TOWARDS UNDERSTANDING THE INTERPLAY BETWEEN CELLULAR STRESSES AND CANCER DEVELOPMENT

by

SHA CAO

(Under the Direction of YING XU)

ABSTRACT

The development of neoplastic cells is hypothesized to be the result of cells responding to a stressful microenvironment such as chronic hypoxia, increased ROS and persistent immune attacks. Distinct levels of oxidative stress, estimated by gene markers in ROS-generating processes, are found to explain well the differences in disease incidences rates associated with different cancer types in different regions of the world. Further, increased levels of ROS could force the cells to induce higher antioxidant synthesis. This process could compete for sulfur resources with SAM synthesis used for DNA methylation, and eventually lead to a globally reduced level of DNA methylation. In metastatic cancer, oxidized cholesterol and its further metabolized derivatives are found to be a key driver of the explosive growth of post-metastatic cancers. My work suggests that it is the change in the O_2 level between the metastasized and the primary sites, i.e., from O_2 poor to O_2 rich, that leads to the substantially increased uptake and de novo synthesis of cholesterol as well as oxidation and further metabolism of cholesterol towards the production of oxysterol and steroidal hormones, all powerful growth signals.

To understand how various stress types may drive the unique biology of cancer, we need to study cancer tissues rather than cancer cell line data since the former contains all the relevant

information but the latter does not. Compared to the cell-based omic data, observed tissue-based gene-expression data are the results of gene-expression levels summed over all cell types, such as cancer cells, multiple immune cell types, fat cells, and normal cells in the tissues. A novel algorithm for de-convoluting tissue-based data to the cell-type specific contributions is developed based on the following information: (1) genes in each cell type are expressed in a coordinated manner, specifically they are grouped into pathways whose genes are co-expressed; and (2) different cell types tend to have different sets of pathways activated.

INDEX WORDS: microenvironment stress; deconvolution; metastatic cancer; cholesterol; oxidative stress; DNA methylation

TOWARDS UNDERSTANDING THE INTERPLAY BETWEEN CELLULAR STRESSES AND CANCER DEVELOPMENT

by

SHA CAO

BS, Beijing Normal University, 2011

MS, University of Georgia, 2015

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

2017

TOWARDS UNDERSTANDING THE INTERPLAY BETWEEN CELLULAR STRESSES AND CANCER DEVELOPMENT

by

SHA CAO

Major Professor:	Ying Xu
Committee:	Jonathan Arnold
	Ping Ma
	Shaying Zhao
	Wenxuan Zhong

Electronic Version Approved:

Suzanne Barbour
Dean of the Graduate School
The University of Georgia
August 2017

DEDICATION

This dissertation is dedicated to my parents: Huaxin Cao and Yulian Tian; my brother Min Cao; and my husband Chi Zhang.

ACKNOWLEDGEMENTS

I want to express my deepest gratitude to my advisor Professor Ying Xu, whose supervision has continually and convincingly conveyed a spirit of adventure and true scholarship, and an excitement in regards to teaching. I can't imagine how much harder my journey to PhD would have been without his contagious optimism and persistent trust in me.

I would like to thank my committee members, Professors Jonathan Arnold, Ping Ma, Shaying Zhao and Wenxuan Zhong, for guiding me through basic cancer biology, systems biology and state of the art statistical learning techniques.

I would like to thank all the members of Computational Systems Biology Lab in UGA and Jilin University, and of the Big Data Group at UGA, for their friendship and support. I would especially like to thank Dr. Victor Olman, who throughout all my six years of study, has always been ready to provide me with free mathematical consultancy.

In addition, I want to thank all of my collaborators, especially Professors Hong Qian and Bernd Schuttler, who opened the door to a whole new world of scientific research topics for me.

I would especially like to thank my family and friends for their love and support. Particularly, my husband, Dr. Chi Zhang, despite hundreds of heated arguments, never doubted his love to me. He is my best friend, my big brother, and my mentor. It is my best fortune to have him stand by me all these years.

TABLE OF CONTENTS

Page

CHAPTER 1

INTRODUCTION

Overall goal of my thesis: characterizing cellular stresses that drive cancer development

The most popular theory about cancer and cancer drivers in the past 40 years is the genomic mutation theory of cancer. The first discoveries of oncogenes by Bishop and Varmus [1] and tumor-suppressor genes by Knudson [2] have had enormous impact on how cancer has been studied in the past four decades. Various driver models have been proposed based on identified oncogenes and tumor-suppressor genes such as the Philadelphia chromosome for *chronic myelogenous leukemia* [3] and APC gene mutations for colon cancer [4]. As of now, hundreds of oncogenes and tumor-suppressor genes have been identified for a variety of cancers [5], which has fueled the idea of cancer being many different diseases and hence the need for personalized treatments. Yet, our ability to cure cancer has not improved substantially in the past four decades, cancer-related mortality rates being 200 per 100,000 people in 1975 *versus* 180 in 2009 [6], a mere 10% decrease. This seems to point to one possible and unfortunate reality: the root causes of cancer may not have been correctly identified, and the current treatments may have been mostly targeted at late developmental events after the disease has evolved substantially and divergently from the root causes.

The goal of my thesis project is to understand the issues of cancer initiation, development and metastasis from the viewpoint that the ever-changing and stressful microenvironment are the ultimate driver of cancer initiation, and its interactions with cancer cells further propel cancer cells to develop, become more malignant, metastasize and lead to human death.

Cancer takes place when normal cells are put under unusual conditions such as hypoxia [7], possibly induced by chronic inflammation [8], or elevated reactive oxygen species (ROS) [9] within localized tissue micro-environments. These conditions, if not causing cell death, would lead to abnormal cellular responses for which cell division may represent the most feasible way for cells to alleviate the pressure for survival [10]. The development of neoplastic cells is the result of cells responding to the stressful microenvironment such as chronic hypoxia, increased ROS, acidity, ECM stress and persistent inflammation. On the other hand, as the neoplastic cells continue to evolve by changing their metabolism to adapt to the challenging microenvironment, they make their environment increasingly more stressful as a side-product of their altered metabolism. Three examples are utilized to demonstrate the interactions.

Hypoxia. During chronic hypoxia with oxygen deprivation, there is a switch between the two energy metabolism methods for ATP production: from oxidative phosphorylation to glycolysis. Since glycolysis is much less efficient for ATP production, cancer cells are forced to increase uptake of glucose to make up for the less efficient energy production, however, there is an intrinsic mismatch between increased glucose influx and the capacity of the glycolysis system. As a result large amount of glucose and its derivatives are accumulated in the cells, which forms mounting pressure for the cells to evolve to remove the accumulation or die. Cell division is believed to be the most feasible way to remove the accumulated glucose derivatives, and would continue as long as the hypoxic condition and metabolite accumulation persist.

Oxidative stress. Previous studies have found that cancer cells tend to have increased ROS levels compared to normal cells, possibly due to the combination of increased metabolic activities, re-oxygenation and mitochondrial malfunction. When the ROS level exceeds the antioxidant capacities, the cells become *oxidatively stressed*. A number of biological processes

are known to respond to the ROS increase to protect the cells. Specifically, ROS can regulate directly or indirectly the activities of some important proteins such as those involved in *GPR* signaling, apoptosis, angiogenesis, immune response and general stress response. Specifically this list includes *ATM*, *ERKs*, *HSF1*, *JAK*, *JNKs*, *NFκB*, *PI3K*, *PKC* (protein kinase C), *PLCγ1* (phospholipase C-γ1) and *STAT*, indicating the global impact of ROS-induced stress, many of which will lead to changes in a range of metabolic activities, hence further altering the cellular and extracellular environments. The current knowledge is that moderately increased ROS is beneficial to the growth of cancer cells while substantially elevated ROS may drive cancers to metastasize [11].

Acidity. It is known that cancer cells tend to create an acidic pericellular environment by releasing higher-than-normal quantities of lactic acid as a result of their altered glucose metabolism, initially induced by their hypoxic environment. As a result, the altered microenvironment induces further changes, and a recent study has established that lactic acidosis in the pericellular space does not only provide a competitive advantage to cancer cells over neighboring normal cells, but also serves as a protector and facilitator for cancer cells to overcome some challenges that the hostile environment imposes on them [12], including evading apoptosis, facilitating invasion and metastasis, reducing immune cells' effectiveness and so on. The same phenomenon can be said about other micro-environmental changes, such as stresses from ECM composition and immune surveillance.

In summary, the evolving cells change the micro- and intracellular environments as side-products of their altered metabolism, which may further drive the evolving cells to become increasingly more malignant, forming a vicious cycle. In light of this, when we study cancer, we

need to take multiple factors into consideration such as (i) hypoxia, (ii) ROS, (iii) acidity, (iv) the composition of the local stroma, and (v) the composition of immune environment.

What has been done in my thesis project

My thesis studied the impact of micro-environmental stress on different stages of cancer development, including cancer initiation, primary cancer and metastatic cancer, which are all based on analyses of tissue expression data, shown in Figure 1.1. The overall theme of my thesis project is: even though different micro-environment stress could exert their effects on pre-cancrous and cancerous cells through various forms, the path human cells take to become more and more malignant is always a vicious cycle of micro-environment stress pushing cancer celles to evolve and cancer cells overcome stress by altering the micro-environment.

I. Oxidative stress, metabolic rate and primary cancer incidences

We wrote a rebuttal letter to a 2015 publication of Tomasetti and Vogelstein [13, 14], which proposes that random mutations arising during DNA replication in normal, noncancerous stem cells are key contributors to cancer, based on their observation that there is a strong and positive correlation between the total number of stem cell divisions and the lifetime cancer risk in a tissue. According to the International Agency for Research on Cancer (IARC) database, Vol-X [15], the variations of cancer risks for different human body parts is large. Moreover, the database also contains cancer risks data of same cancer types for 284 distinct regions in 69 countries spreading five continents, and it is noted that the variations within different populations for the same cancer types are also very significant. We consider that: their observation suggests that there is a baseline cancer-risk level for each tissue type, which is determined by the biology of the tissue. What is more interesting is why for the same cancer type, populations from different parts of the world should vary so much.

Our analyses revealed that (1) a combination of basal metabolic rate and oxidative stress level in a tissue offers a more plausible explanation of the lifetime risk of cancers than their model; and (2) somatic mutations may be predominantly selected to serve as facilitators rather than primary drivers of cancer formation.

II. Oxidative stress, nucleotide synthesis and DNA hypo-methylation

While gene-expression markers can be used to detect stresses, they tend to be too sensitive to other cellular conditions, hence hard to use as reliable predictors for specific stresses, particularly stress levels. Epigenomic markers, such as DNA methylation levels associated with specific genes, are substantially more stable, and it has been well established that (micro)environmental stresses can leave marks in epigenomic patterns. In fact, it has been observed that while the promoter regions of protein-encoding genes in cancer genomes tend to have increased methylation levels, the overall level of DNA methylation in cancer tends to have decreased, in a variety of cancer types, including colon, liver, gastric, ovarian, breast, thyroid, and lung cancer [16-19]. Certain cancers can have over 50% reduction *vs.* their normal controls as observed, including human primary GBMs and glioma cell lines [20]. It has been speculated that such reduced levels of global DNA methylation might have been selected to increase opportunities for the host cancer cells to overcome or to adapt to specific stresses encountered as reduced methylations generally imply increased gene expressions [21].

Currently, some hypothesis try to explain the reason of global DNA hypomethylation in cancer. With dietary methyl (folate, choline, and methionine) deficiency intakes, there are not enough preformed methyl groups to meet the total need for DNA methylation and DNA synthesis, which causes hepatic steatosis, cirrhosis, even hepatic tumorigenesis [22]. Additional methyl groups are synthesized de novo via the one-carbon folate pool. Among folate pool, 5,10-

methylenetetrahydrofolate is preferentially directed toward de novo thymidylate biosynthesis at the expense of homocysteine remethylation during folate deficiency [23]. So, DNA methylation would be sacrificed.

We have identified (1) a possible determinant for global DNA methylation in cancer cells, and (2) a relationship between levels of DNA methylation, nucleotide synthesis and intracellular oxidative stress in cells. We developed a system of kinetic equations to capture the metabolic relations among DNA methylation, nucleotide synthesis, and anti-oxidative stress response, including their competitions for methyl and sulfur groups, based on known information about one-carbon metabolism and trans-sulfuration pathways. We observed a kinetic-based regulatory mechanism that controls reaction rates of the three competing processes when their shared resources are limited, particularly when the nucleotide synthesis rates or oxidative states are high. The combination of this regulatory mechanism and the need for rapid nucleotide synthesis, as well as high production of glutathione dictated by cancer-driving forces, led to the nearly universal observations of reduced global DNA methylation in cancer. Our model provides a natural explanation for differential global DNA methylation levels across cancer types and support the observation that more malignant cancers tend to exhibit reduced DNA methylation levels. Insights obtained from this work provide useful information about the complexities of cancer due to interplays among competing, dynamic biological processes.

III. Oxygen stress, metastatic cancer and cholesterol metabolism

Metastases exhibit progression patterns differently than their primary counterparts, growing substantially faster after dormancy[24, 25]. They respond poorly to existing drugs and are responsible for +90% of all cancer-associated mortalities [26]. While metastasis represents the deadliest stage, little is understood about its unique biology[27, 28]. To date, what drives the

explosive growth of a metastatic cancer and why metastases do not respond to drugs in manners similar to their primary counterparts is unknown. For example, metastatic cancer patients whose primary tumors show complete response to neoadjuvant chemotherapy still have poor outcomes [29]. Metastatic cancer studies focus predominantly on *how* primary cancer cells leave their disease sites, intravasate blood vessels, and extravasate blood vessel to colonize a new location [27, 28]. While recent genome analyses indicate parallel evolution of primary and metastatic tumors [30], mechanistic insights into how these evolutionary processes impart metastasis-site specific growth have yet to be gained. As with primary tumor cells, metastatic cancer cells must adapt to the environment of the organ to which they spread. For example, primary cancer sites are generally hypoxic [31] whereas metastatic sites tend to be blood- and O_2-rich. In addition, immune cells at the primary sites have co-evolved with the cancer from its onset, and are involved throughout a cancer's development [32], referred to as *tumor-associated* immune cells whereas metastatic cells are new to the local immune cells, which act more aggressively towards them.

Cholesterol is the major ingredient of cell membrane biosynthesis. De Novo synthesis via the mevalonate pathway and uptake via various lipoproteins are the two routes for cells to obtain cholesterol. Links between cholesterol and cancer have previously been reported in the literature, such as: (i) epidemiology studies that found connections between blood cholesterol levels and cancer mortality rates [33]; (ii) studies that observed increased cellular cholesterol levels in a few (primary) cancer types, such as breast cancer [34] and prostate cancer [35]; and (iii) studies that link dysregulation or mutations of cholesterol-metabolism genes to cancer occurrence [36]. A few recent cancer-epidemiology studies have detected correlations between long-term usage of cholesterol-lowering drugs such as statins and reduced cancer-associated mortalities [37].

Mechanistic studies on this relationship only start to emerge in the past few years. For example, function-losing mutations in *ABCA1*, the main exporter for cholesterol efflux, are found to be associated with increased cancer occurrences, specifically in colon [38]. One study suggests that the increased membrane-cholesterol level is associated with the activation of the kinase *Akt*, a regulator of apoptosis, and hence increases the chances of cancer cells survivals [39]. While published studies have detected links between cholesterol and cancer progression, no model or understanding has been reported regarding how cholesterol contributes to the explosive growth of metastatic cancers (vs the corresponding primary cancers) except that cholesterol is used to make cell membranes, to the best of our knowledge.

Evolutionary studies strongly suggest that cholesterol (or sterols in general) has co-emerged with O_2 during the early evolution around 2.5 - 3 billion years ago as a "seal" between phospholipids in cell membranes to prevent the toxic O_2 from entering into anaerobic cells [40]. Recent studies have revealed that (a) membrane cholesterol serves as an O_2 sensor and a possible regulator of O_2-entry into the cells by serving as a membrane barrier against O_2 and reactive oxygen species (ROS) [40]; (b) a higher membrane cholesterol-phospholipid ratio gives rise to lower O_2 permeability of cellular membranes [41]; and (c) the plasma membrane-cholesterol levels are found negatively correlated with the amount of changes in cellular O_2 levels of red blood cells when the blood- O_2 level changes [42].

We have identified that the altered O_2 level in primary cancer sites *vs.* metastatic cancer sites represents a key stress that the newly migrated cancer cells must overcome. Specifically, primary cancers tend to be in a hypoxic environment [31], hence their cells have adapted to such environment after years of residence there. Once these cells must escape such an environment and enter new blood-rich, hence O_2 rich environment, substantial changes must be made for adaptation.

My thesis research has identified and characterized oxidized and further metabolized cholesterol derivatives as a key driver of the explosive growth of post-metastatic cancers. This represents the first study on elucidation of drivers of metastatic cancers. My study strongly suggests that it is the change in the O_2 level between the metastasized and the primary sites, i.e., from O_2 poor to O_2 rich, that leads to the substantially increased uptake and *de novo* synthesis of cholesterol as well as oxidation and further metabolism of cholesterol towards the production of oxysterol and steroidal hormones, all powerful growth signals.

IV. De-convolution of tissue-based gene-expression data

To understand what stresses may drive the unique biology of metastatic cancers, we need to study cancer tissue rather than cancer cell line data since the former contains all the relevant information but the latter does not. Compared to the cell-based omic data, tissue-based omic data are substantially information richer but analyses of such data also raised very challenging computational issues, namely, observed gene-expression data are the results of gene-expression levels summed over all cell types, such as cancer cells, multiple immune cell types, fat cells, and normal cells in the tissues. Given is a matrix *X* of measured expression data (or profiles) (row for genes and column for samples) of *M* (human) genes from *K* cell types over *N* samples. A *deconvolution problem* is to estimate S and P so that the total error $\|\epsilon\|$ is minimized with $\mathrm{X_{M\times N}} = \mathrm{S_{M\times K}} \cdot \mathrm{P_{K\times N}} + \epsilon, s.t.: \sum_{t=1}^{N} P_{it} = 1, P_{it} \geq 0;\ i = 1, \ldots, K,$ and S being a *signature* matrix of gene expression data of each cell type and *P* the proportion of the cell type in the mixture.

A few algorithms have been developed to estimate *S* and *P* [43-47]. Some assume that either *S* or *P* is known *a priori*[43, 48-57]. In addition, it is generally assumed that *S*: 1) consists of only cell-type specific genes[43, 52-57] or 2) has a constant expression level or simple variations for each component gene across different tissue samples[54, 58, 59]. Furthermore, it is often assumed that

the to-be-de-convoluted tissue data consists of cancer and one additional cell type[58, 60-62]. These are too restrictive for modeling or solving an actual tissue-based data deconvolution problem consisting of a multiple cell types that cover most of the human genes. We have checked all published algorithms; and found that none are capable to derive for > 2 cell types the detailed gene expression profiles for each cell type for individual tissues; and they all miss one key element: co-expressions among genes in the same pathway are not guaranteed.

I have developed a novel algorithm for de-convoluting tissue-based data to the cell-type specific contributions based on the following information: (1) genes in each cell type are expressed in coordinated manners, specifically they are grouped into pathways whose genes are co-expressed; and (2) different cell types tend to have different sets of pathways activated. From these insights, my algorithm is pathway-centric rather than gene-centric as in virtually all de-convolution algorithms. The challenge in this problem formulation lies in that our current knowledge about what pathways are expressed in which cell type is very limited. I have employed a powerful bi-clustering technique developed in our lab to discover such pathways through systematic analyses of bi-clustering results of gene-expression data of each cell type using clean cell-line based data. This involves enormous amount of very challenging analysis work.

The structure of my thesis

In Chapter 2, we examined the levels of oxidative stress in different organs of the different populations around the world based on our oxidative stress predictor, and discovered the variance of cancer incidences associated with different cancer types and within different population groups could be well explained by the basal metabolic rate intrinsic to different organs and the basal level of oxidative stress associated with different populations in the world; in Chapter 3, we built a mathematical model for DNA methylation associated pathways, the one-

carbon and trans-sulfurration pathways, and found that hypomethylation in cancer cells is resulted from competitions of methyl/sulfur sources between DNA methylation reaction and the nucleotide synthesis/anti-oxidation pathways; in Chapter 4, we conducted a comprehensive study on metastatic cancers in comparison with primary cancers, and discovered highly consistent up-regulation of cholesterol metabolism in metastatic cancers, which is a defense mechanism of for the oxygen-rich environment in the metastatic site, and it turned out cholesterol metabolism could triger metastatic cancer proliferation; in Chapter 5, we developed a novel computational tool for tumor tissue sample deconvolution, aiming to decompose the expression profile measured on a tissue into the sum of expression profiles of its component cells, without losing the sample specificity of these component cells; Chapter 6 is a concluding chapter.

Figures:

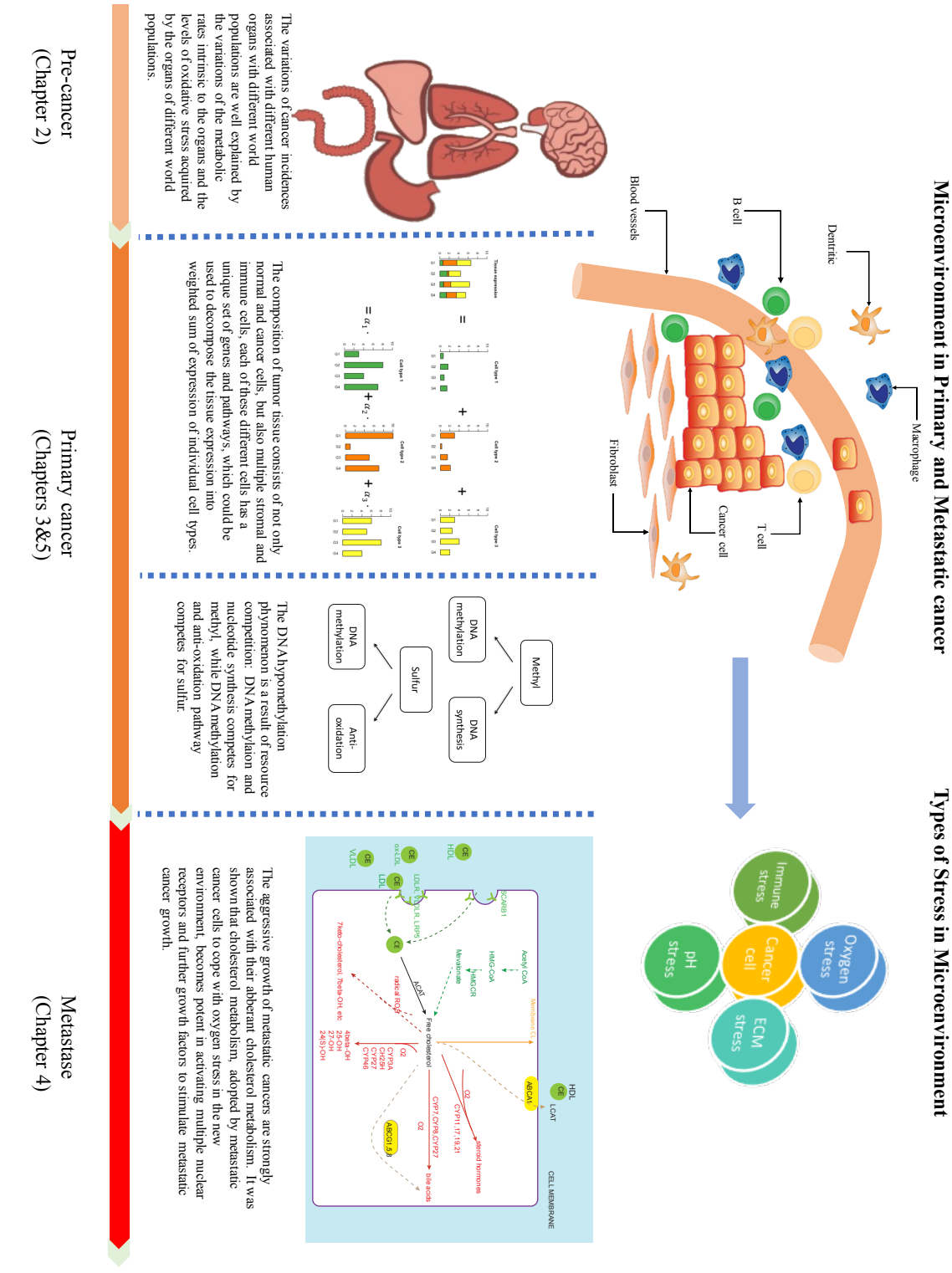


Figure 1.1: An outline of what was done in my thesis project.

CHAPTER 2

SOMATIC MUTATIONS MAY NOT BE THE PRIMARY DRIVERS OF CANCER FORMATION

Cao, S., Zhang, C. and Xu, Y. (2015), Somatic mutations may not be the primary drivers of cancer formation. Int. J. Cancer, 137: 2762–2765. doi:10.1002/ijc.29639. Reprinted here with permission from the publisher.

Abstract

Tomasetti and Vogelstein recently published two articles in Science proposing that random mutations arising during DNA replication in normal, noncancerous stem cells are key contributors to cancer, based on their observation that there is a strong and positive correlation between the total number of stem cell divisions and the lifetime cancer risk in a tissue. Our recent analyses of their and additional data revealed that there is a fundamental disconnection between their observation and their conclusion. In addition, our data suggest that (1) a combination of basal metabolic rate and oxidative stress level in a tissue offers a more plausible explanation of the lifetime risk of cancers than their model; and (2) somatic mutations may be predominantly selected to serve as facilitators rather than primary drivers of cancer formation.

Results

Tomasetti and Vogelstein recently published two articles in Science [13, 14], proposing that "The majority [of cancer] is due to ... random mutations arising during DNA replication in normal, noncancerous stem cells", "[which are] responsible for either initiating the process of tumorigenesis or for driving tumor progression". They then postulate that "This [their proposal] is important not only for understanding the disease but also for designing strategies to limit the mortality it causes." These statements were made based on their observation that there is a strong and positive correlation between the total number of stem cell divisions and the lifetime cancer risk in a tissue. We consider that: (1) there is a fundamental disconnection between their observation and their conclusion; hence their above statements are not supported by their data; and (2) their reports may offer a good opportunity for the research community to have a serious debate regarding the true roles played by somatic mutations in the formation of sporadic cancers, as the current knowledge about cancer versus mutation is predominantly derived based on cell

rather than cancer tissue studies [63]; and knowledge gained from tissue repair studies suggest that cell proliferation in a tissue needs more than just mitogenic/growth signals [64].

There is an alternative and more plausible explanation of their main observation. Tomasetti and Vogelstein made an interesting observation regarding a correlation between the number of cell divisions and the lifetime cancer risk in a tissue. This suggests that there is a baseline cancer-risk level for each tissue type, which is determined by the biology of the tissue. However it represents too much of a leap to suggest, without any supporting data, that it is the replicative mutations associated with cell divisions that are responsible for the lifetime cancer risk of a tissue. The only (indirect) support for their claim is the somatic mutation theory of cancer [65, 66]. While it has been a belief held by many that cancer arises solely as a result of genomic mutations, to the best of our knowledge there have not been any published data demonstrating that a sporadic human cancer can be initiated by somatic mutations alone. Actually, there is an alternative and more plausible explanation of their observation.

We have examined the basal metabolic rate versus the lifetime cancer risk of a tissue type, and found that they have much stronger correlation, as shown in Figure 2.1, than the correlation observed by Tomasetti and Vogelstein. Furthermore, the level of oxidative stress, estimated based on expression levels of some oxidative-stress response genes in datasets unrelated to the current study, explains very well the variations in lifetime cancer risk across different countries. Our model, based on tissue metabolic rate and oxidative stress level, achieves $R^2 = 0.96$ over 13 leading cancer types in 21 countries (see Supplementary Material), in comparison with their model having $R^2 = 0.65$ on their data.

In the gerontological literature, it has long-been proffered that the onset and/or rate of aging have metabolic rate and/or oxidative stress as two key contributors [67, 68], hence (as aging

encompasses a large proportion of the instances of cell death in the life of mature humans), we hypothesize that metabolic rate and/or oxidative stress are key contributors to the total number of cell deaths in a tissue, which is the same as the number of cell divisions for maintaining tissue homeostasis in Tomasetti and Vogelstein's study. We posit that it is the combination of the basal metabolic rate and oxidative stress level of a tissue that determine both the rate of aging (hence the total number of cell divisions) and the lifetime cancer risk. From this perspective, their observation is not a causal but rather a statistical correlation between two events with common causes.

Contributions from environmental factors are hidden and averaged out in their analysis. It is far from being clear from the information provided in [13] why their ERS (extra risk score) can be used to separate replicative mutations from environment-induced mutations (or damages). Our data in Figure 2.1 suggest that the average level of oxidative stress, which affects DNA mutation rates, plays an essential role in the level of the baseline cancer risk for each country considered. Yet, the average level of oxidative stress of an organ varies across different countries (see Supplementary Material), strongly suggesting that environment plays a role in the basal level of oxidative stress and hence the baseline cancer risk of a tissue. It is also noteworthy that the correlation shown in [13] is between the average cancer risk and the number of cell divisions, which has largely removed contributions from environmental factors. Figure 2.2 here shows that the lifetime risk of each cancer type has a large variation across 284 regions of 69 countries in the world, which we can reasonably assume is largely due to environmental factors [69], knowing that different higher eukaryotes, hence different human ethnic groups, all have highly similar mutation rates arising from DNA replications [70]. We have examined cancer risks in three countries and one territory (Brazil, Costa Rica, Cuba, and Puerto Rico) having quite

different risk levels across multiple cancer types. It was found that the correlation between the lifetime cancer risk and the number of cell divisions in each tissue (vs cancer) type as used in [13] is considerably lower than that reported in [13] in each of the four locations, at 0.481, 0.448, 0.416 and 0.496, respectively. Although we do not know to what degree the population of any of these three countries or one territory share genetic background with the population of the U.S.A., we only argue that the high correlation between stem cell division rate and lifetime cancer incidence rate published for the U.S.A. population in [13] doesn't hold in any of these four locations. These data suggest that variations in cancer risks, which can be as large as 100 times the means for some cancer types such as melanoma or ALL (see Figure 2.2) in some countries, could not be explained by the numbers of the basal cell divisions; instead the basal level of oxidative stress, with contributions from the environment, offers a much stronger explanation.

Driver or facilitator: roles played by somatic mutations in cancer formation: Our disagreement with [13, 14] is rooted at a rather fundamental level, i.e., regarding the fundamental roles played by somatic mutations, due to both replication infidelities and induced damages possibly coupled with errors introduced through DNA repair, in cancer formation. While mutations driving cancer has been a popular view in the past few decades, it has not been convincingly established in cancer tissues. Published studies regarding oncogenic mutations versus cell proliferation have been predominantly conducted using cultured cells or mouse models that do not capture the fundamental micro-environmental stresses, including oxidative stresses that can lead to DNA damages [71]. While such models are highly useful for studying mechanistic issues, they may not necessarily be the best or even a correct way to study root causes of a cancer since oncogenic mutations alone cannot activate cell proliferation in a tissue environment as discussed below. Regarding the actual roles played by somatic mutations in evolution, a recent study on bacterial

co-evolution with their antagonistic phage may shed useful lights here [72]. The study convincingly established that it is environmental stresses that lead to increased mutations and hence increased opportunities for the host to gain a competitive advantage and/or survival through selection of certain mutations. Here the ultimate driving force of cellular evolution comes from the need for competitive fitness or survival via stress adaptation through mutation selection. That is: mutation selection serves as a vehicle for better fitness and/or survival. This is in line with Darwin's Theory of Evolution.

Our recent analyses of omic data collected on precancerous and early cancerous colon tissues suggest that cancer evolution follows the same evolutionary principle. Specifically we noted that a substantial fraction of mutations is in genes associated with extracellular matrix (ECM) in precancerous tissues [73] (see Supplementary Material). In parallel, we have also observed that the precancerous tissues tend to have active synthesis of hyaluronic acids (HAs), as well as their extracellular release and fragmentation, possibly triggered by hypoxic and inflammatory conditions coupled with plentiful G6P induced by increased glycolytic activities and metabolic congestions along the glycolytic pathway under a hypoxic condition [74]. It has been well established that short HAs are key signals for tissue repair [75], including signals for cell cycle control, cell growth and survival, and angiogenesis, which are normally released from a damaged ECM but here are produced due to a persistent abnormal condition.

By integrating these two types of data, we proposed a model for explaining the observed mutations in ECM genes in (pre)cancerous tissues [73]. That is: under persistent hypoxic and inflammatory conditions, the underlying epithelial cells are accumulated with glucose metabolites, leading to congestion of their glycolytic pathway, a stressful state. These conditions trigger the production, release and fragmentation of HAs, and hence the activation of some but

not all players involved in the tissue repair machinery. This puts the relevant cells in a partially activated state for tissue repair, another stressful condition, waiting for the additional players to join. The ECM mutations, possibly selected in response to this stress, may represent these awaited players. It has been established that tissue repair, like in tissue development and remodeling, requires changing the ECM structure from being flexible, needed for their normal functions, to very rigid through altering its protein composition as the effect of growth factors can increase by 100-fold when such a change is made [76]. Together, the combination of the partially activated tissue repair system and ECM mutations gives rise to the activation of the HA fragments-induced tissue repair system, enabling cell proliferation, driven by adaptation to the aforementioned stresses.

It is worth noting that oncogenic mutations alone cannot activate cell proliferation in a tissue environment since that requires signals more than growth factors (or equivalently oncogenic mutations), such as signals to ECM [64]. We posit that this HA-enabled tissue repair, which enables cell proliferation, will continue as long as the hypoxic and inflammatory condition does not change, possibly in an intermittent manner due to the randomness in mutation occurrences and HA fragmentation patterns. This process may become more efficient once some oncogenic mutations take place and are selected. It is noteworthy that numerous proto-oncogenes are regulated by HAs, including MYC and HSF [74, 77]. This is further supported by our recent analyses of cancer tissue genomic and transcriptomic data focused on the detailed functional gains and losses by key cancer-associated genes, including TP53 and MYC, which revealed that before the mutations are selected, the relevant functional gains or losses by them tend to have already been accomplished through functional inhibition and/or enhancement (manuscript in preparation). Hence we posit that it is essential to have mutations in ECM related genes for a

cancer to develop in a non-growing tissue. A main point learned here is that somatic mutations tend to serve as facilitators rather than primary drivers of the evolution of sporadic cancers.

A key message of this letter is that there could be an alternative and more plausible explanation of Tomasetti and Vogelstein's observation, as well as of the general roles played by somatic mutations in cancer formation.

Materials and Methods

A. Data used in our study

All the cancer epidemiological data are obtained from the IARC database, Vol-X [15], which covers 284 distinct regions in 69 countries spreading five continents. The lifetime cancer risk is estimated for a person with a lifespan of 80 years based on the methods given in Chapter 7 of [78]. For each cancer type, the median is used if data from multiple regions are available for a country. To combine data from both male and female, an average is taken assuming that the two populations have the same sizes, except for testicular germ cell cancer and ovarian germ cell which usually occur only in male or female. The cancer types marked with "*" in Table S2.2 do not have their lifetime cancer risk readily available from IACR, and are thus estimated based on the closest and available cancer type that covers them and their proportions in the superclass cancer type as reported in [13].

To estimate the percentage of all cancer incidences by the 31 types used in the original paper, we have used the total number of cancer incidents in the USA in 2003-2007 along the ratio of each cancer type (or its superclass cancer type when the data are not available for a specific cancer type such as Duodenum adenocarcinoma) out of all the cancer incidences based on data given in IARC, Vol-X [15] . Overall the 31 cancer types cover not more than 45% of all cancer incidents. As a comparison, our cancer-occurrence variation analysis covers 21 cancer types in the IARC database

(Vol-X) while our causal analysis on cancers covers 13 leading cancer types with gene-expression data available in at least two countries, namely bladder carcinoma, breast carcinoma, colorectal carcinoma, esophagus carcinoma, renal cell carcinoma, hepatocellular carcinoma, lung adenocarcinoma, melanoma, pancreatic adenocarcinoma, prostate adenocarcinoma, osteosarcoma and thyroid carcinoma, which cover 72% of all cancer incidences.

B. Correlations between the number of cell divisions and cancer risk

We have examined 21 cancer types, all the cancer types studied in [13] with publicly available cancer risk data for all five countries: Brazil, Costa Rica, Cuba, Puerto Rico and the USA in the IARC database [15], namely acute myeloid leukemia, chronic lymphoid leukemia, colorectal adenocarcinoma, duodenum adenocarcinoma, esophageal squamous cell carcinoma, gallbladder non-papillary adenocarcinoma, glioblastoma, head and neck squamous cell carcinoma, hepatocellular carcinoma, lung adenocarcinoma (smokers and non-smokers), medulloblastoma, medullary thyroid carcinoma, melanoma, osteosarcoma, ovarian germ cell cancer, pancreatic ductal adenocarcinoma, pancreatic endocrine carcinoma, small intestinal adenocarcinoma, testicular germ cell cancer and thyroid papillary follicular carcinoma. Table S2.2 shows the risks for the 21 cancer types in four countries. The correlation coefficient between the risks across the 21 cancer types and the total number of stem cell divisions in each of the four countries (Brazil, Costa Rica, Cuba, Puerto Rico) are 0.481, 0.448, 0.416 and 0.496, respectively.

