THESIS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

The Origin of Symmetry in the Metabolism of Cancer

From Systems Biology to Translational Medicine

FRANCESCO GATTO

Department of Biology and Biological Engineering
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2015

The Origin of Symmetry in the Metabolism of Cancer
From Systems Biology to Translational Medicine
FRANCESCO GATTO
ISBN 978-91-7597-225-1

Doktorsavhandligar vid Chalmers tekniska högskola
Ny serie nr 3906
ISSN 0346-718X

Department of Biology and Biological Engineering
Chalmers University of Technology
SE-412 96 Gothenburg
Sweden
Telephone +46 (0)31-772 1000

Cover:
Artistic view of the symmetry in metabolism in cancer cells [Own work]

Printed by Chalmers Reproservice
Gothenburg, Sweden 2015

The Origin of Symmetry in the Metabolism of Cancer

From Systems Biology to Translational Medicine

FRANCESCO GATTO
Department of Biology and Biological Engineering
Chalmers University of Technology

ABSTRACT

Why do we not have a cure for cancer yet? Cancer is the malady of the century, the most intensely studied disease of all time. The question is puzzling. It assumes that cancer is a single entity that we can target and eradicate. On the contrary, the current theory on the origin of cancer dictates that each patient bears a cancer that is an exquisite experiment of nature, in which a unique constellation of genetic aberrations confers malignant traits upon the cell, which enable it to proliferate abnormally and survive until death of the host. Nevertheless, the question is legitimate. Cancer is also a single entity because, in spite of the heterogeneity of origins, every individual cancer in its evolution ought to converge in the acquisition of the same malignant traits, e.g. abnormal proliferation and ability to metastasize. I define this phenomenon of convergent evolution as the symmetry of cancer and each of these traits as symmetric, reminiscent of the fact that as diverse as two individual cancers can be in their origin, they can be repositioned along the trait to be identical. This thesis is dedicated to understanding the origin of symmetry of cancer through systems biology. In particular, I focused my interest in a specific malignant trait, the reprogramming of cell metabolism. In order to undertake an unbiased view of this complex system, I adopted a systems level perspective, in which genome-scale changes of gene and protein expression (so-called omics) attributable to cancer were bridged with the network of reactions that form the backbone of human metabolism.

First, we found that any cancer seemed to acquire a symmetric overexpression of nucleotide metabolism, regardless from where it originated (Paper I). However, this symmetry seems rather to represent an adaptation to a metabolic requirement of cellular proliferation and not an obligate metabolic reprogramming to foster cancer evolution. Hence, we sought to characterize those gene expression changes occurring in presence of an oncogenic mutation, again irrespective of the tissue of origin or other confounding factors (Paper II). This analysis revealed that oncogenic mutations independently converge on the deregulation of a sub-network revolving around the metabolism of arachidonic acid and xenobiotics, which we termed AraX.

Second, the symmetry of cancer metabolism broke with the most common form of kidney cancer, clear cell renal cell carcinoma (ccRCC). We reported that a ccRCC-specific set of genetic aberrations is associated with the emergence of a uniquely compromised metabolic network (Paper I). These outstanding features of ccRCC metabolism provided an opportunity for translational medicine. We proved that it is possible to exploit the ccRCC defective network to computationally predict metabolic liabilities that induce selective cell death in ccRCC (Paper III). Moreover, these changes in metabolic regulation unique to ccRCC can be distilled, through an algorithm of our creation, Kiwi (Paper V), in a coordinated regulation of glycosaminoglycan biosynthesis (GAGs) (Paper IV). This is mirrored by an altered profile of GAGs in kidney-proximal fluids, urine and blood, that we prove bearing a strong, accurate, and robust diagnostic value in metastatic ccRCC.

The case of ccRCC and a potential role of inflammation in AraX may raise more doubt than support on the existence of symmetry in the metabolic reprogramming in any cancer cell (Paper VI). Perhaps researchers are simply observing an enhanced plasticity in the adaptation to ever-changing conditions that is induced by mutations, but which is not symmetric under any specific trait and as such not essential to cancer. Yet, I argue that the quest for searching for the symmetry in cancer should not be abandoned. This quest is of paramount importance to unlock the discovery of a cure for cancer.

Keywords: cancer metabolism; omics; systems biology; network modeling; genome-scale metabolic modeling; kidney cancer.

List of publications

This thesis is based on the work contained in the following papers, referred to as Paper I to VI in the text:

I. F. Gatto, I. Nookaew, J. Nielsen, Chromosome 3p loss of heterozygosity is associated with a unique metabolic network in clear cell renal carcinoma.
Proceedings of the National Academy of Sciences of the United States of America **111**, E866-875 (2014).

II. F. Gatto, A. Schulze, J. Nielsen, Metabolic reprogramming resulting from oncogenic mutations converges in the deregulation of arachidonate and xenobiotics metabolism.
Submitted for publication (2015).

III. F. Gatto*, H. Miess*, A. Schulze, J. Nielsen, Flux balance analysis predicts essential genes in clear cell renal cell carcinoma metabolism.
Scientific Reports **5**, 10738 (2015).

IV. F. Gatto, et al., Changes in glycosaminoglycan biosynthesis regulation in clear cell renal cell carcinoma are detectable in patients' plasma and urine and can diagnose metastasis.
Submitted for publication (2015).

V. L. Varemo*, F. Gatto*, J. Nielsen, Kiwi: a tool for integration and visualization of network topology and gene-set analysis.
BMC Bioinformatics **15**, 408 (2014).

VI. F. Gatto, J. Nielsen, In search for symmetries in cancer metabolism.
Submitted for publication (2015).

Additional publications during the doctoral studies not included in this thesis:

VII. A. Mardinoglu, F. Gatto, J. Nielsen, Genome-scale modeling of human metabolism - a systems biology approach.
Biotechnology Journal, **8(9)**, 985-96 (2013).

VIII. F. Mannello, F. Maccari, D. Ligi, M. Santi, F. Gatto, R. J. Linhardt, F. Galeotti, N. Volpi, Breast cyst fluid heparan sulphate is distinctively N-sulphated depending on apocrine or flattened type.
Cell Biochem Funct, **33(3)**, 128-33 (2015).

IX. K. Thorell, Y. Andersson, F. Gatto, et al., Transcriptome analysis reveals differential expression of kynurenine pathway enzymes in Helicobacter pylori-induced gastric inflammation.
Manuscript in preparation (2015).

X. C. Lyssniotis, F. Gatto, et al., Glutamic-oxaloacetic transaminase 1 is essential for the metabolic reprogramming of pancreatic adenocarcinoma.
Manuscript in preparation (2015).

*Authors contributed equally to this work.

Contribution to publications

I. Conceived and designed the study, performed the computational analyses, wrote the paper.

II. Conceived and designed the study, performed the computational analyses, wrote the paper.

III. Conceived and designed the study, performed the computational analyses, wrote the paper.

IV. Conceived and designed the study, performed the computational analyses, wrote the paper.

V. Co-conceived and co-designed the study, co-developed the software.

VI. Wrote the paper.

VII. Contributed to the manuscript draft.

VIII. Performed statistical analyses.

IX. Contributed to the study design, performed computational analyses.

X. Contributed to the study design, performed computational analyses.

Preface

This dissertation is submitted for the partial fulfillment of the degree of doctor of philosophy. It is based on work carried out between March 2012 and July 2015 in the Systems and Synthetic Biology group, Department of Biology and Biological Engineering, Chalmers University of Technology under the supervision of Professor Jens Nielsen. The research was funded by the Knut and Alice Wallenberg Foundation and the Chalmers Foundation.

Francesco Gatto

August 2015

Table of contents

List of figures and tables

Figure copyright

All figures obtained from referred publications were reproduced with permission of the Publisher. Other attributions as follow:

Figure 1-5 - Published with permission of the author and the Publisher. Original source: Perpiñá Tordera M. Complexity in Asthma: Inflammation and Scale-Free Networks. Arch Bronconeumol. 2009;45(9):459-65. © 2008 SEPAR. Published by Elsevier España, S.L. All rights reserved.

Figure 3-1 - "Kidney tubules" by JWSchmidt at en.wikipedia - Transferred from en.wikipedia. Licensed under CC BY-SA 3.0 via Wikimedia Commons - http://commons.wikimedia.org/wiki/File:Kidney_tubules.png#/media/File:Kidney_tubules.png

”*They are the instruments on which we play,*
and what is an instrument without somebody to play on it?”

W.S. Maugham

Abbreviations and symbols

2HG	2-hydroxyglutarate
ATP	Adenosine triphosphate
COSMIC	Catalogue of somatic mutations in cancer
CS	Chondroitin sulfate
FBA	Flux balance analysis
GAG	Glycosaminoglycan
GEM	Genome-scale metabolic model
GO	Gene Ontology consortium
HCV	Human C-virus
HMR	Human metabolic reaction database
HPV	Human papilloma virus
HS	Heparan sulfate
KEGG	Kyoto Encyclopedia of Genes and Genomes
NGS	Next-generation sequencing
PCA	Principal component analysis
PFG-A/B	Posterior fossa group A/B, ependymomas subtypes
(m)(cc)RCC	(Metastatic) (Clear cell) Renal cell carcinoma
ROS	Reactive oxygen species
RSV	Reus sarcoma virus
SNP	Single nucleotide polymorphism
TCA	Tricarboxylic acid cycle

Nomenclature
Genes and proteins in *H. sapiens* are designated following standard nomenclature, capital letters in italics for genes and capital letters (unformatted) for proteins, e.g. *IDH1* and IDH1 respectively.

Chapter I: Introduction to origin of symmetry in cancer

Long before I started my doctoral studies, a question used to whirl my mind every now and then, probably coincidental with the occurrence of a cancer in a person somewhat related to me. Why do we not have a cure for cancer yet? In my regular citizen mindset, the word *cancer* was mostly associated with a lethal condition, of obscure origin, heavily studied and yet poorly treated (by means of puzzling and only partially understood therapies, such as chemo or radio). Now, from a scientist perspective, I figured that our communication to the public on what cancer is and how we deal with it is appalling. Nowadays, a cancer diagnosis does not necessarily correspond to a death sentence. Since money has poured in cancer research, our progresses in understanding this phenomenon have been overwhelming. Lastly, despite the justified caution adopted by scientists when making predictions about cancer, the perspective of an end to cancer does seem at hand. I strongly discourage to interpret these three statements as if cancer was not a foremost hurdle for public health. I feel compelled to report that this pathology currently represents the second leading cause of death worldwide (*1*). Nevertheless, the public perception of cancer appears to be distorted compared to the outstanding scientific and medical advances accomplished in the last decades of research (*2, 3*).

In this introduction, I elaborate the three statements above in the attempt to revert our slanted view of cancer. Hopefully, a clearer picture will emerge. Yet, incomplete. And these knowledge gaps provide the rationale of my contribution to the science of cancer.

The origin of symmetry – From Rous to Weinberg

In 1910, Frances Peyton Rous made an extraordinary discovery (*4*). For the first time in history, Rous reported an origin to cancer[1]. Until then, cancer was observed and described by a number of physicians and scientists, but the reasons beyond its occurrence were elusive and attributed to factors fairly creative in hindsight (Probably the first description of cancer dates back to the 25th century BC, in an ancient Egyptian papyrus (*5*). Curiously, the Egyptians practiced a "surgical" removal of the abnormal mass to treat the disease. Tumor surgery is regarded still today as the most effective treatment for cancer.) In a landmark experiment, Rous extracted a connective tissue tumor (in medical terms, a sarcoma) from a chicken (Fig. 1-1). The chicken sarcoma was grinded and injected to a new host, i.e. an otherwise healthy chicken. Of key importance, Rous filtered the grinded sarcoma prior to injection. The new host developed a sarcoma itself. Given that the "inducing agent" was small enough to pass a filter, Rous concluded that the agent was a virus. The virus went down in history under the name of Rous sarcoma virus (RSV). In 1910, the origin of cancer was found to be a virus.

Figure 1-1. A chicken affected with sarcoma, used as a model organism by Frances Peyton Rous in his experiment to understand the cause of cancer. Reproduced from (*4*).

[1] Always elusive to me was the distinction between the word *cancer* and *tumor*. A tumor is an abnormal mass of cells growing aberrantly in a tissue, such as the pancreas. Cancer is the malignant manifestation of a tumor, in which cells have the potential to migrate from the tissue of origin and colonize distant tissues, a process known as metastasis.

The modern theory on the origin of cancers draws from pivotal experiments by many influential scientists after Rous, most notably Katsusaburo Yamagiwa, and Koichi Ichikawa at Tokyo University (*6*). Despite the hype generated by Rous discovery (Ludwik Gross, a distinguished virologist, would declare at one point that *it would be rather difficult to assume a fundamentally different etiology [than the viral origin] for human tumors*), today we recognize that essentially two tumor types are induced by viruses in humans, namely cervical carcinomas (caused by the human papilloma virus, HPV) and hepatomas (cause by the hepatitis C-virus, HCV). Nevertheless, Rous' experiment was epochal in that it provided a reproducible cause of cancer, paving the way to a clearer understanding on its origin. Yamagiwa's and Ichikawa's contributions follow the same line of argument, when they demonstrated that cancer could be induced by exposure to specific substances (today termed carcinogens). On the clinical side, the implications of these discoveries bore enormous potential. Surgery could be finally replaced by a treatment effectively eliminating the cause of cancer (e.g. the virus), rather than the consequence (e.g. the abnormal tumoral mass)[2].

A major revolution in biology was the resolution of the DNA structure by Francis Crick, Rosalind Franklin, James Watson, and Maurice Wilkins in the 1950s (*7*) and the role of genes in defining the molecular basis of cells. This increased knowledge on genetics enabled to discover that the ability of RSV to cause cancer could be ascribed to a single gene (termed *v-src*), which is carried by the virus. Unexpectedly, the probe that was designed to target the *v-src* sequence recognized instead a DNA sequence in uninfected cells. It became apparent that RSV hijacks and alters a regular human gene to confer cancer properties to the host cell. Importantly, Bruce Ames at University of California, Berkeley subsequently demonstrated that also carcinogens act in a similar fashion. They alter the sequence of genes on the human DNA thereby transforming normal cells into cancer cells (*8*). A permanent change in the DNA sequence is called a mutation, and by the 1970s, the origin of cancer was found to be mutations in the human DNA.

Since these crucial discoveries, a race towards a complete mapping of the mutations at the origin of cancer started in the 1980s. Contrary to initial beliefs, the number of genes in which mutations were linked to carcinogenesis was not just a handful; conversely it seemed possible to constantly finding new cancer-causing mutations. As of April 2015, COSMIC (Catalogue Of Somatic Mutations In Cancer), a manually curated database for cancer mutations, lists 572 genes in which a mutation is *causally* implicated in cancer (*9*). Thanks to significant advances in the sequencing technology, in particular the advent of next-generation sequencing (NGS), it has been recently estimated that even though we have possibly discovered all the most commonly mutated genes, the number of genes that are rarely mutated yet potentially implicated in cancer will continue to rise in the future (*10*). The long-waited promise that a straightforward mechanism could explain the origin of cancer, as postulated when a sole mutation as simple as a single nucleotide polymorphism (SNP) in the *HRAS* gene was found to cause bladder cancer (*11*), was doomed to be abandoned forever. In 1976, Peter Nowell advanced the theory of Darwinian evolution to explain the heterogeneity of mutations observed in different cancers (*12*). He postulated that in a population of normal cells subject to mutations in cancer-associated genes (such as those listed in COSMIC), only those cells (clones) that survive and in turn gain a selective growth advantage compared to the surrounding normal cells have the potential to seed a tumor. In particular, bearing a mutation in these key genes arms the transformed cells with aberrant properties typical of cancer. For example, the above-mentioned

[2] This historical event is reminiscent of those explanations that my former professor of thermodynamics used to provide to justify our interest in this otherwise intangible subject. In the 1920s, it was believed that the *cause* of the incomplete yield in ethanol observed in industrial distillations was the limited volume of reactors, which would not allow enough time for the separation to take place. The engineers at the time designed and concocted increasingly bigger reactors (or abstrusely shaped), with the frustrating result that no higher concentration of alcohol was ever observed. Simple thermodynamics would have determined that the vapor mixture had reached equilibrium, so no further separation between water and ethanol would ever be possible. Nothing beats thermodynamics.

HRAS mutation in bladder cancer permanently alters the structure of the encoded protein, GTPase HRas or the transforming protein p21. p21 normally responds to stimulation by growth factors, which ultimately results in cell division. However, the mutated p21 is altered in such way that it is constitutively activated, regardless of the presence of growth factors, hence eliciting aberrant cell division. On the other hand, Nowell's application of Darwinian evolution to cancer dictates that the environment eventually selects for the mutated clones that initiate a tumor. Indeed, *HRAS* mutations in tissues other than the bladder do not necessarily result in cancer. Already in the 1970s, this powerful concept framed the emerging heterogeneity of cancer mutations in the context of a simple theory, clonal evolution. Drawing from this, in 1982 Bert Vogelstein described the pathogenesis of colorectal cancer as the result of a precise succession of mutational events (Fig. 1-2). Even though this description is today considered an over-simplification, this proof-of-concept laid the basis for the modern theory on the origin of cancer. The modern theory on the origin of cancer prescribes that every tumor is a unique experiment of nature, in which mutations in key genes (also called driver mutations) together with accumulation of mutations in secondary genes (also called passenger mutations) define the fitness of a cancer clone in its strive to survive and proliferate in the host environment, namely the human body (*13, 14*).

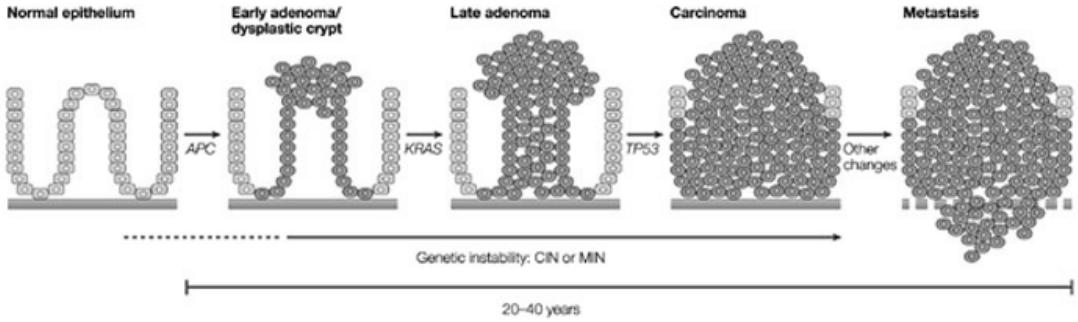


Figure 1-2. Vogelstein's model of mutational events that correlate with each step in the adenoma–carcinoma sequence of colorectal cancerogenesis. Reproduced from (*15*).

In spite of its elegance, the theory of clonal evolution is daunting under a clinical perspective. In recent years, NGS-aided sequencing of an increasing number of cancer genomes has confirmed the broad extent of genomic heterogeneity of the disease. Genomic heterogeneity was shown to encompass both localized and distal tumors in the same patient (*16-18*). Simply put, the fact that every tumor seems to stem from a unique combination of mutations (despite their recurrences) renders a rational treatment of the disease hardly possible. Indeed, the argument of genomic heterogeneity in cancer is one of the pillars in the foundation of personalized (or precision) medicine (*19*). Precision medicine advocates tailoring a treatment to the defined molecular features of a patient's cancer. As for 2015, we are still far from individualization of medical treatments in the clinics, even though notable proofs-of-concepts have been published in the literature (*20*). More properly, the current application of precision medicine is patient stratification, which is not exactly a novel concept. On the contrary, patient stratification is among the oldest approaches used today in cancer treatment (and yet poorly communicated to the general public). The most widespread form of cancer patient stratification is based on histopathology, or in other words, on the tissue of origin and the appearance of the tumor under the lens of a microscope. Under the perspective of patient stratification, the question "Why do we not have a cure for cancer?" looks inevitably dull: a cancer classified as testis cancer is curable in 95% of patients; conversely a triple-negative breast cancer has among the lowest survival rates. I insist on this point because I realized that patient stratification based simply on histopathology (also commonly referred to as the cancer type) is likely the most relevant carrier of information on the biology of cancer itself (*21*), possibly questioning the basic arguments of a dramatically more complex philosophy such as precision medicine.

In this fashion, the need to disentangle cancer heterogeneity by means of detailed patient stratification has been questioned recently (*22*). Using the same arguments as precision medicine,

but reversed, one could argue that, in the end, some characteristics are clearly shared by all cancers despite the genomic heterogeneity. Phenotypic traits as aberrant proliferation and invasion are observed in virtually all cancer types. In a landmark review in 2000, Douglas Hanahan and Robert Weinberg described six phenotypic traits that all cancers seem to acquire, which they called the hallmarks of cancer (*23*) (this list was eventually expanded in 2011 to comprise four emerging traits, in Figure 1-3 (*24*)). A similar argument was put forward also by other researchers in the same year, like Gerard Evan and Karen Voudsen (*25*). In 2004, Vogelstein concluded that the origin of cancer itself must explain the convergence on these phenotypic traits, e.g. the number of mutations found to drive cancer is higher than the number of pathways altered by the mutations (the pathways being ultimately responsible for the acquisition of the cancer phenotype) (*26*). This argument stood solid after the advent of NGS, which enabled the whole genome sequencing of thousands of tumor DNA (*27*). The biological phenomenon in which some phenotypic traits are gained by means of distinct genetic trajectories is known as convergent evolution. In this thesis, I will refer to it also as **convergence** or **symmetry**, reminiscent of the fact that any two cancers as different as they may appear may always be repositioned to look symmetric according to a phenotypic trait. Let me elaborate this definition with an example (*28*). If one were to observe a multiple myeloma (a cancer of plasma cells) and a glioblastoma multiforme (a primary brain tumor), it is easy to recognize how *different* they manifest themselves. Multiple myeloma is typically diagnosed following symptoms of renal failure, anemia, etc, while the most common symptom of glioblastoma multiforme is a neurological deficit or the like. Multiple myeloma is a liquid tumor, glioblastoma multiforme is a solid tumor. Multiple myeloma is more common among African Americans, glioblastoma multiforme has higher incidence among Caucasians, Hispanics, and Asians. Notwithstanding these differences, a multiple myeloma and a glioblastoma multiforme have always a trait in common: an abnormal accumulation of cells due to uncontrolled cell proliferation. If one discarded all other traits and observed only this specific trait, then the two tumors would appear identical. Abnormal proliferation defines an axis of symmetry for cancer, i.e. it is a symmetric trait.

