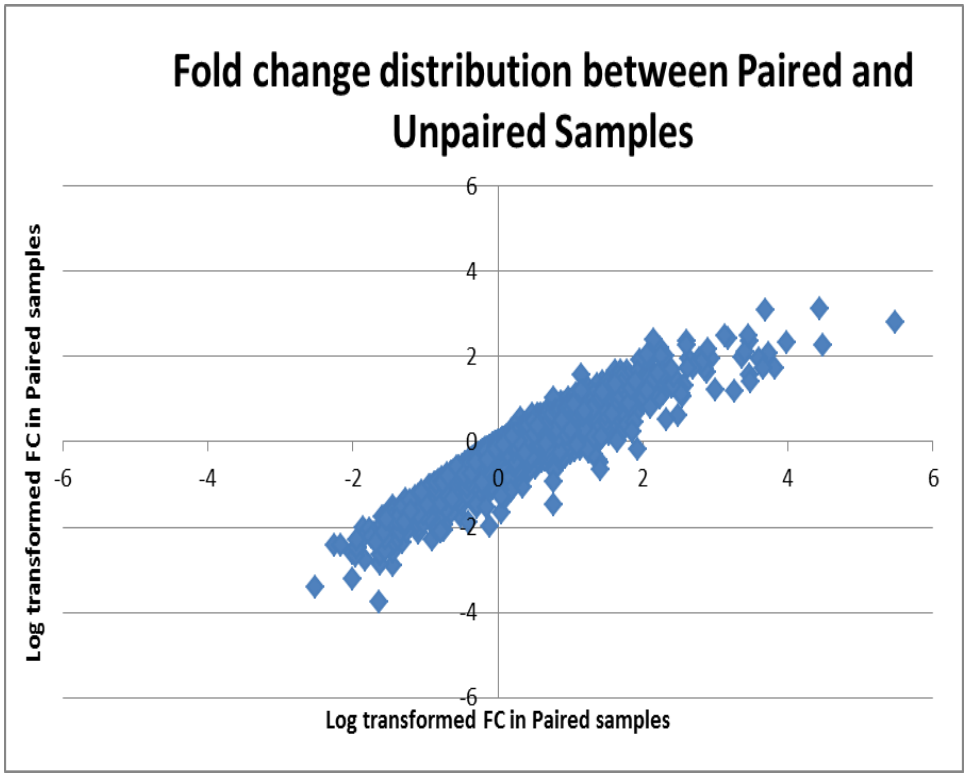
(G)


Fold change distribution between Lung cancer and Melanoma
Log transfornd FC in Melanoma
Log transfornd FC in Lung Cancer
6
4
2
0
-2
-4
-6
-6
-4
-2
0
2
4
6

(H)


Fold change distribution between Lung Cancer and BCC
Log transfornd FC in BCC
Log transfornd FC in Lung Cancer
6
4
2
0
-2
-4
-6
-6
-4
-2
0
2
4
6

(I)


Fold change distribution between Paired and Unpaired Samples
Log transformed FC in Paired samples
Log transformed FC in Paired samples
6
4
2
0
-2
-4
-6
-6
-4
-2
0
2
4
6

(J)

Figure 3.10: Comparison of the gene expression fold changes. (A) between training and testing datasets of lung cancer, and (B-I) between different types of cancer. The distributions of the fold-changes (FC) of individual genes across all genes between cancer and the corresponding control tissues for the seven types of cancers were checked and found to be similar. (J) Comparison of the gene expression fold changes between paired and unpaired breast cancer. (PC: Pearson Correlation, P-value <0.005))

Tables

Table 3.1 : The enriched pathways by all the cancer types in the study. (each cell in the represents the number of genes differentially expressed in the corresponding cancer type and pathway)

genes	PROSTATE	BREAST	KIDNEY	COLON	STOMACH	LUNG	PANCREASE	SUPERF	MROPHF	NODULAR	DN	RGPM	VGPM
hsa05200:Pathways in cancer	18	18	20	28	29	40	56		16		5		40
hsa04010:MAPK signaling pathway	8			17			28		13	19		10	
hsa05010:Alzheimer's disease									13	8			
hsa05012:Parkinson's disease									13	8			
hsa00190:Oxidative phosphorylation									13	7			
hsa05016:Huntington's disease									12	9			
hsa04510:Focal adhesion	16	15	12	14	22	29	38		10			7	22
hsa04310:Wnt signaling pathway	6			10			19		10				18
hsa04810:Regulation of actin cytoskeleton	9				17		32		9	10			26
hsa04530:Tight junction				9		18	16		8	8			16
hsa03010:Ribosome									6			5	22
hsa04512:ECM-receptor interaction	8	12			16	20	20		6				13
hsa04916:Melanogenesis							13		6	7			12
hsa03320:PPAR signaling pathway		11	11	9		9			6				
hsa04270:Vascular smooth muscle contraction	11			13		17			5	6			
hsa04540:Gap junction				7		9			5	6			
hsa04114:Oocyte meiosis				8			13		5				
hsa04722:Neurotrophin signaling pathway							13		5				
hsa05210:Colorectal cancer	6	5	8	7		10	11		5				
hsa04060:Cytokine-cytokine receptor interaction			14		25	28	34						32
hsa04062:Chemokine signaling pathway			12			18	19					7	29
hsa04142:Lysosome			10	8	10		15						26
hsa04120:Ubiquitin mediated												6	23

proteolysis													
hsa04144:Endocytosis			11										21
hsa04110:Cell cycle	5	8	9	17	26	23	28						20
hsa00230:Purine metabolism				14	11	17							18
hsa04650:Natural killer cell mediated cytotoxicity			9				16						18
hsa04514:Cell adhesion molecules (CAMs)		9	11		10	18	14						17
hsa04612:Antigen processing and presentation			7										16
hsa04620:Toll-like receptor signaling pathway						10							16
hsa04660:T cell receptor signaling pathway			7			11							16
hsa04670:Leukocyte transendothelial migration		7	12			13	17						15
hsa04012:ErbB signaling pathway				6			15						14
hsa05416:Viral myocarditis		6	5	9		10	11						14
hsa04664:Fc epsilon RI signaling pathway				6		8							12
hsa04666:Fc gamma R-mediated phagocytosis		6	9			11	15						12
hsa04914:Progesterone-mediated oocyte maturation				6	7	10							12
hsa05322:Systemic lupus erythematosus				7									12
hsa00520:Amino sugar and nucleotide sugar metabolism													11
hsa00982:Drug metabolism			6										11
hsa04115:p53 signaling pathway			11	7	13	14	14						11
hsa04210:Apoptosis			6	7			16						11
hsa04621:NOD-like receptor signaling pathway			5				8						11
hsa04640:Hematopoietic cell lineage			6	6		10	13						11
hsa04662:B cell receptor signaling pathway			5				10						11
hsa00980:Metabolism of xenobiotics by cytochrome P450													10
hsa03050:Proteasome					5		7						10
hsa05332:Graft-versus-host disease													10
hsa04672:Intestinal immune network for IgA production				5	6								9
hsa04920:Adipocytokine		6		5									9

signaling pathway													
hsa05212:Pancreatic cancer	5				8	11	15						9
hsa05217:Basal cell carcinoma					7								9
hsa00010:Glycolysis / Gluconeogenesis		6	10		7	9	11			5			8
hsa04940:Type I diabetes mellitus				5									8
hsa05110:Vibrio cholerae infection						7							8
hsa05320:Autoimmune thyroid disease													8
hsa05330:Allograft rejection													8
hsa00511:Other glycan degradation													7
hsa00591:Linoleic acid metabolism													7
hsa03420:Nucleotide excision repair					5								7
hsa00020:Citrate cycle (TCA cycle)													6
hsa00071:Fatty acid metabolism		6		6									6
hsa00280:Valine, leucine and isoleucine degradation		8	8				6						6
hsa00620:Pyruvate metabolism		8				5							6
hsa00983:Drug metabolism			6										6
hsa03030:DNA replication			5	6	10	8	5						6
hsa04130:SNARE interactions in vesicular transport													6
hsa05310:Asthma													6
hsa05340:Primary immunodeficiency													6
hsa00250:Alanine, aspartate and glutamate metabolism						5							5
hsa00565:Ether lipid metabolism		6		5		7							5
hsa03430:Mismatch repair					6								5
hsa04080:Neuroactive ligand-receptor interaction											7		
hsa04630:Jak-STAT signaling pathway		8				15					7		
hsa04910:Insulin signaling pathway		10					15			9	5		
hsa03040:Spliceosome										8			
hsa04020:Calcium signaling pathway	8			11		16	23			8			
hsa04912:GnRH signaling										7			

pathway													
hsa05414:Dilated cardiomyopathy	8			7		11	17			6			
hsa04720:Long-term potentiation										5			
hsa04730:Long-term depression				7						5			
hsa00030:Pentose phosphate pathway			7				5						
hsa00040:Pentose and glucuronate interconversions													
hsa00051:Fructose and mannose metabolism			7			7	5						
hsa00052:Galactose metabolism							5						
hsa00240:Pyrimidine metabolism			8	7	10								
hsa00260:Glycine, serine and threonine metabolism			5	5			8						
hsa00270:Cysteine and methionine metabolism							7						
hsa00310:Lysine degradation													
hsa00330:Arginine and proline metabolism				6	5	6							
hsa00340:Histidine metabolism		6											
hsa00350:Tyrosine metabolism			5										
hsa00380:Tryptophan metabolism		6				5							
hsa00410:beta-Alanine metabolism		5	5										
hsa00480:Glutathione metabolism			5	6	5		8						
hsa00500:Starch and sucrose metabolism							6						
hsa00510:N-Glycan biosynthesis						5							
hsa00561:Glycerolipid metabolism		7											
hsa00562:Inositol phosphate metabolism						7							
hsa00564:Glycerophospholipid metabolism		7				7							
hsa00590:Arachidonic acid metabolism				6	5								
hsa00600:Sphingolipid metabolism							5						
hsa00630:Glyoxylate and dicarboxylate metabolism		5											
hsa00640:Propanoate		8		5			6						

metabolism													
hsa00650:Butanoate metabolism		6											
hsa00910:Nitrogen metabolism				5									
hsa02010:ABC transporters				6		6							
hsa03018:RNA degradation					5								
hsa03410:Base excision repair					6								
hsa04070:Phosphatidylinositol signaling system						8							
hsa04260:Cardiac muscle contraction	5												
hsa04350:TGF-beta signaling pathway	6			9		12							
hsa04360:Axon guidance		9			11	15	27						
hsa04370:VEGF signaling pathway							9						
hsa04520:Adherens junction	6	6				12	16						
hsa04610:Complement and coagulation cascades			9	8			10						
hsa04742:Taste transduction													
hsa04930:Type II diabetes mellitus							6						
hsa04960:Aldosterone-regulated sodium reabsorption		6					6						
hsa05014:Amyotrophic lateral sclerosis (ALS)			5		5								
hsa05020:Prion diseases						5							
hsa05120:Epithelial cell signaling in Helicobacter pylori infection			6				8						
hsa05130:Pathogenic Escherichia coli infection					6	6	11						
hsa05211:Renal cell carcinoma						9	10						
hsa05213:Endometrial cancer							7						
hsa05214:Glioma				6		7	8						
hsa05215:Prostate cancer			6	7		9	11						
hsa05216:Thyroid cancer							6						
hsa05218:Melanoma			6	8	6	9							
hsa05219:Bladder cancer			5	6	8	10	7						
hsa05220:Chronic myeloid leukemia		6		6		8	11						
hsa05221:Acute myeloid leukemia		5					9						
hsa05222:Small cell lung cancer	5			8	11	14	17						
hsa05223:Non-small cell lung cancer					5	7	7						

hsa05410:Hypertrophic cardiomyopathy (HCM)	8			7		11	15						
hsa05412:Arrhythmogenic right ventricular cardiomyopathy (ARVC)	5			6	8	10	19						