C. Gene expression data analyzed

A total of 57 sets of microarray data from the GEO database [79] are used to predict the basal glucose metabolic rates and the basal oxidative stress in a tissue type across 13 (normal) tissue types . That is 9 cancer types shared with Tomasetti and Vogelstein's analysis, plus bladder, breast, kidney and prostate. Overall, these data consist of 884 normal tissue samples as detailed in Table

S2.3. All the data were measured using Affymetrix Human Genome U133 Plus 2.0 Array and processed by RMA normalization method using the "affy" package in R with default parameter setting. The RMA processed gene expression is then normalized by the total gene expression level of each sample.

D. Estimation of oxidative stress level in a tissue

We have developed a computational method to predict the oxidative stress level for a specified tissue based on the gene expression data measured using Affymetrix Human Genome U133 Plus 2.0 Array in case-controlled experiments. Specifically, we have collected 31 gene-expression datasets consisting of 294 samples with known oxidative stress levels and 219 control samples as the training data to train a predictor using a logistic model, for the level of oxidative stress in a tissue. Each dataset is processed and normalized by the same procedure as described in the previous section. The detailed information of the training data is given in Table S2.4. 136 oxidative-stress response genes (covered by 268 probes) are selected as the discerning features to develop the predictor, which consist of genes annotated to be "response to oxidative stress" by GO [80] and "reductases" from literature [81].

The predictor is trained using logistic regression with variable selection using L1 regularization on the probes. R package "glmnet" is applied to train the predictor and the parameters is selected by highest prediction accuracy under 10-fold cross validation. At the end, 69 probes of 52 genes are selected and used in the final model, which achieves 88.96% prediction accuracy under 10-fold cross validation. Detailed parameters of the model are given in the Table S2.5.

This predictor is used to estimate the oxidative stress level of the normal tissue samples in the current study, which has not overlap with the training data.

E. Correlations between oxidative stress levels and cancer risks for 13 leading cancer types

Figure S2.1 (a – l) show correlations between the estimated oxidative stress levels and the lifetime risks of 13 leading cancers in the world with publicly available gene-expression data, namely bladder carcinoma, breast carcinoma, colorectal carcinoma, esophagus carcinoma, renal cell carcinoma, hepatocellular carcinoma, lung adenocarcinoma, melanoma, pancreatic adenocarcinoma, prostate adenocarcinoma, osteosarcoma and thyroid carcinoma. The figure shows high correlations between the lifetime e risks and predicted oxidative stress levels across different countries for each cancer type.

F. An integrated model to explain cancer risk

We have developed a model that combines the (normalized) gene-expression level of *PFKL* and the predicted oxidative stress level to explain the observed cancer risks as follows:

$$\mathrm{CR_{ij}} = \alpha * \mathrm{PFK_i} + \beta_j * \mathrm{PO_{ij}} + \gamma + \varepsilon_{ij}\,, i = cancer\ type\ index, j = country\ index$$

where CR, PFK and PO are the cumulative risk, gene expression level of PFKL and predicted oxidative stress, α, β_j and γ are regression parameters, i and j are cancer type and country, and ε_{ij} is the error term for the regression model following iid Gaussian, respectively. The three parameters are trained using a regression analysis based on the 13 cancer types: bladder, breast, colon, esophagus, head and neck, kidney, liver, lung, pancreatic, prostate, and thyroid cancer and melanoma and osteosarcoma, which represent the largest set with available epidemiology and gene expression data. Our regression model achieves $R^2 = 0.96$ in fitting the cumulative risk of the 13 cancer types that is considerably higher than $R^2 = 0.65$ in Tomasetti and Vogelstain's regression model, with detailed parameters α, β_j and γ given in Table S2.6.

G. Functional enrichment analysis of the somatic mutations in precancerous colon and colorectal cancer

We have conducted a pathway enrichment analysis by somatic mutations in two datasets in the public domain. The first data set is a sequenced exome dataset of 24 human (pre)cancerous colon samples [82], consisting of 1 polyp with 4 mutations, 8 mild and small adenoma samples harboring 272 mutations, 8 severe and large adenoma samples having 344 mutations and 3 adenocarcinoma samples with 198 mutations; and the second dataset covering 131 colon cancer samples (none hyper-mutated) in TCGA [83], consisting of 18 stage-1 samples with a total of 1,439 mutations, 47 stage-2 samples with 3,683 mutations, 43 stage-3 samples harboring 3,657 mutations and 23 stage-4 samples having 2,061 mutations. We carried out the following pathway/gene set enrichment analysis. We count each mutation once if it is observed in at least one sample in each disease stage; and then conduct a pathway enrichment analysis by the hypergeometric test in DAVID functional annotation tool [84] against KEGG [85], BIOCART [86] and REACTOME [87] databases. A pathway/gene set is considered as enriched by mutations if the p-value for enrichment is < 0.05. We have noted that the gene-sets/pathways associated with cell adhesion and extracellular matrix are significantly enriched throughout all stages of tissues examined here. A detailed list of enriched pathways is listed in Table S2.7.

Figures

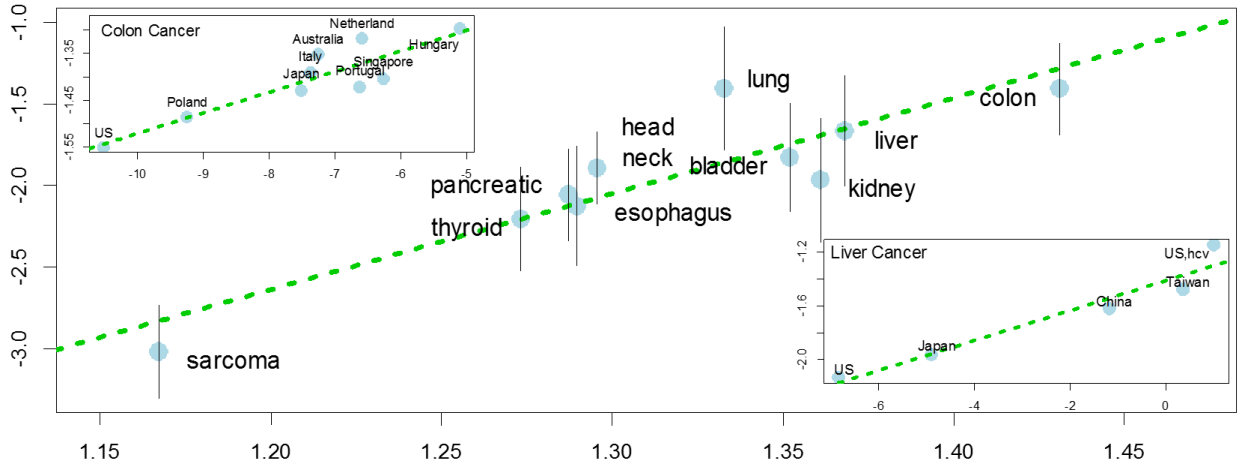


Figure 2.1: Correlation between lifetime cancer risks and basal metabolic rates, along with correlations between variations in risk of each cancer type across different countries and the estimated oxidative stress level in a tissue. All 13 high prevalence cancers for which gene expression data for Pfkl and the oxidative stress-responsive genes that we utilized in our analysis were publicly available are examined here, and they cover 72% of all cancer incidences, versus 45% as covered in [13]. The x-axis for the large box is the gene-expression level of PFKL, the rate-limiting enzyme of glycolysis pathway, used to approximate the glucose metabolic rate, and the y-axis is the average lifetime risk (log10 scale) of cancer for a tissue as in [13] while the x-axis in each inside box is the estimated oxidative stress level and the y-axis is the same as in [1]. All data used throughout this letter are provided in the Supplementary Material.

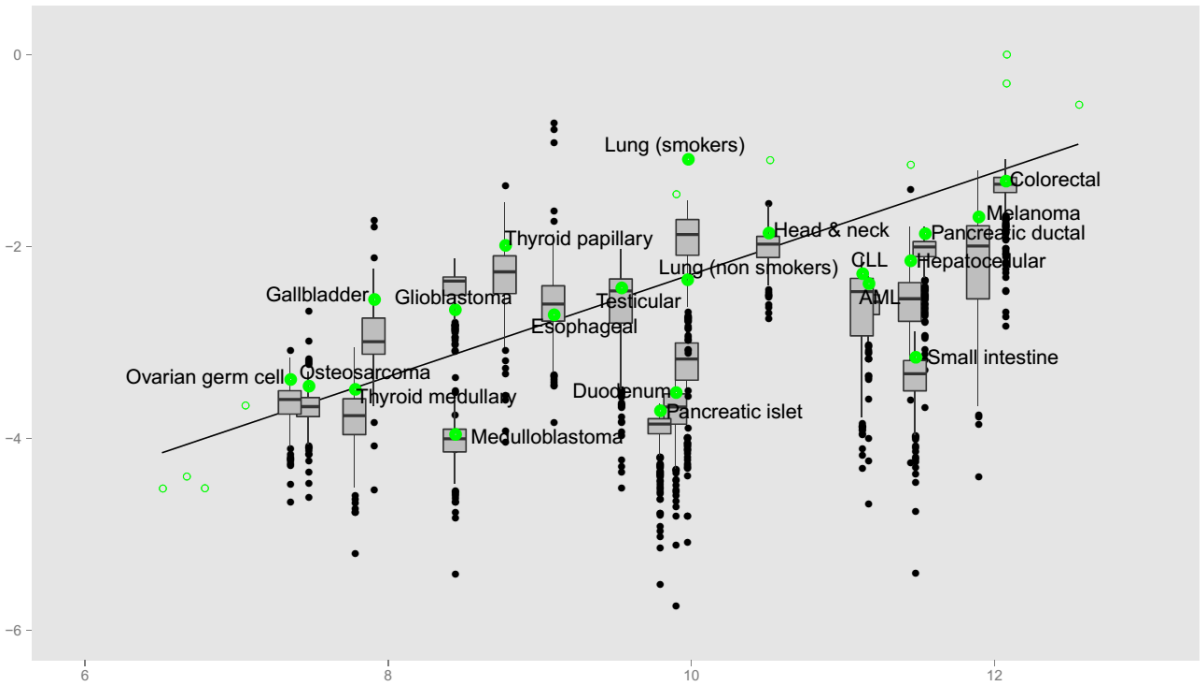
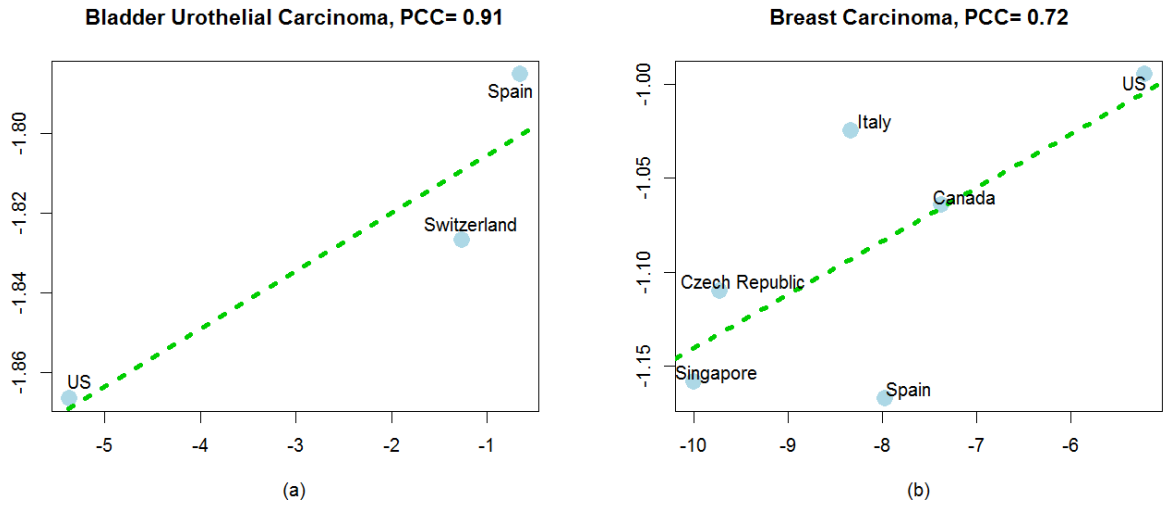
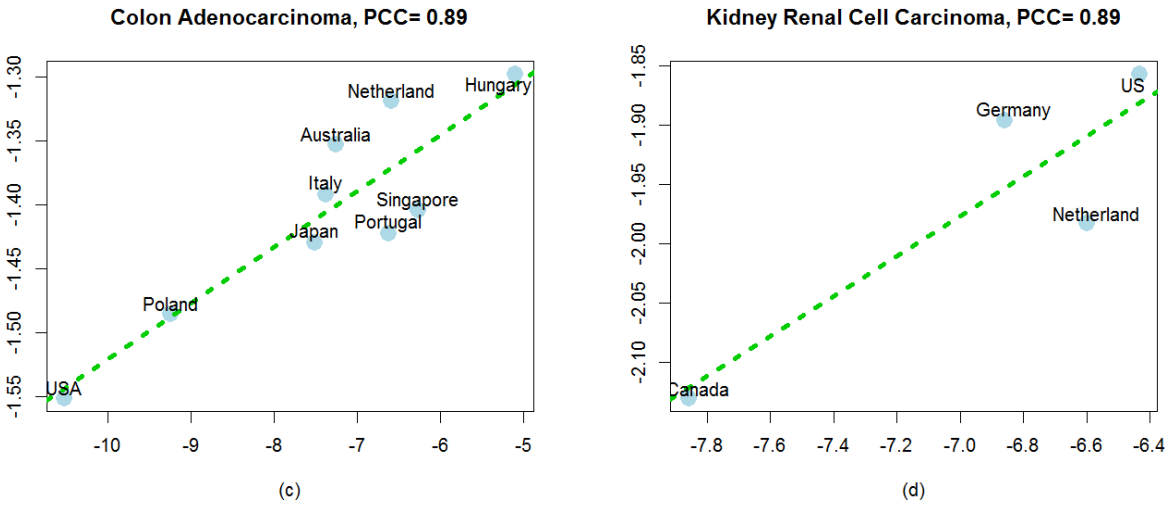
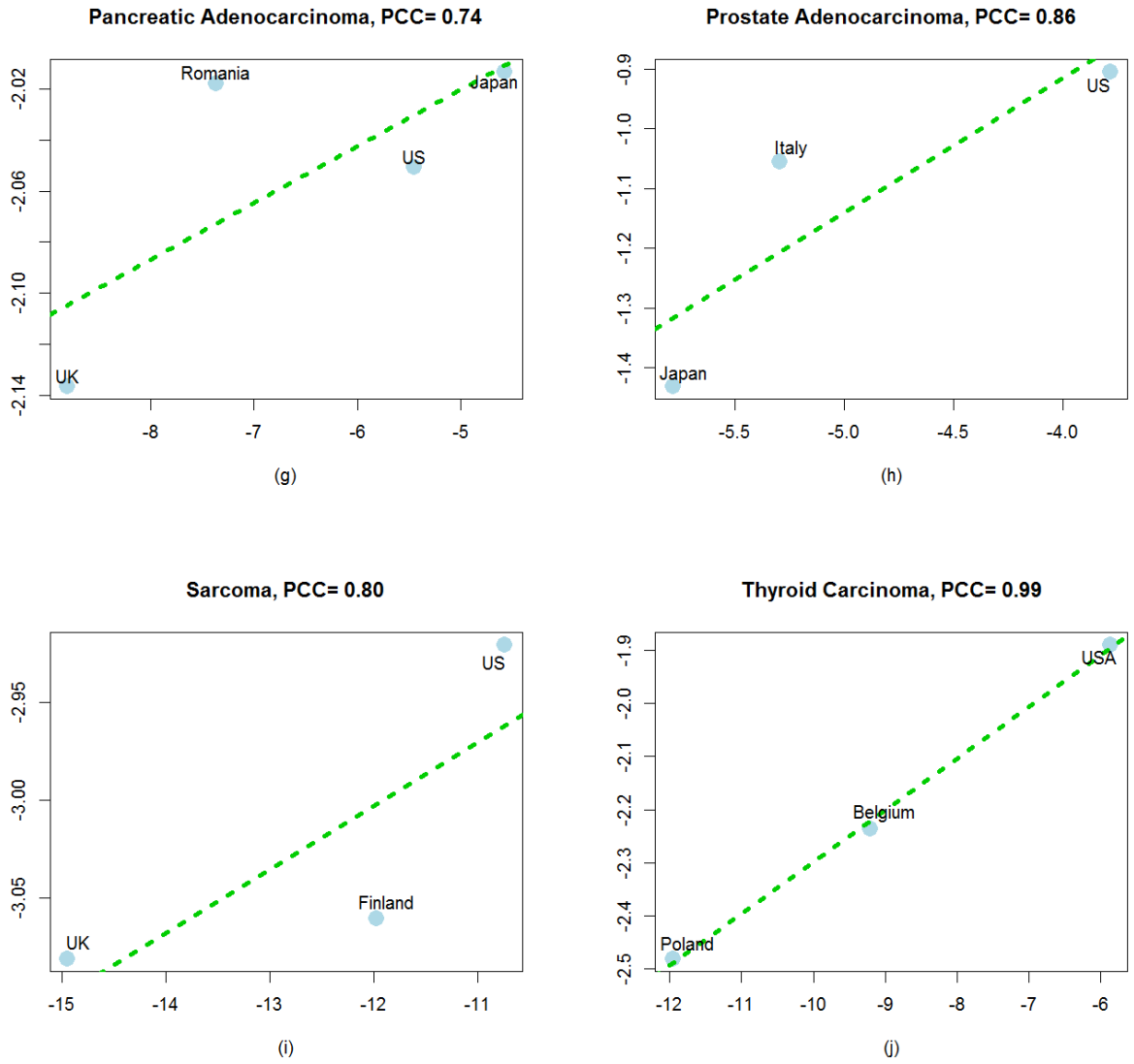


Figure 2.2: Each boxplot shows the distribution of lifetime risks across 284 regions in the world for 21 (solid green) cancer types. The y-axis represents the lifetime risk and the x-axis denotes the number of cell divisions as used in [13], in log10 scale. All the green dots are the same as the dots in Figure 2.1 of [13], and the regression line is based on the original data for the 31 cancer types.


Bladder Urothelial Carcinoma, PCC= 0.91
Spain
Switzerland
US
-1.80
-1.82
-1.84
-1.86
-5
-4
-3
-2
-1
(a)
Breast Carcinoma, PCC= 0.72
US
Italy
Canada
Czech Republic
Singapore
Spain
-1.00
-1.05
-1.10
-1.15
-10
-9
-8
-7
-6
(b)


Colon Adenocarcinoma, PCC= 0.89
Netherland
Hungary
Australia
Italy
Singapore
Portugal
Japan
Poland
USA
-1.30
-1.35
-1.40
-1.45
-1.50
-1.55
-10
-9
-8
-7
-6
-5
(c)
Kidney Renal Cell Carcinoma, PCC= 0.89
US
Germany
Netherland
Canada
-1.85
-1.90
-1.95
-2.00
-2.05
-2.10
-7.8
-7.6
-7.4
-7.2
-7.0
-6.8
-6.6
-6.4
(d)
Liver Hepatocellular Carcinoma, PCC= 0.96
US,hcv
Taiwan
China
Japan
US
-1.2
-1.4
-1.6
-1.8
-2.0
-6
-4
-2
0
(e)
Lung Adenocarcinoma, PCC= 0.9
US
(smoker
US
Germany
Spain
France
Taiwan
-0.5
-1.0
-1.5
-8
-7
-6
-5
-4
-3
(f)


Pancreatic Adenocarcinoma, PCC= 0.74
Romania
Japan
US
UK
-2.02
-2.06
-2.10
-2.14
-8
-7
-6
-5
(g)
Prostate Adenocarcinoma, PCC= 0.86
US
Italy
Japan
-0.9
-1.0
-1.1
-1.2
-1.3
-1.4
-5.5
-5.0
-4.5
-4.0
(h)
Sarcoma, PCC= 0.80
US
Finland
UK
-2.95
-3.00
-3.05
-15
-14
-13
-12
-11
(i)
Thyroid Carcinoma, PCC= 0.99
USA
Belgium
Poland
-1.9
-2.0
-2.1
-2.2
-2.3
-2.4
-2.5
-12
-11
-10
-9
-8
-7
-6
(j)

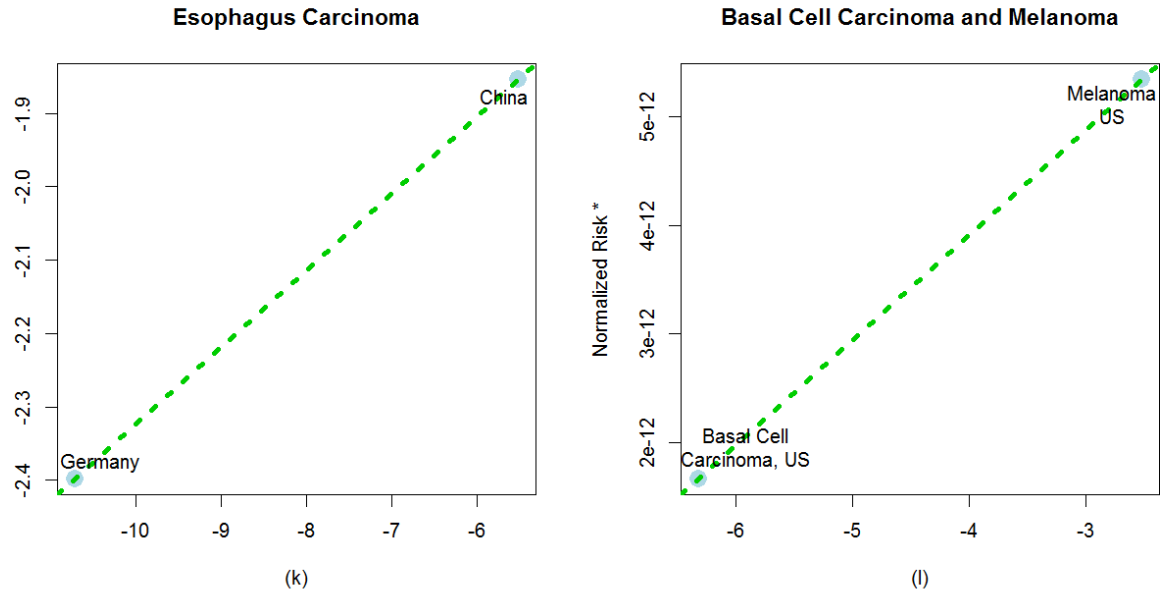


Figure S2.1: In each panel, the *x*-axis is the (average) predicted oxidative stress level in normal tissues where the relevant cancer type occur; and the *y*-axis is the lifetime risk of a cancer type. The dots represent all the countries with both cancer occurrence rates and the relevant gene-expression data available in the IARC and GEO databases. PCC is Pearson correlation coefficient. In (l), the risk of basal cell carcinoma and melanoma is normalized by using the risk / number of normal cells as given in [13] .

Tables

Table S2.1. Four countries with similar race distributions to that of the USA.

Brazil	48.43% white; 43.80% brown, 6.84% black; 0.58% Asian; and 0.28 Amerindian [88]
Costa Rica	65.8% white; 13.65% mestizo; 9.03% immigrants; 6.72% mulatto; 2.40% Amerindian; 1.03% black; 0.21% Asian [89]

Cuba	72% white, 20% black and 8% native American. [90]
Puerto Rico	75.8% white;12.4% black; 0.2% Asian [91]
USA	77.7% white; 13.2% black; 5.3% Asian; 1.2% as American Indian and Alaska Native [92]

Table S2.2. The lifetime risks (in log10 scale) of 21 cancer types in four countries with similar genetic backgrounds to that of the USA.

	Brazil	Costa Rica	Cuba	Puerto Rico
Acute myeloid leukemia	-2.655	-2.893	-3.172	-2.826
Chronic lymphocytic leukemia	-2.883	-3.092	-3.161	-2.944
Colorectal adenocarcinoma	-1.594	-1.703	-1.611	-1.401
Duodenum adenocarcinoma*	-3.570	-3.768	-3.871	-3.786
Esophageal squamous cell carcinoma	-2.287	-2.993	-2.786	-2.580
Gallbladder nonpapillary adenocarcinoma	-2.736	-2.772	-3.175	-3.170
Glioblastoma*	-2.307	-2.568	-2.373	-2.608
Head neck squamous cell carcinoma	-1.759	-2.296	-2.019	-2.018
Hepatocellular carcinoma	-2.873	-2.896	-3.056	-2.568
Lung adnocarcinoma nonsmokers *	-3.402	-3.833	-3.580	-3.584
Lung adenocarcinoma smokers *	-2.108	-2.568	-2.304	-2.316

Medulloblastoma	-3.907	-3.945	-3.922	-4.261
Melanoma	-2.273	-2.469	-2.660	-2.641
Osteosarcoma	-3.537	-3.863	-3.316	-3.781
Ovarian germ cell	-3.714	-3.454	-3.561	-3.583
Pancreatic ductal adenocarcinoma*	-2.210	-2.185	-2.225	-2.330
Pancreatic endocrine islet cell carcinoma*	-4.056	-4.031	-4.071	-4.176
Small intestine adenocarcinoma*	-3.224	-3.421	-3.524	-3.439
Testicular germ cell cancer	-3.160	-2.665	-3.494	-2.757
Thyroid papillary follicular carcinoma	-2.195	-2.171	-2.397	-2.089
Thyroid medullary carcinoma	-3.682	-3.778	-4.262	-3.832

Table S2.3: "Sample size" represents the number of tissue samples. "Accumulated risk" is calculated based on cancer occurrence rates calculated using the formula used in Tomasetti and Vogelstein's analysis. "Oxidative stress" is calculated using the model developed in the following section.

GEO accession number	**Country**	**Tissue type**	**Sample**	**Cumu risk (%)**	**Cumu risk (log10)**	**Average predicted oxidative stress level**
GSE11783	Switzerland	bladder	6	1.49	-1.8268	-1.2635
GSE30522	USA	bladder	2	1.36	-1.8665	-5.3666
GSE7476	Spain	bladder	3	1.64	-1.7852	-0.6505

GSE20711	Canada	breast	2	8.63	-1.064	-7.3728
GSE26457	USA	breast	42	10.13	-0.9944	-5.0013
GSE30010	USA	breast	107	10.13	-0.9944	-5.2755
GSE26457	USA	breast	71	10.13	-0.9944	-5.387
GSE26910	Italy	breast	4	9.45	-1.0246	-8.332
GSE29431	Spain	breast	9	6.8	-1.1675	-7.9668
GSE54002	Singapore	breast	16	6.94	-1.1586	-9.9986
GSE5764	Czech Republic	breast/ lobular	10	7.76	-1.1101	-8.813
GSE5764	Czech Republic	breast (ductal)	10	7.76	-1.1101	-10.6242
GSE14526	Japan	colon	1	3.72	-1.4295	-7.5042
GSE19963	Portugal	colon	4	3.78	-1.4225	-6.6172
GSE20916	Poland	colon	44	3.27	-1.4855	-9.2452
GSE4107	Singapore	colon	1	3.94	-1.4045	-6.2599
GSE33113	Netherland	colon	6	4.8	-1.3188	-6.5836
GSE41328	USA	colon	10	2.81	-1.5513	-10.5123
GSE4183	Hungary	colon	8	5.04	-1.2976	-5.0975
GSE9254	Australia	colon	19	4.44	-1.3526	-7.2539
GSE23194	Italy	colon	12	4.06	-1.3915	-7.3701
GSE26886	Germany	esophagus	19	0.4	-2.3979	-10.7126
GSE45670	China	esophagus	10	1.4	-1.8539	-5.6589

GSE29330	USA	head and neck	5	1.28	-1.8928	-5.3003
GSE6791	USA	head and neck	9	1.28	-1.8928	-3.6784
GSE11045	USA	kidney	3	1.39	-1.857	-6.434
GSE11151	Netherland	kidney	3	1.04	-1.983	-6.5987
GSE12606	Germany	kidney	1	1.27	-1.8962	-6.8596
GSE9489	Switzerland	kidney	13	0.74	-2.1308	-7.8584
GSE13471	USA	liver	5	0.73	-2.1367	-6.824
GSE23343	Japan	liver	7	1.08	-1.9666	-4.8958
GSE24042	China	liver	2	2.37	-1.6253	-0.0823
GSE38663	USA	liver (hcv)	14	7.1	-1.1487	-0.1776
GSE6222	Taiwan	liver	2	3.29	-1.4828	0.3553
GSE10799	Germany	lung	3	3.42	-1.466	-8.219
GSE18842	Spain	lung	45	3.75	-1.426	-4.1123
GSE19804	Taiwan	lung	60	3.06	-1.5143	-6.6337
GSE30219	France	lung	14	4.25	-1.3716	-6.2804
GSE19667	USA	lung	48	4.79	-1.3197	-7.6654
GSE19722	USA	lung	18	4.79	-1.3197	-6.5581
GSE19722	USA	lung (smoker)	28	86.22	-0.0644	-3.1144
GSE16515	USA	pancreatic	16	0.89	-2.0506	-5.4516
GSE15471	Romania	pancreatic	39	0.96	-2.0177	-7.3675

GSE19278	UK	pancreatic	7	0.73	-2.1367	-8.8038
GSE19650	Japan	pancreatic	8	0.97	-2.0132	-4.5748
GSE26910	Italy	prostate	6	8.81	-1.055	-5.2922
GSE3325	USA	prostate	5	12.45	-0.9048	-3.8101
GSE55945	USA	prostate	7	12.45	-0.9048	-3.7625
GSE45016	Japan	prostate	1	3.7	-1.4318	-5.7809
GSE17679	Finland	muscle	5	0.087	-3.0605	-11.9753
GSE34111	UK	muscle	6	0.083	-3.0809	-14.9432
GSE7014	USA	muscle	6	0.12	-2.9208	-10.7408
GSE15605	USA	normal skin	16	30	-0.5229	-6.3149
GSE7553	USA	melanocyte	4	2.03	-1.6925	-2.5187
GSE33630	Belgium	thyroid	45	0.58	-2.2366	-9.2088
GSE53157	Poland	thyroid	3	0.33	-2.4815	-11.9548
GSE6004	USA	thyroid	4	1.29	-1.8894	-5.8687

Table S2.4: Training data used for the oxidative stress predictor including cells treated with different levels of oxidative stress and non-cancerous diseases with distinct oxidative stress levels [93].

GEO accession number	**Data description (Disease data)**	**Data description (Normal data)**
GSE5339	Oxidative stress treatment	Normal control

GSE10896	Oxidative stress treatment	Normal control
GSE13931	Oxidative stress treatment	Normal control
GSE32169	Oxidative stress treatment	Normal control
GSE39156	Oxidative stress treatment	Normal control
GSE39843	Oxidative stress treatment	Normal control
GSE16759	Alzheimer's disease parietal lobe	Healthy parietal lobe
GSE28146	Alzheimer's disease gray matter	Healthy gray matter
GSE29652	Alzheimer's disease astrocyte	Healthy astrocyte
GSE4757	Alzheimer's disease entorhinal cortex	Healthy entorhinal cortex
GSE5281	Alzheimer's disease cortex types	Healthy cortex types
GSE53890	Alzheimer's disease frontal cortex	Healthy frontal cortex
GSE13396	Asthma bronchial epithelial cells	Healthy bronchial epithelial cells
GSE31773	Asthma immune cells	Healthy immune cells
GSE7368	Asthma bronchial epithelial cells	Healthy bronchial epithelial cells
GSE39843	Cystic fibrosis airway	non-cystic fibrosis airway
GSE22459	Fibrosis in kidney transplants	Healthy control
GSE38783	Hypertension stress on endothelial cell	Normal control
GSE24206	Idiopathic pulmonary fibrosis lung	Healthy control

GSE44723	Idiopathic pulmonary fibrosis fibroblast	Healthy fibroblast
GSE22255	Ischemic stroke peripheral blood mononuclear cells	Healthy peripheral blood mononuclear cells
GSE20141	Parkinson's disease SNpc neurons	Healthy SNpc neurons
GSE20146	Parkinson's disease globus pallidus interna	Healthy globus pallidus interna
GSE20153	Parkinson's disease EBV transformed cell lines	Control group EBV transformed cell lines
GSE30792	Parkinson's disease pluripotent stem cell	Control group pluripotent stem cell
GSE7621	Parkinson's disease substantia nigra	Healthy substantia nigra
GSE28133	Retinal detachment	Healthy control
GSE34748	Kidney transplant inflammation	Healthy control
GSE7392	Kidney transplant interstitial fibrosis	Healthy control
GSE27390	Rheumatoid arthritis	Osteoarthritis
GSE36700	Rheumatoid arthritis	Osteoarthritis

Table S2.5: Contributions by the 69 probes of 52 selected genes in the final logistic model for oxidative stress prediction.

Gene/Probe ID	**Coefficient**	**Gene/Probe ID**	**Coefficient**
Intercept	24.05149	NADSYN1_218840_s_at	1.488477

AKR1C1_1562102_at	0.351456	NAPRT1_226707_at	-0.46092
AKR1C1_244266_at	1.177194	NMNAT1_229852_at	-0.9353
AKR1C4_210558_at	1.576839	NMNAT2_1556029_s_at	-0.94922
ALDH1A1_212224_at	-0.97542	NMNAT2_1562818_at	-0.38317
ALDH1L1_205208_at	0.551058	NMNAT2_209755_at	0.067021
ALDH1L1_215798_at	-0.4525	NMNAT3_228090_at	-1.43924
ALDH1L2_1556841_a_at	-1.05907	NMNAT3_243738_at	-0.75224
ALDH9A1_201612_at	2.169183	NMRK1_1562761_at	0.420759
ART1_1570480_s_at	0.069599	NMRK1_219147_s_at	0.128116
ART3_210147_at	0.130363	NT5C_1557303_at	-0.67363
BLVRA_203771_s_at	-0.22258	NT5C1B_1554368_at	-1.36855
CAT_215573_at	-2.52956	NT5C2_236703_at	-1.96505
CBR1_209213_at	1.747679	NT5E_1553995_a_at	-0.62639
CYB5R1_1560043_at	-0.04199	PARP10_228669_x_at	-0.64716
CYB5R1_202263_at	-1.81661	PARP10_229350_x_at	-1.41537
CYB5R3_1554574_a_at	1.868582	PARP8_244008_at	0.28909
CYB5R3_201885_s_at	0.165056	PARP9_223220_s_at	2.177023
CYBB_203923_s_at	0.334607	PGD_1560942_at	-0.09484
CYBB_217431_x_at	-0.13617	PGD_1560943_s_at	-0.79528
CYBB_233538_s_at	-0.40776	PRDX2_201006_at	1.048308
CYP2R1_207786_at	-1.38472	PRDX3_201619_at	-1.45927
CYP39A1_1553977_a_at	0.700801	PRDX6_200844_s_at	-2.92782
CYP4F8_210576_at	-0.64903	PRDX6_200845_s_at	-0.50028

CYP51A1_216607_s_at	-3.26464	RDH14_222203_s_at	-3.3689
DHCR7_201790_s_at	2.360105	SIRT6_219613_s_at	-1.2108
GMPR2_217990_at	-1.84715	SIRT6_233179_x_at	-6.54817
GPX2_202831_at	0.2228	SOD2_215078_at	0.357994
GPX2_239595_at	0.045779	TNKS_216695_s_at	0.014178
GPX3_214091_s_at	0.380668	TNKS2_222563_s_at	-1.01799
HMGCR_202540_s_at	-0.45943	TNKS2_241909_at	-0.65438
IDH1_242956_at	0.663588	TXNRD1_201266_at	3.428549
IDO1_210029_at	2.483813	TXNRD2_211177_s_at	-0.00349
LPO_210682_at	-0.82915	TXNRD3_221906_at	-1.58521
MTHFR_239035_at	-2.88469	TXNRD3_59631_at	-0.36024

Table S2.6. Parameters of the estimated model.

Coefficients	**Value**
γ	-15.478
α	8.741
β_j: Bladder	0.4376
β_j: Breast	0.1302
β_j: Colon	0.1844
β_j: Esophagus	0.2473

β_j: Head and Neck	NA
β_j: Kidney	0.2828
β_j: Liver	0.3549
β_j: Lung	0.197
β_j: Melanoma	NA
β_j: Pancreatic	0.2964
β_j: Prostate	0.2282
β_j: Osteosarcoma	0.2366
β_j: Thyroid	0.234

Table S2.7. Functional groups and pathways significantly enriched (p-value < 0.01) with mutations in precancerous and cancerous colon tissues at different stages.

adenoma (small)	cell adhesion; fibronectins; cell motion; morphogenesis; glycoproteins; extracellular matrix; ECM-receptor interaction; cell cycle
adenoma (large)	glycoproteins; cell adhesion; fibronectin; EGF-like genes; ABC transporters; cadherin; extracellular matrix; actin-binding
colon cancer of dataset 1	cell adhesion; glycoprotein; extracellular matrix; immunoglobulin subtype; cell membrane; EGF-like genes; fibrinogen C terminal; differentiation; von Willebrand factor; laminin G; ECM receptor interaction
colon cancer (stage 1)	glycoprotein; cell adhesion; ion transport; EGF-like region; plasma membrane; cell morphogenesis; fibronectin; cytoskeleton; cadherin; immunoglobulin; laminin; endometrial cnancer; extracellular matrix; collagen' cytoskeleton organization; morphogenesis; cell junction organization; complement control
colon cancer (stage 2)	glycoprotein; cell adhesion; ionic channel' plasma membrane; EGF-like region; ion-binding; ATP-binding; extracellular matrix; fibronectin; laminin; synapse; guanyl nucleotide exchange factor; immunoglobulin subtype; motor protein; cell morphogenesis; ank-repeat; cytoskeletal part; cell motion; actin cytoskeleton; embryonic development

Colon cancer (stage 3)	glycoprotein; cell adhesion; fibronectin; immunoglobulin I-set; plasma membrane; ionic channel; EGF-like region; ion-binding; extracellular matrix; neuron differentiation; cell morphogenesis involved in differentiation; ATP-binding; cytoskeleton; laminin; microtubule; glutamate receptor activity; synapse; motor protein; detection of abiotic stimulus; calcium ion transport; tyrosine-specific protein kinase
colon cancer (stage 4)	glycoprotein; cell adhesion; EGF-like region; fibronectin; ATP-binding; plasma membrane; immunoglobulin; extracellular matrix; ion-transport; metal ion binding; transmission of nerve impulse; neuron differentiation; cytoskeletal part; sarcomere; muscle cell differentiation; leucine-rich repeat; laminin G; cell motion; triple helix and collagen; dynein heavy chain; calmoduling binding; dendrite; tyrosine protein kinase active site; GTPase binding.