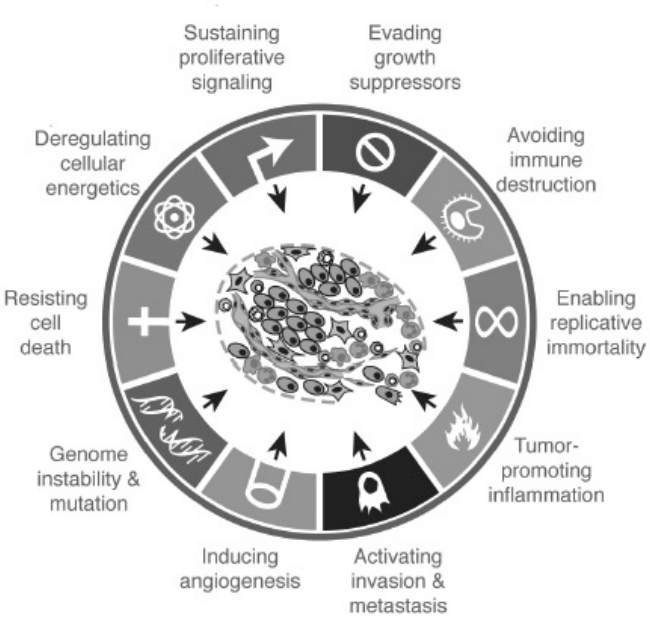


Figure 1-3. The hallmarks of cancer as described by Hanahan and Weinberg in 2011. The two researchers proposed that virtually every cancer acquires these phenotypic traits. Reproduced from (*24*).

NGS however also highlighted some discrepancies. For example, in a whole-genome sequencing study of 183 lung adenocarcinomas, the extent to which mutations could explain the acquisition of the hallmarks was fairly limited (*29*). This conclusion raises some fundamental questions in the role of mutations and the hallmarks. There are indeed other biological mechanisms different than mutations that drive cancer, most notably copy number alterations. Alternatively, it could be that

not all hallmarks are actually consequential the acquisition of mutations, but rather a collateral manifestation that promotes but is not essential to the evolution of cancer. Or perhaps we are still far from knowing the true origin of symmetry in cancer.

Inspired by the philosophy of Weinberg (*22*), I focused my doctoral studies in the search for simplicity in cancer[3]. The contribution of mutations to the origin of symmetry in cancer has therefore become the central question of my research.

Banquet – From Warburg to Cantley

The early NGS era delivered a number of surprises in the search of mutations recurrent in cancer (and as such candidate to elucidate its origin). In two studies published between 2008 and 2009, an unexpected mutation in the cytosolic NADP+-dependent isocitrate dehydrogenase 1 gene (*IDH1*) was found in 12% of glioblastoma multiforme and in 8% of acute myeloid leukemia (a blood cell tumor) (*30, 31*). *IDH1* encodes for an enzyme responsible to catalyze the metabolic reaction that converts isocitrate to 2-oxoglutarate. The discovery of *IDH1* mutations in a significant proportion of cancer cells provided among the first direct connections between the origin of cancer and deregulation of metabolism. Later research would demonstrate that the primary feature being selected for in these tumors is the ability of mutated *IDH1* to produce a distinct metabolite different from 2-oxoglutarate, named 2-hydroxyglutarate (2HG) (*32*). 2HG in turns interferes with cell differentiation and epigenetics to promote cancer progression (*33*).

Altered metabolism, and in particular cellular energetics, has been regarded a phenotypic trait of cancer since long before the discovery of recurrent *IDH1* mutations. The first reports that cancers may reprogram metabolic fluxes to match different requirements for proliferations were published by Otto Warburg in 1920s. Warburg noted that, even in the presence of oxygen, cultured cancer cell lines prefer to ferment glucose into lactic acid, rather than undergo a complete oxidation through the tricarboxylic acid (TCA) cycle, which is more favorable in terms of ATP yield (a phenomenon dubbed aerobic glycolysis or Warburg effect). These repeated observations led him to conclude in 1956 that the origin of cancer stands in the preference of aerobic glycolysis by malignant cells as opposed to normal cells (*34*). As noted above, this theory has been today replaced, in particular considering that aerobic glycolysis has been observed also in normal proliferating cells (*35*). Nevertheless, the importance of metabolism, as prospected by Warburg, has been considerably revived in the last decade. Indeed, besides the discovery of recurrently mutated metabolic enzymes such as *IDH1* (though the list comprises other TCA enzymes, like succinate dehydrogenase in ovarian cancer or fumarate hydratase in renal cell carcinoma (*36*)), evidence started accumulating that implicates renowned driver mutations with metabolic reprogramming. In 2008, an influential discovery occurred in Lewis Cantley's laboratory when the last step of glycolysis, catalyzed by pyruvate kinase (PK), was associated with the expression of the enzyme isoform, PKM2, specific to cancer cells (*37*). The authors proved that PKM2 expression is necessary for aerobic glycolysis, which in turn provides a selective growth advantage to the cancer cell. This discovery provided a first mechanistic explanation for the Warburg effect. PKM2, counter intuitively, limits the flux of glucose-derived carbon to pyruvate. The effect of limiting PK flux results in an increase of the carbon pool for those pathways branching from glycolysis, which are typically anabolic (e.g., nucleotide biosynthesis). Matthew Vander Heiden et al. argued that the increased utilization of glycolysis under normoxia (in other words, the Warburg effect) to boost anabolic pathways is the phenotypic trait being selected for in cancer, as it (intuitively) correlates with the greater metabolic requirements of a proliferating cell (*38*). Many reports challenged this theory, proving instead that the major effect of aerobic glycolysis is to provide reductive power (mostly in the form of

[3] Despite being inspired by his philosophy, I should say that Weinberg does not share a bit of my optimism towards the idea of simplicity (but he argues even stronger against dull complexity). Moreover, he spends bitter words towards systems biology, which, on the other hand, constitutes the spinal cord of the methods I adopted in my research.

NADPH), which cancer cells utilize to maintain a constant pool of reactive oxygen species (ROS) (*39*). In 2011, Rob Cairns et al. revisited metabolic reprogramming in cancer as a mechanism essential to cell survival, rather than growth (*40*). As for today, it is still unclear whether metabolic reprogramming due to cell survival is predominant with respect to meeting the requirements of proliferation. The increased flux in a given metabolic pathway enabled by aerobic glycolysis seems to be context-dependent, attributable to either the cancer type or the evolutionary stage or the microenvironment. Lindsey Boroughs and Ralph Deberardinis recently acknowledged that we might just be observing a different angle of the overwhelming plasticity of cancer, i.e. the ability of malignant cells to rapidly adapt their metabolism to changes in their genome and in their environment (*41*). In this sense, my early research attempted to elucidate the extent to which metabolic reprogramming is symmetric in cancer.

Despite the complexity of metabolic reprogramming, a clear view emerged in the last decade of molecular research. As much as it can be heterogeneous and context-dependent in its realization, the reprogramming of metabolism is still consequential an oncogenic event. For example, the acquisition of a mutation determines changes in metabolism that are either not observed in the parental cell or can be abrogated when the effects of the mutation are averted. As noted above, causally implicated mutations in cancer were found to drive unforeseen effects also at the metabolic level, besides their established role in controlling other cancer hallmarks. Voudsen and colleagues boldly noted that even the profoundly studied and intuitive implications of the most recurrently mutated gene in cancer, *TP53*, are less robust than the effects which *TP53* mutations have on metabolism (*42*). *TP53* is essential to arrest cell cycle and induce apoptosis in the event of genotoxic stress, and loss of these functions is widely regarded to be the key mechanism of cancer initiation. Nevertheless, mice with different *TP53* mutations failed to develop tumors even if the encoded protein effectively lost the above-mentioned functions (*43, 44*). At the same time, in these experiments the protein retained a similar metabolic control to the wild-type protein. This suggests that only *TP53* mutations that result in metabolic reprogramming are indeed oncogenic, and provides the argument for the centrality of metabolism proposed by Voudsen and colleagues. Many studies exploiting genetically engineered mice implicated metabolic reprogramming as a major consequence of the acquisition of common cancer-associated mutations, like *KRAS (45)* or *NFE2L2 (46)*.

Until now, the definition of metabolic reprogramming has been loosely defined as a rewiring of metabolic fluxes, without much detail on which fluxes and to which value. In general, I left this unspecified due to the above-mentioned complexity of observed patterns. Our view on metabolic reprogramming has evolved rapidly in the last years. Possibly, the only symmetry seems to be the Warburg effect. Upon transformation, *in vitro*, *in vivo,* and even *in situ* experiments agree that cancer cells increase the import of glucose and its conversion rate to lactate, seemingly regardless of oxygen levels. To complement this view, it should be acknowledged that increased aerobic glycolysis is associated with proliferating cells in general, and possibly a conserved metabolic switch in the evolution of human cells. Plus, there are cancers in which the Warburg effect is simply not observed. For example, a subset of melanomas defined by overexpression of PPARGC1A (also known as PGC1α) displays a distinctive metabolic state characterized by elevated mitochondrial respiration, as opposed to PGC1α-negative melanomas that are highly glycolytic (*47*). A similar observation regards diffuse large B cell lymphoma, where a tumor subset insensitive to inhibition of B cell receptor signaling also featured a higher rate of mithochondrial respiration (*48*). Besides aerobic glycolysis, molecular biology studies have reported other metabolic switches as symmetric in cancer (Fig. 1-4): aerobic glycolysis (*45, 49*), addiction to glutamine (*50, 51*), *de novo* lipogenesis (*52*), essentiality of one-carbon intermediates (*53*), reliance on autophagy and macropinocytosis (*54, 55*), reactive oxygen species homeostasis (*56*) and dependence on mitochondrial respiration (*57-59*). In line with above, none of these metabolic programs are universally shared among all cancer cells, yet most of these follow an oncogenic event (*60*). For

example, it was observed that the phosphoglycerate dehydrogenase (*PHGDH*) gene is essential in breast cancer cell lines. This gene encodes for the PHGDG enzyme that catalyzes the first step of *de novo* biosynthesis of serine, a non essential amino acid, from 3-phosphoglycerate, a glycolytic intermediate. Possemato and colleagues observed that the gene locus of PHGDH resides in a region of recurrent copy gain in breast cancer, and this correlates with dependence on this pathway for cancer survival (*61*). To follow on the essentiality of this pathway: Voudsen lab proved that *TP53* deficient tumors depends on extracellular serine to control oxidative stress; Zhang et al. demonstrated that initiation of non small lung cell carcinoma relies on glycine metabolism to sustain pyrimidine biosynthesis; Kim et al. showed that serine hydroxymethyltransferase is essential for glioma survival (*62-64*). These researchers employed state-of-art technology to back their conclusions, but the observation that cancer cells have outstanding serine requirements is not exactly novel, and was first reported by Regan and colleagues in 1969[4,5] (*65*). A key difference between earlier and recent studies, however, is that the latter generally proved that the metabolic reprogramming is a product of oncogenesis, thereby providing both an insight in how cancer evolves and a therapeutic window for its treatment. The case of serine is one of the many examples of metabolic reprogramming in cancer uncovered within the last five years. I refer to some excellent reviews for a broader overview of other oncogenic events with consequences in cancer metabolism (*33, 38, 40, 66, 67*).

My interest in metabolism was fostered, perhaps trivially, by the expertise of my research group. Hence, during my doctoral studies, I decided to focus on metabolism to exploit the accumulated knowledge in my group. Once again, the trend in the scientific literature seemed to suggest a personalized metabolic reprogramming depending on where and how the tumor formed. Nevertheless, metabolic reprogramming is an oncogenic event, and virtually all cancers rewire their metabolic fluxes either to survive or to proliferate (or both). In line with the arguments outlined in the previous section, a key objective of my research was to detect if and how distinct cancers are symmetric in the reprogramming of metabolism, and how much of this symmetry should be attributed to an oncogenic event.

Standing in the way of control – From bioinformatics to systems biology

Cancer is a complex system. A complex system is any phenomenon with observable collective behaviors that emerge from the interactions of its components (*68*). An impressive example of a complex system is human society, in which the seamless interaction between billion of individuals has given rise to uncountable collective behaviors, like politics, religion, sport, and technology. The same interactions were responsible for the evolution of society and adaptation to different climates or changes due to natural disasters. In general, a complex system is described by five characteristics (*69*):

1. A network of interacting components, which define the system;
2. Emergence, i.e. the properties that a system possesses by means of its interactions;
3. Self-organization, i.e. the process by which order arises out of an unordered system via the spontaneous interactions of its components;
4. Evolution, i.e. the increase in a system's fitness over time through changes in the interactions or in the components;
5. Adaptability, i.e. the ability of the system to react to changes in the environment.

[4] Intriguingly, Regan et al. proposed to target serine hydratase or treatment with serine antimetabolites as a cure for leukemia, which is essentially the same strategy advocated by all the recent literature here cited.

[5] Even earlier (1959) was the article that described asparagine as the first non essential amino acid to become essential in cancer (see H. Eagle, Amino Acid Metabolism in Mammalian Cell Cultures. *Science* **130**, 432-437 (1959)).

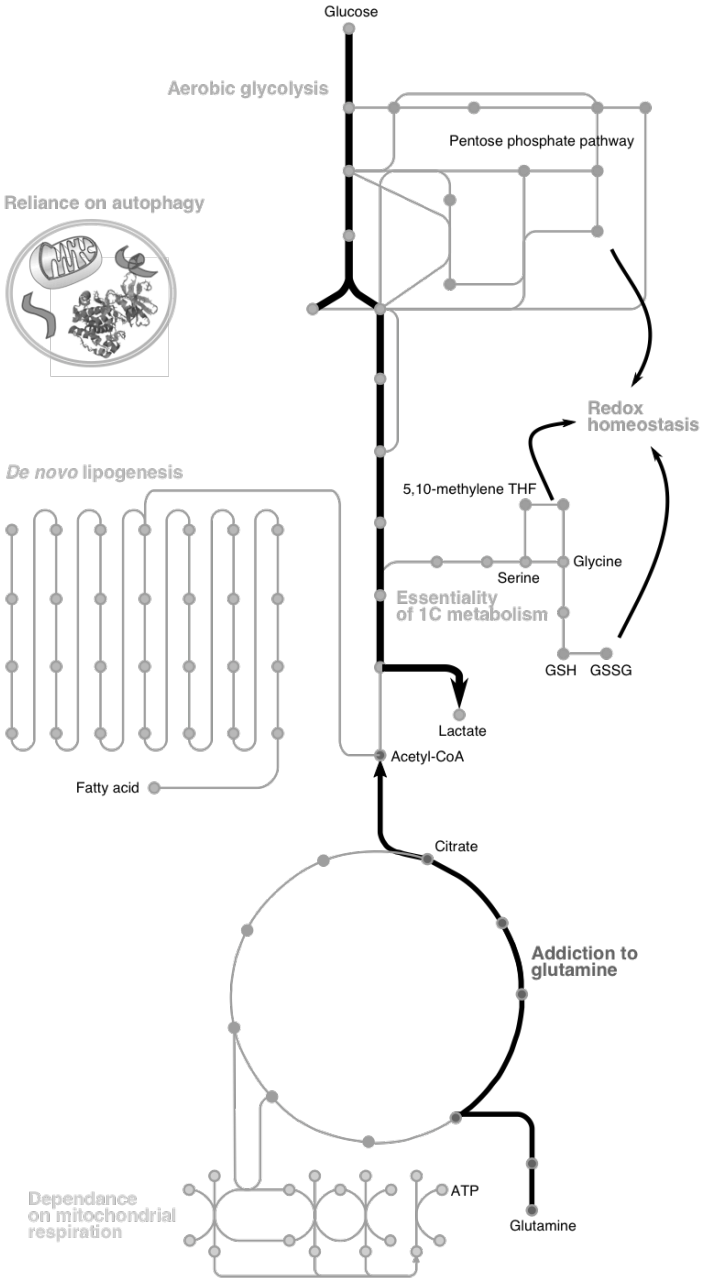


Figure 1-4. Traits of metabolic reprogramming in cancer deemed symmetric in molecular biology. Key: 1C – One-carbon; GSH – Glutathione; GSSG – Glutathione disulfide; THF – Tetrahydrofolate.

Strikingly, very disparate phenomena in our reality can all be classified as complex systems. Swarming is the collective behavior of some animal species in which aggregation typically results in migratory fluxes. Urbanization is a shift of unrelated human beings from rural zones to cities that results in the concentration of goods and services in a relatively small area. Consciousness is putatively emerging from the firing of billions of neurons in the brain with no single neuron directing the overall flow. Notwithstanding the unrelatedness of these systems, they are all the expression of a network. Among the discoveries that had the greatest impact in the science of 21st century stands the realization that the inherent structure of *all* these networks (in mathematical jargon, the topology) is not random. On the contrary, it originates from the same self-organizing principles. Finally, it enables the emergence of some outstanding properties that we all experience, but we rarely appreciate. If networks were purely random, two nodes have the same probability to share a link than any other pair of nodes in the network. The model that describes the origin of random networks was derived by Paul Erdős and Alfréd Rényi in 1959 (*70*). However, real networks are not random, despite the fact that there is no mastermind and no blueprint beyond the

origin of a link between a pair of nodes (e.g. two neurons in the brain). Indeed, Duncan J. Watts and Steven Strogatz realized that real networks featured the popular culture concept of six degrees of separation, i.e. the fact that it takes at most six steps to reach any two nodes in a network (also called small world property). However, such property is not observable in purely random networks described by the Erdős-Rényi model. Hence, in 1998, they developed a model in which a network can emerge randomly from its constituent nodes and yet feature the small world property as experienced in real networks (Watts-Strogatz model (*71*)). The Watts-Strogatz model represented an extraordinary advancement in our understanding on the origin of the networks that stand at the basis of many real complex systems. Nevertheless, it had a severe limitation, which sharply contrasted with one particular feature of real networks: it could not predict hubs, nodes (usually few) with an enormous amount of links compared to the rest. Hubs are everywhere in real networks. A few airports dominate the worldwide air traffic, while most airports only accommodate a handful of lines. In an online social network like Facebook, few users reach the upper limit of 5'000 friendships (yet the great majority of registered users have less than 100). Albert-László Barabási and Réka Albert comprehended that real networks emerge randomly following a power-law distribution, in which the probability to encounter a node with a given number of links (so called degree) decays with the number of limits to the power of an exponent that they found to range between 2 and 3:

$$p(k) = k^{\gamma}$$

where k is the node degree and γ is the exponent of the power law (Fig. 1-5). The fact that the degree decays with a power law rather than an exponential law permits the existence of few but meaningful hubs in the network. In their seminal paper published in 1999, they identified that these networks (named scale-free networks) form from the system nodes by means of just two mechanisms: preferential attachment and growth (*72*). Without getting into much detail, the Barabási-Albert model provides two simple explanations that determine the self-organizing principles beyond real networks. Surprisingly, many disparate complex systems result from networks that follow a power-law distribution. This observation led the researchers to claim that scale-free networks are universal in nature, or at least its properties (*73*).

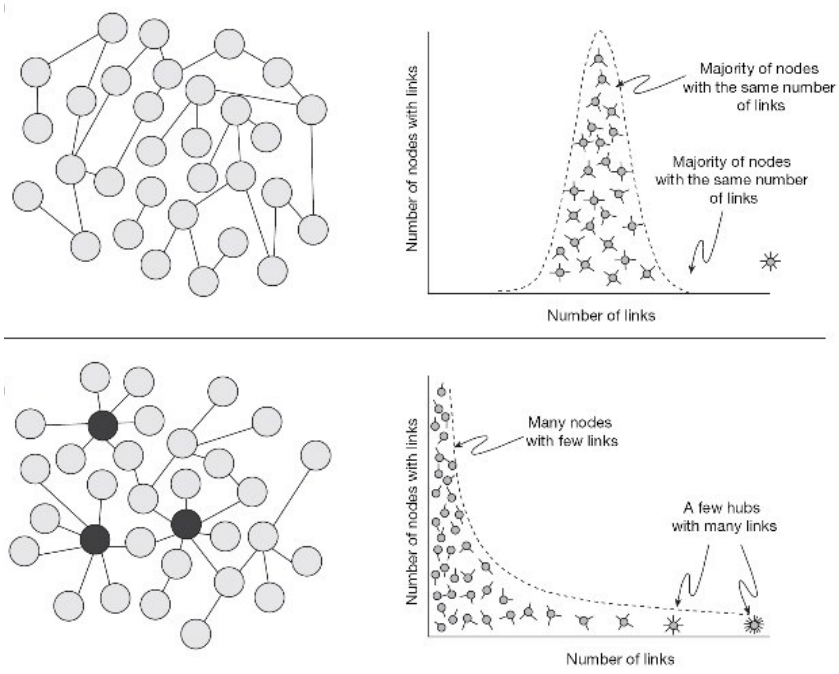


Figure 1-5. Random networks are described by the Erdős-Rényi model (on top), and prescribes that most nodes in the network have approximately the same number of links. Real networks, however, are better resembled by the Barabási-Albert model (bottom), which predicts the existence of few nodes that have a great share of all network links. Reproduced from (*74*).

Biology investigates a number of systems that qualify as complex, such as the cell, ecological communities, epidemics, and human diseases. Biologists observe and strive to explain the different behaviors in each of this system, and they have developed techniques and theories to describe them. I outlined a quite extensive introduction to complex systems just to come to one claim: all these systems share a common point in that they are all expression of a self-organized network, which defines, by emergence, the system's properties and functions (*75*). The area of biology that studies biological systems as the expression of interconnected networks of components is called systems biology. Systems biology is described as a holistic approach that aims at explaining behaviors of interest by modeling the interactions of *all* components within the system (*76*). This discipline is relatively recent, because its ambitions are only met by the simultaneous availability of three ingredients: a comprehensive description of the biological system as a network; large-scale experimental datasets that provide information on the components and/or the interactions in the network; a mathematical model that integrates data with the network thereby providing a platform to explain the system's collective behavior of interest.