Table 3.2: A summary of the non-skin cancer data used in our analysis

Cancer	GEO dataset ID training / testing data	# reference/ #cancer samples
breast cancer	GSE15852 / GSE10810	43/43 (27/31)
colon cancer	GSE6988 / GSE10950	28/53 (24/24)
kidney cancer	GSE15641 / GSE17816	23/49 (9/36)
lung cancer	GSE10072 / GSE7670	49/58 (27/27)
pancreatic cancer	GSE15471 / GSE16515	39/39 (16/36)
prostate cancer	GSE6606 / GSE3933	63/65 (47/62)
stomach cancer	GSE2701 / GSE13911	23/89 (31/38)
Basal Cell Carcinoma	GSE6520 / GSE12542	8/23 (8/8)
Melanoma	GSE12931 / GSE3189	18/34 (7/45)

CHAPTER 4

A SYSTEMS BIOLOGY APPROACH TO ELUCIDATION OF HOW CANCER CELLS AVOID ACIDOSIS[5]

[5]Xu K and Mao X. et al. Submitted to PLoS Computational Biology, 11/9/12

Abstract

The rapid growth of cancer cells fueled by glycolysis produces large amounts of protons in cancer cells, which triggers various mechanisms to transport them out, hence leading to increased acidity in their extracellular environments. It has been well established that the increased acidity will induce cell death of normal cells but not cancer cells. The main question we address here is: how cancer cells deal with the increased acidity to avoid the activation of apoptosis. We have carried out a comparative analysis of transcriptomic data of six solid cancer types, breast, colon, liver, two lung (adenocarcinoma, squamous cell carcinoma) and prostate cancers, and proposed a model of how cancer cells utilize a few mechanisms to keep the protons outside of the cells. The model consists of a number of previously studied, well or partially, mechanisms for transporting out the excess protons, such as through the monocarboxylate transporters, V-ATPases, NHEs and the one facilitated by carbonic anhydrases. In addition we propose a new mechanism that neutralizes protons through the conversion of glutamate to γ-aminobutyrate, which consumes one proton per reaction. We hypothesize that these processes are regulated by cancer related conditions such as hypoxia and growth factors and by pH levels, making these encoded processes not available to normal cells under acidic conditions.

Introduction

One of the key cancer hallmarks is their reprogrammed energy metabolism (Hanahan and Weinberg 2011). That is, glycolysis replaces oxidative phosphorylation to become the main ATP producer. A direct result of this change is that substantially more

lactates, as the terminal receivers of electrons from the glucose metabolism, are produced and transported out of the cells. To maintain the cellular electro-neutrality when releasing lactates, the cells release one proton for each released lactate, the anionic form of lactic acid. This leads to increased acidity in the extracellular environment of the cancer cells. It has been well established that high (extracellular) acidity can induce the apoptotic process in normal cells (Webster, Dischер et al. 1999), leading to the death of these cells. Interestingly this does not seem to happen to cancer cells, hence giving them a competitive advantage over the normal cells and allowing them to encroach the space occupied by the normal cells. Currently it is not well understood of how the cancer cells deal with the increased acidity in their extracellular environments to avoid acidosis.

A number of studies have been published focused on issues related to how cancer cells deal with the increased acidity in both the extracellular and intracellular environments (Wykoff, Beasley et al. 2000; Fang, Gillies et al. 2008; Sonveaux, Vegran et al. 2008; Swietach, Wigfield et al. 2008; Swietach, Wigfield et al. 2008; Neri and Supuran 2011; Hernandez 2012). The majority of these studies were focused on possible cellular mechanisms for transporting out or neutralizing intracellular protons, which are typically on one cancer type. More importantly these studies did not tie such observed capabilities and proposed mechanisms of cancer cells in avoiding acidosis with the rapid growth of cancer as we suspect there is an encoded mechanism that connects the two.

We have carried out a comparative analysis of genome-scale transcriptomic data collected on six types of solid cancers, namely breast, colon, liver, two lung (adenocarcinoma, squamous cell carcinoma) and prostate cancers, aiming to gain a systems level understanding of how the cancer cells keep their intracellular pH level

within the normal range while their extracellular pH level is low. Our analysis of the transcriptomic data on these cancer and their matching control tissues indicate that (i) all the six cancer types utilize the monocarboxylate transporters as the main mechanism to transport out lactates and protons simultaneously, triggered by the accumulation of intracellular lactates; (ii) these transporters are probably supplemented by additional mechanisms through anti-porters such as ATPases to transport protons out along with some cations such as Ca^{2+} or Na^{+} to reduce the intracellular acidity while maintaining the cellular electron-neutrality; and (iii) cancer cells may also utilize another mechanism, i.e., using glutamate decarboxylase to catalyze the decarboxylation of glutamate to a γ-aminobutyric acid (GABA), consuming one proton for each reaction -- a similar process is used by the bacterial *Lactococcus lactis* to neutralize acidity when lactates are produced. Based on these analysis results, we proposed a model that connects these deacidification processes with a number of cancer related genes/cellular conditions, which are probably intrinsic capabilities of fast-growing cells used under hypoxic conditions rather than gained capabilities through molecular mutations.

We believe that our study represents the first systemic study focused on how cancer cells deal with the acidic environment through the activation of the encoded acid resistance mechanisms triggered by cancer associated genes and conditions. These results have established a foundation for a novel model for how cancer cells avoid acidosis.

Results

1. Cellular responses to increased acidity

The degradation of each mole of glucose generates 2 lactates, 2 protons and 2 ATPs, detailed as

$$\text{glucose} + 2\text{ADP} + 2\text{Pi} \rightarrow 2\ \text{lactate} + 2\ H^+ + 2\ \text{ATP} + 2H_2O,$$

showing the source of the increased acidity when glycolysis serves as the main ATP producer in cancer cells (Gatenby and Gillies 2004); in contrast the complete degradation of glucose through oxidative phosphorylation is pH neutral. Clearly these extra protons need to removed or neutralized since otherwise they will induce apoptosis. The monocarboxylate transporter (MCT) has been reported to play a key role in maintaining the pH homeostasis (Feron 2009) with 4 isoforms, MCT1-4, to have crucial physiological roles in proton-linked transportation (Halestrap and Price 1999; Halestrap 2012). Previous studies have reported that a number of genes in the MCT family, namely MCT1, MCT2 and MCT4, are up-regulated in cancer such as breast, colon, lung and ovary cancers. Note that a monocarboxylate transporter transports out lactates and protons with a 1:1 stoichiometry to maintain cellular electron-neutrality (Halestrap and Meredith 2004).

Our transcriptomic data analyses of the six cancer types added to this knowledge that other members of the MCT family also show up-regulation in five out of the six cancer types. The only exception is the prostate cancer, which did not show any increased expression in any member of the proton-linked MCT. Figure 1 shows the transcription up-regulation of two proton-linked MCT member genes, namely MCT1 (SLC16A1) and

MCT4 (SLC16A4) in five cancer types. Specifically MCT4 shows up-regulation in four of the six cancer types, an observation that has not been reported before.

One published study suggests that MCT1 might be regulated by P53 (Boidot, Vegran et al. 2012) in cancer. Another study shows strong evidences that MCT1 and MCT4 are regulated by intracellular hypoxia. We hypothesize that hypoxia may be the main regulating factor of the over-expression of the MCT genes, which may require additional conditions such as pH level or accumulation of lactates as the co-regulating factors. This is consistent with our analysis result of transcriptomic data of cell lines collected under hypoxic condition, where MCT1 and MCT4 genes are up-regulated (see Figure 4.1).

The protons transported out of the cells will increase the acidity of the extracellular environment. Previous studies have shown that (normal) cells tend to adjust the intracellular pH level to a similar pH level of the extracellular environment (Fellenz and Gerweck 1988), possibly to keep the stress level low caused by the large pH gradient between the extracellular and intracellular environments. It has been well established that the increased intracellular acidity will induce apoptosis through directly activating the caspase genes, which bypasses the more upstream regulatory proteins of the apoptosis system such as p53, hence leading to the death of the normal cells that do not seem to have the proper intracellular conditions to deal with the reduced pH.

2. Additional mechanisms for dealing with excess protons in cancer cells

We have examined if other genes relevant to the removal or neutralization of protons in cancer cells systematically across all the human genes. Our main findings are summarized in Figure 1, detailed as follows.

V-ATPase: Transmembrane ATPases import many of the metabolites necessary for cell metabolism and export toxins, wastes and solutes that can hinder the health of the cells (Perez-Sayans, Somoza-Martin et al. 2009). One particular type of ATPase is the V-ATPase that catalyzes ATP hydrolysis to transport solutes out. It pumps out a proton in exchange for an extracellular Na^+ or another cation such as K^+ or Ca^{2+} to maintain the intracellular electro-neutrality. V-ATPases have been found to be up-regulated in multiple cancer types but the previous studies have been mostly focused on using the increased V-ATPases as a biomarker for metastasis (Sennoune, Bakunts et al. 2004) or on utilizing them as potential drug targets as a way to trigger apoptosis, hence causing cancer cell death (Sennoune, Bakunts et al. 2004; Sennoune, Luo et al. 2004; Fais, De Milito et al. 2007).

We have examined the expression levels of the 19 genes that encode the subunits of V-ATPase, the V_0 (transmembrane) domain and the V_1 (cytoplasmic) domain namely ATP6V0A1, ATP6V0A2, ATP6V0B, ATP6V0E1, ATP6V0E2, ATP6AP1 and ATP6AP2 for V0 and ATP6V1A, ATP6V1B1, ATP6V1C1, ATP6V1C2, ATP6V1D, ATP6V1E1, ATP6V1E2, ATP6V1F, ATP6V1G1, ATP6V1G2, ATP6V1G3 and ATP6V1H for V1. We found that multiple V-ATPase genes are up-regulated, indicating

that the V-ATPases are active in transporting the protons out. Interestingly some of the ATPase genes do not show up-regulation and some even show down-regulation in prostate cancer (Figure 1). More detailed examination of the gene expression data indicates that the actual expression levels of the ATPase genes are at the baseline level in both the prostate cancer and the adjacent control issues, hence the fold-change data are not particularly informative. Overall the data on prostate cancer seem to suggest that the acidity level in this cancer type is not substantially elevated. For the other five cancer types, the expression levels of some V-ATPase genes do not show changes in cancer. We found that these genes expression levels are also elevated in the control tissues compared to cell-line data of the corresponding tissue types (data not shown here), which is consistent with previously published data that the elevated acidic level in the extracellular environment can also induce increased expression of the V-ATPase genes in normal tissues (Padilla-Lopez and Pearce 2006). This may explain why some of the V-ATPase genes do not show overexpression in cancer *versus* control tissues.