CHAPTER 3

COMPETITION REGULATION AMONG DNA MEHTYLATION, NUCLEOTIDE SYNTHESIS AND ANTI-OXIDATION IN CANCER VS. NORMAL TISSUES

Cao, S., Zhu, X., Zhang, C., Qian, H., Schuttler HB., Gong, J., and Xu, Y. (2017), Competition between DNA methylation, nucleotide synthesis and anti-oxidation in cancer versus normal tissues.. Cancer Research. (In Press). Authors of articles published in AACR journals are permitted to use their article in support of dissertation.

Abstract

Global DNA hypo-methylation is observed in many cancer types. We present a computational study of genome-scale DNA methylation in 16 cancer types. Two issues are investigated: (1) the possible determinant of the global level of DNA methylation in cancer cells, and (2) the relationship between the DNA methylation level and the nucleotide-synthesis rate as well as the intracellular level of oxidative stress. We have developed a system of kinetic equations to capture the metabolic relations among DNA methylation, nucleotide synthesis, and anti-oxidative stress response, including their competition for methyl and sulfur groups, based on known information about the one-carbon metabolism and the trans-sulfuration pathway. Our main findings are: (i) there is a kinetic-based regulatory mechanism that controls the reaction rates of the three competing processes when their shared resources are limited, particularly when the nucleotide-synthesis rates and/or the oxidative states are high as generally the case in cancer; and (ii) it is the combination of this regulatory mechanism and the need for rapid nucleotide synthesis, as well as high production of glutathione dictated by cancer-driving forces, that leads to the nearly universal observation that cancers have reduced global-scale DNA methylation. Our model provides a natural explanation of why certain cancers have reduced global DNA methylation levels while others do not and why reduced DNA methylation levels tend to be associated with more malignant cancers. The novel insights obtained from this work provide useful information about the complexities of cancer due to interplay among competing, dynamic biological processes.

Introduction

It has been widely observed that cancer genomes tend to have increased DNA methylation levels in the promoter regions of their protein-encoding genes, but intriguingly their global methylation levels (including promoter and non-promoter regions) tend to decrease in comparison with normal

tissue cells. This has been observed in a variety of cancer types, including colon, liver, gastric, ovarian, breast, thyroid, and lung cancer [16-19]. Certain cancers can have over 50% reduction *vs.* their normal controls as observed in human primary GBMs and glioma cell lines [20]. It has been speculated that such reduced levels of global DNA methylation might have been selected to increase opportunities for the host cancer cells to overcome or to adapt to specific stresses encountered as reduced methylations generally imply increased gene expression [21].

A number of studies have been published aiming to explain the possible causes for the altered DNA methylation levels in cancer, predominantly with a focus on tumor suppressor genes. A popular view has been that increased methylation in the promoter regions of such genes will keep the expression of these genes low, hence enabling the survival of the cancerous host cells [94, 95]. A few studies have suggested possible causes of the reduced global DNA methylation in cancer [96]. One proposal is that dietary deficiency in methyl carriers (folate and methionine) could be a reason [22, 95] as insufficient methyl groups in diet have been linked to hepatic steatosis, cirrhosis, and even hepatic tumorigenesis [22]. It was speculated that methyl-carrying molecules entering one-carbon metabolism might be preferentially directed towards *de novo* synthesis of thymidylate needed for nucleotide synthesis at the expense of homocysteine re-methylation, hence resulting in reduced global DNA methylation during folate deficiency [23]. It is noteworthy that the one-carbon metabolism has three exits: the folate cycle, the methionine cycle and the transsulfuration pathway with the latter going to the production of glutathione (GSH), the main antioxidant in human cells. Since cancer cells generally have high oxidative states, they tend to produce more GSH molecules for anti-oxidation and survival, therefore fewer homocysteine molecules will be directed towards DNA methylation compared to normal controls, hence resulting in reduced global DNA methylation as another proposal suggests [97]. All these proposals suggest that the reduced

DNA methylation may be the result of competition among several processes, including nucleotide synthesis and anti-oxidation, for their shared resources but they are all speculative without a detailed mechanistic understanding that offers a reliable explanation of why other processes may outcompete DNA methylation and how this may be regulated.

We present here a computational study to address both the *why* and the *how* questions through (i) quantitative analyses of the well-established competitive relations among the three aforementioned processes using a set of ordinary differential equations; and (ii) an integrative analysis of epigenomic and transcriptomic data of cancer *vs.* control tissues of 16 cancer types to derive cancer type specific relations between DNA methylation and its competing processes. Specifically, we answer: (1) what may have caused the reduced level of global DNA methylation in cancer? And (2) why are the global DNA methylation levels distinct across different cancer types?

We first introduce some basics of the metabolic pathways under study. S-adenosyl methionine (SAM) is the most essential compound for DNA methylation, which consists of a *sulfur-containing group* and a *methyl group*. The sulfur-containing group comes from amino acid methionine, and the methyl group is from amino acid serine carried by folate. While DNA methylation consumes methyl groups, the sulfur-containing group is recycled back to methionine or converted to a GSH precursor. Figure 3.1 shows the detailed pathways for the methionine, folate and GSH metabolisms and their relations, through which metabolites *S*-adenosylhomocysteine (SAH), SAM, methionine and Hcy can be inter-converted. Since folate and GSH pathways do not interact directly (Figure 3.1), we study them separately and call the sub-system consisting of the folate and methionine pathways as F-M and the other with GSH and methionine pathways as G-M. Throughout the paper, we use transsulfuration pathway, anti-oxidation system and GSH pathways interchangeably. In addition, DNA methylation means global level DNA methylation, unless otherwise specified.

Materials and Methods

Data

DNA methylation data measured using HumanMehtylation450 arrays for 14 cancer types and 40 bisulfide sequencing data for eight cancer types are retrieved from the TCGA database. We have also used RNA-seq based gene expression data for 14 cancer types from the TCGA database. Six of these have all three data types available and the numbers of samples available for these three data types are summarized in Supplementary Table S3.3.1.

Method for estimating array-based DNA methylation level

The methylation array data (the HumanMehtylation450) we used cover ~485,000 CpG probes. These probes fall uniquely into one of the six categories: TSS1500, TSS200, 5' UTR, first exon, gene body, and 3' UTR, as summarized in Supplementary Table S3.2. Approximately half of the CpG probes are in more than one category. We have estimated each sample's total methylation level in each of the six CpG categories by summarizing the beta values of all the CpG islands that are located in each of each category, where a beta value is defined as the ratio between the methylated probe intensity and the overall intensity (sum of methylated and un-methylated probe intensities).

Method for estimating sequence-based DNA methylation level

For bisulfide sequencing data, we estimated the total methylation level of the CpG regions located in both gene regions and the LINE-1 regions in the genomes of 39 cancer patients the same way as in [98]. The total methylation level in genes is estimated by averaging the methylation levels across all CpG sites that are located within gene bodies. The total methylation level of gene promoter region is estimated by averaging the methylation levels across CpG sites that are located within 2Kb upstream of gene's transcription starting site. The total methylation level of the LINE-

1 elements is estimated by averaging the methylation levels across all CpG sites that are located within LINE1 elements across the whole genome.

Building equations for each reaction in nucleotide synthesis, DNA methylation and GSH synthesis

For each mono- or bi-substrate reaction under consideration (Figure 3.1), we built a reaction equation based on its kinetics defined by the Michaelis-Menten equation [99], where reaction rate V can be written as:

$$V = \frac{V_{max}[S]}{(K_m + [S])}$$

or

$$V = \frac{V_{max}[S_1][S_2]}{(K_{m,1} + [S_1])(K_{m,2} + [S_2])}$$

Reactions with non-standard forms are taken from [99] and detailed in Supplementary Table S3.3. All the model parameters, V_{max} and K_m, are collected from [99] (see Supplementary Table S3.4). We have noted that the F-M-G system falls into the category of *stiff* systems [100] as the concentrations of different molecular species differ by several orders of magnitudes. Numerical solvers to such ODEs tend to perform poorly on such systems. We have applied the following strategy in an iterative manner by alternatively working on the F-M and G-M systems separately to derive a solution to the F-M-G system, which takes advantage of the fact that there is a natural separation in time scales for the two sub-systems and neither of them is a stiff system.

We started by determining the initial values for the 15 molecular species used in the whole system. By analytically solving for the steady-state concentrations in the F, M, G systems individually using MATLAB function “solve”, we noted that 10 of the 15 species could be represented as analytical functions of the rest five. Though it was not likely that we could represent the steady

state concentrations of all the 15 species using one less variable (i.e. four variables), these reduced number of variables are primarily for reducing the search space for initial values and ease of computation, and theoretically shouldn't affect the identification of the steady state concentrations. We now define the steady-state concentrations of these five species in normal tissues as x_1^0, x_2^0, x_3^0, x_4^0, x_5^0. We will search the vicinity of each of these values, specifically $\frac{x_i^0}{10}$, $\frac{x_i^0}{5}$, x_i^0, $5x_i^0$, and $10x_i^0$ for $5 \geq i \geq 1$. For each of the 5^5 combinations of initial values, we calculate the corresponding values of the 10 dependent variables, and then do the following:

1. Solve the F-M system in steady state by treating the variables in the G system as constants using their current solutions;

2. Update the current solution to variables in the F-M system based on the solution derived in (1);

3. Solve the G system in steady state by treating the variables in F-M system as constants using their current solutions;

4. Update current solution to variables in the G system based on the solution derived in (3);

5. Repeat Steps 2-4 until the first derivatives of all the 15 variables, defined as $\frac{\mathrm{abs}(s_{t+dt}-s_t)}{dt}$, as well as their relative changes, defined as $\frac{\mathrm{abs}(s_{t+dt}-s_t)}{s_t}$, are within a pre-specified threshold, 0.01.

The convergence of the algorithm for finding steady state concentrations of the F-M-G system is established in Methods and Materials.

Relevant enzymes in F-M-G system and marker genes for estimating the levels of folate, serine, cysteine, and methionine

We have used known marker genes and their expression levels for estimating the level of each metabolite and enzyme used in our ordinary differential equations and analysis. Specifically, we have done the following, with detailed gene list given in Supplementary Table S3.5:

1. For folate, we have used expression of genes, *FPGS*, *GGH*, *SLC19A1*, encoding the folate homeostasis mediators as a measure for its concentration [101];

2. Transporters of amino acids cysteine, serine and methionine are taken from [102].

Estimating the expression level of a group of genes in a common pathway

Obtaining an overall expression of a group of genes on multiple samples, X, is to find a one-dimension representation, d, for X, which could be formulated as an optimization problem

$$\min_{\alpha,\mathrm{d}} f(\alpha, d) = ||X - \alpha d^T||_F$$

which is the largest eigen-value of the matrix

$$\sum_{i=1}^{N} x_i^T x_i$$

where x_i is the row vector of X. Note α serves as an ancillary variable in here.

Proof:

$$f(\alpha, d) = \sum_{i=1}^{N} (x_i - \alpha_i d^T)(x_i - \alpha_i d^T)^T = \sum_{i=1}^{N} (x_i x_i^T - 2\alpha_i d^T x_i^T + {\alpha_i}^2 d^T d)$$

$$\frac{\partial f(\alpha, d)}{\partial(\alpha_i)} = -2d^T x_i^T + 2\alpha_i d^T d$$

Let $\frac{\partial f(\alpha,d)}{\partial(\alpha_i)} = 0$, we have $\alpha_i^* = \frac{d^T x_i^T}{d^T d}$. We replace $f(\alpha, d)$ with α_i^*

$$f(\alpha, d) = \sum_{i=1}^{N} (x_i x_i^T - \frac{d^T x_i^T x_i d}{d^T d})$$

To minimize $f(\alpha, d)$ is equivalent to maximizing

$$\sum_{i=1}^{N} \frac{d^T x_i^T x_i d}{d^T d}$$

And that would be the eigenvector corresponding to the largest eigen-value of the matrix

$$\sum_{i=1}^{N} x_i^T x_i$$

Proof of convergence of the algorithm

Let $\vec{x}$ denotes the six species in folate cycle, $\vec{y}$ denotes the four species in methionine cycle, and $\vec{z}$ for the five species in glutathione cycle. The kinetic equation then has the form

$$\frac{d\vec{x}}{dt} = F(\vec{x}, \vec{y}, \vec{z}),\ \frac{d\vec{y}}{dt} = M(\vec{x}, \vec{y}, \vec{z}),\ \frac{d\vec{z}}{dt} = G(\vec{x}, \vec{y}, \vec{z}) = G(\vec{y}, \vec{z}).$$

The last equality holds because $\vec{x}$ does only affect the G through $\vec{y}$. It is because of this special form that allows us to introduce the iterative methods to solve the steady state of the stiff system with a give initial condition.

Define functions f, h, g in such ways that $\overrightarrow{x'} = f(\overrightarrow{z'}), \overrightarrow{y'} = h(\overrightarrow{z'})$ are analytical solutions to the F-M system in the following by treating $\vec{z} = \overrightarrow{z'}$

$$F(\vec{x}, \vec{y}, \overrightarrow{z'}) = \vec{0}$$

$$M(\vec{x}, \vec{y}, \overrightarrow{z'}) = \vec{0}$$

and similarly $\overrightarrow{z''} = g(\overrightarrow{y'})$ is analytical solution to the following system by treating $\vec{y} = \overrightarrow{y'}$

$$G(\vec{y}, \vec{z}) = \vec{0}$$

The original problem of finding $\overrightarrow{x^*}, \overrightarrow{y^*}, \overrightarrow{z^*}$ such that $F(\overrightarrow{x^*}, \overrightarrow{y^*}, \overrightarrow{z^*}) = 0, M(\overrightarrow{x^*}, \overrightarrow{y^*}, \overrightarrow{z^*}) = 0, G(\overrightarrow{y^*}, \overrightarrow{z^*}) = 0$ boils down to find $\overrightarrow{z'}$ such that $\overrightarrow{z'} = g\left(h(\overrightarrow{z'})\right)$, where f, h are functions defined in the context above. The proof is as below.

Let z' be such that $z' = g\left(h(\overrightarrow{z'})\right)$, and let $\overrightarrow{x'} = f(\overrightarrow{z'}), \overrightarrow{y'} = h(\overrightarrow{z'}), \overrightarrow{z"} = g(\overrightarrow{y'})$

Note that $\overrightarrow{z"} = g(\overrightarrow{y'}) = g\left(h(\overrightarrow{z'})\right) = \overrightarrow{z'}$, we then know that

$$F(\overrightarrow{x'}, \overrightarrow{y'}, \overrightarrow{z'}) = 0$$

$$M(\overrightarrow{x'}, \overrightarrow{y'}, \overrightarrow{z'}) = 0$$

$$G(\overrightarrow{y'}, \overrightarrow{z'}) = G(\overrightarrow{y'}, \overrightarrow{z"}) = 0$$

which means such derived $\overrightarrow{x'}, \overrightarrow{y'}, \overrightarrow{z'}$ is solution to the system. It is easy to show that stationary point of the system $\overrightarrow{x^*}, \overrightarrow{y^*}, \overrightarrow{z^*}$ also satisfies the condition that $\overrightarrow{z^*} = g\left(f(\overrightarrow{z^*})\right)$.

In each iteration step in our algorithm of looking for stationary points of the F-M-G system, we start by looking for $\overrightarrow{x'}, \overrightarrow{y'}$ such that for given initial value $\overrightarrow{z'}$

$$F(\overrightarrow{x'}, \overrightarrow{y'}, \overrightarrow{z'}) = 0$$

$$M(\overrightarrow{x'}, \overrightarrow{y'}, \overrightarrow{z'}) = 0$$

The next step is for solved $\overrightarrow{y'}$, look for $\overrightarrow{z"}$ such that

$$G\left(\overrightarrow{y'}, \overrightarrow{z"}\right) = 0$$

As shown above, the iterative process is equivalent as iteratively looking for the fixed point of $g(h(z))$. Since function $g(h(z))$ satisfies the assumptions in Banach Fixed Point Theorem and based on this theorem, we know that the iterative methods of looking for fixed points of $g(h(z))$ could converge, which means that our iterative method of looking for the F-M-G system's stationary point would converge too.

Results

Reduced global DNA methylation in cancer *vs.* normal control

We have examined DNA methylation data of 5,219 tissue samples of cancer *vs.* control tissues of 16 cancer types, out of which 5,179 samples are measured using the array technology and 40 samples measured using the bi-sulfide sequencing technology. The two datasets each cover different cancer types, with the array and sequencing data covering 14 and 8 cancers types, respectively, giving rise to a total of 16 distinct cancer types (see **Materials and Methods**). The array data consist of methylation data of 486,428 CpG islands, covering 99% of RefSeq genes, with an average of 17 CpG sites per gene distributed across its entire promoter region (i.e., TSS1500, TSS200, 5' UTR, and the first exon) and the gene body regions (i.e., gene body and the 3' UTR). The total methylation levels of promoter and gene body regions' CpG islands across different cancers are summarized in Supplementary Figure S3.1. Details regarding how the methylation level of each CpG category is estimated can be found in the **Materials and Methods** section.

We noted that cancer tissues tend to have significantly increased methylation levels in promoter regions (Figure S3.3.1A) and substantially reduced methylation in gene bodies (Figure S3.1B), which account for more than 80% of the CpG islands considered here. Analyses of bisulfide sequencing-based methylation data gives rise to the same conclusion that gene bodies tend to be hypo-methylated (Figure S3.1C) and promoter regions have increased methylation levels. In addition, we have also done similar analyses on DNA methylation in transposable elements, specifically LINE-1, which is accepted as a reliable measure for estimating the DNA methylation level at the genome scale [103]; and observed similar decrease in the methylation level in this

region (Figure S3.1D). We noted that among the 16 cancer types under study, two cancer types PRAD and BRCA have increased levels of methylations than their matching normal tissues.

In the following sections, we built mathematical models to study the detailed reasons and associated mechanisms for the observed hypo-methylation at the global scale in cancer. The units for concentration and time are uM and hr respectively.

The methionine cycle: dependencies of DNA methylation on folate and transsulfuration pathways

We have built a system of kinetic equations based on the known pathway models shown in Figure 3.1, to describe the reaction rates associated with 15 key molecular species in the folate, methionine and GSH (F-M-G) cycles, which consists of 24 enzyme-catalyzed reactions and four transporters, shown in Table 3.1 with detailed information of how each reaction equation is derived given in Supplementary Table S3.3.

We have examined the relationship between the DNA-methylation level and each of its four parameters, namely the levels of folate, serine, cysteine and methionine uptakes, respectively, according to this system of equations with all the four parameters having their values sampled from the normal ranges of their respective physiological values collected from [99]. Specifically, we have derived the numerical solutions of the fixed points of this system of equations for each combination of the parameter values uniformly sampled from the given ranges, using the MATLAB ode solver (see **Materials and Methods**). Figure 3.2 shows the level of the DNA methylation as a function of the levels of folate and serine concentrations (A-C) and as a function of the levels of cysteine and methionine concentrations (D-F).

We have observed that as the levels of serine (a methyl donor) and folate (a methyl carrier) increase, the DNA methylation level goes up; and as the level of methionine (a sulfur donor)

increases, the DNA methylation level goes up similarly. It is noteworthy that changes in the rate of cysteine uptake do not have a significant impact on the DNA methylation rate, since it goes into the downstream of the methionine cycle and it does not directly contribute to the sulfur supply for SAM used in DNA methylation. Here, the DNA-methylation rate, nucleotide synthesis rate and GSH synthesis rate are estimated as the reaction velocities catalyzed by enzymes *DNMT*, *TS* and *GS,* respectively (see Supplementary Table S3.3); the total methyl and the sulfur levels are estimated as the sum of concentrations of all metabolites that carry methyl and sulfur groups, respectively.

This analysis confirms that our system of equations captures the intuition that the level of DNA methylation should be an increasing function of the levels of methyl and sulfurs, provided by folate and serine, and methionine, respectively. In the following sections, we will study how this relationship is affected by other processes when they compete for methyl and sulfurs.

The F-M system: a model of competition for methyl between nucleotide synthesis and DNA methylation

As the ultimate donor of methyl, serine reacts with THF to generate 5,10-methylene-THF, during which methyl in serine is converted to a methylene group. Then this methylene group will go to one of three places: (1) as the methyl group of dTMP by the reaction catalyzed by thymidylate synthase (*TS*); (2) as the methyl group of 5mTHF catalyzed by *MTHFR* and ultimately for DNA methylation; and (3) to THF or 1,10-CH=THF. In our differential equations representing the F-M system, we have used the (steady state) levels of 5mTHF and DHF as estimates for the levels of methyl groups going to DNA methylation and nucleotide synthesis, respectively, since they are the immediate downstream metabolites of 5,10-methylene-THF going into the two processes.

Here, we address the following question: when competing for methyl groups between DNA methylation and nucleotide synthesis, does one process have an encoded priority over the other? We have assessed whether the two processes may have a competitive relationship as defined by our differential equations under a condition that nucleotide synthesis must be done at a rapid rate to mimic the typical situation in cancer. Specifically, we have checked how the DNA methylation rate changes when the rate coefficient V_{max} of *TS* increases, with *TS* being the key enzyme leading to nucleotide synthesis.

Figure 3.3 shows that as the need for nucleotide synthesis goes up, reflected by *TS*'s V_{max} value, methyl groups going to DNA methylation decreases (Figure 3.3A) while those going into nucleotide synthesis increase (Figure 3.3B). So do the DNA methylation rate (Figure 3.3C) and the nucleotide synthesis rate (Figure 3.3D), respectively. More specifically, as V_{max} increase from the low to the high end of its normal range, methyl going to DNA methylation is decreased by 4.8% while that going to nucleotide synthesis is increased by 40.6%. Correspondingly, the DNA methylation rate is decreased by 0.5% while the nucleotide synthesis rate is increased by 36.5%. We predict that this is the result of a regulatory mechanism which controls the competition between the two processes.

To elucidate this regulatory mechanism, we have conducted the following analysis focused on three enzymes *SHMT, TS* and *MTHFR* forming a Y shaped branch structure in the folate cycle with *SHMT* catalyzing the reaction leading to 5,10-m THF, which then branches out to nucleotide synthesis catalyzed by *TS* and to methionine cycle by *MTHFR* (see Figure 3.1). In cancer tissues, the expression level and hence the V_{max} of *TS* tend to increase substantially. We first examined the reaction rate constants of the three enzymes: V_{max}= 5200 uM/hr and K_m = 600 uM for *SHMT*; V_{max}= 5000 uM/hr and K_m = 6.3 uM for *TS*; and V_{max}= 5300 uM/hr and K_m = 50 uM for *MTHFR*.

Hence, the reaction rate of *TS* is close to two orders of magnitude higher than that of *SHMT*, ~91 times higher to be more exact, which will increase as the V_{max} of *TS* increases. We also noted that the substantial increase in the V_{max} of *TS* as done in the above illustrative example leads to only 2.9% increase in the concentration of THF (see Supplementary Table S3.6), indicating that the increase in the reaction rate of *SHMT* is limited by this number, regardless of the level of increase in the serine supply. Because of the tiny increase in this reaction rate, we assume, for the simplicity of discussion, that the rate remains unchanged. This immediately implies that the increased reaction rate of *TS* will take away a portion of the flux to 5mTHF to meet the need of the increased *TS* reaction rate; and the higher the *TS* reaction rate is increased, the higher proportion of the flux to 5mTHF will be diverted towards the reaction catalyzed by *TS*, hence more reduced level of DNA methylation.

In sum, it is the combination of the reaction rate constants in the folate cycle, particularly of three enzymes *SHMT*, *TS* and *MTHFR* along with their relative expression levels that play the key regulatory role in governing the competition for methyl groups between nucleotide synthesis and DNA methylation.

The G-M system: a sulfur-redistribution model: redox balance *vs.* DNA methylation

SAM serves a unique role in the system under study, as it not only contains a methyl group but also a sulfur group that carries the methyl molecule. As discussed earlier, the sulfur group in homocysteine can be recycled back to the methionine cycle and further to SAM, or it can go to the GSH metabolic pathway, indicating that DNA methylation also needs to compete with the GSH pathway for sulfur, in addition to its competition with nucleotide synthesis. This pathway starts with a reaction between homocysteine and serine (Figure 3.1), leading to the generation of cystathionine that is then cleaved by cystathionine lyase to generate α-ketobutyrate and cysteine,

which is then used for GSH production. Here we study how GSH production and DNA methylation may compete for sulfur, which ultimately affects the level of DNA methylation. Similar to the previous section, we use the (steady state) levels of two immediate downstream metabolites, methionine and cystathionine, respectively, to estimate the level of sulfurs going to DNA methylation and GSH synthesis, respectively.

We have examined how the two processes change their activity levels as the level of intracellular H_2O_2 goes up, where H_2O_2 is used to represent the oxidative state since it is the most abundant reactive oxygen species (ROS) in cancer in general [104]. Generally, as the H_2O_2 level goes up, the host cells will increase their GSH production to naturalize the excess H_2O_2 to keep the oxidative stress under control, which will consume sulfurs. Here, we show how an increased demand for sulfur by GSH production affects the level of DNA methylation. We have observed as the need for anti-oxidation and hence GSH synthesis goes up (reflected by the H_2O_2 level), sulfurs going to DNA methylation decrease (Figure 3.3E) while sulfurs going to GSH production increase (Figure 3.3F). So do the DNA methylation rate (Figure 3.3G) and GSH synthesis rate (Figure 3.3H). More specifically, sulfur going to DNA methylation is decreased by 13.6% and those to GSH production is increased by 6.2%. Correspondingly, DNA methylation rate is decreased by 1.9% and the GSH synthesis rate is increased by 2.6%. As in the previous section, we predict that this is the result of a regulatory mechanism that controls the flux of sulfur to different branches when they are limited.

We have conducted an analysis similar to that in the previous section on four enzymes: *SAHH, BHMS, MS* and *CBS* (Figure 3.1), which play key roles in the regulatory mechanism under investigation. As before, we noted: V_{max}= 320 uM/hr and K_m = 6.5 uM for *SAHH;* V_{max}= 2160 uM/hr and K_m = 12 uM for *BHMT;* V_{max}= 500 uM/hr and K_m = 1 uM for *MS;* and V_{max}= 700000 uM/hr and K_m = 1000 uM for *CBS.* From the above example, we observed: as the H_2O_2 level goes

up to the high end we set, the Hcy concentration goes up by 6% (see Supplementary Table S3.6). For the simplicity of discussion, we assume that there is no change in the Hcy concentration considering its tiny increase. All these reveal that the reaction rate of *CBS* goes up substantially and the Hcy flux into MET will go down. Knowing that the reaction rate of *CBS* is significantly higher than those of *BHMT* and *MS* based on their rate constants, we predict that the Hcy flux into MET and hence SAM will go down substantially, which is consistent with the observed change of MET in the above example, as detailed in Supplementary Table S3.6. In addition, H_2O_2 is known to have an inhibitory role on *MS* and *BHMT* (Figure 3.1). Hence the reaction rates of both *MS* and *BHMT* will go down as the H_2O_2 level goes up.

Based on the above, our prediction of the regulatory mechanism is: when the H_2O_2 concentration is not too high, its inhibitory roles on *MS* and *BHMT* will slow down the flux towards DNA methylation, leading to the accumulation of Hcy. As the H_2O_2 concentration further goes up, Hcy concentration continues to increase. Once the Hcy concentration is close to or exceeds the K_m value of *CBS,* the enzyme will instantly dump all the Hcy into the GSH synthesis pathway due to the very high reaction velocity of *CBS*. Overall, it is the combination of the reaction constants of the methionine cycle, particularly those of the four enzymes discussed here and the inhibitory role of H_2O_2 that controls the competition between DNA methylation and GSH synthesis.

We have also conducted a simulation analysis of the F-G-M system as a whole by systematically going through each of the 117 kinetic parameters encoded in the system by individually changing the value of each parameter, specifically through multiplying its default value by 0.1, 0.2, 0.5, 0.6, 0.8, 1.0, 1.2, 1.5, 1.6, 1.8, 2, respectively. The goal is to determine which of these parameters are most impactful on the competition under study. As expected, the kinetic parameters associated with *DNMT* and the uptake of folate, serine and methionine, respectively, are the most impactful,

all resulting in at least 10% change in the DNA methylation level through the above parameter manipulation. Interestingly, other most impactful parameters are those associated with *MTHFR, MS, CBS* and *BHMT*, which are at the core positions of the whole system, where redistributions of methyl and sulfur happen, each of which leads to at least 5% change in the level of DNA methylation.

In sum, our analyses revealed that there are regulatory mechanisms, largely encoded in the relative levels of their enzymes' reaction rate constants, that determine how the three processes compete for two shared resources: methyl and sulfurs.

We have previously demonstrated that the rate of nucleotide synthesis in cancer is dictated by the level of cytosolic Fenton reactions, which also largely determines the level of oxidative stress [105] while the level of cytosolic Fenton reactions is predominantly determined by concentrations of H_2O_2 and iron, for whose accumulation chronic inflammation is largely responsible [106]. Hence we predict that the observed genome-scale hypo-methylation in cancer is the result of these encoded regulatory mechanisms and the urgent need for rapid nucleotide synthesis and GSH synthesis, dictated by the level of Fenton reactions.

With this established framework, we address why different types of cancers may have different levels of global DNA methylation using cancer tissue gene-expression data.

Cancer specific DNA methylation

To understand why different cancer types may have different levels of hypo-methylation *vs.* their normal controls, we have conducted cancer-specific analyses of the integrated model of the above three subsystems, collectively referred to as the *F-M-G system,* through applying the observed gene-expression levels of the relevant enzymes, normalized with respect to their corresponding normal controls. Supplementary Figure S3.2 shows differentially expressed genes across the F-M-

G system. Note that for cases where multiple genes encoding one enzyme or transporter, we estimated the integrated expression level of the gene group as a whole, using a method given in **Materials and Methods**.

We have developed a model for predicting the DNA-methylation level of a given cancer tissue based on the expression levels of selected enzymes in the F-M-G system. We used the Michaelis-Menten equation to capture how the reaction velocity V depends upon the concentrations of the main substrate S and the catalyzing enzyme E:

$$V = \frac{V_{max}[S]}{K_m + [S]} = \frac{k_{cat}[E][S]}{K_m + [S]}$$

Here, we can reasonably assume that the relevant reaction rate constant K_m for each enzyme is the same across different cancer types. Hence, the rate of each reaction under consideration is entirely determined by $[S]$ and $[E]$. The cancer specific enzyme concentration $[E]$ would be that of normal condition multiplied by the fold change in the enzyme's gene expression levels in cancer *vs.* control tissues. For each cancer type, we estimated an "average" fold-change in expression levels of genes encoding the relevant enzymes/transporters between cancer and control samples, as described in **Materials and Methods**. Figure S3.2 shows changes in concentrations of the relevant enzymes and transporters across different cancers *vs.* corresponding controls. Clearly, so estimated enzyme concentrations will give rise to different steady-state concentrations of each metabolite and reaction rates, as shown in the 4th column of Table 3.2.

Our cancer specific DNA methylation prediction is calculated for ten cancer types, which have both methylation array data and RNA-Seq gene-expression data for cancer and control samples. For each cancer type, we have estimated the steady-state concentrations of all the relevant metabolites, particularly the DNA-methylation level using the reaction rate catalyzed by enzyme DNMT. As shown in Table 3.2, our predictions of the DNA methylation levels are highly

consistent with the experimental data in the eight cancer types considered. For the two cancer types where our predictions are not consistent with experimental data, namely BRCA and THCA, we believe that the reason for the hypo-methylation prediction in BRCA and hyper-methylation for THCA is due to the possibility that certain factors that may also contribute to DNA methylation are not included in our model.

We have then studied how DNA methylation levels differ when patients have different levels of GSH synthesis and nucleotide synthesis rates. The analysis is done on the same ten cancer types as above. We noted that samples with higher nucleotide synthesis or GSH synthesis levels tend to have lower DNA methylation rates, consistent with our model that DNA methylation is at an inferior position when competing for shared resources with nucleotide synthesis and GSH synthesis. This is the case for seven out of ten cancer types but not for KIRC, PRAD and THCA as shown in Table 3.3. These three cancer types clearly warrant further studies in order to understand why they behave differently from the other seven cancer types. Note that the levels of GSH synthesis and nucleotide synthesis are estimated using genes involved in glutathione synthesis and RNA polymerases pathways, respectively, with details given in **Materials and Methods**.

We further studied whether DNA methylation is indeed competing with nucleotide synthesis and anti-oxidation system for methyl and sulfur. Particularly, we are interested in the relationships between DNA methylation and nucleotide synthesis/anti-oxidation capacity when methyl/sulfur is limited. To accomplish this, we have introduced a measure of the *average methyl/sulfur availability*: the ratio between nucleotide synthesis/anti-oxidation capacity and methyl/methionine availability with the property: the lower the ratio is, the lower the average availability of methyl/methionine is for the nucleotide synthesis/anti-oxidation capacity. For methyl compound,

we have observed significant correlations between DNA methylation and nucleotide synthesis in those samples with high and low average methyl availability, respectively; and similarly for sulfur, we have also observed significant correlations between DNA methylation and anti-oxidant capacity in those samples with high and low average sulfur availability, respectively. Here, the levels of methyl and methionine, anti-oxidation capacity and nucleotide synthesis are estimated using genes involved in folate and methionine transporters (Supplementary Table S3.5), glutathione synthesis and RNA polymerases pathways, respectively, with details given in **Materials and Methods**.

As shown in Table 3.4, when the average methyl availability is low, most cancer types showed significant negative correlations (p-value cutoff: 0.05) between DNA methylation and nucleotide synthesis, except for KIRC, PRAD and THCA; and when the average sulfur availability is low, most cancer types showed significant negative correlations (p-value < 0.05) between DNA methylation and the anti-oxidant capacity, except for BRCA, KIRC, PRAD and THCA. These data strongly suggest competitive relations between DNA methylation and nucleotide synthesis/anti-oxidation capacity for methyl and sulfur. It has also explained: (1) why our prediction of the methylation level in BRCA is not accurate shown in Table 3.2; and (2) why the DNA methylation levels do not depend on nucleotide synthesis or anti-oxidation capacity as shown in Table 3.4, as methyl and sulfur may be not limited resources in these cancer types, i.e., not rate-liming factor in the cancerous cell division; and DNA methylation does not need to compete for the two resources with other processes. We have also noticed that even when the average methyl/sulfur availability is high, the negative correlations are also significant (p-value < 0.05) for some cancer types. Hence, we posit that even though the average availability of methyl/sulfur is high in thee samples

compared to other samples, these resources are still limited with respect to their cell division rates dictated by cytosolic Fenton reactions (see Discussion), and hence competitions are still there.

Discussion

We have previously developed a model proposing that Fenton reactions, $Fe^{2+} + H_2O_2$ -> $Fe^{3+} + OH^- + \bullet OH$, in cytosol and mitochondria may represent key drivers of cancer initiation and progression at a more basic level than the previously proposed drivers such as genomic mutation [107], epigenomic alteration [108] and metabolic reprogramming [109]. Fenton reactions have been found to take place when concentrations of Fe^{2+} and H_2O_2 are sufficiently high in the same location without involvement of any enzyme. When there are plentiful reducing elements near the reaction sites, such as sulfur, NADH or superoxide, Fe^{3+} can be reduced to Fe^{2+}, hence enabling the reaction to continue indefinitely, which is also referred to as the Harbor-Weiss reaction [110]. Then the reaction can be rewritten as: $O_2^- + H_2O_2 \rightarrow \cdot OH + OH^- + O_2$ with Fe^{2+} as a catalyst since it is not consumed by the (continuous) reaction; and superoxide as the reducing element as our analysis revealed that O_2^- is the most commonly used reducing element in cancer [111]. We have demonstrated statistically that all cancers in the TCGA database have Fenton reactions in their cytosol, mitochondria and extracellular matrix and space [112]. An important implication of this model is that cytosolic Fenton reactions drive *de novo* nucleotide synthesis (and glycolytic ATP production) to produce net protons (H^+) at rates comparable to those of OH^--producing Fenton reactions, hence keeping intracellular pH stable. More specifically, it is the rates of cytosolic Fenton reactions that dictate the rate of nucleotide synthesis and hence the rates of DNA synthesis and cell division in cancer.

This model, in conjunction with the discovery made here, strongly suggests that there is an encoded regulatory mechanism that determines the winners in competition for methyl and sulfurs when

they are of limited availability, and it offers a natural explanation of why cancers in general have reduced genome-scale DNA methylation as well as why different cancers tend to have distinct levels of global DNA methylation, hence addressing an important and open question in cancer biology.

In addition, we have also provided an explanation as why certain cancers tend to have more reduced DNA methylations, i.e., those with higher levels of nucleotide synthesis and oxidative stresses, hence having established that the global DNA methylation level could be used as a predictor for more aggressive cancer types.