Since my research questions revolve around two complex systems, cancer and metabolism, systems biology seemed the natural area of investigation. My work mostly focused on the mathematical modeling part. Most data used in my studies were retrieved from public repositories. They almost exclusively consist of omics data, i.e. the quantitative measurements of (possibly) all elements of the system belonging to a certain class (e.g. all transcripts in a cell, the transcript*ome*; or all genes in a cell, the gen*ome*). A greater availability of omics data was fostered, as mentioned earlier, by the NGS technology. The large scale of these data poses challenge in their storage, distribution, and, compellingly for my purpose, analysis. Bioinformatics is an interdisciplinary branch of computer science and biology that develops algorithms and software in order to treat and understand biological data. Thus, while systems biology laid the framework within which I tackled my research questions, bioinformatics provided the tools. I will not describe the details of the bioinformatics methods employed in the different papers chiefly for two reasons: first, they are widely established tools, from exploratory data analysis (e.g. principal component analysis) to differential expression analysis; second, they are question-dependent, by which I mean that, whenever possible, I adopted the available tool which I believed to confront a given research step best. On the other hand, I will spend some words on the network that defines my biological system of interest, which is the metabolic network. The comprehensive description of its nodes and links is captured by a relatively recent type of mathematical models, called genome-scale metabolic models (GEMs) (*77*).

Genome-scale metabolic models are, in a nutshell, metabolic networks where the comprehensive list of metabolites, reactions, and the associated genes are encoded so to be machine-readable (Fig. 1-6). A GEM of a given system (for example, a human cell) therefore attempts to include an exhaustive list of metabolic reactions occurring in that system, which is generally coded in the genome (hence the genome scale). GEMs are more than metabolic maps, because the represented network can be explored programmatically to answer more insightful questions about metabolism (*78*). In general, there are two main approaches to genome-scale metabolic modeling: topological analysis and simulations (*79*). Topological analysis refers to analyses that regard the properties of the network, for example a simple comparison of the number of nodes in two GEMs or the integration of omics data to highlight the importance of certain parts of the network in a condition. Topological analyses typically describe the metabolic network in different conditions. Simulations, on the other hand, require the formulation of GEMs as a mathematical model that, given some inputs, performs predictions. Virtually all simulations published in GEM literature yield predictions of metabolic fluxes, almost exclusively enabled by variations of a common theoretical framework, called flux balance analysis (FBA) (*80*). In brief, FBA returns distributions of metabolic fluxes such that the mass balance around each metabolite in the network is not violated (i.e. the sum of producing fluxes must equal the sum of consuming fluxes) and subject to some optimization

principle, for example that the flux distribution is wired towards the maximization of a metabolic task, like biomass growth. In systems biology, GEMs are arguably the only models that, first, have organized the unprecedented wealth of biochemical data on metabolism into almost complete metabolic networks; and second, entail a mathematical description of these networks that can be used both to extract information by integrating high-throughput data in the network topology and to simulate phenotypic flux distributions given some constraints (so called constraint-based reconstruction and analysis (*81*)). Applications of FBA flourished in the area of microbiology, and a number of success stories have been reported also for higher organisms (*82*).

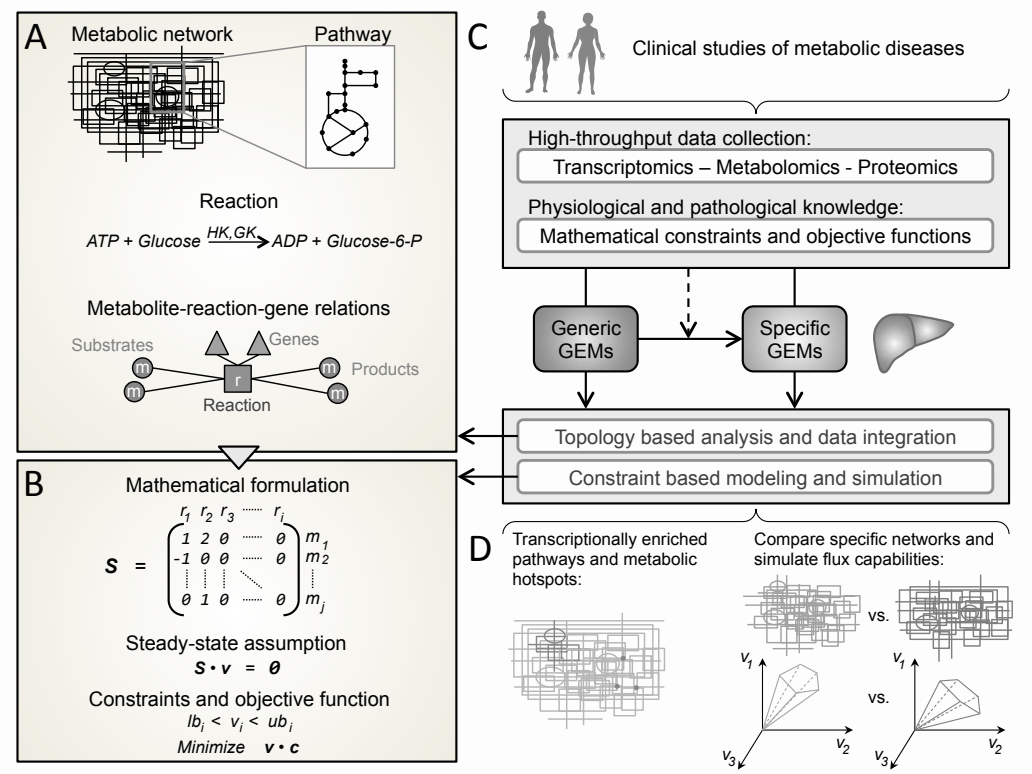


Figure 1-6. An overview on genome-scale metabolic models (GEMs) and their use in systems biology study of metabolic diseases. A) The human metabolic network is exhaustively compiled in terms of its metabolic reactions and their associations with gene products. B) The network is reformulated as a mathematical model, the GEM, in which each reaction is represented by a vector of stoichiometric coefficients according to which metabolites participate. Additional equations provide the standard formulation of a flux balance analysis (FBA) problem. C) Clinical studies provide the necessary knowledge to use the so generated GEM by generating genome-scale data about gene expression, protein expression, and metabolite levels in a physiologically and pathologically defined medical condition. GEMs exploit this knowledge to explore topological properties of metabolism or to simulate metabolic functions. D) Examples of such analyses is the extraction of parts of the metabolic network subject to significant transcriptional regulation (left) or the comparison of predicted flux distributions in two conditions, e.g. disease vs. healthy. Modified with permission from (*79*).

At the start of my PhD, in 2012, use of GEMs for human studies was at its infancy. Indeed, the first human GEMs were published in 2007, Recon1 and EHMN (*83, 84*). A more comprehensive GEM, HMR, was released in 2012, and the model was updated in 2014 as HMR2 (*85, 86*). Recon2, a community-based effort to update Recon1, was also published around the same time (*87*). Since 2011, the centrality of metabolic reprogramming in cancer induced researchers to elucidate for the first time global patterns in metabolism using GEMs (*88*). At the time of writing, six of these studies aimed at engineering cancer metabolism *in silico* so to uncover potential drug targets (*89-94*). The goal was thus to provide an applied use of GEMs for the clinics. In this sense, this research qualifies as translational medicine. Three publications described cancer metabolism at the genome-scale, using GEMs as a scaffold for integrating omics (*95-97*). As such, this research qualifies as systems biology. All these studies relied on topological analyses and simulations adapted from systems biology studies in lower organisms. However, when using human GEMs, a direct translation of FBA is hindered by the absence of an optimization principle for the fluxes. This lack

of tools prompted researchers to design novel methods to explore the space of flux distributions in human GEMs, like random sampling (*98*).

Despite my personal interest in non-applied biological questions, the thesis encompasses both systems biology and translational medicine studies. In the course of my doctorate I started two translational medicine projects (Paper III and IV) that spurred from a system biology study with promising consequences for the clinics (Paper I). The remaining two systems biology studies relied mostly on bioinformatics (Paper I and II). Finally, I contributed in the development of a tool to bridge bioinformatics with topological analysis of GEMs (Paper V).

Chapter II: Systems biology - Symmetries in cancer metabolism

Let us assume that the key of success for a cancer is held in the effective reprogramming of metabolism. Lessons from the other hallmarks of cancer, such as sustained proliferation, have taught researchers that multiple mutations can trigger the same cancer phenotype. Hence, it is reasonable to compare the phenotype of a cancer cell to the phenotype of its putative cell of origin, rather than comparing the mutations acquired along the transformation. Indeed, a change in phenotype, like the reprogramming of metabolism, is indicative that the process was important for the transformation of that cancer cell. A *recurrent* change in phenotype, which is a change independently observed in many unrelated cancer cells, is indicative that the process is important for the transformation of *any* cancer cell. We are required to take a look at all phenotypic traits that changed in many distinct cancer cells compared to their matched normal cells in order to identify those processes important for the transformation because a change was always observed. In other words, we need to test which traits are symmetric in cancer. This test has to be done at a global scale, by exhausting all possible phenotypic traits and all possible cancer cells.

At the current technological state, this test is not feasible. Hence, my collaborators and I applied a number of simplifications and assumptions to leverage on the available technologies. First and foremost, the phenotype of a cell is approximated by its transcriptome (in some studies, proteome). Second, the symmetric phenotypic trait is acquired by the mixed population of cancer and stromal cells that form a tumor rather than the individual cancer cells. Third, a change in phenotype occurs if the probability that the event took place randomly (according to certain assumptions regarding its probability distribution) is deemed too low. Finally, the space of phenotypic traits is restricted to those cellular processes characterized by scientific efforts such as the Gene Ontology Consortium (GO) (*99*), or the Kyoto Encyclopedia of Genes and Genomes (KEGG) (*100*).

With these assumptions in mind, I will explore the results of our first question: confined to metabolism, what phenotypic traits are symmetric in cancer?

Answer #1: Revolving around RNA

Metabolism is a complex system. It is the biological network of anabolic and catabolic reactions that transform nutrients into energy and building blocks for growth in a given organism. In humans, this network (also known as the metabolic network) emerges from interactions of 3765 gene products, according to the latest human GEM (*85*). At the time I first attempted to answer this question, however, the network was slightly smaller, and encoded by 3674 genes (*86*). I will refer to this set of genes as metabolic genes. Hence, if any of these genes show a consistent phenotypic change in all normal cells that transform into cancer, then we are observing the symmetry. In our first study (Paper I), we retrieved metabolic gene expression profiles (RNA-seq data, a NGS technology) from matched cancer-normal pairs, i.e. a resection of both the primary tumor and the healthy tissue was taken from the same patient. In total, 257 cancer-normal pairs, encompassing seven tissues, were evaluated. We measured similarity across samples in terms of metabolic gene expression, here regarded as metabolic phenotype readout, using principal component analysis and correlation-based hierarchical clustering. These analyses led us to three observations:

1. Most cancer samples retain a substantial similarity in the expression of metabolic genes with the normal tissue of origin;
2. If we focus on the subset of metabolic genes that displayed a strong change in expression in cancer vs. normal, then suddenly cancer samples are more similar to each other than to normal samples (Fig. 2-1);
3. A number of cancer samples have a clearly deviating expression of metabolic genes compared to any other sample (Fig. 2-1).

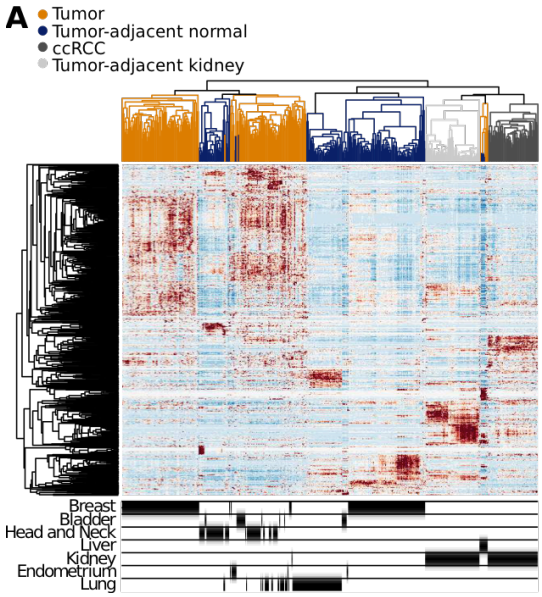


Figure 2-1. Similarity analysis of metabolic gene expression profiles for cancer and tumor-adjacent normal samples. Hierarchical clustering of absolute metabolic gene expression levels (RPKM) for cancer and tumor-adjacent normal samples, featuring only those genes that significantly changed expression across most cancer types upon transformation, thereby subtracting the effect of the tissue of origin. Bottom: corresponding tissue of origin for each sample in the heatmap above.

Observation #1 suggests that the extent of metabolic reprogramming during the transformation from normal to tumor is fairly limited. However, observation #2 argues that those metabolic genes whose expression is reprogrammed across most samples display a similar change, i.e. there is evidence of symmetry in the limited extent of metabolic reprogramming in cancer, regardless of the nature of the sample. Finally, observation #3 finds an exception to this symmetric behavior in a subset of samples. The strange case of the metabolism of these samples has fostered my curiosity and spurred a line of translational research on which I will come back to later in the thesis.

The question for the acquisition of the symmetry in cancer metabolism stimulated also the interest of other researchers. Hu et al. (*101*) also found that cancer retains the metabolic gene expression profile of the normal tissue of origin to a large extent. They calculate that if we assume that a normal tissue (say breast) is 100% dissimilar to another tissue (say prostate), than the corresponding cancers (i.e. breast and prostate cancer, respectively) are only 63% dissimilar on average to the normal tissue of origin, but 83% dissimilar to each other. Therefore, the researchers conclude that there is little place for symmetry in cancer metabolism or, in their words, *cancer-induced changes in the expression of metabolic genes are very heterogeneous across different tumor types*. However, they also report one clear symmetric behavior: most genes involved in the pathway of pyrimidine metabolism are up-regulated (i.e. there is an increase in gene expression) in cancer vs. normal. In the same fashion as Hu et al., we also tried to characterize the pathways beyond the (limited) metabolic reprogramming, though using a different approach. We recognized the importance of the tissue of origin in defining the expression of metabolic genes and thus, in order to find symmetric patterns in cancer, we computed changes in pathway expression for each cancer type separately and then detected if some pathways were consistently regulated across most types (Fig. 2-2). To this end, we adopted a stringent method to report pathway expression changes, called consensus gene-set analysis (*102*). Consistent with Hu et al., we also observe a substantial heterogeneity in the metabolic reprogramming of the different cancer types, with few recurrent patterns. As we conclude in Paper I, *cancer cells orchestrate the expression of metabolic genes in a similar fashion only when it comes to nucleotide, glutamate, and retinol metabolism*. Collectively, both studies agree that if any symmetry occurs in cancer metabolism, this should be searched in the metabolism of nucleotides, in particular pyrimidines.
As I noted in the beginning of this paragraph, the symmetry in cancer metabolism captivated other researchers. Shortly after Hu et al. and we published our results, Nilsson et al. also identified a

number of metabolic genes that all cancers coordinately regulate during the transformation (*103*). Even though the authors did not formally collect these genes into pathways, they report that the top regulated genes belong to glycolysis, anti-oxidant metabolism, glycosylation pathways, nucleotide and deoxynucleotide metabolism. Once again, regulation of nucleotide metabolism is deemed symmetric in cancer.

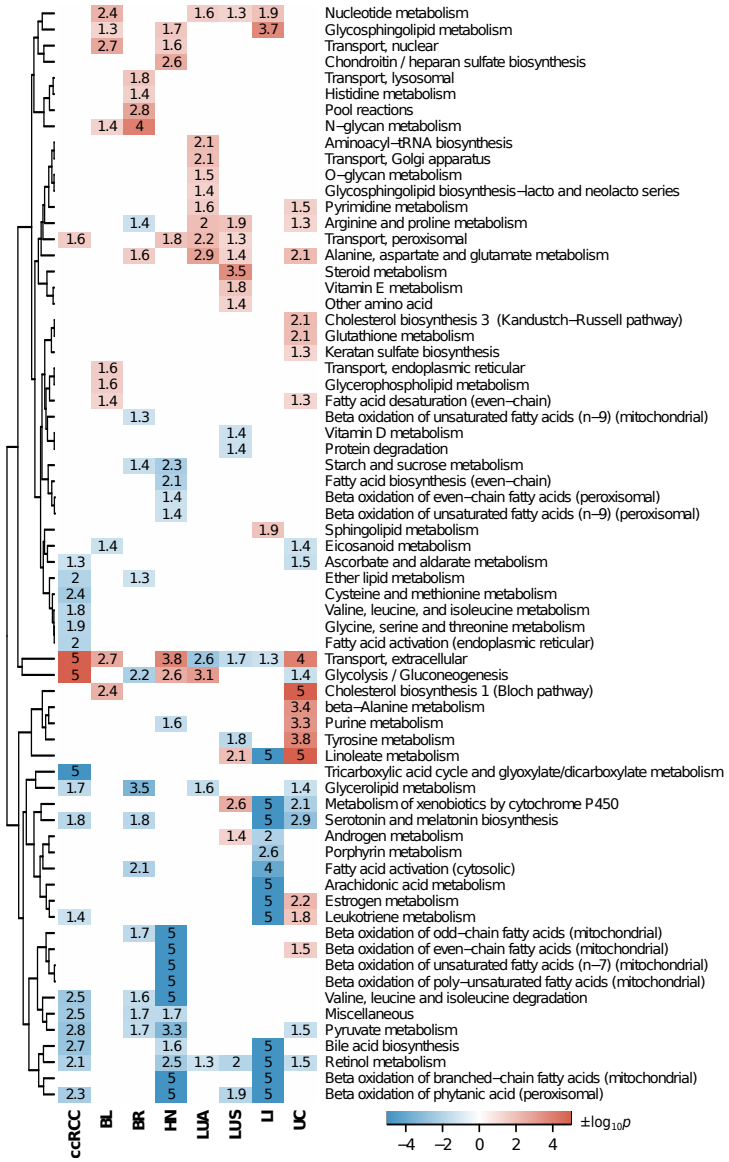


Figure 2-2. Regulation of canonical metabolic pathways in each cancer type according to changes in metabolic gene expression vs. matched tumor-adjacent normal tissues. Each box shows the $\log_{10} p$-value of the gene-set representing a pathway in a certain cancer type, and the color indicates the overall direction of gene expression regulation for the gene-set (red – up, blue – down). Key: ccRCC – Clear Cell Renal Cell Carcinoma; BL – Bladder Urothelial Carcinoma; BR – Breast Invasive Carcinoma; HN – Head and Neck Squamous Cell Carcinoma; LUA – Lung Adenocarcinoma; LUS – Lung Squamous Cell Carcinoma; LI – Liver Hepatocellular Carcinoma; UC – Uterine Corpus Endometrioid Carcinoma.

Taken together, three independent studies suggest that any tumor undertakes an obligate step during the transformation: the metabolism of nucleotides is up-regulated. Considering the global scale of these systems biology studies, one could even argue that nucleotide metabolism is the *only* coordinately regulated metabolic pathway at the transcriptional level in virtually all cancers or, in other words, that is symmetric in cancer. This conclusion is bold in light of observation #3. I will elaborate this discrepancy in the next chapter.

Also, this conclusion raises a number of questions, first and foremost: why only nucleotide metabolism? And within nucleotide metabolism, why preferentially pyrimidines? Why is it up-

regulated and what determines this increase in expression of nucleotide metabolism genes? Is this metabolic reprogramming an adaptive or an oncogenic process? Considering that these studies compared transformed proliferating cells in an abnormal microenvironment to wild-type and mostly quiescent cells in a physiological tissue, is this metabolic reprogramming a feature of cancer due to adaptation to the transformation or is it driven by the genetic aberrations at the origin of cancer, rendering the disease exposed to disruption in this process?

I do not have answers. These questions require mechanistic molecular studies that, on the other hand, are hard to tackle at the same scale as the systems biology studies here cited. However some speculations are due, which we discussed also in a later review (Paper VI). In proliferating cells, nucleotides are continuously synthesized to meet the increased requirements of RNA and DNA due to growth. Nevertheless, this argument should apply to all macromolecular classes (proteins, lipids, etc.) and not only RNA or DNA. Using the same logic, we should have observed up-regulation of the corresponding anabolic pathways as well. As a matter of fact, these other classes are more prominent in terms of cellular composition. In an average mammalian cell, DNA accounts for only 1% of dry weight, RNA slightly more, about 4%. For comparison, proteins accounts for 60%, and lipids for 18% (*104*). Nevertheless, the nucleotide-based macromolecular class of RNA is unique compared to other classes. Contrary to lipids and proteins, nucleotides cannot be readily scavenged from the extracellular environment; in other words, they are a rare nutrient for the cells. In the second place, the metabolism of RNA displays some distinctive patterns in terms of its relation with the growth rate. Experiments in bacteria have shown that when cells grow at higher rates, RNA levels exhibits the greatest relative change, in sharp contrast with DNA and proteins (*105*). This reflects the increasing concentration of ribosomes needed for higher protein synthesis with the growth rate. Whereas there is convincing evidence that DNA concentration is not limiting, the ribosome concentration and the corresponding protein synthesis rate are. Thus, already in 1983, Ehrenberg and Kurland proposed that an energy-efficient metabolic strategy for ribosomes at increasing growth rates is to proportionally increase both the substrate pool and the ribosome concentration (*106*). Since the most prominent phenotypic trait of a cancer cell is the growth advantage over the neighboring normal cells, this growth rate dependent effects on the cell macromolecular composition should be observed during the transformation. In this fashion, the here-in reported convergence on nucleotide metabolism up-regulation seems to support the model of Ehrenberg and Kurland. In other words, this convergence is probably not an oncogenic process, but an evolutionary conserved metabolic strategy that cells adopt when their growth rate increases. An alternative hypothesis that argues in favor of the oncogenic nature of nucleotide metabolism up-regulation may reside in DNA damage. More than (virtually) any other normal cell, cancer cells suffer of significant DNA damage (*107*), which leads to unregulated cell division, and the reincorporation of nucleosides during DNA reparation requires a sanitized pool of these metabolites (as a side note, notable tumor suppressor genes belong to this process as they function in the cellular mechanisms of DNA repair, like *BRCA*). In this case, up-regulation of nucleotide metabolism may serve as a cancer-specific reprogramming and explain the observed convergence.