Then the question is why cancer cells seem to handle the increased acidity better than the normal cells. Our hypothesis is that while pH may play some regulatory role of the expression of the V-ATPase genes, the main regulator of the V-ATPase is probably mTORC1 as it has been suggested recently (Pena-Llopis, Vega-Rubin-de-Celis et al. 2011). mTORC is one of the most important regulators relevant to cell growth, and they generally have dysregulated expressions in cancer. To check on this hypothesis, we have examined the gene expression level of mTORC1 (GBL and FRAP1) in the six cancer types. We see clear up-regulation of this gene in all six cancer types. So overall we

speculate that it is the combined effect of pH and up-regulation of mTORC1 that makes cancer cells more effective in pumping out the excess protons than the normal cells**.**

Na+-H+ Exchanger (NHE): NHE anti-porters represent another class of proteins that can transport out protons and in a cation to maintain intracellular electro-neutrality. We have examined the five genes encoding this class of transporters, and found that these genes are highly up-regulated in the two lung cancer types, which seem to play a complementary role to that of the V-ATPases as their expression-change patterns are highly complementary between NHE genes and the V-ATPase genes in five out six cancer types, specifically up-regulation in breast, colon and liver cancers but not in the two lung cancer types as shown in Figure 4.1. Literature search suggests that NHEs are regulated by both growth factors and pH among a few other factors, which may partially explain why the system is more active in cancer (affected by both growth factors and pH) than in control tissues (affected only by pH).

3. Carbonic anhydrases play roles in pH neutralization in cancer cells

It has been previously suggested that carbonic anhydrases (CAs) play a role in neutralizing the protons in cancer cells. For example, a model of how the membrane-associated CAs facilitate out-transportation of protons has been presented (Swietach, Vaughan-Jones et al. 2007). The key idea of the model is that the CAs catalyze the otherwise slow reaction from $CO_2 + H_2O$ to H_2CO_3, which dissociates into HCO_3^- and H^+ in an acidic extracellular environment, detailed as follows:

$$HCO_3^- + H^+ \rightleftharpoons H_2CO_3 \rightleftharpoons CO_2 + H_2O.$$

The HCO_3^- (bicarbonate) is then transported across the membrane through an NBC transporter (Johnson 2009), where it reacts with a H^+ to form a CO2 and H_2O; and the CO_2 is freely membrane-permeable to get inside the cell, forming a cycle for removing some of the excess H^+. See Figure 4.6 for a more detailed picture of this mechanism.

To check if the model is supported by the data, we found that (1) a number of the membrane-associated CAs (CA9, CA12, CA14) show up-regulation in five out of six cancer types (except for prostate cancer), as shown Figure 4.2; and (2) two of the three NBC genes, NBC2 (SLC4A5) and NBC3 (SLC4A7) show up-regulation in four cancer types. It has been reported that CA9 and CA12 are hypoxia inducible in brain cancer (Proescholdt, Mayer et al. 2005). Hence we assume that all the three above membrane-associated CAs are inducible by hypoxia. In addition, our literature search indicates that the NBC genes are pH inducible (Chiche, Brahimi-Horn et al. 2010).

Interestingly all the cytosolic CAs (CA2, CA3, CA7, CA13) show down-regulation, reflecting that oxidative phosphorylation is not being used as actively in cancer cells as in normal cells.

4. Neutralization of acidity through decarboxylation reactions: a novel mechanism?

Our search for possible mechanisms of cancer cells in deacidification led us to study how *Lactococcus lactis* deals with the lactic acids in their environment. We noted that the bacteria use the glutamate decarboxylases (GAD) to consume one (dissociable)

H^+ during the decarboxylation reaction that it catalyzes (Cotter and Hill 2003), as shown below:

$$OOC\text{-}CH_2\text{-}CH_2\text{-}CH(NH_2)\text{-}COO^{2-} + H^+ \rightarrow CO_2 + OOC\text{-}CH_2\text{-}CH_2\text{-}CH_2NH_2^-$$

or

$$Glu^{2-} + H^+ \rightarrow CO_2 + GABA^-$$

The reaction converts a glutamate to one γ-aminobutyrate (GABA) plus a CO_2. Two human homologues of the GAD, GAD1 and GAD2, are found. Published studies have shown that the activation of the GAD genes leads to GABA synthesis in human brain (Hyde, Lipska et al. 2011), suggesting that the human GAD genes have the same function as the bacterial GAB gene, i.e., catalyzing the reaction for the synthesis of GABA. Most of these studies are in the context of nervous systems in human brains (Kaila 1994; Owens and Kriegstein 2002; Yamada, Okabe et al. 2004). Specifically, GABA is known to serve as a key inhibitory neurotransmitter. In addition, activities of GABA have been identified in human liver (White and Sato 1978). While hypotheses have been postulated about its functions in liver (Lewis and Howdle 2003), no solid evidence has been established about its function there.

We have observed that GAD1 is up-regulated in three out six cancer types under study, namely colon, liver, lung adenocarcinoma, and GAD2 is up-regulated in prostate cancer. It has been fairly well established that glutamate, the substrate of the above reaction catalyzed by GAD, is elevated in cancer in general (DeBerardinis, Lum et al. 2008). Hence it makes sense to assume that the above reaction indeed takes place in cancer. This is supported by our observation that multiple in-take transporters are up-regulated in five

out six cancer types (see Figure 4.3). An even more interesting observation is that multiple genes encoding the out-going transporters of GABA are up-regulated in five out of the six cancer types, indicating that the GABA molecules are not being used by cancer cells but instead serves a way to remove H^+ out of the cells.

Currently no published data are available to implicate which genes encode the main regulator of the GAD genes, to the best of our knowledge. However, our search for possible regulators of the GAD genes in the Cscan database (Zambelli, Prazzoli et al. 2012) revealed that FOS, a known oncogene, can potentially regulate the GAD genes (Wang, Wu et al. 2003). Some experimental data from the ENCODE database (Rosenbloom, Dreszer et al. 2012) show that the expression of the GAD1 gene (NM_000817, NM_013445) is positively co-related with that of the FOS in the HUVEC cell-line. Putting all this information together, we hypothesize that FOS, in conjunction with some pH–associated regulator, regulates the GAD genes, which leads to the synthesis of GABA and reduces one H^+ as a by-product per synthesized GABA; and then the unneeded GABA molecules are transported out of the cells. This may provide another mechanism that cancer cells use to keep their intracellular pH level in the normal range.

5. A model for cancer cells to keep their intracellular pH in the normal range

Overall 44 genes are implicated in our above analyses. Our search results of these genes against Cscan database (Zambelli, Prazzoli et al. 2012) indicate that 28 out of the 44 genes are regulated directly by nine proto-oncogenes, namely BCL3, ETS1, FOS, JUN, MXI1, MYC, PAX5, SPI1 and TAL1; and 17 genes are regulated by two tumor-

suppressors, IRF1 and BRCA1 as shown in Figure 4.4, indicating that there is a strong connection between deacidification and cancer growth.

Figure 4.5 summarizes our overall model for the deacidification mechanisms and the associated conditions that may trigger each mechanism to be activated. Specifically, we hypothesize that hypoxia and growth factors may serve as the main regulatory factors of the deacdification processes in cancer mechanisms, hence making them available only in cancer cells, in conjunction with the cellular pH level,.

Materials and Method

1. Gene expression data for six cancer types

The gene-expression data for the six cancer types, (breast, colon, liver, lung adenocarcinoma, squamous cell lung, prostate), are downloaded from the GEO database (Edgar, Domrachev et al. 2002) of NCBI. For each cancer type, we have applied the following criteria in selecting the dataset used for this study: (1) all the data in each dataset were generated using the same platform by the same research group; (2) each dataset consists of only paired samples, i.e., cancer tissue sample and the matching adjacent noncancerous tissue sample; and (3) each dataset has at least 10 pairs of samples. In the GEO database, only six cancer types have datasets satisfying these criteria. A summary of the 12 datasets, 2 sets for each cancer, is listed in Table 4.1.

2. Identification of differentially expressed genes in cancer *versus* control tissues

For each dataset used in this study, we have used the normalized expression data from the original study. Since we used only paired data, for each dataset a non-parametric

test, Sign Test developed by Frank Wilcoxon (Karas and Savage 1967) for matched pairs, is applied to identify the significant differentially expressed genes in tumor *versus* adjacent normal samples. We consider a gene being differentially expressed if the statistical significance, *p*-value, is less than 0.01.

3. Searching for regulatory relationships in human

To retrieve the transcriptional regulation relationship information about the genes we are interested in this study, we have used a web-based database along with its search engine Cscan (http://www.beaconlab.it/cscan) to predict the common transcription regulators based on a large collection of ChIP-Seq data for several TFs and other factors related to transcription regulation for human and mouse (Zambelli, Prazzoli et al. 2012). It infers regulatory relationships based on ChIP-Seq data collected under 777 different conditions in the hmChip database (Chen, Wu et al. 2011) and transcription factors from the UCSC Genome Browser (Rosenbloom, Dreszer et al. 2012).

4. Cancer related genes

To retrieve cancer related genes, specifically proto-oncogene and tumor suppressor genes for our study, we searched the UNIPROT database (http://www.uniprot.org/keywords/) using keywords, which led to the retrieval of 232 proto-oncogenes (KW-0656) and 194 tumor-suppressor genes (KW-0043) in human.

Concluding Remarks

Based on comparative transcriptomic data analysis results on six cancer types, we have proposed a model of how cancer cells deal with excess protons in both intracellular and extracellular environments. Some of the mechanisms have been reported in the literature but mostly on specific genes or in a fewer cancer types. Our analysis results have confirmed the models previously proposed. In addition we have proposed a new model based on how bacterial *Lactococcus* deals with a similar situation.

As our model is proposed based on transcriptomic data only, further experimental validation on a number of hypotheses are clearly needed, including (i) the main regulators of these processes and their regulatory relationships with pH related regulators, (ii) the new mechanism proposed based on a homologous system in *Lactococcus*, the organism that produces lactates; and (iii) the proposed NBC cotransporter transports in HCO_3^- and Na^+ together but it is not clear how the Na^+ is handled in cancer cells; and similar questions can be asked about the in-transported Ca^{2+} or Na^+ by other deacidification processes. All these require further investigation both experimentally and computationally.

Our overall search procedure for enzymes and transporters that may change the number of protons in a systematic manner proves to be highly effective. For example, the carbonic anhydrases are found to be possibly relevant to the deacidification process from the search, only later we found that this system has been studied and reported in the literature. This result clearly shows the power of this procedure, when coupled with additional searches and analyses of the transcriptomic data, which we believe to be applicable to elucidation of other cancer related processes.

