DECIPHERING THE ROLES OF MICRO-ENVIRONMENTAL STRESS TYPES IN CANCER INITIATION AND PROGRESSION

By

Chi Zhang

(Under the supervision of Ying Xu)

ABSTRACT

What drives a cancer's evolution, including its initiation? This has been the focus of many published studies in the past century. Among the various proposals, the predominating theory in the past four decades has been that cancer is the result of genomic mutations. However, the mutation-centric view of cancer drivers have been challenged by a number of recent studies, partially inspired by the fact that only a very few common mutations have been observed across different tissues of the same cancer type or even different cells in the same tissue.

In this dissertation, I have listed four research topics in my explorative research of the roles of micro-environmental stresses in cancer initiation and progression in the past five years. The topics include: *(1)* Cancer may be a pathway to cell survival under persistent hypoxia and elevated ROS: A model for solid-cancer initiation and early development; (2) Population Dynamics inside Cancer Biomass Driven by Repeated Hypoxia-Reoxygenation Cycles; (3) Elucidation of Drivers of High-Level Production of Lactates throughout a Cancer Development; and (4) A bi-clustering based approach to comprehensively predict the functional gain or loss of somatic mutations.

Through these projects, I have raised one functional model of the role of hypoxia and oxidative stress in cancer initiation, one mechanistical model of the detailed mechanisms that how inflammation induce cancer, identified several possible dyregulations of stromal cells and their association with cancer imitation and progression, and a series of method to model cancer transcriptomics and genomics data.

DECIPHERING THE ROLES OF MICRO-ENVIRONMENTAL STRESS TYPES IN CANCER INITIATION AND PROGRESSION

By

Chi Zhang

B.S., Peking University, China, 2010

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

2015

DECIPHERING THE ROLES OF MICRO-ENVIRONMENTAL STRESS TYPES IN CANCER INITIATION AND PROGRESSION

By

Chi Zhang

Major Professor: Ying Xu

Committee: Shaying Zhao

Liang Liu

Jonathan Arnold

Electronic Version Approved:

Suzanne Barbour

Dean of the Graduate School

The University of Georgia

December 2015

DEDICATION

This dissertation is dedicated to my love for science.

ACKNOWLEDGEMENTS

I am grateful to my advisor, Professor Ying Xu for his support and guidance over the years. He patiently provided the vision, encouragement and advice necessary for me to proceed through the doctoral program and complete my dissertation. Without his guidance and persistent help this dissertation would not have been possible. I would also like to thank the members of my advisory committee: Dr. Shaying Zhao, Dr. Liang Liu and Dr. Jonathan Arnold for their mentorship and guidance during my graduate training. I would like to thank all past and present members of the Computational System Biology lab in the University of Georgia and Joint Center for Systems Biology in Jilin University for their friendship and help. I would like to thank my girlfriend Ms. Sha Cao for standing beside me throughout my Ph.D. studies.

TABLE OF CONTENTS

Page

CHAPTER I

INTRODUCTION AND LITERATURE REVIEW

Background

In the past five years, starting from learning basic cancer biology and statistical analysis of cancer omics data, I have conducted a series of analyses to understand the mechanisms hidden in real cancer tissues and kept extending and deepening my research topics. In this chapter, I will review the current understanding of cancer biology and challenges in cancer research, followed by my research philosophy and the structure of my thesis. My major discoveries in four different but highly related projects will be introduced in CHAPTERS II-V and I will conclude and discuss my contribution to cancer research field in CHAPTER VI.

A General Introduction of Cancer and Cancer Research

Cancer is a group of diseases with uncontrolled cell proliferations and some of the cells can invade and metastasize to other parts of the body [1-4]. Cancer types are classified into the following major types including myeloma, leukemia, lymphoma, sarcoma and different types of carcinoma, which is commonly regarded as solid cancer types [5]. The majority of cancer incidences occur in solid cancer with primary site in dozens of possible human organs such as brain, oral, thyroid, head and neck, lung, liver, stomach, kidney, colon, bladder, prostate and pancreas. Some organ types with complicated tissue organizations

have multiple distinct cancer types such as the Oligodendrogliomas, Meningiomas, and Astrocytomas in brain and small cell and non-small cell lung cancers.

Through the literature review and analysis I have conducted in the past five years, abnormal cell proliferation and potential of metastasis are the two common characteristics shared by all cancer types and are hence defining features of cancer. Other than these two commonalities, there are several characteristics that are generally shared by most cancer types including reprogramming of energy metabolism, evading immune destruction, sustaining proliferative signaling, tumor microenvironment, evading growth suppressors, resisting cell death, enabling replicative immortality, angiogenesis, and activating invasion and metastasis, which are regarded as the hallmarks of cancer (Figure 1) [1, 2]. It is worth to note that my study has been focused on solid cancers, and even though they share multiple commonalities, each solid cancer type has its distinct characteristics suggesting they should be treated as different disease types.

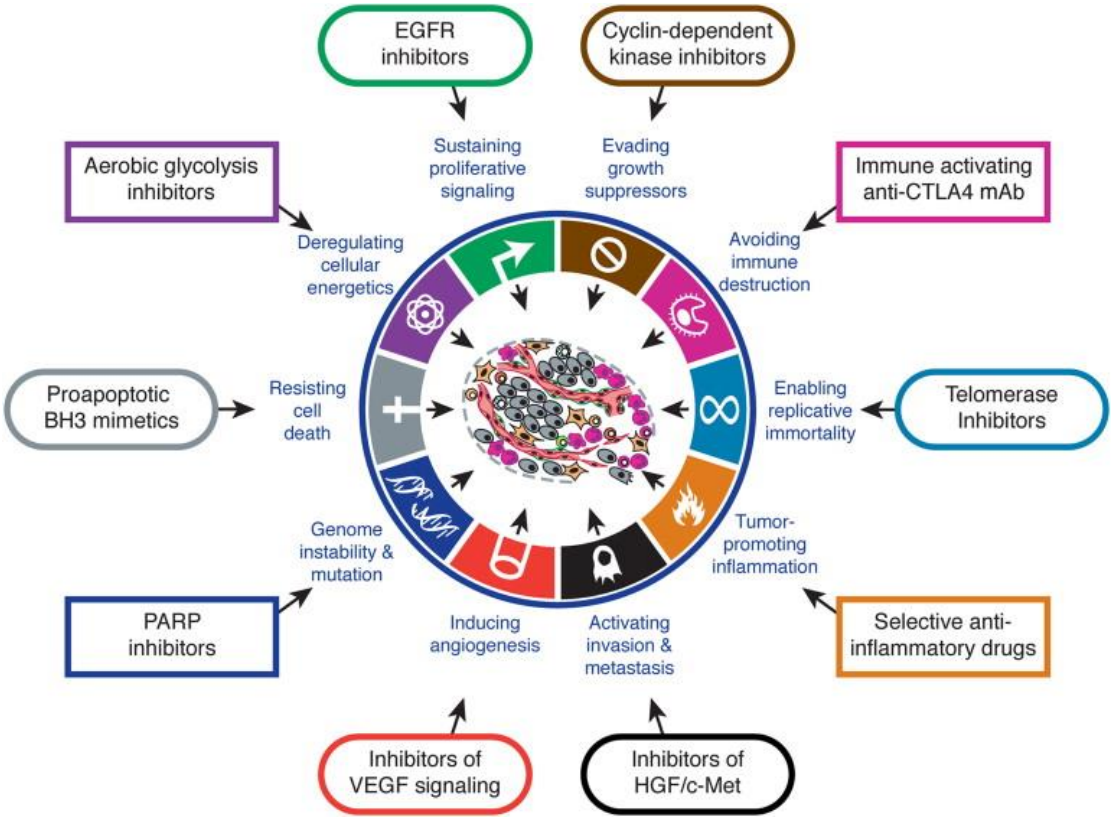


Figure 1.1. Ten hallmarks of cancer from "the Hallmarks of Cancer: next generation", Douglas Hanahan and Robert A. Weinberg [1].

Cancer is a one of the top public health problems in the United States. Around 25% deaths in the United States is related to cancer. In 2014, 1,665,540 new cancer cases and 585,720 cancer deaths are reported in the United States [6]. It is known that five-year survival rates for most of the early stage solid cancers are around 4~10 times for those of advanced stages [5]. Numbers of successful biomarkers for early diagnosis and more general, cancer occurrences, have been developed in the past 30 year. Along with these, a tremendous amount of efforts are invested on finding effective drug targets for cancer at different stages under different circumstances, which rely heavily on the discovered biomarkers. However, a statistical estimation revealed that drugs developed this way generally increase the survival time of corresponding cancer types only by a few months to mostly a few years. The big disappointment is

resulted from the complexity of cancer cells' evolution trajectories. For most of the biomarker detections methods, features including gene expressions, somatic variants are fed into various machine-learning techniques [7-10]. Usually, ones with good sensitivity and specificity would be chosen as candidates, followed by validating experiments performed on cell lines or animals such as mouse. However, fast proliferation allows tumor cells for fast evolution and eventual survivals by evading routes blocked by drugs targeting on those biomarkers, which ceased to guard tumor cell proliferations after a certain period of time [11]. In addition, the interplay between cancer cells and surrounding microenvironment [12], which is of vital importance for cancer evolution but can be hardly captured by current experimental procedures, is usually not taken into account. To better understand the complexity of oncogenesis process, our hope lies on using mathematical and statistical analyses on high throughput omic data collected from real tissue samples, which retains rich information of micro-environmental factors, to identify biological processes and infer the most possible "essential" factors for cancer initiation and progression.

Cancer Initiation Models

Cancer has long been believed as a disease of genomic level disorders, based on which multiple cancer initiation models have been developed that have evolved through the interaction with more diverse studies in related fields in the past 60 years. It is worth to note that even all the mutational cancer initiation models are highly overlapped, but treating them separately can reflect the current emphases in cancer initiation study [13]. In 1950s, with the discovery of first oncogene in bladder cancer, mutation has been believed as the fundamental reason of cancer [14, 15]. Meanwhile, experiments of carcinogens treatment on cell further confirmed the mutation theory [16, 17]. Such a "Mutation Driven" model is

regarded as the first model of carcinogenesis, the essence of which is that cancer is caused by the mutation on certain genes. In 1980s, Knudson's study on retinoblastoma formed the basis of "two-hit" or "multiple-hit" theory, which depict cancer (initiation) is a result of (a sequential of) accumulated mutations [18, 19]. Such a theory lead to the identification and definition of tumor suppressor genes. In 1990s, Fearon and Vogelstein raised a multistep genetic model for the formation of colorectal carcinoma [20], which posits that the formation of colon cancer is by a sequential accumulation of APC, KRAS and TP53 mutations [21, 22]. The key idea of this model is that cancer is caused by a sequential of mutations. In 2000s, with the availability of more epidemiology statistic that show the cancer risk caused by diet, other diseases and environments and the development of new high throughput biotechnologies that can measure whole genome level epigenetic alterations, epigenetic disorders have been observed in all cancers, which raised a series of hypotheses that cancer is caused by epigenetic level disorders, which are results of environmental changes and can induce further mutations [23-25]. More recently, studies in multiple fields suggested that the pure mutation based models are not enough to capture the cancer initiation and progression process and guide the clinical research [26, 27]. Hence lead more researchers start to consider cancer from an evolutionary perspective that the mutations are selected to help the cancer overcome certain micro-environments. The micro-environmental factors such as level of hypoxic and oxidative stress, acidity, and interactions between stroma and cancer that have been neglected, are re-emphasized in current research [28, 29]. Multiple studies revealed that the interactions between stromal cells and cancer cells may take an essential role in the formation of cancer [30, 31]. My analyses in the past five years have targeted multiple dysregulations of stromal cells in precancerous tissues that may

contribute to the carcinogenesis process from several different aspects, the details of which will be discussed in CHAPTER II, III, and VI.

There are a few points worth to be noted here including 1) there are several other theories such as "cancer stem cell" and "Warburg effect" that have also been considered as cancer initiation models [32-35]. These models reflect the oncogenic process from different but parallel aspects comparing to the "mutational" models; 2) all the models are highly overlapped, but treating them separately can reflect the focus of cancer research in different eras. The role of microenvironment factors and their interactions with cancer cells is more emphasized in current research; and 3) current study of cancer initiation process needs a comparative and comprehensive evaluation of the roles of the possible cancer "drivers" in different cancer stages with considering the disease and tissue specificities.

Challenges in Carcinogenesis Research and Possible Computational Solutions

Noting that cancer imitation is a long term evolution process that may encounter interactions among multiple cell types, micro-environmental factors and cellular dysregulations, the study of cancer initiation has the following challenges due to the current experimental conditions and knowledge of the process. It is noteworthy that all the challenges listed here can at least be partially solved by computational modeling of omics data. But such analyses highly depend on the quality of data.

Micro-environmental conditions:

It is worthy emphasizing that cancer, like any evolving organism, is fundamentally driven by its adaptation to the changing microenvironments through selection of certain mutations [36]. In this sense,

the selected mutations provide the footprint information of a cancer evolution. Knowing the encountered stresses in the past could provide fundamentally important information for understanding the phenotypic behavior of a cancer, as well as for elucidating the root causes of the cancer, hence possibly leading to new and more effective ways to treat the cancer. Numerous micro-environmental stresses have been studied in the cancer literature, including lactic acidosis [37], DNA damage[38], imbalance between saturated and unsaturated lipids [39], apoptosis [40], immune attacks [41], imbalance between NAD and NADH [42], dysregulation in ECM component production [43], ER stress [44] and oxidative stress [45, 46].

Unlike genomic or epi-genomic level alterations, the micro-environmental changes are harder to be captured and their impacts are more difficult to be evaluated because 1) there is no direct method to comprehensively capture the micro-environmental alterations in real tissues; 2) such alteration may only exist and function in a certain stage during the cancer initiation; and 3) the micro-environments in real cancer tissue cannot be fully reflected by cell line or animal based experiments [47].

Precancerous stage:

Another challenge is about the transition from pre-cancerous disease stages to cancer. For most tissue data, it is impossible to determine if a collected pre-cancerous tissue such as from cancer prone inflammatory disease or benign tumors will develop malignancy, neither can we assess the exact progression time since the initiation of a collected cancer tissue. For the later question, several studies have reveal the possibility to predict the age a certain tissue by using DNA methylation data [48]. A similar exploratory model can also be applied to predict the relative progression time of a certain cancer tissue [49]. By such a procedure, for any given sample pairs, the probability that one sample is in a

relative more advanced stage comparing to the other can be predicted. It has been proved that such a procedure enables enough prediction power for the distinguished samples pairs. For the first question, comparative analysis among multiple precancerous disease types with known cancer risks (or risk ratios) can identify the possible cancer prone contributing characteristics, based on which a predictor of cancer risks can be trained. However, such analyses demand large scale of training and follow-up experiments to enable sufficient statistical power in the prediction and validation.

Tissue level dysregulations:

The third challenge in analyzing tissue omics data is to computationally decipher the cell types and their relative proportion in the sample tissue, which is termed as de-convolution analysis [50]. Currently, there are dozens of well-established deconvolution methods for different data types including transcriptomics data, genomics data and epigenomics data [51-54]. Most of the methods are supervised method that trains cell specific signatures from cell line based data and apply such signatures to predict the relative proportion of each cell types. For some cell types like the heterogeneous cancer cells in a tumor tissue, no training data is available for such supervised analysis

, an unsupervised method named Non-Negative Factorization (NMF) can be applied to solve the problem [55].

Most existing deconvolution methods assume constant signatures for each cell type. Knowing different cell types may have distinct response under certain micro-environmental conditions, constant signatures are not enough to reflect such alterations. However, noting cellular response to the stresses may follow some rules such as certain stress responsive genes will be up/down regulated together, the

co-expression pattern of such genes can be applied to capture the stress responses, which can be reflected by the co-variance among the genes in large gene expression sets with multiple gene expression conditions. Recently, I have done an exploratory analysis to examine the possibility of considering such "co-variance" in deconvolution analysis by a principle component based approach, with details given in Chapter VI.

Large data size to enable sufficient statistical power:

For most cancer types, there are large data sets contain more than 100 samples available in public domain. However, such data are relatively rarer for precancerous diseases and other specific case-control experiments such as hypoxia/normoxia or oxidative stress/normal treatment experiments. Noting the existence of multiple data sets collected from the same or similar experimental conditions, such data can be combined to achieve enough statistical power, which is generally termed as Meta-analysis. However, such analyses always require specific considerations of the data type specific noise types such as the differences in platforms, sample collection procedures, experimental conditions and other possible batch effect [56, 57]. By the designation of microarray experiment, different probe designs have distinct binging affinity with the corresponding mRNA that may result in totally incomparable expression measures [58]. Also biased sample preparation time may cause different level of mRNA degradations and result in un-estimable noises. Hence for the meta-analysis of gene expression data collected from different species or by different platforms or the data produced under a relative less rigorously defined protocol, combine the results from individual analysis of each data set by meta-analysis methods such as p value combination is recommended [59]. In contrast, certain methods can be applied to remove the batch effects

among the recently produced large scale microarray data sets under well establish experimental protocols and measured by the same platform [60]. For the data measured by Affymetrix UA 133 2.0 plus array, we have applied quantile normalization on the RMA normalized gene expression profile to remove the batch effect among 28 datasets of 11 normal human organ tissues. Dendrogram of the hierarchical clustering computed by the normalized gene expression profile of none housekeeping genes reveals that such a method can remove possible batch effect and remain tissue specific expression signatures in the combined data (See Introduction Appendix Figure).

Unlike simple statistical test based analysis such as differential expression analysis, sophisticated statistical modellings of certain data types are needed to solve such challenges, which raise much higher quality demands for the omics data.

TCGA Data and Other Cancer Omics Data Resources:

In the past seven years, the availability of TCGA data and other large scale data generated from related studies in public domain enable more rational buy sophisticated assumptions in computational study of cancer [61, 62]. For example, genomic sequencing and RNAseq data collected from same cancer tissue samples enable integrative analysis of the mutation and gene expression data to elucidate the transcriptomic level impact of each mutation. Detailed cell specific gene expression signatures can be computed from the large scale gene expression profiles of each immune cell types under different experimental conditions offered by GEO database.

The Cancer Genome Atlas (TCGA) is a collaborative project aiming to produce high quality cancer genomics data for cancer research. Starting from 2006 with long time pilot study, TCGA project has built very standardized pipeline of sample collection, processing, quality control and sequencing data production. Generally, the TCGA data offer multiple omics data types including RNAseq and microarray based transcriptomics data, somatic mutation and copy number variation measured by exon or whole genome sequencing, copy number variation measured by SNP array, proteomics data measured by protein array, microRNA measured by RNAseq and array, DNA methylation data measured by array. New data types are keeping generated by using the stored samples. Rather than the omics data types, TCGA data portal also offer high resolution disease pathological imaging data, information of sample processing procedures, clinical information and pathological report of patients. Level one data of the sequencing data can be accessed with application. Out lab has applied for the access since 2014.

Gene Expression Omnibus (GEO) is a database that stores all the gene expression data as well as DNA methylation data generated in the studies funded by NIH [63]. The current database contain more than ten thousands human data sets measured by different microarray platforms. The GEO database collects data sets from much diverse data sources and has much long time span comparing to TCGA. In the database, there are 4094 data sets covering 115160 samples measured by the most popular human microarray platform – Affymetrix UA133 2.0 plus array and the numbers are keeping updated. It is noteworthy that the GEO database also contains a big collection of data collected from some rarely studied species such as the cancer resistant species naked mole rat and blink mole rat.

Comparing to other data sources, TCGA data have the following advantages: (1) large sample size that enables enough statistical power for most analyses; (2) standard pipelines of sample transportation,

processing, quality control and sequencing data generation that enable comparative analysis among multiple cancer type; (3) complete data sets of level 1-3 sequencing data that contain enough information for complicated modeling analysis such as assumptions of intra-tumor heterogeneity; (4) multiple omic data types are measures from the same cancer tissue that offer possibilities in integrative analysis; and (5) very informative clinical data that can be linked with the molecular characteristics identified in the omics data analysis. In contrast, GEO data sets contain more data sets measured under very diverse experimental conditions but with less standardized data processing method and experimental design can be applied to answer specific questions such as training of predictors for certain micro-environmental stress types. It is worth to note that all the TCGA data are collected from real cancer tissue samples with more than 99% from primary tumor sites. Hence the analysis of pre-cancerous diseases or other disease types should be dependent on other data sources such as GEO data sets. At last, there is another data archive named SRA (Sequence Read Archive) stores the level one sequencing data. Similar to GEO, SRA contains a big collection of data sets measured under various experimental conditions and by different planforms. But the number of experimental conditions and number of data sets in SRA is relative less comparing to GEO because the next generation sequencing data is more advanced and cost more comparing to array based experiments.

Knowledge from Cancer Resistance Species:

Previous studies have shown that some species including hypoxia resistant amphibians and reptiles, longevity rodents, some hibernating species and large animals such as blue whale have certain level of cancer resistance [64-66]. For some species that are feasible for experiments such as turtles, blind mole

rats and naked mole rates, genomic sequencing and chemical treatment experiments have been conducted to characterize the possible cancer resistant mechanisms in such species [67, 68]. Genomic sequencing and anatomical experiment revealed that the naked mole rats tissues continuously produce high density glycosaminoglycan hyaluronic acid, which is believed a chemical protection for the species in its underground habitat. The current knowledge hypothesizes that the large amount of high density hyaluronic acid production can increase the contact proliferation inhibition hence resistance to carcinogenesis in the species [69]. The special necrotic cell death is believed as one of the major cancer resistance mechanism in blind mole rat [68]. Some researchers studying the hypoxia tolerant species has hypothesized that the hypoxia tolerance may associate with certain level of cancer resistance [70]. However, most studies of cancer resistance species only focus on one or several cancer resistance species, hence may lack a general perspective from the commonality of the species. Knowing that cancer initiation is long term selection process under dysregulated micro-environments, study the cancer resistance species in their capability to deal with extreme environments may increase our understanding to carcinogenesis from a different perspective.

Metabolism depression (or suppression), which is defined by the capability to decrease energy demand in response to certain stress type such as starvation or hypoxia [71, 72], has been believed as one commonality in the cancer resistant amphibians and reptiles [65]. The underground living mole rats, species hibernating in enclosed spaces with less oxygen and the blue whale living in deep sea suggest such species may face certain level of oxygen deficiency in their living environments that further indicate the possible association between metabolism depression and cancer resistance.

Topics in My Thesis and the Timeline of My Research

I joined Professor Ying Xu's Computational Systems Biology Lab in the Department of Biochemistry and Molecular Biology of UGA in August of 2010 to pursue my Ph.D. degree in cancer bioinformatics after having received my B.S degree in Mathematics from Peking University, China with two year experience in bioinformatics research as an undergraduate research assistant in a bioinformatics lab. In the 2010-2011, I have majorly involved in several project to study the biological systems of microbial. Even I have not formed such studies into real publications, such studies helped built basic knowledge in system biology. In the fall semester of 2011, I started my first independent cancer project, which is a general comparative analysis among six cancer types. This time I have formed the study into a manuscript. However, with more extensive literature review in the related fields, I found such a study could contribute very limited information to the current understanding of cancer because such a general comparison without a hypothesis driven question can hardly identify significant results and quality of the analyzed data cannot guarantee enough statistical power. During this project, I have done extensive literature reviews about the role of hallmarks in cancer, though which I noticed that most the hallmarks of cancer can be identified in most disease and case-control experiments. Hence the paramount point to the cancer hallmarks should be evaluating the impact of each hallmark and the association among the hallmarks in each data set. Meanwhile, I have involved in two studies majorly conducted by our previous lab member Dr. Kun Xu that led to my first two middle authored academic papers. Through 2011-2013, I have taken multiple statistics classes and conducted several explorative studies of integrative analysis of multiple omic data types, in which I have accumulated some knowledge in computational modeling of cancer omics data. In 2013, with the preliminary studies of the roles of hypoxia and oxidative stress in

cancer progression led by another previous lab member Dr. Juan Cui [73], I conducted one project to elucidate possible detailed roles of hypoxia and oxidative stress in cancer initiation by comparative analysis of the energy consuming genes in the cancer resistance species versus cancer species in response to hypoxia or oxidative stress. In the fall of 2013, I have analyzed around 40 small topics to contribute to Dr. Ying Xu's book "cancer bioinformatics", through which I have summarized dozens of possible future projects. On of such project is formed into my second first-authored paper in 2014 that has identified some cancer tissues may have dynamic cell populations in response to repeated hypoxia and normoxia. In the summer of 2014, I have spent three weeks to visit several bioinformatics and statistics labs in Boston area. With the interests in large biological network analysis, I have conducted my third publication of a co-expression module identification method and its application in elucidating the possible regulators and outcomes of the high lactate production in multiple cancer types. In 2015, I have involved in more collaborations with other lab members and got several co-first authored and middle authored publications. By the August of 2015, I got a chance to thoroughly think about my previous researches and possible research directions in the future. In the long term study on the role of micro-environmental stresses in cancer, I have kept asking myself why we need to focus on the micro-environmental factors.

As most cancer researchers, the ultimate goal in of cancer research is to increase people's health conditions in terms of cancer. In the past five years, I have seen the advantage of computational study that can comprehensively capture the characteristics in real cancer tissue. Meanwhile, I have also witnessed the limitation of the current technologies. Clinical studies have shown that simply consider the mutations are not enough to treat cancer. Highly stressed microenvironments may help the cancer easily develop drug resistance. Even the immunotherapy is believed as a new hope in cancer therapy, the immune attacks

can be highly influenced by the cancer microenvironment. Such micro-environmental factors that have been neglected, are re-emphasized due to the current research demand. However, as mentioned above, there is no technology that can comprehensively identify all the micro-environmental factors. Most of such factors will be fast diminished after the sample dissected from the body. Noting such demands and challenges, by assuming the micro-environmental alterations can be reflected in the omics data of real tissue samples, I have conducted a series of explorative study to quantify possible micro-environmental stress types.

In the beginning of 2015, I started to focus on somatic mutations in cancer and their association with micro-environmental stresses. Although I am clear about the challenges in calling mutations from level one data, my personal interests drive my research more focus on the level 2/3 data analysis. My understanding of this field is that each of the method to predict the role of mutations is benefited by limited by its assumptions. With the development of the high resolution data and more available large data sets, some co-founding factors have been neglected should be emphasized to comprehensively evaluate the role of each mutation. As I have mentioned previously, mutations on the same gene may sever totally different functions in the patients with the mutation. Such differences can be caused by 1) heterogeneous gain or loss of function due to specific mutation patterns, 2) con-current mutations, and 3) interactive effect among multiple mutations. It is worth to note that certain mutation may be selected by one micro-environmental stress in a certain time range during the progression of cancer. Hence the real role of the mutation cannot be reflect in the data if such micro-environmental stress has already vanished when the sample was collected. Hence I conclude the current goal is to develop a computational approach that can cover such possibilities in identification of the roles of each mutation.

In detail, my research in the past five years falls into the following two categories:

A. Model micro-environmental and tissue level alterations and predict their functional roles in cancer initiation and progression: Cancer is a systematic tissue level disease. Current clinical statistics suggest that the most efficient way to control cancer is early diagnosis. Although numerous "driver" mutations have been identified and multiple hypotheses have been made for the cancer initiation and progression process, detailed mechanism level explanations are still lack for how malignancy is developed and progressed from pre-cancerous stage. Considering the cancer initiation and progression is an evolutionary process, i.e. certain cancer associated mutation is selected under certain micro-environment conditions to increase the survival of cancer cells, my first research goal is to comprehensively capture the tissue level alterations in pre-cancerous diseases and early stage cancer, and identify how the dysregulations associated with the malignancy transition. Noting the real micro-environment conditions can be only reflected by the tissue samples collected from cancer patients, my studies are majorly based on statistical modeling of omics of real tissue samples. In the last five years, I have carried out a series of studies and gained the following insights regarding to this issue: (1) the initial stress types in pre-cancerous stages are oxidative stress and hypoxia, which regulate dysregulation of extracellular matrix components further producing initial signals for oncogenesis [74, 75], (2) most of the cancer associated mutations including *TP53*, *PTEN*, *CDKN2A*, *NAV3*, *EGFR*, *VHL* and *PI3KCA* have specific gain or loss of functions corresponding to certain micro-environmental stresses in distinct cancer (tissue) types, suggesting the possible selections of the mutations [76], (3) metabolism of lactate and its role in driving cancer progression [49], and (4) intermittent hypoxia and re-oxygenation may serve as one major pressure driving cancer progression by regulating the relative proportion of different cancer cell populations, i.e.

intra-tumor heterogeneity, in cancer tumor tissue [77]. Specifically, we have conducted a meta-analysis on ~80 transcriptomics data sets of 10 cancer prone and 8 cancer independent inflammatory diseases, 7 neoplastic diseases and benign tumor and 20 cancer types by applying our in-house developed tissue data modeling methods (will be discussed in part B) and revealed the initial stress types in inflammatory induced cancer are the result of dysregulated wounding healing and extra-cellular matrix repair and the key steps in the malignancy transition are shifting from elevated stroma/immune cell proliferation and oxidative stress induced mitochondrion suppression to epithelial (or cancer) cell proliferation and Warburg effect induced mitochondrion suppression [75].

B. Development of computational methods to model genomics and transcriptomics data collected from tissue samples and integrative analysis of multiple omics data types: Since the collected cancer samples may undergo certain unknown micro-environmental conditions such as level of oxygen and immune response and the signals measured from the tissue are the mixture of signals from multiple cell types, specific analyses are needed to study the tissue based data. (1) We have developed predictors for level of oxidative stress and hypoxia by a logistic regression model with L1-regularization on collected training data set and applied the predictors to reflect the association between oxidative stress and cancer risk in an epidemiology work [78]; (2) To identify the alterations of immune and stroma cell proportions in the disease versus normal sample and the gene expression signal from each cell type, I have trained an immune cell deconvolution method that can identify the alterations of 11 immune/stroma cell types by a supervised PC regression approach [75]. (3) For cancer cells without clear gene expression signature, I have conducted an unsupervised deconvolution approach by integrating non-negative matrix factorization of transcriptomics data with somatic mutation data to predict possible intra-tumor heterogeneity [77].

Noting cancer as a systematic disease, I have also developed methods to answer specific biological questions by integrating the multiple omic data types provided in TCGA data set and other public accessible data sources: (4) To answer how the cancer associated mutations are selected in the early stage of cancer, a bi-clustering based approach is developed to predicted the gain or loss of functions of certain mutations by integrating somatic mutation and RNAseq data [76]; (5) With the demand to compute correlations among million level features including exon expressions and methylation data, I have developed a non-parametric gene co-expression network analysis approach for module identification in large biological network that can handle around one million feature [49, 79]; (6) To accurately estimate the metabolism flux changes in disease versus normal samples, a Markov Chain Monte Carlo based method by integrating quantitative metabolomics data, transcriptomics data, enzyme kinetic parameters (accessed from BRENDA database) and metabolic pathway topology is under developing. I have also conducted mathematical proofs to enable correctness for certain developed method. (7) In the bi-clustering method, a Mixed Gaussian model with left truncation is specifically applied to handle the error of low expressed genes in RNAseq data, for which I have provide the mathematical proof of the EM algorithm. (8) To enable the feasibility of the gene co-expression network analysis method on large network, I have theoretically derived a simulatable null distribution of the non-parametric statistics for hub gene identification.

The projects majorly conducted by me are selected and presented in this dissertation by the following orders: **CHAPTER II:** Cancer may be a pathway to cell survival under persistent hypoxia and elevated ROS: A model for solid-cancer initiation and early development; **CHAPTER III:** Population Dynamics inside Cancer Biomass Driven by Repeated Hypoxia-Reoxygenation Cycles; **CHAPTER VI:** Elucidation

of Drivers of High-Level Production of Lactates throughout a Cancer Development; **CHAPTER V:** A bi-clustering based approach to predict the functional gain or loss of somatic mutations (ongoing project); and **CHAPTER VI:** Conclusions and Discussions.

References:

1. Hanahan, D. and R.A. Weinberg, *Hallmarks of cancer: the next generation.* Cell, 2011. **144**(5): p. 646-74.
2. Hanahan, D. and R.A. Weinberg, *The hallmarks of cancer.* Cell, 2000. **100**(1): p. 57-70.
3. Cancer Genome Atlas, N., *Comprehensive molecular portraits of human breast tumours.* Nature, 2012. **490**(7418): p. 61-70.
4. Weinberg, R.A., *The Biology of Cancer*. 2007.
5. Cancer.org. *http://www.cancer.gov/types*.
6. Siegel, R., et al., *Cancer statistics, 2014.* CA Cancer J Clin, 2014. **64**(1): p. 9-29.
7. de Gramont, A., et al., *Pragmatic issues in biomarker evaluation for targeted therapies in cancer.* Nat Rev Clin Oncol, 2015. **12**(4): p. 197-212.
8. Diamandis, E.P., *Cancer biomarkers: can we turn recent failures into success?* J Natl Cancer Inst, 2010. **102**(19): p. 1462-7.
9. Malinowski, D.P., *Multiple biomarkers in molecular oncology. II. Molecular diagnostics applications in breast cancer management.* Expert Rev Mol Diagn, 2007. **7**(3): p. 269-80.
10. Whitfield, M.L., et al., *Common markers of proliferation.* Nat Rev Cancer, 2006. **6**(2): p. 99-106.
11. Holohan, C., et al., *Cancer drug resistance: an evolving paradigm.* Nat Rev Cancer, 2013. **13**(10):

p. 714-26.

12. Quail, D.F. and J.A. Joyce, *Microenvironmental regulation of tumor progression and metastasis.* Nat Med, 2013. **19**(11): p. 1423-37.

13. Vineis, P., A. Schatzkin, and J.D. Potter, *Models of carcinogenesis: an overview.* Carcinogenesis, 2010. **31**(10): p. 1703-9.

14. Armitage, P. and R. Doll, *The age distribution of cancer and a multi-stage theory of carcinogenesis.* Br J Cancer, 1954. **8**(1): p. 1-12.

15. Parada, L.F., et al., *Human EJ bladder carcinoma oncogene is homologue of Harvey sarcoma virus ras gene.* Nature, 1982. **297**(5866): p. 474-8.

16. Doll, R. and R. Peto, *Cigarette smoking and bronchial carcinoma: dose and time relationships among regular smokers and lifelong non-smokers.* J Epidemiol Community Health, 1978. **32**(4): p. 303-13.

17. *IARC monographs on the evaluation of the carcinogenic risk of chemicals to man: some miscellaneous pharmaceutical substances.* IARC Monogr Eval Carcinog Risk Chem Man, 1977. **13**: p. 1-255.

18. Knudson, A., *Retinoblastoma: teacher of cancer biology and medicine.* PLoS Med, 2005. **2**(10): p. e349.

19. Slaga, T.J., *Multistage skin carcinogenesis: a useful model for the study of the chemoprevention of cancer.* Acta Pharmacol Toxicol (Copenh), 1984. **55 Suppl 2**: p. 107-24.

20. Fearon, E.R. and B. Vogelstein, *A genetic model for colorectal tumorigenesis.* Cell, 1990. **61**(5): p. 759-67.

21. Batistatou, A., A. Charalabopoulos, and K. Charalabopoulos, *Molecular basis of colorectal cancer.* N Engl J Med, 2010. **362**(13): p. 1246; author reply 1246-7.

22. Purnak, T., E. Ozaslan, and C. Efe, *Molecular basis of colorectal cancer.* N Engl J Med, 2010. **362**(13): p. 1246; author reply 1246-7.

23. Greaves, M., *Darwinian medicine: a case for cancer.* Nat Rev Cancer, 2007. **7**(3): p. 213-21.

24. Virani, S., et al., *Cancer epigenetics: a brief review.* ILAR J, 2012. **53**(3-4): p. 359-69.

25. Dawson, M.A. and T. Kouzarides, *Cancer epigenetics: from mechanism to therapy.* Cell, 2012. **150**(1): p. 12-27.

26. Gerlinger, M., et al., *Intratumor heterogeneity and branched evolution revealed by multiregion sequencing.* N Engl J Med, 2012. **366**(10): p. 883-92.

27. Michor, F. and V.M. Weaver, *Understanding tissue context influences on intratumour heterogeneity.* Nat Cell Biol, 2014. **16**(4): p. 301-2.

28. Mbeunkui, F. and D.J. Johann, Jr., *Cancer and the tumor microenvironment: a review of an essential relationship.* Cancer Chemother Pharmacol, 2009. **63**(4): p. 571-82.

29. Whiteside, T.L., *The role of immune cells in the tumor microenvironment.* Cancer Treat Res, 2006. **130**: p. 103-24.

30. Zhang, W. and P. Huang, *Cancer-stromal interactions: role in cell survival, metabolism and drug sensitivity.* Cancer Biol Ther, 2011. **11**(2): p. 150-6.

31. Choi, J., et al., *Metabolic interaction between cancer cells and stromal cells according to breast cancer molecular subtype.* Breast Cancer Res, 2013. **15**(5): p. R78.

32. Yoo, M.H. and D.L. Hatfield, *The cancer stem cell theory: is it correct?* Mol Cells, 2008. **26**(5): p.

514-6.

33. Dhawan, P., et al., *Cancer stem cells and colorectal cancer: an overview.* Curr Top Med Chem, 2011. **11**(13): p. 1592-8.

34. Vander Heiden, M.G., L.C. Cantley, and C.B. Thompson, *Understanding the Warburg effect: the metabolic requirements of cell proliferation.* Science, 2009. **324**(5930): p. 1029-33.

35. Lopez-Lazaro, M., *The warburg effect: why and how do cancer cells activate glycolysis in the presence of oxygen?* Anticancer Agents Med Chem, 2008. **8**(3): p. 305-12.

36. Gatenby, R.A. and R.J. Gillies, *A microenvironmental model of carcinogenesis.* Nat Rev Cancer, 2008. **8**(1): p. 56-61.

37. Dhup, S., et al., *Multiple biological activities of lactic acid in cancer: influences on tumor growth, angiogenesis and metastasis.* Curr Pharm Des, 2012. **18**(10): p. 1319-30.

38. Lieberman, H.B., *DNA damage repair and response proteins as targets for cancer therapy.* Curr Med Chem, 2008. **15**(4): p. 360-7.

39. Ackerman, D. and M.C. Simon, *Hypoxia, lipids, and cancer: surviving the harsh tumor microenvironment.* Trends Cell Biol, 2014. **24**(8): p. 472-8.

40. Lowe, S.W. and A.W. Lin, *Apoptosis in cancer.* Carcinogenesis, 2000. **21**(3): p. 485-95.

41. Adam, J.K., B. Odhav, and K.D. Bhoola, *Immune responses in cancer.* Pharmacol Ther, 2003. **99**(1): p. 113-32.

42. Ying, W., *NAD+/NADH and NADP+/NADPH in cellular functions and cell death: regulation and biological consequences.* Antioxid Redox Signal, 2008. **10**(2): p. 179-206.

43. Emery, L.A., et al., *Early dysregulation of cell adhesion and extracellular matrix pathways in*

breast cancer progression. Am J Pathol, 2009. **175**(3): p. 1292-302.

44. Clarke, H.J., et al., *Endoplasmic reticulum stress in malignancy.* Cancer Cell, 2014. **25**(5): p. 563-73.

45. Reuter, S., et al., *Oxidative stress, inflammation, and cancer: how are they linked?* Free Radic Biol Med, 2010. **49**(11): p. 1603-16.

46. Klaunig, J.E., L.M. Kamendulis, and B.A. Hocevar, *Oxidative stress and oxidative damage in carcinogenesis.* Toxicol Pathol, 2010. **38**(1): p. 96-109.

47. Kamb, A., *What's wrong with our cancer models?* Nat Rev Drug Discov, 2005. **4**(2): p. 161-5.

48. Horvath, S., *DNA methylation age of human tissues and cell types.* Genome Biol, 2013. **14**(10): p. R115.

49. Zhang, C., et al., *Elucidation of drivers of high-level production of lactates throughout a cancer development.* J Mol Cell Biol, 2015. **7**(3): p. 267-79.