Further extension of the current study will include detailed metabolisms of serine, methionine and H_2O_2 to make the model more realistic: concentrations of the latter compounds are presently treated as input parameters, rather than treating them in a more realistic manner via explicitly modeling the ways that they are actually brought into cancer cells. In addition, the observation on BRCA and THCA were that the global methylation levels in their tumor samples do not differ significantly from their normal samples, while our model predicts that BRCA is hypo-methylated, and THCA is hyper-methylated. We believe that these discrepancies are due to the assumption in our model that only two factors contribute to the global DNA methylation level, namely the competitions for sulfur and for methyl groups with two other processes. Careful inspection of the data in Table 3.4 revealed that the global methylation levels have no significant negative correlation with neither of the two competing processes, namely nucleotide synthesis and anti-oxidation system in BRCA and THCA, suggesting that other factors, such as hormones, may also affect the global methylation level of DNA. We examined the global methylation levels of different subtypes of breast cancer, and noted that triple-negative breast cancers (free of hormone regulation) are significantly hypo-methylated in tumor (p-value= 0.0129), consistent with our

model prediction. This suggests one possible direction for further development of our model in the future. We believe that our study here offers a good example for studying complex, non-linear relationships among multiple players involved in specific biological processes, leading to novel understanding about previously made perplexing observations, and can be applied to study a suite of such problems in cancer research.

Figures:

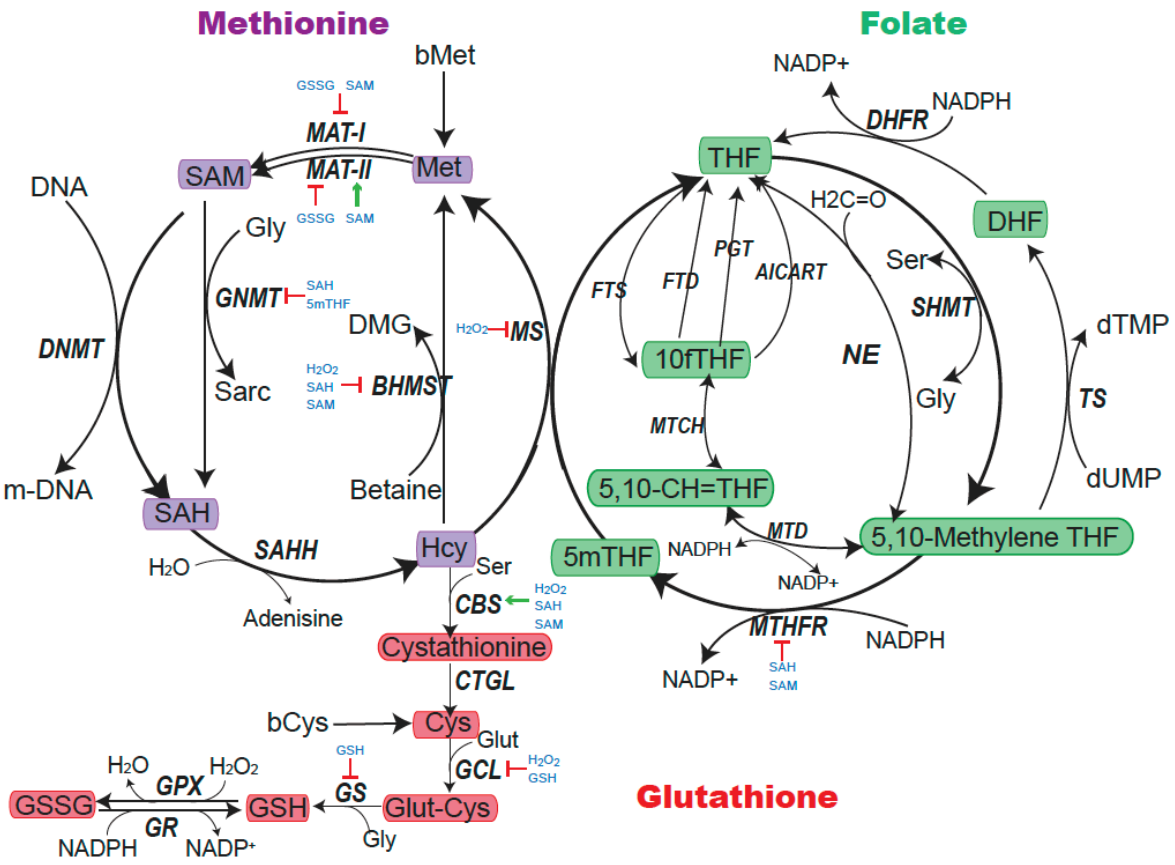


Figure 3.1: The one-carbon metabolic pathway (adapted from [99]) consisting of the folate, the methionine and the transsulphuration pathways. All the reaction substrates are in upright letters and the catalyzing enzymes are in italics. Substrates in purple, green and red represent metabolite variables in three subsystems in our model, and all the other substrates are treated as constants. Metabolites in light blue are those that could activate or inhibit certain reactions. bMET and bCYS represent methionine and cysteine up-taken from the blood circulation, respectively.

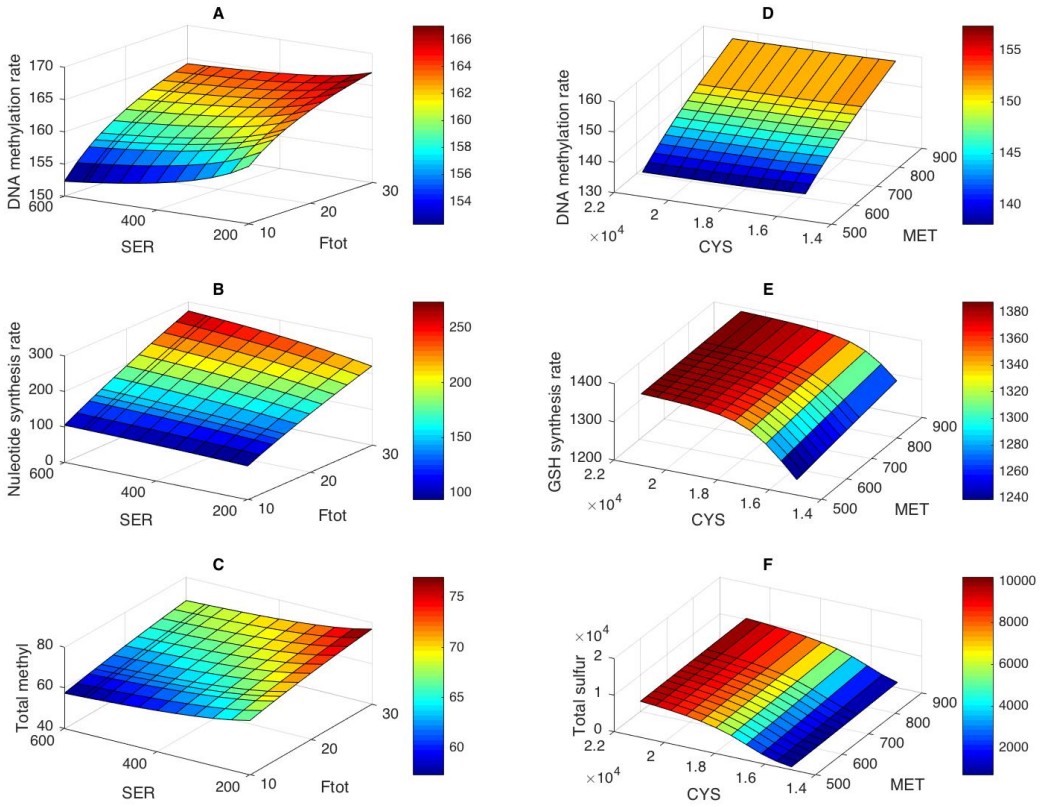


Figure 3.2: Under normal cytosolic conditions, the DNA methylation rate (A), nucleotide synthesis rate (B) and methyl concentration (C) (*z*-axis) in steady states as a function of the serine (*x*-axis) and folate (*y*-axis) concentrations, respectively; and DNA methylation rate (D), GSH synthesis rate (E) and total sulfur concentration (F) (*z*-axis) as a function of the cysteine (*x*-axis) and methionine (*y*-axis) influx from blood circulation, respectively.

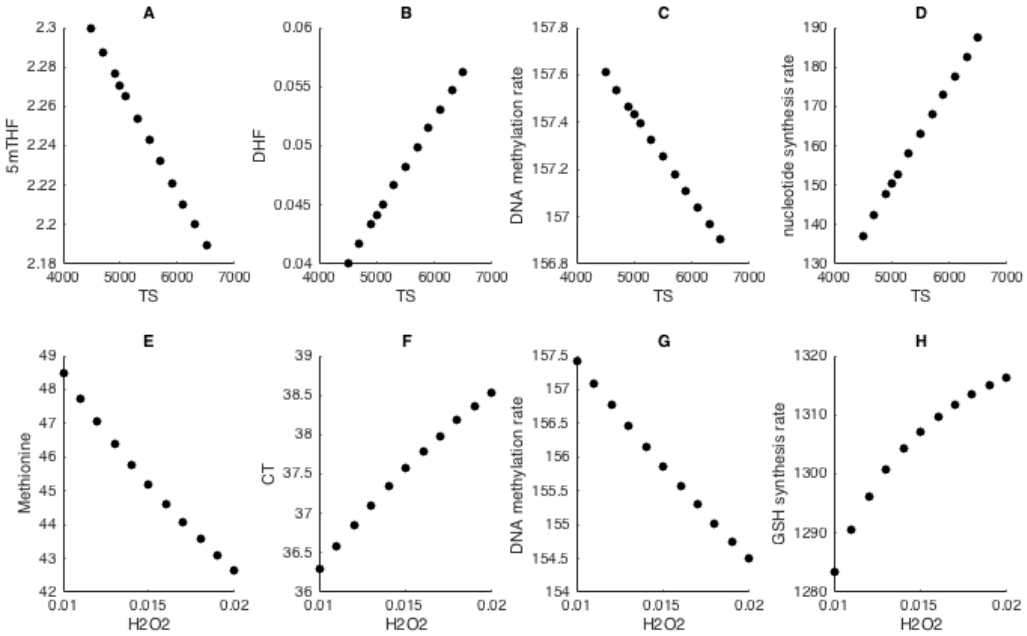
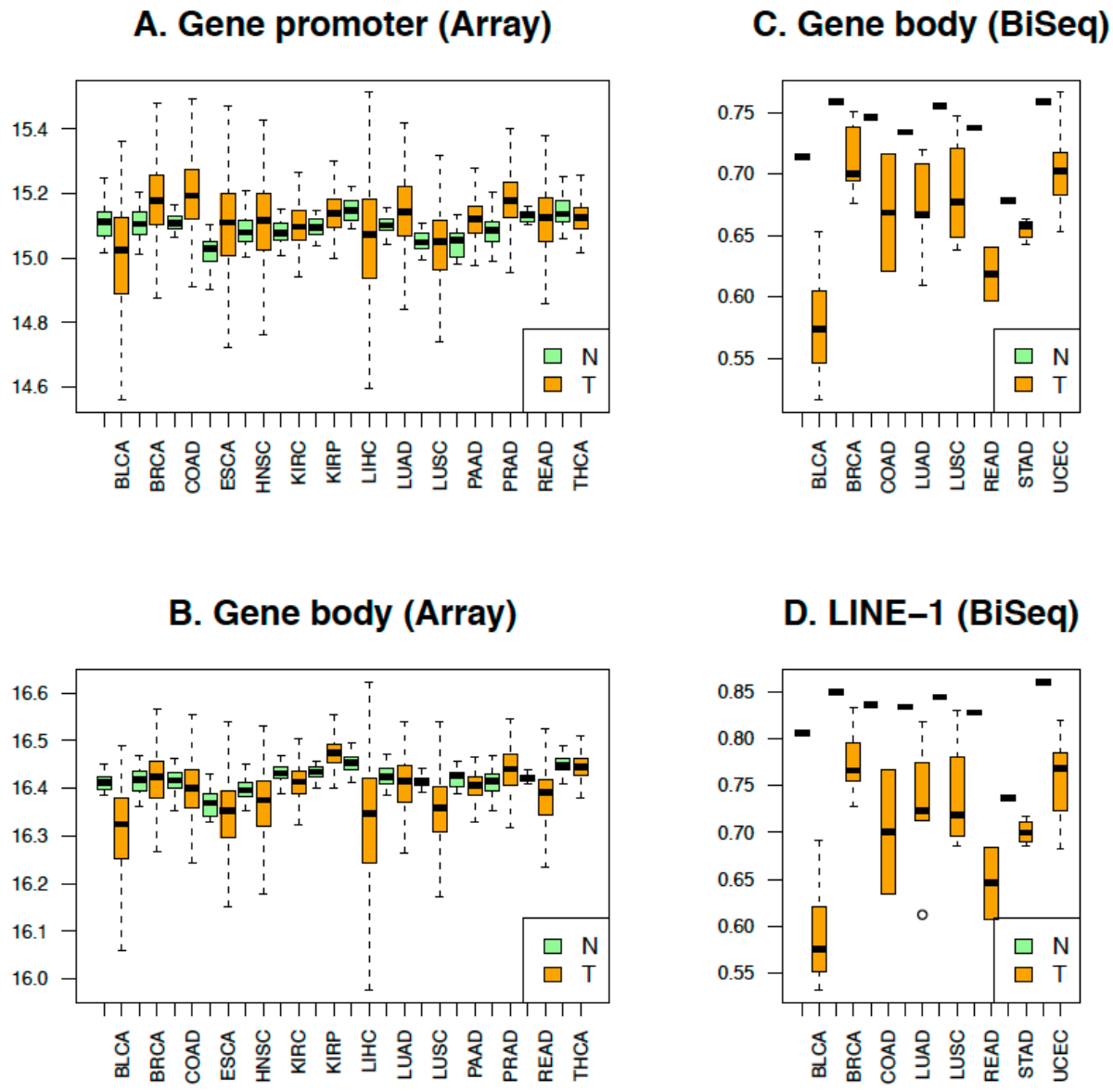
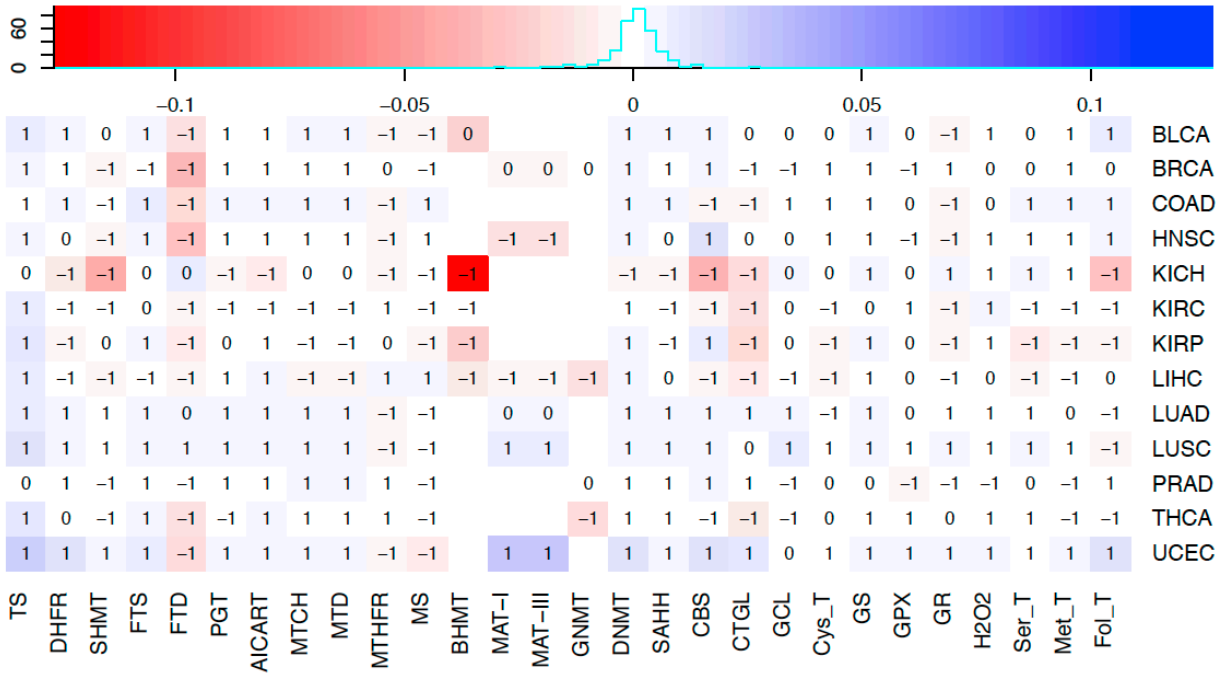


Figure 3.3: (A) The level of 5mTHF (*y*-axis), reflecting the level of methyl going to DNA methylation as a function of the *TS*'s V_{max} value (*x*-axis); (B) The level of DHF (*y*-axis), reflecting the level of methyl going to nucleotide synthesis as a function of the *TS*'s V_{max} value (*x*-axis); (C) the DNA methylation rate (*y*-axis) as a function of the *TS*'s V_{max} value; (D) the nucleotide synthesis rate (*y*-axis) as a function of the *TS*'s V_{max} value; (E) The level of methionine (*y*-axis), representing the level of sulfur going to DNA methylation as a function of the H_2O_2 concentration (*x*-axis); (F) The level of cystathionine (*y*-axis), reflecting the level of sulfur going to GSH synthesis as a function of the H_2O_2 concentration (*x*-axis); (G) The DNA methylation rate (*y*-axis) as a function of the H_2O_2 concentration; (H) The GSH synthesis rate (*y*-axis) as a function of the H_2O_2 concentration.



Supplementary Figure S3.1: (A) Boxplots of the methylation levels in CpG islands in gene promoter regions for 14 cancer types (all array data), with orange and green boxplots for samples of cancer and control tissues, respectively; (B) Boxplots of the methylation levels of CpG islands located inside gene bodies for 14 cancer types; (C) Boxplots of the average methylation levels of CpG islands in gene body regions for eight cancer types (all sequencing data); and (D) Boxplots of the average methylation levels of CpG islands in LINE-1 regions for eight cancer types. Note

that for bi-sulfide sequencing data shown in (C) and (D), the normal control group has only one sample for each cancer type, thus the boxplots look like a black and thick line.



Supplementary Figure S3.2: Differential expression of the relevant enzymes and transporters (with suffix "_T"). Genes (groups) are on the *x*-axis, and cancer types are on the *y*-axis. "1" and "-1" represent significant up- and down-regulation of the relevant enzymes in cancer samples compared with normal control samples, and "0" indicates no significant changes between cancer and normal controls. Entries with missing numbers indicate missing data.

Tables:

Table 3.1: Reaction rates associated with all the metabolites involved in folate, methionine and GSH metabolisms, with the detailed information for each rate along with an explanation, given in Supplementary Table S3.3.

$\frac{d[DHF]}{dt}$	$V_{TS}([dUMP],[CH2F]) - V_{D\square FR}([DHF],[NADPH])$
$\frac{d[5mTHF]}{dt}$	$V_{MTHFR}([CH2F],[NADPH],[SAM]) - V_{MS}([5mTHF],[HCY],[H_2O_2])$
$\frac{d[THF]}{dt}$	$V_{FTD}([10fTHF]) + V_{MS}([5mTHF],[HCY],[H_2O_2])$ $+ V_{PGT}([10fTHF],[GAR]) + V_{ART}([10fTHF],[AICAR])$ $- V_{FTS}([THF],[HCOOH])$ $- V_{SHMT}([SER],[THF],[GLY],[CH2F])$ $- V_{NE}([THF],[H_2C=O],CH2F) + V_{DHFR}([DHF],[NADPH])$
$\frac{d[CH2F]}{dt}$	$V_{SHMT}([SER],[THF],[GLY],[CH2F]) + V_{NE}([THF],[H_2C=O],CH2F)$ $- V_{TS}([dUMP],[CH2F])$ $- V_{MTHFR}([CH2F],[NAD\square H],[SAM])$ $- V_{MHD}([CH2F],[CHF])$
$\frac{d[CHF]}{dt}$	$V_{MHD}([CH2F],[CHF]) - V_{MCH}([CHF],[10fTHF])$
$\frac{d[10fTHF]}{dt}$	$V_{MCH}([CH2F],[10fTHF]) + V_{FTS}([THF],[HCOOH])$ $- V_{PGT}([10fTHF],[GAR]) - V_{ART}([10fTHF],[AICARP])$ $- V_{FTD}([10fTHF])$
$\frac{d[MET]}{dt}$	$V_{BHMT}([HCY],[BET],[SAM],[SAH],[H_2O_2]) + V_{MS}([5mTHF],[HCY],H_2O_2)$ $+ V_{bMetIn}([bMET],[MET]) - V_{MATI}([MET],[SAM],[GSSG])$ $- V_{MATIII}([MET],[SAM],[GSSG])$
$\frac{d[SAM]}{dt}$	$V_{MATI}([MET],[SAM],[GSSG]) + V_{MATIII}([MET],[SAM],[GSSG])$ $- V_{GNMT}([SAM],[SAH],[5mTHF],[GLY])$ $- V_{DNMT}([SAM],[SAH],[DNA])$

$\frac{d[SAH]}{dt}$	$V_{GNMT}([SAM],[SAH],[5mTHF],[GLY]) + V_{DNMT}([SAM],[SAH],[DNA]) - V_{SAAH}([SAH],[HCY])$
$\frac{d[HCY]}{dt}$	$V_{\square AAH}([SAH],[HCY]) - V_{CBS}([HCY],[SAM],[SAH],[SER],[H_2O_2]) - V_{BHMT}([HCY],[BET],[SAM],[SAH],[H_2O_2]) - V_{MS}([5mTHF],[HCY],H_2O_2)$
$\frac{d[GSH]}{dt}$	$V_{GS}([GLY],[GLC],[GSH]) - (2)V_{GPX}([GSH],[H_2O_2]) + (2)V_{GR}([GSSG],[NADPH]) - d_1[GSH] - V_{bGSHOut}([GSH])$
$\frac{d[GSSG]}{dt}$	$V_{GPX}([GSH],[H_2O_2]) - V_{GR}([GSSG],[NADPH]) - d2[GSSG] - V_{bGSSGOut}([GSSG])$
$\frac{d[GLC]}{dt}$	$V_{GCS}([CYS],[GLU],[GSH],[GLC],[H_2O_2]) - V_{GS}([GLY],[GLC],[GSH])$
$\frac{d[CYS]}{dt}$	$V_{CTGL}([CYT]) - V_{GCS}([CYS],[GLU],[GSH],[GLC],[H_2O_2]) + V_{bCYSIn}([bCYS]) - 0.35 * \frac{[CYS]^2}{200}$
$\frac{d[CYT]}{dt}$	$V_{CBS}([HCY],[SAM],[SAH],[SER],[H_2O_2]) - V_{CTGL}([CYT])$

Table 3.2: Cancer specific DNA methylation: observed *vs.* predicted levels. Column 2 is the observed DNA methylation level changes calculated using methylation array data with *p*-values of hyper-, no change or hypo-methylation shown in the third column. Columns 4 and 5 are predicted steady-state DNA methylation levels for cancer and control tissues, where a cancer type is predicted to have hypo-methylation if the predicted level of DNA methylation is lower in cancer compared to that in controls. A prediction is considered to be consistent with experimental data if they both show hypo- or hyper-DNA methylation.

Cancer type	Observed methylation changes	p-value	Predicted cancer (μM/hr)	Predicted control (μM/hr)
BLCA	Hypo	1.33E-08	81.95	94.71
BRCA	No change	5.13E-01	90.27	94.71
COAD	Hypo	2.51E-02	86.99	94.71
HNSC	Hypo	8.39E-04	89.61	94.71
KIRC	Hypo	7.25E-13	82.48	94.71
LIHC	Hypo	5.10E-13	85.76	94.71
LUAD	Hypo	4.11E-02	92.59	94.71
LUSC	Hypo	7.63E-11	85.09	94.71
PRAD	Hyper	1.00E-04	99.83	94.71
THCA	No change	9.13E-02	106.49	94.71

Table 3.3: Comparisons of DNA methylation levels between samples with different levels of anti-oxidation and nucleotide synthesis rates for 10 cancer types. The second column shows *p*-values of Wilcoxin tests for the null hypothesis that DNA methylation levels of patient samples with high GSH synthesis rate is lower than those with relatively lower GSH synthesis activities; the third columns shows *p*-values of Wilcoxin tests for the null hypothesis that the DNA methylation levels of patient samples with high nucleotide synthesis rate is lower than those with relatively lower nucleotide synthesis rates.

Cancer type	Anti-oxidation	Nucleotide synthesis
BLCA	**2.86E-03**	**1.25E-03**
BRCA	1.62E-01	5.54E-02
COAD	**2.60E-02**	**1.21E-02**
HNSC	**2.44E-04**	**1.38E-05**
KIRC	9.58E-01	5.97E-01
LIHC	1.68E-01	**9.69E-06**
LUAD	**8.40E-04**	**6.65E-05**
LUSC	**1.40E-02**	**2.90E-05**
PRAD	5.98E-01	6.94E-01
THCA	1.00E+00	9.34E-01

Table 3.4: The significances of observed negative correlations between: (1) DNA methylation and nucleotide synthesis when high (Methyl_H) and low (Methyl_L) level of methyl is available; and (2) DNA methylation and anti-oxidation capacity when high (Sulfur_H) and low (Sulfur_L) level of sulfur is available. Significant negative correlations (p-value < 0.05) are marked bold.

	Methyl_H	Methyl_L	Sulfur_H	Sulfur_L
BLCA	**5.50E-05**	**5.26E-03**	8.45E-02	**2.20E-02**
BRCA	4.02E-01	**1.96E-02**	2.47E-01	2.10E-01
COAD	6.21E-01	**3.14E-04**	1.03E-01	**2.89E-03**
HNSC	**6.00E-04**	**2.25E-05**	1.62E-01	**1.10E-02**
KIRC	4.69E-01	7.26E-01	9.99E-01	5.14E-01
LIHC	**5.63E-04**	**2.84E-05**	9.32E-01	**5.00E-02**
LUAD	1.02E-01	**3.51E-05**	**1.15E-02**	**2.08E-04**
LUSC	**1.59E-06**	**1.08E-03**	1.18E-01	**1.36E-02**
PRAD	6.90E-01	4.73E-01	9.64E-02	1.82E-01
THCA	9.10E-01	9.60E-01	9.41E-01	9.97E-01

Table S3.1: The numbers of cancer and control tissue samples with DNA methylation data, measured using bisulfide array or sequencing, along with RNA-seq gene expression data of cancer *vs.* control samplers, used in this study. Blanks indicate that the data type is missing.

	Methylation		Gene expression
Cancer	Array	Sequencing	
BLCA	21/259	1/6	19/408
BRCA	96/745	1/5	113/1095
COAD	38/290	1/2	41/285
HNSC	50/517		44/520
KIRC	160/301		72/533
KIRP	45/182		32/290
LIHC	50/204		50/371
LUAD	32/452	1/5	59/515
LUSC	42/359	1/4	51/501
PRAD	49/340		52/497
THCA	56/508		59/505

UCEC		1/5	24/176
READ	7/96	1/2	
STAD		1/4	33/238
ESCA	15/165		13/184
PAAD	9/91		
Total	670/4509	7/33	616/5696

Table S3.2: The numbers of CpG probes falling into six types of regions.

TSS1500	TSS200	1stExon	Body	5'UTR	3'UTR
56,194	44,572	8,745	148,013	24,912	15,379

Table S3.3: Table for rate functions

$V_{TS}([dUMP],[CH2F])$	$\frac{V_{max,TS}[dUMP][CH2F]}{(K_{m,TS,dUMP}+[dUMP])(K_{m,TS,CH2F}+[CH2F])}$
$V_{DHFR}([DHF],[NADPH])$	$\frac{V_{max,DHFR}[DHF][NADPH]}{(K_{m,DHFR,DHF}+[DHF])(K_{m,DHFR,NADPH}+[NADPH])}$
$V_{SHMT}([SER],[THF],[GLY],[CH2F])$	$\frac{V_{max,SHMT,f}[THF][SER]}{(K_{m,SHMT,THF}+[THF])(K_{m,SHMT,SER}+[SER])}-\frac{V_{max,SHMT,r}[CH2F][GLY]}{(K_{m,SHMT,CH2F}+[CH2F])(K_{m,SHMT,GLY}+[GLY])}$
$V_{FTS}([THF],[HCOOH])$	$\frac{V_{max,FTS}[THF][HCOOH]}{(K_{m,FTS,THF}+[THF])(K_{m,FTS,HCOOH}+[HCOOH])}$
$V_{FTD}([10fTHF])$	$\frac{V_{max,FTD}[10fTHF]}{K_{m,FTD}+[10fTHF]}$
$V_{PGT}([10fTHF],[GAR])$	$\frac{V_{max,PGT}[10fTHF][GAR]}{(K_{m,PGT,10fTHF}+[10fTHF])(K_{m,PGT,GAR}+[GAR])}$
$V_{ART}([10fTHF],[AICAR])$	$\frac{V_{max,ART}[10fTHF][AICAR]}{(K_{m,ART,10fTHF}+[10fTHF])(K_{m,ART,AICAR}+[AICAR])}$
$V_{MTCH}([CHF],[10fTHF])$	$\frac{V_{max,MTCH}[CHF]}{(K_{m,MTCH,CHF}+[CHF])}-\frac{V_{max,MTCH}[10fTHF]}{(K_{m,MTCH,10fTHF}+[10fTHF])}$
$V_{MTD}([CH2F],[CHF])$	$\frac{V_{max,MTD}[CH2F]}{(K_{m,MTD,CH2F}+[CH2F])}-\frac{V_{max,MTD}[CHF]}{(K_{m,MTD,CHF}+[CHF])}$
$V_{NE}([THF],[H_2C=O],CH2F)$	$k_1[THF][H_2C=O]-k_2[CH2F]$
$V_{MTHFR}([CH2F],[NADPH],[SAM],[SAH])$	$\frac{V_{max,MTHFR}[CH2F][NADPH]}{(K_{m,MTHFR,CH2F}+[CH2F])(K_{m,MTHFR,NADPH}+[NADPH])}(\frac{10}{10+[SAM]})$
$V_{MS}([5mTHF],[HCY])$	$\frac{V_{max,MS}[5mTHF][HCY]}{(K_{m,MS,5mTHF}+[5mTHF])(K_{m,MS,HCY}+[HCY])}(\frac{ssH_2O_2+K_i}{H_2O_2+K_i})$
$V_{BHMT}([HCY],[BET],[SAM],[SAH],[H_2O_2$	$e^{-0.0021([SAM]+[SAH])}e^{0.0021*102.6}\frac{V_{max,BHMT}[HCY][BET]}{(K_{m,BHMT,HCY}+[HCY])(K_{m,BHMT,BET}+[BET])}(\frac{ssH_2O_2+K_i}{H_2O_2+K_i})$
$V_{MATI}([MET],[SAM],[GSG])$	$\frac{V_{max,MATI}[MET]}{(K_{m,MATI,MET}+[MET])}(0.23+0.8e^{-0.0026[SAM]})(\frac{K_i+66.71}{K_i+[GSG]})$
$V_{MATIII}([MET],[SAM],[GSG])$	$\frac{V_{max,MATIII}[MET]^{1.21}}{(K_{m,MATIII,MET}+[MET]^{1.21})}(1+\frac{7.2[SAM]^2}{K_a^2+[SAM]^2})(\frac{K_i+66.71}{K_i+[GSG]})$

$V_{GNMT}([SAM],[GLY],[SAH],[5mTHF])$	$\frac{V_{max,GNMT}[SAM][GLY]}{(K_{m,GNMT,SAM}+[SAM])(K_{m,GNMT,GLY}+[GLY])}(\frac{1}{1+\frac{[SAH]}{K_i}})(\frac{4.8}{0.35+[5mTHF]})$
$V_{DNMT}([SAM],[SAH])$	$\frac{V_{max,DNMT}[SAM]}{(K_{m,DNMT,SAM}+[SAM])}$
$V_{SAHH}([SAH],[HCY])$	$\frac{V_{max,SAHH,f}[SAH]}{(K_{m,SAHH,SAH}+[SAH])}-\frac{V_{max,SAHH,r}[HCY]}{(K_{m,SAHH,HCY}+[HCY])}$
$V_{CBS}([HCY],[SAM],[SAH],[SER],[H_2O_2],$	$\frac{V_{max,CBS}[HCY][SER]}{(K_{m,CBS,HCY}+[HCY])(K_{m,CBS,SER}+[SER])}(\frac{1.086([SAM]+[SAH])^2}{30^2+([SAM]+[SAH])^2})(\frac{H_2O_2+K_a}{ssH_2O_2+K_a})$
$V_{CTGL}([CYT])$	$\frac{V_{max,CTGL}[CYT]}{K_{m,CTGL}+[CYT]}$
$V_{GCS}([CYS],[GLU],[GSH],[GLC],[H_2O_2],[$	$\frac{V_{max,GCS}\left([CYS][GLU]-\frac{[GLC]}{K_e}\right)}{(K_{m,GCS,CYS}+[CYS])(K_{m,GCS,GLU}+[GLU])+K_{m,GCS,GLU}[CYS]\frac{[GSH]}{K_i}+\frac{[GLC]}{K_p}+\frac{[GSH]}{K_i}}(\frac{ssH_2O_2+K_a}{H_2O_2+K_a})$
$V_{GS}([GLY],[GLC],[GSH])$	$\frac{V_{max,GS}([GLY][GLC]-\frac{[GSH]}{K_e})}{(K_{m,GS,GLY}+[GLY])(K_{m,GS,GLC}+[GLC])+\frac{[GSH]}{K_p}}$
$V_{GPX}([GSH],[H_2O_2])$	$\frac{V_{max,GPX}[GSH][H_2O_2]}{(K_{m,GPX,GSH}+[GSH])(K_{m,GPX,H_2O_2}+[H_2O_2])}$
$V_{GR}([GSG],[NADPH])$	$\frac{V_{max,GR}[GSG][NADPH]}{(K_{m,GR,GSG}+[GSG])(K_{m,GR,NADPH}+[NADPH])}$
GSH export	$\frac{V_{eHGSH}[GSH]}{K_{eHGSH}+[GSH]}+\frac{V_{eLGSH}[GSH]^3}{K_{eLGSH}+[GSH]^3}$
GSH decay	$d_2[GSG]$
GSSG export	$\frac{V_{eHGSG}[GSG]}{K_{eHGSG}+[GSG]}+\frac{V_{eLGSG}[GSG]}{K_{eLGSG}+[GSG]}$
GSSG decay	$d_1[GSG]$
Methionine exchange with blood	$\frac{V_{bMET}[bMET]}{K_{bMET}+[bMET]}-k_{MET}[MET]$
Cystein exchange with blood	$\frac{V_{bCYS}[CYS]}{K_{bCYS}+[bCYS]}-\frac{0.35[CYS]^2}{200}$

Table S3.4: Table for all the rate parameters.

enzymes	Parameter	metabolites	Model	fold change	BLCA	BRCA	COAD	HNSC	KICH	KIRC	KIRP	LIHC	LUAD	LUSC	PRAD	THCA	UCEC
TS	Km,DUMP	dUMP	6.3	VC1	1.0224	1.0143	1.0053	1.0099	1.0000	1.0226	1.0255	1.0226	1.0220	1.0298	1.0000	1.0267	1.0546
TS	Km,2cf	Ch2-THF	14														
TS	Vmax		5000														
DHFR	Km,dhf	DHF	0.5	VC2	1.0060	1.0014	1.0086	1.0000	0.9749	0.9963	0.9910	0.9944	1.0084	1.0128	1.0032	1.0000	1.0297
DHFR	Km,NADPH	NADPH	4														
DHFR	Vmax		2000														
SHMT	Km,ser	serine	600	VC3	1.0000	0.9936	0.9964	0.9858	0.9102	0.9960	1.0000	0.9860	1.0024	1.0015	0.9960	0.9982	1.0176
SHMT	Km,thf	THF	50														

SHMT	cVmax		5200														
SHMT	Km,gly	glycine	10000														
SHMT	Km,2cf	CH2-THF	3200														
SHMT	cVmax		15000000														
cFTS	Km,thf	THF	3	VC4	1.0109	0.9992	1.0189	1.0082	1.0000	1.0000	1.0059	0.9965	1.0065	1.0088	1.0044	1.0070	1.0183
cFTS	Km,coo	HCOOH	43														
cFTS	Vmax		3900														
FTD	Km,10f	10f-THF	20	VC5	0.9676	0.9255	0.9591	0.9344	1.0000	0.9894	0.9795	0.9886	1.0000	1.0154	0.9962	0.9664	0.9625
FTD	cVmax		500														
PGT	Km,10f	10f-THF	4.9	VC6	1.0055	1.0015	1.0114	1.0028	0.9828	0.9980	1.0000	1.0022	1.0078	1.0097	1.0029	0.9992	1.0137
PGT	Km,GAR	GAR	520														
PGT	Vmax		24300														

AICART	Km,10f	10f-THF	5.9	VC7	1.00 38	1.00 47	1.01 10	1.00 53	0.97 84	0.99 87	1.00 58	1.00 62	1.01 15	1.00 87	1.00 14	1.00 96	1.01 28
AICART	Km,aic	AICAR	100														
AICART	Vmax		55000														
MTCH	Km,1cf	CH-THF	250	VC8	1.00 62	1.00 18	1.01 23	1.00 36	1.00 00	0.99 83	0.99 73	0.98 72	1.00 80	1.01 08	1.00 62	1.00 24	1.01 76
MTCH	cVmax		50000 0														
MTCH	Km,10f	10f-THF	100														
MTCH	Vmax		20000														
MTD	Km,2cf	Ch2-THF	2	VC9	1.00 62	1.00 18	1.01 23	1.00 36	1.00 00	0.99 83	0.99 73	0.98 72	1.00 80	1.01 08	1.00 62	1.00 24	1.01 76
MTD	cVmax		80000														
MTD	Km,1cf	CH-THF	10														
MTD	Vmax		60000 0														

NE	k1,thf,hcho	THF	0.03														
NE		HCHO															
NE	k2,2cf	CH2-THF	22														
MTHFR	Km,2cf	CH2-THF	50	VC11	0.9889	1.0000	0.9903	0.9938	0.9856	1.0029	1.0000	1.0078	0.9941	0.9860	1.0058	1.0040	0.9937
MTHFR	Km,NADPH	NADPH	16														
MTHFR	Vmax		5300														
MS	Km,hcy	HCY	1	VC12	0.9912	0.9979	1.0068	1.0048	0.9946	0.9978	0.9930	1.0060	0.9983	0.9958	0.9954	0.9969	0.9824
MS	Km,5mf	5m-THF	25														
MS	Vmax		500														
BHMT	Km,hcy	HCY	12	VC13	1.0000	1.0000	1.0000	1.0000	0.7471	0.9957	0.9457	0.9729	1.0000	1.0000	1.0000	1.0000	1.0000
BHMT	Km,bet	BET	100														
BHMT	Vmax		2160														

MATI	Km,met	methionine	41	VC14	1.00	1.00	1.00	0.96	1.00	1.00	1.00	0.98	1.00	1.02	1.00	1.00	1.06
					00	00	00	98	00	00	00	50	00	22	00	00	60
MATI	Vmax		260														
MATIII	Km,met	methionine	300	VC15	1.00	1.00	1.00	0.96	1.00	1.00	1.00	0.98	1.00	1.02	1.00	1.00	1.06
					00	00	00	98	00	00	00	50	00	22	00	00	60
MATIII	Vmax		220														
GNMT	Km,sam	SAM	32	VC16	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.96	1.00	1.00	1.00	0.96	1.00
					00	00	00	00	00	00	00	86	00	00	00	56	00
GNMT	Km,gly	glycine	130														
GNMT	Vmax		245														
DNMT	Km,sam	SAM	1.4	VC17	1.01	1.00	1.01	1.01	0.99	1.00	1.00	1.01	1.01	1.01	1.00	1.00	1.02
					58	61	20	10	36	39	83	66	19	57	37	20	97
DNMT	Vmax		180														
SAHH	Km,sah	SAH	6.5	VC18	1.00	1.00	1.01	1.00	0.98	0.99	0.99	1.00	1.00	1.01	1.00	1.00	1.01
					67	41	29	00	52	84	47	00	78	26	20	09	31
SAHH	Vmax		320														
SAHH	Km,hcy	HCY	150														

SAHH	Vmax		4530														
CBS	Km,hcy	HCY	1000	VC19	1.0151	1.0088	0.9845	1.0297	0.9186	0.9917	1.0200	0.9889	1.0074	1.0123	1.0090	0.9968	1.0356
CBS	Km,ser	serine	2000														
CBS	Vmax		700000														
CTGL	Km,cyt	cystathionine	500	VC20	1.0000	0.9952	0.9893	1.0000	0.9611	0.9701	0.9570	0.9797	1.0125	1.0000	1.0025	0.9717	1.0279
CTGL	Vmax		1500														
GCS	Km,cys	cystein	100	VC21	1.0000	0.9967	1.0031	1.0000	1.0000	1.0000	1.0000	0.9954	1.0075	1.0192	0.9963	0.9958	1.0000
GCS	Km,glu	glutamine	1900														
GCS	Vmax		3600														
GS	Km,gly	glycine	300	VC22	1.0069	1.0033	1.0028	1.0026	1.0103	1.0000	1.0130	1.0017	1.0073	1.0103	1.0000	1.0057	1.0171
GS	Km,glc	glut-cys	22														

GS	Vmax		5400														
GPX	Km,gsh	GSH	1330	VC23	1.0000	0.9942	1.0000	0.9963	1.0000	1.0006	1.0000	1.0000	1.0000	1.0052	0.9927	1.0009	1.0064
GPX	Km,H2O2	H2O2	0.09														
GPX	Vmax		4500														
GR	Km,gsg	GSSG	107	VC24	0.9924	1.0036	0.9921	0.9908	1.0075	0.9913	0.9922	0.9967	1.0043	1.0074	0.9956	1.0000	1.0146
GR	Km,NADPH	NADPH	10.4														
GR	Vmax		8925														

Table S3.5: The list of genes encoding all the enzymes and folate and amino acids transporters, used for model estimation.