Figures

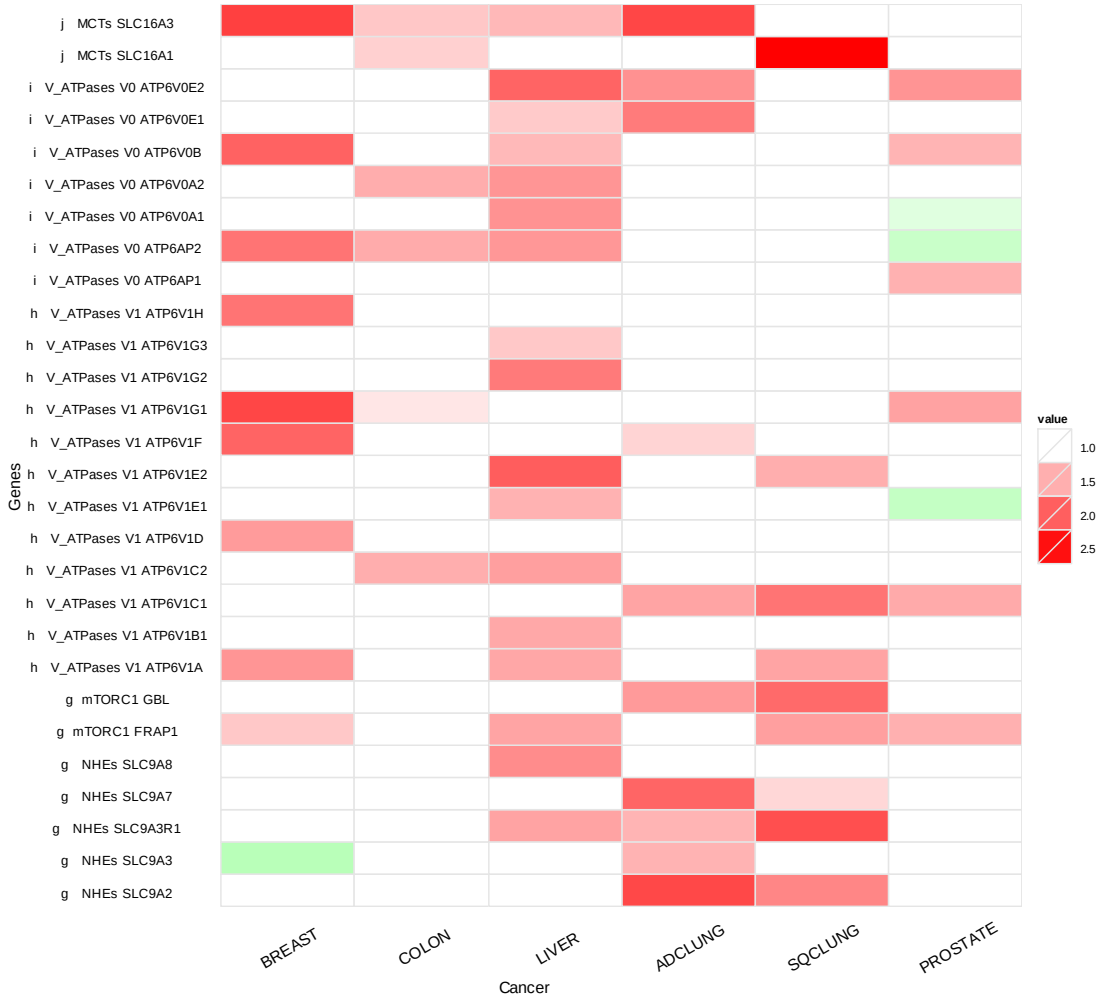


Figure 4.1: Expression level changes of V-ATPase genes in six cancer types in comparison with their matching control tissues. Each entry in the table shows the ratio between a gene's expression levels in cancer and in the matching control, averaged across all the samples (see Methods and Material).

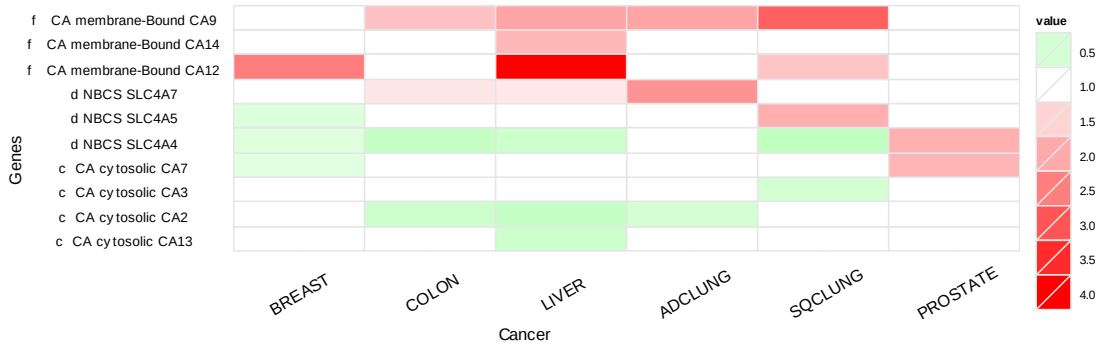


Figure 4.2: Expression level changes of genes involved in carbonic anhydrases (CAs) pH regulation in six cancer tissues in comparison with their matching control tissues.

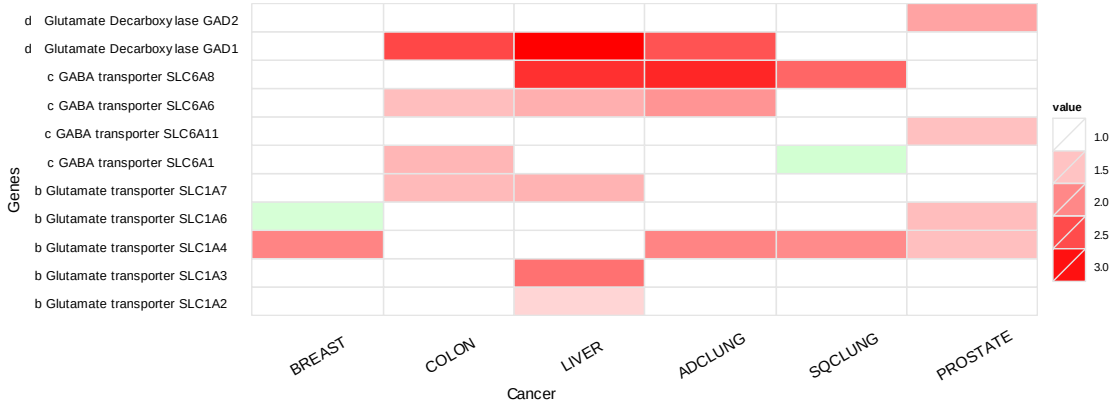


Figure 4.3: Expression level changes of genes involved in the conversion of glutamate to GABA and CO_2, along with the genes encoding the GABA transporters.

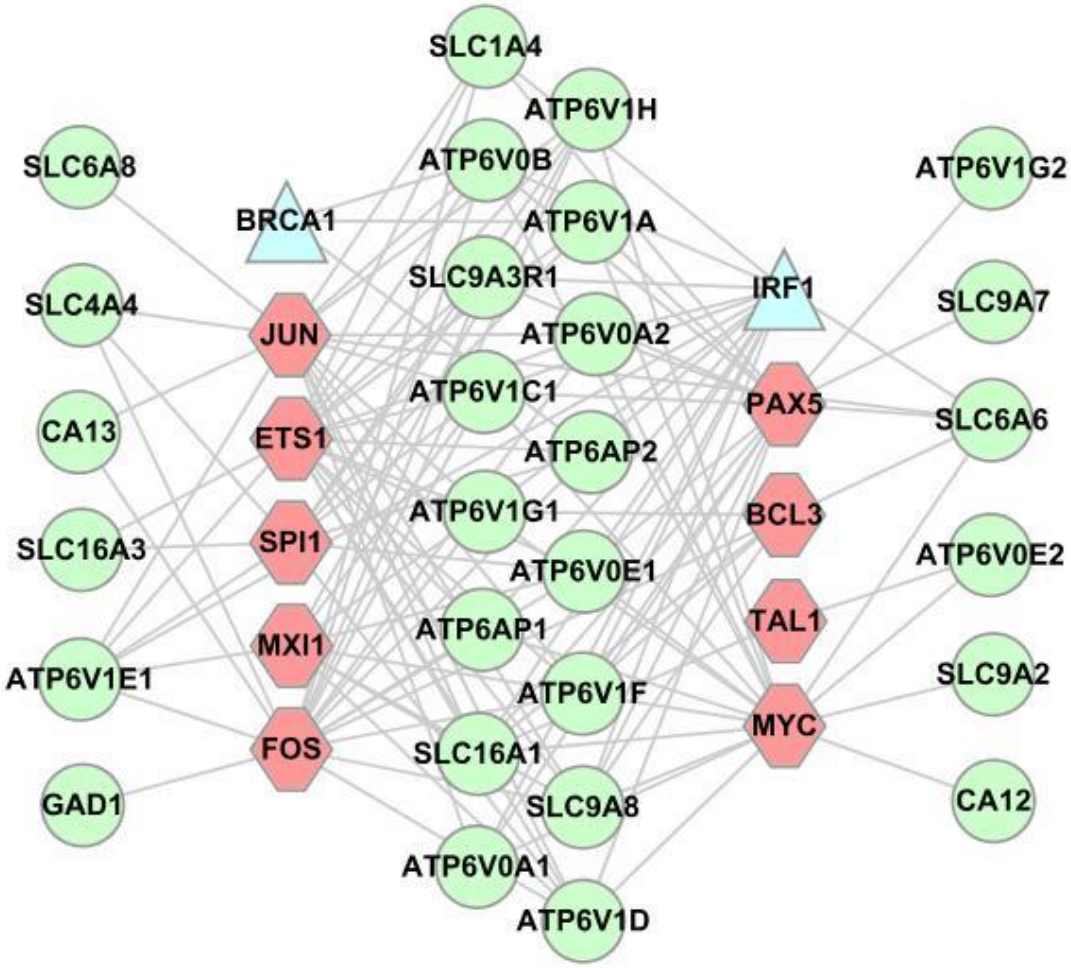


Figure 4.4: Regulatory relationships between genes involved in deacidification and cancer growth. Each circle represents a deacidification related gene, each hexgon represents an oncogene and each triangle a tumor suppressor gene, with each link represents a direct regulatory relationship.

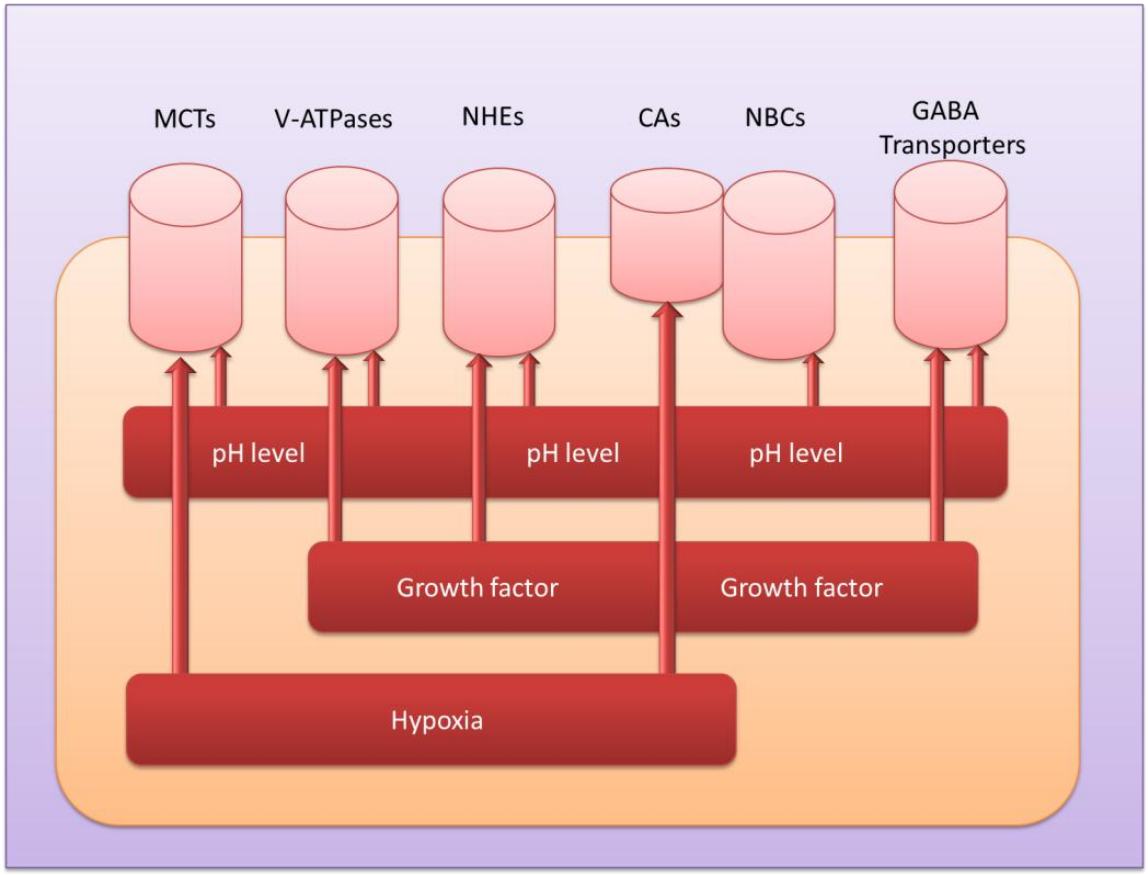


Figure 4.5: A model for deacdification in cancer cells. Each cylinder represents a pump or transporter used to remove protons and possibly other molecules out; and each rectangle bar represents a condition that is a possible regulatory factor for the corresponding pump or transporter.