50. Brereton, R.G., *Tutorial review. Deconvolution of mixtures by factor analysis.* Analyst, 1995.

51. Ahn, J., et al., *DeMix: deconvolution for mixed cancer transcriptomes using raw measured data.* Bioinformatics, 2013. **29**(15): p. 1865-71.

52. Yoshihara, K., et al., *Inferring tumour purity and stromal and immune cell admixture from expression data.* Nat Commun, 2013. **4**: p. 2612.

53. Down, T.A., et al., *A Bayesian deconvolution strategy for immunoprecipitation-based DNA methylome analysis.* Nat Biotechnol, 2008. **26**(7): p. 779-85.

54. Brunet, J.P., et al., *Metagenes and molecular pattern discovery using matrix factorization.* Proc Natl Acad Sci U S A, 2004. **101**(12): p. 4164-9.

55. Gaujoux, R. and C. Seoighe, *Semi-supervised Nonnegative Matrix Factorization for gene expression deconvolution: a case study.* Infect Genet Evol, 2012. **12**(5): p. 913-21.

56. Chen, C., et al., *Removing batch effects in analysis of expression microarray data: an evaluation of six batch adjustment methods.* PLoS One, 2011. **6**(2): p. e17238.

57. Leek, J.T., et al., *Tackling the widespread and critical impact of batch effects in high-throughput data.* Nat Rev Genet, 2010. **11**(10): p. 733-9.

58. Schwanhausser, B., et al., *Global quantification of mammalian gene expression control.* Nature, 2011. **473**(7347): p. 337-42.

59. Won, S., et al., *Choosing an optimal method to combine P-values.* Stat Med, 2009. **28**(11): p. 1537-53.

60. Lazar, C., et al., *Batch effect removal methods for microarray gene expression data integration: a survey.* Brief Bioinform, 2013. **14**(4): p. 469-90.

61. Tomczak, K., P. Czerwinska, and M. Wiznerowicz, *The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge.* Contemp Oncol (Pozn), 2015. **19**(1A): p. A68-77.

62. Chin, L., J.N. Andersen, and P.A. Futreal, *Cancer genomics: from discovery science to personalized medicine.* Nat Med, 2011. **17**(3): p. 297-303.

63. Barrett, T., et al., *NCBI GEO: archive for functional genomics data sets--update.* Nucleic Acids Res, 2013. **41**(Database issue): p. D991-5.

64. Seluanov, A., et al., *Hypersensitivity to contact inhibition provides a clue to cancer resistance of naked mole-rat.* Proc Natl Acad Sci U S A, 2009. **106**(46): p. 19352-7.

65. Ruben, L.N., R.H. Clothier, and M. Balls, *Cancer resistance in amphibians.* Altern Lab Anim,

2007. **35**(5): p. 463-70.

66. Manov, I., et al., *Pronounced cancer resistance in a subterranean rodent, the blind mole-rat, Spalax: in vivo and in vitro evidence.* BMC Biol, 2013. **11**: p. 91.

67. Kim, E.B., et al., *Genome sequencing reveals insights into physiology and longevity of the naked mole rat.* Nature, 2011. **479**(7372): p. 223-7.

68. Gorbunova, V., et al., *Cancer resistance in the blind mole rat is mediated by concerted necrotic cell death mechanism.* Proc Natl Acad Sci U S A, 2012. **109**(47): p. 19392-6.

69. Tian, X., et al., *High-molecular-mass hyaluronan mediates the cancer resistance of the naked mole rat.* Nature, 2013. **499**(7458): p. 346-9.

70. Koumenis, C., *ER stress, hypoxia tolerance and tumor progression.* Curr Mol Med, 2006. **6**(1): p. 55-69.

71. Hochachka, P.W., et al., *Unifying theory of hypoxia tolerance: molecular/metabolic defense and rescue mechanisms for surviving oxygen lack.* Proc Natl Acad Sci U S A, 1996. **93**(18): p. 9493-8.

72. Nilsson, G.E. and P.L. Lutz, *Role of GABA in hypoxia tolerance, metabolic depression and hibernation--possible links to neurotransmitter evolution.* Comp Biochem Physiol C, 1993. **105**(3): p. 329-36.

73. Cui, J., et al., *Hypoxia and miscoupling between reduced energy efficiency and signaling to cell proliferation drive cancer to grow increasingly faster.* J Mol Cell Biol, 2012. **4**(3): p. 174-6.

74. Zhang, C., et al., *Cancer may be a pathway to cell survival under persistent hypoxia and elevated ROS: a model for solid-cancer initiation and early development.* Int J Cancer, 2015. **136**(9): p. 2001-11.

75. Zhang C, Y.F., Dong N, Du W, Tang T, Cao S, Sheng T, Chen X, Xu Y, *A Pan-inflammatory and precancerous disease analysis reveals key biological characteristics in cancer risking chronic inflammatory disease types.* To be submitted, 2015.

76. Zhang C, S.T., Cao S, Ma Q, Xu Y., *A bi-clustering based method to predict gain or loss of function of somatic mutations.* To be submitted, 2015.

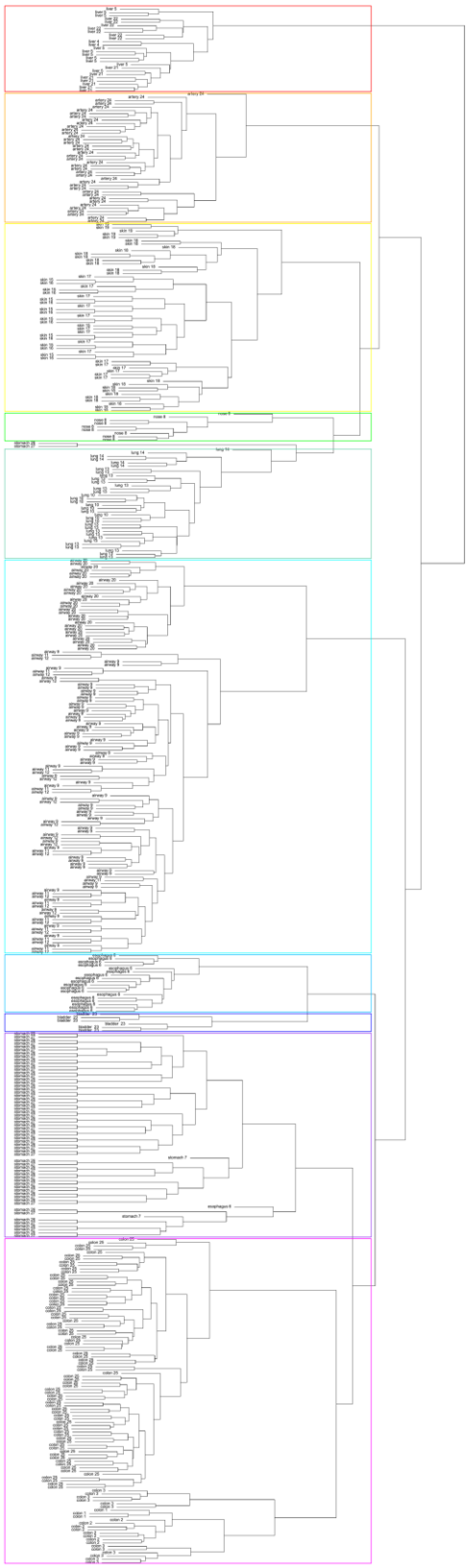
77. Zhang C, C.S.a.X.Y., *Population Dynamics inside Cancer Biomass Driven by Repeated Hypoxia-Reoxygenation Cycles.* Quantitative Biology, 2014.

78. Cao, S., C. Zhang, and Y. Xu, *Somatic mutations may not be the primary drivers of cancer formation.* Int J Cancer, 2015. **137**(11): p. 2762-5.

79. Zhang C, Z.Y., Cao S, Xu Y, *A fast algorithm to compute correlation networks in DNA methylation data*. To be submitted, 2015.

Introduction Appendix Figure:

Figure Legend: We have applied our method to remove the batch effect in a combined data of 28 gene expression data sets of 11 human tissue types. The dendrogam shows that the method can remove the batch effect and keep the tissue specific patterns. In the figure, each colored block indicate one tissue type while the number of each term suggest the data set index. It can be seen the some batch effect among the data sets inside each tissue block has not been fully removed.



CHAPTER II

CANCER MAY BE A PATHWAY TO CELL SURVIVAL UNDER PERSISTENT HYPOXIA AND ELEVATED ROS: A MODEL FOR SOLID-CANCER INITIATION AND EARLY DEVELOPMENT

Chi Zhang[1], Sha Cao[1], Bryan P. Toole[2], and Ying Xu[1, 3*]. 2014. International Journal of Cancer. Volume 136, Issue 9, pages 2001–2011, 1 May 2015.

Abstract

A number of proposals have been made in the past century regarding what may drive sporadic cancers to initiate and develop. Yet, the problem remains largely unsolved as none of the proposals have been widely accepted as cancer-initiation drivers. We propose here a driver model for the initiation and early development of solid cancers associated with inflammation-induced chronic hypoxia and ROS accumulation. The model consists of five key elements: (i) human cells tend to have a substantial gap between ATP demand and supply during chronic hypoxia, which would inevitably lead to increased uptake of glucose and accumulation of its metabolites; (ii) the accumulation of these metabolites will cast mounting pressure on the cells and ultimately result in the production and export of hyaluronic acid; (iii) the exported hyaluronic acid will be degraded into fragments of various sizes, serving as tissue-repair signals, including signals for cell proliferation, cell survival and angiogenesis, which lead to the initial proliferation of the underlying cells; (iv) cell division provides an exit for the accumulated glucose metabolites by using them towards macromolecular synthesis for the new cell, and hence alleviate the pressure from the metabolite accumulation; and (v) this process continues as long as the hypoxic condition persists. In tandem, genetic mutations may be selected to make cell divisions and hence survival more sustainable and efficient, also increasingly more uncontrollable. This model also applies to some hereditary cancers as their key mutations, such as BRCA for breast cancer, generally lead to increased ROS and ultimately to repression of mitochondrial activities and up-regulation of glycolysis, as well as hypoxia; hence the energy gap, glucose-metabolite accumulation, hyaluronic acid production and continuous cell division for survival.

Introduction

The most popular theories about cancer and cancer drivers in the past century include: (1) Warburg's theory as summarized by him in 1960 [1]: "Cancer … has countless secondary causes; But there is only one prime cause, (which) is the replacement of respiration of oxygen in normal body cells by a fermentation of sugar"; (2) the genomic mutation theory of cancer, including mutations that lead to the formation of oncogenes and loss of function in tumor suppressor genes, which has been the most popular in the past four

decades; and (3) microbe-induced cancers such as cervical cancers induced by the infection of human papilloma virus [2] or liver cancer induced by hepatitis viruses [3].

Among these major proposals, Warburg's proposal has been most intriguing and has received considerable renewed interests in the past few years [4, 5]. However the proposal clearly lacks key information that connects the observed energy-metabolism reprogramming to cell proliferation, hence it remains as a proposal rather than a testable model. Various genetic mutation-centric driver models have been proposed since the first discoveries of oncogenes by Bishop and Varmus [6] and tumor-suppressor genes by Knudson [7] about 40 years ago. These models include *APC* gene mutation-based driver model for colon cancer [8] and the Philadelphia chromosome-based model for chronic myelogenous leukemia [9]. As of now, hundreds of "driver mutations" have been predicted for various cancers[10]. One fundamental problem with these mutation-centric driver proposals is that activated oncogenes alone, even coupled with mutations in some tumor-suppressor genes, cannot lead to cell proliferation in a tissue environment since numerous conditions must be satisfied before the cells can divide, including: (1) the cells must be attached to their extracellular matrix [11]; (2) the associated extracellular matrix (ECM) needs to have certain physical properties to support cell division [12, 13]; (3) the cells must overcome the contact-inhibition constraint designed to prevent cell over-growth [14]; (4) cell survival signal(s) need to be present [15]; and (5) the cells need to be in specific morphology [16], among a few other conditions needed for tissue development, remodeling or repair [17]. None of the suggested driver models have proposed mechanisms to overcome these tissue-level constraints, hence making them not applicable as candidates for cancer initiation drivers.

Virus-induced cancer models tend to lack molecular level details. For example, HPV-associated cervical cancers are among the most studied virus-induced cancers, but yet no models have been published that functionally link HPV infection to initiation of cervical cancer, other than a recent publication that observed integration of HPV DNA into the host genome [18]. Overall, no experimentally testable models for cancer initiation have been published.

Here we present a model for the initiation of solid cancers in general. The starting point of the model

is the reprogrammed energy metabolism, which was proposed by Otto Warburg some 50 years ago to be the primary cause of all cancers. This reprogramming between the two energy metabolisms could be the result of chronic hypoxia and/or reactive oxygen species (ROS) accumulation [19, 20], which may be induced by chronic inflammation or by genetic mutations in the case of hereditary cancers. It is worth-noting that chronic inflammation has long been linked to cancer development [21, 22], and it can lead to hypoxia [23-25] and ROS accumulation [26]. These conditions, if not inducing cell death, would lead to increased uptake of glucose from circulation and cellular accumulation of glucose metabolites [27]. However the connection from here to cell proliferation has been elusive. Our model in this paper suggests that the glucose metabolite accumulation will ultimately lead to the production and export of hyaluronic acid under persistent hypoxic conditions [28], which is supported by the available transcriptomic data of both hypoxia-treated cell lines and cancer tissues.

It has been well established in the literature of tissue injury and repair that when a tissue is injured, its ECMs will be fragmented and the hyaluronic acid fragments of different sizes released from the ECM will serve as signals for tissue repair, including signals for cell proliferation, cell survival and angiogenesis [29]. Our analysis has shown that cancer-forming cells have utilized the hyaluronic acid synthesized from the accumulated glucose metabolites, which will be fragmented once exported into extracellular space, to mimic the ECM-released hyaluronic acid fragments as tissue repair signals, hence leading to cell proliferation, which will continue as long as the inflammation-induced hypoxic condition persists.

In addition, a proposal is made regarding how this model relates to genomic mutation-induced constitutive cell proliferation after the cell proliferation is already started. To the best of our knowledge, this represents the first such molecular level model that connects chronic inflammation, hypoxia and/or ROS accumulation to cell proliferation.

Results

Energy gap during hypoxia, metabolite accumulation and evolutionary pressure

Why do human, mouse and rat develop cancer but some other vertebrates such as naked mole rat,

blind mole rat and turtle do not or rarely do? This question has puzzled medical practitioners and researchers for decades. A number of studies have been published, aiming to explain why some organisms such as amphibians, naked and blind mole rats do not develop cancers [30-34]. The proposed mechanisms tend to be organism-specific, such as alterable immune systems being used by amphibians, special abilities to resist ROS in blind mole rats, and a capability in producing long hyaluronic acid polymers by naked mole rats [30-34]. We suspect that there is something more fundamental than these proposals, some common characteristics shared by the basic metabolisms of these organisms, which are distinct from those cancer-prone organisms.

We have examined the above question from the perspective of ATP demand and supply during hypoxia in human, mouse, rat, hypoxia-tolerant rat, naked mole rat, blind mole rat, frog and turtle. These organisms are selected because they are known to either develop cancer or rarely do and their ATP-consumption data is publicly available or can be reliably estimated based on transcriptomic data. The question we address here is: *What percentage of the ATP-consuming proteins is substantially repressed during hypoxia versus normoxia*?

According to the published data, proteins in the following six processes/enzyme families consume on average 84% of the ATPs in vertebrate cells: translation, Na+/K+ ATPase, Ca2+ ATPase, gluconeogenesis, urea synthesis and actin ATPase [35]. Among the eight organisms under consideration, naked mole rat and hypoxia-tolerant rat have both ATP consumption data by these proteins and their corresponding gene-expression data under hypoxia (1-5% oxygen) *versus* normoxia (21% oxygen) [36, 37] in the public domain. In addition, ATP consumption data under hypoxia *versus* normoxia are also publicly available for frog and turtle[38, 39] but without gene-expression data. For human, mouse, rat and blind mole rat, only gene-expression data under hypoxia *versus* normoxia are publicly available (see Supplementary Material Section A for details).

We have derived a regression model between the reduced ATP consumption and the reduced gene-expression levels of the six groups of proteins for naked mole rat and hypoxia-tolerant rat (see the rationale and model in Supplementary Material Sections B and C); and then applied this model to the reduced gene-

expression levels of the same proteins in mouse, rat and blind mole rat, respectively, to estimate their reduced levels of ATP demand during hypoxia. The basic assumption here is that the relationship between gene-expression levels and the activity levels of the six groups of proteins in mouse and rat is highly similar to that in naked mole rat, hypoxia-tolerant rat and blind mole rat. This is a reasonable assumption since these five organisms are closely related except that the latter three have an additional regulatory mechanism that can turn down the gene expression levels of these proteins under hypoxia *versus* normoxia while the former two do not [40-44]. It is worth noting that there is a strong correlation between gene-expression changes and protein abundance changes in metabolic proteins [45] (see details in Supplementary Material Section D).

Using the regression model developed for naked mole rat and hypoxia tolerant rat as well as reduced expression data of mouse, rat and blind mole rat under hypoxia *versus* normoxia, we have predicted the reduction percentages in ATP consumption in mouse, rat and blind mole rat, respectively. Figure 1 shows the prediction result along with the known reduction percentages in the ATP consumption levels in naked mole rat, hypoxia-tolerant rat, turtle and frog. The predicted reductions in ATP consumption are highly consistent with the general knowledge about these organisms. For example, blind mole rats can substantially lower their metabolic rates, hence ATP consumption, when switching from normoxia to hypoxia [46]; and they are known to be more tolerant to hypoxia than naked mole rats since the they can stay viable at 3% oxygen [31] while naked mole rats require a minimum of 10% oxygen, which can occasionally get down to 5% [47]. In addition, naked mole rats are known to be able to live both above the ground and deep in the underground, while rats tend to live much closer to the ground when they are underground. (More supporting data in Supplementary Section D).

Based on this meaningful prediction, we further extrapolate this prediction to include human (see Figure 1). The prediction indicates that under hypoxia, human cells have the least reduction in their ATP consumption.

It is worth noting that this figure is used to make a qualitative (rather than quantitative) statement that human, mouse and rat have substantially smaller reductions in their ATP consumption compared to the

other five organisms when the condition switches from normoxia to hypoxia. To have more accurate estimates of ATP-consumption reductions by these organisms, we need substantially more data under multiple hypoxic conditions *versus* matching normoxia, which are currently not publicly available (the above are all the available data we can find in the public domain). Considering that these organisms use essentially the same glycolytic fermentation processes, we hypothesize that the percentages in the ATP-supply reduction are comparable across these organisms when the condition switches from normoxia to hypoxia. Hence the estimated level of reduction percentage in ATP consumption reflects the gap between the ATP demand and ATP supply in each organism, i.e., the smaller the reduction percentage, the larger the gap.

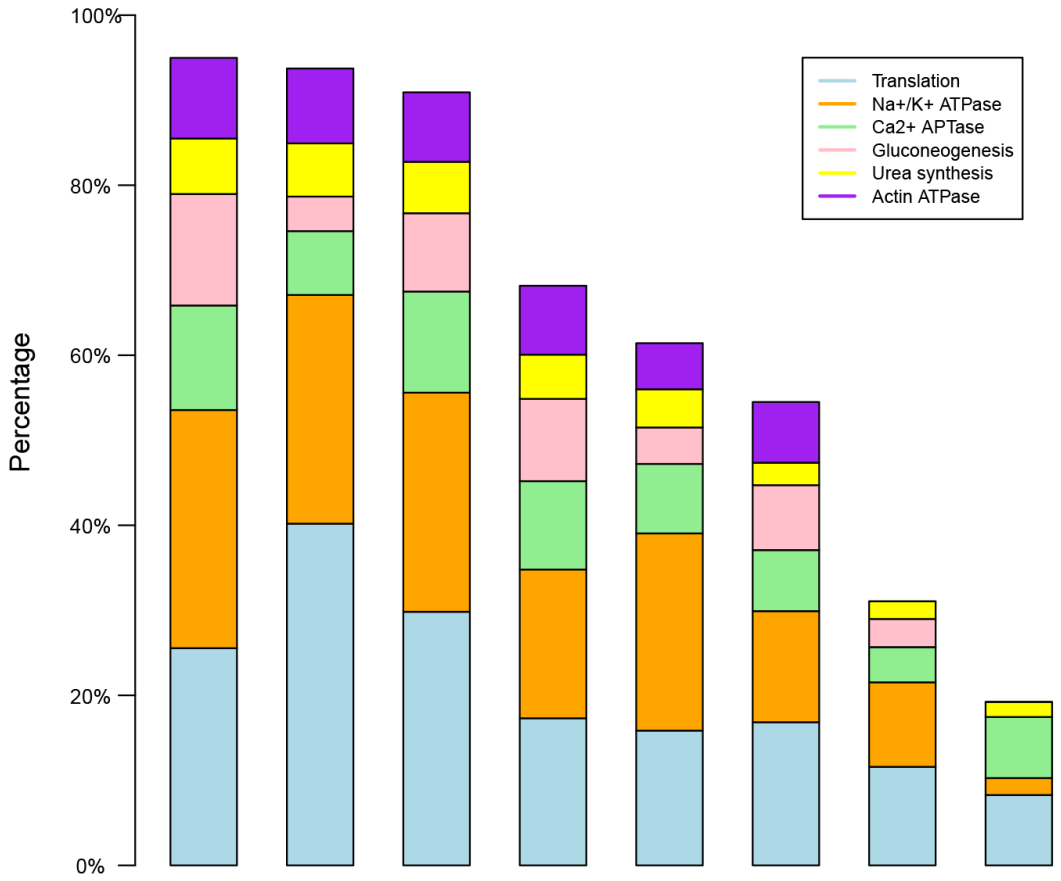


Figure 2.1: For human, rat, mouse, naked mole rat, hypoxia-tolerant rat, blind mole rat, frog and turtle

(from left to right), 1.0 along the y-axis represents the total ATP demand by the six groups of proteins during normoxia. The height of each bar represents the percentage of ATP demand under the matching hypoxic condition, hence the difference between 1.0 and the percentage representing the percentage in reduction. Each bar is divided into six color-coded sections, each representing one group of relevant proteins under consideration. ATP consumption data for turtle, frog, naked mole rat and hypoxia-tolerant rat are collected from the literature and the data for human, rat, mouse and blind mole rat are predicted.

Knowing that frog and turtle can live well under very hypoxic conditions for extended periods of time, it is reasonable to assume that they have no gap between their energy supply and demand under hypoxia [48]. Similarly the energy gap in blind mole rats must be none or small since they can also live well under hypoxia for extended periods of time [42, 44]. Naked mole rat is an interesting case as it not only has reduced energy demand but also has an additional capability to avoid the glucose metabolite accumulation issue that humans have to deal with during hypoxia (see later sections). In contrast, human (and mouse, rat) all have substantial energy gaps during hypoxia, and they cannot live under hypoxia for long [48, 49]. This makes sense since the humans as a population have not lived under hypoxic conditions (e.g., 10% oxygen or lower) for extended periods of time in the past millions of years of evolution, and hence human cells (possibly except for certain tissue types) have not been trained to adapt to such a condition by switching off some portions of the ATP-consuming metabolic reactions to keep the ATP demand within its supply under hypoxia.

Because of the large energy gap, human (and mouse, rat) cells would substantially increase their glucose uptake during hypoxia to make up for the reduced energy efficiency due to the switch from aerobic respiration to anaerobic fermentation, to meet the ATP needs of the hungry cells, which causes these cells to accumulate various glycolytic metabolites as widely observed [27]. The accumulation, we believe, is another result of human evolution. That is, there is a mismatch between the influx rate of glucose, which is regulated by the ATP deficiency, and the maximum flux rate of the glycolytic fermentation pathway, which has been shaped by evolution. Knowing that human cells have not been under hypoxic conditions for long

time, it is reasonable to infer that the glycolytic fermentation system has been serving only as a supplement to the aerobic respiration system for ATP production for very short periods of time, so its capacity has been evolutionarily determined correspondingly. Hence we posit that the maximum flux rate of this system is intrinsically unable to meet the need for dealing with the substantially increased influx of glucose during hypoxia, resulting in the accumulation of glucose metabolites and other molecules (see next section).

Continuous accumulation of the metabolic derivatives, if not removed, will lead to cell death [50, 51]. Thus we propose that the need for removing the accumulated metabolic derivatives forms the initial pressure for the affected cells to evolve; cell division may represent a most feasible way to remove the accumulated glycolytic metabolites by using them towards DNA synthesis. In addition, we propose that it is this capability in switching off certain metabolic activities to keep the ATP demand within the ATP supply under hypoxia that decides if an organism has the potential to develop cancer or not.

In addition, as detailed in Supplementary Material Section E and F, amino acid and fatty acid metabolisms can also contribute to the congestion of glycolysis pathway under hypoxia.

Hyaluronic acid: a key facilitator of cell proliferation

Hyaluronic acid is a long chain of a repeating disaccharide, up to 2 x 10^5 disaccharides, each derived from one D-glucuronic acid (GlcUA) and one D-N-acetylglucosamine (GlcNAc) [52]. Hyaluronic acid is a key component of an ECM structure and mediates cell-ECM signaling. This large polymer has long been known to be associated with cancer development [53, 54]. The most relevant function of the molecule in this context is signaling roles played by its fragments related to tissue repair.

Briefly, when assaulted, an injured tissue releases ECM fragments, among which hyaluronic acid fragments serve as signals for repairing the injured tissue. Most interestingly, hyaluronic acid fragments of different sizes have been found to serve as signals for different purposes, including the induction of inflammation, anti-apoptosis, cell survival, cell-cycle activation, cell proliferation, activation of angiogenesis and cell motility, all related to injury response, maintenance of tissue integrity, and tissue repair [29, 55]. It is worth emphasizing that all short hyaluronic acid fragments (< ~5000 disaccharides)

serve some signaling roles relevant to the above list. To see the connection between glucose metabolite accumulation and cell proliferation, we need to examine the synthesis pathway of hyaluronic acid (Figure 2).

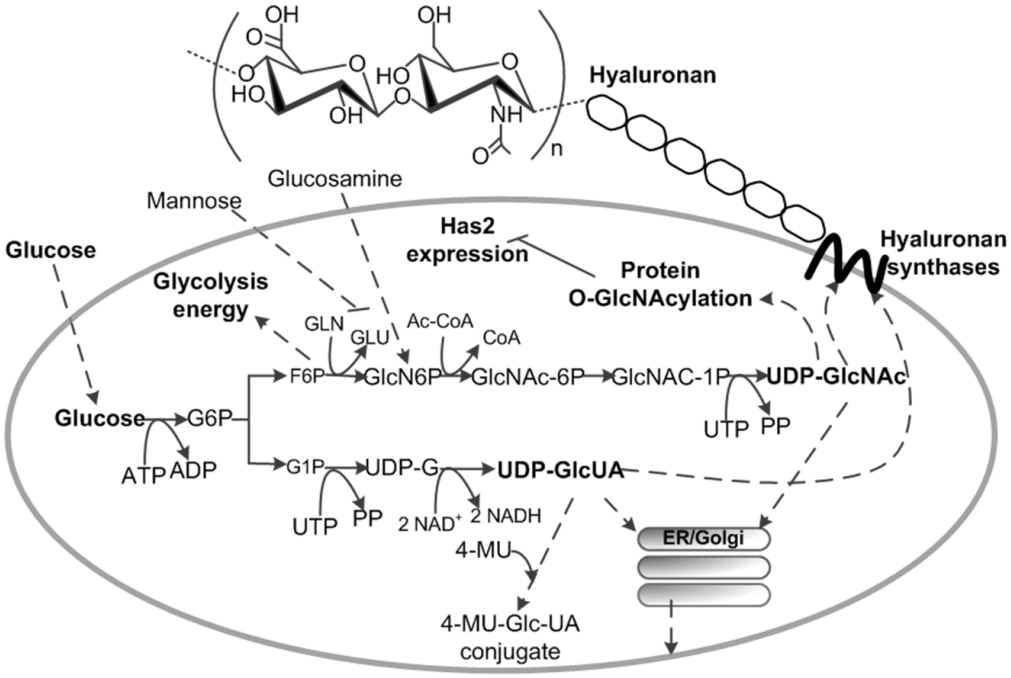


Figure 2.2: The synthesis pathway of hyaluronic acid from UDP-GlcUA and UDP-GlcNAc, both derived from the common precursor glucose (adapted from [56]).

The upper part of the pathway goes from G6P (glucose 6-phosphate) to UDP-GlcNAc and the lower part goes from G6P to UDP-GlcUA. The upper part consists of five enzymes to catalyze the five reactions from left to right in the figure: phosphoglucose isomerase (*GPI*), glutamine-fructose-6-phosphate transaminase (*GFPT*), glucosamine phosphate N-acetyltransferase (*GNPNAT*), phosphoacetyl glucosamine mutase (*PGM3*) and acetylglucosamine pyrophosphorylase (*UAP1*). The lower part consists of three enzymes for the three reaction steps from left to right: phosphoglucomutase (*PGM*), UDP-glucose

pyrophosphorylase (*UGP2*) and UDP-glucose dehydrogenase (*UGDH*). Three hyaluronic acid synthases (*HAS1-3*) are known to synthesize hyaluronic acid from one UDP-GlcNAc and one UDP-GlcUA.

Now we examine the conditions that can trigger the hyaluronic acid synthesis pathway. *GPI* is part of the glycolysis pathway and hence is activated whenever glycolysis is activated. The following three enzymes, *GNPNAT*, *PGM3* and *UAP1,* as part of the hexosamine pathway, can be activated when glucosamine is abundantly available and under hypoxia [57, 58]. *PGM* can be up-regulated by hypoxia [59], so are *UGP2* [60] and *GFPT* [58]. *UGDH* is positively regulated by *TGFβ* (transforming growth factor *β*) [61]. *HAS1*-3 can be activated by various growth factors such as *TGFβ*, *PDGF* (platelet derived growth factor), *KGF* (keratinocyte growth factor), *FGF2* (fibroblast growth factor 2), *EGF* (epidermal growth factor), *IL1β* (interleukin-1 *β*) and *TNFα* (tumor necrosis factor *α*) [56]. In addition, the level of UDP-GlcNAc has been found to control the expression of *HAS2* [56].

In sum, the upper part of the pathway will be activated when there is ample G6P under hypoxia; and the lower part will be activated under hypoxia and the availability of *TGFβ.* So the hyaluronic acid synthesis pathway can be activated by a condition with a plenty of G6P and availability of *TGFβ* under hypoxia. And hyaluronic acid synthases can be activated by *TGFβ* in conjunction with UDP-GlcNAc, the product of the upper part of the pathway. Clearly all these conditions are satisfied for an inflammatory tissue whose cells are accumulated with glucose metabolites under hypoxia, hence strongly suggesting the possibility that hyaluronic acid will be synthesized. It is worth noting that *TGFβ* is generally available in chronic inflammation [62, 63].

Note that the synthesis of hyaluronic acid is done through repeated addition of one glucuronic acid and one N-acetylglucosamine to the nascent polydisaccharide as the molecule is extruded via *ABC* transporters or hyaluronic acid synthase into the extracellular space [64, 65]. The exported hyaluronic acid will be degraded by some of the hyaluronidases (*HYAL1-6*) or by ROS into fragments of different sizes if it is not incorporated into extracellular matrices [66].

We have examined gene-expression data collected on multiple normal human cells (see Supplementary Material Sections G and H) treated with hypoxia between 3 and 48 hours, and found that hyaluronic acid

synthases (*HAS1*) and hyaluronidases (*HYAL1, 3-4*) are up-regulated. These observations are consistent with past studies showing that hypoxia affects hyaluronic acid synthesis and turnover [67]. In addition, we have examined a set of gene-expression data collected on a set of diseased colon tissue samples ranging from precancerous tissues to advanced adenocarcinoma (see Supplementary Material Section H), and found that the hyaluronic acid synthesis pathway is indeed activated in the very early stage of the disease, starting from inflammatory colon disease or at latest colon adenoma. Figure 3 shows a heat-map of the expression level changes of the relevant genes in precancerous/cancer tissues *versus* the normal controls.

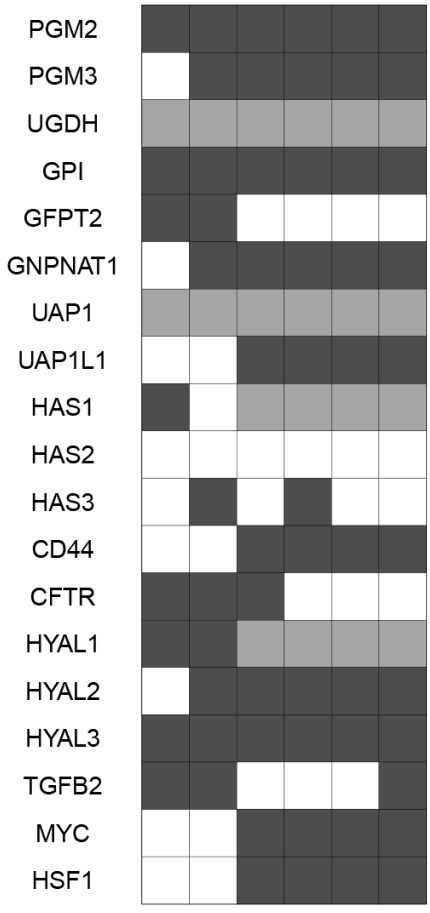


Figure 2.3: Differentially expressed hyaluronic acid metabolism related genes in inflammatory disease, colon adenoma, colon adenocarcinoma tissue samples of stage 1-4 (from left to right). Dark grey and light

grey are for significantly up and down regulation in the disease samples comparing to normal colon samples (by a moderated *t*-test using a significance cutoff level at 0.05[68]), respectively while white represents no significant changes.

From Figure 3, we can see that all genes in the hyaluronic acid synthesis pathway are up-regulated in precancerous tissues except for *UAP1* and *UGDH*. Interestingly, one of *UAP1*'s homologs *UAP1L1* is up-regulated. Hence we predict that *UAP1L1* is used for the same function. For *UGDH,* no homologous genes are found to be up-regulated but we suspect that another gene serving the same purpose is used here in its place since the rate-limiting factor gene *GFPT2* is up-regulated and hyaluronic acid is clearly being produced here since multiple hyaluronidase genes *HAYL 1-3* and one reported hyaluronic acid exporter gene *CFTR* (cystic fibrosis transmembrane conductance regulator) [69] are up-regulated. In addition, *TGFβ* is up-regulated in the precancerous stage; and its expression level goes back to the background level once the down-stream genes of hyaluronic acid synthesis such as *HSF1* (heat shock transcription factor) and *MYC* (v-myc avian myelocytomatosis viral oncogene homolog) are up-regulated, indicating that once the tissue becomes cancerous, cell proliferation will be driven by some factors other rather than hypoxia-induced hyaluronic acid production and fragmentation [70-72]. Very interestingly, highly similar expression patterns of these genes are observed in precancerous, early stage and advanced stage melanoma (see Supplementary Material Section I and Supplementary Figure 3). These are the only two gene-expression datasets covering both precancerous and cancer tissues at different stages we found on the Internet.

The above discussion and Figure 3 strongly suggest a pathway going from inflammation-induced hypoxia to cell proliferation: cells will accumulate glucose metabolites under hypoxia, which will ultimately leads to the synthesis and export of hyaluronic acid; the exported hyaluronic acid will be degraded into fragments. Since all short hyaluronic acid fragments with size ranging from 4 to 5000 disaccharides serve as signals for some aspects of tissue repair [29], the "right" combinations of fragments will be generated in time by some cells, just like the combinations of hyaluronic acid fragments generated

by a damaged ECM, hence starting the signaling process of tissue repair, including cell proliferation, cell survival and angiogenesis. Note that the fragmentation patterns of a damaged ECM should have a high degree of randomness since a tissue can be injured by different causes, hence possibly giving rise to different fragmentation patterns of hyaluronic acid, which all lead to the activation of the tissue repair system. Hence we assume that the probability that the hypoxia-induced hyaluronic acid fragmentations will trigger the tissue repair system is high.

In addition, hyaluronic acid on the cell surface is known to facilitate anchorage-independent proliferation [73, 74], hence bypassing the requirement for cells being connected to their ECM before they can proliferate; similarly it also facilitates the loss of the contact-inhibition constraint for cell proliferation [75]. Essentially hyaluronic acid and their fragments provide all the signals needed for cell proliferation via the utilization of "tissue repair" system. Interestingly several recent studies have shown that hypoxia can stimulate the synthesis of hyaluronic acid [67, 76], hence providing a strong support to our model.

It was recently shown that inhibition of tumor growth caused by the abnormally large hyaluronic acid produced in naked mole rat tissues is reversed by treatment with *HYAL2* [32]. Hence, we posit that hyaluronic acid turnover and the induced cell proliferation provide an exit for the accumulated glycolytic metabolites through stimulation of cell proliferation and subsequent utilization of the metabolites as building blocks for cell division. Our model is also consistent with a recent studies showing that elevated cell sugar concentrations can increase the production of hyaluronic acid [56].

Hypoxia induced signals facilitate additional signals for cell division

In addition to the hypoxia-induced activation of tissue repair system discussed above, persistent hypoxia can lead to the generation of other signals to assist cell proliferation for their survival (i.e., to provide an exit for the accumulated glucose metabolites). For example, hypoxia can lead to the accumulation of necrotic cells that release in-danger signals and trigger the production and release of growth signals such as *FGF2* (fibroblast growth factor 2) and *HDGF* (hepatoma-derived growth factor) [77, 78], hence providing additional signals to enhance the synthesis of hyaluronic acid (see the above section). In

addition, a resistance signal to apoptosis can also be triggered by the necrotic cells to maintain tissue integrity [78].

A recent study suggests that hypoxia may play more direct roles in mediating cell division. Specifically, the authors demonstrated that hypoxia can change the actin cytoskeletal organization, leading to morphological changes of the cells [79] and hence preparing the cells for division. The study also showed that hypoxia leads to increase in cell volume, which we suspect is partially due to the metabolite accumulation discussed in the earlier sections. In addition, previous studies have established that the state of actin filament organization directly controls cell-cycle progression [80, 81]. Furthermore, the level of hypoxia has long been linked to the level of malignancy of a cancer [23, 82]. By integrating all this information, we can see that hypoxia may play a direct and essential role in the modulation of cell division. If this proves to be true, hypoxia probably plays a double role in early carcinogenesis: creating the pressure for cells to evolve to remove the accumulated glucose derivatives and facilitating the removal of the accumulated derivatives through cell division. Clearly this warrants further investigation.

Hypoxia is also known to mediate a few other activities that may facilitate sustained cell survival, such as up-regulation of telomerase [83], genomic instability [84], angiogenesis [85] and cell migration [86], which will lead to more uncontrollable behaviors of the cells. Our recent transcriptomic analysis of precancerous and early stage cancer data strongly suggests that cellular hypoxia takes place no later than any of the cancer hallmark events (see Supplementary Material Section J), hence providing independent evidence in support of the theory we are proposing.

In addition, lactate, generated due to the increased glycolytic fermentation activity due to the reprogrammed energy metabolism, is known to play a number of key roles in driving carcinogenesis such as (a) promotion of chronic inflammation [87], (b) wound over-healing [88], (c) secretion of *VEGF* and tumor angiogenesis [88] and (d) immune escape [89], hence providing additional signals in support of cell proliferation as well as cell survival.

Genomic mutations and cancer development

While genetic mutations are believed by many to be a primary reason for sporadic cancer development, some recent studies start to challenge this view [90, 91]. Within the driver model proposed here, genetic mutations, we believe, may dominantly play a facilitator's role rather than a primary driver's role in sporadic cancer. Specifically we suggest that the change-of-function genetic mutations in many cancer related genes such as *P53* and *RAS* are selected as "permanent" replacements for on-going functions currently accomplished through regulation or other functional means, to facilitate cell division, hence survival, in a more sustained and more efficient manner. We use the following examples to illustrate the idea.