Regardless of whether this process is adaptive or oncogenic, it is still symmetric in cancer. This fact opens a therapeutic window, a term that pharmacologists reserve to the range of dosages of a drug which is effective against a disease (in this case, all cancers) while being acceptably non toxic to normal cells. This therapeutic window has actually already been exploited, since long time. Among the most widespread treatments for cancers of various type is a chemotherapy consisting of so-called antimetabolites. These agents are analogs to human metabolites and they function by interfering with those reactions that use these metabolites as substrate. Gemcitabine, decitabine, and fluorouracil are such drugs. And they are *all* antimetabolites to nucleotides, and specifically pyrimidines. Their clinical use against a number of disparate cancers underscores the symmetry of nucleotide metabolism in cancer. However, it does not necessarily corroborate the above claim about its uniqueness in the landscape of metabolic pathways. Indeed, another prominent class of

antimetabolite drugs is represented by anti-folates (like methotrexate). This, in turn, can be reconciled to some findings reported by Nilsson and colleagues. In their study, they reported that the methylenetetrahydrofolate dehydrogenase (NADP+ dependent) 2 gene (*MTHFD2)* is the most frequently overexpressed in cancer and, more in general, so is the related mitochondrial folate pathway. Taken together, the direct observations in human tumors of the efficacy of inhibition of such (and perhaps *only* such) metabolic pathways seem to support their symmetric character in cancer metabolism. But the fact that cancer can relapse after antimetabolite treatment is also indicative that such symmetric regulation is yet circumventable, hence suggesting that the process is rather adaptive then oncogenic.

The above studies failed to dissipate my beliefs on the existence of one or more metabolic processes that all cancers must reprogram to evolve. Cancer cells are abnormal mutants. They evolve essentially through genomic instability (*14*). Their ever-changing genome exposes them to a high chance that their survival fitness will decrease, in other words that evolution will select *against* their existence. This perfect recipe for death can be reconciled with the evidence of cancer only by a ruthless application of the modern theory for clonal evolution in cancer. The genetic aberrations themselves must simultaneously enable cancer proliferation *and* survival. These symmetries, as well as any other symmetric trait in metabolism, must emerge in connection to the appearance of an oncogenic genetic alteration, most likely a mutation. So what are these phenotypic traits? Which symmetries are oncogenic?

Answer #2: AraX

The posed question cannot be readily explored experimentally. It would require triggering a mutation in a human cell in a healthy tissue capable of driving the neoplasm and measuring the gene expression changes before and after the mutation took place. This should be repeated for a sufficient number of driver mutations and in a sufficient number of backgrounds (i.e. tissues), so that only changes in common to all scenarios can be deemed symmetric. Experimental models such as cancer cell lines or mice models would likely fail to convey a realistic picture, given the role of the human microenvironment in the selection of the processes important for cancer evolution among those enabled by the mutation. Indeed, as much as a cell can attain a growth advantage by means of a mutation also in a Petri dish (one of the historical findings of Howard Temin (*108*)), it is the environment that ultimately selects the mutations conferring the fittest context-dependent growth advantage. In this study (Paper II), we proposed to estimate what are the gene expression changes attributable to the occurrence of a mutation in a human tumor. Considering the nature of the data used in this study, i.e. NGS-derived genomics and transcriptomics of human tumor resections, we held all assumptions elaborated to answer the first question in Paper I. In addition, the estimates for the gene expression changes were inferred by assuming that the expression level observed of a gene can be factorized as the sum of contributions of three distinct features of a tumor: the histopathological classification (e.g. its tissue of origin and morphology); the expression level of a set of validated transcription factors; and finally the factor of our interest, the presence of a mutation in a cancer-associated gene. We retrieved these pieces of information for 1082 tumors, encompassing 13 distinct cancer types.

The fundamentals for this study thus fall within the scope of a typical bioinformatics analysis known as gene expression analysis, that started off with microarray (*109*) and is nowadays almost exclusively performed using NGS in human systems (*110*). We adapted gene expression analysis to identify the gene expression changes attributable to the presence of a mutation, besides the above-mentioned factors, and this was achieved by leveraging on two statistical frameworks: generalized linear modeling (*111*) and model selection (*112*). A first look at the results of the analysis revealed that most gene expression occurring in a tumor could be ascribed to histopathology (as exemplified by the fact that a principal component analysis (PCA) of the 1082 gene expression profiles cluster very well according to the cancer type, Fig. 2-3). This is consistent with what was reported above.

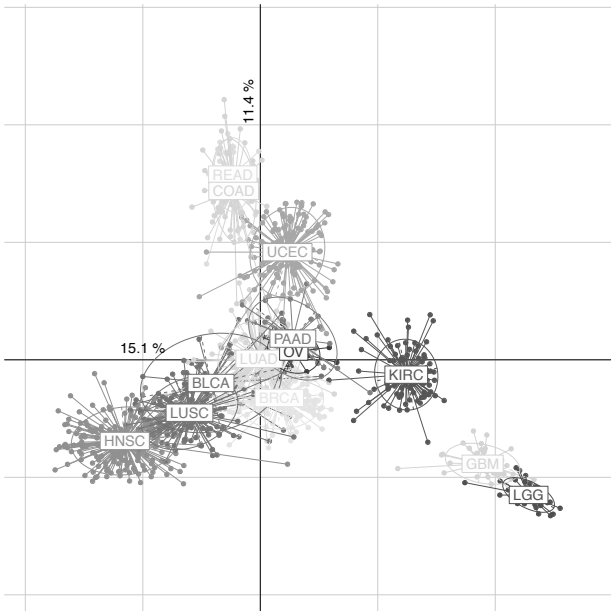


Figure 2-3. Principal component analysis of the 1082 primary tumor samples based on their gene expression profile. Samples were grouped according to the cancer type. Key: BLCA – Bladder adenocarcinoma, BRCA – Breast carcinoma, COAD – Colon adenocarcinoma, GBM – Glioblastoma multiforme, HNSC – Head and neck squamous cell carcinoma, KIRC – Clear cell renal cell carcinoma, LGG – Low grade glioma, LUAD – Lung adenocarcinoma, LUSC – Lung squamous cell carcinoma, OV – Ovarian carcinoma, READ - Rectum adenocarcinoma, PAAD – Pancreatic adenocarcinoma, UCEC – Uterine corpus endometrial carcinoma.

However, a number of gene expression changes were occurring when a cancer-associated mutation was present in the tumor. The number of these changes was particularly significant in the case of 12 mutations: *APC, CASP8, CTNNB1, IDH1, KEAP1, KRAS, NFE2L2, NSD1, PTEN, RB1, STK11,* and *TP53*. At this point, we could finally ask the question: which cellular processes are mostly affected? And is there any process independently affected by all these 12 mutations? In other words: is there any symmetry?

To our surprise, the cellular processes (here identified by GO terms) associated with each mutation were quite heterogeneous and barely overlapping. The second surprise was that those processes where we observed an overlap showed an unexpected enrichment for two families of processes: metabolism and immune system (Fig. 2-4). Not quite the most renowned hallmarks of cancer (see Fig. 1-3). Given the prominence of metabolism and our technological expertise in mining this type of data in the context of the human metabolic network, we decided to focus on the metabolic genes that are regulated in presence of any of these 12 mutations (Fig. 2-5).

Intriguingly, we noticed an area of "high convergence", where multiple mutations independently are associated with the same set of metabolic genes. To figure out the role of this set of genes required a combination of network analysis and literature mining. As a result, we curated a network of metabolic reactions encoded by a majority of this set of genes. This network revolves around the metabolism of arachidonic acid and xenobiotics and it is mediated by the presence of oxygen and glutathione as co-factors (Fig. 2-6). We termed this network AraX.

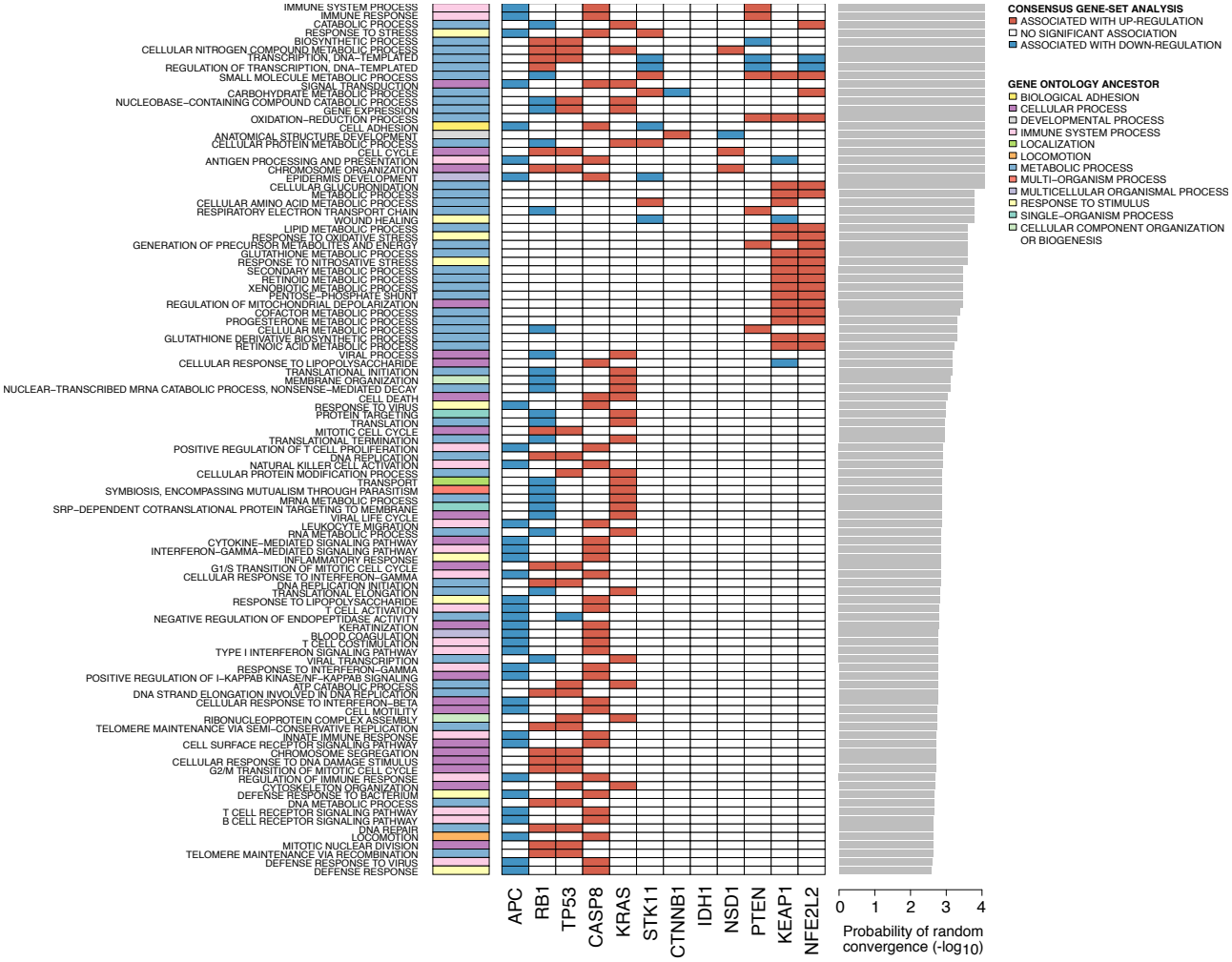


Figure 2-4. Mutations converge on the regulation of GO biological processes that relate primarily to metabolic and immune system processes. Each colored entry indicates that the GO term (row) is a gene-set statistically enriched with up- (red) or down- (blue) regulated genes associated with a mutation (column). GO terms are classified according to the ancestor GO biological process and sorted by the significance of the convergence (barplot on the right).

This finding can be viewed as controversial under different aspects. First, it suggests that the origin of symmetry in cancer due to the mutations is, in essence, a secondary metabolic process. Neither fermentation nor respiration, the processes beyond the Warburg effect, is represented, and so is not nucleotide metabolism. Second, it attributes a prominent role to the metabolism of xenobiotics, which are endogenous or, perhaps more commonly, exogenous substances toxic to the cell (such as drugs), even though these enzymes are generally assumed to be active principally in the liver and induced by exposure to the toxic compounds. Note that the samples in this study are not only untreated, but did not even undergo neoadjuvant therapy, that is a regimen usually administered before the main therapy like surgery. Third, the extent of the regulation of AraX genes by the different mutations is heterogenous and often contradictory. For example, *NQO1* is up-regulated in the presence of mutations of *KEAP1, NFE2L2, PTEN,* and *STK11*, but it is down-regulated with mutations of *CTNBB1* or *NSD1*.

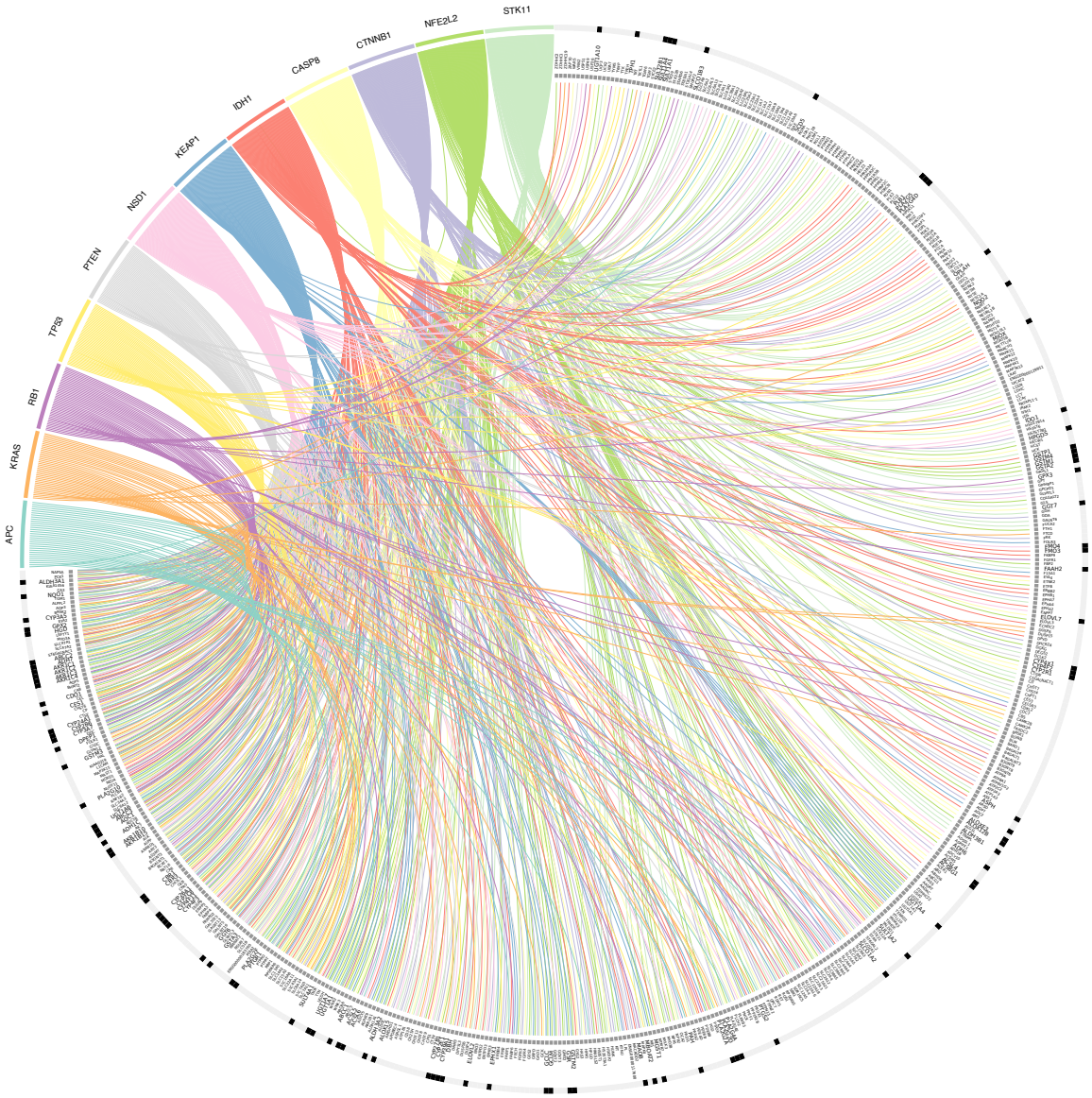


Figure 2-5. The network of associations between cancer mutations and metabolic gene expression reveal a region of high convergence. Metabolic genes are sorted counter-clockwise according to the number of links (i.e. the number of mutations that independently associate with its regulation). Black entries in the outer circle indicate genes belonging to AraX (introduced later).

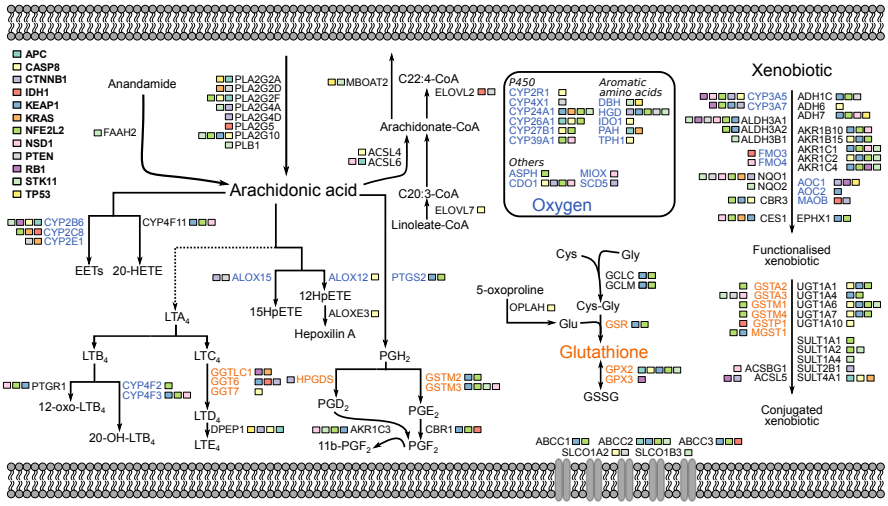


Figure 2-6. A literature curated sub-network of reactions that revolves around arachidonic acid and xenobiotic metabolism (AraX) shows convergence by multiple cancer mutations. The boxes next to each gene indicate which mutations are associated with its regulation.

It is therefore hard to delineate a hypothesis about the function of AraX regulation, if not to conclude that a general deregulation is symmetric in cancer, because independently associated with 12 cancer mutations. This conclusion is powerful because we could not find a better enrichment for the symmetry across 674 canonical pathways listed by Reactome ((*113*), Fig. 2-7), which also includes signaling pathways.

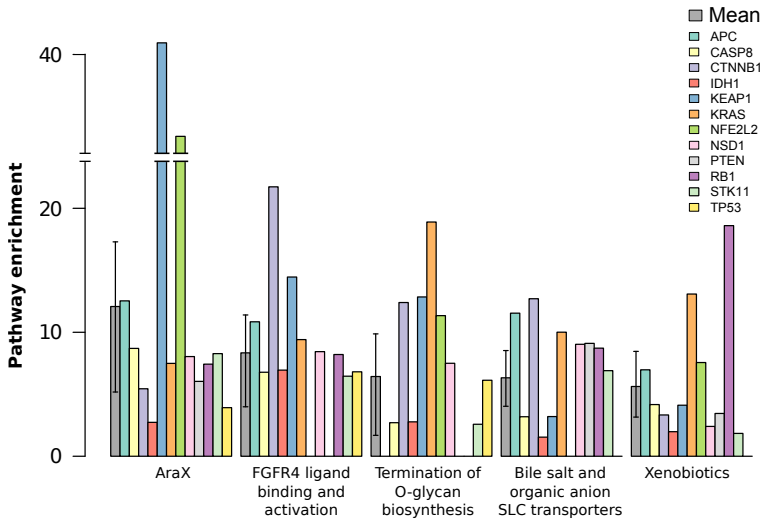


Figure 2-7. Overrepresentation of AraX compared to Reactome pathways by genes associated with a mutation. Each bar indicates the odds ratio for the corresponding mutation. The top five ranked pathways are sorted according to mean overrepresentation (grey bar), where the error bars span the 95% bootstrap confidence interval.

However, this finding must be validated. We reasoned that if AraX deregulation is symmetric in cancer then the extent of its deregulation must ultimately reflect tumor evolution. Hence, bearing a tumor with a highly deregulated AraX should mark lower chances of survival. When we stratified 783 patients according to the level of AraX deregulation compared to the tissue where the tumor originated, we observed that patients with high deregulation scores have far worse prognosis that those with low deregulation scores (Fig. 2-8). The importance of AraX over other metabolic pathways is demonstrated by the fact that no other pathways if deregulated predict survival better than AraX. This strong association with survival is suggestive that AraX is indeed symmetric in cancer because its deregulation seems to mark an evolutionary process driven by known cancer mutations.

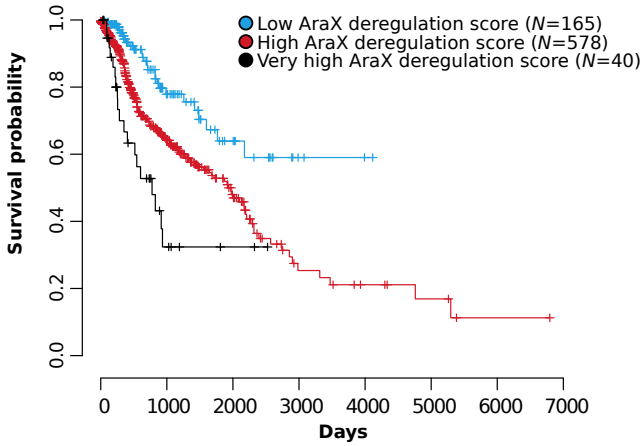


Figure 2-8. Kaplan-Meier survival plots for the subset of 783 samples, for which reference normal samples were available classified with very high (black line), high (dark blue) or low (light blue) deregulation in the AraX pathway.