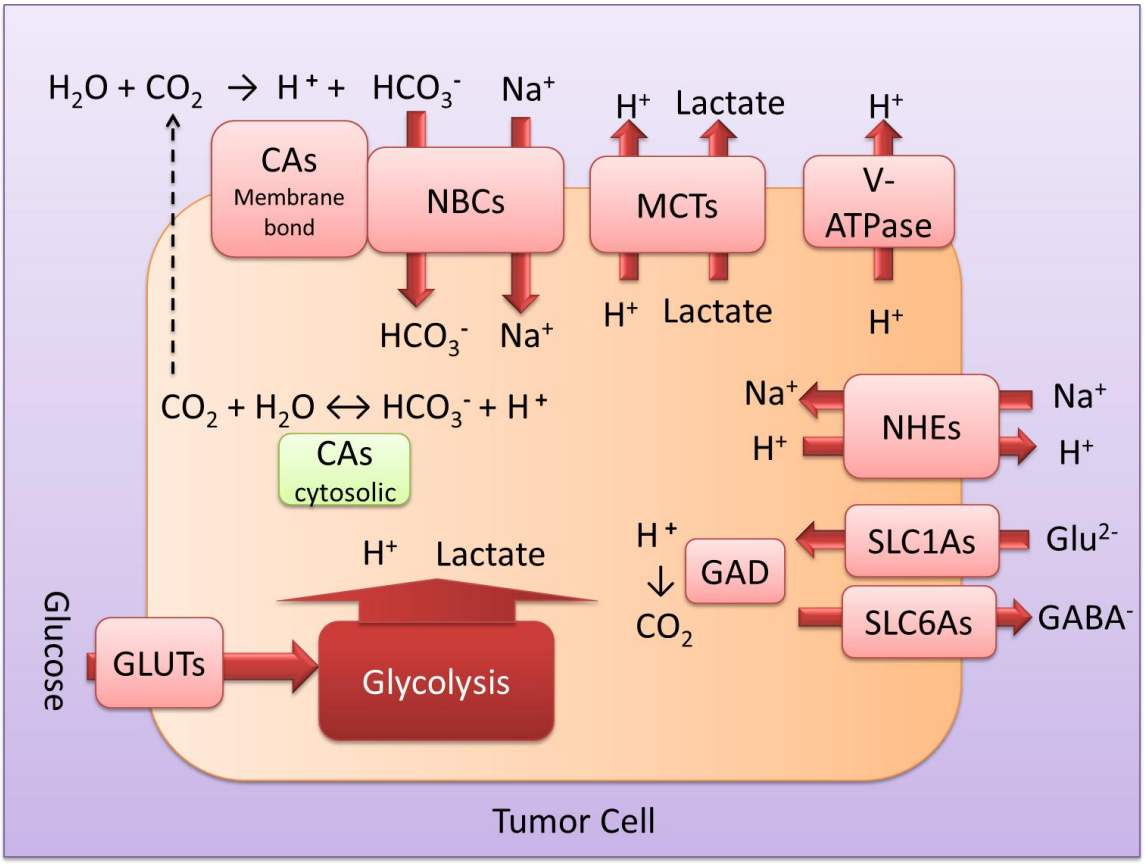


Figure 4.6: Deacdification mechanisms in cancer cells. Each rectangle bar represents a transporter, enzyme or pump family. The red colored rectangles are up-regulated in our study and the green show down-regulation. Dashed arrows indicate CO_2 diffusion across the membrane.

Tables

	set1	set2	pairs
breast cancer	GSE14999	GSE15852	61 / 43
colon cancer	GSE18105	GSE25070	17 / 26
liver cancer	GSE22058	GSE25097	97 / 238
lung adenocarcinoma	GSE31552	GSE7670	31 / 26
lung squamous cell carcinoma	GSE31446	GSE31552	13 / 17
prostate cancer	GSE21034	GSE6608	29 /58

Table 4.1: A summary of the cancer datasets used in our transcriptomic data analysis

CHAPTER 5

DISCUSSION AND PROSPECT

Discussion and Prospect

This systematic study utilizes innovative ideas and cutting-edge computational methods to address biological problems that are related to hot topics in cancer. The idea is to integrate novel computational techniques to find answers to biological questions that are otherwise limited by current experimental techniques. The rapidly increasing pool (Sherlock, Hernandez-Boussard et al. 2001; Barrett, Troup et al. 2007) of large-scale transcriptomic data for various cancer types has provided unprecedented opportunities for computational cancer biologists to study common characteristics across multiple cancer types as well as distinct properties of individual cancer types, which could provide novel insights about different cancer phenotypes at the molecular level.

In this study, I represent a comprehensive study regarding to the issues in the cancer early detection and fundamental cancer biology. For early detection issue, we developed a computational pipeline for the prediction of protein markers in serum for seven cancer types. In addition to individual gene markers, we have focused on gene combinations that can be used to distinguish multiple cancer types and their corresponding reference tissues. The pathway analysis across multiple cancer types has identified a number of abnormally activated or deactivated pathways across multiple cancers and for specific cancers. The information provided on individual genes and pathways, along with

potential serum biomarkers, should provide highly useful information for elucidating pathways in cancer, as well as expediting the search for potential serum biomarkers of specific cancers. The prediction of protein marker is promising as my colleague's urine marker prediction result, generated by a similar prediction process, is validated to be effective with the experimental validation. In addition our computational pipeline is not limited to these seven cancer types biomarkers; it can be applied to different types of cancer as well as any other diseases. Thus, we expect that this will be powerful in aiding biomarker discovery studies.

For fundamental cancer biology issue, we did a systematic analysis on metabolism and followed-up cellular pH regulation topic. Our skin cancer study revealed that both skin cancer types, basal cell carcinoma and melanoma, utilize oxidative phosphorylation as the key energy metabolism in addition to glycolysis and the two skin cancer types are not under hypoxic stress. The boosted energy metabolism in melanoma, powered the up-regulated oxidative phosphorylation with multiple energy resource, along with its high ability to metastasize, become a unique feature and explained why the cancer is so aggressive. Interestingly our predicted metabolic model is examined to inconsistent with Warburg's thesis (Warburg 1956) a former Nobel Prize winning study. However, we are excited to find that according to an independent experimental study (Kluza, Corazao Rozas et al. 2012), a work published on "cancer research" a top journal in cancer research field, we are clear that the melanoma tumors are able to keep functional mitochondrial and used multiple energy resource to compensate from each other as a complementary metabolism to promote cancer cell survival and growth. By this exciting experiment validation, we clearly see the innovation and advantage of the computational result the

experiment design in molecular biology and cancer research. This is not the first study to challenge the generality of the classic Warburg's theory; however we extend the understanding of the metabolic feature of melanoma and hope this finding can potentially inspire the drug development of the melanoma and save patients' life.

Through the microenvironment study, we proposed a model of how cancer cells deal with excess protons in both intracellular and extracellular environments, which are generated due to the reprogrammed energy metabolism. We have proposed a new model based on how bacterial *Lactococcus* deals with a similar situation. Another contribution of the work is that we have proposed possible regulatory mechanisms that allow cancer cells to fully utilize these encoded deacidification mechanisms that are not triggered in normal cells. As homologue based prediction is validated to be effective in many aspects, we are confident our prediction is an innovative way to predict novel oncogenic mechanism in tumor as a key oncogenic feature. The inhibition of pH regulators causes both the pHi and the pHe values to return to normal, with the consequent impairment of tumor growth. This constitutes an novel anti-tumor mechanism for drugs. It has been reported that are few specific, non-toxic compounds, E.g antibodies for CA9 and CA12, that interfere with the pH-regulating proteins that play effective role as anti-cancer drugs (Neri and Supuran 2011). All the key genes involved in tumor deacidification process, including transporters, enzymes and transcription factors, are potentially promising drug target for specific cancer types or even in cancer general.

We believe that our study provides a new and highly effective way to gain new insights and understanding about certain unique characteristics of different cancers in their

formation and progression through mining large-scale gene-expression data across multiple cancer types, but focused on key relevant prospects as cancer early detection and fundamental cancer biology. As the current experiment technology have inevitable limitation on specific issue, the new coming technology especially next generation sequencing will bring much more fruitful information. By continually applying new computational approaches and methodologies, I hope to tackle more biological questions from different angles and new perspectives in cancer research field and hope our research can eventually help doctors to save patients' life.

REFERENCES

Ababneh, M., C. Gotz, et al. (2001). "Downregulation of the cdc2/cyclin B protein kinase activity by binding of p53 to p34(cdc2)." Biochem Biophys Res Commun **283**(2): 507-512.

Akhavan, A., K. H. McHugh, et al. (2006). "Endothelin receptor A blockade enhances taxane effects in prostate cancer." Neoplasia **8**(9): 725-732.

Al Moustafa, A. E., M. A. Alaoui-Jamali, et al. (2002). "Identification of genes associated with head and neck carcinogenesis by cDNA microarray comparison between matched primary normal epithelial and squamous carcinoma cells." Oncogene **21**(17): 2634-2640.

Aota, Y., M. Sumi, et al. (2004). "Type I CD36 deficiency in hematologic disorder." Haematologica **89**(8): EIM17.

Badea, L., V. Herlea, et al. (2008). "Combined gene expression analysis of whole-tissue and microdissected pancreatic ductal adenocarcinoma identifies genes specifically overexpressed in tumor epithelia." Hepatogastroenterology **55**(88): 2016-2027.

Balch, C. M., A. C. Buzaid, et al. (2001). "Final version of the American Joint Committee on Cancer staging system for cutaneous melanoma." J Clin Oncol **19**(16): 3635-3648.

Barrett, T., D. B. Troup, et al. (2007). "NCBI GEO: mining tens of millions of expression profiles--database and tools update." Nucleic Acids Res **35**(Database issue): D760-765.

Bauer, M., J. C. Eickhoff, et al. (2008). "Neutrophil gelatinase-associated lipocalin (NGAL) is a predictor of poor prognosis in human primary breast cancer." Breast Cancer Res Treat **108**(3): 389-397.