The functional form of *PKM2* (pyruvate kinase, muscle 2) is a homo-tetramer and it serves as a key enzyme in glycolysis, which catalyzes the conversion from phosphoenolpyruvate to pyruvate. It has been observed that the vast majority of advanced cancers have loss-of-function mutations in the *PKM2* gene [92]. This suggests that there is an evolutionary pressure for the affected cells to reduce their pyruvate production. It has been shown that oxidation of *PKM2* in specific locations by ROS can increase the possibility of disassociation of the tetramer to dimers or monomers [93], hence reducing its normal function. This suggests the possibility that *PKM2* may gradually lose its functions due to oxidation, and natural selection may have gradually selected such cells since the loss of *PKM2* function may give such cells a survival advantage. Loss-of-function mutations in the *PKM2* gene may be just a permanent replacement for losing the function of the protein currently achieved through function-losing oxidation.

Another example is the loss of the contact-inhibition capability of cells that can terminate cell division when they are in close proximity one to another [94]. The increased hyaluronic acid export can functionally diminish cellular contact-inhibition [75] as discussed earlier, which allows sustained division by the underlying cells even when they get close to each other. At the end, mutations in genes responsible for activating the contact inhibition mechanism, such as *ING4* (inhibitor of growth family, member 4) [95], are selected, thus ensuring the permanent loss of this inhibition capability, as observed in advanced cancers in general [96].

Knowing that the mutation rates in cancer related genes go up as a cancer advances [97], we suggest

that genetic mutations are permanent replacements for on-going functions (or their repression) executed through functional regulation or other means, is generally applicable to cancer development.

Hereditary cancer and the hypoxia-based driver model

While the above inflammation/hypoxia-based driver model is for sporadic cancers, it is natural to ask if this or a similar model may apply to hereditary cancers. To address this issue, we have examined seven major types of hereditary cancers with known key mutations: breast cancer due to *BRCA* (breast cancer, early onset) mutations, kidney cancer due to mutations in *FH* (fumarate hydratase), colon cancer due to *APC* (adenomatous polyposis coli) mutations, retinoblastoma due to *RB1* (retinoblastoma 1) mutations, Li-Fraumeni syndrome due to *P53* mutations, Cowden syndrome due to *PTEN* (phosphatase and tensin homolog) mutations and Von Hippel-Lindau syndrome due to *VHL* (Von Hippel-Lindau tumor suppressor) mutations. A large number of studies have been carried out, focused on linking these mutations to the development of the corresponding hereditary cancer [98-104]. However no general cancer initiation model has been proposed for these hereditary cancers, to the best of our knowledge. Our literature survey did reveal one commonality among all these loss-of-function mutations: they all result in increased production or accumulation of ROS. Potentially, this common role by loss-of-function mutations in the seven genes may prove to be the most essential role in the tumorigenesis of the relevant cancers. Details follow.

Recent studies have shown that *BRCA* mutations in normal breast cells can lead to the generation of hydrogen peroxide as well as increased glycolysis and decreased oxidative phosphorylation [105], revealing the repression of the mitochondrial activities, which forces cells to increase their activity of glycolytic fermentation regardless of being cancer or non-cancer cells.

Regarding *FH*, recent publications have shown that loss-of-function *FH* mutations lead to pseudo-hypoxia and increased ROS constitutively, which further leads to increased glycolysis and decreased oxidative phosphorylation [106], hence leading to induction of glycolytic fermentation pathway in time.

The loss-of-function mutations in the *APC* gene has been found to lead to constitutive activation of the *WNT* signaling pathway [107], which activates a downstream gene called *RAC1*, a GTPase. The activation

of *RAC1* has been shown to lead to ROS production [108]. Hence we posit that the same process leading to the reprogramming of the energy metabolism will take place in time like in the above cases.

The current understanding about the relation between *RB1* mutations and ROS production is that loss-of-function mutations in *RB1* leads to dysregulation of *E2F2*, a component of the transcription factor gene *E2F* involved in cell cycle regulation and DNA synthesis, which drives increased production of ROS [109].

P53 gene mutations have long been linked to the production of ROS [110]. For example, mutations in *P53* can interfere with the normal response of human cells to oxidative stress through attenuating the activation and function of *NFE2* (nuclear factor, erythroid 2) related factor 2, a transcription factor that induces antioxidant response. This effect is manifested by decreased expression of phase 2 detoxifying enzymes *NQO1* (NAD(P)H dehydrogenase, quinone 1) and *HMOX1* (heme oxygenase (decycling) 1) and increased ROS levels; hence the energy-metabolism reprogramming will take in time as above.

The relationship between *PTEN* mutations and ROS production is an interesting one. A recent study has found that mutations in the ATP-binding motif of *PTEN* lead to disruption of the correct subcellular localization of the protein, which results in a significantly decreased nuclear *P53* protein level and transcriptional activity, and enhanced production of ROS [111].

VHL-deficiency was recently found to constitutively activate *NOX* oxidases to maintain the *HIF2α* (hypoxia inducible factor-2α) protein expression while NADPH oxidases of the *NOX* family are the major sources of ROS [112]. Hence the same process of energy-metabolism reprogramming will take place in time.

The increased ROS accumulation associated with the above mutations, possibly in conjunction with some cell type-specific environment, will ultimately lead to repression of the mitochondrial activities, including oxidative phosphorylation, and further to energy-metabolism reprogramming as in the case of hypoxia-induced reprogramming of energy metabolism [113]. In addition, it has been well established that mitochondrial ROS triggers hypoxia-induced transcription [19] and inflammation [26]. Hence, we propose that the same cancer-driver model proposed earlier applies to these hereditary cancers except that the initial trigger is increased ROS instead of persistent hypoxia.

For the same reason, we suspect that aging-induced cancers may also follow this or a similar model as mitochondrial ROS accumulates and inflammatory cells increase, on top of cellular senescence [114] as one ages; and at some point these cells may repress their mitochondrial activities when high enough ROS levels have accumulated, leading to the reprogrammed energy metabolism and associated phenomena discussed above. We believe that the various observed changes in these hereditary cancers as reported in the literature [98-104] contributes to this general driver model proposed here.

Interpretation of naked mole rat data using our model

It is well known that naked mole rats do not or rarely develop cancer [14, 32]. We suspect that their reduced energy demand during hypoxia is one reason, as they will not accumulate as much glycolytic metabolites as in human, mouse and rat. A recent study [32] suggests that naked mole rats may have additional abilities to resist cancer development. This study has found that naked mole rats synthesize and export an unusually long form of hyaluronic acid, multiple times longer than that in human. It was observed that knockdown of the species-specific mutant form of *HAS2* or fragmentation of hyaluronic acid by *HYAL2* leads to cancer development from naked mole rat cells, which otherwise failed to form tumors under the same conditions.

Our model provides a natural and logical explanation of their observation. Specifically, we believe that naked mole rats have evolved systems to move the *HAS2* synthesized very long hyaluronic acid, hence with no overlap with the short hyaluronic acid as signaling molecules for tissue repair, into the circulation without being degraded, which may ultimately get deposited to their skins. The reason that blocking the activation of *HAS2* will lead to cancer development is, we speculate, that it will trigger alternative ways to synthesize hyaluronic acid, such as by *HAS3*, which tend to be much shorter, hence possibly directly serving as signaling molecules. The same can be said about the activation of the degradation enzymes of hyaluronic acid, namely hyaluronidases, which will of course lead to the generation of tissue repair signals. Figure 4 shows a comparison of how human and naked mole rat cells handle the excessive production of glycolytic metabolites.

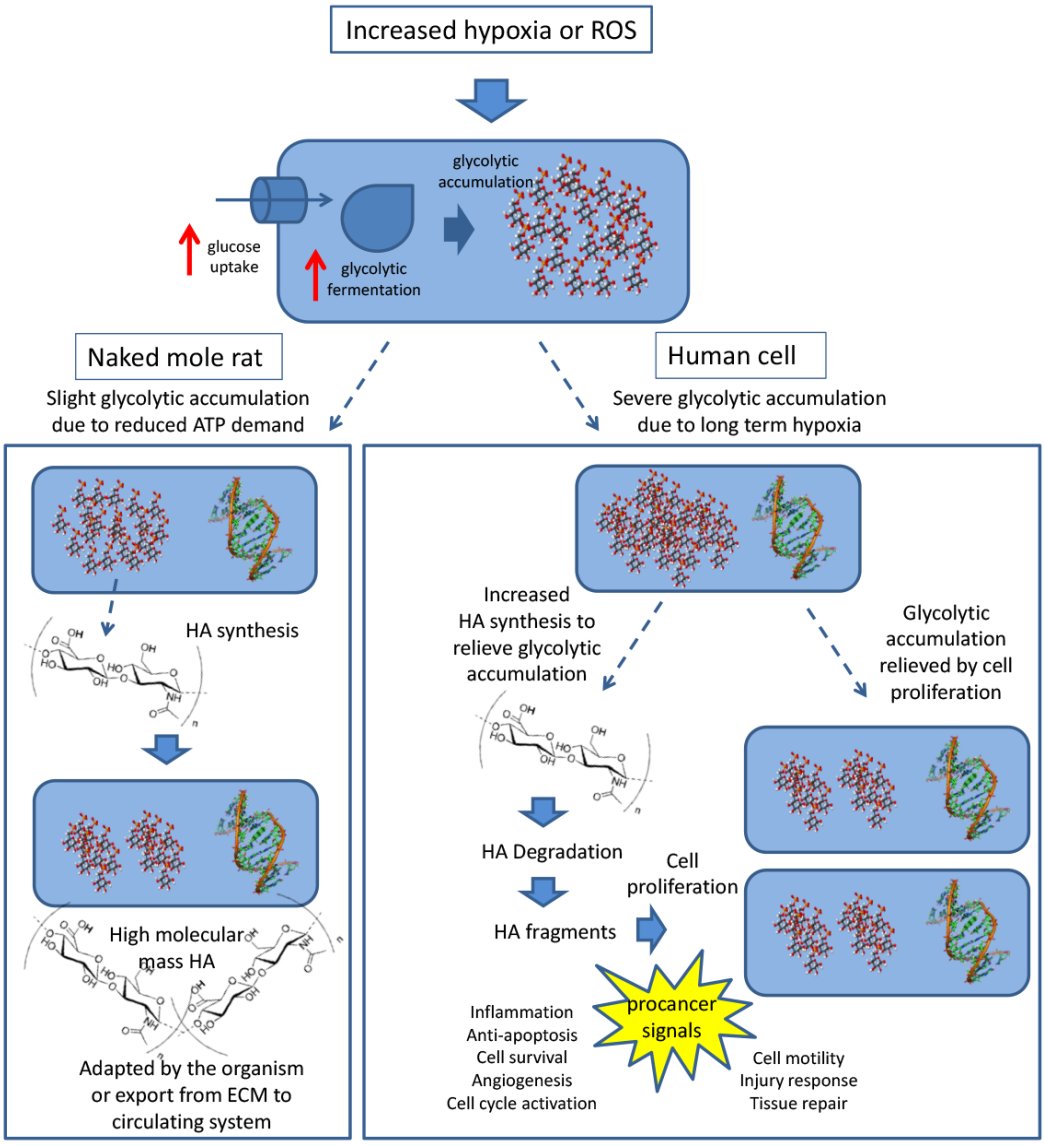


Figure 2.4: A driver model for the early phase of human carcinogenesis *versus* the exit pathway of the excess glucose derivatives in naked mole rats.

Conclusions

It is the accumulation of glycolytic metabolites that puts cells on their way to becoming cancerous under chronic hypoxia and/or increased ROS conditions, coupled with chronic inflammatory condition. The accumulation is the direct result of the energy-metabolism reprogramming as Otto Warburg speculated five decade ago. The pressure for survival casted on the underlying cells with the glucose metabolite accumulation has clearly led the cells to create the conditions that can trigger the synthesis, export and degradation of hyaluronic acid chains and hence essentially make the whole tissue-repair system available to the affected cells for their survival through cell division since this provides an exit for the accumulated glucose metabolites. If the hypoxic and/or ROS condition persists, this process will continue, hence cell proliferation on a continuous basis. Along the way, some genetic mutations may be selected to provide permanent replacement for various on-going functions to make the proliferation and hence survival more sustainable and possibly more efficient, particularly knowing that hyaluronic acid can directly activate a number of proto-oncogenes such as *HSF1* and *MYC* [70-72].

While energy-metabolism reprogramming may be the key reason for a cancer to start, two “limitations” in our cellular systems shaped by the past evolution may be the fundamental reason of why human can develop cancer while organisms like blind mole rats or turtle do not: (1) the ATP demand could not drop to the level of ATP supply when energy-metabolism reprogramming takes place; and (2) there is an intrinsic mismatch between the increased glucose influx triggered by the ATP-deficiency and the maximum flux of the pathway, both of which are due to the lack of “training” in the past. Overall, our model proposes a molecular level mechanism for how energy-metabolism reprogramming will lead to continuous cell proliferation for survival, i.e., the initiation of a cancer, hence providing an explanation of Warburg’s speculation in 1960s: "Cancer … has countless secondary causes; but there is only one prime cause, (which) is the replacement of respiration of oxygen in normal body cells by a fermentation of sugar".

Acknowledgement

The authors want to thank Professor J. David Puett of the University of Georgia, Professor Yusuf Hannun of the SUNY at Stony Brook and Professor Kenneth B. Storey of the Carleton University for their insightful comments and suggestions. YX thanks Georgia Research Alliance and the University of Georgia Research Foundation for the Endowment provided to his chair position.

References

1. Warburg, O., *On the Origin of Cancer Cells.* Science, 1956. **123**(3191): p. 309-314.
2. Munoz, N., et al., *Epidemiologic classification of human papillomavirus types associated with cervical cancer.* N Engl J Med, 2003. **348**(6): p. 518-27.
3. Perz, J.F., et al., *The contributions of hepatitis B virus and hepatitis C virus infections to cirrhosis and primary liver cancer worldwide*. J Hepatol, 2006. **45**(4): p. 529-38.
4. DeBerardinis, R.J., et al., *The biology of cancer: metabolic reprogramming fuels cell growth and proliferation.* Cell Metab, 2008. **7**(1): p. 11-20.
5. Vander Heiden, M.G., L.C. Cantley, and C.B. Thompson, *Understanding the Warburg effect: the metabolic requirements of cell proliferation.* Science, 2009. **324**(5930): p. 1029-33.
6. Stehelin, D., et al., *DNA related to the transforming gene(s) of avian sarcoma viruses is present in normal avian DNA.* nature, 1976. **260**: p. 170-173.
7. Knudson, A.G., *Mutation and Cancer: Statistical Study of Retinoblastoma*. Proc Natl Acad Sci U S A., 1971. **68**(4): p. 820-823.
8. Nishisho, I., et al., *Mutations of chromosome 5q21 genes in FAP and colorectal cancer patients.*

Science, 1991. **2253**(5020): p. 665-669.

9. Nowell, P.C. and D.A. Hungerford, *A minute chromosome in human chronic granulocytic leukemia.* Science, 1960. **142**(1497).

10. Greenman, C., et al., *Patterns of somatic mutation in human cancer genomes.* Nature, 2007. **446**(7132): p. 153-8.

11. Zhu, X. and R.K. Assoian, *Integrin-dependent activation of MAP kinase: a link to shape-dependent cell proliferation.* Mol Biol Cell, 1995. **6**(3): p. 273-82.

12. Senoo, H. and R. Hata, *Extracellular matrix regulates cell morphology, proliferation, and tissue formation.* Kaibogaku Zasshi, 1994. **69**(6): p. 719-33.

13. Wells, R.G., *The role of matrix stiffness in regulating cell behavior.* Hepatology, 2008. **47**(4): p. 1394-400.

14. Seluanov, A., et al., *Hypersensitivity to contact inhibition provides a clue to cancer resistance of naked mole-rat.* Proc Natl Acad Sci U S A, 2009. **106**(46): p. 19352-7.

15. Hanahan, D. and R.A. Weinberg, *Hallmarks of cancer: the next generation.* Cell, 2011. **144**(5): p. 646-74.

16. Huang, S. and D.E. Ingber, *Shape-dependent control of cell growth, differentiation, and apoptosis: switching between attractors in cell regulatory networks.* Exp Cell Res, 2000. **261**(1): p. 91-103.

17. Beachy, P.A., S.S. Karhadkar, and D.M. Berman, *Tissue repair and stem cell renewal in carcinogenesis.* Nature, 2004. **432**(7015): p. 324-31.

18. Akagi, K., et al., *Genome-wide analysis of HPV integration in human cancers reveals recurrent,*

focal genomic instability. Genome Res, 2014. **24**(2): p. 185-99.

19. Chandel, N.S., et al., *Mitochondrial reactive oxygen species trigger hypoxia-induced transcription.* Proc Natl Acad Sci U S A, 1998. **95**(20): p. 11715-20.

20. Bartrons, R. and J. Caro, *Hypoxia, glucose metabolism and the Warburg's effect.* J Bioenerg Biomembr, 2007. **39**(3): p. 223-9.

21. Elinav, E., et al., *Inflammation-induced cancer: crosstalk between tumours, immune cells and microorganisms.* Nat Rev Cancer, 2013. **13**(11): p. 759-71.

22. Shacter, E. and S.A. Weitzman, *Chronic inflammation and cancer.* Oncology (Williston Park), 2002. **16**(2): p. 217-26, 229; discussion 230-2.

23. Vaupel, P. and A. Mayer, *Hypoxia in cancer: significance and impact on clinical outcome.* Cancer Metastasis Rev. , 2007. **26**(2): p. 225-39.

24. Eltzschig, H.K. and P. Carmeliet, *Hypoxia and inflammation.* N Engl J Med, 2011. **364**(7): p. 656-65.

25. Bartels, K., A. Grenz, and H.K. Eltzschig, *Hypoxia and inflammation are two sides of the same coin.* Proc Natl Acad Sci U S A, 2013. **110**(46): p. 18351-2.

26. Gupta, S.C., et al., *Upsides and downsides of reactive oxygen species for cancer: the roles of reactive oxygen species in tumorigenesis, prevention, and therapy.* Antioxid Redox Signal, 2012. **16**(11): p. 1295-322.

27. Frezza, C., et al., *Metabolic profiling of hypoxic cells revealed a catabolic signature required for cell survival.* PLoS One, 2011. **6**(9): p. e24411.

28. Yevdokimova, N.Y., *Elevated level of ambient glucose stimulates the synthesis of high-molecular-*

weight hyaluronic acid by human mesangial cells. The involvement of transforming growth factor beta1 and its activation by thrombospondin-1. Acta Biochim Pol, 2006. **53**(2): p. 383-93.

29. Stern, R., A.A. Asari, and K.N. Sugahara, *Hyaluronan fragments: an information-rich system.* Eur J Cell Biol, 2006. **85**(8): p. 699-715.

30. Ruben, L.N., R.H. Clothier, and M. Balls, *Cancer resistance in amphibians.* Altern Lab Anim, 2007. **35**(5): p. 463-70.

31. Manov, I., et al., *Pronounced cancer resistance in a subterranean rodent, the blind mole-rat, Spalax: in vivo and in vitro evidence.* BMC Biol, 2013. **11**: p. 91.

32. Tian, X., et al., *High-molecular-mass hyaluronan mediates the cancer resistance of the naked mole rat.* Nature, 2013. **499**(7458): p. 346-9.

33. Krivoruchko, A. and K.B. Storey, *Forever young: mechanisms of natural anoxia tolerance and potential links to longevity.* Oxid Med Cell Longev, 2010. **3**(3): p. 186-98.

34. Caulin, A.F. and C.C. Maley, *Peto's Paradox: evolution's prescription for cancer prevention.* Trends Ecol Evol, 2011. **26**(4): p. 175-82.

35. Rolfe, D. and G. Brown, *Cellular energy utilization and molecular origin of standard metabolic rate in mammals.* PHYSIOLOGICAL REVIEWS, 1997. **77**(3): p. 731-758.

36. Buttgereit, F. and M.D. Brand, *A hierarchy of ATP-consuming processes in mammalian cells.* Biochem J, 1995. **312 (Pt 1)**: p. 163-7.

37. Nathaniel, T.I., et al., *Effect of hypoxia on metabolic rate, core body temperature, and c-fos expression in the naked mole rat.* Int J Dev Neurosci, 2012. **30**(6): p. 539-44.

38. Hochachka, P.W., et al., *Unifying theory of hypoxia tolerance: molecular/metabolic defense and*

rescue mechanisms for surviving oxygen lack. Proc Natl Acad Sci U S A, 1996. **93**(18): p. 9493-8.

39. St-Pierre, J., M.D. Brand, and R.G. Boutilier, *The effect of metabolic depression on proton leak rate in mitochondria from hibernating frogs.* J Exp Biol, 2000. **203**(Pt 9): p. 1469-76.

40. Kim, E.B., et al., *Genome sequencing reveals insights into physiology and longevity of the naked mole rat.* Nature, 2011. **479**(7372): p. 223-7.

41. Azpurua, J. and A. Seluanov, *Long-lived cancer-resistant rodents as new model species for cancer research.* Front Genet, 2012. **3**: p. 319.

42. Shams, I., A. Avivi, and E. Nevo, *Hypoxic stress tolerance of the blind subterranean mole rat: expression of erythropoietin and hypoxia-inducible factor 1 alpha.* Proc Natl Acad Sci U S A, 2004. **101**(26): p. 9698-703.

43. Kano, M., et al., *A meta-clustering analysis indicates distinct pattern alteration between two series of gene expression profiles for induced ischemic tolerance in rats.* Physiol Genomics, 2005. **21**(2): p. 274-83.

44. Gesser, H., K. Johansen, and G.M. Maloiy, *Tissue metabolism and enzyme activities in the rodent Heterocephalus glaber, a poor temperature regulator.* Comp Biochem Physiol B, 1977. **57**(4): p. 293-6.

45. Schwanhausser, B., et al., *Global quantification of mammalian gene expression control.* Nature, 2011. **473**(7347): p. 337-42.

46. Widmer, H.R., et al., *Working underground: respiratory adaptations in the blind mole rat.* Proc Natl Acad Sci U S A, 1997. **94**(5): p. 2062-7.

47. Edrey, Y.H., et al., *Endocrine function and neurobiology of the longest-living rodent, the naked*

mole-rat. Exp Gerontol, 2011. **46**(2-3): p. 116-23.

48. Bickler, P.E. and L.T. Buck, *Hypoxia tolerance in reptiles, amphibians, and fishes: life with variable oxygen availability.* Annu Rev Physiol, 2007. **69**: p. 145-70.

49. Ramirez, J., L. Folkow, and A. Blix, *Hypoxia Tolerance in Mammals and Birds: From the Wilderness to the Clinic.* Annu Rev Physiol., 2007. **69**: p. 113-43.

50. Kubasiak, L.A., et al., *Hypoxia and acidosis activate cardiac myocyte death through the Bcl-2 family protein BNIP3.* Proc Natl Acad Sci U S A., 2002. **99**(20): p. 12825-12830.

51. Schaffer, J., *Lipotoxicity: when tissues overeat.* Curr Opin Lipidol., 2003. **14**(3): p. 281-7.

52. Stern, R., *Hyaluronan in cancer biology*. 1st ed. 2009, San Diego, CA: Academic Press/Elsevier. xxvii, 426 p., 12 p. of plates.

53. Toole, B., *Hyaluronan-CD44 Interactions in Cancer: Paradoxes and Possibilities.* Clin Cancer Res, 2009. **15**(24): p. 7462-7468.

54. Sironen, R., et al., *Hyaluronan in human malignancies.* Exp Cell Res, 2011. **317**(4): p. 383-91.

55. Jiang, D., J. Liang, and P.W. Noble, *Hyaluronan in tissue injury and repair.* Annu Rev Cell Dev Biol, 2007. **23**: p. 435-61.

56. Tammi, R.H., et al., *Transcriptional and post-translational regulation of hyaluronan synthesis.* FEBS J, 2011. **278**(9): p. 1419-28.

57. Fantus, I.G., et al., *The Hexosamine Biosynthesis Pathway.* Contemporary Diabetes, 2006. **The Diabetic Kidney**: p. 117-133.

58. Guillaumond, F., et al., *Strengthened glycolysis under hypoxia supports tumor symbiosis and hexosamine biosynthesis in pancreatic adenocarcinoma.* Proc Natl Acad Sci U S A, 2013.

110(10): p. 3919-24.

59. Pelletier, J., et al., *Glycogen Synthesis is Induced in Hypoxia by the Hypoxia-Inducible Factor and Promotes Cancer Cell Survival.* Front Oncol, 2012. **2**: p. 18.

60. Pescador, N., et al., *Hypoxia promotes glycogen accumulation through hypoxia inducible factor (HIF)-mediated induction of glycogen synthase 1.* PLoS One, 2010. **5**(3): p. e9644.

61. Bontemps, Y., et al., *Specific protein-1 is a universal regulator of UDP-glucose dehydrogenase expression: its positive involvement in transforming growth factor-beta signaling and inhibition in hypoxia.* J Biol Chem, 2003. **278**(24): p. 21566-75.

62. Melillo, G., *Hypoxia: jump-starting inflammation.* Blood, 2011. **117**(9): p. 2561-2.

63. Wahl, S.M., *Transforming growth factor beta (TGF-beta) in inflammation: a cause and a cure.* J Clin Immunol, 1992. **12**(2): p. 61-74.

64. Medina, A.P., J. Lin, and P.H. Weigel, *Hyaluronan synthase mediates dye translocation across liposomal membranes.* BMC Biochem, 2012. **13**: p. 2.

65. Schulz, T., U. Schumacher, and P. Prehm, *Hyaluronan export by the ABC transporter MRP5 and its modulation by intracellular cGMP.* J Biol Chem, 2007. **282**(29): p. 20999-1004.

66. Agren, U.M., R.H. Tammi, and M.I. Tammi, *Reactive oxygen species contribute to epidermal hyaluronan catabolism in human skin organ culture.* Free Radic Biol Med, 1997. **23**(7): p. 996-1001.

67. Gao, F., et al., *Hypoxia-induced alterations in hyaluronan and hyaluronidase.* Adv Exp Med Biol., 2005. **566**: p. 249-56.

68. Tusher, V.G., R. Tibshirani, and G. Chu, *Significance analysis of microarrays applied to the*

ionizing radiation response. Proc Natl Acad Sci U S A, 2001. **98**(9): p. 5116-21.

69. Schulz, T., et al., *Cystic fibrosis transmembrane conductance regulator can export hyaluronan.* Pathobiology, 2010. **77**(4): p. 200-9.

70. Xu, H., et al., *Effect of hyaluronan oligosaccharides on the expression of heat shock protein 72.* J Biol Chem, 2002. **277**(19): p. 17308-14.

71. Tsatas, D., et al., *EGF receptor modifies cellular responses to hyaluronan in glioblastoma cell lines.* J Clin Neurosci, 2002. **9**(3): p. 282-8.

72. Savani, R.C., et al., *Increased inflammation, hyaluronan & respiratory distress in mice overexpressing the hyaluronan receptor RHAMM in macrophages.* The FASEB Journal, 2007. **21**: p. 374.8.

73. Kosaki, R., K. Watanabe, and Y. Yamaguchi, *Overproduction of hyaluronan by expression of the hyaluronan synthase Has2 enhances anchorage-independent growth and tumorigenicity.* Cancer Res, 1999. **59**(5): p. 1141-5.

74. Toole, B.P., *Hyaluronan promotes the malignant phenotype.* Glycobiology, 2002. **12**(3): p. 37R-42R.

75. Itano, N., et al., *Abnormal accumulation of hyaluronan matrix diminishes contact inhibition of cell growth and promotes cell migration.* Proc Natl Acad Sci U S A, 2002. **99**(6): p. 3609-14.

76. Hashimoto, K., et al., *Hypoxia-induced hyaluronan synthesis by articular chondrocytes: the role of nitric oxide.* Inflamm Res, 2006. **55**(2): p. 72-7.

77. Stenfeldt, A.-L. and C. Wennerås, *Danger signals derived from stressed and necrotic epithelial cells activate human eosinophils.* Immunology, 2004. **112**(4): p. 605-614.

78. Zong, W.-X. and C.B. Thompson, *Necrotic death as a cell fate.* Genes And Development, 2006. **20**: p. 1-15.

79. Kayyali, U.S., et al., *Cytoskeletal Changes in Hypoxic Pulmonary Endothelial Cells Are Dependent on MAPK-activated Protein Kinase MK2.* J. Biol. Chem., 2002. **277**: p. 42596-42602.

80. Assoian, R.K. and X. Zhu, *Cell anchorage and the cytoskeleton as partners in growth factor dependent cell cycle progression.* Current Opinion in Cell Biology, 1997. **9**(1): p. 93-98.

81. Théry, M. and M. Bornens, *Cell shape and cell division.* urr. Opin. Cell Biol. , 2006. **18**(6): p. 648-657.

82. Bedogni, B., et al., *The hypoxic microenvironment of the skin contributes to Akt-mediated melanocyte transformation.* Cancer Cell, 2005. **8**(6): p. 443-54.

83. Nishi, H., et al., *Hypoxia-Inducible Factor 1 Mediates Upregulation of Telomerase (hTERT).* Mol. Cell. Biol., 2004. **24**(13): p. 6076-6083.

84. Bristow, R. and R. Hill, *Hypoxia and metabolism. Hypoxia, DNA repair and genetic instability.* Nat Rev Cancer, 2008. **8**(3): p. 180-92.

85. Pugh, C.W. and P.J. Ratcliffe, *Regulation of angiogenesis by hypoxia: role of the HIF system.* Nature Medicine, 2003. **9**: p. 677 - 684.

86. Fujiwara, S., et al., *Silencing hypoxia-inducible factor-1alpha inhibits cell migration and invasion under hypoxic environment in malignant gliomas.* Int J Oncol. , 2007. **30**(4): p. 793-802.

87. Yabu, M., et al., *IL-23-dependent and -independent enhancement pathways of IL-17A production by lactic acid.* Int. Immunol. , 2011. **23**(1): p. 29-41.

88. Beckert, S., et al., *Lactate stimulates endothelial cell migration.* Wound Repair and Regeneration,

2006. **14**: p. 321-324.

89. Hirschhaeuser, F., U.G.A. Sattler, and W. Mueller-Klieser, *Lactate: A Metabolic Key Player in Cancer.* Cancer Research, 2011. **71**: p. 6921-6925.

90. Satgé, D., *Analysis of Somatic Mutations in Cancer Tissues Challenges the Somatic Mutation Theory of Cancer.* eLS, 2013.

91. Soto, A.M. and C. Sonnenschein, *The somatic mutation theory of cancer: growing problems with the paradigm?* Bioessays, 2004. **26**(10): p. 1097-107.

92. Mazurek, S., et al., *Pyruvate kinase type M2 and its role in tumor growth and spreading.* Semin Cancer Biol., 2005. **15**(4): p. 300-308.

93. Anastasiou, D., et al., *Inhibition of pyruvate kinase M2 by reactive oxygen species contributes to cellular antioxidant responses.* Science, 2011. **334**(6060): p. 1278-1283.

94. Sgambato, A., et al., *Multiple functions of p27Kip1 and its alterations in tumor cells: a review.* J. Cell. Physiol., 2000. **183**(1): p. 18-27.

95. Kim, S., et al., *A screen for genes that suppress loss of contact inhibition: identification of ING4 as a candidate tumor suppressor gene in human cancer.* Proc Natl Acad Sci U S A, 2004. **101**(46): p. 16251-6.

96. Kim, S., A.L. Welm, and J.M. Bishop, *A dominant mutant allele of the ING4 tumor suppressor found in human cancer cells exacerbates MYC-initiated mouse mammary tumorigenesis.* Cancer Res, 2010. **70**(12): p. 5155-62.

97. McFarland, C.D., et al., *Impact of deleterious passenger mutations on cancer progression.* Proc Natl Acad Sci U S A, 2013. **110**(8): p. 2910-5.

98. Easton, D.F., D. Ford, and D.T. Bishop, *Breast and ovarian cancer incidence in BRCA1-mutation carriers. Breast Cancer Linkage Consortium.* Am J Hum Genet, 1995. **56**(1): p. 265-71.

99. Toro, J.R., et al., *Mutations in the fumarate hydratase gene cause hereditary leiomyomatosis and renal cell cancer in families in North America.* Am J Hum Genet, 2003. **73**(1): p. 95-106.

100. Morin, P.J., et al., *Activation of beta-catenin-Tcf signaling in colon cancer by mutations in beta-catenin or APC.* Science, 1997. **275**(5307): p. 1787-90.

101. Hogg, A., et al., *Detection of heterozygous mutations in the RB1 gene in retinoblastoma patients using single-strand conformation polymorphism analysis and polymerase chain reaction sequencing.* Oncogene, 1992. **7**(7): p. 1445-51.

102. Malkin, D., et al., *Germ line p53 mutations in a familial syndrome of breast cancer, sarcomas, and other neoplasms.* Science, 1990. **250**(4985): p. 1233-8.

103. Liaw, D., et al., *Germline mutations of the PTEN gene in Cowden disease, an inherited breast and thyroid cancer syndrome.* Nat Genet, 1997. **16**(1): p. 64-7.

104. Iliopoulos, O., et al., *Tumour suppression by the human von Hippel-Lindau gene product.* Nat Med, 1995. **1**(8): p. 822-6.

105. Martinez-Outschoorn, U., et al., *BRCA1 mutations drive oxidative stress and glycolysis in the tumor microenvironment: implications for breast cancer prevention with antioxidant therapies.* Cell Cycle., 2012. **11**(23): p. 4402-13.

106. Sudarshan, S., et al., *Fumarate Hydratase Deficiency in Renal Cancer Induces Glycolytic Addiction and Hypoxia-Inducible Transcription Factor 1α Stabilization by Glucose-Dependent Generation of Reactive Oxygen Species.* Mol Cell Biol, 2009. **29**(15): p. 4080-4090.

107. Sunaga, N., et al., *Constitutive activation of the Wnt signaling pathway by CTNNB1 (beta-catenin) mutations in a subset of human lung adenocarcinoma.* Genes Chromosomes Cancer., 2001. **30**(3): p. 316-21.

108. Myant, K.B., et al., *ROS Production and NF-κB Activation Triggered by RAC1 Facilitate WNT-Driven Intestinal Stem Cell Proliferation and Colorectal Cancer Initiation.* Cell Stem Cell, 2013. **12**(6): p. 761-773.

109. Bremner, R. and E. Zacksenhaus, *Cyclins, Cdks, E2f, Skp2, and More at the First International RB Tumor Suppressor Meeting.* Cancer Res, 2010. **70**(15): p. 6114-8.

110. Jain, A.K., D.A. Bloom, and A.K. Jaiswal, *Nuclear import and export signals in control of Nrf2.* J Biol Chem, 2005. **280**(32): p. 29158-68.

111. He, X., et al., *Naturally occurring germline and tumor-associated mutations within the ATP-binding motifs of PTEN lead to oxidative damage of DNA associated with decreased nuclear p53.* Hum Mol Genet., 2011. **20**(1): p. 80-9.

112. Block, K., et al., *The NADPH Oxidase Subunit p22phox Inhibits the Function of the Tumor Suppressor Protein Tuberin.* Am J Pathol, 2010. **176**(5): p. 2447-2455.

113. Mauro, C., et al., *NF-kappaB controls energy homeostasis and metabolic adaptation by upregulating mitochondrial respiration.* Nat Cell Biol, 2011. **13**(10): p. 1272-9.

114. Campisi, J., et al., *Cellular senescence: a link between cancer and age-related degenerative disease?* Semin Cancer Biol. , 2011. **21**(6): p. 354-9.

CHAPTER III

POPULATION DYNAMICS INSIDE CANCER BIOMASS DRIVEN BY REPEATED HYPOXIA-REOXYGENATION CYCLES

Chi Zhang, Sha Cao, and Ying Xu. 2014. Quantitative Biology. Volume 2, Issue 3, pp 85-99

Reprinted here with permission of the publisher.

This is a pre-copyedited, author-produced PDF of an article accepted for publication in "Quantitative Biology" following peer review. The version of record is available online at: http://link.springer.com/article/10.1007%2Fs40484-014-0032-8

Abstract

A computational analysis of genome-scale transcriptomic data collected on ~1,700 tissue samples of three cancer types: breast carcinoma, colon adenocarcinoma and lung adenocarcinoma, revealed that each tissue consists of (at least) two major subpopulations of cancer cells with different capabilities to handle fluctuating O_2 levels. The two populations have distinct genomic and transcriptomic characteristics, one accelerating its proliferation under hypoxic conditions and the other proliferating faster with higher O_2 levels, referred to as the *hypoxia* and the *reoxygenation* subpopulations, respectively. The proportions of the two subpopulations within a cancer tissue change as the average O_2 level changes. They both contribute to cancer development but in a complementary manner. The hypoxia subpopulation tends to have higher proliferation rates than the reoxygenation one as well as higher apoptosis rates; and it is largely responsible for the acidic environment that enables tissue invasion and provides protection against attacks from T-cells. In comparison, the reoxygenation subpopulation generates new extracellular matrices in support of further growth of the tumor and strengthens cell-cell adhesion to provide scaffolds to keep all the cells connected. This subpopulation also serves as the major source of growth factors for tissue growth. These data and observations strongly suggest that these two major subpopulations within each tumor work together in a conjugative relationship to allow the tumor to overcome stresses associated with the constantly changing O_2 level due to repeated growth and angiogenesis. The analysis results not only reveal new insights about the population dynamics within a tumor but also have implications to our understanding of possible causes of different cancer phenotypes such as diffused *versus* more tightly connected tumor tissues.

Introduction

It is well established that malignant tumor tissues have heterogeneous cell populations, determined based on genomic mutation data [1-3] and morphological information of cells within a tumor [4]. However, very little is known about (i) the typical composition of the subpopulations, if there is such a thing, within a cancer tissue; (ii) how the subpopulations change throughout the development of a cancer; and (iii) if different subpopulations may contribute in different ways to the development of a cancer. Gaining a detailed understanding about these issues will not only offer new insights about the mechanisms of a cancer's development but also potentially lead to improved capabilities for treating the disease.

The key biological question addressed here is: how do cancer cells cope with the fluctuation in the O_2 level, as a result of repeated cycles of hypoxia and reoxygenation [6] caused by tumor growth and angiogenesis throughout the entire development of a cancer. A natural hypothesis is to test if each tumor may consist of different subpopulations with different capabilities to cope with the fluctuating O_2 levels.

Detection and characterization of distinct subpopulations within a cancer tissue experimentally represent a very challenging problem. While single-cell genome sequencing can provide some information about different cell types within a tissue [3-5], it is currently not feasible to derive subpopulation-level information within a tissue and mention how such subpopulation composition changes with time. Fortunately, the availability of genome-scale omic data collected on multiple cancer tissues has made it possible to tackle this problem using computational techniques.

We have recently carried out a computational study to infer and characterize the major subpopulations in terms of their capabilities to deal with the constantly changing O_2 levels within each tissue sample of three cancer types: breast, colon and lung cancers, which all have large numbers of both transcriptomic and matching genomic data in the public domain. Specifically by de-convoluting the gene-expression data of hypoxia and angiogenesis related marker genes collected on tissue samples into contributions by individual subpopulations each consisting of cells with similar expression patterns, we discovered that all tissue samples under consideration each consist of at least three substantial and distinct subpopulations.

One subpopulation increases its fraction when the tissue becomes more hypoxic, referred to as the *hypoxia* population while another increases its proportion as the tissue has a higher O_2 level, named the *reoxygenation* subpopulation. The third subpopulation is one with relatively longer hypoxia and reoxygenation cycles, suggesting slower growth. Of all three cancer types, the fraction of the hypoxia subpopulation within a cancer tissue generally increases as the underlying cancer advances. We focus on the first two subpopulations in this study.