Catalyzing enzymes	*TS*	*TYMS*
	DHFR	*DHFR*
	SHMT	*SHMT1*
	FTS	*MTHFD1, MTHFD1L*
	FTD	*ALDH1L1*
	PGT	*GART*
	AICART	*ATIC*
	MTCH	*MTHFD1, MTHFD2, MTHFD2L*
	MTD	*MTHFD1, MTHFD2, MTHFD2L*
	MTHFR	*MTHFR*
	MS	*MTR*
	BHMT	*BHMT*
	MAT-I	*MAT1A*
	MAT-III	*MAT1A*
	GNMT	*GNMT*
	DNMT	*DNMT1, DNMT3A, DNMT3B*
	SAHH	*AHCY*
	CBS	*CBS*
	CTGL	*CTH*
	GCL	*GCLC, GCLM*
	GS	*GSS*

	GPX	*GPX1, GPX2, GPX3, GPX4, GPX5, GPX6, GPX7, GPX8*
	GR	*GSR*
Transporters	Serine	*SLC1A4, SLC1A5, SLC7A10, SLC3A2, SLC7A11, SLC6A14, SLC38A2, SLC6A18, SLC6A19, SLC38A4, SLC7A7, SLC3A2, SLC38A2, SLC43A1, SLC6A15, NTT73, SLC6A18*
	Methionine	*SLC7A5, SLC38A1, SLC43A2, SLC7A5, SLC3A2, SLC7A6, SLC3A2, SLC7A7, SLC3A2, SLC38A2, SLC43A1, SLC6A15, NTT73, SLC6A18, SLC38A4*
	Cystein	*SLC1A1,SLC1A4,SLC1A5,SLC7A10,SLC7A11,SLC3A2,SLC38A1,SLC6A14,SLC7A7,SLC3A2,SLC7A8,SLC3A1,SLC7A9,SLC38A2,SLC6A19,SLC6A18,SLC38A4*
	Folate	*FPGS, GGH, SLC19A1*

T**able S3.6:** Table for all the simulation data.

Vmax_TS	nulcotide synthesis rate	DNA methylation rate	5mTHF	5,10-m-THF	5,10-CH=THF	10fTHF	THF	DHF	H	M	SAH	SAM
4500	137.234	157.611	2.299	0.585	0.314	4.064	7.698	0.040	1.248	48.486	19.940	9.856
4700	142.504	157.539	2.288	0.581	0.312	4.054	7.722	0.042	1.250	48.478	19.949	9.819
4900	147.713	157.467	2.276	0.578	0.311	4.045	7.746	0.043	1.251	48.470	19.957	9.783
5000	150.295	157.431	2.271	0.576	0.310	4.041	7.758	0.044	1.251	48.466	19.962	9.766
5100	152.863	157.395	2.265	0.574	0.309	4.037	7.770	0.045	1.252	48.463	19.966	9.748
5300	157.954	157.324	2.254	0.571	0.308	4.028	7.793	0.047	1.253	48.455	19.974	9.713
5500	162.988	157.253	2.243	0.568	0.307	4.019	7.816	0.048	1.255	48.447	19.983	9.678
5700	167.965	157.182	2.232	0.564	0.305	4.010	7.838	0.050	1.256	48.439	19.991	9.644
5900	172.887	157.112	2.221	0.561	0.304	4.001	7.861	0.051	1.257	48.432	19.999	9.610
6100	177.754	157.041	2.210	0.558	0.302	3.993	7.883	0.053	1.258	48.424	20.008	9.576

6300	182.567	156.972	2.200	0.555	0.301	3.984	7.905	0.055	1.259	48.417	20.016	9.543
6500	187.328	156.902	2.189	0.551	0.300	3.976	7.927	0.056	1.260	48.409	20.024	9.510

H2O2	nucleotide synthesis rate	DNA methylation rate	H	M	SAH	SAM	CT	Cys	GC	GSH	GSSG
0.011	149.142	157.093	1.265	47.734	19.386	9.601	36.583	126.698	8.468	1160.400	39.892
0.012	147.991	156.769	1.278	47.041	18.866	9.448	36.849	115.084	8.519	1153.925	44.025
0.013	146.841	156.457	1.289	46.385	18.392	9.304	37.101	105.196	8.559	1126.223	47.578
0.014	145.692	156.155	1.300	45.764	17.958	9.168	37.340	96.748	8.592	1087.565	50.652
0.015	144.545	155.862	1.309	45.175	17.559	9.040	37.567	89.490	8.619	1044.380	53.339
0.016	143.398	155.577	1.318	44.618	17.191	8.918	37.782	83.208	8.641	1000.396	55.712

CHAPTER 4

CHOLESTEROL AND CYP ENZYMES: A POWERFUL COMBINATION FOR DRIVING CELL PROLIFERATION IN METASTATIC CANCER

Introduction

Metastatic cancer is clinically known to grow substantially faster with significantly decreased volume doubling time compared to its primary counterpart [24, 25]; and it responds poorly to the currently available treatment regiments designed predominantly for primary cancers. Unfortunately, very little is known of why metastatic cancers tend to behave this way. The question to be addressed is whether metastatic cancers may have additional drivers in addition to what drives primary cancers possess. Here we propose a model that links accelerated cell proliferation of metastatic (vs corresponding primary) cancers to oxidized cholesterols and further metabolized products.

Links between cholesterol and cancer have previously been reported in the literature, such as: (i) epidemiology studies that found connections between blood cholesterol levels and cancer mortality rates [33]; (ii) studies that observed increased cellular cholesterol levels in a few (primary) cancer types, such as breast cancer [34] and prostate cancer [35]; and (iii) studies that link dysregulation or mutations of cholesterol-metabolism genes to cancer occurrence [36]. A few recent cancer-epidemiology studies have detected correlations between long-term usage of cholesterol-lowering drugs such as statins and reduced cancer-associated mortalities [37]. Mechanistic studies on this relationship only have started to emerge in the past few years. For example, function-losing mutations in *ABCA1*, the main exporter for cholesterol efflux, are found to be associated with

increased cancer occurrences, specifically in colon [38]. One study suggests that the increased membrane-cholesterol level is associated with the activation of the kinase *Akt*, a regulator of apoptosis, and hence increases the chance of cancer cells survival [39].

While published studies have detected links between cholesterol and cancer progression, no model or understanding has been reported regarding how cholesterol contributes to the explosive growth of metastatic cancers (vs the corresponding primary cancers) except that cholesterol is used to make cell membranes, to the best of our knowledge. Our preliminary data have provided strong evidence for a possible causal relationship between the increased (oxidized) cholesterol level and accelerated cell proliferation of metastatic cancers. And this evidence ties very well with the general understanding about the relationship between O_2 and cholesterol.

Evolutionary studies strongly suggest that cholesterol (or sterols in general) has co-emerged with O_2 during the early evolution around 2.5 - 3 billion years ago as a "seal" between phospholipids in cell membranes to prevent the toxic O_2 from entering into anaerobic cells [40]. Recent studies have revealed that (a) membrane cholesterol serves as an O_2 sensor and a possible regulator of O_2-entry into the cells by serving as a membrane barrier against O_2 and reactive oxygen species (ROS) [40]; (b) a higher membrane cholesterol-phospholipid ratio gives rise to lower O_2 permeability of cellular membranes [41]; and (c) the plasma membrane-cholesterol levels are found negatively correlated with the amount of changes in cellular O_2 levels of red blood cells when the blood- O_2 level changes [42]. These studies suggest that while cholesterol has evolved numerous functions in human cells, its original function as an O_2 barrier in cell membrane, as suggested by the evolutionary studies, may have been kept, and possibly play a major role in cell proliferation of metastatic cancers as our study has shown.

Results

The purpose of this study is to gain a general understanding about the distinct driving forces of metastatic cancer in comparison with the corresponding primary cancer. To accomplish this, we have collected 20 sets of microarray-based transcriptomic data of metastatic and corresponding primary cancers from a public database (see Supplementary Table S4.1). These data cover 980 tissue samples of 12 primary-metastatic cancer combinations, namely breast-to-bone metastases; prostate-to-bone metastases; breast-to-brain metastases; breast-to-liver, colon-to-liver, pancreas-to-liver and prostate-to-liver metastases; and bone-to-lung, breast-to-lung, colon-to-lung, kidney-to-lung and pancreas-to-lung metastases. The first question addressed is: *are there genes/gene sets significantly up-regulated in metastatic versus the corresponding primary cancers?* This leads to the identification of dozens of genes (gene sets) involved in cholesterol uptake, synthesis and metabolism towards the production of oxysterols, bile acids and steroid hormones. This initial result led us to ask and address the two main questions of the paper: (i) *what drives metastatic cancers to increase their cholesterol influx*, and (ii) *what consequence does the increased flux of cholesterol have in metastatic cancers*? The following provides our analysis results related to these two questions.

Throughout this paper, we utilized one-sided *t* test for differential gene expression, and Gene Set Enrichment Analysis (GSEA) procedure for gene set enrichment analysis [113]. Considering that most of the collected datasets have relatively small sample sizes (see Supplementary Table S4.1), we utilized a meta-analysis approach [114], which combines the analysis results over all the 20 datasets using Fisher's method, for both differential expression analysis and gene set enrichment analysis, and is adjusted for false discovery rate (FDR) (See Supplementary Methods and Materials).

O_2-Level Difference between Metastatic and Primary Sites: Stresses and Responses

Metastatic cancers tend to have higher O_2 level and stronger oxidative stress response: We have estimated the O_2 level and the level of responses to oxidative stresses in metastatic vs corresponding primary cancers, based on expression data of the relevant marker genes.

Hypoxia is a key characteristic of primary cancers, and it plays a pivotal role in tumorigenesis [115-117]. We noted that *HIF1α* is down-regulated moderately (*p*-value=0.2) in metastatic *versus* primary cancers. In addition, *HIF1AN*, an inhibitor of *HIF1A*, is significantly up-regulated (*p*-value=1.19E-5) in metastatic cancers (in comparison with corresponding primary cancers). We also examined gene sets whose proteins utilize oxygen, namely "biological oxidation", "oxygen and reactive oxygen species metabolic process" and "oxidoreductase activity" in the Msigdb database [113] , and found that all three sets of genes are significantly up-regulated with p-values 0, 0 and 3.01E-5, respectively. These together strongly suggest an increased O_2 level [118] in metastatic cancers vs corresponding primary cancers. The detailed list of all the genes discussed throughout this section is given in Supplementary Table S4.2.

In addition, multiple antioxidant enzymes are up-regulated in metastatic cancers. Specifically, *SOD1-3* (superoxide dismutases) are known to catalyze the conversion of superoxide (O_2-) to hydrogen peroxide (H_2O_2) and O_2, where H_2O_2 can be further converted to water by *CAT* (catalase) and *GPX1-5* (glutathione peroxidases). Glutathione (GSH) is the key thiol-based redox buffer, and glutamate cysteine ligase (encoded by *GCLC* and *GCLM*) is the main (rate-limiting) synthesis enzyme. *SOD1, SOD2, GCLC* and *GPX3* are all significantly up-regulated (with *p*-values = 0.037, 1.7E-3, 0.024, 1.06E-9, respectively), and the detailed statistical significance values of the other genes mentioned here can be found in Supplementary Table S4.2. Based on these observations, we

posit that metastatic cancers have reduced capacities for coping with O_2, due to their extended residence in their primary and hypoxic sites, possibly having lost some of such capacities.

Evidences suggest cellular membrane damages: Two lines of evidence suggest that metastatic cancers generally have damaged cell membranes: (i) genes in response to membrane damages are up-regulated; and (ii) the catabolism of the oxidized products of phospholipids, a key component of cell membrane, is up-regulated.

It is known that O_2 can oxidize membrane cholesterol (and phospholipids) in an oxidative microenvironment through lipid peroxidation, leading to continuous membrane damage and loss of membrane cholesterol [119, 120]. Lipid peroxidation can take place autonomously or can be catalyzed by lipoxygenases such as *ALOX15, ALOX15B, ALOX12* and *ALOX5* [121]. We noted that *ALOX5* is up-regulated in metastatic cancers significantly with p-value 3.68E-4, while the other three genes are not changed significantly. In addition, arachidonic acid and linoleic acid metabolism are both significantly up-regulated with *p*-values 0 and 7.84E-7, respectively, which are known products of oxidation-induced membrane-phospholipid catabolism[122, 123].

PON1 of the paraoxonase family plays a protective role against membrane peroxidation by a joint activity with HDL (high-density lipoprotein) for membrane repair [124]; and *PON2* is known to be able to counteract lipid peroxidation on plasma membrane [125]. The anti-oxidation property of the last member of this family, *PON3*, is usually applied on lipoproteins [125], though. It has been established that increased abundance of 4-hydroxy-2-nonenal (HNE), the most studied lipid-peroxidation product [126], can significantly increase the protein levels of the *SQSTM1*, *HMOX1* and *PRDX1* genes [127]. Here we indeed observe that *PON1, PON2, PON3, SQSTM1, HMOX1* all show significant up-regulation with *p*-values 9.62E-10, 2.13E-4, 1.31E-11, 1.00E-9, 1.07E-7, respectively, strongly suggesting active oxidation activities of plasma membrane in metastatic

cancer sites. The strong correlation between O_2-sensing genes and membrane damage-related genes are observed, and could be found in Supplementary Table S4.3.

Increased expression of cholesterol uptake and/or synthesis genes: The most striking observation made on the 20 datasets is that genes responsible for obtaining cholesterol through either *de novo* biosynthesis or uptake from circulation of cholesterol-carrying lipoprotein particles are significantly up-regulated in metastatic cancers, revealing that metastatic cancer cells have increased needs for cholesterol.

Four types of cholesterol-carrying lipoprotein particles, namely high, low and very low-density lipoproteins (HDL, LDL and VLDL) and chylomicrons have been observed and extensively studied. Each of these particles carries different amounts of cholesterol: 5% in chylomicron, 25% in VLDL, 47% in HDL and 61% in LDL [128], respectively. Cholesterol in HDL and oxidized LDL can be transported into cells from circulation via the scavenger receptor class B1 (*SRB1*) [129]; LDL and chylomicron via low-density lipoprotein receptor (*LDLR*) and low-density lipoprotein receptor-related protein 5 (*LRP5*); and VLDLs via *VLDLR*s. *CD36* can uptake HDLs, (oxidized) LDLs and VLDLs [130]. *SRB1, LDLR, CD36* and *LRP5* are significantly up-regulated with *p*-values 2.04E-13, 5.01E-5, 2.68E-6 and 1.50E-5, respectively. Gene sets related to chylomicron transport and cholesterol biosynthesis are both up-regulated with *p*-values 0 and 0.006, respectively.

SREBP has been established as the main regulator of cholesterol biosynthesis and *LDLR*- and *SRB1*-based cholesterol uptake [131]. The protein has two encoding genes *SREBF-1* and *SREBF-2*. Previous studies have shown that oxidative stress can regulate *SREBP* in human, and O_2 can regulate it in fission yeast[132, 133], although no studies have confirmed that O_2 can directly regulate *SREBP* in human. Another study has demonstrated that toxins-induced membrane damage

can induce the activation of *SREBF*s, as a way to activate membrane biogenesis [134]. Another regulator of the cholesterol-carrying lipoprotein receptors is *PDZK1* (PDZ domain containing 1), whose activation prevents the degradation of *SRB1* [135]. It has been found that *PDZK1* can be regulated by estrogen receptor α, *ESR1* [136] and by *PPARA* (the peroxisome proliferator-activated receptor α) [137], which are both active in metastatic cancers as shown later. *SREBF1* is not significantly changed, but *SREBF2* and *PDZK1* are up-regulated in metastatic cancers with *p*-values 5.66E-3 and 6.85E-7, respectively. Overall, metastatic cancers have up-regulated *SREBP2* and *PDZK1*, possibly due to increased O_2 level and/or oxidative stress, leading to an increased uptake of cholesterol-carrying lipoproteins, as well as biosynthesis of cholesterol as shown earlier. Furthermore, our analyses revealed strong statistical correlation between the aforementioned regulators and cholesterol influx genes. In addition, strong positive correlations are observed between the oxygen-sensing and membrane damage response genes and cholesterol influx regulators, as detailed in Supplementary Table S4.3.

To provide further evidence that there is indeed an increased influx of cholesterol, we noted that numerous genes relevant to cholesterol efflux are up-regulated. Note that excess cellular cholesterol can be either converted to cholesteryl esters by acyl-coenzyme A:cholesterol acyltransferase *SOAT1* and *SOAT2* or removed via cholesterol efflux through ATP-binding cassette (*ABC*) exporters such as *ABCA1* [138]. In addition, bile-acid synthesis represents another key exit for excess cholesterol, in which *ABCB11* and *SLC10A1* serve as two bile-acid exporters [139]. Here *ABCB11*, *SLC10A1*, *ABCA1* are significantly up-regulated with *p*-values 5.58E-11, 8.23E-4, 4.76E-9, respectively, and the bile-acid synthesis and metabolism pathways are both significantly up-regulated (see Supplementary Table S4.2 for *p*-values).

These data strongly suggest that (1) metastatic cancers in general have elevated (steady state) cholesterol levels and increased cholesterol flux (both influx and efflux) in comparison with their primary counterparts; and (2) substantial portions of the arriving cholesterols are not used by the cells. Based on all this information, we posit that membrane damage is a key reason for the continuous influx of new cholesterol via uptake or synthesis, a process possibly regulated by *SREBP* and some by *PDZK1*.

A Powerful Combination of Cholesterol and CYPs for Cell Proliferation: a Side-Effect of Cell Survival

Increased expression of CYP (and other cholesterol metabolic) genes as a defense against increased O_2: In addition to the antioxidants mentioned in the previous sections, *CYP*s represent another large class of enzymes that can consume cellular O_2 by oxidizing cholesterol s[140]. For example, a number of *CYP* genes are known to oxidize cholesterol to a variety of biochemically active oxysterols, including *CYP3A4,5,7* and *CYP27A1* [120], where the class of the *CYP3A* enzymes are known to produce 4β-OHC using cholesterol as substrates [141, 142], and *CYP27A1* to produce 27-OHC [143]. Another class of *CYP* genes are steroidogenic, which are responsible for steroid hormone syntheses [144], such as *CYP11 (CYP11A1, CYP11B1, CYP11B2), CYP17A1, CYP19A1* and *CYP21A2*. The correspondence between *CYP* genes and oxysterols are summarized in Supplementary Table S4.4.

Our data analyses revealed that *CYP3A4,5,7* and *CYP27A1* are all significantly up-regulated with *p*-values = 3.78E-18, 3.66E-13, 7.91E-4, 2.46E-14, respectively. For steroidogenic *CYP*s, *CYP21A2* is significantly up-regulated with *p*-value 3.72E-3, while the others are not significantly changed.

A number of genes are known to be involved in further metabolism of oxysterols towards steroidogenic products that can directly bind with and activate various nuclear receptors (*NRs*) or even growth factor receptors. Specifically, *HSD3B1* (hydroxy-delta-5-steroid dehydrogenase 3β) and *HSD3B2*, both being able to metabolize various oxysterols to progesterone, testosterone and other steroidogenic metabolites, respectively, are up-regulated in metastatic cancers. *HSD17B1,3,7*, in conjunction with *HSD3B1*-2, play key roles in steroidogenesis. *SRD5A1-2* (steroid-5α-reductase) are responsible for converting testosterone into androgen, a more potent growth hormone with higher binding affinity with the androgen receptor, are also up-regulated (See Supplementary Table S4.2 for significance values). The steroid hormone pathway is also significantly up-regulated with *p*-value 0.005. Statistical correlation between oxygen-sensing genes and *CYP*s are outlined in Supplementary Table S4.3.

To verify that cholesterols are indeed brought into the relevant compartments of metastatic cancer cells and metabolized, we have examined a number of genes relevant to intracellular transportation of cholesterols. Earlier discussion has shown that the increased cholesterol influx has activated the bile-acid synthesis pathway in endoplasmic reticulum (ER). Mitochondria represent the main organelle where cholesterol is oxidized to oxysterols and further metabolized to steroid hormones. A prerequisite for these productions is the transport of cholesterol into mitochondria inner membrane. *StAR* (or *STARD1,* steroidogenic acute regulatory protein) is one such transporter. A number of *STARD* genes (*StAR*-related lipid transfer protein in late endosome), close relatives of *StAR* such as *STARD3* and *STARD6,* are also capable of moving cholesterol from the outer to the inner membrane of mitochondria [145, 146]. *STARD3* is significantly up-regulated in metastatic cancers with *p*-value 7.16E-4. In addition, it is known that *STARD1* can be activated by oxidative stress [147]. Hence we postulate that the other *STARD* genes may also be up-regulated and

activated by oxidative stress. And again, the statistical correlations between oxygen sensing related genes and *STARD* genes are significantly positive (see Supplementary Table S4.3), suggesting a possible causal relationship.

Increased expression of nuclear receptors imply their increased activation: Both *CYP* genes and their enzymatic products have close functional relationships with various *NRs*. On one hand, *NRs*, serving as transcription factors, can regulate key processes related to cholesterol homeostasis [148]. For example, it has been shown that cholesterol accumulation can trigger the up-regulation of *CYP27A1* through the regulation by *PPAR* (peroxisome-proliferator-activated receptor) [149], even though the details of how an up-regulated *CYP27A1* leads to the production of 27-OHC are not well understood yet[150]. In addition, *NR5A1* (steroidogenic factor 1) is known to positively regulate the steroidogenic *CYP*s [151]. On the other side, the metabolic products of some *CYP*s can serve as ligands of *NR*s and activate them as transcription factors upon binding. Generally, the activation of an expressed *NR* requires the binding with its (cognate) ligand(s), leading to its dimerization, homo- or hetero-dimers, and then translocation to the nucleus to execute its function as a transcription regulator. Steroid hormones are potent ligands for steroid hormone receptors *PGR*, *ER*, *AR* and *ESRRA* [152]. Some of the hydroxyl-cholesterols, such as 22(R)-OHC and 27-OHC, are known to be able to activate *NR5A1* [153] and *ESR1* [154]. 27-OHC can also bind with and activate *LXR, ESR1* and *PGR* [155, 156]. Bile acids are known to be able to activate *FXR*, whose physiological function is a bile-acid sensor that prevents intracellular over-accumulation of bile acids and stimulates its export [157]. In addition, 4β-OHC, the product of *CYP3A,* can activate *LXR* upon binding [158]. Supplementary Table S4.5 summarizes the known relationships between cholesterol metabolites and *NRs;* and statistical associations between *NRs* and various *CYPs* could be found in Supplementary Table S4.3.

LXRα (liver X receptor, encoded by *NR1H3*), *FXR* (farnesoid X receptor, encoded by *NR1H4*), *ESR1-2* (estrogen receptor 1 and 2), *AR, NR5A1* and *PPARA,D,G* are among the overly-expressed *NR* genes in metastatic cancers, with *p*-values 2.38E-7, 1.16E-6, 2.51E-8, 4.11E-4, 1.85E-11, 1.16E-6, 8.05E-6, 5.81E-4 and 8.52E-3, respectively. While the observed up-regulation of *NR* genes does not directly imply the activation of their protein products, the increased production of oxysterols and steroidogenic products as discussed in the previous Section, in conjunction with our quantitative metabolic profiling on 27-OHC (see Figure 4.1), offer a strong indirect evidence.

Increased expression of growth factor receptors and cell cycle genes imply increased proliferation rates: The regulatory relationships between *NR*s (and some cholesterol metabolites) and growth factor receptors are presented in Table 4.1, and the statistical correlations between the *NR*s and their regulated *GF*s/*GFR*s, as well as those between the growth factors and their receptors, are described in Supplementary Table S4.3.

Activated growth factor/receptors such as *EGFR, HER2, ERBB3* can directly trigger cell growth [159]. Other growth factor receptors such as *FGFR1* and *FGFR3* can enhance cell proliferation [160-162]. *TERT* (telomerase) is known to enable cells' immortality [163]; and *MET* has been recognized as a proto-oncogene [164]. We have examined the expression levels changes of these genes along with cell cycle genes, and found them to be all up-regulated in metastatic cancers with the detailed *p*-values given in Supplementary Table S4.2. In addition, the cell cycle pathway is significantly up-regulated with *p*-value 0. All these suggest accelerated cell proliferation in metastatic vs corresponding primary cancers in general.

A Driver Model for Accelerated Cell Proliferation

We have developed a model for accelerated cell proliferation in metastatic vs corresponding primary cancers, based on observations presented above, along with some experimental validation

in support of the model. A key assumption here is that migrated cancer cells have reduced capacities for coping with O_2, whose levels are higher in the metastatic vs the corresponding primary sites. Hence their migrations to the new locations led to the oxidized and therefore damaged plasma membranes of the metastatic cells. In response to the damaged cell membranes due to lipid peroxidation as well as to increased cellular O_2 levels, the affected cells increase their influx of cholesterol needed for membrane repair and up-regulate their antioxidant enzymes, particularly *CYP*s, for consumption of O_2; some of the arriving cholesterol is moved to mitochondria (or ER) and oxidized to oxysterols (or bile acid) either by *CYP* enzymes or through auto-oxidation, and some further metabolized into steroidogenic products; the resulted oxysterols, bile acids and steroidogenic products will activate a variety of *NRs* as transcription factors, which further induce a series of growth-related activities, including the activation of growth factor receptors, *TERT*, and increased rates of cell cycle, hence cell proliferation. 27-OHC and bile acids are even able to activate growth factor/receptors directly. This process continues as long as lipid peroxidation and cholesterol influx (and efflux) continue at a certain level throughout the entire development of a metastatic cancer.

Figure 4.2 shows our driver model for accelerated cell proliferation of metastatic cancers, comprised of 11 steps, each supported by known functional relationships among genes at the two ends of each link, their correlated expression patterns and/or published findings. The correlation coefficients between observations at the two ends of each link in Figure 4.2 are calculated and detailed in Figure 4.3 and Supplementary Table S4.3. Specifically, Figure 4.3 shows the correlation coefficients between each pair of functionally linked gene sets in metastatic *versus* primary cancer tissues in each of the examined dataset. It is clear that each predicted functional link in our model fits the metastatic cancer data better than the primary cancers, hence providing an indication of the

overall quality of our model. The detailed correlation data, generated using a meta-analysis over all the 20 datasets of the paired gene sets, are given in Table S4.3.

One key remaining issue about the driver model is: *can we rule out the possibility that the increased cholesterol uptake or synthesis is induced to support of the increased need for making more cell membranes by the accelerated cell proliferation of the metastatic cancers*? To answer this question, we have calculated the statistical correlations between cholesterol uptake/synthesis genes and cell-proliferation genes, and compared them with those between the cholesterol uptake/synthesis genes and genes responsive to membrane damages. The rationale is that both cell proliferation and membrane damage can lead to the increase in cholesterol influx, and a comparison between the two sets of statistical correlations could provide information about which of these two causes are predominantly responsible for the observed increase in cholesterol influx. We have compared the correlations between three groups of genes in both primary and metastatic cancer samples: genes involved in cholesterol synthesis and uptake (CL), genes involved in membrane damage response (MDR), and DNA polymerases and checkpoint regulators involved in cell cycle [165] (CYCLE) (see Supplementary Methods and Materials for detailed gene lists), and found that the correlations between CL and MD is significantly higher than the correlations between CL and CYCLE with p-value =0.04 in metastatic cancers while the correlations between CL and MD is lower than the correlations between CL and CYCLE with p-value 0.13 in primary cancers. Here, the significance for a derived correlation between the expression data of two gene sets is calculated as the significance of the Spearman correlation between the first principle components of the two gene expression submatrices containing the relevant genes [166]. To compare the significance values for CL *vs* MDR and values for CL *vs* CYCLE, a non-parametric one-sided Wilcox test is conducted (See Supplementary Methods and Materials). Our analyses

revealed that while increased cholesterol influx in primary cancers (*versus* corresponding normal controls) is largely due to cell proliferation, the increase in metastatic cancers is predominantly due to continuous cell-membrane damages, which represents a fundamental difference between primary and metastatic cancers.

To support our model, we have carried out limited experimental validation as given in the following.

Increased cell proliferation when treated with HDL cholesterol: SW620 [167], a metastatic cancer cell is used to examine if HDL cholesterol may have any effect on the growth of a metastatic cancer. The cell population, cultured in cholesterol-containing HDL medium, was split equally into two halves along with the evenly split medium and one of culture contains cells with *SRB1* siRNA treatment (see Supplementary Methods and Materials for details of siRNA transfection). Figure 4.4 shows the growth curves of the two cell populations, revealing approximately 43% reduction in growth rate by the cells with *SRB1* interference RNAs*,* hence providing supporting evidence that HDL cholesterol, up-taken via *SBR1*, plays a key role in the accelerated growth of metastatic cancer.

Metabolic profiling shows increased oxysterols in metastatic versus primary cancers: Knowing that *CYP27A1* is up-regulated in metastatic cancers, we have examined through metabolomic analyses if its product, 27-OHC, and indeed shows increased abundances in metastatic *versus* primary cancer tissue samples. Specifically, we have measured the quantities of 27-OHC in primary colon and gastric cancer *versus* their matching normal tissues, liver metastases along with primary liver cancer and matching normal tissues, respectively. Figure 4.1 shows the measured quantities of 27-OHC in three pooled samples: five primary colon *versus* matching normal colon tissues, five primary gastric *versus* matching normal gastric tissues, and five

metastatic liver cancers (from colon and breast cancers) along with five primary liver *versus* five matching normal liver tissues (see Supplementary Table S4.6). IRB approval is obtained regarding the use of human subjects. Clearly, the quantity of 27-OHC is about one order of magnitude higher in metastatic cancer tissues than those in primary and normal tissues. The observed changes in oxidized cholesterol metabolites are highly consistent with the observed expression changes of the relevant genes, hence providing a strong supporting evidence for the validity of our gene-expression data-based model prediction.

Concluding Remarks

It has been established that higher membrane cholesterol-phospholipid ratios give rise to lower O_2 permeability of the (plasma) cellular membrane [41]. In addition, it has been shown that high membrane cholesterol content can result in deficiency of oxidative phosphorylation in mitochondria [168]. These strongly indicate that membrane cholesterol serves as a defense against O_2 entry into the cells, which was proposed to be the original function of cholesterol (or sterol in general) when it first emerged some two billion years ago by an evolutionary biology study [40]. Based on this information, we speculate that the membrane cholesterol level of human cells is O_2-level dependent in cancer, as suggested previously for normal cells [41]. We infer that primary cancer cells may have lost some of their membrane cholesterols and maintain lower membrane cholesterol contents since these cells have been in hypoxic environments for an extended time before they metastasize to new locations. After having been in such environments for long, their cellular metabolisms may have partially evolved to become less O_2-dependent and even possibly anaerobic. When moving to a blood-rich and hence O_2-rich location, these cells may need to quickly increase their membrane cholesterol level as well as their antioxidant defenses against the increased O_2. This may represent the initial driver for increased need for cholesterol, which is

consistent with established knowledge that oxidative stress can lead to accumulation of cholesterol [132]. Then the cell membrane of metastatic cancers goes through continuous lipid peroxidation, membrane damage and loss of cholesterol. This may be the key reason for the continuous need for additional cholesterol. We hypothesize that *SREBP* may be the initial trigger of cholesterol influx in response to O_2-level increase in human.

We have, for the first time, proposed a driver model for the accelerated cell proliferation of metastatic cancers compared to the corresponding primary cancers, in which oxidized cholesterol has a key role. Ultimately it is the increased influx of cholesterol, largely induced by the increased O_2 level and associated membrane damage that accelerate the growth of such cancer cells. This model is based on 20 sets of transcriptomic data covering 12 types of 980 cancer tissues and validated experimentally on its key steps. Based on the diversity of the metastatic cancer types and the large sample size studied here, we suggest that this model is applicable to metastatic cancer in general. The new insights gained and information derived here not only offer fundamentally novel understanding about metastatic cancers but could also lead to new directions in terms of developing new and more effective drugs for intervention of metastatic cancers and slowing down their explosive growth.

Methods and materials

Cell lines and culture: SW620 is lymph node metastasis of colon cancer and is in stock in Dr. Yuan Yuan's lab. The SW620 cells were cultivated at 37℃ in an atmosphere of 95% air and 5% CO2 with RPMI-1640 medium (Gibco) supplemented with 10% fetal bovine serum, 100 units/ml penicillin, 100 mg/ml streptomycin, and 20 mM L-glutamine.

Antibodies: All antibodies used in this study, including anti-*SRB1*, anti-*EGFR* and anti-*p-EGFR,* are purchased from Abcam Inc. (Cambridge, MA, USA).

In vitro cell growth assay: The SW620 cells were prepared at a concentration of 1 × 104 cells/uL. Aliquots (100 uL) were dispensed into 96-well plates. They were incubated for 12, 24, 36 or 48 hours, and the 3-(4,5-dimethylthiazol-2-yl)-2,5-diphenyltetrazolium bromide (MTT) assay was performed by adding 20 ml of MTT (5 mg/ml; Promega) for 4 hours. Then supernatants were removed. A total of 150 ml of dimethylsulfoxide (Sigma, St Louis, Missouri, USA) was added to each well. Fifteen minutes later, the absorbance value (optical density (OD)) of each well was measured at 490 nm with a microplate reader. All experiments were repeated three times.