Becker, K. G., S. L. White, et al. (2000). "BBID: the biological biochemical image database." Bioinformatics **16**(8): 745-746.

Behrens, P., M. Mathiak, et al. (2003). "Stromal expression of invasion-promoting, matrix-degrading proteases MMP-1 and -9 and the Ets 1 transcription factor in HNPCC carcinomas and sporadic colorectal cancers." Int J Cancer **107**(2): 183-188.

Belbin, T. J., B. Singh, et al. (2002). "Molecular classification of head and neck squamous cell carcinoma using cDNA microarrays." Cancer Res **62**(4): 1184-1190.

Ben-Gal, I., A. Shani, et al. (2005). "Identification of transcription factor binding sites with variable-order Bayesian networks." Bioinformatics **21**(11): 2657-2666.

Bhattacharjee, A., W. G. Richards, et al. (2001). "Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses." Proc Natl Acad Sci U S A **98**(24): 13790-13795.

Bing, C., Y. Bao, et al. (2004). "Zinc-alpha2-glycoprotein, a lipid mobilizing factor, is expressed in adipocytes and is up-regulated in mice with cancer cachexia." Proc Natl Acad Sci U S A **101**(8): 2500-2505.

Boeckmann, B., A. Bairoch, et al. (2003). "The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003." Nucleic Acids Res **31**(1): 365-370.

Boidot, R., F. Vegran, et al. (2012). "Regulation of monocarboxylate transporter MCT1 expression by p53 mediates inward and outward lactate fluxes in tumors." Cancer Res **72**(4): 939-948.

Box, N. F., D. L. Duffy, et al. (2001). "MC1R genotype modifies risk of melanoma in families segregating CDKN2A mutations." Am J Hum Genet **69**(4): 765-773.

Bozikov, K. and I. Taggart (2006). "Metastatic basal cell carcinoma: is infiltrative/morpheaform subtype a risk factor?" Eur J Dermatol **16**(6): 691-692.

Bridges, J. P., S. E. Wert, et al. (2003). "Expression of a human surfactant protein C mutation associated with interstitial lung disease disrupts lung development in transgenic mice." J Biol Chem **278**(52): 52739-52746.

Broutet, N., M. Plebani, et al. (2003). "Pepsinogen A, pepsinogen C, and gastrin as markers of atrophic chronic gastritis in European dyspeptics." Br J Cancer **88**(8): 1239-1247.

CancerFact (2006). www.cancer.org. A. C. Society.

Catalona, W. J., D. S. Smith, et al. (1995). "Evaluation of percentage of free serum prostate-specific antigen to improve specificity of prostate cancer screening." JAMA **274**(15): 1214-1220.

Chan, J. M., M. J. Stampfer, et al. (2002). "Insulin-like growth factor-I (IGF-I) and IGF binding protein-3 as predictors of advanced-stage prostate cancer." J Natl Cancer Inst **94**(14): 1099-1106.

Chandran, U. R., R. Dhir, et al. (2005). "Differences in gene expression in prostate cancer, normal appearing prostate tissue adjacent to cancer and prostate tissue from cancer free organ donors." BMC Cancer **5**: 45.

Chang, C.-c. and C.-J. Lin (2001). LIBSVM: a Library for Support Vector Machines.

Chekerov, R., I. Klaman, et al. (2006). "Altered expression pattern of topoisomerase IIalpha in ovarian tumor epithelial and stromal cells after platinum-based chemotherapy." Neoplasia **8**(1): 38-45.

Chen, L., G. Wu, et al. (2011). "hmChIP: a database and web server for exploring publicly available human and mouse ChIP-seq and ChIP-chip data." Bioinformatics **27**(10): 1447-1448.

Chen, S., Y. Xu, et al. (2006). "Androgen receptor phosphorylation and stabilization in prostate cancer by cyclin-dependent kinase 1." Proc Natl Acad Sci U S A **103**(43): 15969-15974.

Chen, W. B., W. Lenschow, et al. (2002). "Smad4/DPC4-dependent regulation of biglycan gene expression by transforming growth factor-beta in pancreatic tumor cells." J Biol Chem **277**(39): 36118-36128.

Chen, X., S. T. Cheung, et al. (2002). "Gene expression patterns in human liver cancers." Molecular biology of the cell **13**(6): 1929-1939.

Chen, X., S. Y. Leung, et al. (2003). "Variation in gene expression patterns in human gastric cancers." Mol Biol Cell **14**(8): 3208-3215.

Chiche, J., M. C. Brahimi-Horn, et al. (2010). "Tumour hypoxia induces a metabolic shift causing acidosis: a common feature in cancer." J Cell Mol Med **14**(4): 771-794.

Cotter, P. D. and C. Hill (2003). "Surviving the acid test: responses of gram-positive bacteria to low pH." Microbiol Mol Biol Rev **67**(3): 429-453, table of contents.

Cui, J., Y. Chen, et al. (2011). "An integrated transcriptomic and computational analysis for biomarker identification in gastric cancer." Nucleic Acids Res **39**(4): 1197-1207.

Cui, J., F. Li, et al. (2011). "Gene-expression signatures can distinguish gastric cancer grades and stages." PLoS One **6**(3): e17819.

Cui, J., Q. Liu, et al. (2008). "Computational prediction of human proteins that can be secreted into the bloodstream." Bioinformatics **24**(20): 2370-2375.

D'Errico, M., E. de Rinaldis, et al. (2009). "Genome-wide expression profile of sporadic gastric cancers with microsatellite instability." Eur J Cancer **45**(3): 461-469.

DeBerardinis, R. J., J. J. Lum, et al. (2008). "The biology of cancer: metabolic reprogramming fuels cell growth and proliferation." Cell Metab **7**(1): 11-20.

Dennis, G., Jr., B. T. Sherman, et al. (2003). "DAVID: Database for Annotation, Visualization, and Integrated Discovery." Genome Biol **4**(5): P3.

Dhanasekaran, S. M., T. R. Barrette, et al. (2001). "Delineation of prognostic biomarkers in prostate cancer." Nature **412**(6849): 822-826.

Diamandis, E. P. (2004). "Analysis of serum proteomic patterns for early cancer diagnosis: drawing attention to potential problems." J Natl Cancer Inst **96**(5): 353-356.

Dunham, W. (2007). "Report sees 7.6 million global 2007 cancer deaths." Reuters, New York.

Edgar, R., M. Domrachev, et al. (2002). "Gene Expression Omnibus: NCBI gene expression and hybridization array data repository." Nucleic Acids Res **30**(1): 207-210.

Eguchi, T., T. Takaki, et al. (2007). "RB silencing compromises the DNA damage-induced G2/M checkpoint and causes deregulated expression of the ECT2 oncogene." Oncogene **26**(4): 509-520.

Fais, S., A. De Milito, et al. (2007). "Targeting vacuolar H+-ATPases as a new strategy against cancer." Cancer Res **67**(22): 10627-10630.

Fang, J. S., R. D. Gillies, et al. (2008). "Adaptation to hypoxia and acidosis in carcinogenesis and tumor progression." Semin Cancer Biol **18**(5): 330-337.

Fellenz, M. P. and L. E. Gerweck (1988). "Influence of extracellular pH on intracellular pH and cell energy status: relationship to hyperthermic sensitivity." Radiat Res **116**(2): 305-312.

Feron, O. (2009). "Pyruvate into lactate and back: from the Warburg effect to symbiotic energy fuel exchange in cancer cells." Radiother Oncol **92**(3): 329-333.

Ford, D., D. F. Easton, et al. (1998). "Genetic heterogeneity and penetrance analysis of the BRCA1 and BRCA2 genes in breast cancer families. The Breast Cancer Linkage Consortium." Am J Hum Genet **62**(3): 676-689.

Fu, L. and C. C. Lee (2003). "The circadian clock: pacemaker and tumour suppressor." Nat Rev Cancer **3**(5): 350-361.

Garber, M. E., O. G. Troyanskaya, et al. (2001). "Diversity of gene expression in adenocarcinoma of the lung." Proc Natl Acad Sci U S A **98**(24): 13784-13789.

Garcia-Caballero, M., E. Neugebauer, et al. (1988). "Increased histidine decarboxylase (HDC) activity in human colorectal cancer: results of a study on ten patients." Agents Actions **23**(3-4): 357-360.

Gatenby, R. A. and R. J. Gillies (2004). "Why do cancers have high aerobic glycolysis?" Nat Rev Cancer **4**(11): 891-899.

Groundwater, P., S. A. Beck, et al. (1990). "Alteration of serum and urinary lipolytic activity with weight loss in cachectic cancer patients." Br J Cancer **62**(5): 816-821.

Guilford, P., J. Hopkins, et al. (1998). "E-cadherin germline mutations in familial gastric cancer." Nature **392**(6674): 402-405.

Guo, Q. M. (2003). "DNA microarray and cancer." Current opinion in oncology **15**(1): 36-43.

Halestrap, A. P. (2012). "The monocarboxylate transporter family--Structure and functional characterization." IUBMB Life **64**(1): 1-9.

Halestrap, A. P. and D. Meredith (2004). "The SLC16 gene family-from monocarboxylate transporters (MCTs) to aromatic amino acid transporters and beyond." Pflugers Arch **447**(5): 619-628.

Halestrap, A. P. and N. T. Price (1999). "The proton-linked monocarboxylate transporter (MCT) family: structure, function and regulation." Biochem J **343 Pt 2**: 281-299.

Han, H., D. J. Bearss, et al. (2002). "Identification of differentially expressed genes in pancreatic cancer cells using cDNA microarray." Cancer research **62**(10): 2890-2896.

Hanahan, D. and R. A. Weinberg (2000). "The hallmarks of cancer." Cell **100**(1): 57-70.

Hanahan, D. and R. A. Weinberg (2011). "Hallmarks of cancer: the next generation." Cell **144**(5): 646-674.

Hart, T. C., M. C. Gorry, et al. (2002). "Mutations of the UMOD gene are responsible for medullary cystic kidney disease 2 and familial juvenile hyperuricaemic nephropathy." J Med Genet **39**(12): 882-892.

Hedenfalk, I. A., M. Ringner, et al. (2002). "Gene expression in inherited breast cancer." Advances in cancer research **84**: 1-34.

Herb, F., T. Thye, et al. (2008). "ALOX5 variants associated with susceptibility to human pulmonary tuberculosis." Hum Mol Genet **17**(7): 1052-1060.

Hernandez, A. (2012). "Proton dynamics in cancer." Curr Pharm Des **18**(10): 1317-1318.

Hippo, Y., H. Taniguchi, et al. (2002). "Global gene expression analysis of gastric cancer by oligonucleotide microarrays." Cancer research **62**(1): 233-240.

Ho, C. C., P. H. Huang, et al. (2002). "Up-regulated caveolin-1 accentuates the metastasis capability of lung adenocarcinoma by inducing filopodia formation." Am J Pathol **161**(5): 1647-1656.

Hong, C. S., J. Cui, et al. (2011). "A computational method for prediction of excretory proteins and application to identification of gastric cancer markers in urine." PLoS One **6**(2): e16875.

Hughes-Davies, T. H. (1998). "CDKN2A mutations in multiple primary melanomas." N Engl J Med **339**(5): 347-348.

Hyde, T. M., B. K. Lipska, et al. (2011). "Expression of GABA signaling molecules KCC2, NKCC1, and GAD1 in cortical development and schizophrenia." J Neurosci **31**(30): 11088-11095.