A number of distinct characteristics are revealed about the two subpopulations through our analyses, which are consistent across all the samples of the three cancer types. Specifically, the hypoxia subpopulation has substantially higher genomic mutation rates than the reoxygenation one. It also grows faster and relies more heavily on glycolysis than the reoxygenation population. In comparison, the reoxygenation subpopulation has cell-cell and cell-ECM adhesion genes as well as growth factors substantially over-expressed.

To the best of our knowledge, all the results reported here represent novel discoveries about the subpopulation composition and dynamics within cancer tumors throughout their development, which may have important implications to development of improved treatments for cancers.

Results

Transcriptomic data analyses of three cancer types: breast invasive carcinoma (BRCA), colon adenocarcinoma (COAD) and lung adenocarcinoma (LUAD), have been carried out to gain a detailed understanding about the subpopulation composition and dynamics within cancer tissues as they progress. It is generally understood that a cancer tumor goes through multiple rounds of hypoxia and reoxygenation due to tumor growth and angiogenesis throughout a cancer's development. Specifically as a tumor grows, its cells tend to become increasingly more hypoxic; once the level of hypoxia exceeds certain thresholds, cells start to release signals for angiogenesis [8, 9], which can lead to increased blood flow and hence oxygen into the tissue. Then the cells will go through a growth phase under relatively normoxic conditions before they become hypoxic again, and repeat this cycle. In this paper, we examine how the

repeated hypoxia-reoxygenation cycle affects the evolution of cancer subpopulations with distinct characteristics.

Biclustering analysis of transcriptomic data reveals sizes of subpopulations change in accordance with O_2 levels

We have carried out a biclustering analysis of gene-expression data collected on tissue samples of three cancer types, based on 242 marker genes related to hypoxia, angiogenesis signaling, angiogenesis process, markers of endothelial cells (blood vessel cells), and cell proliferation (see Supplementary Table S1 for the gene names). Our goal is to determine if some substantial subsets of the tissue samples of each cancer type may share similar expression patterns among some (to be determined) subsets of the 242 genes. Our hypothesis is that samples in the same phase of the hypoxia-reoxygenation cycle may share similar expression patterns of some hypoxia and reoxygenation related genes. If true, a biclustering analysis over all the samples of the same cancer type vs the 242 genes should be able to detect such subsets of samples sharing similar expression patterns. Our in-house software QUBIC [10] is used for the analysis, where all samples of each cancer type are represented as a two dimensional matrix with each row representing one of the 242 genes, each column representing one sample and each entry being the expression level of the corresponding gene in the relevant sample (see METHODS for details).

Four substantial biclusters are observed consistently across all three cancer types with statistical significance < p=1E-50, given by the biclustering method. A permutation test is conducted, which supports the identified biclusters with the estimated statistical significance levels <p=1E-5 (see Supplementary Figure S2A). These biclusters cover at least 40% samples for each cancer type. Further analyses of the expression patterns and functions of the involved genes in the four clusters revealed that one cluster of samples generally have hypoxia-associated genes up-regulated, hence referred to as the *hypoxia* cluster; another cluster tends to have angiogenesis-signal receptor and endothelial marker genes up-regulated, hence called the *reoxygenation* cluster; and the third cluster shows up-regulation in both hypoxia-associated and angiogenesis-signaling genes, referred to as the *angiogenesis* cluster while the

fourth cluster shows no expression changes in any substantial subset of the 242 genes with statistical significance. We speculate that the samples in the fourth cluster may have relatively long hypoxia-reoxygenation cycles, i.e., being slow growing tumors, rather than no such cycles since it is known that all cancers go through angiogenesis; we call these samples as the *background* group. Supplementary Figure S2B illustrates the biclusters identified in each cancer type.

The above result is consistent with our general understanding about the repeated hypoxia-reoxygenation cycles that a tumor tends to go through as it advances [12, 13]. In the remaining of the paper, we will carry out further analyses of the first two clusters. Our first question is: *do the two clusters of samples each represent the same cell population displaying different gene-expression patterns under different conditions, or each represents a mixture of multiple subpopulations of different cell types?*

To rule out the possibility that the two clusters are samples in different stages, a simple statistical test using Fisher exact test was conducted, which revealed that neither of the two clusters of samples show preference to any specific stage (p=0.1718, 0.1799, and 0.2791 for breast, colon and lung cancer respectively). This is conducted through building a 2 by 4 table with rows representing the two clusters and the four columns representing four cancer stages, and each entry denoting the number of samples in a specific stage in each cluster (see Supplementary Table S3), hence ruling out the possibility of the two clusters being stage dependent. We then checked the genomic mutation rates in the two clusters, and found that the hypoxia cluster has substantially higher mutation rates than the reoxygenation cluster in breast cancer (p=4.612e-06, by Mann-Whitney test), as shown in the boxplots of Figure 1A, and colon cancer (p=0.02215) and lung cancer (p=5.433e-07), as shown in Supplementary Figures S4A and S5A. Hence we conclude that the two clusters of samples represent (at least) two different cell types, one having substantially higher genomic mutations than the other, which leaves two possibilities: the two clusters each represent a distinct (homogeneous) subtype of the disease, or samples in each cluster all consist of a mixture of multiple cell types and the fraction of each cell type is approximately the same across all such samples (hence giving rise to their similar expression patterns across samples in the same cluster) but different for samples in different clusters. A mathematical analysis is carried out to determine

which of these two may better represent the available transcriptomic and genomic data.

Before we move on, we noted a clear correspondence among the identified clusters across the three cancer types with all the corresponding clusters sharing similar expression patterns. These shared expression patterns across each set of corresponding clusters are particularly outstanding and discerning across non-corresponding clusters over 62 out of the 242 genes. Hence we use this reduced set of 62 genes, referred to as *hypoxia-reoxygenation marker* (HRM) genes, for the following analysis. See Supplementary Table S6 for the names of these genes. Figure 1B and Supplementary Figures S4B and S5B show the expression patterns of the HRM genes in the identified biclusters in different cancer types.

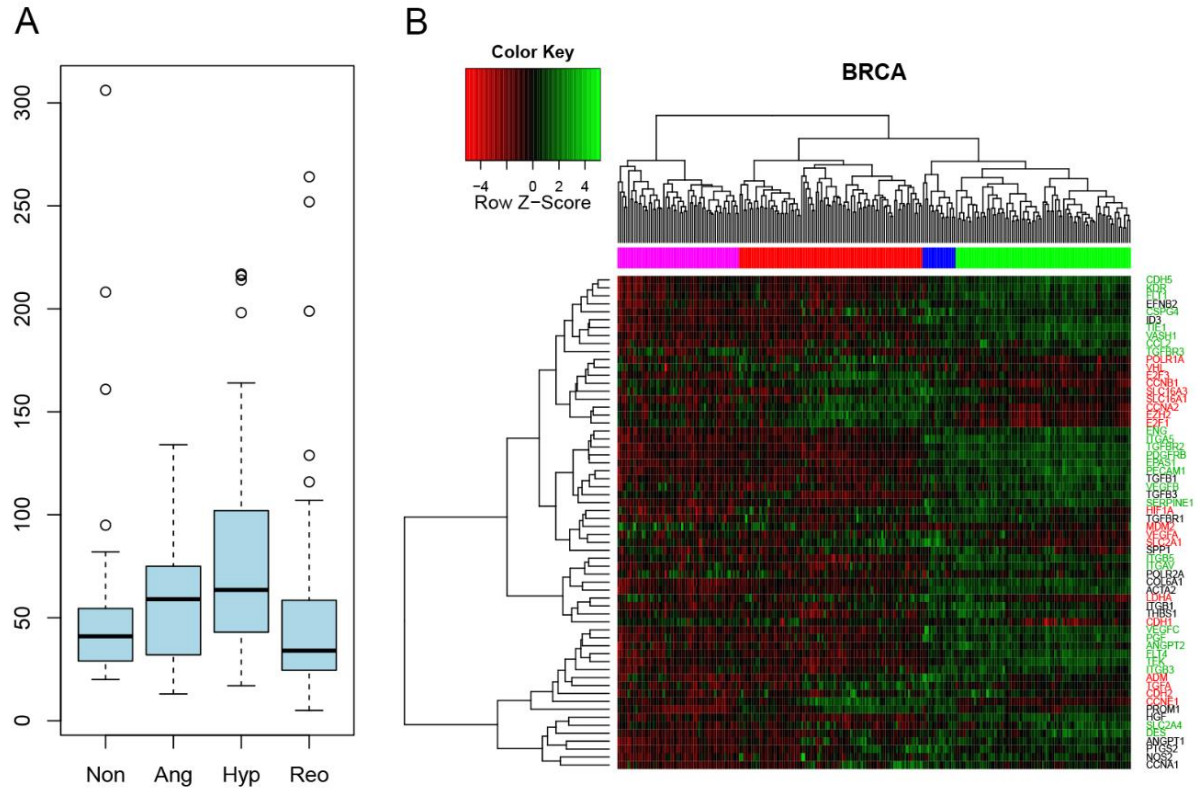


Figure 3.1: (A) Distribution of the numbers of mutations in the four major biclustering classes of breast cancer, where the four boxplots, from left to right, are for the background, angiogenesis, hypoxia and reoxygenation cluster, respectively. (B) Gene expression profiles of the HRM genes in the four major biclustering classes of breast cancer. Samples from the background, hypoxia, angiogenesis, and reoxygenation clusters are color-coded using the pink, red, blue and green bars on top of the figure,

respectively. Hypoxia and reoxygenation markers are labeled by red and green, respectively.

Subpopulations of different cancer cell types are consistently observed across all cancer tissues

i. Nonnegative matrix factorization reveals the existence subpopulations within each tissue sample

A number of studies have reported gene-expression patterns associated with hypoxia and reoxygenation conditions [14-17]. However, no reports have been published, to the best of our knowledge, regarding how such patterns change as a cancer advances and the implications of such changes to a cancer's development. We have carried out a nonnegative factorization of the sample-expression matrix to determine if all tissues of a specific cancer type may each consist of subpopulations of different cell types, each of which has its distinct characteristics. Specifically, we check if the expression matrix for each cancer type can be represented as the product of two nonnegative matrices with one representing the gene-expression signatures of the to-be-identified cell subpopulations across all the samples and the other representing the (to-be-determined) proportion of each cell type in the total cell population in each sample.

This problem can be formulated as a *nonnegative matrix factorization* (NMF) problem of the sample-expression data matrix $X_{m\times n}$ for each cancer type, i.e., $X_{m\times n}$ can be approximately represented as the product of two nonnegative matrices, $S_{m\times k} \cdot P_{k\times n}$, so the 2-norm of the residual, $||X\text{-}SP||^2$, is minimized, with $k \ll \min\{m, n\}$ being a to-be identified value, under the constraint that $\sum_i P_{i\times j}$ is approximately 1.0 for each $1 \leq j \leq n$, where m is the number of marker genes (i.e., m = 62) and n is the number of samples; S denotes a signature matrix with k columns, each representing one signature; P is a proportion matrix with k rows representing the proportion of each signature; and $||X||^2$ represents the 2-norm of vector X. A solution strategy is given in METHODS for this constrained optimization problem, which is used to solve the above for a fixed $k > 0$.

We noted that when k = 2, the residual matrix E is quite significant, but drops substantially when k = 3 and further reduction is minimal when k goes to 4 and beyond (see Supplementary Figure S7) for each of the three cancer types. This strongly suggests that there are three independent, significant and

consistent signatures of three cell subpopulations within each cancer tissue, across all samples for each cancer type.

We have examined the distribution of the proportions of each cell type across all samples of each cancer type, and found that three distributions are unimodal for all three cell types, as shown in Supplementary Figure S8. This strongly suggests that the cell-type specific fraction across all samples of the same cancer type forms one continuous distribution rather than multiple ones, hence suggesting that each sample consists of three subpopulations of distinct cell types, whose fractions change as the hypoxia level changes.

Interestingly, the three cancer types under consideration share very similar expression patterns in each of their three signatures: one signature with highly expressed hypoxia markers, one signature having highly expressed reoxygenation markers and a signature with low expression levels across the vast majority of the HRM genes as shown in Figure 2 and Supplementary Figure S9. Hence we call the three cell types the *hypoxia*, *reoxygenation* and *background* cells, respectively. Further analyses show that there is a strong correspondence between the three cell subpopulations and the biclustering results given in the previous section. Specifically, samples in the hypoxia and reoxygenation biclusters have substantially higher fractions of the hypoxia and the reoxygenation cell subpopulation, respectively, compared to the other clusters. And the angiogenesis samples tend to have more balanced fractions between the hypoxia subpopulation and the reoxygenation subpopulation, respectively.

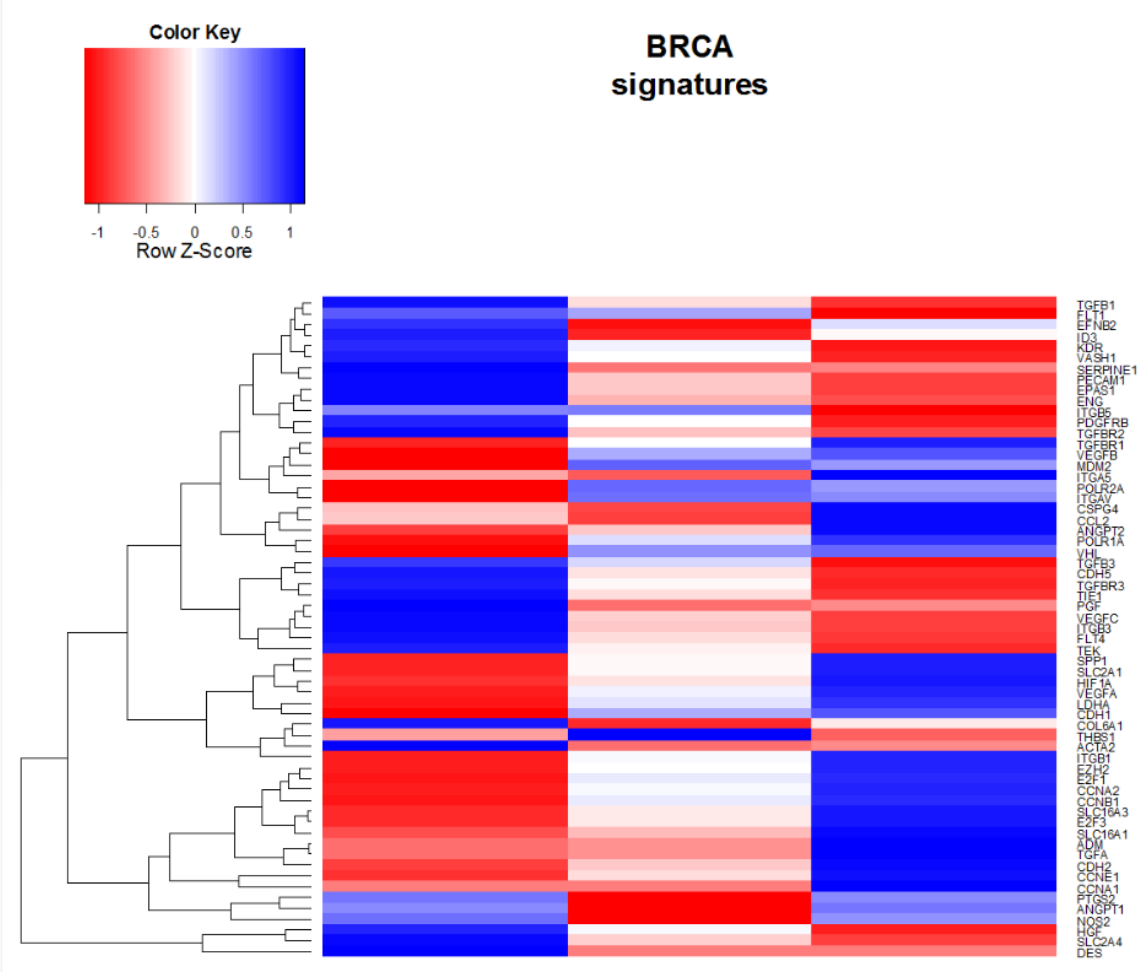


Figure 3.2: Gene-expression signatures for three cell subpopulations in breast cancer. In the figure, the columns represent the gene expression pattern of the reoxygenation, background and hypoxia subpopulation, from left to right. The color code represents down or up-regulation with the detailed definition given in the top of the figure.

As expected, we noted a significant negative (positive) correlation between the proportion of the hypoxia (reoxygenation) subpopulation in a tissue and its expression levels of the oxygen-consuming enzymes (excluding any from the 62 HRM genes) (Supplementary Figures S10A and S10B). This revealed a dynamics among the relative sizes of the three subpopulations within a cancer tissue, as a function of the oxygen level, which changes due to repeated growth and angiogenesis. Specifically, the hypoxia subpopulation increases as the hypoxia level goes up while the other two subpopulations clearly

shrink due to cell death possibly induced by the incompatible O_2 condition and/or by cell-cell competition within the same growing tissue [20], which has been widely observed in Drosophila where more viable cells can directly induce cell death in less fit cells in the neighborhood and replace them by the daughter cells of their own [21]. The same can be said about the fraction decreases of the hypoxia subpopulation in tissues as they become increasingly less hypoxic. This dynamics within a tissue is not apparent from the outside since the overall size of a tumor may not change when the relative proportions of different subpopulations change. One direct supporting evidence of the proposed change of subpopulation proportions is that there is a strong correlation between the proportion of the hypoxia subpopulation and the genomic mutation rate, as discussed in the previous section that the hypoxia cell type has higher genomic mutation rates, as shown in Figures 1 and 4 and Supplementary Figures S4 and S5.

ii. Biological processes specific to different subpopulations

We noted that the following HRM genes are consistently overly expressed in the hypoxia subpopulations across all three cancer types: (1) hypoxia markers *HIF1A* and *VHL*; (2) hypoxia-response genes *VEGFA, SLC2A1, ADM, TGFA, LDHA, SLC16A1* and *SLC16A3*; and (3) cell-cycle genes *CCNE1, CCNA2, CCNB1, E2F1, POLR1A, E2F3, EZH2* while the following HRM genes are consistently up-regulated in the reoxygenation subpopulation across all cancer types: angiogenic factor receptors and related genes such as *TIE1, DES, SERPINE1, HGF* and *FLT4*. A detailed list of these genes is given in Supplementary Table S11.

A pathway enrichment analysis was conducted over the up- or down-regulated genes, respectively, in the hypoxia and the reoxygenation subpopulations, separately, with the detailed lists of differentially expressed genes in each subpopulation and in each cancer type given in Supplementary Table S12. The enriched pathways are shown in Figure 3 and Supplementary Table S13.

Cell proliferation, glycolysis, DNA synthesis, amino acid transporters and cholesterol transporter genes are up-regulated in the hypoxia subpopulation: The following genes show strong positive correlations between their expression levels and the proportion of the hypoxia subpopulation: *CCNA2,*

CCNB1, *CCND1*, *CCNE1*, and *CCNE2*, DNA polymerase *POLR1A*, cyclin dependent kinases *CDK1* and *CDK2*, and cell cycle related transcription factor *E2F1*, *E2F2*, and *E2F3*, as well as glucose transporters *SLC2A1*, *SLC2A4*, glycolysis enzymes *ENO1*, *GAPDH*, *GPI* and *HK2*, lactate dehydrogenase *LDHA*, amino acid transporter genes, and cholesterol sensor and transporter genes *SCARB1* and *LDLR*, hence indicating accelerated cell proliferation by the hypoxia subpopulation with the increase in the hypoxia level. These observations are consistent with published studies showing that more hypoxic cancers tend to be more aggressive [23, 24]. In addition, acidity-response genes *CA9* and *CA12*, lactate transporter genes *SLC16A1* and *SLC16A3*, sodium/hydrogen exchanger genes *SLC9A3* and *SLC9A6*, and a number of V-ATPase genes exhibit positive correlations with the proportion of the hypoxia subpopulation in a tissue, indicating increased acidity level in the vicinity of the hypoxia subpopulation.

Over-expressions in cell-matrix interaction and cell-cell adhesion, cell-cell signaling, immune responses, nuclear receptors and growth factor genes are distinct characteristics of the reoxygenation subpopulation: The following genes show strong positive correlations with the proportion of the reoxygenation subpopulation in a cancer tissue: the ECM-component proteins and cell-adhesion proteins such as cell-matrix adhesion proteins, collagens, angiogenesis proteins, extracellular proteins and cell surface proteins, as well as multiple nuclear receptors such as *NR1D2*, *NR1H2*, *NR2F2*, *NR3C1*, *NR4A1*, *NR4A2*, *NR4A3*, and *NR5A2*, growth factors such as *CLEC11A*, *IGF1*, *FGF1*, *FGF13*, *FGF7* and *PDGFC*, and cell signaling proteins, plus immune response genes such as killer T-cell marker genes *CD1D*, *CD8* and *CD28*, helper T-cell marker *CD4*, and tumor associated macrophage (TAM) markers *CD68*, *CD163* and *CD206* (*MRC1*). In addition, oxygen-consuming enzymes in steroid hormone synthesis pathways are up-regulated, which may produce ligands that can bind to the over-expressed nuclear receptors to activate growth factors for cell proliferation [25, 26]. It is worth noting that the helper T-cells and TAMs (tumor associated macrophages) can release growth factors and regulate angiogenesis, ECM components, adaptive immunity and metastasis [27].

Based on this information, we posit that the two subpopulations play distinct and complementary roles in supporting tumor development and expansion. Specifically, the reoxygenation subpopulation

provides growth signals; make new ECMs as the foundation for further tumor growth; enhance cell-cell and cell-ECM interactions to serve as scaffolds for keeping the tumor cells connected; and bring in new blood supply in support of further growth while the hypoxia subpopulation releases angiogenesis signals as a result of increased hypoxia and create an acidic environment through releasing lactic acids to facilitate tumor cells to encroach into the neighboring areas [28] and to weaken T-cells attack [29]. We suspect that the reoxygenation cells may serve a role to keep all the cells connected since they have increased cell-cell adhesion activities while the hypoxia cells have decreased cell-cell adhesions. For this reason, we call the hypoxia and reoxygenation types the *loose* and the *sticky* type, respectively.

From our gene-expression data analyses, the reoxygenation subpopulation is clearly the slower growing subpopulation in comparison with the hypoxia cells, confirming a general evolutionary hypothesis: *greater pressures drive organisms to evolve faster* [30-32] as the hypoxia group is clearly under greater pressures.

Over-expressed superoxide reductases in hypoxia subpopulation *versus* peroxide reductases in reoxygenation subpopulation suggest that the two subpopulations have different sources of ROS: It has been established that reactive oxygen species (ROS) can be produced by different mechanisms including repeated cycles of hypoxia followed by reoxygenation and cell proliferation [33]. Interestigly the two cell populations have different sets of antioxidant enzymes up-regulated, namely anti-superoxide enzymes such as *XDH*, *SOD2* and *IDO1* in the hypoxia subpopulation and peroxide reductases such as *GPX2*, *GPX3*, and *GPX5* in the reoxygenation population. It is known that superoxide is mainly produced by the leakage of mitochondria under hypoxic conditions while peroxide can be produced by repeated cycles of hypoxia-reoxygenation [34-36]. Hence we posit that as a cancer tissue evolves, superoxides and peroxides are alternatively produced by the two dominating subpopulations, leading to the over-expression of different classes of anti-oxidants.

Over-expressed mesenchymal marker and matrix metalloproteinase genes reveal the invasiveness of the reoxygenation subpopulation: We have checked the expressions of the mesenchymal marker genes in the two subpopulations, and found that mesenchymal marker genes *CDH2*, *FN1*, *VIM* and *VTN* and

matrix metalloproteinase genes *MMP2*, *MMP3*, *MMP14*, *MMP16*, *MMP17*, *MMP19* and *MMP28* are consistently up-regulated in the reoxygenation subpopulation but not in the hypoxia population across all tissue samples, suggesting that the reoxygenation cells represent the more invasive population between the two. Clearly, this, in conjunction with the earlier data, revealed that faster growing and more invasive are independent behaviors of a cancer.

Up-regulated apoptosis genes in the hypoxia population and the *ATM/ATR* DNA repair pathway in the reoxygenation population are observed: Apoptosis (activation) genes such as *BAK1*, *CASP3*, *CASP9*, *CYCS*, *DIABLO* and proteasome genes are up-regulated in the hypoxia population. This, in conjunction with an earlier discussion, reveals that the hypoxia subpopulation has both high proliferation and apoptosis rates. In comparison, the reoxygenation population has both lower proliferation and apoptosis rates.

The observed up-regulation of the DNA repair genes ATM and ATR suggests more activated DNA repair activities in the reoxygenation population, which is consistent with previous reports [37].

Over-expressed steroid hormone receptors are observed in the reoxygenation population in all three cancer types, such as *ESR1, AR* and *PGR* (see Supplementary Figure S14A). In contrast, these genes are down-regulated in the hypoxia population. A possible explanation is that their ligand steroid hormones are produced through cholesterol oxidation, which has reduced activities due to reduced oxygen levels. At the tissue level, breast cancer samples fall into three groups each having a distinct expression pattern by these genes (see Supplementary Figure S14A). Interestingly the other two cancer types do not have such a grouping. Knowing that the expression levels of *ESR1* and *PGR* are the defining features of breast cancer subtypes, we are tempted to speculate that breast cancer subtypes may be related to the oxygen level of the tumor, which warrants further investigation.

Figure 3 summarizes the main distinct as well as common characteristics of the two subpopulations, as detailed in the above.

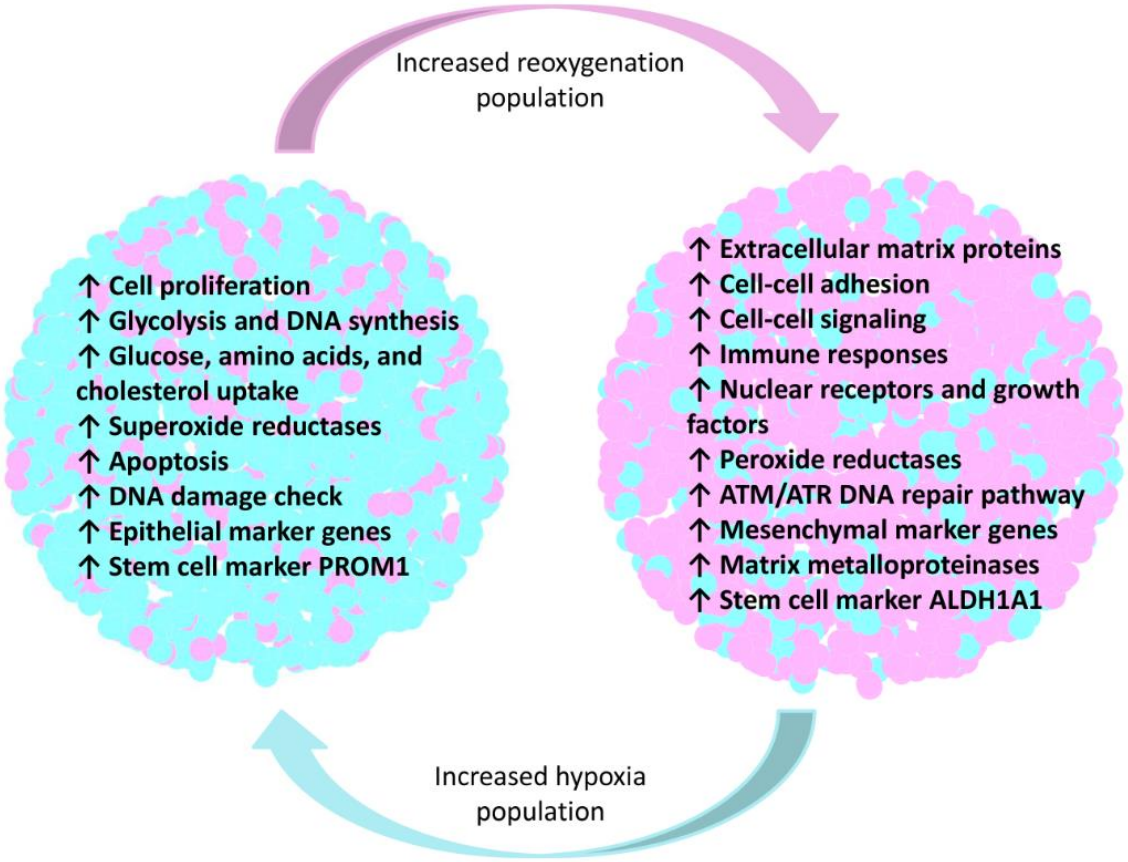


Figure 3.3: Key characteristics of the two subpopulations.

iii. Population dynamics over cancer stages reveal cancer evolution possibly driven by repeated hypoxia-reoxygenation cycles

The impacts of hypoxia and tumor angiogenesis on cancer evolution have been discussed in previous studies, including increased genome instability [13], drug resistance [38] and selection of specific characteristics of cancer stem cells [39]. It has been established that a number of cancer micro-environmental factors such as the hypoxic level, acidic level, oxidative stress, immune responses and ECM composition could be altered in response to the repeated hypoxia-angiogenesis cycles [40-42], leading to the selection of cell subpopulations with certain characteristics, which better fit the microenvironments, hence driving a cancer's evolution.

Selection of genomic mutations by the repeated hypoxia-reoxygenation cycles: By comparing the mutation rates of cancer tissue samples at different stages, we noted that the mutation rate does not increase significantly as a cancer advances (Supplementary Figure S10). It is interesting to note that highly mutated cancer genomes are rarely observed in advanced cancers when compared with early stage cancers. Hence we speculate that while increased genomic mutations may provide larger pools for natural selection and hence for cell survival, cancer cells may have a maximal level of tolerance in the level of mutations; that is, cells with highly mutated genomes are eliminated to keep the cells viable and capable to cope with the changing and complex microenvironments.

To examine the possible roles played by the repeated hypoxia-reoxygenation on cancer genome evolution, we have carried out a regression analysis to correlate the proportions of subpopulations in a cancer tissue to the mutation rates across all samples of each cancer type. The regression result, as shown in Figure 4, revealed that the mutation rate is positively correlated with the proportion of the hypoxia subpopulation and negatively correlated with that of the reoxygenation subpopulation in a tissue sample. This indicates that the mutation rate increases when a tissue becomes more hypoxic and the rate decreases when the tissue is more reoxygenated, which is true for all the cancer types under consideration. Overall, we speculate that there may be a general relationship between the mutation rate and the cell subpopulation with a specific optimal O_2 level, possibly adjusted by the organs; and each subpopulation has its intrinsic mutation rate. Hence when the proportion of a subpopulation changes, the mutation rate of a whole cancer tissue changes accordingly. Detailed regression parameters are listed in supplementary table S15.

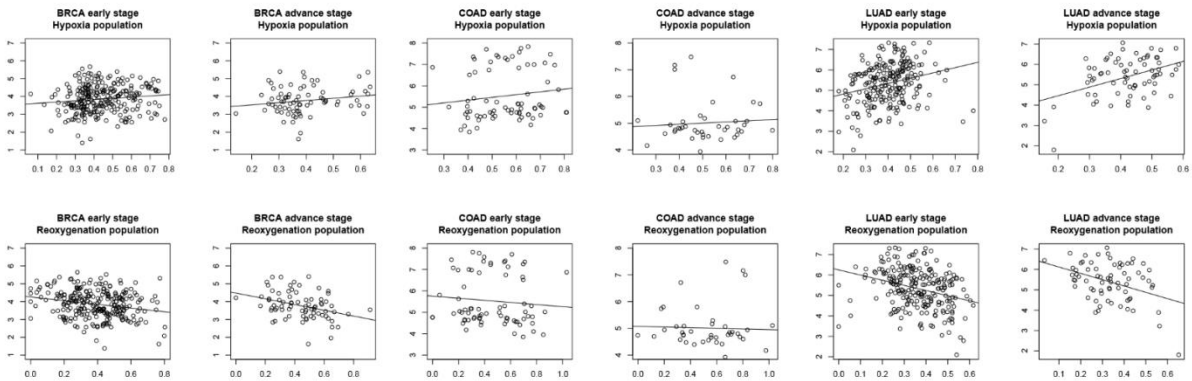


Figure 3.4: Gene mutation rate *versus* the proportion of a specific cell type. In each panel, the x-axis represents the proportion of a subpopulation and the y-axis represents the logarithm of the number of mutations averaged over all samples. The slope of the regression line suggests the correlation between the proportion of a subpopulation and the corresponding mutation rate.

Biological processes selected by repeated hypoxia-reoxygenation: We have carried out the following analysis to determine which biological processes may be positively selected by cancer evolution, i.e., defined as those that are strongly correlated with the proportion of a specific subpopulation (see METHODS for details).

The following pathways are positively selected by the repeated hypoxia-reoxygenation: apoptosis, cell proliferation, DNA synthesis, proteasome and WNT signaling, while cell-matrix junction, ECM components, immune responses such as T- and B-cell activation, cell defense system, cell-cell signaling, steroid hormone receptors, regulation of cell proliferation, platelet homeostasis, collagen and lipid rafts are negatively selected (See details in Figure 5). These findings indicate that cancer evolution in general positively selects the main characteristics of the hypoxia subpopulation and negatively select key features of the reoxygenation subpopulation.

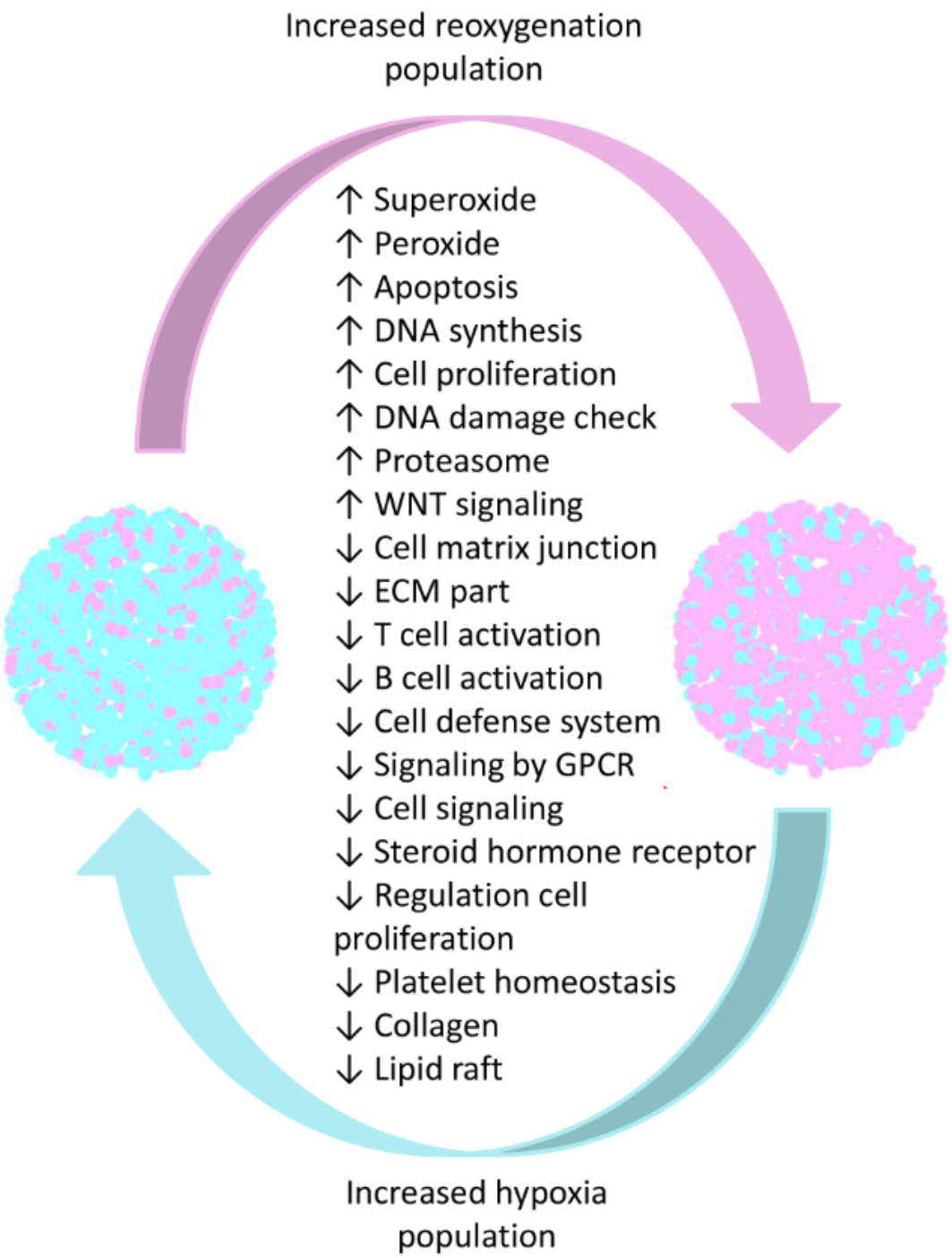


Figure 3.5: Biological processes selected by repeated hypoxia-reoxygenation.

Proportion distributions of the two subpopulations *versus* cancer stages: we have examined changes of proportions of subpopulations in a cancer tissue *versus* cancer stages and the T, N, M status,

representing cancer progression level, the primary tumor size, the number of regional lymph nodes that cancer has spread to, and the state of distant metastasis, respectively. Interestingly, the proportion of the hypoxic subpopulation in a tumor increases monotonously with respect to the cancer stage N, as shown in Figure 6, indicating a possible correlation between the relative fractions of subpopulations and the number of local lymph nodes that have been infected.

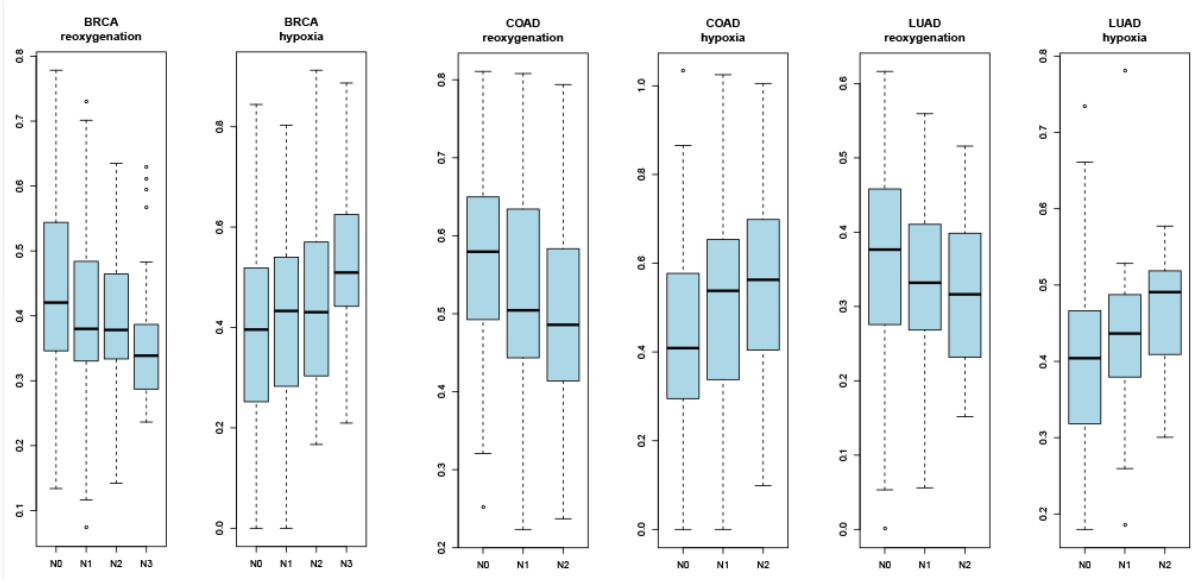


Figure 3.6: Distributions of proportions of the hypoxia and the reoxygenation subpopulations *versus* the N stages. In each panel, each bar from left to right represents the distribution of the proportion of a subpopulation in stage N0, N1 and N2 (and N3 in breast cancer) samples.