The question is: why AraX? I am not sure if I have satisfactory answers at this point. We discussed at length in Paper II that the importance of AraX in cancer may reside in the functions of its individual components, given the associations found, for example, between xenobiotic-metabolizing enzymes, genotoxicity (i.e. DNA damage), and cancer initiation (*114*). Undeniably,

the presage of immune system infiltration is hinted by many enzymes in AraX. For example, the involvement of arachidonic acid-derived metabolites suggest active pro-tumorigenic inflammation (*115*). This fact nicely reconciles the major regulation of immune system processes that we observed in correlation with the 12 cancer mutations (refer to Fig. 2-4).

But is this the end of the story, as we know it? Obviously not. The fact that many cancer hallmarks were not recovered by means of this study points to the limitations of this approach in discovering the symmetry. I believe that our study shed some light on the origin of the symmetry in cancer, but it is apparent that we are just scratching the surface. I will frame the symmetry as we know it within the current view of cancer in Chapter IV.

Chapter III: Translational medicine – The case of kidney cancer

The kidney is the dump of the human body. It contributes to the regulation of whole-body homeostasis primarily by excreting waste products of metabolism through filtration, which occurs in functional units called nephrons. The portion of the nephrons that first communicates with the blood to be filtrated is the proximal tubule, composed of epithelial cells as shown in Fig. 3-1.

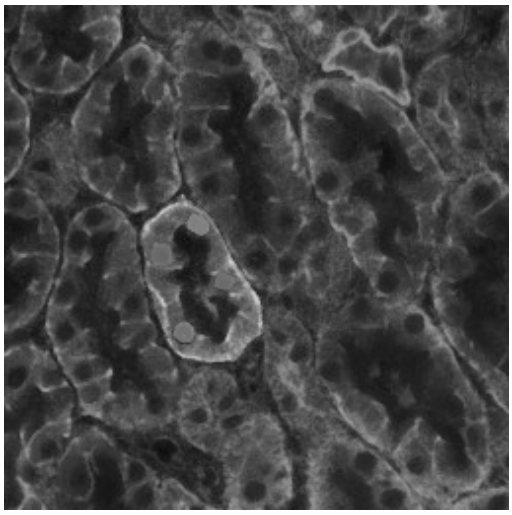


Figure 3-1. Microscopy of kidney tissue showing tubules. One tubule is highlighted to show epithelial cells (light grey), cell nuclei (grey dots) and the tubule lumen (dark center).

This introduction serves one point: these cells give origin to the most common form of kidney cancer, renal cell carcinomas (RCC) (*116*). Why focusing on RCC among all cancer types for which we sought symmetries? One reason. Remember that a subset of samples that clustered away from any other tumor sample when we searched for similarities in the expression of metabolic genes (Fig. 2-1)? They all happened to be RCCs.
Why?

Answer #3: Loss of heterozygosity in metabolic genes

Observe Figure 3-2. Each dot in the PCA is a tumor sample, placed in the two Cartesian coordinates according to the similarity in the *change* of expression from the adjacent normal tissue. Note also that not all genes were used in the PCA, but only metabolic genes and, among these, only those that we deemed commonly regulated among multiple cancer types. So if all tumor samples were regulating these genes with the same expression change, i.e. a symmetric change, then they would cluster altogether. And indeed most do, the dark grey samples, possibly in the regulation of nucleotide metabolism. But the RCCs do not. The direction of the regulation for these metabolic genes is *opposite*.

To be exact, these RCCs have a particular phenotype. They accumulate large amount of glycogen and lipids, so to appear clear under the microscope. Hence, they are called clear cell renal cell carcinomas, ccRCC. It is not surprising that our collaborator in Malmö, a molecular pathologist active in the treatment of ccRCC, told me once "*we (physicians) are all very well aware of the outstanding metabolism of ccRCC*". But he also agreed that there was no evidence for this awareness. And I would add that none could predict this cancer type to be the *sole* one to deviate its metabolic expression profile.

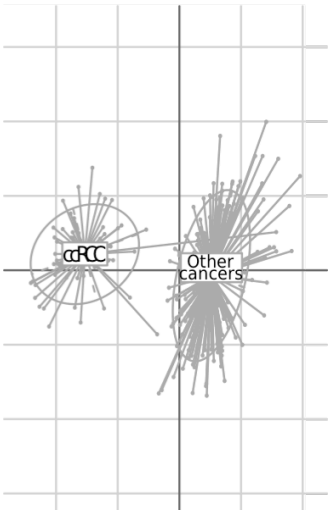


Figure 3-2. Similarity analysis of metabolic gene expression profiles for cancer and tumor-adjacent normal samples. Principal component analysis of $\log_2$ gene expression fold-change vs. matched tumor-adjacent normal samples for ccRCC (light grey, left) and other cancer type samples (dark grey, right).

How is ccRCC metabolism different then? We reconstructed a ccRCC metabolic network in the form of GEM by using protein evidence for each reaction-encoding gene in the generic human GEM. Then we compared the ccRCC network to four other cancer type networks and simply looked to what was in common and what was not at the gene level (Fig. 3-3). One figure stood out clearly. If one looked at the number of metabolic genes that were part of all networks except to a given cancer type network (say bladder carcinoma, BL), then this number ranges 17 – 82 genes (for the example of BL, this number was 33 genes). In the case of ccRCC, the number of such metabolic genes was 169. Compared to the GEM of the normal kidney, 159 out of these 169 genes were part of the normal kidney metabolic network. In other words, ccRCC appears to lose functions for 159 genes that are otherwise part of the kidney metabolism and part of the metabolic network of four unrelated cancer types. We verified that these genes mostly belonged to the metabolism of glycerophospholipids, oxidative phosphorylation, inositol metabolism and, strikingly, nucleotide metabolism.

The symmetry breaks with ccRCC.

We ascribed the uniqueness of ccRCC metabolic networks to recurrent copy number alterations in the chromosome 3p (Fig. 3-4). Here is also located the most commonly mutated tumor suppressor gene in ccRCC, the von Hippel-Lindau (*VHL*) tumor suppressor gene (*117*). We observed that 14 metabolic genes displayed a simultaneous loss of heterozygosity, transcriptional down-regulation and decrease in protein staining in ccRCC compared to the normal kidney. Among these, *ABHD5, CHDH, GPD1L, IMPDH2,* and *PDHB* are located within 3p14.3 and 3p22.3, a region that showed significant decrease in gene copy number in the range of 75% - 81% of samples. Remarkably, these deletions explained many defects previously unveiled in ccRCC metabolic regulation: *ABHD5* and *GPD1L* are involved in glycerophospholipid metabolism; *CHDH* is the first step in choline-dependent one-carbon metabolism; *PDHB* commits pyruvate in the TCA cycle and hence regulates oxidative phosphorylation; and *IMPDH2* is a key step in purine biosynthesis.

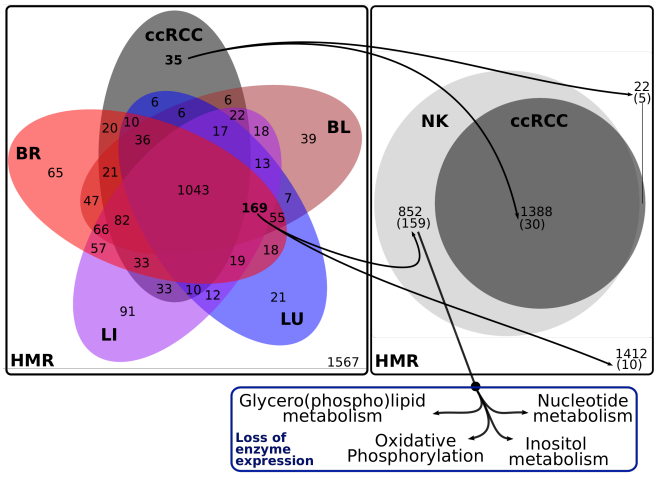


Figure 3-3. Venn diagram for the metabolic genes present in the ccRCC-specific GEM compared to other reconstructed GEMs (other cancer-type GEMs, left, kidney cell in tubules GEM, right). Metabolic genes absent in ccRCC-specific GEM but present both in the kidney cell in tubules GEM and other cancer type GEMs were used to enrich canonical pathways. Key as Fig. 2-3.

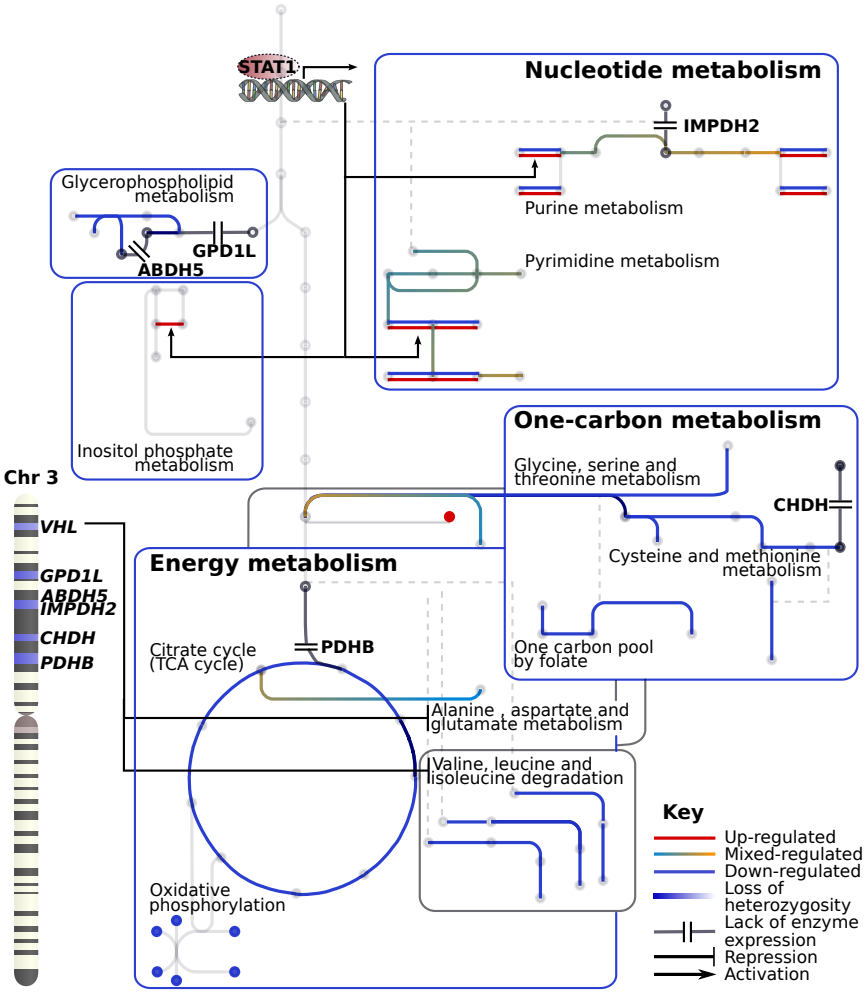


Figure 3-4. An overview of the metabolic features unique to ccRCC in the landscape of cancer metabolic regulation. The figure shows reporter pathways (represented by edges) and metabolites (represented by nodes) transcriptionally regulated only in ccRCC vs. matched tumor-adjacent normal tissue; and sub-networks (represented by rectangles) that feature lack of gene redundancy only in ccRCC metabolic network. The mechanisms that contribute to this metabolic phenotype are summarized. First, loss of *VHL* represses expression of metabolic genes in alanine, aspartate, glutamate, and branched-chain amino acids metabolism. Second, potential activation of *STAT1* up-regulates redundant genes in nucleotide biosynthesis and inositol metabolism. Third, loss of heterozygosity in metabolic genes adjacent to *VHL* affect several pathways previously identified as down-regulated or deficient only in ccRCC (represented by double bar).

Our results suggested that genetic alterations linked to *VHL* loss, a mutational event remarkably exclusive to ccRCC, shape a unique metabolic network in this disease. This may partially explain the symmetry break: there occurs an exceptional genetic event such as loss of *VHL* to determine a different flavor of metabolic reprogramming.

The here-proposed model of metabolic reprogramming in ccRCC remains to be validated. However, it opened some exciting avenues for treatment tailored to ccRCC. For example, does the fact that ccRCC features a uniquely compromised metabolic network expose it to disruption, without affecting the normal kidney or other human cells?

Answer #4: Five liabilities induced by metabolism

The above question requires exploring the limits of the metabolic network of ccRCC. This collection of metabolic reactions indeed defines the potential use of nutrients to fuel production of energy-rich metabolites (like ATP) and macromolecules essential for cell growth. The network imposes stoichiometric constraints on this potential: if one believes in the law of mass conservation, than a reaction in the network defines *exactly* how much product can be synthesized per unit of substrate simply due to stoichiometry. For example, hexokinases are enzymes in the network that mediate the conversion of glucose to glucose-6-phosphate (Fig. 1-6A). This reaction has a 1:1 stoichiometry, i.e. each molecule of glucose yield exactly one molecule of glucose-6-phosphate. The quest for the limits of a metabolic network essentially reduces to selecting a metabolic function of interest (e.g. a product, like glucose-6-phosphate or, more interestingly, a unit of biomass; a process, like ATP production) and verifying which stoichiometric constraints exerted by the network impede the fulfillment of such metabolic function of interest. If our interest is production of glucose-6-phosphate, the limits of the metabolic network are (trivially) the import of glucose and the availability of hexokinases.

More complicated metabolic functions, like biomass production, require a dedicated computational framework. There are way too many reactions that lead to the fulfillment of complicated metabolic functions. The metabolic network of a generic human cell accounts for over 8,000 reactions, according to the latest human GEM, HMR2. These reactions have different stoichiometry, different requirements of substrates and co-factors, different thermodynamically favored directions. However, many of them contribute to the production of biomass, either directly by synthesizing biomass precursors (e.g. membrane lipids) or indirectly by fueling energy production (e.g. the TCA cycle). A common computational framework when dealing with "complete" metabolic network is called flux balance analysis (FBA, previously introduced in Chapter I). FBA can scan all the different routes emerging from the metabolic network that fulfill a metabolic function of interest. It operates under the assumptions of mass-conservation and steady state, i.e. a newly synthesized unit of product must be consumed at the same rate so that it does not accumulate over time. This framework also requires defining which nutrients are available to the network and it assumes for some reaction a thermodynamically privileged direction.

In the view of our study (Paper III), exploring the limits of the metabolic network of ccRCC using FBA translated into selecting biomass production as the metabolic function of interest and manipulating the network to identify which reactions impede the fulfillment of this function. If our hypothesis was correct, that is the ccRCC network is severely compromised, then manipulating the network should have dramatic effect on biomass production. A simple network manipulation is a single-gene deletion. In FBA, a single-gene deletion is simulated by eliminating from the network the reaction(s) encoded by that gene. Hence, using this approach, we could scan each single-gene deletion and determine if there was any gene that ablated biomass production simply due to interference with the ccRCC network. In other words, we could explore the metabolic liabilities induced by the network of ccRCC.

Despite the fact that this strategy has found some utility in the simulation of microbial metabolism (*118*), the accuracy of this method in human cancer cells has never been really tested. Hence, we sought to benchmark our predictions with large-scale experiments of gene essentiality to gauge FBA accuracy. These experiments employed a library of siRNAs, targeting around 200 genes that participate in the human metabolic network. We compared the predictions of FBA simulations *in silico* to the experimental observations *in vivo* for two cancer types, ccRCC and prostate adenocarcinoma, under a number of different set of parameters (nutrients available, use of measured fluxes to constraint the rate of import or secretion of nutrients).

Unfortunately, the accuracy of FBA was quite limited (Table 3-1).

Table 3-1. Statistical measure of accuracy of the flux balance analysis predictions on gene essentiality compared to *in vitro* results for different set of constraints, media, and cancer types. Key: TP – true positive (essential *in silico* and *in vitro*); FN – false negative (non essential *in silico*, essential *in vitro*); FP – false positive (essential *in silico*, non essential *in vitro*); TN – true negative (non essential *in silico*, non essential *in vitro*); MCC – Matthews Correlation Coefficient.

Cancer type	FBA constraints	Medium	TP	FN	FP	TN	Fisher exact test *p*-value	MCC
Clear cell renal cell carcinoma	Topology	FBS	2	18	1	135	0.043	0.226
		HAM	5	15	12	124	0.046	0.174
	Topology + Exchange fluxes	FBS	6	14	11	125	0.010	0.235
		HAM	6	14	15	121	0.032	0.186
Prostate adenocarcinoma	Topology	FBS	2	12	12	186	0.233	0.082
		HAM	2	12	14	184	0.285	0.068
	Topology + Exchange fluxes	FBS	2	12	19	179	0.635	0.039
		HAM	2	12	27	171	1	0.005

At the same time, it revealed that at least in the case of ccRCC the predictions were statistically meaningful. But the numbers were small, the accuracy was limited, and this fostered my skepticism about the utility of FBA. One cannot deny the data though. We confirmed that the statistical significance of FBA predictions in ccRCC were robust with respect to all parameters. The only way to clarify whether these predictions could still be a product of chance (in spite of the low Fisher's statistics, Table 3-1) was to survey the ccRCC metabolic network at the sites where essential genes for growth were claimed. In other words, what is so unique in the network that renders ccRCC (and only ccRCC) liable to disruption?

We ran some subsequent simulations, which convinced me that FBA might have correctly identified several genes essential *in vitro* (Fig. 3-5). For example, *AGPAT6* silencing was associated with substantial cell death in 4 of 5 ccRCC cell lines and identified as essential by FBA. When we explored the network around the acylation of glycerol-3-phosphate to 1-acyl-glycerol-3-phosphate, which is encoded by *AGPAT6* and 3 additional genes, we noticed that the only expressed protein in ccRCC, according to the Human Protein Atlas (*119*), is AGPAT6 (Fig. 3-5A). Since this reaction is crucial for biosynthesis of glycerolipids needed for biomass growth, ablation of the only available isoenzyme would unavoidably lead to cell death. Intriguingly, when we knocked down three of these essential genes (*RRM2B, GCLC,* and *GSS,* shown in Figure 3-5) in a normal kidney epithelial cell line, where ccRCC is thought to originate, the viability of normal cells was not significantly affected (Fig. 3-6).

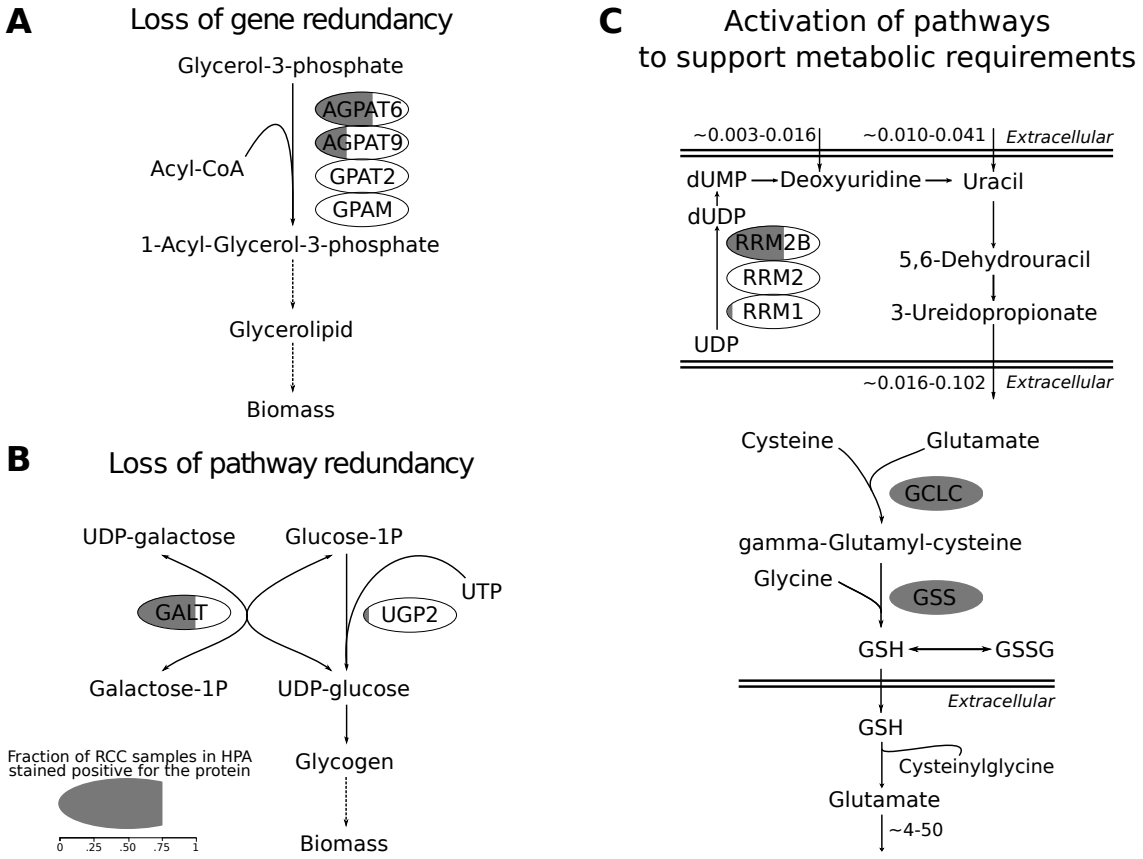


Figure 3-5. *In silico* elucidation of the mechanisms of essentiality for the five genes selectively essential in ccRCC. A) *AGPAT6* is essential only in ccRCC because of loss of gene redundancy. In ccRCC, the repression of *AGPAT9*, *GPAT2*, and *GPAM* in glycerolipid metabolism renders the pathway solely dependent on AGPAT6 to produce essential lipids for biomass. B) *GALT* is selectively essential because of loss of pathway redundancy in ccRCC. Low or no expression of UGP2 forces the flux through GALT to produce glycogen in ccRCC. C) *RRM2B, GCLC* and *GSS* are essential only in ccRCC because of specific metabolic requirements of ccRCC cells that activate the corresponding pathway (flux rates are shown in fmol $cell^{-1}$ h^{-1}). Top: the measured secretion rate of 3-ureidopropionate in ccRCC cell lines is not matched by the observed uptake rate of its direct precursors, uracil and deoxyuridine. This forces a flux active in the catabolism of UDP (part of the pyrimidine degradation pathway) to compensate for the observed 3-ureidopropionate secretion rate. One of the pathway steps is uniquely catalyzed by RRM2B, given that the other genes associated to this reaction (RRM1 and RRM2) are not expressed in ccRCC. Bottom: ccRCC cell lines secrete glutamate at a high rate and the only flux distribution that fits glutamate secretion in the ccRCC metabolic network requires the cleavage of extracellular glutathione (GSH). Extracellular GSH is in turn derived from *de novo* GSH intracellular synthesis that is catalyzed by GSS and GCLC. Noteworthy, the reduction of reactive oxygen species like H_2O_2 by GSH is a metabolic function preserved in the predicted flux distribution. For each protein, the grey shading represents the fraction of ccRCC samples in which the protein is expressed according to the Human Protein Atlas.