Ikewaki, N., H. Tamauchi, et al. (2007). "Decrease in CD93 (C1qRp) expression in a human monocyte-like cell line (U937) treated with various apoptosis-inducing chemical substances." Microbiol Immunol **51**(12): 1189-1200.

Iwasaki, J. K., D. Srivastava, et al. (2010). "The molecular genetics underlying basal cell carcinoma pathogenesis and links to targeted therapeutics." J Am Acad Dermatol.

Jain, N., J. Thatte, et al. (2003). "Local-pooled-error test for identifying differentially expressed genes with a small number of replicated microarrays." Bioinformatics **19**(15): 1945-1951.

Jemal, A., R. Siegel, et al. (2010). "Cancer statistics, 2010." CA Cancer J Clin **60**(5): 277-300.

Jerant, A. F., J. T. Johnson, et al. (2000). "Early detection and treatment of skin cancer." Am Fam Physician **62**(2): 357-368, 375-356, 381-352.

Jiang, W. G., A. Douglas-Jones, et al. (2003). "Expression of peroxisome-proliferator activated receptor-gamma (PPARgamma) and the PPARgamma co-activator, PGC-1, in human breast cancer correlates with clinical outcomes." Int J Cancer **106**(5): 752-757.

Jiang, X., J. Tan, et al. (2008). "DACT3 is an epigenetic regulator of Wnt/beta-catenin signaling in colorectal cancer and is a therapeutic target of histone modifications." Cancer Cell **13**(6): 529-541.

Johnson, D. E. C., J. R. (2009). Bicarbonate Transport Metabolons. Drug Design of Zinc-Enzyme Inhibitors: Functional, Structural, and Disease Applications. C. T. W. Supuran, J. Y. Hoboken,New Jersey, Wiley**:** 415–437.

Jones, J., H. Otu, et al. (2005). "Gene signatures of progression and metastasis in renal cell cancer." Clin Cancer Res **11**(16): 5730-5739.

Kaila, K. (1994). "Ionic basis of GABAA receptor channel function in the nervous system." Prog Neurobiol **42**(4): 489-537.

Kane, R., C. Godson, et al. (2008). "Chordin-like 1, a bone morphogenetic protein-4 antagonist, is upregulated by hypoxia in human retinal pericytes and plays a role in regulating angiogenesis." Mol Vis **14**: 1138-1148.

Kanehisa, M. and S. Goto (2000). "KEGG: kyoto encyclopedia of genes and genomes." Nucleic Acids Res **28**(1): 27-30.

Karas, J. and R. Savage (1967). "Publications of Frank Wilcoxon (1892-1965)." Biometrics **23**(1): 1-10.

Kholodnyuk, I. D., S. Kozireva, et al. (2006). "Down regulation of 3p genes, LTF, SLC38A3 and DRR1, upon growth of human chromosome 3-mouse fibrosarcoma hybrids in severe combined immunodeficiency mice." Int J Cancer **119**(1): 99-107.

Ki, D. H., H. C. Jeung, et al. (2007). "Whole genome analysis for liver metastasis gene signatures in colorectal cancer." Int J Cancer **121**(9): 2005-2012.

Kluza, J., P. Corazao Rozas, et al. (2012). "INACTIVATION OF THE HIF-1alpha/PDK3 SIGNALING AXIS DRIVES MELANOMA TOWARD MITOCHONDRIAL OXIDATIVE METABOLISM AND POTENTIATES THE THERAPEUTIC ACTIVITY OF PRO-OXIDANTS." Cancer Res.

Klymkowsky, M. W. and P. Savagner (2009). "Epithelial-mesenchymal transition: a cancer researcher's conceptual friend and foe." Am J Pathol **174**(5): 1588-1593.

Kolesar, J., W. Huang, et al. (2009). "Evaluation of mRNA by Q-RTPCR and protein expression by AQUA of the M2 subunit of ribonucleotide reductase (RRM2) in human tumors." Cancer Chemother Pharmacol **64**(1): 79-86.

Koren, R., L. Rath-Wolfson, et al. (2004). "Prognostic value of Topoisomerase II in female breast cancer." Oncol Rep **12**(4): 915-919.

Kosari, F., Y. W. Asmann, et al. (2002). "Cysteine-rich secretory protein-3: a potential biomarker for prostate cancer." Cancer Epidemiol Biomarkers Prev **11**(11): 1419-1426.

Landi, M. T., T. Dracheva, et al. (2008). "Gene expression signature of cigarette smoking and its role in lung adenocarcinoma development and survival." PLoS One **3**(2): e1651.

Lapointe, J., C. Li, et al. (2004). "Gene expression profiling identifies clinically relevant subtypes of prostate cancer." Proc Natl Acad Sci U S A **101**(3): 811-816.

Le Jan, S., C. Amy, et al. (2003). "Angiopoietin-like 4 is a proangiogenic factor produced during ischemia and in conventional renal cell carcinoma." Am J Pathol **162**(5): 1521-1528.

Lee, H. J., E. K. Lee, et al. (2006). "Ectopic expression of neutrophil gelatinase-associated lipocalin suppresses the invasion and liver metastasis of colon cancer cells." Int J Cancer **118**(10): 2490-2497.

Lee, H. Y., M. K. Kim, et al. (2005). "Serum amyloid A stimulates matrix-metalloproteinase-9 upregulation via formyl peptide receptor like-1-mediated signaling in human monocytic cells." Biochem Biophys Res Commun **330**(3): 989-998.

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

Lewis, M. and P. D. Howdle (2003). "The neurology of liver failure." QJM **96**(9): 623-633.

Li, H., Q. Lu, et al. (2007). "[Expression of fatty acid binding protein in human breast cancer tissues]." Xi Bao Yu Fen Zi Mian Yi Xue Za Zhi **23**(4): 312-316.

Li, J. J. (2006). Laughing gas, Viagra, and Lipitor : the human stories behind the drugs we use. Oxford ; New York, Oxford University Press.

Lilja, H., D. Ulmert, et al. (2008). "Prostate-specific antigen and prostate cancer: prediction, detection and monitoring." Nat Rev Cancer **8**(4): 268-278.

Lo, B. K., M. Yu, et al. (2010). "CXCR3/ligands are significantly involved in the tumorigenesis of basal cell carcinomas." Am J Pathol **176**(5): 2435-2446.

Luo, J., T. Dunn, et al. (2002). "Gene expression signature of benign prostatic hyperplasia revealed by cDNA microarray analysis." The Prostate **51**(3): 189-200.

Masuda, T. A., H. Inoue, et al. (2003). "Cyclin-dependent kinase 1 gene expression is associated with poor prognosis in gastric carcinoma." Clin Cancer Res **9**(15): 5693-5698.

Matei, D., T. G. Graeber, et al. (2002). "Gene expression in epithelial ovarian carcinoma." Oncogene **21**(41): 6289-6298.

Mechtcheriakova, D., A. Wlachos, et al. (2007). "Sphingosine 1-phosphate phosphatase 2 is induced during inflammatory responses." Cell Signal **19**(4): 748-760.

Miyoshi, Y., T. Funahashi, et al. (2003). "Association of serum adiponectin levels with breast cancer risk." Clin Cancer Res **9**(15): 5699-5704.

Morales, A., A. Cuellar, et al. (1999). "Synthesis of steroids in pancreas: evidence of cytochrome P-450scc activity." Pancreas **19**(1): 39-44.

Mulla, C. M., E. Geras-Raaka, et al. (2009). "High levels of thyrotropin-releasing hormone receptors activate programmed cell death in human pancreatic precursors." Pancreas **38**(2): 197-202.

Neri, D. and C. T. Supuran (2011). "Interfering with pH regulation in tumours as a therapeutic strategy." Nat Rev Drug Discov **10**(10): 767-777.

Ning, P. F., L. P. Sun, et al. (2004). "[Expression of pepsinogen C in gastric cancer and its precursor]." Zhonghua Yi Xue Za Zhi **84**(10): 818-821.

Nozoe, T., M. Honda, et al. (2003). "p34cdc2 expression is an independent indicator for lymph node metastasis in colorectal carcinoma." J Cancer Res Clin Oncol **129**(9): 498-502.

Ohlsson, G., J. M. Moreira, et al. (2005). "Loss of expression of the adipocyte-type fatty acid-binding protein (A-FABP) is associated with progression of human urothelial carcinomas." Mol Cell Proteomics **4**(4): 570-581.

Owens, D. F. and A. R. Kriegstein (2002). "Is there more to GABA than synaptic inhibition?" Nat Rev Neurosci **3**(9): 715-727.

Padilla-Lopez, S. and D. A. Pearce (2006). "Saccharomyces cerevisiae lacking Btn1p modulate vacuolar ATPase activity to regulate pH imbalance in the vacuole." J Biol Chem **281**(15): 10273-10280.

Pan, Y. H. and B. J. Bahnson (2007). "Structural basis for bile salt inhibition of pancreatic phospholipase A2." J Mol Biol **369**(2): 439-450.

Park, S. S., J. E. Kim, et al. (2005). "Caveolin-1 is down-regulated and inversely correlated with HER2 and EGFR expression status in invasive ductal carcinoma of the breast." Histopathology **47**(6): 625-630.

Parr, C., G. Watkins, et al. (2004). "The hepatocyte growth factor regulatory factors in human breast cancer." Clin Cancer Res **10**(1 Pt 1): 202-211.

Pau Ni, I. B., Z. Zakaria, et al. (2010). "Gene expression patterns distinguish breast carcinomas from normal breast tissues: The Malaysian context." Pathol Res Pract.

Pedraza, V., J. A. Gomez-Capilla, et al. (2010). "Gene expression signatures in breast cancer distinguish phenotype characteristics, histologic subtypes, and tumor invasiveness." Cancer **116**(2): 486-496.

Pei, H., L. Li, et al. (2009). "FKBP51 affects cancer cell response to chemotherapy by negatively regulating Akt." Cancer Cell **16**(3): 259-266.

Pena-Llopis, S., S. Vega-Rubin-de-Celis, et al. (2011). "Regulation of TFEB and V-ATPases by mTORC1." EMBO J **30**(16): 3242-3258.

Perez-Sayans, M., J. M. Somoza-Martin, et al. (2009). "V-ATPase inhibitors and implication in cancer treatment." Cancer Treat Rev **35**(8): 707-713.

Piovan, E., V. Tosello, et al. (2005). "Chemokine receptor expression in EBV-associated lymphoproliferation in hu/SCID mice: implications for CXCL12/CXCR4 axis in lymphoma generation." Blood **105**(3): 931-939.

Polsky, D. and C. Cordon-Cardo (2003). "Oncogenes in melanoma." Oncogene **22**(20): 3087-3091.

Polsky, D. and S. Q. Wang. (2011). "Skin Cancer Facts." Retrieved 3/31, 2011, from www.skincancer.org.

Polyak, K. and R. A. Weinberg (2009). "Transitions between epithelial and mesenchymal states: acquisition of malignant and stem cell traits." Nat Rev Cancer **9**(4): 265-273.

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

Proescholdt, M. A., C. Mayer, et al. (2005). "Expression of hypoxia-inducible carbonic anhydrases in brain tumors." Neuro Oncol **7**(4): 465-475.

Rai, A. J. and D. W. Chan (2004). "Cancer proteomics: Serum diagnostics for tumor marker discovery." Ann N Y Acad Sci **1022**: 286-294.

Ries LAG, M. D., Krapcho M. (2008). "SEER Cancer Statistics Review, 1975-2005." from http://seer.cancer.gov/csr/1975_2005/.