Relationship between the two subpopulations and circulating tumor cells: To understand if the metastasized cells may have come from one of or both the subpopulations, we have compared the gene-expression signatures of the two subpopulations with those of circulating tumor cells (CTC) [43, 44] collected from cancer patients' circulation. Specifically we have compared the signatures of CTCs originated from melanoma and pancreatic cancers (the only CTC data we can find and use here) with those of breast, colon and lung. Our analysis found that the cell proliferation, DNA synthesis, reoxygenation signatures, oxygen consuming enzymes, and cell-cell adhesion genes are over-expressed while apoptosis marker and hypoxia marker genes are under-expressed in CTCs compared to their

corresponding primary cancers, indicating that the up-regulated genes in the CTCs are consistent with those of the reoxygenation subpopulation in multiple gene categories while they show over-expressed cell-proliferation genes, which resemble more the hypoxia subpopulation. It is clear that additional information such as genomic mutation data is needed to draw a conclusion of whether the CTCs are predominantly from either, both or none of the two major subpopulations in a cancer tissue.

Discussion

An application of the NMF method to gene-expression data collected on a large number of cancer tissue samples allows us to probe the internal structure of a cancer tissue in terms of the subpopulation composition and dynamics and study how the relative sizes of the subpopulations change with the changing O_2 level, which is due to a repeated cycle of tissue growth and tumor angiogenesis throughout the entire process of a cancer development. It is noteworthy that there could be other ways to define subpopulations within a cancer tissue when examined from other angles. Here we look at the subpopulation composition and dynamics from the perspective of how the cells respond to repeated cycles of hypoxia and angiogenesis.

While it is not intuitively straightforward, our analysis provides strong evidence that the internal composition of the subpopulations, specifically the hypoxia and the reoxygenation subpopulations, in a cancer tissue changes. Various known mechanisms can be used to explain the predicted changes in the subpopulation dynamics in a tissue. This includes (1) increased apoptosis rates of one subpopulation under unfavorable O_2 conditions, while the dead cells can be replaced by new cells of another subpopulation, which generally has increased growth rates under the same condition as our data have shown; and (2) cell-cell competition as discussed earlier can directly replace viable but less fit cells (of the opposite subpopulation) by daughters of the more fit cells.

While the higher genomic mutation rates provide a solid evidence for the changing fraction of the hypoxia subpopulation in a cancer tissue as a function of the O_2 level, it remains to be fully understood how exactly the high mutation rates benefit or are even needed by the development of a cancer tissue.

One possible benefit, as discussed earlier, could be that the higher mutations may provide a richer pool of survival pathways, of course at the possible expense of higher death rates of such cells. More in-depth studies are needed to further investigate this issue.

The discovery of the conjugative relationship between the hypoxia and reoxygenation subpopulations, as discussed earlier, clearly adds a new dimension to cancer study. Even though previous studies have found that cancer tissues generally have heterogeneous cells, virtually nothing is known about their mutually beneficial relationship, potentially mutually dependent relationship among the subpopulations as we aim to understand in our future study. One potential direction is to investigate how the ratios between the sizes of the hypoxia and the reoxygenation subpopulations may be relevant to or even affect the form of a cancer biomass, such as diffused *versus* all neoplastic cells sticking together as one or a few tumors, as we would intuitively imagine that the higher percentage of the loose cells of the hypoxia subpopulation in a cancer, the more diffused the cancer cells may become in an organ, which clearly requires further study and validation. Our observation also suggests possible links between the activation of the steroidal hormone receptors and the hypoxia level, which also requires further analyses, particularly in relevance to breast cancer subtypes.

The discovery made on cancer subpopulation composition and dynamics could also have important implications to the study of the metastasis process since one may intuitively imagine that the loose cells of the hypoxia subpopulation may have higher chances to escape from their bases.

Conclusion

Our computation-based analyses of transcriptomic data of a total of 1,697 tissue samples of three cancer types revealed that the alternating hypoxia and reoxygenation cycles, plus the transitions between them resulted from tumor growth and repeated tumor angiogenesis, drive the changes in the relative proportions of two major subpopulations of cancer cells within a cancer tissue, i.e., the hypoxia and the reoxygenation subpopulations, which provide complementary beneficial conditions and material to facilitate a sustained growth of a neoplastic tumor under substantially different O_2 levels. The new

knowledge gained here, along with the computational techniques employed, should help to open new doors in cancer research, allowing studies of cancer development processes and mechanisms in a more realistic setting.

Material and Methods

Data used in the analyses

We applied our analysis on the TCGA RNAseqV2 and somatic mutation data. Three cancer types, namely breast invasive carcinoma (BRCA), colon adenocarcinoma (COAD) and lung adenocarcinoma (LUAD), were selected and analyzed with the following consideration: (1) they all have normal control tissue samples; (2) their sample sizes are relatively large with 994 samples for the BRCA set, 233 samples for the COAD set, and 470 samples for the LUAD set [45, 46]; and (3) the large variance in their expression levels of the hypoxia and reoxygenation marker genes suggests possible repeated cycles of hypoxia and reoxygenation.

Microarray gene-expression data of CTCs and their matching primary cancer samples were collected from the GEO database. Data set GSE18670 (pancreatic cancer, Affymetrix Human Genome U133 Plus 2.0 Array) and GSE38495 (melanoma, Illumina Genome Analyzer IIx) were analyzed to get gene markers of CTCs. Both of the two sets have 6 CTC samples, along with 6 and 20 matching primary cancer samples, respectively.

Hypoxia, angiogenesis and reoxygenation marker genes

It has been reported that repeated hypoxia-reoxygenation cycles may directly regulate the cell cycle, glucose metabolism and angiogenesis [47]. 242 genes related to various biological processes such as cell cycle, hypoxia response, glycolysis, glucose and lactate transport, and marker genes of angiogenesis signaling, angiogenesis process, endothelial cells, pericytes and reoxygenation are collected from the literature and public databases such as David [48], KEGG [49] and MSigDB [50], which represent the list of all directly relevant processes to hypoxia and reoxygenation. The detailed gene list is given in

Supplementary Table S1.

Biclustering analysis to identify repeated hypoxia-reoxygenation process

Repeated hypoxia has been observed using electron-paramagnetic resonance in cancer tissues [51]. Transcriptomic level alterations of several key maker genes have been observed in hypoxia and reoxygenation experiments [14-17]. In this analysis, we assume that each tissue sample is collected at one time point from the repeated hypoxia and reoxygenation process, hence their gene expression data can reflect the hypoxia level of a tissue when it was collected, and are comparable across different samples. The current understanding is that the typical response time of hypoxia and reoxygenation is approximately a few days to a few weeks, which is much longer than the hours long response time of transcriptomic alterations [52, 53]. Hence the transcriptomic level response to repeated hypoxia and reoxygenation cycle can be reflected by the transcriptomic data collected on the tissue samples.

We applied our in-house biclustering software QUBIC 1.0 that can simultaneously cluster the samples and the gene-expression data for identification of genes sharing similar expression patterns across some to-be-identified subset of samples, and find all such patterns above some similarity and size thresholds [10]. We have applied this tool to the sample-expression matrices for all the three cancer types for identification of all major biclusters. The tool is executed using the default parameters.

Nonnegative matrix factorization of a transcriptomic profile of multiple samples

Nonnegative matrix factorization has been applied to cancer stratification [54, 55], in which the transcriptomic profile of selected genes across multiple samples of a selected cancer type is modeled as follows:

$$X_{m\times n} = S_{m\times k} \cdot P_{k\times n} + E,\ k \ll \min\{m, n\},$$

in which $X_{m\times n}$ denotes a gene-expression matrix with m genes and n samples; $S_{m\times k}$ denotes a

nonnegative signature matrix with m rows and k columns, each representing one signature; $P_{k\times n}$ is the nonnegative proportion matrix with k rows representing the proportion of each signature; and E is the residual matrix, the 2-norm of which is d

enoted by $||E||^2$. The goal is find $S_{m\times k}$ and $P_{k\times n}$ that minimize $||E||^2$ for given k<< m, n.

The "Brunet" method in the NMF package [55] of R is applied with default setting to calculate the nonnegative matrix factorization on a given gene-expression matrix over 62 HRM genes. In detail, with an input matrix X and a given rank k, the S matrix and the P matrix are iteratively calculated using the following recurrence relationship:

$$\begin{cases} P_{ak} \leftarrow P_{ak} \dfrac{\sum_{i=1}^{m} \dfrac{S_{ia} X_{ik}}{(S \cdot P)_{ik}}}{\sum_{i=1}^{m} S_{ia}} \\ S_{ka} \leftarrow S_{ka} \dfrac{\sum_{j=1}^{n} \dfrac{P_{aj} X_{kj}}{(S \cdot P)_{kj}}}{\sum_{j=1}^{n} P_{aj}} \end{cases}$$

The stopping criterion of this iterative process is when the algorithm reaches convergence. Rather than directly evaluating the convergence of $||E||^2$, the Brunet method evaluates the stability of a connectivity matrix C of size m × m, in which C_{ij}=1 if samples i and j belong to the same cluster, and C_{ij}=0 if they belong to the different clusters [55]. We normalize P using $P = \frac{P_{estimated} * J}{\sum_{ij} P_{i\times j}}$ to ensure the constraint that $\sum_i P_{i\times j}$ is approximately 1.0 for each $1 \le j \le n$. The distribution of $\sum_i P_{i\times j}$ for each cancer type is shown in Supplementary Figure S16.

The rank, k, of the signature and proportion matrices is determined based on the number of significant principle components of the gene-expression matrix X, i.e. the number of significant eigenvalues of the variance matrix of X [56]. Supplemental Figure S5 shows the distribution of the eigenvalues of each cancer type. In all the three cancer types, k = 3 is selected in our analyses based on the eigenvalue distributions.

To determine whether the identified biclusters are due to different cancer subtypes or the changing proportions of different subpopulations within tissues in response to the micro-environmental conditions,

we have examined the proportion of each signature, i.e. each column of the matrix P. If the expression-sample matrix is composed of samples from different cancer subtypes, proportions of different cancer subtypes should in general follow different distributions and hence result in multimodal distributions of the fitted curves for the proportion distributions across all the samples for each cancer type, as shown in Supplementary Figure S12 (derived from data set GSE32646). Generally speaking, if such a proportion distribution is unimodal, the samples are most likely from one population [57]. Based on this argument, we can predict whether the observed expression patterns of the HRM genes are due to cancer subtypes or population dynamics of the same set of subpopulations.

Basic tools used in our data analyses

We have used Pearson correlation coefficients between gene-expression levels and the estimated proportion of each subpopulation to assess the level of correlation between the expression patterns of specific genes and the sizes of different subpopulations. Student t distribution is assumed when testing the statistical significance of each derived correlation coefficient, using 0.05 as the significance threshold. In the current study, we focus mainly on the hypoxia subpopulation and the reoxygenation subpopulation.

A gene is considered as up-regulated in one subpopulation if and only if the gene's expression level is significantly positively correlate with the size of the subpopulation and negatively correlate with the size of the other subpopulation. Similarly defined are down-regulated genes.

Mutation rate is defined by the # mutations/24,981, where 24,981 is the total number of mutated genes found in all TCGA data. For each cluster, the mutation rate is defined by the average mutation rate of the samples in the cluster.

For categorical data of cancer stages, Spearman correlation coefficients are used to assess correlations between gene expression levels and cancer stages; and the permutation test is used to assess the statistical significance of each correlation coefficient, using 0.05 as the significance threshold.

Pathway enrichment is assessed using a hypergeometric test (statistical significance cutoff = 0.001) [58]. 2,775 pathways from the Msigdb database including the GO terms and the Msigdb canonical gene

sets are used in our pathway-enrichment analysis.

Differentially expressed genes in CTCs *versus* their matching primary cancers are determined using the SAM's differential expression analysis method [59] (statistical significance cutoff = 0.05).

Acknowledgement

Authors would like to thanks Dr. Wenxuan Zhong, Dr. Ping Ma and Mr. Xin Xing from the Department of Statistics in the University of Georgia for their helpful discussion in the NMF method of this research project. YX would like to thank the continuous financial support from the Eminent Scholar Program of the Georgia Research Alliance.

References

1. Xu, X., et al., *Single-cell exome sequencing reveals single-nucleotide mutation characteristics of a kidney tumor.* Cell, 2012. **148**(5): p. 886-95.
2. Gerlinger, M., et al., *Intratumor heterogeneity and branched evolution revealed by multiregion sequencing.* N Engl J Med, 2012. **366**(10): p. 883-92.
3. Hou, Y., et al., *Single-cell exome sequencing and monoclonal evolution of a JAK2-negative myeloproliferative neoplasm.* Cell, 2012. **148**(5): p. 873-85.
4. Axelson, H., et al., *Hypoxia-induced dedifferentiation of tumor cells--a mechanism behind heterogeneity and aggressiveness of solid tumors.* Semin Cell Dev Biol, 2005. **16**(4-5): p. 554-63.
5. Navin, N., et al., *Tumour evolution inferred by single-cell sequencing.* Nature, 2011. **472**(7341): p. 90-4.
6. Malec, V., et al., *HIF-1 alpha signaling is augmented during intermittent hypoxia by induction of the Nrf2 pathway in NOX1-expressing adenocarcinoma A549 cells.* Free Radic Biol Med, 2010. **48**(12): p. 1626-35.
7. Denko, N.C., *Hypoxia, HIF1 and glucose metabolism in the solid tumour.* Nat Rev Cancer, 2008. **8**(9): p. 705-13.

8. Liao, D. and R.S. Johnson, *Hypoxia: a key regulator of angiogenesis in cancer.* Cancer Metastasis Rev, 2007. **26**(2): p. 281-90.

9. Dewhirst, M.W., Y. Cao, and B. Moeller, *Cycling hypoxia and free radicals regulate angiogenesis and radiotherapy response.* Nat Rev Cancer, 2008. **8**(6): p. 425-37.

10. Li, G., et al., *QUBIC: a qualitative biclustering algorithm for analyses of gene expression data.* Nucleic Acids Res, 2009. **37**(15): p. e101.

11. Yuan, G., et al., *Induction of HIF-1alpha expression by intermittent hypoxia: involvement of NADPH oxidase, Ca2+ signaling, prolyl hydroxylases, and mTOR.* J Cell Physiol, 2008. **217**(3): p. 674-85.

12. Hanahan, D. and R.A. Weinberg, *Hallmarks of cancer: the next generation.* Cell, 2011. **144**(5): p. 646-74.

13. Luoto, K.R., R. Kumareswaran, and R.G. Bristow, *Tumor hypoxia as a driving force in genetic instability.* Genome Integr, 2013. **4**(1): p. 5.

14. Dewhirst, M.W., *Relationships between cycling hypoxia, HIF-1, angiogenesis and oxidative stress.* Radiat Res, 2009. **172**(6): p. 653-65.

15. Polotsky, V.Y., et al., *Intermittent and sustained hypoxia induce a similar gene expression profile in human aortic endothelial cells.* Physiol Genomics, 2010. **41**(3): p. 306-14.

16. Dewhirst, M.W., *Intermittent hypoxia furthers the rationale for hypoxia-inducible factor-1 targeting.* Cancer Res, 2007. **67**(3): p. 854-5.

17. Toffoli, S. and C. Michiels, *Intermittent hypoxia is a key regulator of cancer cell and endothelial cell interplay in tumours.* FEBS J, 2008. **275**(12): p. 2991-3002.

18. Wheaton, W.W. and N.S. Chandel, *Hypoxia. 2. Hypoxia regulates cellular metabolism.* Am J Physiol Cell Physiol, 2011. **300**(3): p. C385-93.

19. Scott, B., et al., *Role of oxygen consumption in hypoxia protection by translation factor depletion.* J Exp Biol, 2013. **216**(Pt 12): p. 2283-92.

20. Gatenby, R.A., et al., *Cellular adaptations to hypoxia and acidosis during somatic evolution of*

breast cancer. Br J Cancer, 2007. **97**(5): p. 646-53.

21. Bondar, T. and R. Medzhitov, *p53-mediated hematopoietic stem and progenitor cell competition.* Cell Stem Cell, 2010. **6**(4): p. 309-22.
22. Gillies, R.J., D. Verduzco, and R.A. Gatenby, *Evolutionary dynamics of carcinogenesis and why targeted therapy does not work.* Nat Rev Cancer, 2012. **12**(7): p. 487-93.
23. Vaupel, P. and A. Mayer, *Hypoxia in cancer: significance and impact on clinical outcome.* Cancer Metastasis Rev, 2007. **26**(2): p. 225-39.
24. Vaupel, P., *Hypoxia and aggressive tumor phenotype: implications for therapy and prognosis.* Oncologist, 2008. **13 Suppl 3**: p. 21-6.
25. Lai, L.C., *Role of steroid hormones and growth factors in breast cancer.* Clin Chem Lab Med, 2002. **40**(10): p. 969-74.
26. Evangelou, A.I., et al., *Steroid hormones, polypeptide growth factors, hormone refractory prostate cancer, and the neuroendocrine phenotype.* J Cell Biochem, 2004. **91**(4): p. 671-83.
27. Quatromoni, J.G. and E. Eruslanov, *Tumor-associated macrophages: function, phenotype, and link to prognosis in human lung cancer.* Am J Transl Res, 2012. **4**(4): p. 376-89.
28. Hirschhaeuser, F., U.G. Sattler, and W. Mueller-Klieser, *Lactate: a metabolic key player in cancer.* Cancer Res, 2011. **71**(22): p. 6921-5.
29. Fischer, K., et al., *Inhibitory effect of tumor cell-derived lactic acid on human T cells.* Blood, 2007. **109**(9): p. 3812-9.
30. Greaves, M. and C.C. Maley, *Clonal evolution in cancer.* Nature, 2012. **481**(7381): p. 306-13.
31. Gutierrez, A., et al., *beta-Lactam antibiotics promote bacterial mutagenesis via an RpoS-mediated reduction in replication fidelity.* Nat Commun, 2013. **4**: p. 1610.
32. Tompkins, J.D., et al., *Error-prone polymerase, DNA polymerase IV, is responsible for transient hypermutation during adaptive mutation in Escherichia coli.* J Bacteriol, 2003. **185**(11): p. 3469-72.
33. Millar, T.M., V. Phan, and L.A. Tibbles, *ROS generation in endothelial hypoxia and*

reoxygenation stimulates MAP kinase signaling and kinase-dependent neutrophil recruitment. Free Radic Biol Med, 2007. **42**(8): p. 1165-77.

34. Jastroch, M., et al., *Mitochondrial proton and electron leaks.* Essays Biochem, 2010. **47**: p. 53-67.
35. Zulueta, J.J., et al., *Release of hydrogen peroxide in response to hypoxia-reoxygenation: role of an NAD(P)H oxidase-like enzyme in endothelial cell plasma membrane.* Am J Respir Cell Mol Biol, 1995. **12**(1): p. 41-9.
36. Tas, F., et al., *Oxidative stress in breast cancer.* Med Oncol, 2005. **22**(1): p. 11-5.
37. Kim, B.M., et al., *Reoxygenation following hypoxia activates DNA-damage checkpoint signaling pathways that suppress cell-cycle progression in cultured human lymphocytes.* FEBS Lett, 2007. **581**(16): p. 3005-12.
38. Sullivan, R., et al., *Hypoxia-induced resistance to anticancer drugs is associated with decreased senescence and requires hypoxia-inducible factor-1 activity.* Mol Cancer Ther, 2008. **7**(7): p. 1961-73.
39. Louie, E., et al., *Identification of a stem-like cell population by exposing metastatic breast cancer cell lines to repetitive cycles of hypoxia and reoxygenation.* Breast Cancer Res, 2010. **12**(6): p. R94.
40. Kim, Y., et al., *Hypoxic tumor microenvironment and cancer cell differentiation.* Curr Mol Med, 2009. **9**(4): p. 425-34.
41. Teppo, S., et al., *The hypoxic tumor microenvironment regulates invasion of aggressive oral carcinoma cells.* Exp Cell Res, 2013. **319**(4): p. 376-89.
42. Weljie, A.M. and F.R. Jirik, *Hypoxia-induced metabolic shifts in cancer cells: moving beyond the Warburg effect.* Int J Biochem Cell Biol, 2011. **43**(7): p. 981-9.
43. Sergeant, G., et al., *Pancreatic cancer circulating tumour cells express a cell motility gene signature that predicts survival after surgery.* BMC Cancer, 2012. **12**: p. 527.
44. Ramskold, D., et al., *Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells.* Nat Biotechnol, 2012. **30**(8): p. 777-82.

45. Cancer Genome Atlas, N., *Comprehensive molecular characterization of human colon and rectal cancer.* Nature, 2012. **487**(7407): p. 330-7.

46. Cancer Genome Atlas, N., *Comprehensive molecular portraits of human breast tumours.* Nature, 2012. **490**(7418): p. 61-70.

47. Cui, J., et al., *Hypoxia and miscoupling between reduced energy efficiency and signaling to cell proliferation drive cancer to grow increasingly faster.* J Mol Cell Biol, 2012. **4**(3): p. 174-6.

48. Huang da, W., B.T. Sherman, and R.A. Lempicki, *Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources.* Nat Protoc, 2009. **4**(1): p. 44-57.

49. Kanehisa, M., et al., *Data, information, knowledge and principle: back to metabolism in KEGG.* Nucleic Acids Res, 2014. **42**(Database issue): p. D199-205.

50. Subramanian, A., et al., *Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.* Proc Natl Acad Sci U S A, 2005. **102**(43): p. 15545-50.

51. Matsumoto, S., et al., *Imaging cycling tumor hypoxia.* Cancer Res, 2010. **70**(24): p. 10019-23.

52. Weis, S.M. and D.A. Cheresh, *Tumor angiogenesis: molecular pathways and therapeutic targets.* Nat Med, 2011. **17**(11): p. 1359-70.

53. Carmeliet, P. and R.K. Jain, *Angiogenesis in cancer and other diseases.* Nature, 2000. **407**(6801): p. 249-57.

54. Gao, Y. and G. Church, *Improving molecular cancer class discovery through sparse non-negative matrix factorization.* Bioinformatics, 2005. **21**(21): p. 3970-5.

55. Brunet, J.P., et al., *Metagenes and molecular pattern discovery using matrix factorization.* Proc Natl Acad Sci U S A, 2004. **101**(12): p. 4164-9.

56. Lee, D.D. and H.S. Seung, *Learning the parts of objects by non-negative matrix factorization.* Nature, 1999. **401**(6755): p. 788-91.

57. Kong, X.Z., et al., *Molecular cancer class discovery using non-negative matrix factorization with sparseness constraint.* Advanced Intelligent Computing Theories and Applications: With Aspects

of Theoretical and Methodological Issues, 2007. **4681**: p. 792-802.

58. Evangelou, M., et al., *Comparison of methods for competitive tests of pathway analysis.* PLoS One, 2012. **7**(7): p. e41018.

59. Tusher, V.G., R. Tibshirani, and G. Chu, *Significance analysis of microarrays applied to the ionizing radiation response.* Proc Natl Acad Sci U S A, 2001. **98**(9): p. 5116-21.

CHAPTER IV

ELUCIDATION OF DRIVERS OF HIGH-LEVEL PRODUCTION OF LACTATES THROUGHOUT A CANCER DEVELOPMENT

Chi Zhang, Chao Liu, Sha Cao, and Ying Xu. 2015. Journal of Molecular Cell Biology Volume 7, Issue 3. Pp. 267-279

Abstract

Lactates play key roles in facilitating and protecting the development of a cancer in the majority of or possibly all cancer types. While its beneficial effects to cancer development have been extensively studied, very little is known about what derive the high-level production of lactates in a cancer throughout its entire development across different cancers. Here we present a novel computational analysis of tissue transcriptomic data of nine primary cancer types, plus a few precancerous and metastatic cancers, to address this issue. Our approach is to identify stress types, known to play key roles in a cancer development, that show strong co-expressions with LDHA, the gene encoding the lactate dehydrogenase-A, at different stages of a cancer development. A number of interesting observations are made through our co-expression analyses, including (i) all nine primary cancer types show similar association patterns between stresses and LDHA, namely the strengths of the associations increase from early to intermediate-stage cancer tissues but then make a substantial down turn at the most advanced stage universally; (ii) while the detailed stress types associated with LDHA may vary across different cancer types, stresses induced by apoptosis and adaptive immune responses are present universally across all cancer types, suggesting that these two stresses are possibly two key drivers to keep the high-level production of lactates; and (iii) there is a clear distinction between stress types associated with LDHA in precancerous tissues vs early-stage cancer tissues while in contrast, no such distinction exists between advanced stage primary cancers and metastatic cancers. We anticipate that the analysis results here can provide highly useful information for designing personalized treatments for different cancers at different stages as stopping lactate production could have devastating effects on a cancer development.

Introduction

Warburg effect represents one of the most profound observations about cancer, initially made by German biochemist and Nobel Laureate Dr. Otto Warburg in 1920s [1]. It basically states that cancer cells tend to maintain high levels of glycolytic fermentation activities regardless of the cellular oxygen level. Substantial amounts of efforts have been invested into the study of this very intriguing and counterintuitive observation in the past century, particularly in the past decade. A number of proposals have been made regarding why cancer cells tend to use this ancient and less energy-efficient pathway even when mitochondria are functional and not limited by the lack of oxygen. These include that (i) the glycolytic fermentation pathway can produce intermediates for macromolecular synthesis needed for cell proliferation [2]; (ii) the pathway can provide NADPH, a key element needed for macromolecular synthesis [3]; and (iii) the pathway is faster in ATP production compared to the respiration pathway so the cells can keep up with the rate of cell proliferation [4]. Among all the proposals, maintaining a high-level production of lactates may represent a most interesting one.

A number of recent studies have provided strong evidence of how lactates (or lactic acids) can benefit cancer throughout its entire development, including (i) induction of apoptosis in the neighboring non-dividing cells [5], hence enabling cancer encroachment into their territories; (ii) enhancing anti-apoptotic activities in cancer cells [5]; (iii) suppressing the attacking activities of cytotoxic T cells [6]; (iv) fueling their associated macrophages (TAMs) [7] in exchange for TAMs' supporting roles in their growth; (v) promoting angiogenesis [8]; (v) reduced production of ROS (reactive oxygen species) [9, 10], which could be highly damaging, among others [5, 11]. Interestingly, cancers indeed tend to maintain a high-level production of lactates throughout their entire development, even including those cancers that do not rely glucose as an energy source such as prostate cancer [12]. In this study, we address the following question: *what drives a cancer to maintain its high-level production of lactates throughout its entire development*, through mining transcriptomic data of cancer vs control tissue samples.

Systematic analyses of transcriptomic datasets of four precancerous diseases, nine primary cancer types with normal controls and three metastases were carried out to infer what may influence the high-level production of lactates and the associated regulation at different stages of different cancer types. Specifically,

genes co-expressed with LDHA are identified in samples at each stage of each cancer type across all stages and all cancer types under consideration. A novel method is developed to infer an LDHA-centric co-expression network, organized as multiple gene modules, each covering functionally associated genes, over samples of each cancer type. The analysis gives rise to a collection of LDHA-centric co-expression networks, each at a specific stage of a cancer type, which collectively provide a dynamic view of how the co-expression networks, and hence the possible causes as well as associated regulatory mechanisms for maintaining the high-level production of lactates, change as a cancer advances. Comparative analyses of these stage-dependent networks across different cancer types offered new insights about the commonalities as well as distinct characteristics of such networks, and hence associated regulatory machineries of high-level production of lactates, across all cancer types examined.

Results

LDHA-centric co-expression networks at different stages across multiple cancer types

Increased lactate production has been widely observed based on over-expression of LDHA across different cancer types [11], which is confirmed by our own analyses, i.e., the gene-expression of LDHA is consistently up-regulated in cancer *versus* controls with high statistical significance (see Supplementary Figure 1 and Supplementary Table 1) across all the cancer types under consideration.

A main goal here is to infer the possible causes for cancer cells to keep high-level production of lactates and elucidate the associated transcription regulatory systems. Our analyses consist of two key components: (i) prediction of the relative order of progression among the tissue samples of each cancer type, which each have an assigned stage by pathologists and publicly available along with the gene-expression data; and (ii) identification of genes co-expressed with LDHA and function-based grouping of genes. Note that (i) enables our analysis to go beyond the four-stage categorization of cancer samples typically provided by cancer pathologists, allowing us to sequentially order samples based on similarities of their gene-expression patterns. The rationale is that the provided stage information offers only a rough estimate of where a cancer is in terms of its development, largely determined based on morphological and limited genetic biomarker information, which hence may not be necessarily very accurate in deciding the actual level of development

of individual samples. Our method makes its prediction using molecular data, making it more function-based, more reliable and more useful.

We have predicted the relative progression order of the samples in each dataset considered in this study, using the method given in METHODS. The idea is to identify a (small) subset of genes and then map all the given samples, based on their expression levels of the selected genes, to a linear order so that the original stage relationship (e.g., samples of stage X should be in general placed before samples of stage X+1) among all the samples is minimally violated by the ordering. Using this method, we have predicted the progression order of the samples under consideration, with an agreement level ranging from 71.62% to 96.01%, with the pathologist-assigned stages (see METHODS), as shown in Table 1. The prediction results for all the nine primary cancer types are provided in Supplementary Figure 2 and Supplementary Table 2. We then partitioned the ordered list of samples of each cancer type into four groups of consecutively ordered samples, referred to as four *predicted stages*, using a method given in METHODS, and then predicted a LDHA-centric co-expression network based on samples in each predicted stage (of each cancer type). The reason for using the predicted rather than provided stage is that our predicted stages better capture the similarities of expression patterns of the selected genes, hence biologically more meaningful. In addition, this procedure enables to sample the datasets under consideration to ensure that they each have similar sizes to avoid any issues associated with uneven data sizes. Additional information is provided in METHODS regarding the overall quality of the predicted stages.

Cancer type	**Concordance rate**
BLCA	84.11%
BRCA	73.32%
COAD	71.62%
HNSC	79.55%
KIRC	90.69%
LUAD	82.29%

PAAD	85.16%
PRAD	88.47%
THCA	96.01%

Table 4.1. Concordance rates between the predicted and assigned stages by pathologists of each cancer type. Details about the calculation of these values are described in METHODS.

To elucidate what factors may be responsible for the high-level production of lactates, we have examined five transcriptional factors (TFs) (see next section), possibly relevant to LDHA, and a number of cellular and micro-environmental stresses that are known to be cancer associated, referred to as *cancer associated stresses* (CAS) in our construction of co-expression networks centered on LDHA to infer their possible relevance to lactate production. The stresses considered are dysregulated glycolysis, hypoxia, oxidative stress, ER stress, DNA damage, extracellular matrix (ECM) component dysregulation, dysregulated innate and adaptive immune responses, NAPDH deficiency, NAD+/NADH imbalance, apoptosis and proliferating stresses. Based on these selected stresses, one LDHA-centric co-expression network is constructed, consisting of these stress associated genes, five TFs and LDHA, for each stage of each cancer type, using the method detailed in METHODS. Figure 1 shows the co-expression networks of different stages of colon cancer, with detailed statistical significance values for all the relevant co-expressions are given in Tables 3 and 4. Similar sets of co-expression networks for each of the eight other cancer types are given in Supplementary Figure 3 and Supplementary Table 3. Table 2 summarizes these networks using two measures: the network size and the number of known CAS types (CAS), along with the statistical significance value of each associated stress type in Table 4.

We noted that these networks tend to (1) have their sizes increase from stage 1 through stage 3 but consistently have reduced sizes in the last stage across all cancer types (except for breast cancer that has reduced sizes in the last two stages); (2) the CAS genes exhibit similar patterns in terms of their size distributions to those of (1); (3) there does not seem to be any significant difference in terms of the two above measures across precancerous, primary cancer and metastatic cancer tissues; (4) while these two numbers have a wide range of distributions across different cancer types, there do not seem to be any clear

relationships between the sizes and specific phenotypes of cancers such as their survival or recurrence rates; and (5) the size of a network and the number of CAS genes in cancer samples are consistently substantially higher than those of the normal controls. Further investigation may lead to new insights about how these numbers may contribute the development of a cancer.

Cancer	Category	Stage	Network size	#CAS genes
Ulcerative colitis	Precancerous	*	188	32
Inflammatory bowel disease	Precancerous	*	898	120
Early stage adenoma	Precancerous	*	404	51
Advanced stage adenoma	Precancerous	*	701	131
Normal Bladder	Normal	*	225	33
Bladder cancer	Primary cancer	I	445	110
Bladder cancer	Primary cancer	II	1011	260
Bladder cancer	Primary cancer	III	1035	213
Bladder cancer	Primary cancer	IV	906	210
Normal Breast	Normal	*	329	49
Breast cancer	Primary cancer	I	246	109
Breast cancer	Primary cancer	II	310	127

Breast cancer	Primary cancer	III	323	160
Breast cancer	Primary cancer	IV	285	167
Breast cancer	Primary cancer	V	238	133
Normal Colon	Normal	*	84	25
Colon cancer	Primary cancer	I	541	151
Colon cancer	Primary cancer	II	652	184
Colon cancer	Primary cancer	III	747	218
Colon cancer	Primary cancer	IV	411	130
Normal Head and neck	Normal	*	308	56
Head and neck cancer	Primary cancer	I	287	80
Head and neck cancer	Primary cancer	II	769	159
Head and neck cancer	Primary cancer	III	1206	305
Head and neck cancer	Primary cancer	IV	1408	293
Head and neck cancer	Primary cancer	V	817	205
Normal Kidney	Normal	*	431	57

Kidney carcinoma	Primary cancer	I	684	121
Kidney carcinoma	Primary cancer	II	1228	160
Kidney carcinoma	Primary cancer	III	1016	231
Kidney carcinoma	Primary cancer	IV	831	147
Kidney carcinoma	Primary cancer	V	593	155
Normal Lung	Normal	*	229	53
Lung adenocarcinoma	Primary cancer	I	107	70
Lung adenocarcinoma	Primary cancer	II	352	118
Lung adenocarcinoma	Primary cancer	III	512	108
Lung adenocarcinoma	Primary cancer	IV	546	196
Lung adenocarcinoma	Primary cancer	V	440	129
Normal Thyroid	Normal	*	123	20
Thyroid cancer	Primary cancer	I	321	72
Thyroid cancer	Primary cancer	II	373	121

Thyroid cancer	Primary cancer	III	353	91
Thyroid cancer	Primary cancer	IV	1040	268
Thyroid cancer	Primary cancer	V	465	111
Normal Prostate	Normal	*	139	27
Prostate cancer	Primary cancer	I	58	31
Prostate cancer	Primary cancer	II	135	56
Prostate cancer	Primary cancer	III	631	178
Prostate cancer	Primary cancer	IV	490	105
Pancreatic cancer	Primary cancer	I	748	146
Pancreatic cancer	Primary cancer	II	1123	254
Colon->liver	Metastases	*	365	42
Colon->lung	Metastases	*	596	107
Prostate->bone	Metastases	*	881	184

Table 4.2. A **summary of LDHA-centric co-expression networks over different stages of different cancer types**. Here, the network size is the number of genes in the network; the #CAS genes is the number of marker genes of cancer associated stress types in the network (see METHODS).

Co-expression network analyses of colon cancer

Because of the availability of precancerous and metastasis data in colon cancer in addition to primary cancer data, we use the cancer as an example to provide a detailed analysis of their co-expression networks focused on which TFs may be involved in regulation of LDHA and which stresses may contribute to these networks; and have a summarized analysis of all the other cancer types in the next section.

It has been well established that HIF1A and MYC are two key TFs of LDHA [13]. In addition, ATF1, CREB1 and SP1 have been reported to be able to regulate LDHA [14]. Here we examine the potential roles of these TFs on transcriptional regulation of LDHA in cancers under consideration. Specifically, a hypergeometric test is conducted to assess the p-value of the enrichment score by genes co-expressed with LDHA among genes regulated by each TF under consideration. The rationale is that a TF is predicted to be a regulator of LDHA if the number of genes regulated by the TF is significantly higher than the expected number of such genes among those co-expressed with LDHA by chance (see METHODS for detail). Table 3 shows the number of genes co-expressed with LDHA and simultaneously regulated by a specific TF, along with the estimated p-values over samples of specific stages of colon cancer.

Disease Type	**ATF1**	**CREB1**	**HIF1A**	**MYC**	**SP1**
Ulcerative colitis	0 (p=1)	0 (p=1)	0 (p=1)	**13 (p=0.0253)**	0 (p=1)
Inflammatory bowel disease	1 (p=0.4961)	3 (p=1)	**6 (p=0.0275)**	27 (p=1)	12 (p=0.4075)
Early adenoma of colon	0 (p=1)	3 (p=0.7378)	1 (p=1)	13 (p=0.8239)	4 (p=1)
Advanced adenoma of colon	0 (p=1)	1 (p=1)	3 (p=0.2467)	28 (p=0.3969)	9 (p=0.5722)
Normal Colon	1 (p=0.0639)	0 (p=1)	0 (p=1)	1 (p=1)	0 (p=1)
Colon cancer (stage I)	0 (p=1)	4 (p=0.1661)	**5 (p=0.0126)**	**34 (p=1.3e-03)**	10 (p=0.0908)

Colon cancer (stage II)	0 (p=1)	2 (p=1)	**5 (p=0.0253)**	**44 (p=6.4e-05)**	9 (p=0.4357)
Colon cancer (stage III)	1 (p=0.4358)	2 (p=1)	2 (p=0.7001)	**64 (p=7.1e-10)**	10 (p=0.4662)
Colon cancer (stage IV)	0 (p=1)	4 (p=0.0818)	3 (p=0.0229)	**22 (p=4.1e-02)**	**9 (p=0.0482)**
Colon->liver metastasis	0 (p=1)	1 (p=1)	3 (p=0.0632)	15 (p=0.4651)	6 (p=0.2957)
Colon->lung metastasis	0 (p=1)	2 (p=1)	2 (p=0.6559)	24 (p=0.4221)	9 (p=1)

Table 4.3. **A summary of predicted TFs of LDHA in colon cancer.** In the table, "p" represents the p-value of a gene being a possible TF of LDHA in a specific set of cancer samples, and all data marked in bold denote predictions with p-values < 0.05.

The results shown in Table 3 suggest that (a) none of the five TFs seem to regulate LDHA in normal colon tissues; (b) in precancerous tissues, MYC may regulate lactate production in ulcerative colitis samples; and HIF1A may regulate it in inflammatory bowel disease; (c) in primary cancer tissues, HIF1A seems to regulate the lactate production in the early-stage samples; SP1 regulates the production in the advanced-stage samples while MYC may jointly regulate it in samples across all stages; (d) the level of the regulation by MYC, measured by the p-values, increases as a cancer advances from early to the intermediate stages but decreases in the most advanced stage; and (e) none of the five TFs seem to regulate the lactate production in colon metastases, indicating that it may be regulated by different regulatory mechanisms, possibly driven by fundamentally different causes.

From Figure 1, we can see that multiple stress types, represented by their marker genes, are correlated with the high expression level of LDHA, hence the high production of lactates. It is noteworthy that a different number of genes associated with a specific stress type may be included in the co-expression

networks for different stages of the cancer. Our interpretation of the varying number of genes associated with a specific stress type is that the number reflects the strength of the association. That is: the higher the number of genes associated with a specific stress type, the stronger the stress is associated with the production of lactates. Overall, we have the following observations from Figure 1: (i) stresses induced by apoptosis and adaptive immune responses are associated with the high-level production of lactates throughout all stages of colon cancer, including precancerous and metastasis samples while the strength of the association may change; (ii) the network is small in ulcerative colitis containing only a few genes associated with innate immune responses, ER stress and ECM associated stresses (in addition to those in (i)); and notably, the activities of glycolysis and cell proliferation continue to increase throughout different stages of colon cancer, starting from here to most advanced stage cancers; (iii) the network is substantially larger for the inflammatory bowel disease than (ii), consisting of sizeable gene groups associated with ECM, hypoxia, ER stress, apoptosis and adaptive immunity related stresses; interestingly, the inflammatory bowel disease seems to have the most severe hypoxia and ECM dysregulation among all the stages of the disease; (iv) in colon adenomas, the dominating stresses seem to come from cell proliferation and DNA damages; and the strengths of these associations increase from early to advanced adenoma; in addition, ER stress and NAD/NADH imbalance show strong associations with LDHA in advanced adenomas; (v) in primary colon cancers, the co-expression networks clearly have more associated stress types than precancerous tissues; and the dominating stresses tend to come from cell proliferation, DNA damage, ECM component, apoptosis, oxidative and ER stresses; (vi) while the association strengths of these stresses tend to increase as the disease advances, the most advanced stage primary cancer tissues show a different stress pattern, with both the strengths and the number of associated stress types going down substantially, which seem to suggest that fundamental changes have taken place at this stage and the main causes for the high-level production of lactates have switched to something largely different from the earlier stage cancers; and (vii) in both lung and liver metastases of colon cancer, the sizes of the co-expression networks, especially the number of genes strongly co-expressed with LDHA, tend to be smaller than those in primary cancers; and the stress with the largest increasing strength is oxidative stress, suggesting the possibility that oxidative stress may represent the key stress that drives the high-level production of lactates in metastasized cancers.