Taken together, it seems that ccRCC shapes the metabolic network uniquely, and it is vulnerable to disruption in at least 5 different sites. These liabilities are noteworthy because they specifically arise from the emergent stoichiometric constraints of the ccRCC network and cannot be attributed either to functions present exclusively in the kidney nor to metabolic requirements of proliferating cells, as suggested by the knock-down experiments in the normal kidney epithelial cell line.

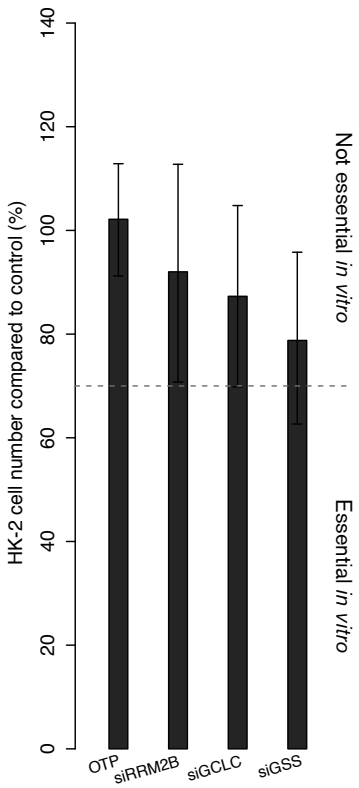


Figure 3-6. Toxicity for *GCLC, GSS, RRM2B* knockouts in a normal kidney epithelial cell line, HK-2. Cells were transfected with siRNA targeting RRM2B, GCLC, and GSS and a non-targeting scrambled siRNA, OTP, was used as negative control. Each bar represents the mean cell number reduction relative to control together with the 95% highest density interval of two experiments performed in triplicate.

FBA is an approach essentially limited to characterizing the uniqueness of the topological properties of ccRCC metabolic network. However our first key observation of the uniqueness of ccRCC metabolism was at the level of gene expression regulation. In Paper I, we sought to decipher why ccRCC is unique, but can we provide a better understanding about which expression changes are unique?

Answer #5: Shine on your crazy glycosaminoglycans

I refer, one more time, to Figure 3-1. The gene expression in ccRCC metabolism showed some unique patterns, which we could not reconcile with the fact that ccRCC arise in the kidney, nor with the observed higher infiltration of stromal and immune system cells (*120*). The expression changes were vast, and we mainly focused on characterizing them in terms of metabolic pathways (refer to Fig. 2-2). In this follow-up study (Paper IV), we sought to obtain additional information. First of all, we retrieved more data, passing from 65 pairs of ccRCC-normal kidney samples to 481 ccRCC vs. 74 tumor-adjacent kidney samples. Second, we explored different algorithms. One way to mine metabolism is represented by so-called reporter metabolites (*121*). This approach abstracts the need to arbitrarily define the gene content of a pathway. Instead, it lists all the genes that "surround" a metabolite, by encoding for the reactions that either produce or consume it. Then, it calculates a score for the metabolite, which condenses the statistical confidence that the metabolite-associated genes changed expression in the condition of interest. This score follows an inverse normal distribution, and therefore one can calculate the statistical confidence that the metabolite is significant because surrounded by differential regulation of the associated genes. These metabolites are called reporter metabolites. Although widely informative, this approach has a drawback: it treats metabolites in isolation with each other, therefore the scientist cannot understand whether two reporter metabolites are closely connected in the metabolic network unless he has prior knowledge about them. One may know that succinate is closely related to fumarate, but maybe not that it is

also related to γ-aminobutyrate (better known as the neurotransmitter GABA). (They are actually just two reactions away.) This led to the development of the Kiwi algorithm (Paper V), whose scope was precisely to fish the sub-network of connected metabolites that are relevant in a condition of interest. The algorithm can be applied broadly to any biological network in which the user has computed a statistic for each node. This statistic is in turn correlated to the confidence that the node is relevant in a certain condition (*122*). For our case, we limited our analysis to metabolites in the human metabolic network represented in HMR2.

Running Kiwi in the comparison ccRCC vs. normal kidney recovered expression changes in metabolic genes already unveiled by our previous study (*120*). However, we also observed the emergence of an unanticipated sub-network of metabolites (Fig. 3-7A). This component comprised precursors of chondroitin sulfate (CS) that were connected to precursors of heparan sulfate (HS). The former were distinctively characterized by up-regulation of the associated genes, the latter by down-regulation. Further manual inspection of this part of human metabolism revealed that these metabolites represent two branches of a larger metabolic pathway known as glycosaminoglycan (GAG) biosynthesis (Fig. 3-7B). As pointed out earlier, this pathway displayed an astonishingly coordinated regulation in ccRCC compared to the normal kidney. We confirmed the pattern of gene expression changes in two independent published datasets, which also confronted ccRCC vs. normal kidney samples (*123, 124*) (Fig. 3-7C). Moreover, the coordinated character of this regulation in ccRCC was not observed in six other common epithelial cancer types (Fig. 3-7D). Taken together, this was suggestive that a unique metabolic event is taking place during the progression of ccRCC in the kidney, which entails a coordinated alteration of GAG biosynthesis.

The exceptionality of this metabolic event in ccRCC called for further inspection. What if these gene expression alterations were actually reflected by changes in GAG levels and/or composition in ccRCC? The unique character of this regulation may imply that specific alterations in GAG levels might occur in ccRCC. Speculatively, these unique alterations might be even appreciated in kidney-proximal fluids. Prompted by this, we decided to analyze GAGs in the plasma and urine of subjects with metastatic ccRCC (mccRCC) as opposed to healthy individuals. We selected metastatic patients because we thought that at this stage we would have better chances to detect changes. This choice stemmed from the consideration that GAGs have been previously implicated in metastatic processes when located at the extracellular level (*125, 126*). Hence, our initial cohort included 34 patients with mccRCC plus 16 healthy individuals. In 21 of 34 mccRCC patients, only plasma samples were collected. CS and HS concentration and their disaccharide composition were quantified in the samples using liquid chromatography with on-line electrospray ionization mass spectrometry (*127, 128*). In total, 18 independent GAG properties were measured in every fluid sample (note that the GAG charge is the sum of all sulfated disaccharide fractions). The collection of all these data points defines a GAG profile. In line with our speculation, we observed marked differences in both the plasma and urine GAG profile of subjects with mccRCC (Fig. 3-8).

Some GAG properties were so significantly altered that we considered the possibility of designing an unprecedented diagnostic biomarker for mccRCC. It should be noticed that no such biomarker has entered the clinical routine as for 2015 (*129*). We utilized a method called penalized Lasso (*130*) to pull out the most relevant GAG properties in the comparison mccRCC vs. healthy. Lasso works as a regression for a certain dependent variable, in our case the clinical outcome (mccRCC or healthy, a binary value), on defined independent variable, in our case the GAG profile either in the plasma or in the urine. The regression is then penalized to return only those independent variables that are predictive of the dependent variable, i.e. the GAG properties most predictive of mccRCC. This method features an internal cross-validation to avoid over-fitting and returns robust predictors. Then, the coefficients of the regression were used to develop a formula designed to yield higher values in case of mccRCC.

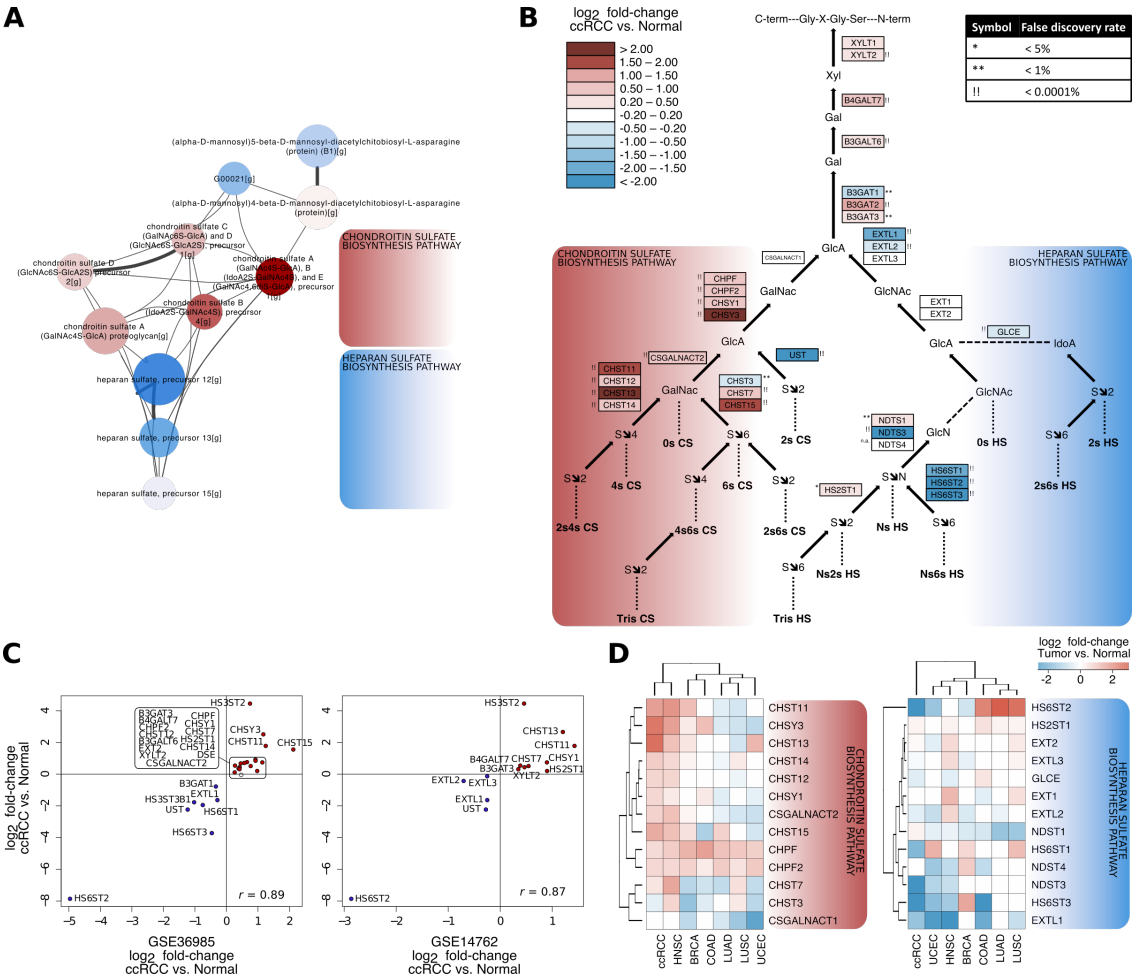


Figure 3-7. Coordinated regulation of glycosaminoglycan biosynthesis in ccRCC vs. normal kidney. A) Genome-scale metabolic modeling using Kiwi reveals a coordinated transcriptional regulation in a subnetwork of metabolites belonging to the chondroitin sulfate and heparin sulfate biosynthetic pathway. The node color indicates the general direction of regulation of the genes associated with the metabolite (red – up-regulation; blue – down-regulation). B) Pathway-view of glycosaminoglycan biosynthesis in ccRCC. Each box shows the enzyme(s) carrying out a given reaction in the pathway. The color represents the $\log_{10}$ fold-change in ccRCC vs. normal for the enzyme-coding gene, while the symbol next to each box reports the significance for the corresponding gene regulation (in terms of false discovery rate). The pathway has been drawn according to KEGG gene associations (Note that genes related to dermatan sulfate biosynthesis or sulfation at C3 in heparan sulfate are not shown, the latter event being rarely observed (*131*)). Solid arrows indicate addition of a molecule, dashed lines indicate conversion of a molecule, and dotted lines indicate the final disaccharide composition up to that point. C) Correlation of gene expression $\log_2$ fold-changes in the glycosaminoglycan biosynthesis pathway between TCGA samples (y-axis) and two independent studies (GSE36986 and GSE14762, (*123, 124*)). D) Gene expression $\log_2$ fold-changes in the glycosaminoglycan biosynthesis pathway in ccRCC as opposed to other cancers vs. matched normal tissues. Key: HNSC – Head and neck squamous cell carcinoma; BRCA – Breast invasive carcinoma; COAD – Colon adenocarcinoma; LUAD – Lung adenocarcinoma; LUSC – Lung squamous cell carcinoma; UCEC – Uterine corpus endometrial carcinoma.

We developed three formulas, based on either the GAG profile in the plasma, or in the urine, or both:

$$\text{Plasma score} = \frac{[6s\ CS] + CS_{tot}}{\frac{3}{10}\frac{[4s\ CS]}{[6s\ CS]} + [Ns\ HS]}$$

$$\text{Urine score} = \frac{[Ns6s\ HS] + 60 \cdot \text{Charge}\ HS}{[4s\ CS]}$$

$$\text{Combined score} = \text{mean}(\text{Plasma score}, \text{Urine score})$$

where terms in brackets represent the fraction of the disaccharide for the corresponding GAG (the abbreviations describe different sulfation patterns for CS and HS as per Fig. 18B), CS_{tot} is the total concentration of CS (in mg/mL) and Charge HS is the total fraction of sulfated disaccharides of HS.

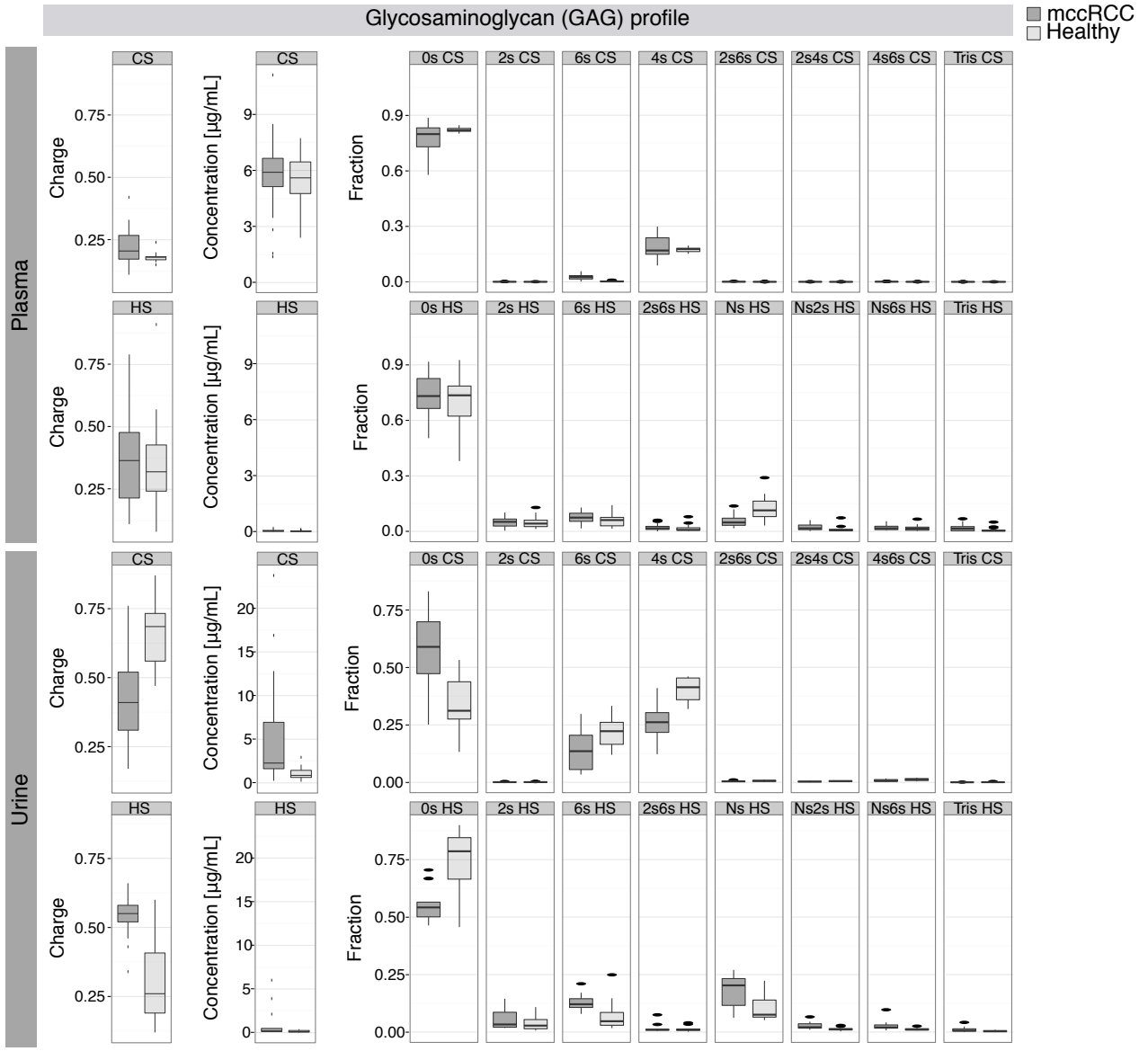


Figure 3-8. The glycosaminoglycan plasma and urine profile of mccRCC patients is markedly distinct than healthy individuals. The glycosaminoglycan profile of mccRCC patients (dark grey boxplots) and healthy individuals (light grey boxplots) in the plasma (top) and urine (bottom). Each profile comprises 18 independent measurements of GAGs (9 related to chondroitin sulfate, CS, and 9 related to heparan sulfate, HS), which refer to the total concentration and the disaccharide composition.

Not only were all three scores substantially higher in subjects with mccRCC compared to healthy individuals (Fig. 3-9A), but also a classification based on these scores had an area-under-curve ranging from 0.966 for the urine score to 1 for the plasma and combined scores, i.e. perfect classification (Fig. 3-9B). In other words, it seems possible to profile GAGs in subjects' plasma and urine and classify a subject as either healthy or with mccRCC.

We decided to validate whether these scores were inflated to achieve perfect separation by means of the method utilized to design the formulas (even though I had already noted that the GAG properties were inherently selected by Lasso to be robust predictors). We gathered a second validation cohort, blindly and independent from the first cohort, consisting of 18 patients with mccRCC and 9 healthy individuals. For 11 of 18 mccRCC patients, only plasma samples were

obtained. To our delight, the performance of the scores was even higher in the validation cohort (Fig. 3-9C and D).

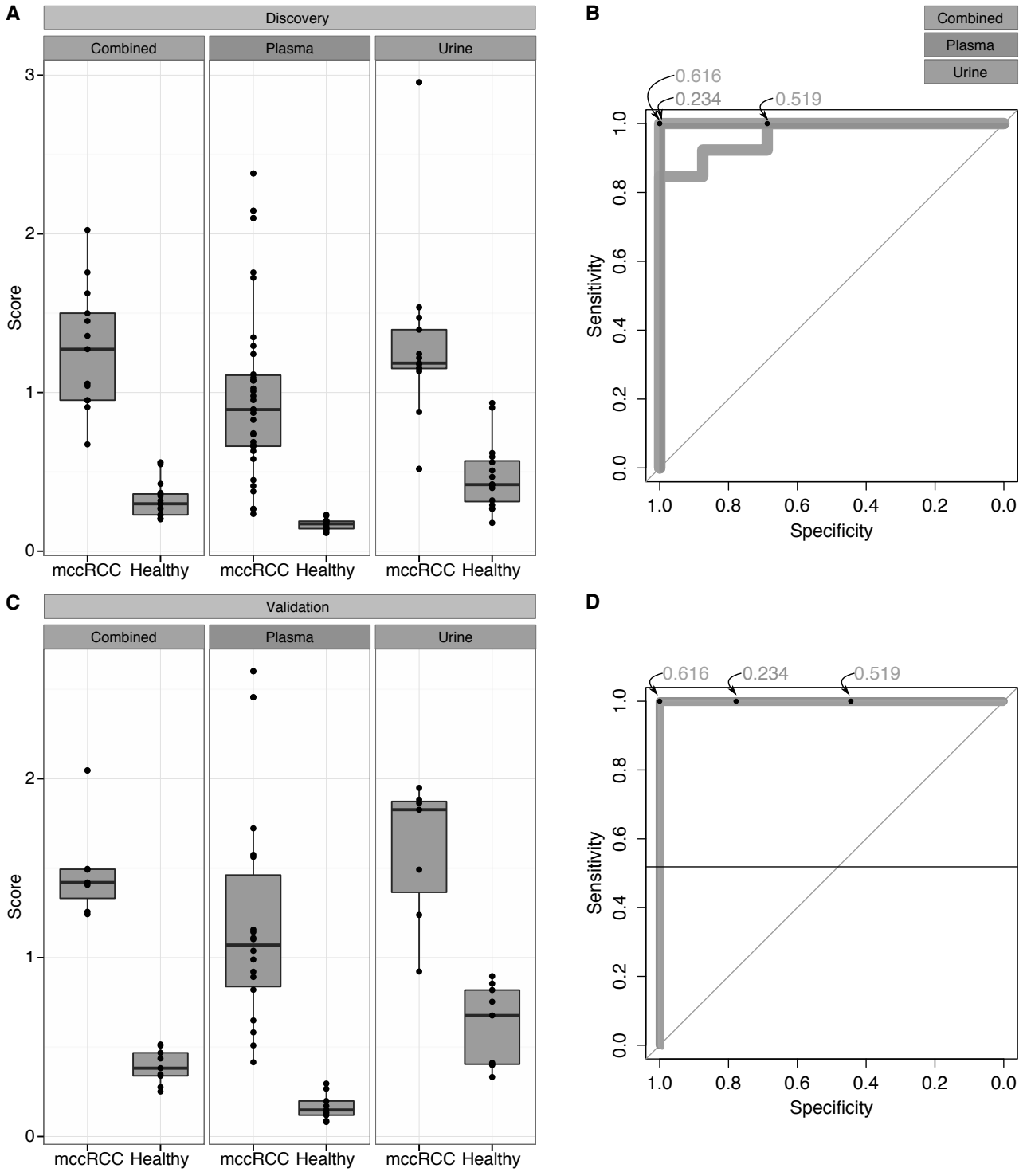


Figure 3-9. The glycosaminoglycan profile can be summarized in three scores (based on measurements in the plasma, urine, or both) that can accurately predict occurrence of mccRCC. A) Plasma, urine, and combined scores in mccRCC patients (dark grey boxplots) and healthy individuals (light grey boxplots) belonging to the discovery cohort (34 samples vs. 17, respectively). B) Receiver-operating-characteristic (ROC) curves in the classification of samples of the discovery cohort as either mccRCC or healthy based on the combined, plasma, and urine scores. For each marker, an optimal cut-off scores that maximizes the negative predictive value is indicated. C) Plasma, urine, and combined scores in mccRCC patients (dark grey boxplots) and healthy individuals (light grey boxplots) belonging to the validation cohort (18 samples vs. 9, respectively). D) ROC curves in the classification of samples of the validation cohort as either mccRCC or healthy based on the combined, plasma, and urine scores.