Rogers, H. W., M. A. Weinstock, et al. (2010). "Incidence estimate of nonmelanoma skin cancer in the United States, 2006." Arch Dermatol **146**(3): 283-287.

Rosenbloom, K. R., T. R. Dreszer, et al. (2012). "ENCODE whole-genome data in the UCSC Genome Browser: update 2012." Nucleic Acids Res **40**(Database issue): D912-917.

Scatolini, M., M. M. Grand, et al. (2010). "Altered molecular pathways in melanocytic lesions." Int J Cancer **126**(8): 1869-1881.

Sennoune, S. R., K. Bakunts, et al. (2004). "Vacuolar H+-ATPase in human breast cancer cells with distinct metastatic potential: distribution and functional activity." Am J Physiol Cell Physiol **286**(6): C1443-1452.

Sennoune, S. R., D. Luo, et al. (2004). "Plasmalemmal vacuolar-type H+-ATPase in cancer biology." Cell Biochem Biophys **40**(2): 185-206.

Sherlock, G., T. Hernandez-Boussard, et al. (2001). "The Stanford Microarray Database." Nucleic Acids Res **29**(1): 152-155.

Shimada, S., K. Yamaguchi, et al. (2002). "Pancreatic elastase IIIA and its variants are expressed in pancreatic carcinoma cells." Int J Mol Med **10**(5): 599-603.

Simon, R. (2003). "Diagnostic and prognostic prediction using gene expression profiles in high-dimensional microarray data." Br J Cancer **89**(9): 1599-1604.

Sloan, E. K., K. L. Stanley, et al. (2004). "Caveolin-1 inhibits breast cancer growth and metastasis." Oncogene **23**(47): 7893-7897.

Someya, S., T. Yamasoba, et al. (2008). "The role of mtDNA mutations in the pathogenesis of age-related hearing loss in mice carrying a mutator DNA polymerase gamma." Neurobiol Aging **29**(7): 1080-1092.

Sommer, A. and B. Haendler (2003). "Androgen receptor and prostate cancer: molecular aspects and gene expression profiling." Curr Opin Drug Discov Devel **6**(5): 702-711.

Sonveaux, P., F. Vegran, et al. (2008). "Targeting lactate-fueled respiration selectively kills hypoxic tumor cells in mice." J Clin Invest **118**(12): 3930-3942.

Soulitzis, N., I. Karyotis, et al. (2006). "Expression analysis of peptide growth factors VEGF, FGF2, TGFB1, EGF and IGF1 in prostate cancer and benign prostatic hyperplasia." Int J Oncol **29**(2): 305-314.

Su, L. J., C. W. Chang, et al. (2007). "Selection of DDX5 as a novel internal control for Q-RT-PCR from microarray data using a block bootstrap re-sampling scheme." BMC Genomics **8**: 140.

Swietach, P., R. D. Vaughan-Jones, et al. (2007). "Regulation of tumor pH and the role of carbonic anhydrase 9." Cancer Metastasis Rev **26**(2): 299-310.

Swietach, P., S. Wigfield, et al. (2008). "Tumor-associated carbonic anhydrase 9 spatially coordinates intracellular pH in three-dimensional multicellular growths." J Biol Chem **283**(29): 20473-20483.

Swietach, P., S. Wigfield, et al. (2008). "Cancer-associated, hypoxia-inducible carbonic anhydrase IX facilitates CO2 diffusion." BJU Int **101 Suppl 4**: 22-24.

Tang, L., D. L. Dai, et al. (2006). "Aberrant expression of collagen triple helix repeat containing 1 in human solid cancers." Clin Cancer Res **12**(12): 3716-3722.

Tong, Z., A. B. Kunnumakkara, et al. (2008). "Neutrophil gelatinase-associated lipocalin: a novel suppressor of invasion and angiogenesis in pancreatic cancer." Cancer Res **68**(15): 6100-6108.

Tonin, P. N., T. J. Hudson, et al. (2001). "Microarray analysis of gene expression mirrors the biology of an ovarian cancer model." Oncogene **20**(45): 6617-6626.

Ulazzi, L., S. Sabbioni, et al. (2007). "Nidogen 1 and 2 gene promoters are aberrantly methylated in human gastrointestinal cancer." Mol Cancer **6**: 17.

van der Horst, G. T., M. Muijtjens, et al. (1999). "Mammalian Cry1 and Cry2 are essential for maintenance of circadian rhythms." Nature **398**(6728): 627-630.

Varadhachary, A., J. S. Wolf, et al. (2004). "Oral lactoferrin inhibits growth of established tumors and potentiates conventional chemotherapy." Int J Cancer **111**(3): 398-403.

Varis, A., A. Zaika, et al. (2004). "Coamplified and overexpressed genes at ERBB2 locus in gastric cancer." Int J Cancer **109**(4): 548-553.

Wang, C. S., K. H. Lin, et al. (2004). "Overexpression of SPARC gene in human gastric carcinoma and its clinic-pathologic significance." Br J Cancer **91**(11): 1924-1930.

Wang, S., M. G. Hasham, et al. (2003). "Upregulation of Cdc2 and cyclin A during apoptosis of endothelial cells induced by cleaved high-molecular-weight kininogen." Am J Physiol Heart Circ Physiol **284**(6): H1917-1923.

Wang, W. S., P. M. Chen, et al. (2006). "Matrix metalloproteinase-7 increases resistance to Fas-mediated apoptosis and is a poor prognostic factor of patients with colorectal carcinoma." Carcinogenesis **27**(5): 1113-1120.

Wang, Y. Y., S. X. Wu, et al. (2003). "Effects of c-fos antisense oligodeoxynucleotide on 5-HT-induced upregulation of preprodynorphin, preproenkephalin, and glutamic acid decarboxylase mRNA expression in cultured rat spinal dorsal horn neurons." Biochem Biophys Res Commun **309**(3): 631-636.

Warburg, O. (1956). "On the origin of cancer cells." Science **123**(3191): 309-314.

Webster, K. A., D. J. Discher, et al. (1999). "Hypoxia-activated apoptosis of cardiac myocytes requires reoxygenation or a pH shift and is independent of p53." J Clin Invest **104**(3): 239-252.

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

Wei, D., W. Gong, et al. (2005). "Drastic down-regulation of Kruppel-like factor 4 expression is critical in human gastric cancer development and progression." Cancer Res **65**(7): 2746-2754.

White, H. L. and T. L. Sato (1978). "GABA-transaminases of human brain and peripheral tissues--kinetic and molecular properties." J Neurochem **31**(1): 41-47.

WHO (2006). *Cancer*, *World Health Organization*.

Wilcoxin, F. (1947). "Probability tables for individual comparisons by ranking methods." Biometrics **3**(3): 119-122.

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.

Wu, Y., K. McRoberts, et al. (2007). "Neuromedin U is regulated by the metastasis suppressor RhoGDI2 and is a novel promoter of tumor formation, lung metastasis and cancer cachexia." Oncogene **26**(5): 765-773.

Wykoff, C. C., N. J. Beasley, et al. (2000). "Hypoxia-inducible expression of tumor-associated carbonic anhydrases." Cancer Res **60**(24): 7075-7083.

Xu, K., J. Cui, et al. (2010). "A comparative analysis of gene-expression data of multiple cancer types." PLoS One **5**(10): e13696.

Yamada, J., A. Okabe, et al. (2004). "Cl- uptake promoting depolarizing GABA actions in immature rat neocortical neurones is mediated by NKCC1." J Physiol **557**(Pt 3): 829-841.

Yanagisawa, K., B. J. Xu, et al. (2003). "Molecular fingerprinting in human lung cancer." Clin Lung Cancer **5**(2): 113-118.

Zambelli, F., G. M. Prazzoli, et al. (2012). "Cscan: finding common regulators of a set of genes by using a collection of genome-wide ChIP-seq datasets." Nucleic Acids Res **40**(Web Server issue): W510-515.

Zhang, J., X. Jin, et al. (2005). "The functional polymorphism in the matrix metalloproteinase-7 promoter increases susceptibility to esophageal squamous

cell carcinoma, gastric cardiac adenocarcinoma and non-small cell lung carcinoma." Carcinogenesis **26**(10): 1748-1753.

Zuo, L., J. Weger, et al. (1996). "Germline mutations in the p16INK4a binding domain of CDK4 in familial melanoma." Nat Genet **12**(1): 97-99.

APPENDICES

Appendix Tables

Appendix Table A2.1.1 The detailed list of 100 *k*-gene 100 combinations for **Figure 2.1** (a)(b) breast cancer

http://csbl.bmb.uga.edu/publications/materials/kunxu/PAPER_1/BREAST_CANCER/breast_tn_blood_1to4.htm

Appendix Table A2.1.2. The detailed list of 100 *k*-gene combinations for **Figure 2.1** (c) (d) early stage breast cancer

http://csbl.bmb.uga.edu/publications/materials/kunxu/PAPER_1/BREAST_CANCER/BREAST_STAGE_1_TO_4.htm

Appendix Table A2.2. The detailed list of 100 *k*-gene combinations for **Figure 2.2** (a)(b) colon cancer

http://csbl.bmb.uga.edu/publications/materials/kunxu/PAPER_1/COLON_CANCER/colon_1to4_marker.htm

Appendix Table A2.3. The detailed list of 100 *k*-gene combinations for **Figure 2.3** (a)(b) kidney cancer

http://csbl.bmb.uga.edu/publications/materials/kunxu/PAPER_1/KIDNEY_CANCER/kidney_1to4_chart.htm

Appendix Table A2.4. The detailed list of 100 *k*-gene combinations for **Figure 2.4** (a)(b) lung cancer

http://csbl.bmb.uga.edu/publications/materials/kunxu/PAPER_1/LUNG_CANCER/LUNG_CANCER_1to4.htm

Appendix Table A2.5. The detailed list of 100 *k*-gene combinations for **Figure 2.5** (a)(b) pancreatic cancer

http://csbl.bmb.uga.edu/publications/materials/kunxu/PAPER_1/PANCREATIC_CANCER/PANCREASE_CANCER_1TO4xlsx.htm

Appendix Table A2.6. The detailed list of 100 *k*-gene combinations for **Figure 2.6** (a)(b) prostate cancer

http://csbl.bmb.uga.edu/publications/materials/kunxu/PAPER_1/PROSTATE_CANCER/PROSTATE_1to4_MARKER_CHART.htm

Appendix Table A2.7.1. The detailed list of 100 *k*-gene combinations for **Figure 2.7** (a)(b) stomach cancer

http://csbl.bmb.uga.edu/publications/materials/kunxu/PAPER_1/STOMACH_CANCER/STOMACH_1_to_4_CHART.htm

Appendix Table A2.7.2. The detailed list of 100 *k*-gene combinations for **Figure 2.7** (c)(d) early stage stomach cancer

http://csbl.bmb.uga.edu/publications/materials/kunxu/PAPER_1/STOMACH_CANCER/stamach_stage_1to4.htm

Appendix Table A3.1. Differentially expressed genes in skin cancer.
http://csbl.bmb.uga.edu/publications/materials/kunxu/PAPER_2/table_1.htm

Appendix Table A3.2. The top signatures for the melanoma and BCC.
http://csbl.bmb.uga.edu/publications/materials/kunxu/PAPER_2/table_2.htm

Appendix Table A3.3. The Common Gene Shared by the Datasets
http://csbl.bmb.uga.edu/publications/materials/kunxu/PAPER_2/table_3.htm