Among the identified stresses associated with LDHA, the following has been previously reported to be able to induce lactate production: (1) when under hypoxia, cells are forced to metabolize glucose anaerobically, leading to lactate production [15]; (2) oxidative stress is due to the over-production of ROS, which may lead to repression of mitochondrial activities, including oxygenic respiration for ATP production, hence leading to increased glycolytic fermentation and lactate production; in addition, lactate has been shown to be a potential antioxidant agent [10], hence possibly triggered by ROS as a counterbalancing force; (3) fast proliferating cells may lead to NADPH deficiency, which can trigger lactate production [3, 16]; (4) lactic acidosis can induce up-regulation of anti-apoptotic proteins, which correlate with drug resistance pathways in multiple cancer types [17], suggesting a functional link between cell survival and lactate production; (5) lactic acids are known to weaken immune responses through multiple mechanisms, such as inhibiting the differentiation of monocytes to dendritic cells and interrupting the normal functions of activated T-cells [11, 18]; (6) acidic environment can lead to the degradation of extracellular matrices, and promote cancer invasion and metastasis [19]; (7) lactates can neutralize ROS, hence reducing the main damaging agent to DNA [10]; (8) the production of lactate requires NADH, which converts NADH to NAD+; and the relative concentrations between NADH and NAD+ can regulate the activity of LDHA, positively or negatively [20]; and (9) ER stress can lead to multiple stress types, of which, oxidative stress is a key one [21]. This partial list of relationships between various stresses and lactate production, all verified through experimental studies, provide strong evidence that our predicted associations between stresses and lactate production represent actually causal relations.

LDHA-centric co-expression networks in other cancer types

Network commonalities across multiple cancer types

Similar analyses of the LDHA-centric co-expression networks have also been carried out on other (primary) cancer types, namely bladder (BLCA), breast (BRCA), head and neck (HNSC), lung (LUAD), kidney (KIRC), pancreatic (PAAD), prostate (PRAD) and thyroid cancers (THCA). We summarize the key commonalities among the networks of different cancer types: (1) like in colon cancer, apoptotic and adaptive immune stresses show strong presence at each stage of all examined cancer types; (2) ER stress is

highly co-expressed with LDHA in all cancer types except for thyroid cancer; (3) cell proliferation-induced stress shows an increasing strength of association with LDHA as a cancer advances before the last stage samples across all cancer types except for lung cancer; (4) unlike the other stress types, oxidative stress shows increasing association strengths from early to the advanced stage for all cancer types, suggesting that overcoming oxidative stress may represent a key cause for maintaining a high-level production of lactates, knowing that ROS tends to increase with the progression of the disease [22]; and (5) like colon cancer, six out of the eight cancer types show a general trend with their network sizes increasing as the cancer advances to the advanced stage and then making a sharp down-turn at the advanced stage cancer samples. Figure 2 shows the co-expression networks over samples at the intermediate stage across the nine cancer types, while Supplementary Figure 3 gives the co-expression network for each stage of each of the eight cancer types.

As in colon cancer, we have also analyzed the role of the five TFs in the eight primary cancer types, with the detailed analysis results given in Supplementary Table 3. Overall, we noted that (i) MYC seems to regulate LDHA in early through intermediate-stage samples of BLCA, HNSC, all-stage samples of BRCA, PRAD and PAAD, and the advanced KIRC samples. Interestingly, MYC seems to have roles in normal lung and kidney tissues, but the statistical significance of the functional roles is apparently lower than in the cancer samples; (ii) SP1 seems to regulate LDHA in the advanced stage samples across seven out of eight cancer types, namely BLCA, BRCA, HNSC, KIRC, LUAD, THCA and PAAD; and (iii) HIF1A may regulate LDHA in majority of the cancer samples of BRCA, KIRC and PAAD.

Distinct network characteristics of individual cancer types

A number of distinct features of the networks for each of the nine primary cancer types are observed, namely: (1) hypoxia-induced stress is present in the networks for BRCA, KIRC and advanced stage BLCA, HNSC, LUAD and PAAD samples; (2) substantial level of ECM dysregulation associated stress is observed in BLAC, HNSC, LUAD, THCA and PRAD; (3) NADPH-deficiency induced stress has strong presence in BLAC, BRCA, COAD, KIRC, PAAD and early stage HNSC; (4) DNA damage-induced stress has strong presence in BLAC, COAD, LUAD, PAAD, intermediate-stage HNSC, advanced-stage KIRC and PRAD samples; and (5) unlike other cancer types having increased stress induced by cell proliferation, LUAD has

increased stress induced by innate immune responses and ECM dysregulation in its intermediate-stage samples.

Cancer	Category	Stage	Glycolysis	Hypoxia	ER Stress	Oxidative stress	ECM dysregulation	NADPH deficiency	NAD+/NADH imbalance
Ulcerative colitis	Precancerous	*	**0.0427**	1	**0.0192**	1	0.1119	1	1
Inflammatory bowel disease	Precancerous	*	**4.00E-04**	**<1e-5**	**<1e-5**	1	**<1e-5**	0.4625	1
Early stage adenoma	Precancerous	*	**0.0022**	0.1807	1	1	**8.00E-04**	0.0548	0.153
Advanced stage adenoma	Precancerous	*	**<1e-5**	0.0966	**2.00E-04**	**2.00E-04**	0.3454	**<1e-5**	0.2647
Normal Bladder	Normal	*	0.2088	1	1	0.2375	1	0.3996	0.1072
Bladder cancer	Primary cancer	I	**<1e-5**	1	0.3036	1	**0.0371**	0.0791	1
Bladder cancer	Primary cancer	II	**<1e-5**	0.2649	**0.007**	0.3715	**<1e-5**	**0.0018**	0.7789
Bladder cancer	Primary cancer	III	**<1e-5**	**0.0099**	**2.00E-04**	**0.0278**	**1.00E-04**	**0.012**	0.6143
Bladder cancer	Primary cancer	IV	**<1e-5**	**0.0237**	0.4692	**0.0012**	**0.0497**	**3.00E-04**	0.7826
Normal Breast	Normal	*	**0.0387**	1	1	0.1568	0.6316	1	0.6802
Breast cancer	Primary cancer	I	**2.00E-04**	1	**<1e-5**	0.0584	1	**0.0139**	1
Breast cancer	Primary cancer	II	**<1e-5**	**0.002**	**1.00E-04**	0.0564	1	0.4687	1
Breast cancer	Primary cancer	III	**<1e-5**	**0.0034**	**<1e-5**	**0.0073**	0.3825	**0.0041**	1
Breast cancer	Primary cancer	IV	**<1e-5**	**0.0421**	**<1e-5**	**0.0226**	0.4333	**0.0058**	0.7438

Breast cancer	Primary cancer	V	**<1e-5**	**0.0235**	**<1e-5**	**0.0081**	**0.0291**	**0.0148**	0.0577
Normal Colon	Normal	*	**0.0418**	1	1	**0.0057**	1	**0.0164**	0.345
Colon cancer	Primary cancer	I	**<1e-5**	0.0568	0.2829	**0.0011**	**0.0461**	**<1e-5**	0.7659
Colon cancer	Primary cancer	II	**<1e-5**	0.1517	**1.00E-04**	**0.0011**	0.1549	**0.0034**	0.0938
Colon cancer	Primary cancer	III	**<1e-5**	0.6184	**<1e-5**	**<1e-5**	1	**0.0015**	0.1608
Colon cancer	Primary cancer	IV	**<1e-5**	0.121	**<1e-5**	**0.0021**	0.1311	0.1108	0.0591
Normal Head and Neck	Normal	*	**0.0317**	1	0.4527	1	1	1	1
Head and neck cancer	Primary cancer	I	**<1e-5**	0.5227	**0.0027**	**0.0019**	**0.0068**	**0.0019**	1
Head and neck cancer	Primary cancer	II	**<1e-5**	0.0551	0.2275	**0.0233**	0.0914	**0.0014**	0.3652
Head and neck cancer	Primary cancer	III	**2.00E-04**	0.1788	**0.0041**	**0.011**	**0.0022**	0.1419	0.8182
Head and neck cancer	Primary cancer	IV	**<1e-5**	0.1704	**<1e-5**	**0.0035**	**0.0031**	0.3941	0.3455
Head and neck cancer	Primary cancer	V	**4.00E-04**	**0.0188**	**0.0326**	**0.0081**	**<1e-5**	0.6062	0.3914
Normal Kidney	Normal	*	**<1e-5**	0.4617	0.7561	0.6952	1	1	1
Kidney carcinoma	Primary cancer	I	**<1e-5**	**<1e-5**	0.3175	**0.024**	**<1e-5**	**0.0346**	1
Kidney carcinoma	Primary cancer	II	**<1e-5**	**0.0114**	0.4183	**0.0113**	0.1372	**<1e-5**	0.373
Kidney carcinoma	Primary cancer	III	**<1e-5**	**0.0273**	**<1e-5**	**<1e-5**	1	**<1e-5**	0.4282
Kidney carcinoma	Primary cancer	IV	**<1e-5**	**1.00E-04**	**<1e-5**	**0.0166**	**0.0013**	**<1e-5**	0.0975
Kidney carcinoma	Primary cancer	V	**<1e-5**	**0.0072**	**<1e-5**	**1.00E-04**	0.8417	**0.0047**	1

Normal Lung	Normal	*	0.1304	0.3777	0.1655	1	1	1	0.6802
Lung adenocarcinoma	Primary cancer	I	**0.003**	0.402	**0.0313**	**2.00E-04**	0.0655	0.2869	0.1093
Lung adenocarcinoma	Primary cancer	II	**<1e-5**	0.0189	**0.0369**	**0.0012**	**0.0396**	1	0.1216
Lung adenocarcinoma	Primary cancer	III	**8.00E-04**	0.5778	0.2919	0.3184	**<1e-5**	1	1
Lung adenocarcinoma	Primary cancer	IV	**0.0018**	**0.0422**	0.6531	0.2285	**0.0054**	1	0.3007
Lung adenocarcinoma	Primary cancer	V	**<1e-5**	0.1253	**6.00E-04**	**<1e-5**	**0.0087**	**0.0054**	0.0636
Normal Thyroid	Normal	*	0.0806	0.164	1	0.397	0.2335	0.5285	1
Thyroid cancer	Primary cancer	I	**<1e-5**	1	1	0.2061	**<1e-5**	1	0.4418
Thyroid cancer	Primary cancer	II	**3.00E-04**	0.2841	**0.0654**	0.017	**0.0413**	0.0571	1
Thyroid cancer	Primary cancer	III	1	1	1	1	**0.0287**	1	0.1715
Thyroid cancer	Primary cancer	IV	**0.0477**	1	1	1	**<1e-5**	0.65	1
Thyroid cancer	Primary cancer	V	**0.0144**	1	1	0.6627	0.1671	0.0791	1
Normal Prostate	Normal	*	**0.0023**	1	0.2263	1	1	1	1
Prostate cancer	Primary cancer	I	0.3039	1	0.1295	0.327	**0.0115**	0.1199	0.3828
Prostate cancer	Primary cancer	II	**<1e-5**	1	**0.0283**	**0.0067**	**9.00E-04**	**0.001**	1
Prostate cancer	Primary cancer	III	**1.00E-04**	0.0908	**0.0082**	**0.0156**	**1.00E-04**	1	0.2476
Prostate cancer	Primary cancer	IV	**<1e-5**	0.2022	**6.00E-04**	1	1	**0.0284**	1
Pancreatic cancer	Primary cancer	I	**<1e-5**	0.1577	**0.0054**	1	**<1e-5**	**0.0334**	1

Pancreatic cancer	Primary cancer	II	**<1e-5**	**0.0369**	**0.0026**	0.3952	**<1e-5**	**0.0416**	1
Colon->liver	Metastases	*	**<1e-5**	0.1445	**0.034**	**0.0035**	0.0609	0.1423	1
Colon->lung	Metastases	*	**<1e-5**	**0.009**	**<1e-5**	**<1e-5**	0.0607	**<1e-5**	0.056
Prostate->bone	Metastases	*	**<1e-5**	0.1612	**<1e-5**	**0.0014**	**0.0025**	0.2328	1

Cancer	**Category**	**Stage**	**Acidosis**	**Apoptosis**	**Innate immune response**	**Adaptive immune response**	**Cell proliferation**	**DNA damage**
Ulcerative colitis	Precancerous	*	1	**0.016**	0.1981	**<1e-5**	1	1
Inflammatory bowel disease	Precancerous	*	0.1192	**<1e-5**	0.2091	**<1e-5**	**0.0065**	1
Early stage adenoma	Precancerous	*	1	**0.0012**	**0.0042**	**3.00E-04**	**0.0085**	1
Advanced stage adenoma	Precancerous	*	**0.0268**	**<1e-5**	0.4202	**<1e-5**	**<1e-5**	**0.0095**
Normal Bladder	Normal	*	0.3249	0.548	1	0.1782	1	0.6558
Bladder cancer	Primary cancer	I	0.4705	**1.00E-04**	0.3154	0.0727	**<1e-5**	**0.0032**
Bladder cancer	Primary cancer	II	**0.0374**	**<1e-5**	**0.0789**	**<1e-5**	**<1e-5**	**0.0025**
Bladder cancer	Primary cancer	III	0.6561	**<1e-5**	**0.0174**	**<1e-5**	**<1e-5**	**0.0017**
Bladder cancer	Primary cancer	IV	**0.0024**	**<1e-5**	<1e-5	**<1e-5**	**0.0032**	0.4669
Normal Breast	Normal	*	0.4358	0.0505	1	0.1124	0.8765	1

Breast cancer	Primary cancer	I	0.2164	**<1e-5**	0.3923	**<1e-5**	**<1e-5**	**<1e-5**
Breast cancer	Primary cancer	II	0.1232	**0.0012**	1	**<1e-5**	**<1e-5**	0.0724
Breast cancer	Primary cancer	III	0.4971	**<1e-5**	**0.0202**	**<1e-5**	**<1e-5**	0.0457
Breast cancer	Primary cancer	IV	0.2278	**<1e-5**	**0.018**	**<1e-5**	**<1e-5**	0.1504
Breast cancer	Primary cancer	V	0.1606	**1.00E-04**	**7.00E-04**	**<1e-5**	**<1e-5**	0.1927
Normal Colon	Normal	*	1	0.1092	0.6628	0.7499	1	0.4155
Colon cancer	Primary cancer	I	**0.0048**	**<1e-5**	0.7498	**0.0014**	**<1e-5**	**6.00E-04**
Colon cancer	Primary cancer	II	**0.0534**	**<1e-5**	0.4833	**<1e-5**	**<1e-5**	0.5473
Colon cancer	Primary cancer	III	0.3301	**<1e-5**	1	**<1e-5**	**<1e-5**	**<1e-5**
Colon cancer	Primary cancer	IV	**0.007**	**<1e-5**	1	**0.0109**	**<1e-5**	**0.0185**
Normal Head and Neck	Normal	*	0.415	**7.00E-04**	1	3.00E-04	**0.0256**	0.7213
Head and neck cancer	Primary cancer	I	5.00E-04	**<1e-5**	**0.0423**	**<1e-5**	**<1e-5**	0.5722
Head and neck cancer	Primary cancer	II	**0.0115**	**<1e-5**	**0.0282**	**6.00E-04**	**0.0048**	0.2242
Head and neck cancer	Primary cancer	III	**0.0357**	**<1e-5**	**0.021**	**<1e-5**	**<1e-5**	**<1e-5**
Head and neck cancer	Primary cancer	IV	**0.0097**	**<1e-5**	0.172	**<1e-5**	**<1e-5**	**0.014**
Head and neck cancer	Primary cancer	V	0.1154	**<1e-5**	**0.0076**	**<1e-5**	**<1e-5**	0.3233
Normal Kidney	Normal	*	0.5264	1	1	1	1	1
Kidney carcinoma	Primary cancer	I	0.1479	**<1e-5**	**<1e-5**	0.0198	**0.01**	0.7378

Kidney carcinoma	Primary cancer	II	0.0819	**3.00E-04**	**0.0024**	0.182	**0.0183**	0.4155
Kidney carcinoma	Primary cancer	III	**<1e-5**	**<1e-5**	0.7772	**<1e-5**	**0.0011**	**<1e-5**
Kidney carcinoma	Primary cancer	IV	**0.0117**	**<1e-5**	0.22	**0.0125**	**0.0206**	0.1338
Kidney carcinoma	Primary cancer	V	**1.00E-04**	**<1e-5**	0.2989	**1.00E-04**	**0.0056**	**0.006**
Normal Lung	Normal	*	1	0.2521	1	0.7516	0.3507	0.1624
Lung adenocarcinoma	Primary cancer	I	0.1319	**<1e-5**	0.3077	**1.00E-04**	**<1e-5**	**0.0097**
Lung adenocarcinoma	Primary cancer	II	**0.0302**	**<1e-5**	**0.0026**	**0.001**	**<1e-5**	**<1e-5**
Lung adenocarcinoma	Primary cancer	III	0.2849	**<1e-5**	**<1e-5**	**<1e-5**	1	1
Lung adenocarcinoma	Primary cancer	IV	1	**1.00E-04**	**<1e-5**	**<1e-5**	0.0899	**0.055**
Lung adenocarcinoma	Primary cancer	V	0.1679	**<1e-5**	1	**5.00E-04**	**1.00E-04**	**6.00E-04**
Normal Thyroid	Normal	*	1	0.7846	0.0743	1	0.6086	1
Thyroid cancer	Primary cancer	I	0.4813	**1.00E-04**	**<1e-5**	**<1e-5**	0.6621	1
Thyroid cancer	Primary cancer	II	0.3319	**3.00E-04**	**0.0747**	**0.0024**	**0.0016**	0.064
Thyroid cancer	Primary cancer	III	1	**<1e-5**	**<1e-5**	**<1e-5**	**0.0074**	1
Thyroid cancer	Primary cancer	IV	0.1907	**<1e-5**	**<1e-5**	**<1e-5**	0.379	1
Thyroid cancer	Primary cancer	V	0.4705	**<1e-5**	**<1e-5**	**<1e-5**	0.3018	1
Normal Prostate	Normal	*	1	**0.004**	1	**0.0265**	**0.0299**	0.2238
Prostate cancer	Primary cancer	I	1	**9.00E-04**	0.0979	**0.0052**	**0.0189**	0.458
Prostate cancer	Primary cancer	II	1	**0.0433**	**0.0154**	**0.0083**	0.1992	0.3048

Prostate cancer	Primary cancer	III	0.4482	**<1e-5**	1	**6.00E-04**	**<1e-5**	**0.0254**
Prostate cancer	Primary cancer	IV	0.6263	**<1e-5**	1	0.1832	**<1e-5**	**<1e-5**
Pancreatic cancer	Primary cancer	I	0.2089	**<1e-5**	**0.0385**	**<1e-5**	**2.00E-04**	**0.0052**
Pancreatic cancer	Primary cancer	II	**0.0047**	**<1e-5**	0.8924	**<1e-5**	**<1e-5**	**0.0129**
Colon->liver	Metastases	*	**0.0162**	**0.0172**	0.444	**0.0011**	**1.00E-04**	**0.0334**
Colon->lung	Metastases	*	**0.0147**	**<1e-5**	0.6437	**<1e-5**	**<1e-5**	**1.00E-04**
Prostate->bone	Metastases	*	0.0586	**<1e-5**	0.2965	**<1e-5**	**<1e-5**	**<1e-5**

Table 4.4. **Stresses strongly associated with the LDHA co-expression networks for different cancer types**. The numerical values shown here are the p-values of the enrichment of LDHA co-expressed genes in each stress type related gene set, assessed using a hypergeometric test (See METHODS). The bold letters represent associations with significance level $p < 0.05$.

Metabolism and lactate production

In addition to the above analyses that link stresses to the production of lactates, we have also examined metabolic pathways known to be essential to cancer development in terms of whether their expressions may be correlated with that of LDHA, i.e., which may drive the high-level production of lactates. For example, cancers tend to increase their carbon flux to the pentose-phosphate pathway for the biosynthesis of NADPH among other functions, needed for the synthesis of macromolecules, in which enhanced lactate production as a side product is commonly observed [23]. 83 KEGG metabolic pathways [24] are analyzed in terms of their co-expressions with LDHA. Specifically, genes co-expressed with LDHA were first identified, and then pathway enrichment analyses were carried out to identify metabolic pathways enriched with such co-expressed genes. This is done using a method similar to the analysis in the above sections, i.e., a hypergeometric test for pathway enrichment plus a FDR-based correction for false discovery rate.

Table 5 summarizes the metabolic pathways that show strong correlations with LDHA at different stages of the nine (primary) cancer types, suggesting the possibility that their activations may also contribute to the high-level production of lactates. A few observations are noteworthy: (1) sugar metabolisms such as glucose (glycolysis), fructose, mannose, and galactose metabolisms are highly

correlated with LDHA in most stages of most cancer types; (2) cell autophagy related proteasomes are co-expressed with LDHA throughout all stages of BLCA, BRCA, COAD, LUAD, PRAD and PAAD, the intermediate stage of HNSE and KIRC, and the early-stage samples of THCA, which are consistent with our knowledge that the lactate accumulation can induce autophagy [25]; (3) nucleotide synthesis of purine and pyrimidine and pentose phosphate pathway both co-express with LDHA in the intermediate-stage samples of all the cancer types except for THCA and PAAD. This observation is consistent with PKM2 induced up-regulation of the pentose phosphate pathways and lactate production observed in fast proliferating cancer cells [23, 26]; and (4) oxidative phosphorylation pathway co-expresses with LDHA in the intermediate and advanced-stage samples of BRCA, COAD, HNSC, KIRC, LUAD and in the advanced-stage samples of PRAD.

	BLCA	BRCA	COAD	HNSC	KIRC	LUAD	THCA	PRAD	PAAD
Aminoacyl tRNA biosynthesis	M	E-M		M-A	M-A		E	A	
Fructose and mannose metabolism	ALL	M-A	ALL	E&A	ALL			M-A	ALL
Galactose metabolism	M-A	A	E	M-A	M-A	A			E
Glycolysis, gluconeogenesis	ALL	ALL	ALL	ALL	ALL	ALL	E&A	ALL	ALL
Nucleotide excision repair	E	ALL	A	M	M	E&A		A	A
Oxidative phosphorylation		M-A	M-A	M-A	M-A	M-A		A	
Pentose phosphate pathway	M-A	ALL	ALL	M	M-A	A		M-A	
Proteasome	ALL	ALL	ALL	M	M	ALL	E	ALL	ALL

Purine metabolism	M-A	M	M	M	M	E-M		A	
Pyrimidine metabolism	E-M	M	M	M	M	E-M		A	
Pyruvate metabolism	M	ALL	E-M		M	E	E	M-A	
RNA degradation	E	M-A	M-A		A	M		A	

Table 4.5. **Metabolic pathways co-expressed with LDHA.** Here E, M, A, and ALL denote early, intermediate, advanced and all stages, respectively.

Regulation of LDHA vs other glycolytic enzymes

One natural question now is: *are other enzymes in the glycolysis pathway, such as ENO1, PGK1, GAPDH, HK2 and PKM2, transcriptionally regulated through the same mechanism as that of LDHA*? To address this issue, we have carried out co-expression network analyses similar to the above but centered on each of these glycolytic enzymes. The detailed information of the co-expression networks are shown in Supplementary Table 4. By comparing the LDHA-centric co-expression networks with the corresponding networks centered on each of these enzymes, we made a number of observations. Specifically, the gene-groups associated with each of the following stress types for the LDHA-centric networks are similar to the corresponding networks for the majority of the other glycolytic enzymes: hypoxia, ER stress, ECM dysregulation, and NADPH deficiency, while for the following stress types, the LDHA networks are substantially larger than the corresponding networks for the other enzymes: oxidative stress, apoptosis, adaptive immune response, cell proliferation and DNA damage. The first observation suggests that the transcriptional regulation of LDHA is probably the same as or highly similar to those of the other glycolytic genes under the first group of stresses, while the second observation strongly suggests under the second group of stresses, the transcriptional regulations for LDHA and the other glycolytic genes are different.

Discussion

Lactate represents a key ingredient in cancer development as it serves as both a facilitator and a protector for cancer evolution. However it remains largely unknown what drives the high-level production of lactates throughout the development of the majority of or possibly all cancers. Here we presented a novel

approach to study the possible causes. Different from published studies aiming to elucidate possible causes of a specific observation about cancer, we did not try to address this issue from a mechanistic perspective, e.g., to infer the regulators of the observed high production level of lactates, which may prove to be exceedingly challenging, knowing the very complex and convoluted relationships in the networks of regulators and regulated genes, which may change dynamically. Instead, we took an evolutionary approach by identifying all the stress types (and carbon metabolisms) that are known to play key roles at different stages of a cancer and found their strong co-expressions with lactate production. The result provides a high-level, conceptually clear roadmap of what may contribute to or drive the production of lactate. Our interpretation of the derived LDHA-centric co-expression networks with stresses is that cancer cells have "learned" to trigger specific stresses at different developmental stages, which will lead to lactate production as a response, as encoded in our genomes, hence providing the protection and facilitation needed by cancer.

The derived networks for different cancer types revealed that different stress types may have driven the high-level production of lactates, possibly due to the distinct microenvironments associated with as well as the genomic programs encoded in different cancer cell types. We fully anticipate that more detailed analyses of these stresses of different cancer types will reveal distinct causes of specific cancer types in addition to the induction of lactate production, which we plan to do as a follow-up study. The most striking commonality as we observed across eight primary cancer types is the substantially reduced overall stresses associated with lactate production in the advanced stage cancer samples, suggesting that these cancers are making a fundamental transition at this stage. Further analyses of (i) what drive this transition and (ii) what this transition leads to may give rise to fundamentally new understanding of cancer as an evolving system.

Study of cancer evolution, as we have done here, from a stress perspective represents a novel approach to cancer studies. Our first effort along this direction has generated novel and exciting results. Further analyses of the derived stress-associated co-expression networks centered on LDHA here or on other genes/pathways in other analyses will require abilities to infer the underlying regulatory mechanisms of how a specific stress or stresses lead to the production of lactates, which will be done in our follow-up study of the work presented here.

Conclusions

Our stress-based analyses have generated new insights about what may drive the high production of lactates, a key beneficial element to cancer development, which may be applicable to all cancer types. The method presented here represents the first such analysis, to the best of our knowledge, which could be applied to other studies for elucidation of causal relationships in cancer and other diseases. The results derived through our stress-associated co-expression networks can be possibly used in personalized plans for treating specific cancers at a particular stage through removing or minimizing the beneficial effects of lactates, which we believe will be devastating to the relevant cancer.

Material and Methods

Data

TCGA RNAseqV2 data of nine primary cancer types, namely bladder urothelial carcinoma (BLCA), breast invasive carcinoma (BRCA), colon adenocarcinoma (COAD), head and neck squamous cell carcinoma (HNSC), kidney renal clear cell carcinoma (KIRC), lung adenocarcinoma (LUAD), pancreatic adenocarcinoma (PAAD), prostate adenocarcinoma (PRAD) and thyroid carcinoma (THCA), were selected and analyzed in this study. These cancer types were selected based on the following consideration: (1) they have transcriptomic data publicly available and their sample sizes are relatively large, with 182 and 18 samples for BLCA and normal bladder controls, 994 and 106 samples for BRCA and normal breast tissues, 233 and 21 samples for COAD and normal colon, 303 and 37 samples for HNSC and normal head and neck, 480 and 71 samples for KIRC and normal kidney tissues, 470 and 58 samples for LUAD and normal lung, 494 and 58 samples for THCA and normal thyroid, 195 and 45 samples for PRAD and normal prostate, and 56 samples for PAAD with 2 normal control [27]; and (2) the data come with the stage information for each cancer sample.

Microarray data of pre-cancerous diseases and metastasis measured by Affymetrix Human Genome U133 Plus 2.0 Array were collected from the GEO database. Data sets GSE4183 and GSE38713 of precancerous colon inflammatory diseases, GSE37364 of precancerous colon adenomas primary cancer samples, and GSE41258 and GSE32269 of cancer metastases were analyzed. Supplementary Table 5 lists

the detailed information of these data sets.

Prediction of the relative order of cancer progression and staging

A regression model is developed to map a given set of cancer tissue samples, along with their pathologists-assigned stages, to a linear order based on their gene-expression data so that the partial order provided by the assigned stages is maximally preserved. Specifically, the following elastic-net penalized linear regression model is fitted:

$$\min \frac{1}{2\mathrm{N}} \sum_{i=1}^{N} \left(y_i - \beta_0 - x_i^T \beta\right)^2 + \lambda \left[\frac{(1-\alpha)\left|\left|\beta\right|\right|_2^2}{2} + \alpha \left|\left|\beta\right|\right|_1 \right],$$

where N is the number of samples, y_i is the numeric-valued stage, x_i^T is the gene expression profile of sample i. β_0 and vector β are linear coefficients, λ is a penalizing parameter, and α is an elastic-net mixing parameter, with range [0,1]. Details can be found in [28]. The R codes of this method is given in the Supplementary Appendix.

In order to determine parameter value α, 100 permutations on the sample labels are performed. In each permutation, a sequence of α from 0 to 1 with step 0.1 is employed. Then, a 10-fold cross-validation is done to select an α with the highest concordance rate, which is defined as the percentage of partial ordering that are preserved by the predicted sample orders compared with the provided pathologic stages. The results not only give the best parameter value for each cancer type based on cross-validation, but also show that the trained linear predictor is very significant since samples with the original order gives the best prediction with the highest concordance rate compared with all the permutations.

To assign stages to the linearly ordered samples of a cancer type, we split the sequence of the ordered samples into sub-lists of a fixed size. We fix the sample size to 40 for all the assigned stages to eliminate possible bias due to samples of different sizes in further co-expression analysis. Supplementary Table 2 lists the samples each with the pathologist-assigned stage. Supplementary Figure 2 shows the predicted vs the assigned stages of the tissues samples of the nine primary cancer types under consideration.

Co-expression module analysis

Most of the current co-expression network construction methods tend to focus on identification of the global properties of a network. Inclusion of local co-expression modules (or subnetworks) represents one way to extend the current methods for co-expression network construction. Here we have developed a novel computational method to identify co-expression modules that are each co-expressed with a specific target gene, where a *co-expression module* is defined as a maximal set of genes among which co-expressed gene pairs exist significantly more frequently than those by chance. To identify the co-expression modules for a given gene, we used a novel measure, *mutual rank (MR)*, to rank genes based on the number of genes co-expressed with them vs such numbers of their neighboring genes. Consider a set of genes and their pairwise co-expression relationships, represented as nodes and edges, respectively, the mutual rank between gene (node) *i* and gene (node) *j* is defined as [29]:

$$\mathrm{MR(i,j)} = \mathrm{MR(j,i)} = \sqrt{Rank(i \to j) \cdot Rank(j \to i)}\,,$$

where $Rank(i \to j)$ is the rank of the distance from gene *i* to gene *j* among all distances from *i* to other genes in the given co-expression network, where distance between *i* and *j* is measured using the Pearson correlation coefficient between the expression patterns of the two genes. In general, a smaller MR value generally has a higher rank for the relevant genes (nodes). We use this measure to identify *hub* genes, i.e., genes with substantially more neighbors than other genes in the neighborhood.

Searching for hub genes: For a given node i and a threshold value T, define $\mathrm{N_{MR}}(i,T) = \left|\left\{j \,\middle|\, j \neq i,\ MR(i,j) < T^{\frac{1}{2}}\right\}\right|$. Note that the distance between a gene and a hub gene should have a relatively top rank, and a top ranked $Rank(j \to hub)$ should generally have a relatively low increase rate of $\mathrm{MR(hub,j)}$. Hence for a certain threshold T, a hub gene should have more MRs smaller than $\mathrm{T}^{\frac{1}{2}}$ compared to its neighboring genes. Supplementary Figure 4 shows the rates of increase of the MRs of hub and non-hub genes are substantially different. Based on this observation, we have developed the following two procedures, one non-parametric and one parametric, to identify hub genes in a co-expression network.

A non-parametric procedure*:* For a given network with N nodes having a co-expression (weight) matrix S, we assess the significance of the N_{MR} value for any node *i* by comparing it with the N_{MR} value computed with randomly assigned co-expression levels.

a) Repeat the following for N times: randomly shuffle the index of each gene, assign the weight matrix S and re-compute N_{MR} of node *i;* and create a histogram of the computed N_{MR} values.
b) Estimate the nominal P-value for node *i*'s N_{MR} value and threshold T based on the above estimated histogram.

A parametric procedure: While a non-parametric procedure is preferable, it becomes computationally prohibitive for large networks (e.g., co-expression networks with $>10^6$ nodes, which may happen when alternatively spliced genes are considered when analyzing human transcriptomes). To overcome this issue, we have developed a parametric model based on an assumption about the distribution of the edge weights of the given network. We first demonstrate that the MR value follows a distribution as shown below, and then present a parametric procedure for estimating the P-value of a node *i*'s N_{MR}. Specifically we have mathematically proved that

$$\mathrm{P}(Rank(i \to j) = x, Rank(j \to i) = y) =$$

$$\int C_{N-1}^{x-1}(1-F(k))^{x-1}F(k)^{N-x} \cdot C_{N-1}^{y-1}(1-F(k))^{y-1}F(k)^{N-y} \cdot f(k)\, dk \qquad (*)$$

in which $f(k)$ and $F(k)$ denote the pdf and cdf of the null distribution of the original distances, respectively. As in this work, we use the Pearson correlation of the expression of each gene pair as the distance and take the null distribution by $\mathrm{f(r)} = \frac{(1-\mathrm{r}^2)^{\frac{\mathrm{n}-4}{2}}}{\mathrm{B}\left(\frac{1}{2},\frac{\mathrm{n}-2}{2}\right)}$. By the distribution given in (*), the significance level of the $N_{MR}(i,T)$ can be estimated (See details in Supplementary Methods).

Construction of a co-expression network: With the identified hub genes in a global network, the local co-expression network of a gene X (e.g. LDHA in the current study) can be constructed as follows:

1) Identify genes that are significantly co-expressed with X and put them into the network, where substantial co-expressions are determined using Pearson correlation coefficients and student *t* tests by using Pearson's product moment correlation coefficient with significance level p = 1e-4 for FDR correction or MR < 100.

2) For each hub gene in the network, genes that are significantly co-expressed with the hub, as defined in (1), are put into the network.
3) A significant link is determined by the significance as defined in (1).

Stress related gene sets and enrichment analysis

Thirteen stress types including altered glucose metabolism, hypoxia, ER stress, oxidative stress, dysregulated ECM, NADPH deficiency, NAD+/NADH imbalance, acidosis, apoptosis, innate immune response, adaptive immune response, cell proliferation induced stress and DNA damage are considered. Gene sets corresponding to the stress types are collected from the literature and the Msigdb gene sets, which cover KEGG, Biocarta, Reactome pathways and GO entries [30]. Marker genes for altered glucose metabolism, hypoxia, ER stress, dysregulated ECM, apoptosis, innate immune response, adaptive immune response, cell proliferation induced stress and DNA damage can be found in public database/literature while marker genes for lactate acidosis and oxidative stress have been studied in our previous work [31] and other studies [32]. Enzymes that catalyze the irreversible reactions of NADP+/NADPH and NAD+/NADH are used as markers for NADPH deficiency and NAD+/NADH imbalance. More detailed marker genes for these stresses are given in Supplementary Table 6.

Pathway enrichment is assessed using a hypergeometric test.

Reliability assessment of employed methodologies

We have conducted the following analyses to assess the prediction reliability of two key methods used in our study, namely (i) a method for predicting the relative order of cancer samples in terms of their relative level of progression and (ii) a method for assessing the statistical significance of the predicted co-expression.

For (i), we used gene-expression data of colon cancer to assess the quality of the predicted ordering, where colon cancer is selected since it has the lowest concordance between predicted and pathologist-assigned cancer stages (see Table 1). The idea of our assessment is to compare a number of characteristics of the predicted co-expression networks based on the predicted *versus* the assigned stages, and assess which of the two offers better *continuity* of each of these characteristics as a cancer progresses (as defined by the

two approaches).

The colon sample set used contains 34 samples at stage 1, 89 samples at stage 2, 64 samples at stage 3, and 32 samples at stage 4. We used different sampling sizes: R=15, 20, 25 and 30, and generated a set of co-expression networks for a fixed R value and do this for 100 times to assess the quality of the predicted networks based on the two staging approaches.

First we noted that when using R = 25 and 30, the predicted co-expression networks based on the two staging approaches have similar network characteristics as discussed in the RESULTs section except that the average network size, 588, of using predicted stages is considerably larger than that based on pathologist-assigned stages, 397, suggesting that samples at each assigned stage have a higher level of consistency (vs those at each pathologist-assigned stage) in terms of their stress patterns. In addition, 5-25% of the sampling results did not give rise to any co-expression networks having better than the p-value cutoff based on the pathologist-assigned stages when using R=15 and 20.

Based on these data, we consider that our predicted stages offer better characterization of disease progression. Supplementary Figure 5 shows the network size and the number of CAS associated genes in predicted co-expression networks based on pathologist-assigned stages.

For (ii), we have assessed the significance of our predicted LDHA-centric networks by randomly selecting 100 genes that are not closely related to cancers to build a co-expression network centered on each of these genes (*versus* on LDHA as we have done before) and then compare the network size and numbers of CAS genes in the 100 networks with those in the LDHA-centric networks for each of the nine cancer types. We found that the network size and number of CAS genes for the LDHA-centric networks is much larger than those of co-expression networks centered on each of the 100 randomly selected genes. Supplementary Figure 6 shows the comparisons between these numbers, while the detailed list of the 100 randomly selected genes are given in Supplementary Table 7.

Acknowledgements

The authors want to thank Mr. Xin Ying and Professor Wenxuan Zhong of the University of Georgia for their suggestions in the statistical method part. Z.C. thanks Professor Hong Qu from the center for

bioinformatics at Peking University for her mentoring of the gene co-expression network analysis in 2007-2010. Y.X. thanks Georgia Research Alliance and the University of Georgia Research Foundation for the Endowment provided to his chair position.

Funding Information

This work is supported by Georgia Research Alliance, United States and the Technology Development Plan Project of Shandong Province, China (Grant No. 2014GSF1181).

References

1. Koppenol, W.H., P.L. Bounds, and C.V. Dang, *Otto Warburg's contributions to current concepts of cancer metabolism.* Nat Rev Cancer, 2011. **11**(5): p. 325-37.
2. Vander Heiden, M.G., L.C. Cantley, and C.B. Thompson, *Understanding the Warburg effect: the metabolic requirements of cell proliferation.* Science, 2009. **324**(5930): p. 1029-33.
3. Lunt, S.Y. and M.G. Vander Heiden, *Aerobic glycolysis: meeting the metabolic requirements of cell proliferation.* Annu Rev Cell Dev Biol, 2011. **27**: p. 441-64.
4. Pfeiffer, T., S. Schuster, and S. Bonhoeffer, *Cooperation and competition in the evolution of ATP-producing pathways.* Science, 2001. **292**(5516): p. 504-7.
5. Webb, B.A., et al., *Dysregulated pH: a perfect storm for cancer progression.* Nat Rev Cancer, 2011. **11**(9): p. 671-7.
6. Fischer, K., et al., *Inhibitory effect of tumor cell-derived lactic acid on human T cells.* Blood, 2007. **109**(9): p. 3812-9.
7. Bronte, V., *Tumor cells hijack macrophages via lactic acid.* Immunol Cell Biol, 2014. **92**(8): p.