Small interfering RNA transfections: The SW620 cells were seeded at a density of 3×105 cells/well in 6-well plates. After 24 hours, cells were transfected with standard small interfering RNA (siRNA) with 3'-overhangs using Lipofectamine 2000 reagent according to the manufacturer's protocols. The *SRB1* siRNA sequence used was: CCA UGA CCC UGA AGC UCA U. The control sequence used was: AAT TCT CCG AAC GTG TCA CGT. *SRB1* and control siRNAs were produced from Genechem Co. (Shanghai, China). Gene silencing effect was verified by Immunoblotting (see Figure S4.1).

Tissue collection and storage: All tissues used in this study were put into a liquid nitrogen tank within 30 minutes of resection from the patients, and were then put into a freezer at -80°C for long term storage. Informed consent on sample collection and use in the study was obtained from patients.

Quantitative metabolic profiling of 27-OHC on human cancer tissues: The tissue samples were weighed and each sample was placed in a 10mL EP tube. Hydrolysis was performed by adding 1mL of 1M KOH in ethanol to each tube, followed by ultrasound 1min and placement in a water bath at 60°C for 40min. Following hydrolysis 1mL ultrapure water and 3mL hexane were added to each sample, the tubes were capped and the samples were vortexed for 1min and shaken for

10min. Samples were centrifuged at 3,500 rpm for 5 min at room temperature. The supernatant was transferred to a 10mL glass tube and set aside. Each sample was extracted with 3 mL hexane again. The supernatant was transferred to the 10mL glass tube with the initial sample, and then dried under nitrogen. Prior to GC-MS analysis, 27-hydroxycholesterol was converted to TMS ether. The residue was dissolved with 200 μL 99:1 TMCS:BSTFA and derived at 60 ℃ for 30min. GC-MS was performed using a Shimadzu GCMS-QP2010 instrument equipped with an HP-5ms column (30 m × 0.25 mm inner diameter, 0.25μm phase thickness; Agilent J&W). The gas chromatography program was 180 °C for 1 min, followed by a temperature gradient of 20 °C/min to 290 °C and a final elution at 290 °C for 13.5 min. Helium was used as the carrier gas at a constant flow rate of 1 mL/min.27-hydroxycholesterol was monitored in selected ion monitoring (SIM) mode and ion used for quantitation was m/z 456 [169]. The linearity was within the concentration range of 100-30000ng/mL.

Gene expression data: All transcriptomic data used in this study was collected from the GEO database [170]. Initially, a total of 37 sets of genome-scale transcriptomic data of metastatic cancer and corresponding primary cancer tissues were retrieved. The following criteria were then applied to select quality datasets for our analysis: (1) each dataset must consist of at least five samples of metastatic cancers and five samples of primary cancers to ensure statistical significance of our analyses; (2) each dataset must have an associated publication in a scientific journal; and (3) the dataset must include specific information of which organ the metastases are located. After applying these rules, 20 datasets are retained and used in our analyses (see Supplementary Table S4.1). It is worth-noting that data normalization has been done by the GEO Dataset contributors and the normalizations procedures by each contributor might be different, since these datasets come from various platforms. Platforms, normalization methods and references for all the datasets can be

found in Supplementary Table S4.1. After retrieving these datasets, log2 transformations are ensured to stand for the normality assumption of gene expression data distributions. Additional 11 datasets with only metastatic cancer samples are also retrieved from the same source, considering the relatively small sizes of the existing metastatic cancer sample sets used above. These data are not used in differential gene expression analysis or GSEA, and used solely to compare the correlations of cholesterol influx genes with cell cycle genes as well as with membrane-damage response genes.

Western Blot. After cell treatments, cells were extracted and protein was quantified as described previously [171]. Aliquots (50 μg) of each lysate were separated by electrophoresis on SDS-PAGE gels and transferred to nitrocellulose membranes. Membranes were blocked with 5% non-fat milk in TBST (10 mM Tris-HCl pH 7.4, 100 mM NaCl, 0.5% Tween-20) for 2 h at room temperature and incubated overnight at 4 °C in 5% non-fat milk in TBST containing primary antibodies: rabbit anti-scavenger receptor type B-1 (anti-SRB1; Abcam), rabbit anti-EGFR (anti-EGFR; Cell Signaling), rabbit anti-phospho-EGFR (anti–p-EGFR Tyr1068; Cell Signaling), and rabbit anti-β-actin (sc-1616-R; Santa Cruz). After the appropriate secondary antibodies were added for 30 min at room temperature, the proteins were detected with enhanced chemiluminescence reagent (SuperSignal Western Pico Chemiluminescent Substrate; Pierce, USA) and visualized with the Electrophoresis Gel Imaging Analysis System (DNR Bio-Imaging Systems, Israel).

Differential gene-expression analysis: One-sided *t* test was used to test the hypothesis: if a gene expression is up-regulated ($p1$) or down-regulated ($p2$) between metastatic *versus* the corresponding primary cancer samples in each dataset. Then, for each gene, Fisher's method [114] is applied to combine, respectively, the $p1$ and $p2$ values. For cases where a gene has multiple probes, the probe with the highest fold-change is used in our analysis. To control the false

discovery rate, the Holm method [172] is used. All the statistical analyses are done using R. A total of 2,122 genes are found to be significantly up-regulated and 2,409 genes down-regulated in metastatic cancers (vs corresponding primary cancers) using 0.01 as significance cut-off, among all 19,983 human genes we examined.

Gene set enrichment analysis: The GSEA approach is utilized for gene set enrichment analysis, and the *p*-values for positive and negative enrichments are combined using Fisher's method, and *p*-values are all adjusted for FDR. In total, 2,963 gene sets are retrieved from the Molecular Signatures Database under the c2, c5 and c6 collection [173].

Co-expression analysis: The correlations shown in Figure 4.3 are calculated for pairs of gene groups as follows: for the two groups of genes at the two ends of each functional link in Figure 4.2, only the significantly up-regulated genes (significance threshold = 0.05) in metastatic samples compared with primary samples are considered. Then the correlation matrices for the two subsets are calculated in primary and metastatic cancer samples respectively, and the highest correlations in the matrices are deemed as the correlations between the two set of genes in primary and metastatic cancer samples correspondingly. If the sizes of metastatic cancer samples and primary cancer samples are different for a dataset, we will sample the larger set so the same number of samples will be randomly selected from both sets. To avoid accidental biases, we will conduct the sampling 500 times for each uneven sample sets and use the median of the 500 correlation coefficients as the final correlation between the two gene sets in the larger set. The correlation between two genes' expression patterns is calculated as follows. Pearson correlation coefficient (PCC) is obtained for each gene pair, and Fisher's z transformation of the PCC is standardized with respect to the standard deviation (divided by $\frac{1}{\sqrt{N-3}}$, with N being the sample size), which gives rise to an approximate normal distribution. Then these distributions are added

over the 20 datasets under consideration. A pair of genes is deemed to be co-expressed over the 20 datasets if the aforementioned sum is greater than 0 tested using one-sided test. This process is repeated for each gene pair, and the final *p*-values for all the tests are adjusted for multiple testing.

Compare correlations across gene groups: In each of the 20 datasets, we compared the correlations between three groups of genes in both primary and metastatic cancer samples: genes involved in cholesterol synthesis and uptake (all genes in gene set "REACTOME_CHOLESTEROL_BIOSYNTHESIS" plus *SCARB1, LDLR, VLDLR, LRP5, CD36*), denoted by CL; genes involved in membrane damage response (*HIF1A, HIF1AN, HMOX1, ALOX15, ALOX15B, ALOX12, ALOX5, PON1, PON2, SQSTM1*), denoted by MDR; and DNA polymerases and checkpoints involved in cell cycle (*POLA1, POLA2, POLB, POLD1, POLD2, POLD3, POLD4, POLE2, POLE3, POLE4, POLG, POLG2, POLI, POLK, POLL, POLM, POLN, POLQ, POLE, CCNA2, CCND1, CCND2, CCNE1, CCNE2, CCNB1, CDK1, CDK4, CDK3*), denote by CYCLE. We have tested the hypothesis: "the correlation between the first principle components of the expression matrices of the two gene groups is not 0", on the primary samples and metastatic samples in each dataset, respectively. We also used datasets containing only metastatic cancer samples, considering that sample sizes of the metastatic cancer datasets are relatively small. Then the resulting *p*-values of the correlations between CL and MD and those between CL and CYCLE are compared to assess which correlation is more significant.

Figures

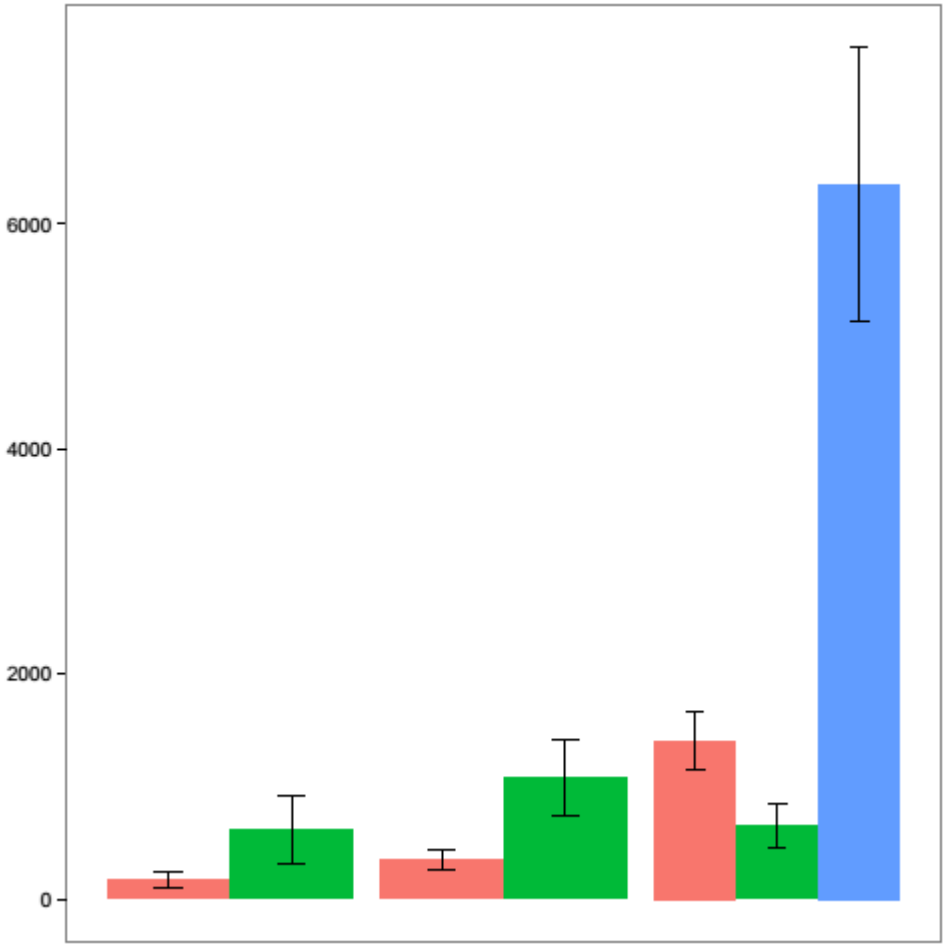


Figure 4.1: The abundances of 27-OHC in, from left to right, normal and primary colon cancer tissues; normal and primary gastric cancer tissues; and normal, primary liver cancer and liver metastases tissues.

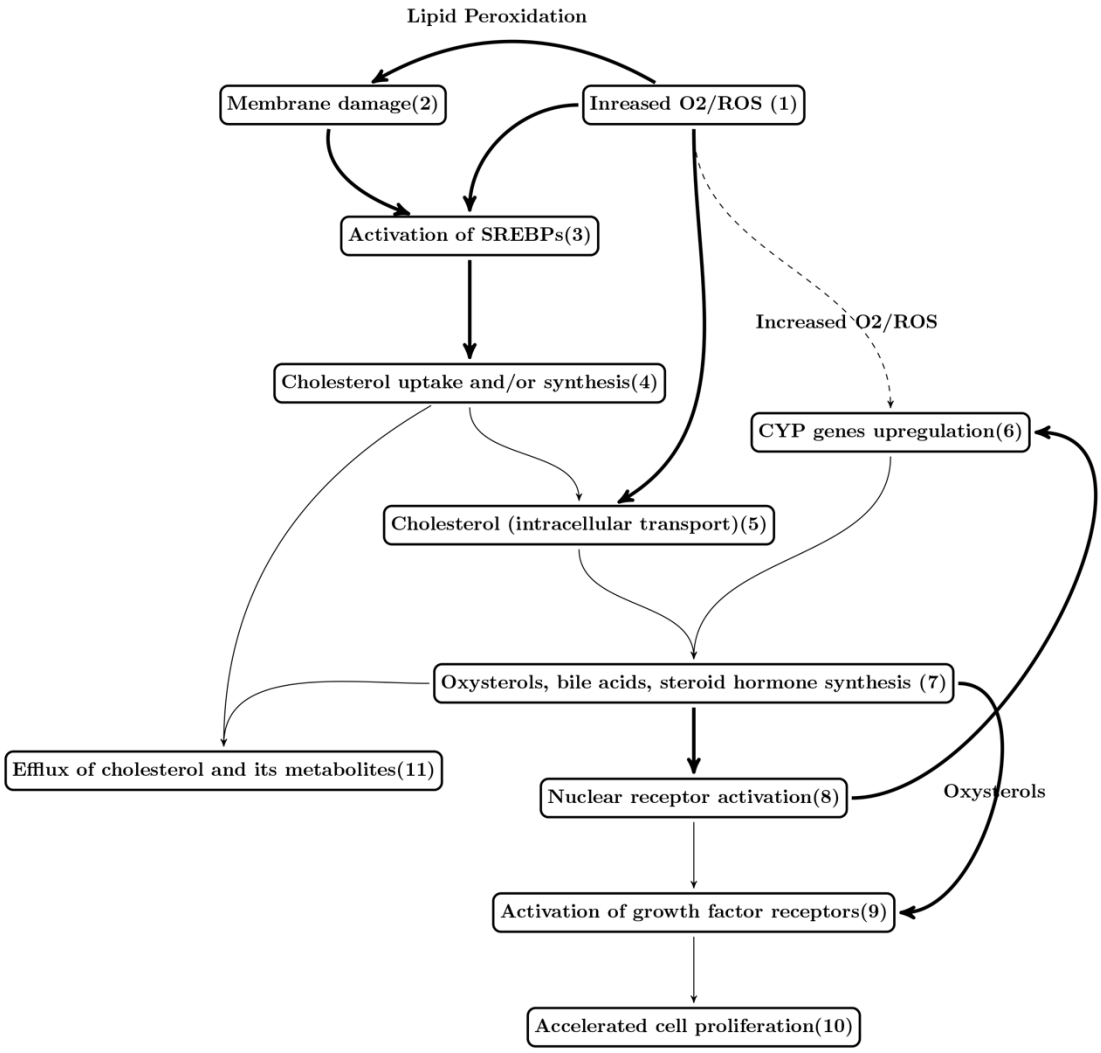


Figure 4.2: An oxidized cholesterol-based driver model for the accelerated cell proliferation of metastatic cancers. The numbers in the parentheses serve as step labels in the diagram. The arrows in bold are supported by both literature and correlation analysis; the thin arrows are supported by literature; and the dashed arrows are only supported by correlation analysis.

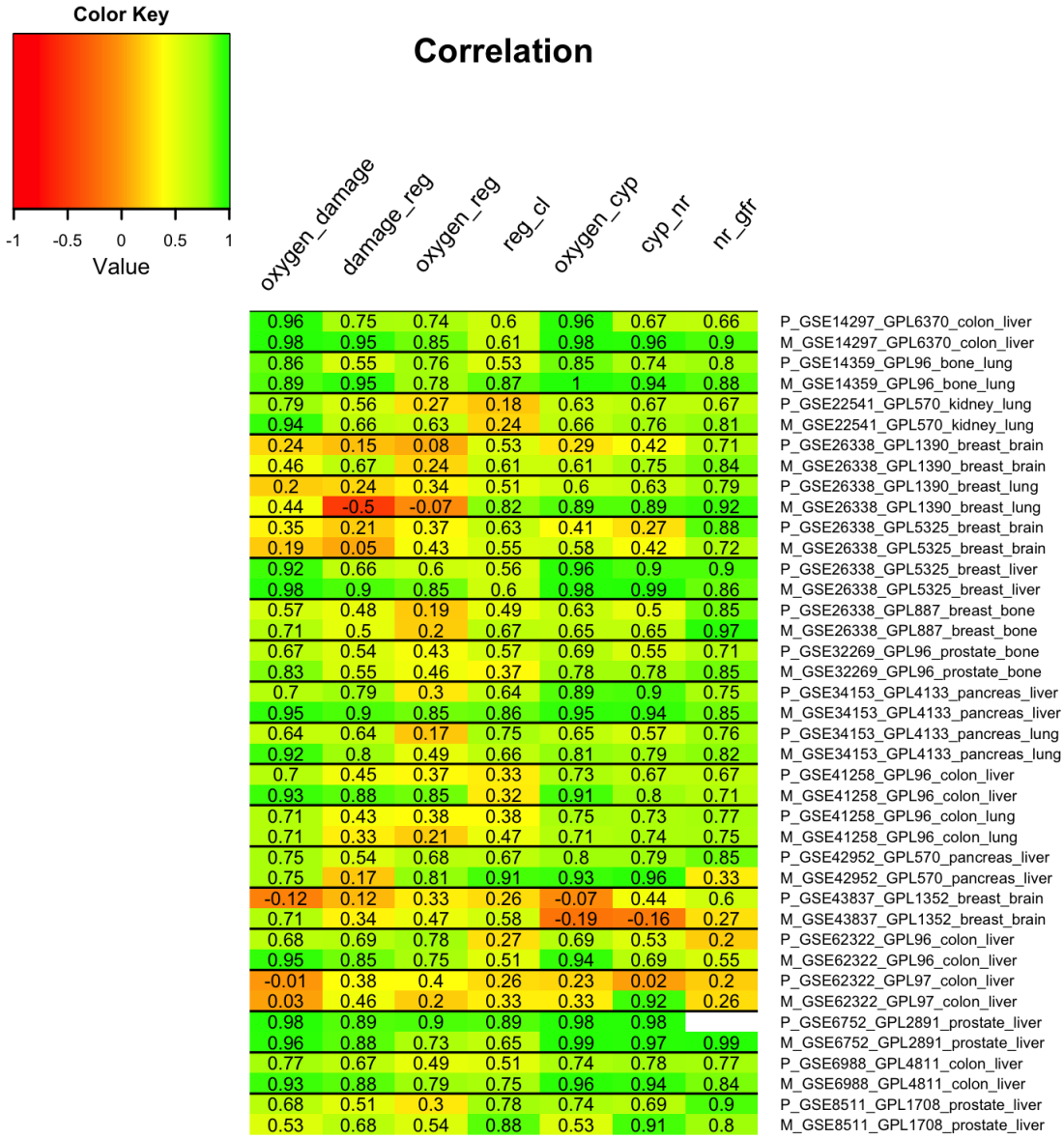


Figure 4.3: A heat-map of correlation coefficients for all functional links in Figure 4.2 calculated for both primary and metastatic cancer samples in 20 datasets. The y-axis of the figure represents the 20 pairs of primary and metastatic cancer datasets, with each pair represented by two consecutive rows, the first for primary and the second for corresponding metastatic cancers. The eight columns from left to right are correlation coefficients for links (1,2), (2,3), (1,3), (3,4), (1,6), (7,8) and (8,9), respectively.

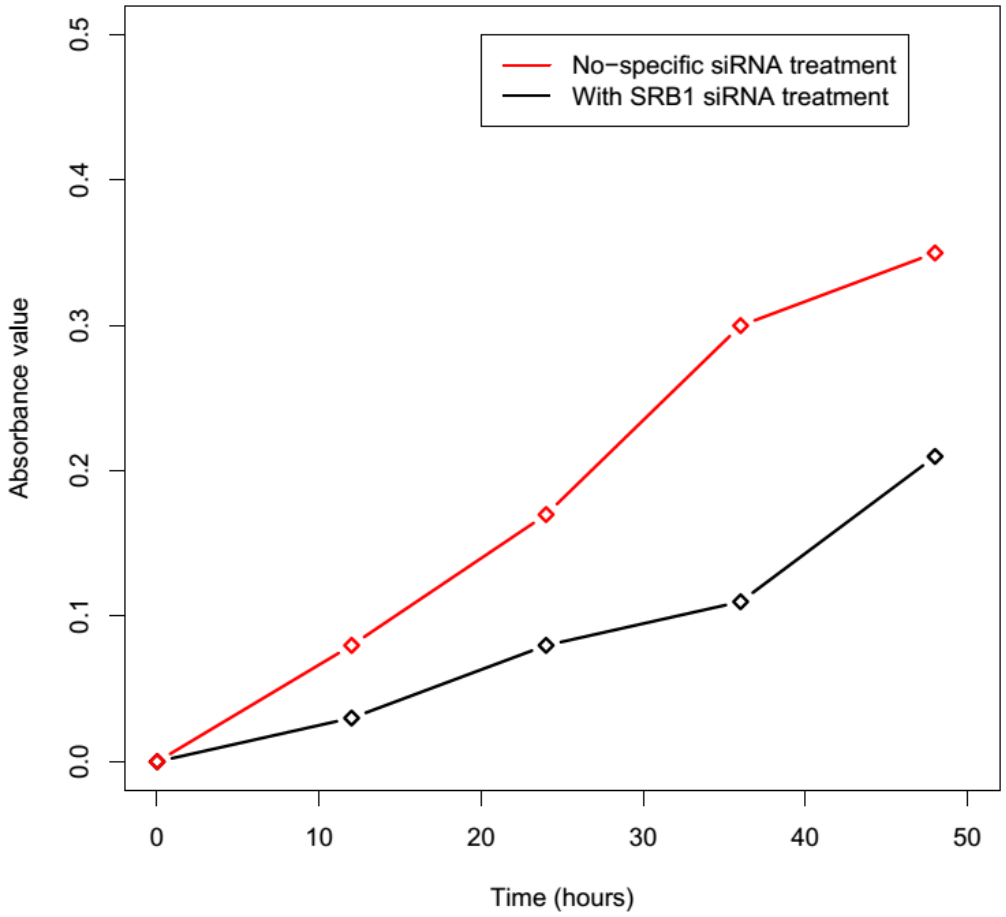


Figure 4.4: Proliferation rates of SW620 cells with *SRB1* siRNA (cube line) *versus* non-specific siRNA (diamond line) measured at 1, 12, 24, 36 and 48 hour after the treatment.

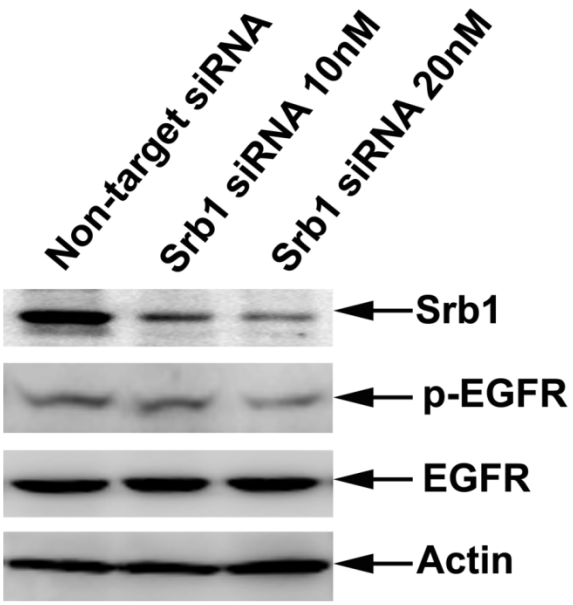


Figure S4.1: *SRB1,* phosphorylated *EGFR, EGFR* and *Actin* protein abundance comparisons, from left to right, under condition ranging from non-targeted siRNA, *SRB1* siRNA 10nM and 20nM.

Tables

Table 4.1: Known relationships between *NRs*/cholesterol metabolites and growth (related) factors/receptors or pathways.

NR/cholesterol metabolites	**Growth-related proteins activated**	**References**
LXR	*FGF19*, *FGFR1-4*	[161, 174-176]
FXR	*FGF19, FGF21, FGFR1-4*	[161, 174-178]

ER	*EGFR*, *HGF*, *MET*, *TERT* (telomerase), cell cycle	[164, 179-187]
AR	*EGFR, ERBB2, ERBB3, TERT,* cell cycle	[179-182, 185, 186, 188]
PGR	*EGFR*, *ERBB2*, *ERBB3*	[189]
ESRRA	*ERBB2*	[190]
27-OHC	*ERBB2*	[191]
bile acids	*EGFR*	[192]

Table S4.1: Detailed information about the datasets used in our analysis and model building. From 21-31, datasets are used only for comparing the correlation of cholesterol influx with cell proliferation and with membrane damage response genes, since they contain only metastatic site samples and they have relatively large sample sizes.

ID	Dataset	Metastatic cancer type	# Primary cancer samples	# Metastatic cancer samples	Platform technology	Normalization method	Reference
1	GSE26338/GPL887	breast->bone	35	5	Agilent Technologies DNA microarrays	Lowess	[193]
2	GSE32269	prostate -> bone	22	29	Affymetrix Human Genome U133A Array	MAS5	[194]

3	GSE26338/GPL 1390	breast -> brain	201	8	Agilent Technologies DNA microarrays	Lowess	[193]
4	GSE26338/GPL 5325	breast -> brain	19	9	Agilent Technologies DNA microarrays	Lowess	[193]
5	GSE43837	breast->brain	19	19	Affymetrix Human X3P Array	MAS5	[195]
6	GSE26338/GPL 5325	breast -> liver	19	5	Agilent Technologies	Lowess	[193]

					DNA microarrays		
7	GSE14297	colon -> liver	18	18	Illumina Sentrix-6 V2 BeadChips	Variance stabilization and spline normalization	[196]
8	GSE41258	colon -> liver	186	47	Affymetrix Human Genome U133A Array	MAS 5	[197]
9	GSE62322/GPL96	colon -> liver	20	19	Affymetrix Human Genome U133A Array	MAS 5	[198]

10	GSE62322/GPL97	colon -> liver	20	19	Affymetrix Human Genome U133B Array	MAS 5	[198]
11	GSE6988	colon -> liver	53	29	Human 17K cDNA-GeneTrack	GenePix 4.1 software	[199]
12	GSE34153	pancreas->liver	14	20	Agilent-014850 Whole Human Genome Microarray 4x44K G4112F	Lowess normalization	[200]
13	GSE42952	pancreas -> liver	12	7	Affymetrix Human Genome	RMA	[201]

					U133 Plus 2.0 Array		
14	GSE6752	prostate -> liver	10	5	GE Healthcare/Amersham Biosciences CodeLink™ UniSet Human 20K I Bioarray	CodeLink	[202]
15	GSE8511	prostate -> liver	12	6	Agilent-012391 Whole Human Genome Oligo Microarray G4112A	linear-lowess	[203]

16	GSE14359	bone->lung	10	8	[HG-U133A] Affymetrix Human Genome U133A Array	MAS5	[204]
17	GSE26338/GPL1390	breast -> lung	201	6	Agilent Technologies DNA microarrays	Lowess	[193]
18	GSE41258	colon -> lung	186	20	Affymetrix Human Genome U133A Array	MAS 5	[197]
19	GSE22541	kidney->lung	24	24	Affymetrix Human Genome	RMA	[205]

					U133 Plus 2.0 Array		
20	GSE34153	pancreas->lung	14	8	Agilent-014850 Whole Human Genome Microarray 4x44K G4112F	Lowess	[200]
21	GSE14017	breast->bone	0	10	Affymetrix Human Genome U133 Plus 2.0	RMA	[206]
22	GSE14018	breast->bone	0	10	Affymetrix Human Genome U133 Plus 2.0	RMA	[206]

23	GSE46141	breast->bone	0	5	Rosetta/Merck Human RSTA Custom Affymetrix 2.0 microarray	RMA	[207]
24	GSE56493	breast->bone	0	5	Rosetta/Merck Human RSTA Custom Affymetrix 2.0 microarray	RMA	[208]
25	GSE14017	breast->brain	0	15	Affymetrix Human Genome U133 Plus 2.0	RMA	[206]

26	GSE14018	breast->brain	0	15	Affymetrix Human Genome U133 Plus 2.0	RMA	[206]
27	GSE14018	breast->liver	0	5	Affymetrix Human Genome U133 Plus 2.0	RMA	[206]
28	GSE14018	breast->lung	0	16	Affymetrix Human Genome U133A Array	RMA	[206]
29	GSE20565	breast->ovary	0	35	Affymetrix Human Genome	GCRMA	[209]

					U133 Plus 2.0 Array		
30	GSE46141	breast->liver	0	16	Rosetta/Merck Human RSTA Custom Affymetrix 2.0 microarray	RMA	[207]
31	GSE56493	breast->liver	0	27	Rosetta/Merck Human RSTA Custom Affymetrix 2.0 microarray	RMA	[208]

Table S4.2: Significance values for gene-expression changes and gene set enrichment analysis discussed in the main text. The first column is categories of the genes and gene sets, and the second column is gene symbols or gene set names from Msigdb [113]. The third column is the (meta) *p*-value for testing the hypotheses "Gene expression is increased in metastatic *vs* primary site or gene set is positively enriched", and the fourth column is the (meta) *p*-value for testing the hypotheses "Gene expression is decreased in metastatic vs primary site or gene set is negatively enriched". All the *p*-values are adjusted for FDR using the Holm method [172] (see Supplementary Methods and Materials). CL: cholesterol.

	Gene Symbol/Gene set name	**p-val(up)**	**pval(down)**
Oxygen and ROS	*HIF1A*	1.00E+00	2.03E-01
	HIF1AN	1.19E-05	1.00E+00
	REACTOME_BIOLOGICAL_OXIDATIONS	0.00E+00	9.66E-01
	OXYGEN_AND_REACTIVE_OXYGEN_SPECIES_METABOLIC_PROCESS	0.00E+00	8.89E-01
	OXIDOREDUCTASE_ACTIVITY_GO_0016705	3.01E-05	9.79E-01
	SOD1	3.73E-02	1.00E+00
	SOD2	1.71E-03	1.35E-01
	SOD3	1.00E+00	2.75E-02
	CAT	8.46E-02	6.56E-01

	GPX1	1.00E+00	1.00E+00
	GPX2	1.00E+00	1.00E+00
	GPX3	1.06E-09	1.00E+00
	GPX4	1.00E+00	1.00E+00
	GPX5	1.00E+00	1.00E+00
	GCLC	2.41E-02	1.00E+00
	GCLM	1.00E+00	1.00E+00
Membrane damage response	*ALOX15*	1.00E+00	1.00E+00
	ALOX15B	1.00E+00	1.00E+00
	ALOX12	1.00E+00	1.00E+00
	ALOX5	3.69E-04	1.00E+00
	KEGG_ARACHIDONIC_ACID_METABOLISM	0.00E+00	9.95E-01
	KEGG_LINOLEIC_ACID_METABOLISM	7.84E-07	1.00E+00
	PON1	9.62E-10	1.00E+00
	PON2	2.13E-04	1.00E+00
	PON3	1.31E-11	1.00E+00
	SQSTM1	1.00E-09	1.00E+00
	HMOX1	1.07E-07	1.00E+00
CL influx	*SCARB1*	2.04E-13	1.00E+00
	LDLR	5.01E-05	2.00E-05
	LRP5	1.50E-05	1.00E+00

	CD36	2.68E-06	3.23E-03
	REACTOME_CHOLESTEROL_BIOSYNTHESIS	6.28E-03	9.73E-01
	REACTOME_CHYLOMICRON_MEDIATED_LIPID_TRANSPORT	0.00E+00	1.00E+00
	HMGCR	1.00E+00	1.93E-09
CL flux regulator	*SREBF1*	1.00E+00	1.00E+00
	SREBF2	5.66E-03	1.00E+00
	PDZK1	6.85E-07	1.00E+00
CL efflux	*ABCB11*	5.58E-11	1.00E+00
	SLC10A1	8.24E-04	1.00E+00
	ABCA1	4.76E-09	1.00E+00
CL oxidation	*KEGG_PRIMARY_BILE_ACID_BIOSYNTHESIS*	1.98E-03	4.16E-01
	REACTOME_BILE_ACID_AND_BILE_SALT_METABOLISM	1.05E-05	8.81E-01
	CYP3A4	3.78E-18	1.00E+00
	CYP3A5	3.66E-13	1.00E+00
	CYP3A7	7.91E-04	1.00E+00
	CYP27A1	2.46E-14	1.00E+00
	CYP11A1	1.00E+00	1.00E+00
	CYP11B1	1.00E+00	1.00E+00

	CYP11B2	3.10E-01	1.00E+00
	CYP17A1	1.80E-01	1.00E+00
	CYP19A1	1.00E+00	1.00E+00
	CYP21A2	3.72E-03	1.00E+00
	HSD3B1	1.00E+00	1.00E+00
	HSD3B2	1.00E+00	1.00E+00
	HSD17B1	1.00E+00	2.97E-03
	HSD17B3	1.00E+00	1.00E+00
	HSD17B7	1.57E-02	1.00E+00
	SRD5A1	8.60E-03	1.00E+00
	SRD5A11	8.60E-03	1.00E+00
	KEGG_STEROID_HORMONE_BIOSYNTHESIS	7.88E-06	9.01E-01
CL intracellular flux	*STAR*	1.00E+00	1.00E+00
	STARD3	7.17E-04	1.00E+00
	STARD6	1.00E+00	1.00E+00
Nuclear receptor	*PGR*	1.00E+00	1.00E+00
	ESR1	2.51E-08	1.98E-07
	ESR2	4.12E-04	1.66E-07
	AR	1.85E-11	1.00E+00
	ESRRA	1.85E-01	1.00E+00
	NR5A1	1.16E-06	1.00E+00

	NR1H2	1.00E+00	1.00E+00
	NR1H3	2.38E-07	1.00E+00
	NR1H4	2.50E-06	1.00E+00
Cell proliferation and cell cycle	*FGF19*	1.00E+00	1.00E+00
	FGF21	7.04E-05	1.00E+00
	FGFR1	5.50E-07	4.03E-17
	FGFR2	3.81E-01	8.61E-06
	FGFR3	2.61E-01	1.05E-01
	FGFR4	3.49E-01	1.00E+00
	EGFR	3.86E-08	1.00E+00
	ERBB2	1.00E+00	1.00E+00
	ERBB3	1.01E-05	1.00E+00
	TERT	3.95E-04	1.00E+00
	CELL_CYCLE_PROCESS	0.00E+00	1.96E-01
	HGF	1.02E-03	1.00E+00
	MET	1.70E-05	1.00E+00

Table S4.3: The statistical significance of correlations between gene expression. The first column is the names of categories of genes and gene sets that correlations are calculated for; the next two columns are corresponding key genes, and the fourth column is (meta) *p*-value for testing the hypothesis "correlation is significantly positive", and the fifth column is the FDR adjusted *p*-value using the Holm method (See Methods and Materials).