However, a truly diagnostic marker would return a decreased score in the case of patients previously diagnosed with mccRCC but currently declared as disease-free. To test this possibility, we analyzed the GAG profile of 8 such subjects. Reassuringly, we observed a decreasing trend in all scores, particularly pronounced for the plasma score (Fig. 3-10). Using our initial discovery cohort, we selected a score cut-off that maximized the negative predictive value, so that our test could be indicative of healthy status if the score in the subject is below the cut-off. Such test would be useful to monitor mccRCC recurrence during follow-up of patients like those represented in this last cohort of 8 subjects. In our experiment, the plasma test would correctly classify 7 of 8 subjects as healthy.

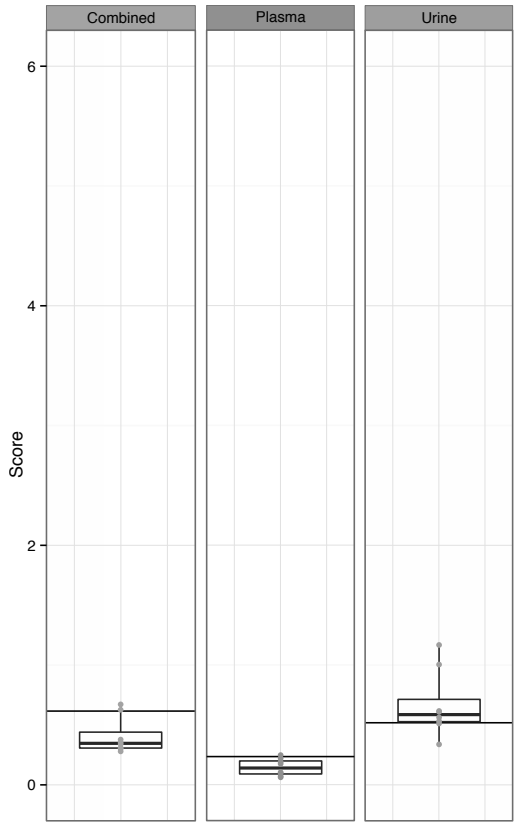


Figure 3-10. Combined, plasma, and urine scores in subjects previously diagnosed with mccRCC but with no evidence of disease at the time of sampling. The horizontal lines represent the optimal cut-off scores at which a subject is classified as either mccRCC or healthy according to maximum negative predictive value.

In Paper IV, we also performed analysis of covariance on some confounding factors, such as age, BMI, dietary intake or treatment regimen. We showed that the clinical outcome is always best predicted by the scores, independent of the confounding factors. Taken together, these results are very promising. They suggest that a unique metabolic event in mccRCC may be translated to an actual accessible diagnostic biomarker for the disease. In the future, it might be possible to establish these markers for a diverse range of diagnostic tools in the clinical management of ccRCC, with considerable benefits for general healthcare.

I would like to conclude this section, and with it the chapter, with one speculative note on the symmetry that I learnt from the case of ccRCC. As discussed in Paper IV and corroborated by our data, the plasma of healthy individuals has a stable GAG composition, typically not affected by any tissue (refer to Fig. 3-8). However, there is an obvious systemic alteration of GAG composition concomitant to metastatic ccRCC. The resulting GAG profile is remarkably similar to the GAG composition of lymphocytes (*132*). This led us to speculate that the infiltration of the immune system in mccRCC could lie behind the observed coordinated regulation of GAG biosynthesis. This, in turn, was probably the most relevant and exclusive metabolic event in ccRCC at the system

level. In the same way, AraX was the most relevant and exclusive metabolic event induced by oncogenic mutation at the system level. It seems to me that both events entail a prominent role of the immune system. What if the only oncogenic aspect of metabolic reprogramming that any cancer must acquire is a beneficial engagement of the immune system? Let me rephrase this. What if the only metabolic trait symmetric in cancer is not the Warburg effect, to name one, but rather a redistribution of fluxes beneficial to the immune system?

Chapter IV: The future of cancer care

My last four years or so of research are now turning to a conclusion in what I would likely deem a bitter end. This bitterness arises from the nagging feeling that I have missed something crucially important in the conceptualization of the origin of symmetry in cancer. Let me be more specific. I believe that the premises beyond the origin of symmetry in cancer that I described in Chapter I are sound: cancer *is* the collection of diseases defined by the acquisition of malignant phenotypic traits like abnormal proliferation and convincing evidence shows how this symmetry is acquired by disparate cancers via converging evolutionary trajectories. However, in my attempt to elucidate whether reprogramming of metabolism figures among these symmetric phenotypic traits, I stumbled upon results that rather argue the opposite. This is puzzling. Of course, I acknowledge the multifaceted appearance of cancer metabolism in virtue of the fact that metabolism is meant to be plastic and adaptive. Yet, I regard it plausible that the transformation should lead to a symmetric reprogramming of metabolism, if anything to support the life of a mutant proliferating cell. Looking back at my data and methods, I may certainly detect flaws or assumptions that can be tinkered, but I doubt that this would largely affect the results, and with that the general interpretations that my collaborators and I drew. These interpretations, as mentioned before, collectively seem to argue against the symmetry of metabolic reprogramming. This is the bitter end. Now I question: is this true or, as systems biologists, what have we missed? How does this speak for cancer in general and, as human beings, how does this affect us in terms of hazard for our health? How does this speak for the way we do science? In this last chapter, I will address, or better said conjecture, these questions separately.

From here we go sublime

I have insisted in the introductory section that our results point against the symmetry of metabolic reprogramming in cancer. This is perhaps confusing, considering that in Chapter II I first suggested that up-regulation of nucleotide metabolism is symmetric and second that deregulation of AraX is not only symmetric but also oncogenic. The reason why I am skeptical is twofold. In the case of nucleotide metabolism, I have already discussed that this process seems to be symmetric because cells are proliferating. In the case of AraX, this result hardly reconciles the accumulated evidence from molecular biology studies on cancer metabolism. On the other hand, I would like to refute the thesis that molecular biology unfolded a proof of symmetry in the metabolism of cancer cells, as contended by some review papers in the field (Paper VI).

The contribution to our understanding on how metabolism is regulated in cancer enabled by molecular biology is enormous. With reference to Fig. 1-4, a number of phenotypic traits in metabolism were declared symmetric at different times. Nevertheless, we and others (*133*) recognized that even the trait most commonly declared symmetric, aerobic glycolysis, is not observed in some cancers. A consensus model that accommodates the limits of the symmetry for these phenotypic traits has been constantly challenged by newer discoveries and remains therefore elusive (*134*). I have already stressed that this observation does not preclude the existence of symmetry in the reprogramming of cancer metabolism, but it poses a boundary to the breadth of the conclusions that can be derived in molecular biology. Systems biology studies like ours sought to uncover the origin of symmetry by approximating the definition of phenotypic trait in metabolism. This approximation is necessary simply because the current technology either has not enough resolution to precisely interrogate a phenotypic trait or because it is not sufficiently scalable to interrogate the phenotype at the systems level. We stratified these early studies into five levels of approximation. A first level of approximation is to define a trait by the presence of genetic or epigenetic alterations in its corresponding metabolic pathway. Ciriello et al. condensed these alterations across 3,299 tumor samples from 12 tumor types into 31 oncogenic signatures (*135*). None of these signatures were found to specifically enrich metabolic pathways or processes, however, all signatures collectively encompassed 5 signaling pathways known to control metabolic

reprogramming. A second level of approximation is adopted in studies like ours, such as Hu et al. (*101*) and Nilsson et al. (*103*), in which a trait is defined by the expression level of the genes belonging to its corresponding metabolic pathway. As mentioned earlier, the consensus between three independent studies is that metabolic reprogramming in a tumor appears to be limited to the up-regulation of nucleotide metabolism. A third level of approximation defines a trait by the expression of sufficiently connected components with the metabolic network, like metabolites and reactions, rather than relying on canonical and arbitrary definition of metabolic processes. In this fashion, two studies by Ågren et al. (*136*) and Wang et al. (*137*) independently showed that cancers are symmetric in the formation of reaction sub-networks that revolve around the metabolism of eicosanoids. A fourth level of approximation defines a trait by the abundance of related metabolites, which are informative of the metabolic state of that trait. A global view on the systems traits can be provided by untargeted metabolomics (*138*), but as for today no study has simultaneously addressed multiple cancer types in a tumor vs. healthy control comparison. However, Patel and Ahmed recently reviewed a progresses in individual cancer types spanning 12 tissues (*139*) and ascribed recurrent, but by no means symmetric, perturbations in the level of metabolites in glycolysis, TCA cycle, choline metabolism, and fatty acid metabolism. Finally, a fifth level of approximation defines a trait by the flux through sufficiently connected components of the metabolic network. This ultimately represents the quintessential experiment to gauge metabolic reprogramming, but the only two systems biology studies in this sense are severely limited by the scalability of the technology. But still Yuneva et al. (*140*) and Fan et al. (*141*) observed that reprogramming of central carbon metabolism is not symmetric, in that the flux redistribution induced by mutations could not be purely ascribable to cancer regardless of genetic and/or micro-environmental heterogeneity.

So are there any symmetries? Boldly, one could argue that regulation of nucleotide and eicosanoid metabolism was alternatively deemed symmetric in some of the above-mentioned systems biology studies. Nevertheless, the fact that this was not consistently observed may induce the reader to refute it with same argument according to which we would not dare declaring aerobic glycolysis symmetric. In fact, this argument does not apply yet. Contrary to aerobic glycolysis where experiments provided solid demonstrations of its absence in certain tumor types, no study has explicitly proved that altered regulation of nucleotide and eicosanoid metabolism is *not* symmetric. Even in our study (*120*), in which we differentiated ccRCC from all other cancer types because it does not up-regulate nucleotide metabolism, we still reported a mixed regulation perhaps meant to compensate the defects in the network introduced by its outstanding genetic make-up. So, until convincing dismissal, these two metabolic pathways may still constitute the origin of the notorious symmetry in the metabolism of cancer. However, I conjecture that this is not so interesting. I have already speculated at length why I do not regard reprogramming of nucleotide metabolism central to the evolution of cancer. And in the case of eicosanoids, the only study that has so far attributed to their deregulation an oncogenic character indispensable to any cancer was ours, in which eicosanoids are featured in the AraX pathway (Paper II). There are numerous proofs that eicosanoids can function to promote cancer progression (*115*), and there exist even anti-tumorigenic drugs that target their metabolism, like aspirin (yes, aspirin) in colorectal cancer (*142-144*). These proofs do not nor are intended to demonstrate its symmetry. I have not found any study in which the cancer phenotype was reverted in a diversity of tumor models by restoring the regulation of the eicosanoid pathway, or more generally the AraX pathway. In the absence of data, how can I claim that this is probably not so interesting? I offer two explanations. First, I am biased by the fact that the wealth of molecular biology studies that I surveyed in the last years rarely reported the activation of AraX metabolism as a response to oncogenesis. At the same time, this lack may just reflect an inherent bias of the experimentalists in what they regard interesting or trendy in metabolism. (Anecdote: a seminal paper in cancer metabolism is the demonstration by Ying et al. that oncogenic KRAS primarily acts to redirect central carbon metabolism towards anabolism in pancreatic cancer (*45*). Their focus on metabolism stemmed from the gene expression profiling of cells following oncogenic activation of KRAS, which revealed that most overexpressed genes are

metabolic. Nevertheless, most of these genes belonged to steroid biosynthesis. It is unclear why the authors decided to focus on central carbon metabolism for the rest of their paper.) Thus I give a second explanation. The deregulation of eicosanoid metabolism, and in general of AraX, seems not to emerge from an active reprogramming induced by the mutations to confer the cells with a selective evolutionary advantage, but rather by a passive adaptation to an environment that the presence of mutant cells have remodeled. This could still qualify as an oncogenic process, because it requires the presence of a mutant population in an otherwise physiologically normal environment. I should specify that when I talk about remodeling of the environment I primarily refer to the infiltration of the immune system, in other words inflammation. If this hypothesis is correct, cancer cells do not require reprogramming AraX at all. However, they will *always* respond to inflammation by altering the *same* pathway (i.e. AraX). Maybe, as I speculated at the end of the previous chapter, the oncogenic aspect of metabolic reprogramming in cancer is simply to be at the service of the immune system.

Or are we missing something? One could discard the whole argumentation above and point to the fact that the way we dealt with the quest for symmetries is out of focus. In support of this, we proposed and ranked five major weaknesses that affected these early studies in systems biology:

1. First and foremost, the phenotype of a tumor is only approximated by its gene expression profile. In the end, the central dogma recognizes that the proteins are the ultimate effectors for the phenotype of a cell (well, the dogma neglects the outstanding role of small molecules in many biological functions). Although recent estimates put transcription on the front seat when it comes to controlling protein levels (accounting for ~70% of the variance (*145*)), translation and degradation play a non neglectable role. Even under this approximation, proteins exert their function and define a cell phenotype by means of interactions within each other and with the environment (*75, 146*). These interactions depend on the availability of certain compounds at a given time, potential modifications of protein active sites (or even distant sites) via post-translational modifications, and probably a number of processes that we do not fully understand or, more likely, not know at all (*147*). Probably this last point is the most compelling. We rely and attempt to corroborate the paradigm introduced by Watson and Crick on molecular biology, the central dogma. This imposes tremendous limitations on our ability to describe and interpret a phenotype. Most of these results will undoubtedly collapse or necessitate to be revisited in light of the future paradigm shift.
2. The sample size of these studies is still limited, with only thousands of cancer samples across tens of cancer types. It is worth reminding that some 15 million *new* cases of cancer are diagnosed every year, classified in over 100 types. Any claim about a symmetric property should be framed with these numbers in mind.
3. Mutations are not the only drivers of cancer evolution. This is not exactly surprising to cancer researchers, but I find this fact very fascinating. Two recent studies reported some interesting data on this. The first one is about ependymomas. Ependymomas are common brain tumors in childhood. In particular, the posterior fossa group A (PFG-A) is a subtype of ependymoma that is more common in infants, lethal in most cases. It is hard to conciliate the initiation of such cancers with the occurrence of driver mutations, because there is realistically no time to accumulate them in infants. Indeed, the authors reported that PFG-A ependymomas were genetically bland. However, there was a major difference with the other ependymoma subtype, PFG-B: the DNA was drastically methylated in the loci of genes that are target of the Polycomb repressive complex 2, which indirectly represses expression of differentiation genes. Even without any mechanistic proof, this study is suggestive that an entire cancer type could be triggered by specific epigenetic alterations and not by mutations (*148*). The second one is about normal skin. The authors found that skin cells display a surprisingly high number of driver mutations, despite being physiologically normal. For example, as many as 83 clones per square centimeter of skin

positively selected for mutations in *NOTCH* genes, a gene family causally implicated in cancer due to their role in the regulation of stem cell biology. Even though these are aged and UV-exposed cells and even considering that the mutation burden is still at the lower end for most skin cancers, the fact that normal cells carry so many cancer-causing mutations is sufficient to question "*what combinations of events are sufficient for transformation*" (*149*).

4. The technology is limiting. This is a fact that will always impinge the spectrum of scientific questions that can be legitimately answered. In this case, the advancement in our understanding on which and to which extent a gene is expressed is outstanding compared to twenty years ago or so. Yet, we still relied on pictures of the transcriptome that are static, approximated, and related to a variegated population of human cells.
5. In close relation with the previous point, these questions can and should be addressed experimentally. Once research engineers will provide a scalable and practical technology to design the experiments outlined before, the reliance on statistical associations to infer links between mutations and phenotypes will appear cumbersome and become obsolete.

A natural question I posed myself is if there is a future for the specific question of symmetry in cancer metabolism. I believe that any study that takes in account one or more of the points above as part of the experimental design can certainly throw more light on the existence of the symmetry. Also, there are two specific research questions that can be addressed immediately to clarify some of the concerns I raised here. First, it is conceivable to perform a meta-analysis of the expression profile of nucleotide metabolism of cancer cells and physiologically healthy proliferating cells, possibly at the resolution of single cells and ideally adjusted by tissue of origin. Second, one could envision "normalizing" the expression of AraX using small molecules in a number of tumor models with the aim of first validating whether tumor evolution is so halted and second estimate the role of inflammation. There are some hurdles in this experiment, such as how to "normalize" the expression of a pathway and which are good tumor models (with particular consideration to the special nature of human immune system). However, I maintain that the reach of the scientific question is worth the time investment.

Yesterday was dramatic, today is ok

I pay tribute to the beginning of my thesis by attempting to communicate how scientific research contributed to make cancer a less frightening disease. My obligate premise is that I still would not want to be diagnosed with *any* form of cancer, but this goes along with a long list of bad diseases. I think that, when talking about cancer, people care about only one thing: a cure. This topic is way too vast to be even briefly elaborated and, also, I lack sufficient expertise on the matter. I will focus on two aspects about curing cancer: how research shaped a new perspective on the cure of cancer and the role of systems biology.

Precision or personalized medicine is possibly the hottest trend in clinical cancer research. It stems from the observation of the daunting complexity of cancer. The idea beyond precision medicine is to target the specific molecular features characterizing a patient's tumor, and these are exquisitely acquired by the unique genetic make-up of the tumor in that patient. Currently, we are far from being able to individualize treatments. Hence much research goes into stratification, which essentially recognizes that the tumor genetic make-up is in the end not so unique to each patient and the molecular features are in the end not that specific to an individual tumor. I have already argued that this represents to me a conceptual fallacy, in that this definition of stratification may well apply to decades of clinical management of cancer based on histopathological classification, which makes the concept of precision medicine a bit ridiculous. What is not ridiculous is the realization that different drug regimens can be utilized to target the molecular features typical of a patient segment. There are classic examples that are frequently reported to corroborate this realization: gefinitib in lung cancer patients with *EGFR* mutations (*150*), vemurafenib in melanoma patients with *BRAF*

mutations (*151*), and imatinib in gastrointestinal stromal tumor patients with *cKIT* mutations (*152*). In general, targeted therapies have transformed the chances of survival of cancer patients. A collaborator of ours brought to our attention that the prognosis of metastatic ccRCC has dramatically improved since the introduction of tyrosine-kinase inhibitors (*153*), and I think it is fair to say that this positive trend applies to many different cancer types. Compellingly, the only stratification here applied is the tissue of origin and the tumor grading and/or staging, so this success should not be attributed to precision medicine (*154*). (Anecdote: a phase III trial in advanced squamous-cell non-small-cell lung cancer recently reported the superiority of nivolumab as opposed to the current standard of care (*155*). This is a novel targeted therapy that inhibits the signaling mediated by the programmed death 1 (PD-1) receptor, which is engaged by ligands often expressed by this cancer type. Based on this, precision medicine would argue that a stratification based on the ligand expression could separate the patients that respond better to the therapy. Too bad that the superiority of nivolumab was independent of the expression of the ligand, which was neither prognostic nor predictive of benefit). I am not suggesting that precision medicine did not produce a relevant progress in cancer care, in that I believe that whatever improves the survival chances of a patient *is* an undisputable progress. However, I argue that at this stage its conceptual impact far supersedes its factual impact. At the opposite extreme stands the position of researchers who stratify all patients based on the sole feature of bearing a tumor (apologies for the contradiction in terms). This position is bashed by proponents of precision medicine, on the wake of the (perceived?) limitations of untargeted therapies, like chemotherapy. I propose that researchers in this position produced encouraging progresses, which, contrary to precision medicine, have the potential to reach and benefit all cancer patients. My favorite example is the anti-tumorigenic activity of MTH1 inhibition. MTH1 is a protein responsible for sanitizing damaged bases prior incorporation into the DNA. Gad et al. found this activity to be essential only in cancer cells, possibly because bases are damaged at much higher rate than DNA and this process is continuous throughout cancer evolution. Accordingly, inhibition of MTH1 was found to eradicate cancer in patient-derived mouse xenografts (*156*). This work demonstrates that there is still a window to kill cancer by attacking the phenotype rather than its genetic background.

In my opinion, the greatest breakthrough for cancer care is not precision medicine but an impressive advancement in medical technologies. NGS is starting to revolution healthcare. Thanks to progresses in the quality, speed, and cost-effectiveness of this technology, medical doctors have in their hands an instrument that is primarily changing the way they diagnose disease. A pivotal clinical study in 2013 reported the successful use of NGS in patients so referred for evaluation for a possible Mendelian disorder (*157*). However applications of NGS to inform clinical diagnostics were found also in non genetic diseases. An illuminating example is the case of a 14-year old boy with severe neurological conditions in which the diagnostic workup was unrevealing. NGS of the cerebrospinal fluid yielded 3,063,784 sequence reads of which 475 reads did not belong to the human genome, but rather mapped to the *Leptospiraceae* family. This corresponds to 0.02% of all reads. The patient was treated for neuroleptospirosis and subsequently discharged home close to his premorbid functional status (*158*). In cancer, NGS will on one hand provide information on the presence of cancer predisposition genes in a certain individual, hence aiding in the identification of population at risk (*159*), and on the other hand eventually fulfill the promise of precision medicine, by informing oncologists on clinically-actionable genomic aberrations in the patients' tumors (*160, 161*). However, advancements in medical technologies for cancer care go beyond NGS. Particular credit deserves in my opinion nanotechnology, in which some prototypes may replace the need for sophisticated cancer biology in order to effectively treat patients (*162*). Nonetheless, significant contributions came from the more classical fields of biomedical engineering as well as biochemistry, in the form of devices and assays that can aid cancer diagnosis and follow-up. I selected three representative examples. The first is a microdevice that tackles the long-standing problem of drug inefficacy in human tumors. Indeed, despite considerable effort is spent in finding druggable targets and demonstrating its efficacy in preclinical models (such as mice), the dominant

reason for failure in drug development is lack of efficacy in humans. Jonas et al. (*163*) designed a microdevice that simultaneously tests multiple drugs in the living tumor, in which it is implanted via a minimally invasive biopsy, and returns the individual drug or combinations of them that produced the best response. The second example is a diagnostic device developed by a company called Miroculus. This consists in a blood test, in which circulating microRNAs are profiled and this profile is matched to cancer signatures via machine learning to return real time the occurrence of a certain cancer type. Even if the technology is not published, it is based on the breakthrough discovery made in 2008 by Mitchell et al., who showed that circulating microRNAs are diagnostic in cancer patients (*164*). The third example is a biochemical assay that has the same scope as the above-mentioned implantable microdevice. In this case, cells extracted from a primary tumor are exposed to different drug treatments and a quick kinetic tracer analysis allows computing the induction of apoptosis accomplished by the different drugs, within 24 hours from the biopsy (*165*).