647-9.

8. Beckert, S., et al., *Lactate stimulates endothelial cell migration.* Wound Repair Regen, 2006. **14**(3): p. 321-4.

9. Sattler, U.G. and W. Mueller-Klieser, *The anti-oxidant capacity of tumour glycolysis.* Int J Radiat Biol, 2009. **85**(11): p. 963-71.

10. Groussard, C., et al., *Free radical scavenging and antioxidant effects of lactate ion: an in vitro study.* J Appl Physiol (1985), 2000. **89**(1): p. 169-75.

11. Hirschhaeuser, F., U.G. Sattler, and W. Mueller-Klieser, *Lactate: a metabolic key player in cancer.* Cancer Res, 2011. **71**(22): p. 6921-5.

12. Liu, Y., *Fatty acid oxidation is a dominant bioenergetic pathway in prostate cancer.* Prostate Cancer Prostatic Dis, 2006. **9**(3): p. 230-234.

13. Doherty, J.R. and J.L. Cleveland, *Targeting lactate metabolism for cancer therapeutics.* J Clin Invest, 2013. **123**(9): p. 3685-92.

14. Jungmann, R.A., D. Huang, and D. Tian, *Regulation of LDH-A gene expression by transcriptional and posttranscriptional signal transduction mechanisms.* J Exp Zool, 1998. **282**(1-2): p. 188-95.

15. Firth, J.D., B.L. Ebert, and P.J. Ratcliffe, *Hypoxic regulation of lactate dehydrogenase A Interaction between hypoxia-inducible factor 1 and cAMP response elements.* Journal of Biological Chemistry, 1995. **270**(36): p. 21021-21027.

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

17. Wu, H., et al., *Central role of lactic acidosis in cancer cell resistance to glucose deprivation‐induced cell death.* The Journal of pathology, 2012. **227**(2): p. 189-199.

18. Fischer, K., et al., *Inhibitory effect of tumor cell–derived lactic acid on human T cells.* Blood, 2007. **109**(9): p. 3812-3819.

19. van Horssen, R., et al., *Cancer cell metabolism regulates extracellular matrix degradation by invadopodia.* Eur J Cell Biol, 2013. **92**(3): p. 113-21.

20. Halprin, K.M. and A. Ohkawara, *Lactate Production and Lactate Dehydrogenase in the Human Epidermis1*. The Journal of Investigative Dermatology, 1966. **47**(3): p. 222-226.

21. Malhotra, J.D. and R.J. Kaufman, *Endoplasmic reticulum stress and oxidative stress: a vicious cycle or a double-edged sword?* Antioxid Redox Signal, 2007. **9**(12): p. 2277-93.

22. Pani, G., T. Galeotti, and P. Chiarugi, *Metastasis: cancer cell's escape from oxidative stress.* Cancer and Metastasis Reviews, 2010. **29**(2): p. 351-378.

23. Cairns, R.A., I.S. Harris, and T.W. Mak, *Regulation of cancer cell metabolism.* Nat Rev Cancer, 2011. **11**(2): p. 85-95.

24. Kanehisa, M., *The KEGG database.* Novartis Found Symp, 2002. **247**: p. 91-101; discussion 101-3, 119-28, 244-52.

25. Wu, W., et al., *Lactate down-regulates matrix systhesis and promotes apoptosis and autophagy in rat nucleus pulposus cells.* J Orthop Res, 2014. **32**(2): p. 253-61.

26. Wong, N., J. De Melo, and D. Tang, *PKM2, a Central Point of Regulation in Cancer Metabolism.* Int J Cell Biol, 2013. **2013**: p. 242513.

27. Cancer Genome Atlas Research, N., et al., *The Cancer Genome Atlas Pan-Cancer analysis project.* Nat Genet, 2013. **45**(10): p. 1113-20.

28. Friedman, J., T. Hastie, and R. Tibshirani, *Regularization Paths for Generalized Linear Models via Coordinate Descent.* J Stat Softw, 2010. **33**(1): p. 1-22.

29. Obayashi, T., et al., *COXPRESdb: a database of comparative gene coexpression networks of eleven species for mammals.* Nucleic Acids Res, 2013. **41**(Database issue): p. D1014-20.

30. Subramanian, A., et al., *Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.* Proc Natl Acad Sci U S A, 2005. **102**(43): p. 15545-50.

31. Xu, K., et al., *Elucidation of How Cancer Cells Avoid Acidosis through Comparative Transcriptomic Data Analysis.* Plos One, 2013. **8**(8).

32. Hayes, J.D. and M. McMahon, *NRF2 and KEAP1 mutations: permanent activation of an adaptive response in cancer.* Trends Biochem Sci, 2009. **34**(4): p. 176-88.

Figures

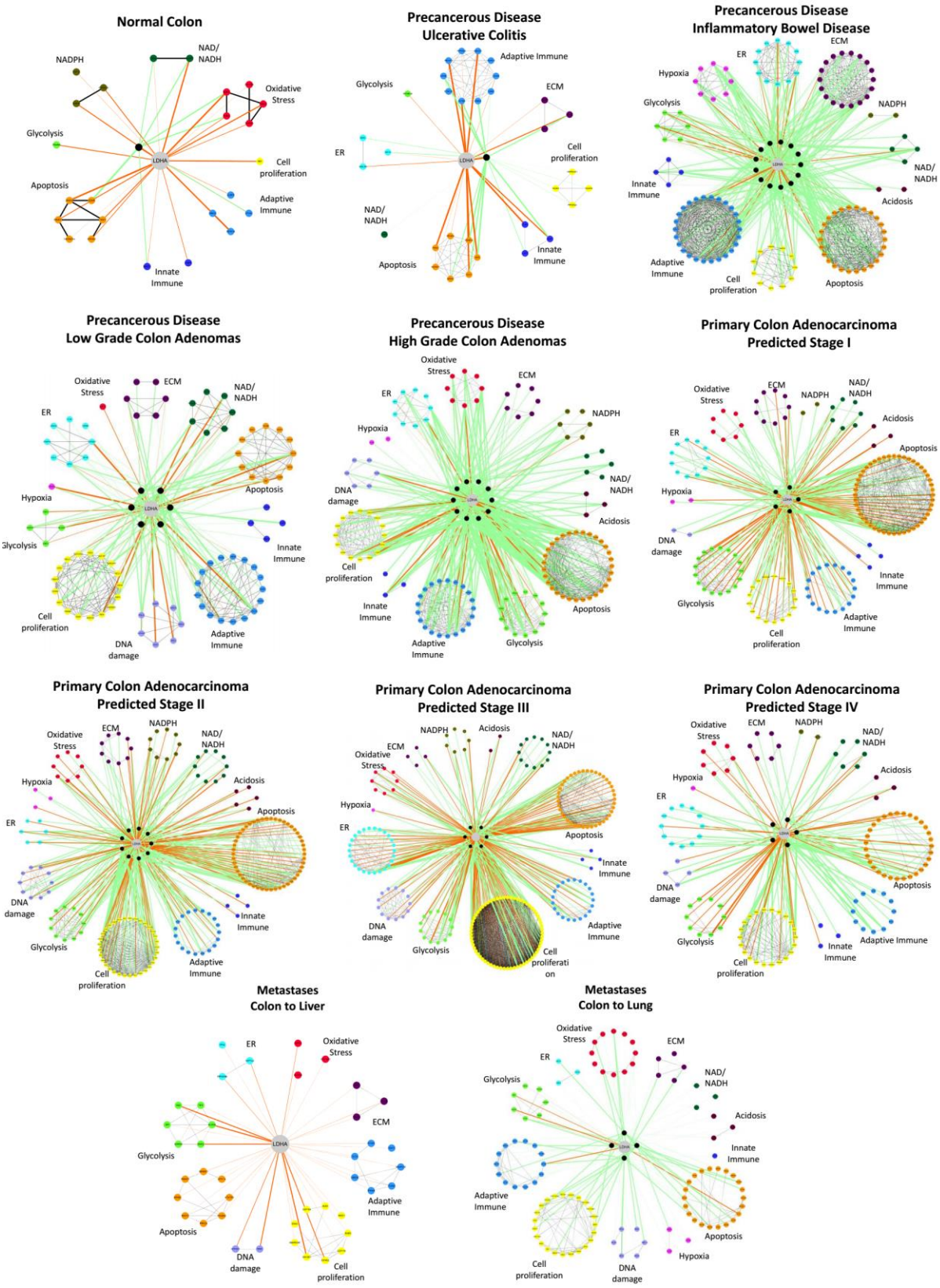

Normal Colon
NADPH
NAD/ NADH
Oxidative Stress
Glycolysis
Cell proliferation
Apoptosis
Adaptive Immune
Innate Immune
Precancerous Disease Ulcerative Colitis
Adaptive Immune
Glycolysis
ECM
ER
Cell proliferation
NAD/ NADH
Innate Immune
Apoptosis
Precancerous Disease Inflammatory Bowel Disease
ECM
ER
Hypoxia
Glycolysis
NADPH
NAD/ NADH
Innate Immune
Acidosis
Adaptive Immune
Cell proliferation
Apoptosis
Precancerous Disease Low Grade Colon Adenomas
Oxidative Stress
ECM
NAD/ NADH
ER
Apoptosis
Hypoxia
Glycolysis
Innate Immune
Cell proliferation
DNA damage
Adaptive Immune
Precancerous Disease High Grade Colon Adenomas
Oxidative Stress
ECM
ER
NADPH
Hypoxia
NAD/ NADH
DNA damage
Acidosis
Cell proliferation
Apoptosis
Innate Immune
Adaptive Immune
Glycolysis
Primary Colon Adenocarcinoma Predicted Stage I
Oxidative Stress
ECM
NADPH
NAD/ NADH
Acidosis
ER
Apoptosis
Hypoxia
DNA damage
Innate Immune
Glycolysis
Adaptive Immune
Cell proliferation
Primary Colon Adenocarcinoma Predicted Stage II
Oxidative Stress
ECM
NADPH
NAD/ NADH
Hypoxia
Acidosis
ER
Apoptosis
DNA damage
Innate Immune
Glycolysis
Adaptive Immune
Cell proliferation
Primary Colon Adenocarcinoma Predicted Stage III
NADPH
Acidosis
NAD/ NADH
ECM
Oxidative Stress
Hypoxia
ER
Apoptosis
Innate Immune
DNA damage
Adaptive Immune
Glycolysis
Cell proliferation
Primary Colon Adenocarcinoma Predicted Stage IV
Oxidative Stress
ECM
NADPH
NAD/ NADH
Hypoxia
Acidosis
ER
Apoptosis
DNA damage
Glycolysis
Innate Immune
Adaptive Immune
Cell proliferation
Metastases Colon to Liver
ER
Oxidative Stress
ECM
Glycolysis
Adaptive Immune
Apoptosis
DNA damage
Cell proliferation
Metastases Colon to Lung
Oxidative Stress
ECM
ER
Glycolysis
NAD/ NADH
Acidosis
Innate Immune
Adaptive Immune
Apoptosis
Hypoxia
Cell proliferation
DNA damage

Figure 4.1. LDHA-centric co-expression networks of colon cancer at different stages. In each network, the grey node at the center is the LDHA gene and the black nodes around LDHA are hub genes identified in the network, where a hub gene is defined as genes with substantially more neighbors than other genes in the neighborhood (see Method for details). Genes related to each stress type are distinctly colored and are defined as stress module (see METHODS). Co-expressions (see METHODS) between LDHA and other genes, between the hub and the other genes, and among genes in each stress module are represented by orange, green and black edges, respectively. The significance of the enrichment of each stress module in each network is given in Table 4. The thickness of each edge is explained in the Supplementary Method. The number of hub genes in each network is given in Supplementary Table 1.

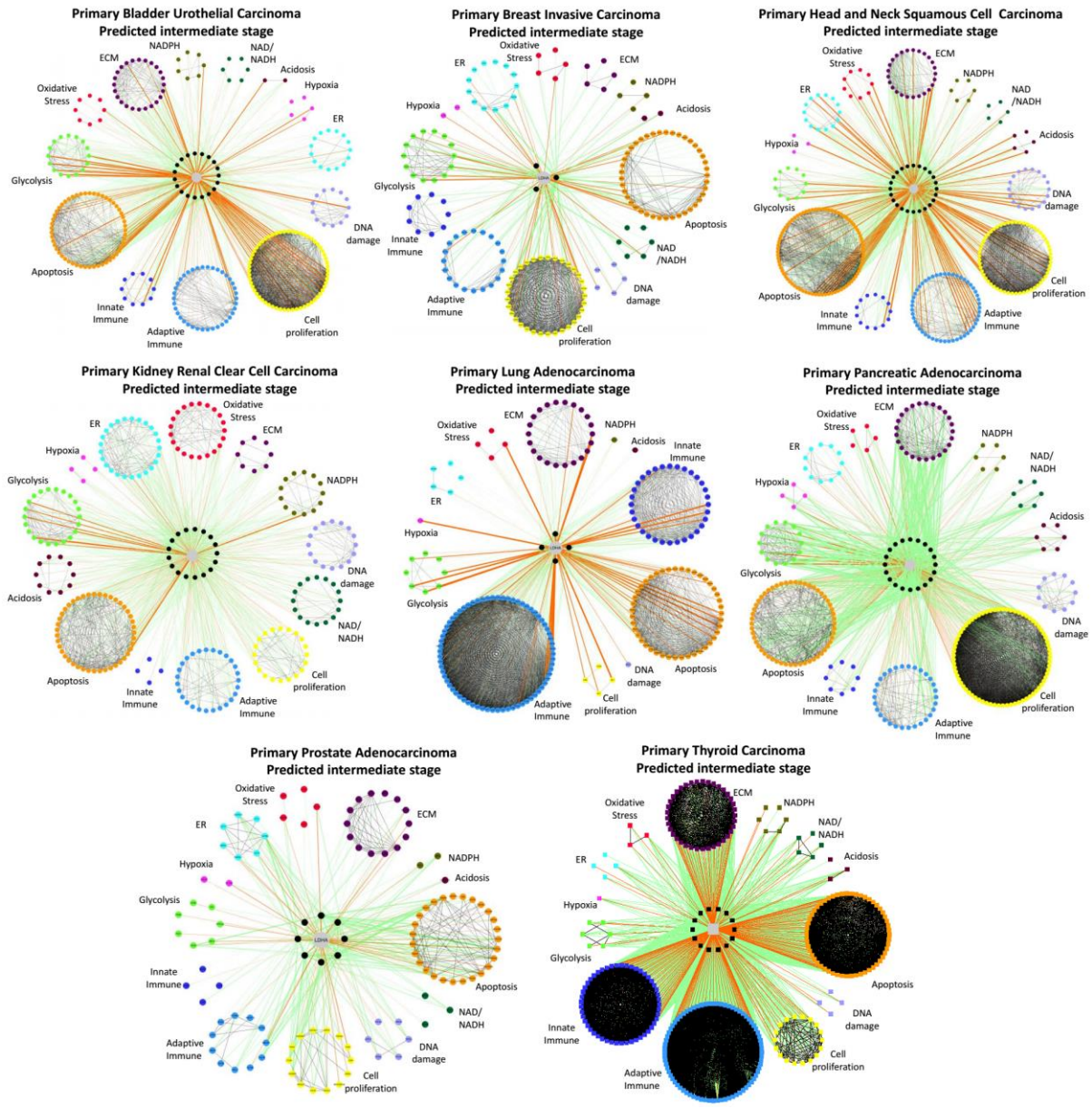


Figure 4.2. A summary of LDHA-centric co-expression networks for the intermediate-stage samples in nine cancer types. Co-expression between LDHA and other genes, between the hub genes and other genes, and among genes in each network are represented by orange, green and black edges, respectively. The significance of the enrichment of each stress module in the networks here is given in Table 4. The thickness of each edge is defined in the Supplementary Method.

Supplementary Methods

Co-expression module analysis

Most of the current co-expression network construction methods tend to focus on identification of the global properties of a network. Inclusion of local co-expression modules (or subnetworks) represents one way to extend the current methods for co-expression network construction. Here we have developed a novel computational method to identify co-expression modules that are each co-expressed with a specific target gene, where a *co-expression module* is defined as a maximal set of genes among which co-expressed gene pairs exist significantly more frequently than those by chance. To identify the co-expression modules for a given gene, we used a novel measure, *mutual rank (MR)*, to rank genes based on the number of genes co-expressed with them vs such numbers of their neighboring genes. Consider a set of genes and their pairwise co-expression relationships, represented as nodes and edges, respectively, the mutual rank between gene (node) *i* and gene (node) *j* is defined as follows, which is borrowed from ***'s work[29]:

$$\mathrm{MR(i,j)} = \mathrm{MR(j,i)} = \sqrt{Rank(i \to j) \cdot Rank(j \to i)}\,,$$

where $Rank(i \to j)$ is the rank of the distance from gene *i* to gene *j* among all distances from *i* to other genes in the given co-expression network, where distance between *i* and *j* could be any distance or weight applied to measure the co-expression strength between the two genes. We will show the null distribution of MR in our method is independent to the null distribution of the distance measure. In general, a smaller MR

value generally has a higher rank for the relevant genes (nodes). We use this measure to identify *hub* genes, i.e., genes with substantially more neighbors than other genes in the neighborhood, which may have more small MR values with other genes comparing to non-hub genes.

Searching for hub genes (property): For a given node i and a threshold value T, define $\mathrm{N}_{\mathrm{MR}}(i,T) = \left|\left\{j \,\middle|\, j \neq i,\ MR(i,j) < T^{\frac{1}{2}}\right\}\right|$. Note that the distance between a gene and a hub gene should have a relatively top rank, and a top ranked $Rank(j \to hub)$ should generally have a relatively low increase rate of $\mathrm{MR}(\mathrm{hub}, \mathrm{j})$. Hence for a certain threshold T, a hub gene should have more MRs smaller than $\mathrm{T}^{\frac{1}{2}}$ compared to its neighboring genes. By such a property, we specifically define the growth function of the MR by the following definitions. The growth of the MR of hub genes are smaller than the MR growth of other genes.

Definition 1. Define $MR_{i,(j)}, j = 1 \ldots N$ as the sorted $MR_{i,j}$ by increasing order.

Definition 2. Define the growth function of the $MR_{i,(j)}$ in from index m to n by step p as:

$$Growth\left(MR_{i,(j)}, step\right) = \{G_1, G_2, \ldots G_n\}$$

$$G_k = \frac{MR_{i,(j_{k+})} - MR_{i,(j_{k-})}}{\mathrm{j_{k+}} - \mathrm{j_{k-}} + 1}$$

$$\mathrm{j_{k-}} = \max(k - \frac{step}{2}, 0), \mathrm{j_{k+}} = \min(k + \frac{step}{2}, n)$$

Figure 2 shows the growth rate (defined in theorem 3) of MRs of hub and non-hub genes (nodes in randomly generated networks) are substantially different. Based on this property, we have developed the following two procedures, one non-parametric and one parametric, to identify hub genes in a co-expression network.

The idea is to simulate the null distribution of MR of randomized correlation matrix, hence the null distribution of MR growth, which could be applied to calculate the significance of the growth rate of MRs from real data as shown in Figure 1.

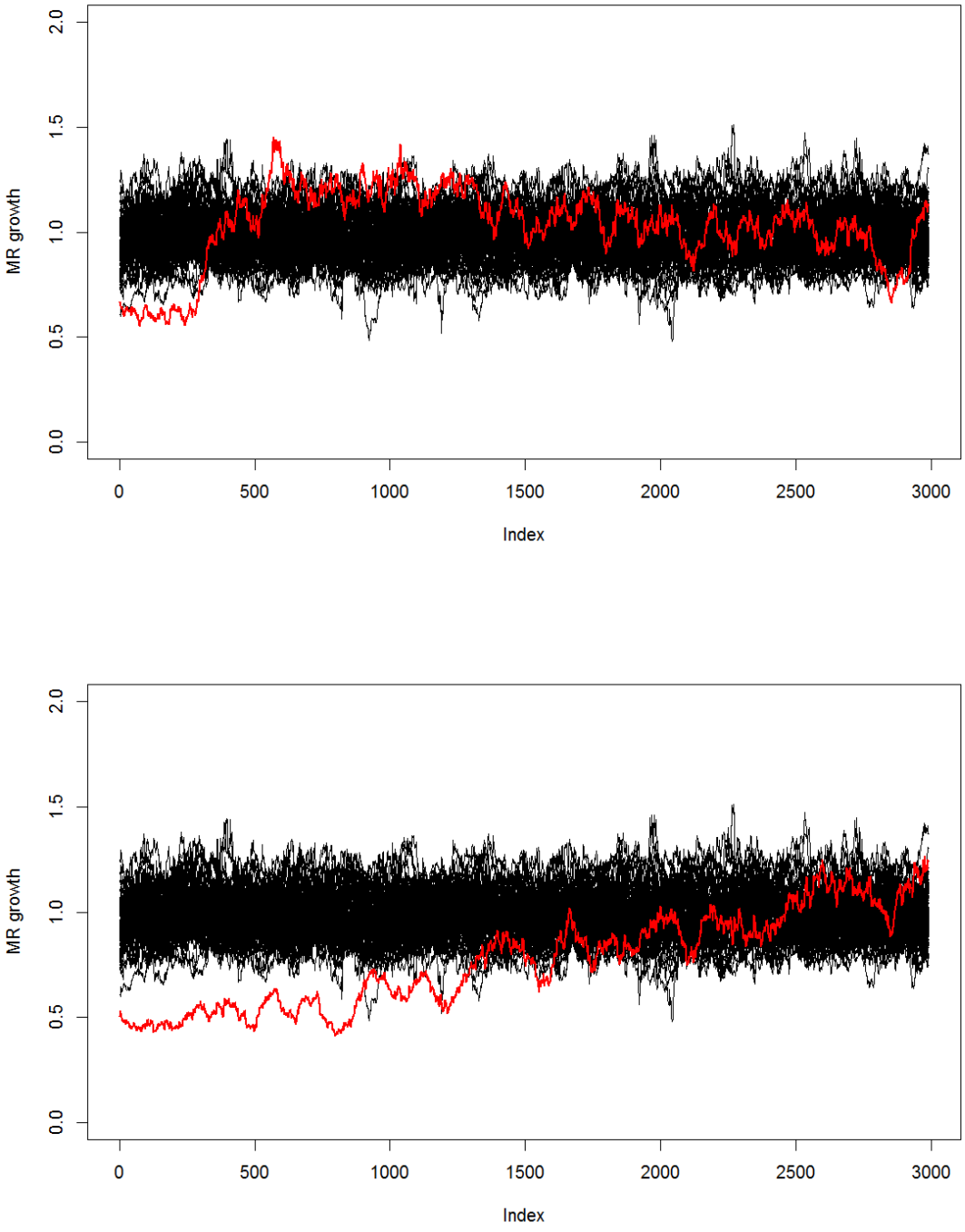


Figure 4.1S. The growth of MR of the randomly simulated correlation matrix (black) and a hub gene in one E coli gene coexpression network (red). The growth is defined as in definition 2. It is worth

to note that he top 300 MRs in the first plot and the top 1200 MRs in the second plot are significant lower than the null distribution, suggesting the size of modules of the two hubs are around 300 and 1200, respectively.

Before we move on, we have the following definition for a randomly assigned distance matrix and a randomly generated distance matrix D':

Definition 3. A randomly assigned distance matrix D is defined by:

$$\mathrm{D_{i,j}} = \mathrm{D_{j,i}} > 0, i \neq j, \mathrm{i} = 1..\mathrm{N}, \mathrm{j} = 1..\mathrm{N}$$

$$\mathrm{D_{i,i}} = 0, \mathrm{i} = 1..\mathrm{N}$$

$$\mathrm{D_{i,j}} \sim f(x)\ iid\ for\ different\ i\ or\ j$$

where $f(x)$ is the null (partial) distribution of the distance.

Definition 4. A randomly generated distance matrix D' corresponding to the original data matrix X is defined by:

1. Shuffle the sample index of each row of X to generated the randomly generated data matrix X'
2. D' is computed as the distance matrix of X'

(I think for Pearson Correlation Coefficients, which is the inner product of two vectors, should satisfy this definition if the vectors are independently generated. If not, the covariance among the Di,j should be very weak)

(Or we need to prove that the distance matrix calculated from a randomly generated gene expression profile should be very close to a randomly assigned distance matrix)

Identification of the hub genes:

A non-parametric procedure*:* For a given network with N nodes having a co-expression (weight) matrix S, we assess the significance of the N_{MR} value for any node *i* by comparing it with the N_{MR} value computed with randomly assigned co-expression levels.

c) Repeat the following for N times: randomly shuffle the index of each gene, assign the weight matrix S and re-compute $MR\ and\ MR\ growth$ of node *i;* and create a histogram of the computed $MR\ growth$ values.

d) Estimate the nominal P-value for node *i*'s $MR\ growth$ and threshold T based on the above estimated histogram.b

A parametric procedure: While a non-parametric procedure is preferable, it becomes computationally prohibitive for large networks (e.g., co-expression networks with $>10^6$ nodes, which may happen when alternatively spliced genes are considered when analyzing human transcriptomes). To overcome this issue, we have developed a parametric model based on an assumption about the distribution of the edge weights of the given network. We first demonstrate that the MR value follows a distribution as shown below, and then present a parametric procedure for estimating the P-value of a node *i*'s N_{MR}. Specifically we have mathematically proved that

$$\mathrm{P}(Rank(i \to j) = x, Rank(j \to i) = y) =$$

$$\int (C_{N-1}^{x-1}(1 - F(k))^{x-1}F(k)^{N-x} + 1) \cdot (C_{N-1}^{y-1}(1 - F(k))^{y-1}F(k)^{N-y} + 1) \cdot f(k)\, dk \quad (*)$$

in which $f(k)$ and $F(k)$ denote the pdf and cdf of the null distribution of the original distances, respectively. By formula (*), we have further proved that the distribution $\mathrm{P}(Rank(i \to j) = x, Rank(j \to i) = y)$ does not depend on the distribution of S_{ij}. The empirical null distribution of $\mathrm{MR(i, j)}$ can be simulated by

$$MR_{i,j} = \sqrt{(X_{i,j,1} + 1) \cdot (X_{i,j,2} + 1)}$$

$$X_{i,j,1} \sim Binom(n - 1, \mathrm{p_{ij}}), X_{i,j,2} \sim Binom(n - 1, \mathrm{p_{ij}})$$

$$\mathrm{p_{ij}} \sim U(0,1)$$

Parametric hub genes identification procedure

We have also developed a parametric model based on an assumption about the distribution of the edge weights of the given network. Specifically, we have mathematically proven that the distribution of $\mathrm{P}(Rank(i \to j) = x, Rank(j \to i) = y)$ is

$$\int (C_{N-1}^{x-1}(1 - F(k))^{x-1}F(k)^{N-x} + 1) \cdot (C_{N-1}^{y-1}(1 - F(k))^{y-1}F(k)^{N-y} + 1) \cdot f(k)\, dk$$

in which $f(k)$ and $F(k)$ denote the pdf and cdf of the null distribution of the original distances, respectively.

Lemma 1:

For the N*N randomly assigned matrix S, $\text{Rank}(i \to j) \perp Rank(j \to i) | S_{ij}$.

Proof: With given S_{ij}, $\text{Rank}(i \to j)$ only depends on $S_{i\cdot}, \cdot \neq j$ while $Rank(j \to i)$ only depends on $S_{\cdot j}$, $\cdot \neq i$. Since $S_{i\cdot} \perp S_{\cdot j}$ for $i \cdot \neq ij$ and $\cdot j! = ij \;\rightarrow\; \text{Rank}(i \to j) \perp Rank(j \to i) | S_{ij}$ □

Proposition 1:

For the N*N matrix randomly assigned from the distribution with pdf f and cdf F, the joint distribution of $Rank(i \to j)\ and\ Rank(j \to i)$ is:

$$\text{P}(Rank(i \to j) = x, Rank(j \to i) = y) =$$

$$\int (C_{N-1}^{x-1}(1 - F(k))^{x-1} F(k)^{N-x} + 1) \cdot (C_{N-1}^{y-1}(1 - F(k))^{y-1} F(k)^{N-y} + 1) \cdot f(k)\, dk$$

Proof:

$$Prob\big(Rank(i \to j) = x \big| S_{ij} = k\big) =$$

$$Prob\big(S_{ij_m} > k, S_{ij_n} < K, m = 1 \dots x-1, n = x+1 \dots N\big) =$$

$$C_{N-1}^{x-1}(1 - F(k))^{x-1} F(k)^{N-x} + 1$$

,

$$, in\ which\ j_m\ and\ j_n\ are\ permuting\ of\ 1 \dots j-1, j+1 \dots N$$

By lemma 1,

$$P\big(Rank(i \to j), Rank(j \to i) | S_{ij}\big) =$$

$$P\big(Rank(i \to j) | S_{ij}\big) * P\big(Rank(j \to i) | S_{ij}\big)$$

$$\mathrm{P}\big(Rank(i \to j) = x, Rank(j \to i) = y \big| S_{ij} = k\big) =$$

$$(C_{N-1}^{x-1}(1 - F(k))^{x-1}F(k)^{N-x} + 1) \cdot (C_{N-1}^{y-1}(1 - F(k))^{y-1}F(k)^{N-y} + 1)$$

Hence

$$\mathrm{P}(Rank(i \to j) = x, Rank(j \to i) = y) =$$

$$\int (C_{N-1}^{x-1}(1 - F(k))^{x-1}F(k)^{N-x} + 1) \cdot (C_{N-1}^{y-1}(1 - F(k))^{y-1}F(k)^{N-y} + 1) \cdot f(k)\, dk$$

□

It is noteworthy that the distribution $\mathrm{P}(Rank(i \to j) = x, Rank(j \to i) = y)$ does not depend on the distribution of S_{ij}, hence the empirical null distributions of $\mathrm{N_{MR}}(i, T)$ can be simulated in more general cases.

Proposition 2: The distribution $\mathrm{P}(Rank(i \to j) = x, Rank(j \to i) = y)$ does not depend on the distribution of S_{ij}

Proof: Since S_{ij} follows the pdf f, by Proposition 1:

$$\mathrm{P}\big(Rank(i \to j) = x, Rank(j \to i) = y \big| S_{ij} = k\big) =$$

$$(C_{N-1}^{x-1}(1 - F(k))^{x-1}F(k)^{N-x} + 1) \cdot (C_{N-1}^{y-1}(1 - F(k))^{y-1}F(k)^{N-y} + 1)$$

Hence the $\mathrm{MR(i,j)}^2 = Rank(i \to j) \cdot Rank(j \to i)$ with given $S_{ij} = k$ follows a multiplication of two identical binomial distributions (plus 1) with n= N-1 and P=1-F(S_{ij}) ~U(0,1). To randomly generate the null distribution of MR, we randomly draw $\tilde{\mathrm{p}}$ from U(0,1) and randomly generate $Rank(i \to j)$ and $Rank(j \to i)$ from two identical binomial distribution $\mathrm{Bin}(\tilde{\mathrm{p}}, \mathrm{n} - 1)$ and then $\mathrm{MR(i,j)} =$

$\sqrt{Rank(i \to j) \cdot Rank(j \to i)}$. Hence the $P(Rank(i \to j) = x, Rank(j \to i) = y)$ does not depend on the distribution of S_{ij}.

□

Theorem 1:

Empirical null distribution of MR(i, j) can be simulated by

$$MR_{i,j} = \sqrt{(X_{i,j,1} + 1) \cdot (X_{i,j,2} + 1)}$$

$$X_{i,j,1} \sim Binom(n-2, p_{ij}), X_{i,j,2} \sim Binom(n-2, p_{ij})$$

$$p_{ij} \sim U(0,1)$$

Proof:

By Proposition 1 and Proposition 2, the random numbers simulated as described in the theorem follows the empirical null distribution of MR(i, j).

□

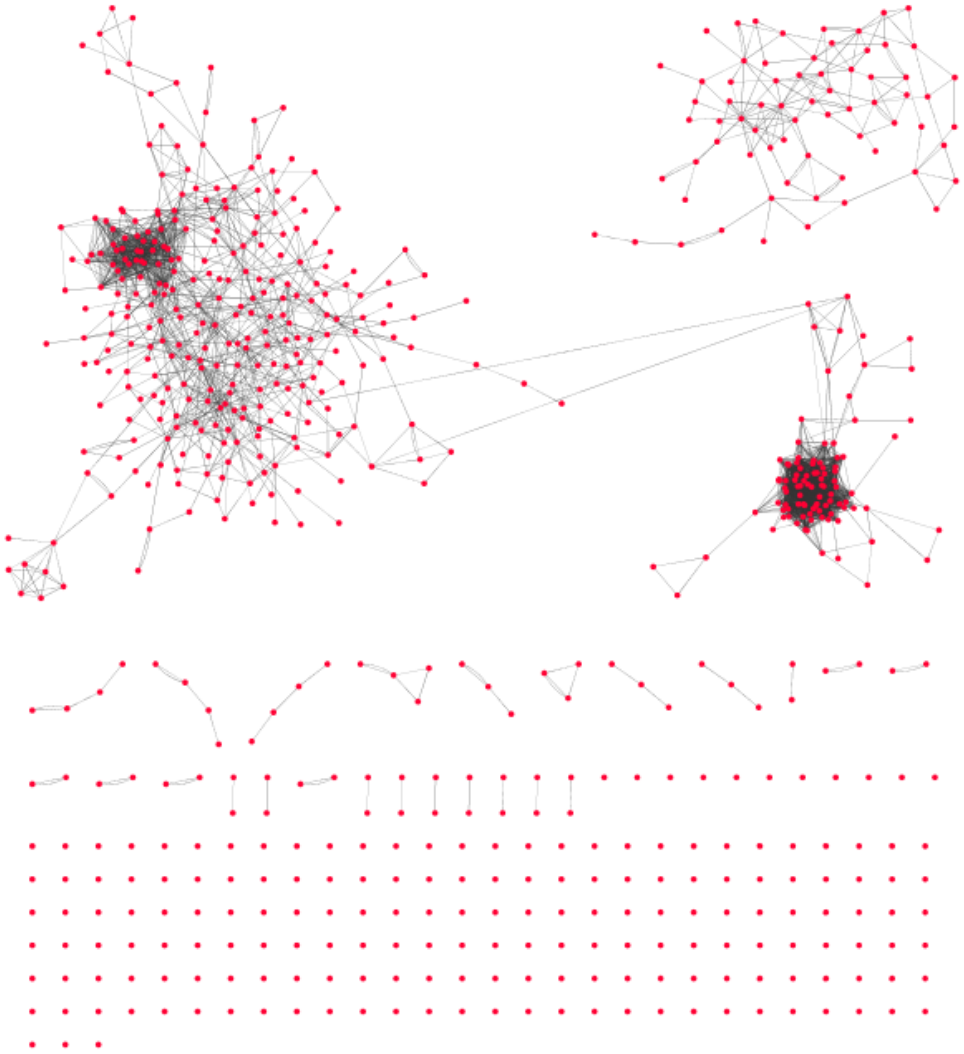


Figure 4.2S. The identified hub genes in Ecoli gene network by using ~400 sets of E coli microarray data. Links are defined by PCC>0.6 & MR<80.

Theorem 2:

The growth G_k does not depend on k (ignore the left and right tails) under fixed *step* for the $MR_{i,j}$ generated by the procedure in Theorem 1, i.e. the null distribution of MR.

Proof:

By Theorem 1. □

Construction of a co-expression network

With the identified hub genes in a global network, the local co-expression network of a gene X (e.g. LDHA in the current study) can be constructed as follows:

4) Identify genes that are significantly co-expressed with X and put them into the network, where substantial co-expressions are determined using Pearson correlation coefficients and student *t* tests by using Pearson's product moment correlation coefficient with significance level p = 1e-4 for FDR correction or MR < 100.
5) For each hub gene in the network, genes that are significantly co-expressed with the hub, as defined in (1), are put into the network.
6) A significant link is determined by the significance as defined in (1).

Pseudo Code for the algorithm to identify hub genes by using MR

1: Input: A give dependence function D(x,y), higher value of D suggests larger dependence; a data matrix X with N features and M samples

2: Initialize K<<N but depends on N, like N/100 or N^(1/2), ROUNDS=5000 for number of randomized MR values will be generated for significance testing; Initialize ROUNDS by N matrix MR_R for the randomized MR

3: Initialize N by K integer matrix RANK_E (the largest intermediate result will be stored in memory, so

K is adjusted to control the memory consumption)

4: For i <-1 to N do

4: For j<-1 to N do

5: Compute D(X[i,], X[j,])

6: Identify the top K values in D(X[i,],*) and the corresponding index $j^i_{(1)}, j^i_{(2)}, \dots, j^i_{(K)}$ and store the index in the ith row in MR_E

7: For i<-1 to ROUNDS do

8: For j<-1 to N do

9: Generate random number p from U(0,1)

10: Generate two independent random numbers X1, X2 from Binom(p,N-1)

11: if(X1>M) X1=M; if(X2>M) X2=M;

11: MR_R[i,j] = $\sqrt{(X1+1)*(X2+1)}$

12: MR_R_sorted[i,]=sort(MR_R[i,])

12: For j<-1 to floor(M*0.9)

13: compute the growth rate of MR_R_sorted[i,] by Definition 2 and store growth rate in GR_MR_R

14: For i<-1 to N do

15: Compute MR[i, *] by

16: For j <-1 to N do

17: if(i==j) MR[i,j]=infinite;

18: else

19: if(j is in RANK_E[i,]) Rank_j_i=which(RANK_E[i,]==j)

20: else Rank_j_i=M;

21: if(i is in RANK_E[j,]) Rank_i_j=which(RANK_E[j,]==i)

22: else Rank_i_j=M;

23: MR[i,j]= $\sqrt{\text{Rank\_i\_j} * \text{Rank\_j\_i}}$

24: Test the significance of MR[i, *] by

25: MR_sorted=sort(MR[i,])

25: For j<-1 to floor(M*0.9)

26: compute the growth rate of MR_sorted[i,] by Definition 2

27: compare the MR_sorted[i,] with MR_R_sorted as shown in Figure 1 and assess the significance of the MR_sorted (if MR_sorted is significantly smaller than MR_R_sorted)

28: if(significant in 27) i is a significant hub genes and the size of the modules corresponding to i can be estimate the range that MR_sorted is significantly smaller than MR_R_sorted.

In the pseudo code: 1-3 are algorithm initialization, 4-6 compute RANK_E matrix for further MR computation, 7-13 generate null distribution of MR and MR growth rate, 14-28 compute MR and MR growth for each feature i in the given data and test the significance of i for the hub property of i.

CHAPTER V

A BI-CLUSTERING BASED APPROACH TO PREDICT THE FUNCTIONAL GAIN OR LOSS OF SOMATIC MUTATIONS (ONGOING PROJECT)

Abstract

While the traditional view of cancer-contributing mutations as oncogenes or tumor-suppressors has provided a useful framework for linking mutations to cancer, it is clearly over-simplifying as the majority of mutations do not "drive" cancer nor inhibit apoptosis/cell cycle progression; instead they play many other roles to facilitate adaptation to specific stresses. In this work, we have developed a bi-clustering based approach by integrating the somatic mutation and RNAseq data to comprehensively predict the heterogeneous gain or loss of function led by possible mutation patterns of a certain mutation or collective effect of multiple mutations. By applying the method to 18 known cancer associated mutations in 20 cancer types in TCGA data, heterogeneous gain or loss of functions of each mutation associated to certain mutation patterns in each cancer type will be comprehensively predicted.