	Gene/gene set 1	**Gene/gene set 2**	***p*-value**	***p*-value(adjusted)**
Cholesterol influx regulator and cholesterol influx genes	*SREBF1*	*MSMO1*	4.04E-02	1.00E+00
	SREBF1	*CYP51A1*	8.04E-02	1.00E+00
	SREBF2	*MSMO1*	4.87E-05	4.82E-03
	SREBF2	*CYP51A1*	9.53E-09	1.32E-06
	SREBF1	*SCARB1*	2.73E-11	4.36E-09
	SREBF1	*LDLR*	1.07E-11	1.72E-09
	SREBF1	*LRP5*	2.75E-04	2.47E-02
	SREBF1	*CD36*	7.77E-02	1.00E+00
	SREBF2	*SCARB1*	4.58E-07	5.86E-05
	SREBF2	*LDLR*	7.79E-11	1.23E-08
	SREBF2	*LRP5*	1.02E-09	1.50E-07
	SREBF2	*CD36*	1.32E-05	1.40E-03
	PDZK1	*SCARB1*	9.68E-11	1.52E-08
Membrane damage response genes and cholesterol influx regulators	*PON1*	*SREBF1*	2.77E-04	2.47E-02
	PON1	*SREBF2*	1.05E-03	8.06E-02
	PON1	*PDZK1*	2.12E-29	3.87E-27
	PON2	*SREBF1*	8.68E-03	5.38E-01
	PON2	*SREBF2*	1.05E-03	8.06E-02
	PON2	*PDZK1*	7.69E-06	8.68E-04
	PON3	*SREBF1*	1.03E-01	1.00E+00

	PON3	*SREBF2*	6.27E-02	1.00E+00
	PON3	*PDZK1*	4.72E-18	8.35E-16
	SQSTM1	*SREBF1*	1.64E-06	2.00E-04
	SQSTM1	*SREBF2*	1.37E-09	2.01E-07
	SQSTM1	*PDZK1*	1.95E-10	2.99E-08
	HMOX1	*SREBF1*	3.33E-01	1.00E+00
	HMOX1	*SREBF2*	1.17E-01	1.00E+00
	HMOX1	*PDZK1*	1.13E-07	1.49E-05
	PRDX1	*SREBF1*	2.27E-01	1.00E+00
	PRDX1	*SREBF2*	2.90E-03	1.97E-01
	PRDX1	*PDZK1*	1.35E-02	7.53E-01
	ALOX15	*SREBF1*	2.33E-02	1.00E+00
	ALOX15	*SREBF2*	1.78E-03	1.24E-01
	ALOX15	*PDZK1*	1.44E-01	1.00E+00
	ALOX15B	*SREBF1*	1.12E-02	6.51E-01
	ALOX15B	*SREBF2*	5.81E-07	7.38E-05
	ALOX15B	*PDZK1*	1.85E-04	1.70E-02
	ALOX12	*SREBF1*	4.06E-01	1.00E+00
	ALOX12	*SREBF2*	1.30E-02	7.41E-01
	ALOX12	*PDZK1*	1.50E-09	2.20E-07
	ALOX5	*SREBF1*	3.10E-04	2.73E-02
	ALOX5	*SREBF2*	7.34E-04	5.87E-02

	ALOX5	*PDZK1*	3.21E-09	4.56E-07
NR and growth factors/receptors	*ESR1*	*EGFR*	7.90E-21	1.42E-18
	ESR1	*HGF*	7.21E-28	1.31E-25
	ESR1	*MET*	5.45E-15	9.42E-13
	ESR1	*TERT*	1.37E-10	2.12E-08
	ESR2	*EGFR*	1.84E-18	3.28E-16
	ESR2	*HGF*	3.25E-15	5.66E-13
	ESR2	*MET*	3.70E-19	6.63E-17
	ESR2	*TERT*	5.08E-06	5.84E-04
	AR	*EGFR*	1.35E-10	2.10E-08
	AR	*ERBB2*	7.87E-09	1.10E-06
	AR	*ERBB3*	2.07E-04	1.89E-02
	AR	*TERT*	6.29E-02	1.00E+00
	PGR	*EGFR*	8.52E-06	9.46E-04
	PGR	*ERBB2*	1.20E-05	1.29E-03
	PGR	*ERBB3*	1.33E-04	1.25E-02
	ESRRA	*ERBB2*	9.24E-12	1.50E-09
	NR1H2	*FGF19*	1.07E-05	1.16E-03
	NR1H2	*FGFR1*	4.34E-08	5.91E-06
	NR1H2	*FGFR2*	9.03E-05	8.76E-03
	NR1H2	*FGFR3*	5.84E-07	7.38E-05
	NR1H2	*FGFR4*	5.13E-14	8.62E-12

	NR1H3	*FGF19*	2.64E-01	1.00E+00
	NR1H3	*FGFR1*	2.12E-09	3.07E-07
	NR1H3	*FGFR2*	1.26E-05	1.35E-03
	NR1H3	*FGFR3*	4.82E-05	4.82E-03
	NR1H3	*FGFR4*	1.15E-01	1.00E+00
	NR1H4	*FGF19*	3.71E-01	1.00E+00
	NR1H4	*FGF21*	4.97E-12	8.10E-10
	NR1H4	*FGFR1*	8.04E-06	9.00E-04
	NR1H4	*FGFR2*	2.94E-10	4.44E-08
	NR1H4	*FGFR3*	1.73E-05	1.82E-03
	NR1H4	*FGFR4*	2.48E-02	1.00E+00
Cholesterol metabolites and NRs	*ABCA1*	*EGFR*	6.47E-15	1.11E-12
	ABCG1	*EGFR*	2.72E-05	2.83E-03
	ABCG5	*EGFR*	6.42E-07	8.02E-05
	ABCB11	*EGFR*	7.40E-17	1.30E-14
	SLC10A1	*EGFR*	2.39E-12	3.92E-10
	CYP27A1	*ERBB2*	2.65E-01	1.00E+00
	CYP27A1	*NR1H2*	7.00E-03	4.41E-01
	CYP27A1	*NR1H3*	1.77E-12	2.93E-10
	CYP27A1	*ESR1*	4.41E-24	7.98E-22
	CYP3A4	*NR1H2*	1.61E-07	2.10E-05
	CYP3A4	*NR1H3*	2.80E-05	2.88E-03

	CYP3A5	*NR1H2*	3.13E-06	3.73E-04
	CYP3A5	*NR1H3*	4.54E-06	5.26E-04
	CYP3A7	*NR1H2*	8.26E-04	6.45E-02
	CYP3A7	*NR1H3*	4.89E-08	6.60E-06
	CYP11A1	*ESR1*	9.60E-15	1.64E-12
	CYP11A1	*NR1H2*	1.37E-03	9.88E-02
	CYP11A1	*NR1H3*	6.20E-02	1.00E+00
GFs and corresponding GFRs	*HGF*	*MET*	4.77E-14	8.06E-12
	FGF19	*FGFR1*	1.04E-02	6.21E-01
	FGF19	*FGFR2*	5.82E-02	1.00E+00
	FGF19	*FGFR3*	1.31E-01	1.00E+00
	FGF19	*FGFR4*	6.73E-07	8.34E-05
	FGF21	*FGFR1*	4.03E-06	4.71E-04
	FGF21	*FGFR2*	1.36E-06	1.67E-04
	FGF21	*FGFR3*	1.27E-04	1.20E-02
	FGF21	*FGFR4*	3.19E-03	2.14E-01
Membrane damage response genes and cholesterol influx genes	*PON1*	*SCARB1*	2.38E-14	4.04E-12
	PON1	*LDLR*	3.57E-09	5.03E-07
	PON1	*LRP5*	7.54E-06	8.59E-04
	PON2	*SCARB1*	2.75E-09	3.96E-07
	PON2	*LDLR*	1.98E-07	2.57E-05
	PON2	*LRP5*	1.12E-03	8.29E-02

	PON3	*SCARB1*	1.27E-12	2.10E-10
	PON3	*LDLR*	1.66E-10	2.56E-08
	PON3	*LRP5*	3.60E-05	3.68E-03
	SQSTM1	*SCARB1*	3.03E-10	4.55E-08
	SQSTM1	*LDLR*	9.51E-14	1.59E-11
	SQSTM1	*LRP5*	3.57E-07	4.60E-05
	PON1	*MSMO1*	3.41E-04	2.93E-02
	PON1	*CYP51A1*	3.54E-04	3.01E-02
	PON2	*MSMO1*	1.09E-08	1.50E-06
	PON2	*CYP51A1*	2.88E-10	4.38E-08
	PON3	*MSMO1*	2.62E-06	3.17E-04
	PON3	*CYP51A1*	3.64E-06	4.29E-04
	SQSTM1	*MSMO1*	2.01E-02	1.00E+00
	SQSTM1	*CYP51A1*	2.86E-09	4.09E-07
O_2 level and membrane damage response genes	*HIF1AN*	*PON1*	2.85E-06	3.42E-04
	HIF1AN	*PON2*	2.04E-03	1.41E-01
	HIF1AN	*PON3*	3.52E-02	1.00E+00
	HIF1AN	*SQSTM1*	4.89E-11	7.78E-09
	HIF1AN	*HMOX1*	2.13E-02	1.00E+00
	HIF1AN	*PRDX1*	6.10E-01	1.00E+00
	HIF1AN	*ALOX15*	8.40E-05	8.23E-03
	HIF1AN	*ALOX15B*	1.06E-02	6.27E-01

	HIF1AN	*ALOX12*	4.92E-04	4.04E-02
	HIF1AN	*ALOX5*	7.62E-08	1.01E-05
	HIF1AN	*SREBF1*	2.79E-01	1.00E+00
	HIF1AN	*SREBF2*	6.50E-04	5.27E-02
	HIF1AN	*PDZK1*	1.63E-02	8.95E-01
O_2 level and cholesterol intracellular transporters	*HIF1AN*	*STAR*	1.09E-03	8.14E-02
	HIF1AN	*STARD3*	1.34E-03	9.81E-02
	HIF1AN	*STARD6*	9.29E-05	8.92E-03
O_2 level and CYP genes	*HIF1AN*	*CYP11A1*	7.33E-04	5.87E-02
	HIF1AN	*CYP3A4*	7.72E-16	1.35E-13
	HIF1AN	*CYP3A5*	1.78E-04	1.66E-02
	HIF1AN	*CYP3A7*	9.27E-03	5.66E-01
	HIF1AN	*CYP27A1*	4.75E-02	1.00E+00
	HIF1AN	*CYP11B1*	3.73E-04	3.14E-02
	HIF1AN	*CYP11B2*	5.81E-08	7.79E-06
	HIF1AN	*CYP17A1*	3.10E-04	2.73E-02
	HIF1AN	*CYP19A1*	4.46E-04	3.70E-02
	HIF1AN	*CYP21A2*	1.62E-03	1.15E-01

Table S4.4: Oxidized cholesterol products and associated genes.

Oxidized cholesterol metabolites	**Catalyzing enzymes/pathway**
27-OHC	*CYP27A1*
Steroid hormones	steroid hormone synthesis pathway
22-OHC	*CYP11A1*
4β-OHC	*CYP3A4,5,7*
Bile acids	Bile acid synthesis pathway

Table S4.5: Known relationships between cholesterol metabolites and *NRs*.

CYP gene or product	**Relationship**	**NRs**
27-OHC	activates	*ESR1, LXR*
CYP27A1	is regulated by	*PPAR,* cholesterol
22(R)-OHC	activates	*ESR1, LXR*
CYP11A1, CYP17A1, CYP19A1, CYP21A2	are regulated by	*NR5A1*
4β-OHC	activates	*LXR*
CYP3A4,5,7	are regulated by	*unknown*
bile acid	activates	*FXR*
bile-acid producing *CYPs*	are regulated by	*FXR*
steroid hormone	activates	*ER, AR, PGR*
steroidogenic *CYPs*	are regulated by	*NR5A1*

Table S4.6: Clinical information of tissue samples for quantitative metabolic profiling.

Disease	ID	Gender	Age
liver metastase from colon	W0598	M	50
liver metastase from colon	G0504	M	60
liver metastase from breast	G0270	F	54
liver metastase from colon	W0225	M	43
liver metastase from colon	G0480	M	64
primary liver cancer	G0004	M	46
primary liver cancer	G0018	M	53
primary liver cancer	G0014	F	61
primary liver cancer	G0008	M	55
primary liver cancer	G0005	M	60
Primary gastric cancer	W0388	F	56
Primary gastric cancer	W0415	F	64
Primary gastric cancer	W0392	M	38
Primary gastric cancer	W0390	M	54
Primary gastric cancer	W0417	M	41
primary colon cancer	W0424	M	70
primary colon cancer	W0399	F	82
primary colon cancer	W0427	M	55
primary colon cancer	W0422	F	58
primary colon cancer	W0400	F	62

CHAPTER 5

DEVELOPMENT OF A DECONVOLUTION ALGORITHM FOR TISSUE-BASED GENE EXPRESSION DATA

Introduction

Traditionally, cancer research has been mostly conducted on cancer cell lines cultured in man-made environments. While large amounts of data have been published about such studies, their true relevance to cancer biology remains largely unknown, at least for certain aspects of cancer biology. For example, autophagy has been widely considered to have key roles in cancer development based on cancer cell line studies [210]. Our recent analyses of cancer tissue gene-expression data clearly showed that macro-autophagy, the most studied autophagy in cancer, is consistently repressed in cancer tissues across 11 cancer types [211]; in comparison, it is up-regulated in cancer cell lines when treated with nutrient deprivation and/or metabolic stress conditions [212, 213]. Examples like this strongly suggest the necessity in studying cancer tissues in addition to cancer cell lines in order to understand the real biology of cancer. Actually, it has been well established that it is essential to study the microenvironment where cancer cells originate and evolve [214, 215], which consists of multiple types of immune, stromal cells, fat, endothelial and blood cells along with the extracellular matrix (ECM). It is this environment that dictates how a disease evolves.

Tissue data provide substantially more information than cell-line data and offer new opportunities to study cancer biology and evolution in its actual microenvironment, when multiple tissue samples of the same cancer type are analyzed together. However, it is very challenging to do

information discovery from tissue data because of their compositional complexity – each dataset represents a mixture of gene-expression data from multiple cell types. Hence, meaningful tissue-data analyses require to first sort out the detailed contributions to the observed tissue-level data by different cell types, like experimentally using laser-directed microdissection to put cells of different types into separate bins. Once such data are de-convoluted, one can start addressing issues concerning specific genes and pathways in certain cell types and interactions among different cell types. For example, one can possibly rigorously examine the Warburg effect in cancer tissue cells by asking: "Do some cancers utilize the electron transport chain to produce ATPs in addition to glycolysis?" or "Are there (fundamental) differences between the Warburg effect observed in cancer tissue cells and normal proliferating cells as answers to this question have been conflicting?" These issues can be examined only when gene-expression levels are accurately estimated for each involved cell type. Actually, tissue data deconvolution is not only very useful, but also necessary for correct data interpretation. For example, different cancer tissues may have varying proportions of cancer cells, which makes direct comparison between expression levels of individual genes in two tissues challenging without data deconvolution.

A number of large databases for cancer tissue omic data have been developed. Among them, TCGA is the most comprehensive [216]. It currently consists of 34 cancer types and 11,000+ tissue samples, having for each transcriptomic, genomic, epigenomic and some proteomic data. Numerous studies of TCGA data have been published [217], but the majority of them treat the tissue data as coming from a single source rather than multiple ones. While such studies have revealed useful information, analyses of tissue data without proper deconvolution will be limited to cell type-specific genes or may lead to questionable results. The reality is: while cancer tissue data are rapidly generated, analysis capabilities of such data fall far behind.

The computational challenge in solving the tissue data deconvolution problem stems from the reality: each cell type has a very large number of complex relations among its expressed genes and pathways, which are preserved under different conditions. To make deconvolution results meaningful, some or many of these relations, e.g., co-expression among functionally closely related genes, must be captured and enforced in a deconvolution problem formulation. This, as one can imagine, is a daunting task.

Classical methods of profile deconvolution assume that a mixed profile is a linear combination of a predetermined number of pure constituent profiles. Written in matrix form, the measured mixed profiles X are a product of S, a matrix of gene expression profiles of each constituent, weighted by the fractions P of each cell type in the mixture, and a de-convolution problem is defined as to estimate S and P so that the total error $\|\epsilon\|$ is minimized with $\mathrm{X_{M\times N}} = \mathrm{S_{M\times K}} \cdot \mathrm{P_{K\times N}} + \epsilon$, *s.t.*: $\sum_{t=1}^{N} P_{it} = 1, P_{it} \geq 0;\ i = 1, \ldots, K$. Different algorithms use different methods of deconvolution. Estimating both signature and proportion matrix is a very loosely constrained problem, which requires assumptions on the structures of the gene expression matrix [43-47]. Some assume that the fractions, X, are known, and (average) cell type expression profiles are derived [48-51]. Others derive cell type uniquely express genes and/or cell type specific expression profiles to estimate the cell type proportions matrix P [43, 52-57, 218]. These approaches do not bring insight into the cancer patient-specific variations, as they assume that all tumor sample share a constant expression profiles, up to varying component abundances. Methods to derive patient-specific expression profiles were developed, which requires input of matched/unmatched normal tissue expression profiles, and assumes there is only one component in addition to tumor cells, called “normal”, are within the tumor tissue [58, 60-62], however, no knowledge regarding the immune cells are derived.

Model setup

There are a few serious issues with the current deconvolution formulation: (i) it requires that each gene in each cell type has a constant expression across all tissues under consideration, which is clearly too restrictive and unrealistic; (ii) no information is included to enforce any co-relations among genes in the same cell type; (iii) none of the algorithms attempted to deal with the special challenges in de-convoluting cancer tissue data as mentioned earlier. Some authors have limited their deconvolution algorithms to specialized applications, such as assessing the purity of cancer tissues [219]. It will represent a major step forward if a highly effective deconvolution technique becomes available that can tease out contributions by different cell types in tissue-based RNA data.

Our developed deconvolution tool overcame all theaw limitations, as it is capable of 1) deriving cell type proportions and expression profiles simultaneously; 2) deriving cancer patient specific expression profiles for all the cell types; 3) covering major types of immune cells; with the only input as any number of mixture tissue expression.

We have collected a large amount of gene expression data for the following cell types as our traning data: B-cell, breast, colon, dendritic cell, endothelial, fibroblast, liver, macrophage, neutrophil and T-cell, measured using Affymetrix U133 Plus 2.0 Array, which are retrieved from GEO [220], totaling 406, 513, 745, 410, 638, 398, 341,412, 477, and 445 samples, respectively. Over 20,000 gene sets retrieved from Msigdb [113], covering ~20,000 genes, are used in our study. All expression data are normalized using MAS5, log2 transformed, and quantile-normalized. The pipeline of our method includes the following steps: 1) estimating the number of uniquely expressed genes for each cell type; 2) detecting cell type specific pathways; 3) their expression signatures; 4) inferring the cell type proportions; 5) deconvolution of cell specific contribution of a tissue sample.

Estimating the numbers of expressed and uniquely expressed genes in each cell type: We estimated which genes are typically expressed in each of the ten cell types using the above data. For each gene in each cell type, we fit a bimodal Gaussian mixture model against the density distribution of the gene's expression. Samples falling under the peak with smaller mean values are considered as not expressing the gene. Genes are considered *not expressed* in a cell type if < 25% of the samples of the cell type express the gene. An expressed gene is considered as *unique* to a cell type if the majority (> 75%) of the samples of the cell type express the gene and < 25% of the samples of the other cell types express the gene. The numbers of uniquely expressed genes in the cell types are given as the first number in the parenthesis following each cell type: B-cell (72, 1446), breast (67, 1687), colon (40, 1680), dendritic cell (9, 1465), endothelial (39, 1557), fibroblast (134, 1236), liver (317, 1549), macrophage (15, 1319), neutrophil (19, 456) and T-cell (83, 1650).

Detecting cell-type specific pathways: We have examined each of the ~20,000 gene sets in Msigdb, and if (subset of) genes in a set are expressed and co-expressed for a cell type, this (subset of) genes are considered as this cell type's specific co-expressed gene cluster, this their co-expression signatures are derived using sparse non-negative matrix factorization. The procedure is performed on all cell types for all the gene sets, and we identified all of the pathways specific to each cell type, listed as the second number inside the above parentheses. Hence, pathways refer to such gene clusters throughout the paper.

Estimating the expression signature of each pathway: Given a certain cell type, for each of its specific pathway, we fit the expression matrix (genes in row, samples in column) of its genes with two non-negative rank 1 matrices, and call the matrix on the left as signature for this pathway. Binding all the signatures into a large matrix, which each column containing the expression signature for one pathway, and locations in the column where genes do not belong to the pathway

are set as zero. As pathways inevitably overlap, the signature for genes who occur in multiple pathways are not accurate. We solved this problem with fine tuning the signature matrix by performing a hierarchical alternative least square (HALS) matrix decomposition on the training expression matrix for the cell type [221], using the "coarse" signature matrix as initial point. HALS adopts a column-wise updating scheme, which fits well into our problem, as for each iteration, it would be very convenient to keep the zero valued elements in the old signature matrix as zero in the updated signature matrix.

Inference of proportions of individual cell types in a cell mixture: The proportions of individual cell types in a cell mixture can be estimated using the expression levels of genes uniquely expressed in each cell type. The challenge lies in that the observed expression levels for such genes have two contributions: expression levels of the genes in each cell type and the proportions of the cell type in the cell mixture. Hence, we need to have a way to decouple the two. For each cell type, we are able to find a number of uniquely expressed genes whose expression fall within a narrow range; and hence their total expression in the mixture offer a reliable measure for estimating the proportion of each cell type. Specifically, an expression matrix comprised of the mean expression values of all such unique genes for each cell type is constructed. Then, the problem of estimating the proportion of each cell type in a mixture can be formulated as to find a solution $\{p_i, i = 1, \dots, K\}$ that minimizes the following sparse non-negative least square regression problem: where $Y_{Q\times 1}$are the expression data for all the Q cell type-specific low-variability genes in a given mixture; $e_{ij}, i = 1, \dots K; j = 1, \dots, D_i$, is the mean expression value of the jth gene in the ith cell type; and $p_i, i = 1, \dots, K$, is the proportion of the ith cell type in the mixture, λ is regularization parameter, which is selected by cross-validation.

$$\arg\min_{\mathrm{p_t}}\left(\left\|Y_{Q\times 1}-\begin{pmatrix} e_{11} & \cdots & 0 \\ \vdots & & \vdots \\ e_{1D_1} & & 0 \\ \vdots & \ddots & \vdots \\ 0 & & e_{K1} \\ \vdots & \cdots & \vdots \\ 0 & & e_{KD_K} \end{pmatrix}\begin{pmatrix} p_1 \\ \vdots \\ p_K \end{pmatrix}\right\|_F+\lambda\sum_{t=1}^{K}|p_t|\right), s.t.\sum_{t=1}^{K}p_t=1, p_t\geq 0$$

Deconvolution of cell specific contribution of a tissue sample: Given a vector of summed expression levels for G genes from ten cell types, the de-convolution problem is defined as to estimate $\{M_{ti}\}$ so that the following is minimized:

$$\arg\min_{\mathrm{M_{ti}}}\left(\left\|X-\sum_{t=1}^{K}\begin{pmatrix} S_{t11} & \cdots & S_{tC_t1} \\ \vdots & \ddots & \vdots \\ S_{t1G} & \cdots & S_{tC_tG} \end{pmatrix}\left(p_t\begin{pmatrix} M_{t1} \\ \vdots \\ M_{tC_t} \end{pmatrix}\right)\right\|_F+\lambda\sum_{t=1}^{K}\sqrt{\beta_t}\sqrt{\sum_{j=1}^{C_t}M_{tj}^2}\right), s.t. M_{ti}\geq 0$$

where $S_{ti}=\{S_{tij}, j=1,\dots G\}$ is the signature of the ith pathway in cell type t. It has non-zero values only in rows that correspond to genes in the pathway; p_t is the estimated proportion of cell type t in the mixture; and M_{ti} is the expression level of the ith pathway in cell type t. λ, $\beta 1$,…,βK are group lasso regularization parameters. Group lasso penalizes the coefficients for the same cell type on a group level, meaning the entire group (cell type) may drop out of the model. This is suitable to remove cell types not present in the tissue.

Simulation study

Assessment of the performance of deconvolution algorithm. One thousand simulated mixture sample expression are generated as follows. For each mixture sample, ten expression datasets each from the ten aforementioned cell types were randomly selected from the corresponding GEO datasets along with a proportion vector $p_i, i=1,\dots,10$ generated using$(p_1,\dots,p_{10})\sim Dirichlet(10)$. For each gene g, its expression in a mixture is determined by $E(g)=\sum_{i=1}^{10}p_iE_i(g)+\epsilon_g$, where $E_i(g)$ is the expression level of g in cell type i, and ϵ_g is an

additive error following a Gaussian distribution. Figure 5. 1 shows the actual *versus* the predicted proportions for the nine cell types, highlighting the quality of the de-convolution algorithm. Figure 5. 2 shows for six mixture samples, the expression profiles of cell type breast before and after deconvolution, where the actual proportions of breast cells vary from 0.3-0.8. Overall, the median correlations between the deconvoluted and actual expression across all the genes are all above 0.85.

Discussion

There are two challenges with the current formulation: 1) more convincing validation methods are needed; 2) it cannot handle newly emerged pathways/subtypes. Clearly, the current method validation purely depends on simulation data, which has the linear structure as we assumed in the model. Validation on real bio-specimens are needed.

New pathways may emerge to contribute to tissue-based data due to multiple reasons: (i) genomic mutations, (ii) epigenomic alteration, and (iii) abnormal conditions that may trigger unusual pathways. The reason that our cell-line based study missed a pathway is that it was not included in the master list. We will infer new pathways based on co-expression patterns among genes in the residual matrix. For a given set of such co-expressed genes in the residual matrix, we assess if they belong to one or a few missing pathways through bi-clustering analysis across all gene-expression data for each relevant cell type using cell line data first. We predict each maximal bi-cluster as a *candidate pathway* if it covers a substantial fraction of all the cell-line experimental conditions in the relevant gene-expression datasets under study and has sufficiently high statistical significance (both parameters to be determined), using our in-house and widely used bi-clustering tool QUBIC [222]. For each predicted new pathway (i.e., a cluster identified via bi-clustering analysis), we derive its signature using the procedure outlined in **Preliminary Study**. Then we check if some

residuals still have significant co-expression patterns, and continue the above till no such residuals left or no further improvements can be made. If it was the latter, we conclude that the missing pathways are not detectable using the cell-line data, and will conduct a similar bi-clustering analysis over tissue-based data under study. The main difference is that (1) only cell types that account for major proportions in the tissues may have their co-expressed genes detectable via tissue-based clustering; and (2) we need to predict which cell types each missing pathway may belong. For (1), only cancer cell or 1 - 2 cell types that are most strongly associated with cancer cells are considered, as discussed earlier. For (2), we assign each predicted pathway to cell types if the genes show the strongest correlation with other genes of the cell types.

Conclusion

Direct interpretation of tissue-based gene-expression data without sorting out the actual expression levels of each gene in each cell type could lead to incorrect results, particularly missing subtle but important changes and interplay among different cell types. Thus, we developed an algorithm for deconvolution of tissue-based gene-expression data to cell type-specific contributions for each gene. The informational basis of our planned algorithm is: each cell type (a) expresses a unique set of genes; and (b) has a unique combination of expressed pathways, each of which defines a condition-independent covariance among the expression levels of its participating genes. The current formulation deconvolution method enables us to answer a lot of interesting cancer biology questions, which otherwise are not approachable. To name a few, do cancer tissues of the same type but different grades, going from un-differentiated to highly differentiated, tend to have the proportion of any specific cell type going up or down monotonically; Do such tissues tend to have the expression levels of specific pathways in a specific cell type going up or down monotonically; Do such tissues tend to have co-expression levels among specific pathways intracellularly going

up or down monotonically; Do such tissues tend to have co-expression levels among specific pathways across different cell types going up or down monotonically. All these questions can be answered straightforwardly once the deconvolution results are available.

Figures

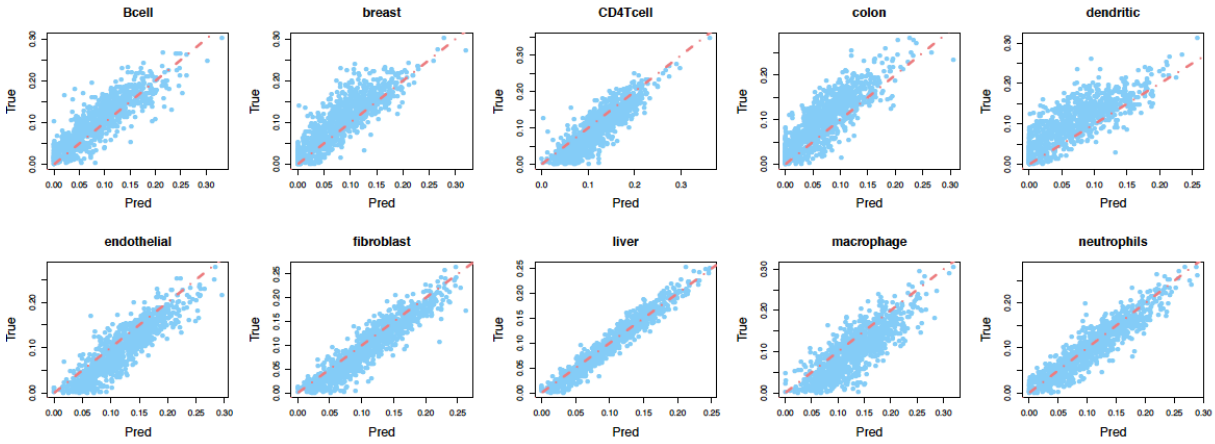


Figure 5.1: Predicted (x-axis) vs. actual (y-axis) proportions of cell types across 100 simulated mixtures.

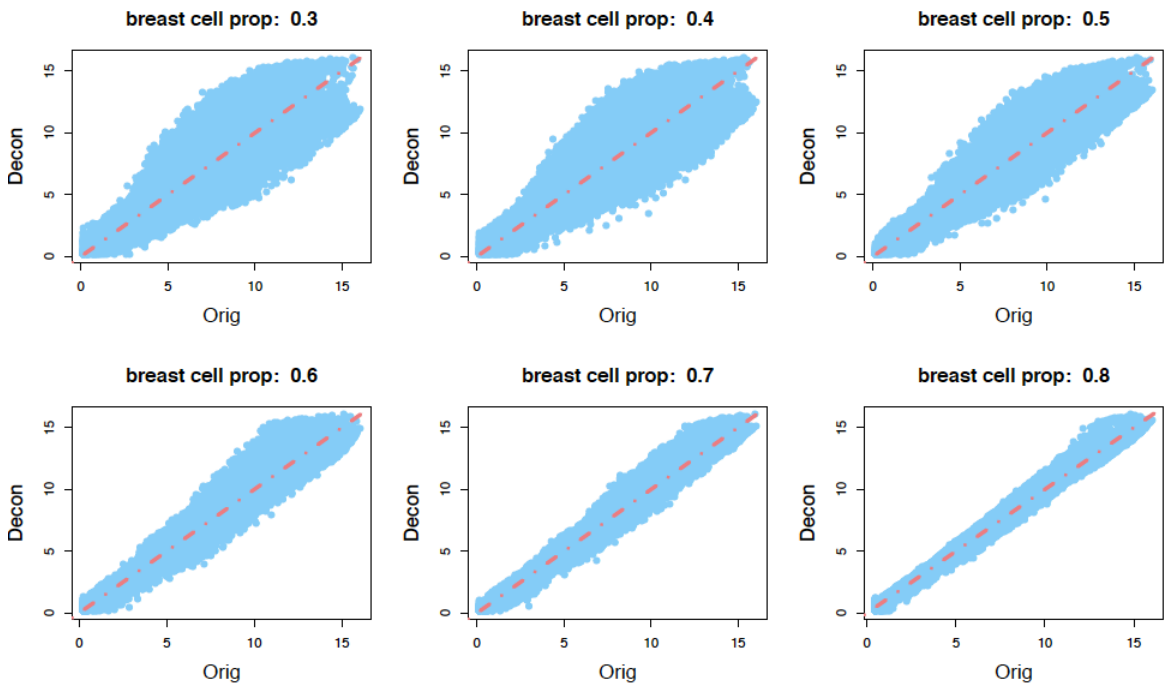


Figure 5.2: The tissue specific expression profiles of breast cell components with different levels of abundances. The x-axis represents the true expression profiles of breast cells in tissues, while y-axis the deconvoluted profiles.

CHAPTER 6

CONCLUSIONS

In the past six years, my research has covered the following areas: (1) elucidation of key drivers of post-metastatic cancer evolution [223]; (2) modeling of complex behaviors of competing processes in cancer tissues [224]; (3) development and application of effective algorithms for de-convoluting tissue-based gene-expression data to cell-type specific contributions [225]; (4) derivation of novel roles of elevated mRNA expressions and DNA methylation in cancer cell survival [226, 227]; and (5) development of a causal framework for assessing the functional roles played by somatic mutations in cancer development [228]. In addition, I have been involved in multiple projects as a contributing author, including (i) elucidation of functional roles of micro-environmental stresses in cancer initiation and progression [31, 47, 229, 230]; (ii) characterizations of aberrations in different stages of cancer [231, 232]; (iii) cancer biomarker identification [233]; (iv) genomic mutation annotations in gastric cancer [234]; and (v) development of pipeline for RNA-Seq data processing [235, 236]. There have been two tracks for my research: one focused on development of quantitative techniques essential to solving general and challenging cancer biology questions; and one focused on addressing important cancer biology questions, through data mining and quantitative modeling. The advantage of this two-track approach includes: (i) having a good understanding about key cancer biology problems will guide me to focus on the most important tool development problems; and (ii) working directly on cancer biology problems will enable me to work directly with cancer biologists.

Three important biological questions are examined through data mining for my thesis projects. First of all, we discovered that: (1) a combination of basal metabolic rate and oxidative stress level in a tissue well explained the variations of lifetime risk of cancers of different types across different populations; and (2) somatic mutations may be predominantly selected to serve as facilitators rather than primary drivers of cancer formation. Secondly, we identified (1) a possible determinant for global DNA methylation in cancer cells, and (2) competitive relationship between DNA methylation and nucleotide synthesis for methyl resource, and competitive relationship between DNA methylation and intracellular oxidative stress for sulfur resource in cancer cells. Thirdly, we have identified that the altered O_2 level in primary cancer sites *vs.* metastatic cancer sites represents a key stress that the newly migrated cancer cells must overcome and cholesterol is selected by the metastatic cancer cells as oxygen barrier, and the oxidation derivatives of which is a key driver for the explosive growth of post-metastatic cancers.

A novel computational tool to enable the study of cancer microenvironment using tissue expresson data is developed. Our algorithm for de-convoluting tissue-based data to the cell-type specific contributions are based on the following information: (1) genes in each cell type are expressed in coordinated manners, specifically they are grouped into pathways whose genes are co-expressed; and (2) different cell types tend to have different sets of pathways activated.

REFERENCES

1. Stehelin, D., Varmus, H. E., Bishop, J. M. & Vogt, P. K, *DNA related to the transforming gene(s) of avian sarcoma viruses is present in normal avian DNA.* nature, 1976. **260**: p. 170-173.
2. Alfred G. Knudson, J., *Mutation and Cancer: Statistical Study of Retinoblastoma.* Proc Natl Acad Sci U S A., 1971. **68**(4): p. 820-823.
3. Nowell P, H.D., *A minute chromosome in chronic granulocytic leukemia.* Science, 1960. **132**(3438).
4. I Nishisho, Y.N., Y Miyoshi, Y Miki, H Ando, A Horii, K Koyama, J Utsunomiya, S Baba, P Hedge, *Mutations of chromosome 5q21 genes in FAP and colorectal cancer patients.* Science, 1991. **2253**(5020): p. 665-669.
5. P. Andrew Futreal , L.C., Mhairi Marshall , Thomas Down , Timothy Hubbard , Richard Wooster , Nazneen Rahman & Michael R. Stratton, *A census of human cancer genes.* Nature Reviews Cancer, 2004. **4**: p. 177-183.
6. Siegel, R., Naishadham, D. and Jemal, A., *Cancer statistics, 2013.* CA: A Cancer Journal for Clinicians, 2013. **63**(1): p. 11-30.
7. Vaupel P, M.A., *Hypoxia in cancer: significance and impact on clinical outcome.* Cancer Metastasis Rev. , 2007. **26**(2): p. 225-39.
8. Eltzschig, H.K. and P. Carmeliet, *Hypoxia and inflammation.* N Engl J Med, 2011. **364**(7): p. 656-65.

9. Chandel, R.B.H.a.N.S., *Mitochondrial reactive oxygen species regulate hypoxic signaling.* Curr Opin Cell Biol., 2009. **21**(6): p. 894-899.

10. 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.

11. Pani, G., T. Galeotti, and P. Chiarugi, *Metastasis: cancer cell's escape from oxidative stress.* Cancer Metastasis Rev, 2010. **29**(2): p. 351-78.

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

13. Tomasetti, C. and B. Vogelstein, *Variation in cancer risk among tissues can be explained by the number of stem cell divisions.* Science, 2015. **347**(6217): p. 78-81.

14. Tomasetti, C. and B. Vogelstein, *Cancer risk: role of environment-response.* Science, 2015. **347**(6223): p. 729-31.

15. Forman, D., et al., *Cancer Incidence in Five Continents, Vol. X* 2013.

16. Ehrlich, M., *DNA hypomethylation in cancer cells.* Epigenomics, 2009. **1**(2): p. 239-59.

17. Li, J., et al., *The prognostic value of global DNA hypomethylation in cancer: a meta-analysis.* PLoS One, 2014. **9**(9): p. e106290.

18. Kuchiba, A., et al., *Global methylation levels in peripheral blood leukocyte DNA by LUMA and breast cancer: a case-control study in Japanese women.* Br J Cancer, 2014. **110**(11): p. 2765-71.

19. Nojima, M., et al., *Global, cancer-specific microRNA cluster hypomethylation was functionally associated with the development of non-B non-C hepatocellular carcinoma.* Mol Cancer, 2016. **15**(1): p. 31.

20. Cadieux, B., et al., *Genome-wide hypomethylation in human glioblastomas associated with specific copy number alteration, methylenetetrahydrofolate reductase allele status, and increased proliferation.* Cancer Res, 2006. **66**(17): p. 8469-76.

21. Yang, X., et al., *Gene body methylation can alter gene expression and is a therapeutic target in cancer.* Cancer Cell, 2014. **26**(4): p. 577-90.

22. Davis, C.D. and E.O. Uthus, *DNA methylation, cancer susceptibility, and nutrient interactions.* Exp Biol Med (Maywood), 2004. **229**(10): p. 988-95.

23. Field, M.S., et al., *Nuclear enrichment of folate cofactors and methylenetetrahydrofolate dehydrogenase 1 (MTHFD1) protect de novo thymidylate biosynthesis during folate deficiency.* J Biol Chem, 2014. **289**(43): p. 29642-50.

24. Klein, C.A., *Parallel progression of primary tumours and metastases.* Nat Rev Cancer, 2009. **9**(4): p. 302-12.

25. Oda, T., et al., *Growth rates of primary and metastatic lesions of renal cell carcinoma.* Int J Urol, 2001. **8**(9): p. 473-7.

26. Chaffer, C.L. and R.A. Weinberg, *A Perspective on Cancer Cell Metastasis.* Science, 2011. **331**(6024): p. 1559-1564.

27. Lambert, A.W., D.R. Pattabiraman, and R.A. Weinberg, *Emerging Biological Principles of Metastasis.* Cell, 2017. **168**(4): p. 670-691.

28. Steeg, P.S., *Targeting metastasis.* Nat Rev Cancer, 2016. **16**(4): p. 201-218.

29. Callari, M., et al., *Subtype-Specific Metagene-Based Prediction of Outcome after Neoadjuvant and Adjuvant Treatment in Breast Cancer.* Clin Cancer Res, 2016. **22**(2): p. 337-45.