Introduction

Considering cancer initiation and progression as an evolutionary process, the cancer associated gene mutations, or "driver mutations", are selected to enhance the cancer's survival by increasing the capability of the tissue to overcome certain stress types in tissue microenvironments. Gain or loss of function of multiple cancer associated mutations including TP53, KRAS, VHL, and PI3KCA are well acknowledged [1-4]. Such observations indicate that same mutation may serve different roles in different cancer tissues. On the other hand, same mutation may be selected by different reasons [5]. It is worth to note that most of the "driver" mutations, especially the suppressor genes as TP53 and VHL always involve in large set of pathways with distinct cellular functions, suggesting possible heterogeneous gain or loss of functions can be led by the mutation. Unfortunately, such heterogeneous gain or loss of functions are excluded in the assumption of most of the current mutation identification method [6, 7]. To our best knowledge, there is no method assumes that the mutation may cause heterogeneous functional gain or loss in a certain subset of the mutation samples.

In this work, we have developed a bi-clustering based method to predict such heterogeneous gain or loss of functions led by specific mutations on a single gene or interactive effect of multiple mutations. For a given mutated gene and a cancer type, we get all the transcriptomic data of the N samples of the cancer from the TCGA database, with N_1 samples having mutations and N_2 having no mutations in the gene, with $N_1 + N_2 = N$. The expression pattern of gene X is considered to be affected by the mutated gene if the distribution of X's expression levels over the N samples is significantly different from its distribution over the N_2 samples without mutations, for any X in the genome. Currently, two types of significant differences

are considered: the N_1 samples significantly enrich the highest (or lowest) expressed genes among the N samples; or the distribution of X's expressions over the N samples has a distinct peak that is significantly enriched by the N_1 samples. Our goal here is to determine which pathways are enriched by genes with altered expression distributions due to a specific mutated gene in some subset of the N_1 samples. This problem can be formulated as a bi-clustering problem and solved using our own program QUBIC [8], with each bi-cluster consisting of a set of samples sharing common functional losses/gains, measured using pathways with altered expression patterns, which will be followed with a pathway enrichment analysis coupled with a statistical significance assessment. The result of this analysis for each target mutated gene is a set of pathways in an integrated set of the KEGG, MisgDB, TransPath and GO pathways, with altered expression patterns over some subsets of the N_1 samples due to the mutated gene [9-11]. We have applied our method on TCGA data to examine the possible heterogeneous gain or loss of function of 18 cancer associated mutations in 20 cancer types. The predicted functional changes, loss or gain, by p53 mutations and APC mutations in colorectal adenocarcinoma are highlighted in the result part. It is worth to note this work is still ongoing. This chapter will be developed into a formal academic paper.

Results

Problem formulation and analysis pipeline

To identify the heterogeneous gain or loss of function (GoLoF) of a certain mutation, we developed a bi-clustering based method by integrative analysis of high throughput transcriptomics data and somatic mutation prolife. Noting that acquiring or losing of specific functions can be reflected on the gene

expression level of the function related genes and the samples with a specific gain or loss of function is an unknown subset among all the samples, we formulate the problem as a bi-clustering problem, i.e. simultaneously identifying gene expression patterns that are significantly associated to a sub set of the mutation samples. In detail, the method predicts the possible GoLoF of a given mutation by the following four steps as shown in Figure 1: (I) analysis of transcriptomic profile to filter out the mutation associated gene expressions; (II) discretizing the expression level of the mutation associated gene into possible GoLoF associated expression patterns; (III) bi-clustering analysis to identify possible GoLoF and corresponding samples and (IV) assessment of the significance of the identified GoLoF by protein structure and known protein-protein interaction (PPI) network. Detailed computation procedures and parameters are given in Method part and Supplementary Method.

Step I (Identify the mutation associated gene expression)

Mixed Gaussian model with left truncated assumption is first fit to the RSEM normalized gene expression level to identify the number of peaks in each gene's expression profile. For a given mutation, the mutation associated gene is determined by if the gene is significantly differential expressed between the mutation samples and non-mutations samples or at least one peak in the gene expression profile is significantly associated with the mutation.

Step II (Data discretization for possible GoLoF associated expressions)

The expression profile of each mutation associated gene is then discretized into 1/0 values indicating

a specific gene expression pattern associated with the mutation. For the mutation associated gene expression with a single peak, a Kolmogorov statistics based approach is applied to identify the mutation associated over/under expression. For the gene expression with multiple peaks, fisher exact test is applied to identify the mutation associated peak. For each mutation associated over/under expression or peak, samples with the pattern are assigned by 1 while the other rest are assigned by 0. The discretized data are merged with the mutation profile of other genes and input to the bi-cluster step.

Step III (Bi-clustering to identify GoLoF and corresponding samples)

We apply our in-house bi-clustering method QUBIC 1.0 to identify the bi-clusters by using the discretized input data. To ensure the significance of the identified bi-clusters, relatively strict parameters are applied in the computation (See More in Method part). The identified bi-clusters contain other mutations may correspond to possible interactive effect of multiple mutations, the large bi-clusters cover most of the mutation samples are considered as the general effect of the mutation while the rest may correspond to possible GoLoF. Functions of the genes and the mutation types and positions of the samples in each of the identified bi-clusters are then examined to elucidate the detailed GoLoF of each bi-cluster and further assess the significance of the prediction.

Step IV (Assessment of prediction significance)

Significance of the identified GoLoF is assessed on three levels by: 1) examining consistence of the mutation types or mutation positions on protein secondary or tertiary structure level; 2) examining pathways

enrichment of the genes in the identified bi-clusters and 3) comparing the predicted GoLoF with the known PPI network of the mutation. For the predicted interactive effect of multiple mutations, pathway enrichment and comparison with PPI are applied to assess the prediction significance.

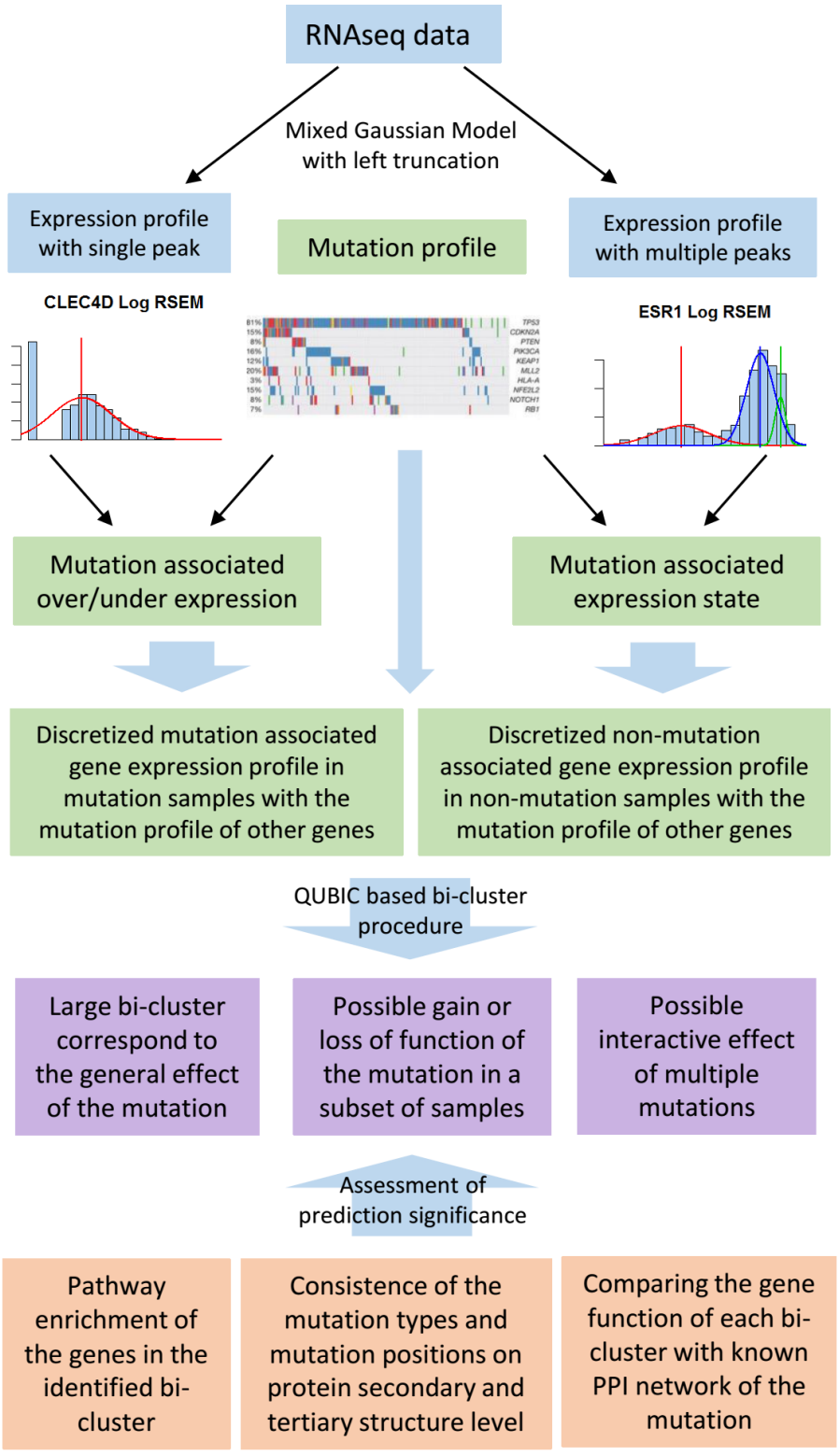


Figure 5.1. Analysis pipeline

Analysis of pan-cancer patterns of GoLoF and interactive effect of multiple mutations

We have applied our method on the TCGA data of 20 cancer types to identify possible GoLoF and interactive effect of multiple mutations for 14 known cancer associated genes and 4 frequently mutated genes among the examined samples. The selected cancer types and mutations are plotted in Figure 3.

After identify the bi-clusters may correspond to certain gain or loss of functions, the detailed functions are captured by pathway enrichment analysis on the identified genes in each bi-cluster and the mutations are analyzed by several tests for the consistence of mutation patterns. Currently, we have considered the following mutation patterns and their combinations that may possibly lead to a heterogeneous functional gain or loss: 1) mutations that are significantly enriched on a certain exon or several exons, 2) mutations that are significantly enriched on one or several mutation sites, 3) mutations that are significantly enriched on one or several functional regions on the protein tertiary structure, 4) mutations that are significantly enriched on one or several mutation types, 5) con-current mutations, and 6) collect effect of two or multiple mutations.

Functional loss/gain by p53 mutations in colorectal adenocarcinoma tissues.

A bi-clustering analysis was conducted on the transcriptomic data of all the colon cancer samples in TCGA, 121 of which having p53 mutations. A few dozen of bi-clusters were identified, each representing a subset of the samples with specific altered functions due to p53 mutations. Some bi-clusters each represent one type of functional change while others are more complex and need further analyses. We now highlight six such bi-clusters, each consisting of a set of functionally related pathways. Cluster #1 enriches the

following down-regulated pathways: mitochondrial lumen, membrane enclosed lumen, organelle lumen, and nuclear lumen, which is consistent with published results that over-expressed p53 increases lumen sizes [12]. Cluster #2 enriches a large number of immune pathways, predominantly innate immunity functions, which is consistent with a recent publication in Science [13]. Cluster #3 enriches apoptosis related down-regulated pathways, which corresponds to the widely studied functional loss by p53 mutations. Cluster #4 enriches cytoskeletal reorganization related pathways, which is consistent with the published results that p53 can regulate cytoskeletal reorganization [14]. Cluster #5 enriches multiple homeostasis related pathways, which is again supported by the previous publications [15]. Cluster #6 enriches G-proteins and associated functions, which is consistent with published studies [16]. While more detailed analyses are clearly needed to further elucidate why specific subsets of the 121 colon cancer samples have certain functional gains or losses while other samples do not, this study clearly indicates that we are on the right track to possibly derive all the functional changes caused by the selected p53 mutations.

Functional loss/gain by APC mutations in colorectal adenocarcinoma tissues.

Of the same set of colon cancer samples, 159 have APC mutations. Over ten bi-clusters are detected associated with APC mutations. Somewhat surprisingly, the largest cluster are associated with a variety of immune functions. Our literature search indeed found that APC plays a key role in innate immunity as reported by multiple papers [17]. Other clusters each consist of pathways associated with cell-cell adhesion and cell-extracellular matrix, which are both consistent with published literature [18, 19]. This analysis seems to suggest that losing the immune-related functions by APC mutations may be a key reason for cancer

development in colon, rather than the cell-cell adhesion and beta-catenin regulation function as widely speculated in the cancer literature [20].

Materials and Methods

Data analyzed in this study

20 cancer types namely namely bladder urothelial carcinoma, breast invasive carcinoma, cervical squamous cell carcinoma, endo-cervical adenocarcinoma, colon adenocarcinoma, esophageal carcinoma, glioblastoma multiforme, head and neck squamous cell carcinoma, kidney chromophobe, kidney renal clear cell carcinoma, kidney renal papillary cell carcinoma, liver hepatocellular carcinoma, lung adenocarcinoma, lung squamous cell carcinoma, ovarian serous cystadenocarcinoma, pancreatic adenocarcinoma, prostate adenocarcinoma, rectum adenocarcinoma, skin cutaneous melanoma, stomach adenocarcinoma, thyroid carcinoma, and uterine corpus endometrial carcinoma with available RNAseq and somatic mutation data are accessed from TCGA databased. Level 3 RSEM normalized RNAseq gene expression and level 2 somatic mutation data are used for the analysis. Noting the somatic mutation profile are tested by different platforms and analyzed by different procedures, we class the somatic mutation data of the selected cancer types into three quality classes, i.e., high quality class for the data contain at least one manually corrected mutation data; medium quality class for the data contain at least two automatically identified mutation data by two centers; and low quality class for the data contain only one automatically identified mutation data. To ensure the reliability of the analyzed mutations, for the data of high and medium quality class, a mutation profile is selected if and only if the detailed mutated nucleotide(s) is identified in at least two data sets.

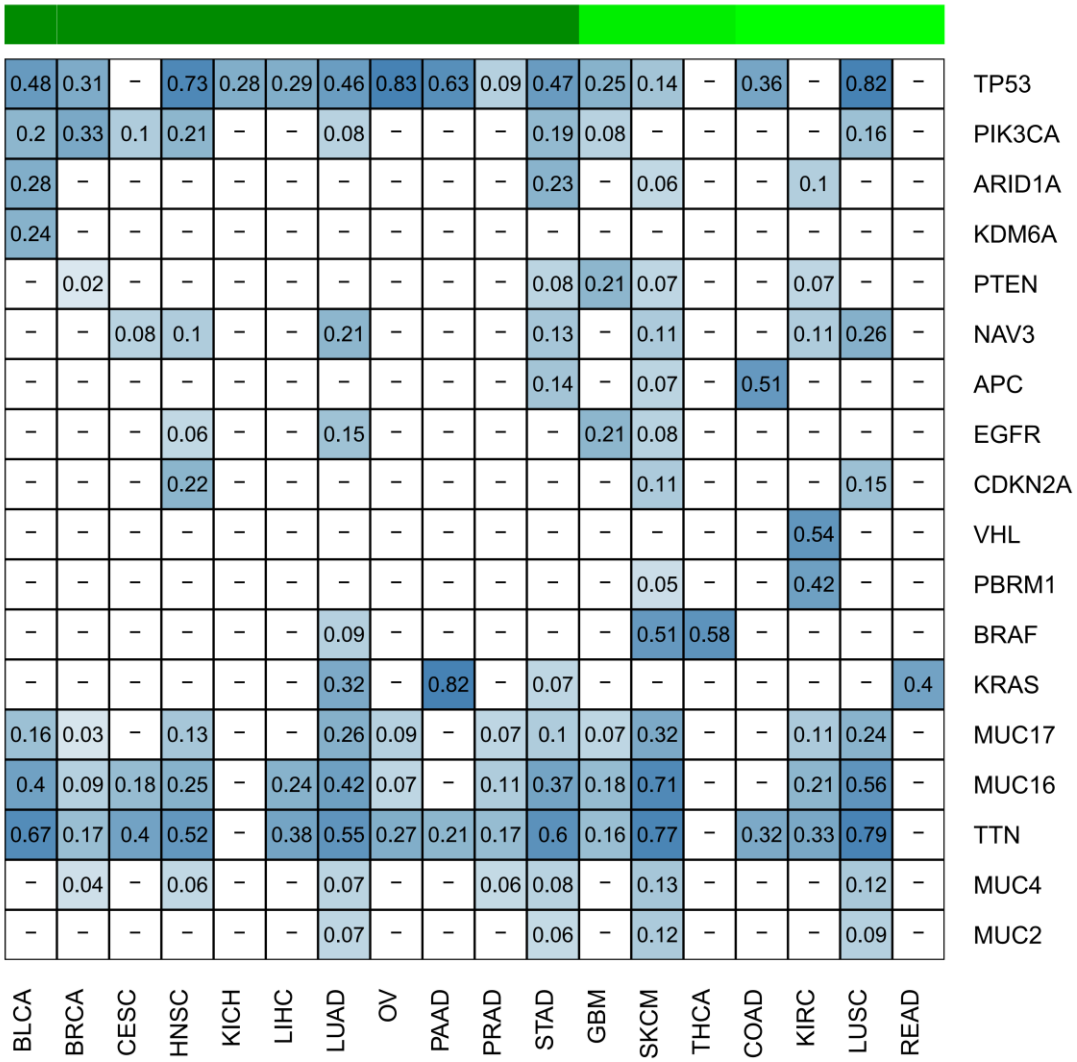


Figure 5.2. Analyzed mutations and cancer types. The column color bar reflect the quality level (dark green of high quality and light green for low quality) of the mutation profile of each cancer type as defined in Method part.

Log-mixed Gaussian model with left truncation assumption to model the expression profile of each gene measured by RSEM normalized expression level

In order to model the patterns in gene expression profile that correspond to possible gain or loss of functions of certain mutations, we specifically apply a log-mixed Gaussian model with left truncation assumption to fit the gene expression level measured by RSEM normalized expression level. Mixed Gaussian distribution has been applied to model the multiple signals in the data with normal errors. Noting the error at the zero and low expressions measured by RSEM does not fit normal distribution, we specifically introduce a left truncation assumption in fitting the RSEM gene expression profile by log-mixed Gaussian model. In detail, denotes the observed expression value of gene X over N conditions as $X = (x_1, x_2, ..., x_N)$. We have $x \in X$ follows a mixture of K normal distributions corresponding to K possible peaks and the density function of X is:

$$p(X \mid \Theta) = \prod_{j=1}^{N} p(x_j \mid \Theta) = \prod_{j=1}^{N} \sum_{i=1}^{K} a_i p_i\left(x_j \mid \theta_i\right) = \prod_{j=1}^{N} \sum_{i=1}^{K} a_i \frac{1}{\sqrt{2\pi}\sigma_i} e^{\frac{-(x_j-\mu_i)^2}{2\sigma_i^2}} = L(\Theta \mid X),$$

Parameters Θ can be estimated by EM algorithm with given X:

$$\Theta^* = \arg\max_{\Theta} L(\Theta \mid X)$$

With left truncation assumption, assuming we have M positive observed expression value and N-M left censored observations of gene X through the N conditions, a latent variable Z is introduced to estimate Θ by the following Q function and EM algorithm:

$$Q(\Theta,\Theta^{t-1}) = \sum_{i=1}^{K}\sum_{j=1}^{M}\log(a_i p_i(x_j \mid u_i,\sigma_i))p(y_j = i \mid x_j,\Theta^{t-1})$$

$$+\sum_{j=M+1}^{N}\int\sum_{i=1}^{K}\log(a_i p_i(Z_j \mid u_i,\sigma_i))\,p(y_j = i \mid Z_j,\Theta^{t-1})\,p(Z_j \mid x_j,\Theta^{t-1})dZ_j$$

$$=\sum_{i=1}^{K}\sum_{j=1}^{M}\log(a_i)\,p(y_j = i \mid x_j,\Theta^{t-1})+\sum_{i=1}^{K}\sum_{j=1}^{M}\log(p_i(x_j \mid u_i,\sigma_i))\,p(y_j = i \mid x_j,\Theta^{t-1})$$

$$+\sum_{i=1}^{K}\sum_{j=M+1}^{N}\int p(Z_j \mid x_j,\Theta^{t-1})dZ_j\,p(y_j = i \mid Z_j,\Theta^{t-1})\log(a_i)$$

$$+\sum_{i=1}^{K}\sum_{j=M+1}^{N}\int\log(p_i(x_j \mid u_i,\sigma_i))p(Z_j \mid x_j,\Theta^{t-1})dZ_j\,p(y_j = i \mid x_j,\Theta^{t-1})$$

$$=\sum_{i=1}^{K}\sum_{j=1}^{M}\log(a_i)\,p(y_j = i \mid x_j,\Theta^{t-1})+\sum_{i=1}^{K}\sum_{j=1}^{M}\log(p_i(x_j \mid u_i,\sigma_i))\,p(y_j = i \mid x_j,\Theta^{t-1})$$

$$+\sum_{i=1}^{K}\sum_{j=1}^{N}\log(a_i)\,p(y_j = i \mid Z_j,\Theta^{t-1})+\sum_{i=1}^{K}\sum_{j=M+1}^{N}\frac{1}{2\sigma_i^2}\,p(y_j = i \mid x_j,\Theta^{t-1})*$$

$$[E(Z_j^2 \mid u_i^{t-1},\sigma_i^{t-1},Z_j < Z_{cut})-2u_i E(Z_j \mid u_i^{t-1},\sigma_i^{t-1},Z_j < Z_{cut})+u_i^2]$$

The M step is then:

$$\frac{\partial Q}{\partial a_i}=0 \;\Rightarrow\; a_i^t=\frac{1}{N}\Big(\sum_{j=1}^{M}P(i \mid x_j,\Theta^{t-1})+\sum_{j=M+1}^{N}P(i \mid Z_j,Z_{cut},\Theta^{t-1})\Big)$$

$$\frac{\partial Q}{\partial u_i}=0 \Rightarrow\; u_i^t=\frac{\sum_{j=1}^{M}x_j P(i \mid x_j,\Theta^{t-1})+\sum_{j=M+1}^{N}(u_i^{t-1}-\sigma_i^{t-1}H(\frac{Z_{cut}-u_i^{t-1}}{\sigma_i}))P(i \mid Z_j,Z_{cut},\Theta^{t-1})}{\sum_{j=1}^{M}P(i \mid x_j,\Theta^{t-1})+\sum_{j=M+1}^{N}P(i \mid Z_j,Z_{cut},\Theta^{t-1})}$$

$$\frac{\partial Q}{\partial \sigma_i}=0 \Rightarrow$$

$$\sigma_i^{t^2}=\frac{\sum_{j=1}^{M}P(i \mid x_j,\Theta^{t-1})(x_j-u_i^{t-1})^2+\sigma_i^{t-1^2}\sum_{j=M+1}^{N}(1-\frac{Z_{cut}-u_i^{t-1}}{\sigma_i}*H(\frac{Z_{cut}-u_i^{t-1}}{\sigma_i}))*P(i \mid Z_j,Z_{cut},\Theta^{t-1})}{\sum_{j=1}^{M}P(i \mid x_j,\Theta^{t-1})+\sum_{j=M+1}^{N}P(i \mid Z_j,Z_{cut},\Theta^{t-1})}$$

where $P(i \mid Z_j,Z_{cut},\Theta^{t-1})=\dfrac{P(-\infty < Z_j < Z_{cut} \mid u_i^{t-1},\sigma_i^{t-1})}{\sum_{i=1}^{K}P(-\infty < Z_j < Z_{cut} \mid u_i^{t-1},\sigma_i^{t-1})}$, $H(x)=\dfrac{\phi(x)}{\Phi(x)}$, $\phi(x)$ and $\Phi(x)$ are

the pdf and cdf of standard normal distribution.

$$E(Z_j \mid u_i^{t-1}, \sigma_i^{t-1}, 0 \le Z_j < Z_{cut})$$
$$= u_i^{t-1} + \sigma_i^{t-1} E(\varepsilon_j \mid \varepsilon_j < \frac{Z_{cut} - u_i^{t-1}}{\sigma_i^{t-1}})$$
$$= u_i^{t-1} + \sigma_i^{t-1} \int_{-\infty}^{\frac{Z_{cut} - u_i^{t-1}}{\sigma_i^{t-1}}} w\phi(w)dw / \Phi(\frac{Z_{cut} - u_i^{t-1}}{\sigma_i^{t-1}})$$
$$= u_i^{t-1} + \sigma_i^{t-1} \phi(\frac{Z_{cut} - u_i^{t-1}}{\sigma_i^{t-1}}) / \Phi(\frac{Z_{cut} - u_i^{t-1}}{\sigma_i^{t-1}})$$
$$= u_i^{t-1} + \sigma_i^{t-1} H(\frac{Z_{cut} - u_i^{t-1}}{\sigma_i^{t-1}})$$

The same, we have

$$E(Z_j^2 \mid u_i^{t-1}, \sigma_i^{t-1}, 0 \le Z_j < Z_{cut}) =$$
$$u_i^{t-1^2} + \sigma_i^{t-1^2} - \sigma_i^{t-1}(Z_{cut} + u_i^{t-1}) H(\frac{Z_{cut} - u_i^{t-1}}{\sigma_i^{t-1}})$$

Identify the mutation associated gene expression

For the gene expression profile with one peak, a Kolmogorov statistics based approach is applied to identify the mutation associated over/under expression for each mutation by the following mathematical procedures:

1. For the expression profile of gene X in n samples $\{x_1, x_2, \dots, x_n\}$, order the genes according to the increasing or decreasing rank of the gene expression level to form $G_X^O = \{x_{(1)}, x_{(2)}, \dots, x_{(n)}\}$.

2. To identify the association between the over expression of X and the mutation profile of gene Y, we define the score S as $S(G_X^O, Y^{Mut}, i) = \sum_{\{j:x_{(j)} \ge x_{(i)}, Y_{(j)}^{Mut}=1\}} a - \sum_{\{j:x_{(j)} \ge x_{(i)}, Y_{(j)}^{Mut}=0\}} b$, in which Y^{Mut} is the mutation profile of Y defined by $Y_{(j)}^{Mut} = 1$ for mutation and $Y_{(j)}^{Mut} = 0$ for non − mutation of gene Y in sample (j), a and b are two positive parameters. Similarly, $S(G_X^O, Y^{Mut}, i) = \sum_{\{j:x_{(j)} \le x_{(i)}, Y_{(j)}^{Mut}=1\}} a - \sum_{\{j:x_{(j)} \le x_{(i)}, Y_{(j)}^{Mut}=0\}} b$ are defined to identify the association between the under

expression of X and the mutation of Y.

3. Association between the over/under expression of X and the mutation profile of Y is reflected by the statistics $M = \max_i\{S(G_X^O, Y^{Mut}, i)\}$. Permutation test is applied to test the significance M with p.value<0.005 as the significance level.

4. For the significant association between over expression of X and mutation Y, define $D_{XY,O} = \{d_{X,1}, d_{X,n}, \dots, d_{X,n}\}$ by $d_{X,i} = 1$ if $x_i \geq M$ and $Y_i^{Mut} = 1$ and $d_{X,i} = 0$ for other cases. Similarly, for the significant association of under expression of X and mutation Y, $D_{XY,U} = \{d_{X,1}, d_{X,n}, \dots, d_{X,n}\}$, where $d_{X,i} = 1$ if $x_i \leq M$ and $Y_i^{Mut} = 1$ and $d_{X,i} = 0$ for other cases.

For the gene expression profile $\{x_1, x_2, \dots, x_n\}$ with multiple peaks identified, each sample is discretized by assigning the index of each peak based on maximal likelihood. Association between each peak and certain mutation are tested by fisher exact test with p.value<0.005 as the significance level. Define $D_{XY,I} = \{d_{X,1}, d_{X,n}, \dots, d_{X,n}\}$ if peak I is significantly associated with the mutation of Y, where $d_{X,i} = 1$ if x_i is assigned to peak I by maximal likelihood and $Y_i^{Mut} = 1$, and $d_{X,i} = 0$ if x_i is assigned to other peaks.

Data discretization and bi-clustering analysis

For each mutation Y, combine the identified $D_{XY,U}, D_{XY,O}$, and $D_{XY,I}$ with the gene mutation profiles of other mutations by row to form a 0-1 discretized matrix D_Y. We first remove the columns corresponding to non-mutated Y and then remove the rows with too small or large number of from D_Y. The discretized matrix D_Y are then input into our in-house bi-clustering method QUBIC1.0 with parameters set as –f 0.25

-c 0.7/0.8. It is worth to note that all the 1 in each row of D_Y reflects a gene expression pattern associated with the mutation of Y, which is possible caused by gain or loss of function of certain mutation sites or integrative effect of Y and other mutations.

Pathway enrichment analysis

Pathway enrichment is assessed by using a hypergeometric test (statistical significance cutoff = 0.005) against 2801 pathways collected from the Msigdb database including the GO terms and the Msigdb canonical gene sets are used in our pathway-enrichment analysis [9].

References

1. Kim, W.Y. and W.G. Kaelin, *Role of VHL gene mutation in human cancer.* J Clin Oncol, 2004. **22**(24): p. 4991-5004.
2. van Oijen, M.G. and P.J. Slootweg, *Gain-of-function mutations in the tumor suppressor gene p53.* Clin Cancer Res, 2000. **6**(6): p. 2138-45.
3. Rios, J.J., et al., *Somatic gain-of-function mutations in PIK3CA in patients with macrodactyly.* Hum Mol Genet, 2013. **22**(3): p. 444-51.
4. Cirstea, I.C., et al., *Diverging gain-of-function mechanisms of two novel KRAS mutations associated with Noonan and cardio-facio-cutaneous syndromes.* Hum Mol Genet, 2013. **22**(2): p. 262-70.
5. Petitjean, A., et al., *TP53 mutations in human cancers: functional selection and impact on cancer*

prognosis and outcomes. Oncogene, 2007. **26**(15): p. 2157-65.

6. Mwenifumbo, J.C. and M.A. Marra, *Cancer genome-sequencing study design.* Nat Rev Genet, 2013. **14**(5): p. 321-32.
7. Mutation, C. and C. Pathway Analysis working group of the International Cancer Genome, *Pathway and network analysis of cancer genomes.* Nat Methods, 2015. **12**(7): p. 615-21.
8. Li, G., et al., *QUBIC: a qualitative biclustering algorithm for analyses of gene expression data.* Nucleic Acids Res, 2009. **37**(15): p. e101.
9. Subramanian, A., et al., *Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.* Proc Natl Acad Sci U S A, 2005. **102**(43): p. 15545-50.
10. Gene Ontology, C., *Gene Ontology Consortium: going forward.* Nucleic Acids Res, 2015. **43**(Database issue): p. D1049-56.
11. Krull, M., et al., *TRANSPATH: an information resource for storing and visualizing signaling pathways and their pathological aberrations.* Nucleic Acids Res, 2006. **34**(Database issue): p. D546-51.
12. Wan, S., et al., *Overexpression of p53 increases lumen size and blocks neointima formation in porcine interposition vein grafts.* Mol Ther, 2004. **9**(5): p. 689-98.
13. Menendez, D., M. Shatz, and M.A. Resnick, *Interactions between the tumor suppressor p53 and immune responses.* Curr Opin Oncol, 2013. **25**(1): p. 85-92.
14. Guo, F., et al., *p19Arf-p53 tumor suppressor pathway regulates cell motility by suppression of*

phosphoinositide 3-kinase and Rac1 GTPase activities. J Biol Chem, 2003. **278**(16): p. 14414-9.

15. Olovnikov, I.A., J.E. Kravchenko, and P.M. Chumakov, *Homeostatic functions of the p53 tumor suppressor: regulation of energy metabolism and antioxidant defense.* Semin Cancer Biol, 2009. **19**(1): p. 32-41.

16. Buckbinder, L., et al., *The p53 tumor suppressor targets a novel regulator of G protein signaling.* Proc Natl Acad Sci U S A, 1997. **94**(15): p. 7868-72.

17. So, A.Y., et al., *Regulation of APC development, immune response, and autoimmunity by Bach1/HO-1 pathway in mice.* Blood, 2012. **120**(12): p. 2428-37.

18. Brabletz, T., et al., *Variable beta-catenin expression in colorectal cancers indicates tumor progression driven by the tumor environment.* Proc Natl Acad Sci U S A, 2001. **98**(18): p. 10356-61.

19. Fearnhead, N.S., M.P. Britton, and W.F. Bodmer, *The ABC of APC.* Hum Mol Genet, 2001. **10**(7): p. 721-33.

20. Fodde, R., *The APC gene in colorectal cancer.* Eur J Cancer, 2002. **38**(7): p. 867-71.

CHAPTER VI

Conclusion and Discussion

The above projects show the major results of my research in the past five years, specifically focusing on modeling the micro-environment stresses and elucidating their roles in different stages of cancer. Through the analyses, I have raised one model of how the hypoxia and oxidative stress contribute to the cancer initiation; elucidated a possible cancer tissue's response to cycled hypoxia and re-oxygenation; developed a computational method to comprehensively predict the possible gain or loss of functions of certain mutations; and developed a method to identify modules in large co-expression (correlation) network and applied the method to elucidate the possible reasons that keep the high lactate production through different stages of multiple cancer types. The major results of the individual work can be found in the conclusions of each chapter. Except these major works, I have also carried some small analyses, some of which are based on my personal interests. I select some topics relating to the four major projects and discuss the details below.

Through the first project, we have raise a possible model that the long term elevation of hypoxia or oxidative stress level, which can be led by chronic inflammation or constant contact with carcinogens, can cause possibly activate the biosynthesis and exportation of the long chain glycosaminoglycan hyaluronic acid. Hyaluronic acids in the extracellular matrix can be degraded by certain enzymes or reactive oxygen species (ROS). Published results suggest the chopped hyaluronic acid pieces may serve as possible signals to aggregate the inflammation and induce multiple oncogenic signals. With these results identified in 2014, we have reviewed more literature about the role of hyaluronic acids in inflammation and cancer and

analyzed multiple data sets of various disease types including cancer, precancerous diseases and chronic inflammatory diseases. These preliminary results push us to thoroughly think about possible "oncogenic signals" we have analyzed. Recently, we have carried a comprehensively Meta-analysis on ~200 case/control data sets of 10 cancer prone chronic inflammatory diseases, 8 cancer independent inflammatory diseases, 7 precancerous neoplastic diseases and 23 cancer types. Such analyzed revealed that all the "oncogenic signals" including the pathways of angiogenesis, inflammation, ROS production, tissue repair, cell proliferation, immune response, mitochondrion, apoptosis and signaling for angiogenesis, cell proliferation, and immune response, which are defined as the "hallmarks of cancer", are highly and consistently dysregulated in both cancer and cancer prone inflammation. Such a result suggests that the cancer associated dysregulations are already happened in the cancer prone inflammations. By checking the gene co-expression networks, differences between cancer and inflammation are identified: (1) in cancer prone inflammatory diseases, functional analysis and co-expression analysis suggest that the decreased expression of mitochondrion genes are major iron-sulfur clustering genes, which are highly associated with the elevated oxidative stress. Literature review suggests Fenton Reaction driven by elevated oxidative stress in mitochondrion may possible serve as the reason of the decreased mitochondrion in inflammatory disease. But in cancer, the decreased mitochondrion is more associated with the shifted metabolism that termed Warburg Effect; (2) strong cell proliferation signals associated with tissue repair genes are identified in inflammations. The cell cycle are slightly over expressed in inflammation that are more associated with immune and stroma cell markers, suggesting the possible increased cell proliferation are driven by tissue repair process. The highly up regulated cell cycle markers in cancer is less associated with a certain pathway except the Warburg Effect; (3) other than the elevated synthesis of hyaluronic acid, the biosynthesis

enzymes of other glycosaminoglycan types including chondroitin sulfate and heperan sulfate as well as other ECM components are highly over expressed in cancer prone inflammation but not in the cancer independent diseases; and (4) oxidative stress level is highly elevated in cancer prone diseases, which is probably caused by increased macrophages and neutrophils regulated by Th1 CD4 T cells through interferon gamma signaling.

These observations suggest that the highly dysregulated micro-environment with certain cancer associated stresses such as oxidative stress and hypoxia can be commonly observed in the cancer prone inflammations. With the knowledge of the diverse functions of most cancer "driver" genes such TP53, VHL, PI3KCA in regulating the cellular response to such stresses, we fully speculate it is such stresses that consecutively drive the evolution of cancer and select and optimize the mutations of cancer. To fully understand the carcinogenesis process, with such stress types identified, we further focus on linking the mutations to such stress types aiming to comprehensively elucidate the role of most mutations in cancer initiation and progression process. The detailed analysis of these part is shown in CHAPTER V.

To examine the possible functions the cancer "non-associated" mutations, we have conducted an analysis of two sets of public genomes of precancerous and cancerous colon tissues. The first set consists of 20 precancerous and early cancer samples of colon: one polyp having 4 mutations, eight small adenoma samples harboring 272 mutations, eight large adenoma samples having 344 mutations and three adenocarcinoma samples with 198 mutations while second consists of 131 samples of colon adenocarcinomas. We have analyzed only the passenger mutations predicted by the original authors to find out if useful information can be derived from these passenger mutations, through which we have identified pathways enriched by these mutations among tissues in each disease stage, and made the following

observations all with high statistical significance: (i) mutations in small adenoma enrich functional groups related to the composition of the ECM, cell-ECM interaction, cell-cell adhesion, cell morphology and cell cycle control; (ii) mutations in large adenoma enrich the following in addition to (i): *EGF*-like domain, *ABC* transporters, cadherin, and actin binding; (iii) mutations in stage-1 adenocarcinoma enrich functional groups associated with ion transporters, plasma membrane, immunoglobulins and complement control; in addition, they also enrich functional groups associated with cell adhesion, ECM composition, cytoskeletal structure and ATP binding, which are true for mutations in stage 2-4 samples; and (iv) mutations in stages 2-4 samples enrich functional groups related to cytoskeletal reorganization; cell motion; differentiation; embryonic development; and a tyrosine kinase.

Substantial amount of new biology can be derived from these analysis results. For example, the observation that mutations enrich in genes involved in ECM composition throughout the entire development of a cancer is most interesting. Previous studies have established that during a tissue development, remodeling or repair, the ECM structure needs to become substantially more rigid compared to a non-developing tissue as the effect of growth factors can increase by 100 fold when such a change is made through altering its component protein composition. This observation, in conjunction with the earlier discussion regarding the initial cell-proliferation signals coming from hyaluronic acid fragments, strongly suggest the following: the randomly generated hyaluronic acid fragments may lead to the activation of some but not all players involved in the tissue repair machinery, at least not at the same level of coordinated activities of these players as in normal tissue repair, which puts the relevant cells in a partially activated state for tissue repair, a stressful state, waiting for the additional players to join. The ECM mutations, selected in response to the stress, may represent these awaited players, i.e., which open the doors needed

for the cells to divide without full signals for tissue repair. Overall, the analysis showed that substantially more information can be derived about the evolving cells compared to the studies focused on "driver" mutations. The algorithm introduced in CHAPTER V will be applied to 50-60 mutations in 20 cancer types to elucidate the possible roles of these mutations in each analyzed cancer.

Another possible topic is about the ECM gene mutations of stromal cells in early stage of carcinogenesis. In CHAPTER III, we have identified certain cell populations in response to cycled hypoxia and reoxygenation. Further study revealed that one cell type shows multiple stroma cell markers. Another analysis revealed that the samples with higher relative proportion of this cell type have less number of total mutations but more enriched ECM gene mutations. With the more proportion of ECM mutations identified in precancer and early stage cancer and the ECM components are majorly synthesized by stromal cells, we hypothesize that during the initiation of cancer, ECM mutations in stromal cells (like fibroblast cell) may take a role in driving the progression of pre/early stage cancer. However, unlike cancer cells, fibroblast cell can be renewed from fibrocytes from blood. Hence the mutations in fibroblast cells will easily diminish once the mutations no longer increase the tissue's fitness level. Single cell and spatial genomics/transcriptomics data can be helpful to solve this issue.