30. Brastianos, P.K., et al., *Genomic Characterization of Brain Metastases Reveals Branched Evolution and Potential Therapeutic Targets.* Cancer Discov, 2015. **5**(11): p. 1164-77.

31. 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.* International Journal of Cancer, 2015. **136**(9): p. 2001-2011.

32. Quail, D.F. and J.A. Joyce, *Microenvironmental regulation of tumor progression and metastasis.* Nature medicine, 2013. **19**(11): p. 1423-1437.

33. Krycer, J.R. and A.J. Brown, *Cholesterol accumulation in prostate cancer: a classic observation from a modern perspective.* Biochim Biophys Acta, 2013. **1835**(2): p. 219-29.

34. Danilo, C. and P.G. Frank, *Cholesterol and breast cancer development.* Curr Opin Pharmacol, 2012. **12**(6): p. 677-82.

35. Thysell, E., et al., *Metabolomic Characterization of Human Prostate Cancer Bone Metastases Reveals Increased Levels of Cholesterol.* PLoS ONE, 2010. **5**(12).

36. Baenke, F., et al., *Hooked on fat: the role of lipid synthesis in cancer metabolism and tumour development.* Dis Model Mech, 2013. **6**(6): p. 1353-63.

37. Nielsen, S.F., B.G. Nordestgaard, and S.E. Bojesen, *Statin Use and Reduced Cancer-Related Mortality.* N Engl J Med, 2012. **367**: p. 1792-1802.

38. Smith, B. and H. Land, *Anticancer activity of the cholesterol exporter ABCA1 gene.* Cell Rep, 2012. **2**(3): p. 580-90.

39. Li, Y.C., et al., *Elevated levels of cholesterol-rich lipid rafts in cancer cells are correlated with apoptosis sensitivity induced by cholesterol-depleting agents.* Am J Pathol, 2006. **168**(4): p. 1107-18; quiz 1404-5.

40. Galea, A.M. and A.J. Brown, *Special relationship between sterols and oxygen: were sterols an adaptation to aerobic life?* Free Radic Biol Med, 2009. **47**(6): p. 880-9.

41. Subczynski, W.K., J.S. Hyde, and A. Kusumi, *Oxygen permeability of phosphatidylcholine--cholesterol membranes.* Proc Natl Acad Sci U S A, 1989. **86**(12): p. 4474-8.

42. Buchwald, H., et al., *Effect of plasma cholesterol on red blood cell oxygen transport.* Clin Exp Pharmacol Physiol, 2000. **27**(12): p. 951-5.

43. Gaujoux, R. and C. Seoighe, *Semi-supervised Nonnegative Matrix Factorization for gene expression deconvolution: A case study.* Infection, Genetics and Evolution, 2012. **12**(5): p. 913-921.

44. Lähdesmäki, H., et al., *In silico microdissection of microarray data from heterogeneous cell populations.* BMC Bioinformatics, 2005. **6**: p. 54-54.

45. Repsilber, D., et al., *Biomarker discovery in heterogeneous tissue samples -taking the in-silico deconfounding approach.* BMC Bioinformatics, 2010. **11**(1): p. 1-15.

46. Venet, D., et al., *Separation of samples into their constituents using gene expression data.* Bioinformatics, 2001. **17**(suppl 1): p. S279-S287.

47. Zhang, C., S. Cao, and Y. Xu, *Population dynamics inside cancer biomass driven by repeated hypoxia-reoxygenation cycles.* Quantitative Biology, 2014. **2**(3): p. 85-99.

48. Erkkila, T., et al., *Probabilistic analysis of gene expression measurements from heterogeneous tissues.* Bioinformatics, 2010. **26**(20): p. 2571-7.

49. Ghosh, D., *Mixture models for assessing differential expression in complex tissues using microarray data.* Bioinformatics, 2004. **20**(11): p. 1663-1669.

50. Shen-Orr, S.S., et al., *Cell type-specific gene expression differences in complex tissues.* Nat Methods, 2010. **7**(4): p. 287-9.

51. Stuart, R.O., et al., *In silico dissection of cell-type-associated patterns of gene expression in prostate cancer.* Proceedings of the National Academy of Sciences of the United States of America, 2004. **101**(2): p. 615-620.

52. Abbas, A.R., et al., *Deconvolution of Blood Microarray Data Identifies Cellular Activation Patterns in Systemic Lupus Erythematosus.* PLoS ONE, 2009. **4**(7): p. e6098.

53. Bolen, C., M. Uduman, and S. Kleinstein, *Cell subset prediction for blood genomic studies.* BMC Bioinformatics, 2011. **12**(1): p. 258.

54. Gong, T., et al., *Optimal deconvolution of transcriptional profiling data using quadratic programming with application to complex clinical blood samples.* PLoS One, 2011. **6**(11): p. e27156.

55. Newman, A.M., et al., *Robust enumeration of cell subsets from tissue expression profiles.* Nature methods, 2015. **12**(5): p. 453-457.

56. Qiao, W., et al., *PERT: A Method for Expression Deconvolution of Human Blood Samples from Varied Microenvironmental and Developmental Conditions.* PLoS Comput Biol, 2012. **8**(12): p. e1002838.

57. Zuckerman, N.S., et al., *A Self-Directed Method for Cell-Type Identification and Separation of Gene Expression Microarrays.* PLoS Computational Biology, 2013. **9**(8): p. e1003189.

58. Gosink, M.M., H.T. Petrie, and N.F. Tsinoremas, *Electronically subtracting expression patterns from a mixed cell population.* Bioinformatics, 2007. **23**(24): p. 3328-34.

59. Zhong, Y. and Z. Liu, *Gene expression deconvolution in linear space.* Nature methods, 2011. **9**(1): p. 8-9.

60. Quon, G., et al., *Computational purification of individual tumor gene expression profiles leads to significant improvements in prognostic prediction.* Genome medicine, 2013. **5**(3): p. 29.

61. Ahn, J., et al., *DeMix: deconvolution for mixed cancer transcriptomes using raw measured data.* Bioinformatics, 2013: p. btt301.

62. Clarke, J., P. Seo, and B. Clarke, *Statistical expression deconvolution from mixed tissue samples.* Bioinformatics, 2010. **26**(8): p. 1043-1049.

63. Stehelin, D., et al., *DNA related to the transforming gene(s) of avian sarcoma viruses is present in normal avian DNA.* Nature, 1976. **260**(5547): p. 170-173.

64. Alberts, B., et al., *Extracellular Control of Cell Division, Cell Growth, and Apoptosis.* 2002.

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

66. Vogelstein, B., et al., *Cancer genome landscapes.* Science, 2013. **339**(6127): p. 1546-58.

67. Genova, M.L. and G. Lenaz, *The Interplay between Respiratory Supercomplexes and Ros in Aging.* Antioxid Redox Signal, 2015.

68. Arum, O., et al., *Do altered energy metabolism or spontaneous locomotion 'mediate' decelerated senescence?* Aging Cell, 2015.

69. Notani, P.N., *Global variation in cancer incidence and mortality.* CURRENT SCIENCE-BANGALORE-, 2001. **81**(5): p. 465-474.

70. Drake, J.W., et al., *Rates of spontaneous mutation.* Genetics, 1998. **148**(4): p. 1667-1686.

71. COOKE, M.S., et al., *Oxidative DNA damage: mechanisms, mutation, and disease.* The FASEB Journal, 2003. **17**(10): p. 1195-1214.

72. Paterson, S., et al., *Antagonistic coevolution accelerates molecular evolution.* Nature, 2010. **464**(7286): p. 275-8.

73. Xu, Y., J. Cui, and D. Puett, *Cancer Bioinformatics.* Springer, 2014. **Chapter 4**.

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.* International Journal of Cancer, 2014: p. n/a-n/a.

75. 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.

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

77. Xu, H., et al., *Effect of hyaluronan oligosaccharides on the expression of heat shock protein 72.* Journal of Biological Chemistry, 2002. **277**(19): p. 17308-17314.

78. Curado, M.P., I.A.f.R.o. Cancer, and W.H. Organization, *Cancer Incidence in Five Continents Vol. IX.* 2008.

79. Barrett, T., et al., *NCBI GEO: mining tens of millions of expression profiles—database and tools update.* Nucleic acids research, 2007. **35**(suppl 1): p. D760-D765.

80. Subramanian, A., et al., *Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.* Proceedings of the National Academy of Sciences of the United States of America, 2005. **102**(43): p. 15545-15550.

81. Hayes, J.D. and M. McMahon, *NRF2 and KEAP1 mutations: permanent activation of an adaptive response in cancer.* Trends in Biochemical Sciences, 2009. **34**(4): p. 176-188.

82. Nikolaev, S.I., et al., *A single-nucleotide substitution mutator phenotype revealed by exome sequencing of human colon adenomas.* Cancer Res, 2012. **72**(23): p. 6279-89.

83. Cancer Genome Atlas Research, N., *Comprehensive genomic characterization defines human glioblastoma genes and core pathways.* Nature, 2008. **455**(7216): p. 1061-8.

84. Huang da, W., B.T. Sherman, and R.A. Lempicki, *Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists.* Nucleic Acids Res, 2009. **37**(1): p. 1-13.

85. Ogata, H., et al., *KEGG: Kyoto Encyclopedia of Genes and Genomes.* Nucleic Acids Res, 1999. **27**(1): p. 29-34.

86. Nishimura, D., *Biocarta.* Biotech Software & Internet Report, 2001. **2**(3).

87. Croft, D., et al., *Reactome: a database of reactions, pathways and biological processes.* Nucleic Acids Res, 2011. **39**(Database issue): p. D691-7.

88. http://www.sidra.ibge.gov.br/bda/tabela/protabl.asp?c=262&i=P&nome=on&notarodape=on&tab=262&unit=0&pov=3&opc1=1&poc2=1&OpcTipoNivt=1&opn1=2&nivt=0&orc86=3&poc1=1&orp=6&qtu3=27&opv=1&poc86=2&sec1=0&opc2=1&pop=1&opn2=0&orv=2&orc2=5&qtu2=5&sev=93&sev=1000093&opc86=1&sec2=0&opp=1&opn3=0&sec86=0&sec86=2776&sec86=2777&sec86=2779&sec86=2778&sec86=2780&sec86=2781&ascendente=on&sep=43344&orn=1&qtu7=9&orc1=4&qtu1=1&cabec=on&pon=1&OpcCara=44&proc=1&opn7=0&decm=99.

89. http://www.inec.go.cr/Web/Home/pagPrincipal.aspx.

90. Marcheco-Teruel, B., et al., *Cuba: exploring the history of admixture and the genetic basis of pigmentation using autosomal and uniparental markers.* PLoS genetics, 2014. **10**(7): p. e1004488.

91. http://www2.census.gov/geo/maps/dc10_thematic/2010_Profile/2010_Profile_Map_Puerto_Rico.pdf.

92. http://quickfacts.census.gov/qfd/states/00000.html.

93. Pham-Huy, L.A., H. He, and C. Pham-Huy, *Free radicals, antioxidants in disease and health.* International journal of biomedical science: IJBS, 2008. **4**(2): p. 89.

94. Esteller, M., *CpG island hypermethylation and tumor suppressor genes: a booming present, a brighter future.* Oncogene, 2002. **21**(35): p. 5427-40.

95. Crider, K.S., et al., *Folate and DNA methylation: a review of molecular mechanisms and the evidence for folate's role.* Adv Nutr, 2012. **3**(1): p. 21-38.

96. Ehrlich, M., *DNA methylation in cancer: too much, but also too little.* Oncogene, 2002. **21**(35): p. 5400-13.

97. Tsun, Z.Y. and R. Possemato, *Amino acid management in cancer.* Semin Cell Dev Biol, 2015. **43**: p. 22-32.

98. Warden, C.D., et al., *COHCAP: an integrative genomic pipeline for single-nucleotide resolution DNA methylation analysis.* Nucleic Acids Research, 2013. **41**(11): p. e117-e117.

99. Reed, M.C., et al., *A mathematical model of glutathione metabolism.* Theor Biol Med Model, 2008. **5**: p. 8.

100. Higham, D.J. and L.N. Trefethen, *Stiffness of odes.* BIT Numerical Mathematics, 1993. **33**(2): p. 285-303.

101. Figueiredo, J.C., et al., *Genes involved with folate uptake and distribution and their association with colorectal cancer risk.* Cancer causes & control : CCC, 2010. **21**(4): p. 597-608.

102. Sahoo, S., et al., *Membrane transporters in a human genome-scale metabolic knowledgebase and their implications for disease.* Front Physiol, 2014. **5**: p. 91.

103. Yang, A.S., et al., *A simple method for estimating global DNA methylation using bisulfite PCR of repetitive DNA elements.* Nucleic Acids Res, 2004. **32**(3): p. e38.

104. Bienert, G.P., J.K. Schjoerring, and T.P. Jahn, *Membrane transport of hydrogen peroxide.* Biochimica et Biophysica Acta (BBA) - Biomembranes, 2006. **1758**(8): p. 994-1003.

105. Richter, Y. and B. Fischer, *Nucleotides and inorganic phosphates as potential antioxidants.* J Biol Inorg Chem, 2006. **11**(8): p. 1063-74.

106. Imlay, J.A., S.M. Chin, and S. Linn, *Toxic DNA damage by hydrogen peroxide through the Fenton reaction in vivo and in vitro.* Science, 1988. **240**(4852): p. 640-2.

107. Greenman, C., et al., *Patterns of somatic mutation in human cancer genomes.* Nature, 2007. **446**(7132): p. 153-158.

108. De Carvalho, D.D., et al., *DNA methylation screening identifies driver epigenetic events of cancer cell survival.* Cancer cell, 2012. **21**(5): p. 655-667.

109. Hsu, P.P. and D.M. Sabatini, *Cancer cell metabolism: Warburg and beyond.* Cell, 2008. **134**(5): p. 703-707.

110. Mello Filho, A.C. and R. Meneghini, *In vivo formation of single-strand breaks in DNA by hydrogen peroxide is mediated by the Haber-Weiss reaction.* Biochimica et Biophysica Acta (BBA)-Gene Structure and Expression, 1984. **781**(1): p. 56-63.

111. Ray, G. and S.A. Husain, *Oxidants, antioxidants and carcinogenesis.* Indian J Exp Biol, 2002. **40**(11): p. 1213-32.

112. Sun, H., et al., *Targeting Fenton's reaction and its role in the regulation of cancer tissue's pH level: a computational approach (In preparation).* http://csbl.bmb.uga.edu/~zhangchi/FentonReaction/, 2016.

113. Subramanian, A., et al., *Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles.* Proceedings of the National Academy of Sciences of the United States of America, 2005. **102**(43): p. 15545-15550.

114. Fisher, R.A., *STATISTICAL METHODS FOR RESEARCH WORKERS.* Oliver and Boyd, 1925.

115. 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.* International Journal of Cancer, 2015. **136**(9): p. 2001-2011.

116. Zhang, C., et al., *Cancer may be a pathway to cell survival in response to persistent hypoxia and elevated ROS: A model for solid cancer initiation and early development.* International Journal of Cancer, 2014. **Accepted.**

117. Hanahan, D. and Robert A. Weinberg, *Hallmarks of Cancer: The Next Generation.* Cell, 2011. **144**(5): p. 646-674.

118. Wenger, R.H., *Cellular adaptation to hypoxia: O2-sensing protein hydroxylases, hypoxia-inducible transcription factors, and O2-regulated gene expression.* FASEB J, 2002. **16**(10): p. 1151-62.

119. Liu, Z.Q. and H.Y. Shan, *Cholesterol, not polyunsaturated fatty acids, is target molecule in oxidation induced by reactive oxygen species in membrane of human erythrocytes.* Cell Biochem Biophys. , 2006. **45**(2): p. 185-93.

120. Murphy, R.C. and K.M. Johnson, *Cholesterol, Reactive Oxygen Species, and the Formation of Biologically Active Mediators.* J Biol Chem, 2008. **283**(23): p. 15521-5.

121. Yamamoto, S., *"Enzymatic" lipid peroxidation: reactions of mammalian lipoxygenases.* Free radical biology & medicine, 1991. **10**(2): p. 149.

122. Spiteller, G., *Linoleic acid peroxidation—the dominant lipid peroxidation process in low density lipoprotein—and its relationship to chronic diseases.* Chemistry and Physics of Lipids, 1998. **95**(2): p. 105-162.

123. Brash, A.R., *Arachidonic acid as a bioactive molecule.* Journal of Clinical Investigation, 2001. **107**(11): p. 1339-1345.

124. Ferretti, G., et al., *HDL-paraoxonase and Membrane Lipid Peroxidation: A Comparison Between Healthy and Obese Subjects.* Obesity, 2010. **18**(6): p. 1079-1084.

125. Hagmann, H., et al., *Breaking the chain at the membrane: paraoxonase 2 counteracts lipid peroxidation at the plasma membrane.* FASEB J, 2014.

126. Esterbauer, H., R.J. Schaur, and H. Zollner, *Chemistry and biochemistry of 4-hydroxynonenal, malonaldehyde and related aldehydes.* Free Radical Biology and Medicine, 1991. **11**(1): p. 81-128.

127. Ishii, T., et al., *Role of Nrf2 in the Regulation of CD36 and Stress Protein Expression in Murine Macrophages: Activation by Oxidatively Modified LDL and 4-Hydroxynonenal.* Circulation Research, 2004. **94**(5): p. 609-616.

128. AOCS. *PLASMA LIPOPROTEINS COMPOSITION, STRUCTURE AND BIOCHEMISTRY.* Available at: http://lipidlibrary.aocs.org/Lipids/lipoprot/index.htm 2014.

129. Valacchi, G., et al., *Scavenger receptor class B type I: a multifunctional receptor.* Ann N Y Acad Sci., 2011. **1229**: p. E1-7.

130. Calvo, D., et al., *Human CD36 is a high affinity receptor for the native lipoproteins HDL, LDL, and VLDL.* J Lipid Res, 1998. **39**(4): p. 777-88.

131. Lopez, D. and M.P. McLean, *Sterol regulatory element-binding protein-1a binds to cis elements in the promoter of the rat high density lipoprotein receptor SR-BI gene.* Endocrinology, 1999. **140**(12): p. 5669-5681.

132. Sekiya, M., et al., *Oxidative stress induced lipid accumulation via SREBP1c activation in HepG2 cells.* Biochem Biophys Res Commun, 2008. **375**(4): p. 602-7.

133. Hughes, A.L., B.L. Todd, and P.J. Espenshade, *SREBP pathway responds to sterols and functions as an oxygen sensor in fission yeast.* Cell, 2005. **120**(6): p. 831-42.

134. Gurcel, L., et al., *Caspase-1 Activation of Lipid Metabolic Pathways in Response to Bacterial Pore-Forming Toxins Promotes Cell Survival.* Cell, 2006. **126**(6): p. 1135-1145.

135. Liang, K. and N.D. Vaziri, *Down-regulation of hepatic high-density lipoprotein receptor, SR-B1, in nephrotic syndrome.* Kidney Int. , 1999. **56**(2): p. 621-626.

136. Ghosh, M.G., D.A. Thompson, and R.J. Weigel, *PDZK1 and GREB1 are estrogen-regulated genes expressed in hormone-responsive breast cancer.* Cancer Res, 2000. **60**(22): p. 6367-75.

137. Tachibana, K., et al., *Regulation of the human PDZK1 expression by peroxisome proliferator-activated receptor alpha.* FEBS Lett, 2008. **582**(28): p. 3884-8.

138. Wang, N., et al., *ATP-binding cassette transporter A1 (ABCA1) functions as a cholesterol efflux regulatory protein.* Journal of Biological Chemistry, 2001. **276**(26): p. 23742-23747.

139. Debruyne, P.R., et al., *The role of bile acids in carcinogenesis.* Mutat Res, 2001. **480-481**: p. 359-69.

140. Bjorkhem, I., *Do oxysterols control cholesterol homeostasis?* Journal of Clinical Investigation, 2002. **110**(6): p. 725-730.

141. Diczfalusy, U., et al., *4β-hydroxycholesterol, an endogenous marker of CYP3A4/5 activity in humans.* British journal of clinical pharmacology, 2011. **71**(2): p. 183-189.

142. Nylén, H., *Studies on the oxysterols 4alpha-and 4beta-hydroxycholesterol.* 2011.

143. Umetani, M. and P.W. Shaul, *27-Hydroxycholesterol: the first identified endogenous SERM.* Trends Endocrinol Metab, 2011. **22**(4): p. 130-5.

144. Payne, A.H. and D.B. Hales, *Overview of Steroidogenic Enzymes in the Pathway from Cholesterol to Active Steroid Hormones.* Endocr Rev., 2004. **25**(6): p. 947-70.

145. Charman, M., et al., *MLN64 mediates egress of cholesterol from endosomes to mitochondria in the absence of functional Niemann-Pick Type C1 protein.* J Lipid Res, 2010. **51**(5): p. 1023-34.

146. Bose, H.S., et al., *StAR-like activity and molten globule behavior of StARD6, a male germ-line protein.* Biochemistry, 2008. **47**(8): p. 2277-88.

147. Korytowski, W., et al., *Deleterious cholesterol hydroperoxide trafficking in steroidogenic acute regulatory (StAR) protein-expressing MA-10 Leydig cells: implications for oxidative stress-impaired steroidogenesis.* J Biol Chem, 2013. **288**(16): p. 11509-19.

148. Olefsky, J.M., *Nuclear Receptor Minireview Series.* Journal of Biological Chemistry, 2001. **276**(40): p. 36863-36864.

149. Quinn, C.M., et al., *Expression and regulation of sterol 27-hydroxylase (CYP27A1) in human macrophages: a role for RXR and PPARgamma ligands.* Biochem J, 2005. **385**(Pt 3): p. 823-30.

150. Honkakoski, P. and M. Negishi, *Regulation of cytochrome P450 (CYP) genes by nuclear receptors.* Biochem J, 2000. **347**(Pt 2): p. 321-37.

151. Val, P., et al., *SF-1 a key player in the development and differentiation of steroidogenic tissues.* Nucl Recept, 2003. **1**(1): p. 8.

152. Beato, M. and J. Klug, *Steroid hormone receptors: an update.* Hum Reprod Update, 2000. **6**(3): p. 225-36.

153. Lala, D.S., et al., *Activation of the orphan nuclear receptor steroidogenic factor 1 by oxysterols.* Proc Natl Acad Sci U S A, 1997. **94**(10): p. 4895-900.

154. Andrew, J. and W. Jessup, *Oxysterols and atherosclerosis.* Atherosclerosis, 1999. **142**(1): p. 1-28.

155. DuSell, C.D., et al., *27-Hydroxycholesterol Is an Endogenous Selective Estrogen Receptor Modulator.* Mol Endocrinol., 2008. **22**(1): p. 65–77.

156. Fu, X., et al., *27-hydroxycholesterol is an endogenous ligand for liver X receptor in cholesterol-loaded cells.* Journal of Biological Chemistry, 2001. **276**(42): p. 38378-38387.

157. Ory, D.S., *Nuclear receptor signaling in the control of cholesterol homeostasis: have the orphans found a home?* Circ Res, 2004. **95**(7): p. 660-70.

158. Bodin, K., et al., *Metabolism of 4β-Hydroxycholesterol in Humans.* Journal of Biological Chemistry, 2002. **277**(35): p. 31534-31540.

159. Yarden, Y., *The EGFR family and its ligands in human cancer: signalling mechanisms and therapeutic opportunities.* European Journal of Cancer, 2001. **37, Supplement 4**(0): p. 3-8.

160. Plowright, E.E., et al., *Ectopic expression of fibroblast growth factor receptor 3 promotes myeloma cell proliferation and prevents apoptosis.* Blood, 2000. **95**(3): p. 992-8.

161. Inglis-Broadgate, S.L., et al., *FGFR3 regulates brain size by controlling progenitor cell proliferation and apoptosis during embryonic development.* Dev Biol, 2005. **279**(1): p. 73-85.

162. Stachowiak, E.K., et al., *Nuclear accumulation of fibroblast growth factor receptors in human glial cells--association with cell proliferation.* Oncogene, 1997. **14**(18): p. 2201-11.

163. Shay, J.W., et al., *Telomerase and cancer.* Human Molecular Genetics, 2001. **10**(7): p. 677-685.

164. Bottaro, D.P., et al., *Identification of the Hepatocyte Growth Factor Receptor as the c-met Proto-Oncogene Product.* Science, 1991. **251**(4995): p. 802-804.

165. Funk, J.O., *Cell cycle checkpoint genes and cancer.* eLS, 2005.

166. Langfelder, P. and S. Horvath, *Eigengene networks for studying the relationships between co-expression modules.* Bmc Systems Biology, 2007. **1**.

167. Fogh, J., W.C. Wright, and J.D. Loveless, *Absence of HeLa cell contamination in 169 cell lines derived from human tumors.* J Natl Cancer Inst, 1977. **58**(2): p. 209-14.

168. Campbell, A. and S.P. Chan, *Mitochondrial membrane cholesterol, the voltage dependent anion channel (VDAC), and the Warburg effect.* Journal of Bioenergetics and Biomembranes, 2008. **40**(3): p. 193-197.

169. Fu, X., et al., *27-Hydroxycholesterol is an endogenous ligand for LXR in cholesterol-loaded cells.* Journal of Biological Chemistry, 2001.

170. Edgar, R., M. Domrachev, and A.E. Lash, *Gene Expression Omnibus: NCBI gene expression and hybridization array data repository.* Nucleic Acids Res, 2002. **30**(1): p. 207-10.

171. Zhang, Y., et al., *Reversal of P-glycoprotein-mediated multi-drug resistance by the E3 ubiquitin ligase Cbl-b in human gastric adenocarcinoma cells.* J Pathol, 2009. **218**(2): p. 248-55.

172. Holm, S., *A simple sequentially rejective multiple test procedure.* Scandinavian journal of statistics, 1979: p. 65-70.

173. 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.

174. Chiang, J.Y., *Bile acids: regulation of synthesis.* J Lipid Res, 2009. **50**(10): p. 1955-66.

175. Quinet, E.M., et al., *LXR ligand lowers LDL cholesterol in primates, is lipid neutral in hamster, and reduces atherosclerosis in mouse.* Journal of Lipid Research, 2009. **50**(12): p. 2358-2370.

176. Wu, X.L., et al., *FGF19-induced Hepatocyte Proliferation Is Mediated through FGFR4 Activation.* Journal of Biological Chemistry, 2010. **285**(8): p. 5165-5170.

177. Liu, S.Q., et al., *Endocrine protection of ischemic myocardium by FGF21 from the liver and adipose tissue.* Sci Rep, 2013. **3**: p. 2767.

178. Cyphert, H.A., et al., *Activation of the farnesoid X receptor induces hepatic expression and secretion of fibroblast growth factor 21.* Journal of Biological Chemistry, 2012. **287**(30): p. 25123-25138.

179. Pignon, J., et al., *Androgen receptor controls EGFR and ERBB2 gene expression at different levels in prostate cancer cell lines.* Cancer Res, 2009. **69**(7): p. 2941-9.

180. JavanMoghaddam, S., *Cell Cycle Regulatory Roles of Estrogen Receptor alpha (ERα) in Breast Cancer Cells.* 2011.

181. Schiewer, M.J., M.A. Augello, and K.E. Knudsen, *The AR dependent cell cycle: Mechanisms and cancer relevance.* Mol Cell Endocrinol, 2012. **352**(1-2): p. 34-45.

182. Razandi, M., et al., *Proximal events in signaling by plasma membrane estrogen receptors.* J Biol Chem, 2003. **278**(4): p. 2701-12.

183. Jiang, J.-G., et al., *Transcriptional Regulation of the Hepatocyte Growth Factor Gene by the Nuclear Receptors Chicken Ovalbumin Upstream Promoter Transcription Factor and Estrogen Receptor.* Journal of Biological Chemistry, 1997. **272**(7): p. 3928-3934.

184. Park, M., et al., *Sequence of MET protooncogene cDNA has features characteristic of the tyrosine kinase family of growth-factor receptors.* Proceedings of the National Academy of Sciences, 1987. **84**(18): p. 6379-6383.

185. Lue, N. and C. Autexier, *Telomerases: Chemistry, Biology and Clinical Applications*. 2012, WILEY. 2002.

186. Smith, L.L., H.A. Coller, and J.M. Roberts, *Telomerase modulates expression of growth-controlling genes and enhances cell proliferation.* Nat Cell Biol., 2003. **5**(5): p. 474-479.

187. Shou, J., et al., *Mechanisms of Tamoxifen Resistance: Increased Estrogen Receptor-HER2/neu Cross-Talk in ER/HER2–Positive Breast Cancer.* Journal of the National Cancer Institute, 2004. **96**(12): p. 926-935.

188. Ni, M., et al., *Targeting androgen receptor in estrogen receptor-negative breast cancer.* Cancer Cell, 2011. **20**(1): p. 119-31.

189. Li, X. and B.W. O'Malley, *Unfolding the action of progesterone receptors.* J Biol Chem, 2003. **278**(41): p. 39261-4.

190. Deblois, G., et al., *Transcriptional Control of the ERBB2 Amplicon by ERRα and PGC-1β Promotes Mammary Gland Tumorigenesis.* Cancer Research, 2010. **70**(24): p. 10277-10287.

191. Torres, C.G., et al., *27-hydroxycholesterol induces the transition of MCF7 cells into a mesenchymal phenotype.* Oncology reports, 2011. **26**(2): p. 389-397.

192. Werneburg, N.W., et al., *Bile acids activate EGF receptor via a TGF-alpha-dependent mechanism in human cholangiocyte cell lines.* Am J Physiol Gastrointest Liver Physiol, 2003. **285**(1): p. G31-6.

193. Harrell, J.C., et al., *Genomic analysis identifies unique signatures predictive of brain, lung, and liver relapse.* Breast Cancer Res Treat, 2012. **132**(2): p. 523-35.

194. Cai, C., et al., *ERG induces androgen receptor-mediated regulation of SOX9 in prostate cancer.* J Clin Invest, 2013. **123**(3): p. 1109-22.

195. McMullin, R.P., et al., *A BRCA1 deficient-like signature is enriched in breast cancer brain metastases and predicts DNA damage-induced poly (ADP-ribose) polymerase inhibitor sensitivity.* Breast Cancer Res, 2014. **16**(2): p. R25.

196. Stange, D.E., et al., *Expression of an ASCL2 related stem cell signature and IGF2 in colorectal cancer liver metastases with 11p15.5 gain.* Gut, 2010. **59**(9): p. 1236-1244.

197. Sheffer, M., et al., *Association of survival and disease progression with chromosomal instability: a genomic exploration of colorectal cancer.* Proc Natl Acad Sci U S A, 2009. **106**(17): p. 7131-6.

198. Del Rio, M., et al., *Gene expression signature in advanced colorectal cancer patients select drugs and response for the use of leucovorin, fluorouracil, and irinotecan.* J Clin Oncol, 2007. **25**(7): p. 773-80.

199. Ki, D.H., et al., *Whole genome analysis for liver metastasis gene signatures in colorectal cancer.* International Journal of Cancer, 2007. **121**(9): p. 2005-2012.

200. Chaika, N.V., et al., *Differential expression of metabolic genes in tumor and stromal components of primary and metastatic loci in pancreatic adenocarcinoma.* PLoS One, 2012. **7**(3): p. e32996.

201. Van den Broeck, A., et al., *Molecular markers associated with outcome and metastasis in human pancreatic cancer.* J Exp Clin Cancer Res, 2012. **31**: p. 68.

202. Chandran, U.R., et al., *Gene expression profiles of prostate cancer reveal involvement of multiple molecular pathways in the metastatic process.* BMC Cancer, 2007. **7**: p. 64.

203. Poisson, L., J. Taylor, and D. Ghosh, *Integrative set enrichment testing for multiple omics platforms.* BMC Bioinformatics, 2011. **12**(1): p. 459.

204. Fritsche-Guenther, R., et al., *Therapeutic potential of CAMPATH-1H in skeletal tumours.* Histopathology, 2010. **57**(6): p. 851-861.

205. Wuttig, D., et al., *CD31, EDNRB and TSPAN7 are promising prognostic markers in clear-cell renal cell carcinoma revealed by genome-wide expression analyses of primary tumors and metastases.* International Journal of Cancer, 2012. **131**(5): p. E693-E704.

206. Zhang, X.H., et al., *Latent bone metastasis in breast cancer tied to Src-dependent survival signals.* Cancer Cell, 2009. **16**(1): p. 67-78.

207. Kimbung, S., et al., *Claudin-2 is an independent negative prognostic factor in breast cancer and specifically predicts early liver recurrences.* Mol Oncol, 2014. **8**(1): p. 119-28.

208. Tobin, N.P., et al., *Molecular subtype and tumor characteristics of breast cancer metastases as assessed by gene expression significantly influence patient post-relapse survival.* Ann Oncol, 2015. **26**(1): p. 81-8.

209. Meyniel, J.P., et al., *A genomic and transcriptomic approach for a differential diagnosis between primary and secondary ovarian carcinomas in patients with a previous history of breast cancer.* BMC Cancer, 2010. **10**: p. 222.

210. Mathew, R., V. Karantza-Wadsworth, and E. White, *Role of autophagy in cancer.* Nat Rev Cancer, 2007. **7**(12): p. 961-7.

211. Chi Zhang, T.S., Sha Cao, Samira Issa-Boube, Tongyu Tang, Xiwen Zhu, Ning Dong, Wei Du, Ying Xu, *Autophagy in Cancer Cells vs Cancer Tissues: two different stories*. [Book] Targeting Autophagy in Cancer Therapy, ed. J. Yang. 2016: Springer.

212. Dengjel, J., et al., *Autophagy promotes MHC class II presentation of peptides from intracellular source proteins.* Proc Natl Acad Sci U S A, 2005. **102**(22): p. 7922-7.

213. Mathew, R., V. Karantza-Wadsworth, and E. White, *Assessing metabolic stress and autophagy status in epithelial tumors.* Methods Enzymol, 2009. **453**: p. 53-81.

214. Balkwill, F.R., M. Capasso, and T. Hagemann, *The tumor microenvironment at a glance.* J Cell Sci, 2012. **125**(Pt 23): p. 5591-6.

215. Whiteside, T.L., *The tumor microenvironment and its role in promoting tumor growth.* Oncogene, 2008. **27**(45): p. 5904-12.

216. 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.

217. Chin, L., et al., *Making sense of cancer genomic data.* Genes Dev, 2011. **25**(6): p. 534-55.

218. Miller, J., et al., *Strategies for aggregating gene expression data: The collapseRows R function.* BMC Bioinformatics, 2011. **12**(1): p. 322.

219. Yadav, V.K. and S. De, *An assessment of computational methods for estimating purity and clonality using genomic data derived from heterogeneous tumor tissue samples.* Brief Bioinform, 2015. **16**(2): p. 232-41.

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

221. Cichocki, A., R. Zdunek, and S.-i. Amari. *Hierarchical ALS algorithms for nonnegative matrix and 3D tensor factorization*. in *International Conference on Independent Component Analysis and Signal Separation*. 2007. Springer.

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

223. Cao, S., et al., *Oxidized Cholesterol Plays a Key Role in Driving the Rapid Growth of Metastatic Cancer.* (In preparation), 2016.

224. Cao, S., et al., *Competition Regulation among DNA Methylation, Nucleotide Synthesis and Anti-Oxidation in Cancer vs. Normal Tissues.* . (submitted), 2016.

225. Cao, S., F. Yao, and Y. Xu, *De-convolution of tissue-based gene-expression data to cell type specific contributions and application to cancer tissue gene-expression data analyses.* (In preparation), 2016.

226. Cao, S., et al., *Two major contributors to cancer tissue-based gene-expression data: biological functions and anti-oxidation, and application to reliable prediction of gene-expression levels of protein complexes* (In preparation), 2016.

227. Cao, S., Y. Zhou, and Y. Xu, *Overcoming of micro-environmental stresses on methylation level regulation in cancer cells* (In preparation), 2016.

228. Cao, S., C. Zhang, and Y. Xu, *Somatic mutations may not be the primary drivers of cancer formation.* International Journal of Cancer, 2015. **137**(11): p. 2762-2765.

229. Zhang, C., et al., *Elucidation of drivers of high-level production of lactates throughout a cancer development.* Journal of molecular cell biology, 2015. **7**(3): p. 267-279.

230. Liu, C., et al., *Stresses drive a cancer's initiation, progression and metastasis: Critical comments on the book" Cancer Bioinformatics".* Journal of bioinformatics and computational biology, 2015. **13**(02): p. 1571002.

231. Zhang, C., et al., *Computational analysis of the impact of Autophagy in different stages of cancer progression*, in *Autophagy and Cancer*, J.-M. Yang, Editor. 2016, Springer: Springer.

232. Song, T., et al., *The Method for Breast Cancer Grade Prediction and Pathway Analysis Based on Improved Multiple Kernel Learning.* Journal of Bioinformatics and Computational Biology, 2016. **Accepted.**

233. Coothankandaswamy, V., et al., *Amino acid transporter SLC6A14 is a novel and effective drug target for pancreatic cancer.* Br J Pharmacol, 2016.

234. Cui, J., et al., *Comprehensive characterization of the genomic alterations in human gastric cancer.* International journal of cancer, 2015. **137**(1): p. 86-95.

235. Zhang, C., et al., *LTMG-QB: A probabilistic model based biclustering method for single cell transcriptomic data analysis.* (Ready for submission), 2016.

236. Chou, W.-C., et al., *Analysis of strand-specific RNA-seq data using machine learning reveals the structures of transcription units in Clostridium thermocellum.* Nucleic acids research, 2015. **43**(10): p. e67-e67.