What about systems biology? How did it contribute to a progress in cancer care? Simply put, it did not. In my opinion, systems biology has delivered proofs-of-concept, but its major contributions stand in fundamental research. For example, a study often cited for its elegance was performed by Graham et al. (*166*). In this work, the researchers dissected how glucose deprivation leads to cell death. They demonstrated that a positive feedback loop involving cross-talk between metabolism and phospho-tyrosine signaling is activated upon glucose withdrawal, even in cancer cells where active tyrosine kinases are constitutively expressed. In the same fashion, another example is provided by the work by Komurov et al. (*167*), in which the authors investigated why a subtype of breast cancer cells acquired resistance to lapatinib. They discovered that resistant cells activated a response network induced glucose deprivation to counteract the drug toxicity. Some clinical applications in systems biology exist in the form of proofs-of-concept, even though it is not clear to me the boundary within which these projects qualify as systems biology. Systems biology *must* rely on emerging properties of networks of interacting components, as clearly demonstrated by the studies by Graham et al. and Komurov et al. The Institute for Systems Biology (Seattle, United States) is promoting the 100k wellness project (*168*), aimed at collecting a vast amount of biological and clinical data from a hundred thousand individuals. This data is crunched to deliver predictions on individual health status, to undertake actionable healthcare and critically inform the physicians. It is unknown at this stage how the researchers in the institute intend to mine this wealth of data, but I nourish little expectation on the fact that interactions between data points will be taken in account, as systems biology would dictate.

My view of cancer has been transformed by four year of scientific research. I have a deeper understanding on how cancer works, but I honestly cannot form an opinion on how this affects my expectations when cancer is under the lens of a clinician rather than a scientist. I know that some forms of cancer can be cured, as well as others are unescapably lethal. If I were diagnosed with cancer today, I would feel that my expertise is of little help in the task of encouraging myself that there is hope. The contrast between the scientific reality and the clinical reality of cancer is too overwhelming for me. Due to this, in the course of my PhD I have shied away from making any claim on potential cures for cancer. I would find that ludicrous. I do not know whether the cure of *any* cancer will come from precision medicine, rather than nanotechnology, molecular or systems biology, or a radical paradigm shift on the origin of cancer. But I have hope in science.

Dear science

I consider the doctoral thesis concluded with the previous section. However, I would like to take advantage of this last part to write a brief essay on science as I see it. My thesis is that scientific knowledge is accumulating at a pace lower than its full potential, due to (1) misconduct of academic research, (2) problematic dissemination of knowledge, and (3) irrelevance of most scientific research. The arguments in support of my thesis are largely, if not entirely, overlapping with the

view of scientific knowledge elaborated by Karl Popper, which I borrow and revisit to provide a concise line of argumentation, allowing me also to be brief.

It is helpful to give the definition of scientific knowledge. Knowledge is the ability of declaring something true or not, and truth is what corresponds to reality. I will not elaborate further what I intend by reality, as I believe that the intuitive understanding of this word by the reader is sufficient to comprehend my thesis. I insist on the distinction that Popper makes between truth and certainty, because it stands at the heart of what I perceive as vast misconduct in the world of academic research. We can conjecture that we know something because we can elaborate a theory that can be said true, in that it corresponds to reality, but we can never be certain about it. Certainty can never be achieved. This point to me is trivial, but I am continuously amazed when I am confronted with discordance on this matter, so I propose some arguments that will become handy also later on. Reality as we know it is in fact a purely sensorial experience, in that our senses represent the mean and the limit through which we can interface with reality. Arthur Schopenhauer called this sensorial interface a deceptive veil of Maya. I turn instead to the Kantian interpretation, which prescribes that our knowledge of reality is the imposition on reality of rules and theories produced by the cues we attain via our senses. Nothing is certain because *we* are the artifices of the rules and theories that shape reality. We build knowledge by producing rules and theories that do not contradict reality as we perceive it, in other words that are true. What is scientific knowledge then? It is the accumulation of knowledge through criticism. The duty of a scientist is to critically examine whether the rules and theories that he/she or others have proposed for a certain aspect of reality are true, attempt to refute them, and replace them with rules and theories that are true until subsequent refutation. Note that I believe that my only disagreement with Popper's philosophy of science stands in the idea that all we can do as scientists is to *prove* a theory as false, and never as true.

This brings me to the first part of my thesis, a critique towards misconduct of academic research. My first point is that the above arguments should be convincing that any academic research that proposes to demonstrate a theory or a rule as true or, even worse, as certain is a wrongdoing. (Dogmatism in science is a gruesome symptom of idiocy.) Indeed, every experiment we design, every observation we make, every logical statement we infer to perpetuate scientific knowledge has the sole scope to refute a theory or a rule. What is commonly perceived as a failed experiment is our success in the refutation of a theory. My second point is therefore that academic research is misconducted whenever a scientist ignores a refutation on purpose. The reason why a scientist would willingly ignore a refutation is that failure in the process of refutation produces so-called corroborative evidence. Corroborative evidence supports the theory rather than refuting it. However, dare not to use corroborative evidence as a proof of truth. It simply states our incapacity to realize a better knowledge of reality. Of course, we can be content with our current knowledge. But if we stop attempting to refute a theory or a rule, than we actively choose not to seek for truth anymore, hence we cease being scientists. If in the past we had been content with the Newtonian theory of gravitation, which has accumulated corroborative evidence for centuries thanks to its excellent predictions, we would have never realized that it is factually wrong. Again, we can be content with our current knowledge, most people are. Engineers are a great example of people who do no seek further truth but exploit the current theories to build technology. (A great deal of technologies is based on Newtonian laws.) The myth that we can prove a theory as true by means of corroborative evidence is best described as a fallacy, in my opinion, by Nicholas Taleb's theory of Black Swans. This theory convincingly re-elaborates the consequences of an instinctive psychological human bias known as the confirmation bias. Whenever we ignore a refutation in favor of corroborative evidence, we commit a fallacy and foster misconduct in academic research.

The second part of my thesis is that, regrettably, many scientific journals seem to be unaware about the process of accumulation of scientific knowledge, despite being the main channel for its dissemination. Since scientists can only prove the refutation of a theory, any request to demonstrate

its truth is ill defined. The role of scientific editors and peer reviewers should be the role of any scientist: a constant attempt to refute what proposed by the authors. Hence, any request for new experiments, any instance of doubt about novelty, and any suggestion of rejection are justified solely in view that the theory or rule offered by the authors can be proven false. As a young practitioner, all too often I witnessed that this does not remotely apply to a peer review process. My most dreadful experiences came from reviews in which fellow scientists object that the statistical confidence on a certain hypothesis is not sufficient to prove a theory or rule as true. This statement is heretical under so many aspects. The purpose of statistical confidence is to *score* our propensity to *reject* a certain hypothesis behind a theory that we believe is true. Both the fact that we are measuring a propensity and the fact that statistical confidence can only reject a hypothesis are sufficient to discard such review as idiotic. More on statistical confidence. A scientist uses tests that reveal a statistical confidence on a certain hypothesis to convince himself/herself that what he/she observes is not attributable to chance, in other words to reject the theory that chance can generate such hypothesis. There are two implications here. First, statistical confidence measures a propensity. It is down to the scientist to gauge this metric and decide whether the theory alternative to chance is more likely to be true. The second point descends straightway. Statistical confidence is a mean, not a purpose. Any observation adjusted to achieve statistical confidence is to repudiate. This act entices the misconduct of ignoring a refutation in favor of corroborative evidence. Trivially, the correct way to proceed is to refute the current theory or rule and replace it with a theory or rule that matches reality.

The third part of my thesis might sound controversial, but I believe it derives quite straightforward from the arguments I presented above. Scientific knowledge must start from a theory that we, as scientists, attempt to critically refute. If we fail in this process than we are corroborating the theory. I argue that most scientific research is dedicated to corroboration. This research did not propose any advancement in our search for truth and it is hence irrelevant. If we succeed in the refutation, than we ought to propose a new theory, which other fellow scientists and we are designated to criticize. Most of the time that scientific research incurs in this stage, which I do not believe to happen often, the scientists propose a variation of the current theory that is better than the refuted theory in its adherence to reality, and that we can thus endorse as true. Now, one may judge a variation of the current theory as a relevant advancement in scientific knowledge. I do not doubt that this process constitutes the backbone of accumulation of scientific knowledge. However, I can hardly find excitement in that. It seems to me that a variation to a refuted theory was somehow embedded in the current theory, and just awaited to be discovered. There is a third way. It occurs when the scientists cannot propose a variation of the refuted theory as a replacement to the current theory. The evidences simply do not support any current theory or rule that we believe to be true, and they are perceived as anomalies. Thomas Kuhn termed these situations in science as crises. Eventually, brilliant personalities in the history of science propose a radically novel theory that renders the anomalies adherent to reality. In the vocabulary of Kuhn, these moments are the paradigm shifts that we commonly refer to as scientific revolutions. In this fashion, science is articulated by accumulation of scientific knowledge through variations of the current theory until the theory produces inexplicable anomalies that are solved by scientific revolutions. In this sense, I regard scientific research relevant only if it fuels a scientific revolution. As a consequence, I value scientific research relevant if it has the ambition to (a) successfully refute the current theory and (b) propose a radically new theory adherent to reality because it solves a crisis.

In my short experience of scientist, I have formulated my personal recipe in the attitude towards scientific research. There are fundamentally two main ingredients: skepticism and passion. Skepticism is an extreme of the rational criticism necessary to the practice of science. It underlies an assumption of suspicion, because it acknowledges the fact that we scientists are humans susceptible to fallacies. Passion is fundamental because I realized that in the actual practice of science it is rare to find yourself in the situation where your research is relevant, as previously

defined. However, we can undertake an emotional involvement in our quest to ameliorate the adherence of reality of the current theory. Most of the time we fail in this scope, and this causes frustration, which is irrational in consideration that we just succeeded in refuting a theory that is not true, but which is comprehensible in light of our passion. Sometimes we succeed, and passion rewards us, even though we are most likely simply perpetuating the current theory with a minor incidental variation. But, rarely, we might be causing a scientific revolution.

Acknowledgements

Jens, my gratitude towards you unfolds in so many different ways. Under a purely pragmatic point of view, the economical investment that you placed to educate me, to support my research ideas and my collaborations, and occasionally to entertain me, deserves my deepest and humblest gratitude. Under less pragmatic points of view, I thank you also for the following things: bearing with me, in particular those times that my stubbornness, which I awkwardly attribute to my refusal to enslave my rationality, forced you to take unease positions and those times that my lack of elegant manners (I did not know how to formulate this) perturbed the serenity of your working environment; motivating me, and I am fully aware that you consider this a personal mission, but that does not make it less honorable; believing in me, especially when it was time to take in consideration my PhD "application". For all this and a lot more, thank you.

I had the privilege to collaborate with several scientists, to whom I am most indebted for their incredibly clever insights, their effectiveness and productivity, and their passion for our research. This list must start with my co-supervisor Intawat, who I also thank for his indispensable guidance in my early steps in science (your first teaching was what SNP stands for, just to give an idea of my vast ignorance. By the way, sorry for never following up on that nice research idea). I temporarily turn to acknowledge our international collaborators. I am most grateful to Almut Schulze (now at the University of Würzburg) and her then-PhD student Heike Miess at the Cancer Research UK's London Research Institute. A great team formed in our attempt to design a diagnostic test for kidney cancer. My acknowledgements go first and foremost to Umberto Basso and his team at the IOV – IRCCS of Padova, who first accepted to embrace this project. I thank Nicola Volpi and his team at the University of Modena and Reggio Emilia, who must be commended for the outstanding quality of their experiments and timing in the delivery of the results. I was proud to collaborate with Sven Lundstam and Ulrika Stierner at the Sahlgrenska University Hospital in Göteborg and with Martin Johansson and Helén Nilsson at the Skåne University Hospital in Malmö, who I also thank for their early curiosity and interest in our research. On a different project, I thank Costas Lyssiotis and Lewis Cantley at the Weill Cornell Medical College in New York City for involving us in one of the trickiest puzzles of cancer metabolism I faced during my PhD. Back to in-house collaborations. I acknowledge my (professional) collaboration with Leif, because you are freaking (took me a while to replace the word I had in mind) smart and for, in the words of Adil, "delivering *always* a work of excellent quality". Secondly, I thank Kaisa (one day our work will be published!) who has also the merit of conveying an unconditioned passion for science. Lastly, I have had an unwritten collaborator for endless constructive discussions in *any* of my works, Adil. Thank you.

Beyond those people who actively contributed to the findings in this thesis stands an army of people that led me on the right track. Within our research group, I cannot forget to acknowledge the aid I got from Tobias in my early days of PhD, as well as Sergio, Amir, Natapol, and Fredrik. A special mention goes to Rasmus, whose insights can be at the least defined as brilliant. Many troubles I had with IT were dazzlingly solved by Shaq, and the staff at C3SE, thanks for everything. I thank Rahul and Petri for eye-opening discussions and, surprise surprise, Sakda, Marie, Ximena, and Julia for lab (!) support (yes, I labeled some samples once or twice). I acknowledge Dina, Verena, Ivan, and José for undertaking the effort to revise at different times our purely computational manuscripts. I am indebted for their continuous pragmatic support to Erica and Martina as well as the rest of the administrative staff in the department, Helena and Anna on top. Outside the group, my ignorance in statistics (which I impute to the incipient neurodegeneration of the professor of statistics in my M. Eng. program that caused me to study actual statistics *during* the exam, in a day which I will always remember for its paroxysm), I was saying, my ignorance in statistics led me to reach for the department of mathematics here at Chalmers multiple times. Hence I thank Erik Kristiansson and his lab members, Mariana, Viktor, Anna, José, and Fredrik, and Olle Nerman for their advices. I am thankful to Pontus Hjortskog, David Jensen, Kenny Nilsson, and Luuk van Egeraat, undergraduate students in Göteborg University who built a database for the AraX project. Finally, I recall crucial discussions with Martin Eilers (University of Würzburg), Börje Haraldsson (Göteborg University), Pernille Hojman and Bente Pedersen (Centre of Inflammation and Metabolism in Copenhagen), Maria Pinhal (Federal University of Sao Paulo), Eduard Reznik (Memorial Sloan Kettering Cancer Center in New York City), for which I am most grateful.

A deep thank goes to the Alice and Knut Wallenberg Foundation and the Chalmers Foundation for their financial support to our research. I am also (morally) indebted with these funding agencies for sponsoring

my participation in various conferences and courses: Assar Gabrielssons Fond, Stiftelsen ÅForsk, Nils Pihlblad Stiftelsen, and the (defunct) Department of Chemical and Biological Engineering in Chalmers.

This concludes the formal part of my acknowledgements, which I prioritized because I feel morally bounded. Nevertheless, the acknowledgements that follow should be perceived equally important because they refer to instances that made my life in Sweden an amazing experience.

This thesis is a tribute to music, and it is constellated by references that can only partially satisfy my gratitude towards it.

One realizes how cool is to be a Sysbio member when confronted with the multitude of people in the group in occasions like the group outings, PhD and housewarming parties, the Christmas and Spring parties, ski trips, etc. I think that in each of these events I formulated the thought that this group of people is fantastic. I am so freaking (I like this replacement adjective) thankful. I would like to extend my gratitude to the Indbio members, especially Heidi, Emma, Helén, and Josh. These following Sysbio guys deserve a special mention for entertaining me beyond standard expectation: Stefan, Bouke, Verena, Sakda, FK, Avlant, Eugene, José, Shaq, Min, Marina, Marie, Natapol, Saeed, Raphael, Julia, Magnus, JC, and Ximena. A heartwarming thanks to my unforgettable office mates Adil and Knuf, my rescuing mentor Dina, my Italian buddy Martina, my sailor-in-arms Kaisa, my joyful penguin Tobias, my lost in translation mate Amir. I wish to thank also this people related to Sysbio yet proudly distinct: Isa, Lea, Eleonore, and Bart. And finally I am way too grateful to this group of people for the extra fun beyond office walls: Alexandra, Anastasia, Ed, Michi, Jullesson, and my unauthorized elder brother Florian.

Outside these walls I had the luck and privilege to bind with great people who rendered this Swedish experience memorable. Thanks to the happy GoBiGgers, Anna, Fredrik, Johan, and Viktor. Thanks to all the international students I made friend with, with honorable mention to Chelsea, Mattia, Patrick and Stefania. Finally, thanks to my "Swedish" gang, I am so proud of you guys, in particular Lukasz, Dagmara, Sylvain, Elena, and Mitra.

My life in Sweden was blessed by my friendship with five people, who had a great impact, and these I would like to acknowledge individually. Petri, you are the greatest flatmate I have ever had: your presence made me a bit less Italian and a bit more Estonian, which I repaid by making you far from being Italian and way more Estonian. Rachele, non so cosa darei per la gioia di rasare a zero i riccioli rossi, ma è alla loro vista che mi rincuoro di essere tornato a casa: grazie per il tuo perenne supporto. Kees, you defied my criticism more than any other person I have met in my life, but I am more grateful for being the longest standing friend I had the fortune to encounter in Sweden. Elias, you are so immensely Swedish (98%), thank you from the bottom of my heart. Eddy, it is just too hard for my mind to verbalize my gratitude: I can only write that I am grateful for everything you did with me and *for* me, and in particular for everything you did for me without me knowing.

La mia famiglia merita un ringraziamento speciale, per il suo eterno e incondizionato supporto. Grazie mamma e grazie papà. E ai miei amici in Italia ne riservo un altro, che traduco pari pari dalla conclusione della tesi di laurea magistrale: "È immensa e incommensurabile la mia devozione a voi che siete la mia perenne fonte di perserveranza, curiosità, creatività, passione e, in fin dei conti, vita". Mi date la linfa.
Grazie a tutti di tutto.

References

1. A. Jemal *et al.*, Global cancer statistics. *CA: a cancer journal for clinicians* **61**, 69-90 (2011).
2. L. F. Rutten, B. W. Hesse, R. P. Moser, K. D. McCaul, A. J. Rothman, Public perceptions of cancer prevention, screening, and survival: comparison with state-of-science evidence for colon, skin, and lung cancer. *Journal of cancer education : the official journal of the American Association for Cancer Education* **24**, 40-48 (2009).
3. K. I. Baghurst, P. A. Baghurst, S. J. Record, Public perceptions of the role of dietary and other environmental factors in cancer causation or prevention. *Journal of epidemiology and community health* **46**, 120-126 (1992).
4. P. Rous, A Transmissible Avian Neoplasm. (Sarcoma of the Common Fowl.). *The Journal of experimental medicine* **12**, 696-705 (1910).
5. S. Mukherjee, *The emperor of all maladies : a biography of cancer*. (Scribner, New York, ed. 1st Scribner trade paperback, 2011), pp. xviii, 573, 512 p., 578 p. of plates.
6. K. Yamagiwa, K. Ichikawa, Experimental study of the pathogenesis of carcinoma. *CA: a cancer journal for clinicians* **27**, 174-181 (1977).
7. J. D. Watson, F. H. Crick, Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature* **171**, 737-738 (1953).
8. B. N. Ames, Identifying environmental chemicals causing mutations and cancer. *Science* **204**, 587-593 (1979).
9. S. A. Forbes *et al.*, COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic acids research* **43**, D805-811 (2015).
10. M. S. Lawrence *et al.*, Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* **505**, 495-501 (2014).
11. E. Taparowsky *et al.*, Activation of the T24 bladder carcinoma transforming gene is linked to a single amino acid change. *Nature* **300**, 762-765 (1982).
12. P. C. Nowell, The clonal evolution of tumor cell populations. *Science* **194**, 23-28 (1976).
13. R. A. Weinberg, *The biology of cancer*. (Garland Science, New York, NY, US, ed. Second edition., 2014), pp. xx, 875 pages.
14. M. Gerlinger *et al.*, Cancer: Evolution Within a Lifetime. *Annu Rev Genet* **48**, 215-236 (2014).
15. H. Rajagopalan, M. A. Nowak, B. Vogelstein, C. Lengauer, The significance of unstable chromosomes in colorectal cancer. *Nature reviews. Cancer* **3**, 695-701 (2003).
16. S. Nik-Zainal *et al.*, The life history of 21 breast cancers. *Cell* **149**, 994-1007 (2012).
17. M. Gerlinger *et al.*, Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *The New England journal of medicine* **366**, 883-892 (2012).
18. B. E. Johnson *et al.*, Mutational analysis reveals the origin and therapy-driven evolution of recurrent glioma. *Science* **343**, 189-193 (2014).
19. L. Chin, J. N. Andersen, P. A. Futreal, Cancer genomics: from discovery science to personalized medicine. *Nature medicine* **17**, 297-303 (2011).
20. R. Chen *et al.*, Personal Omics Profiling Reveals Dynamic Molecular and Medical Phenotypes. *Cell* **148**, 1293-1307 (2012).
21. K. A. Hoadley *et al.*, Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. *Cell* **158**, 929-944 (2014).
22. R. A. Weinberg, Coming full circle-from endless complexity to simplicity and back again. *Cell* **157**, 267-271 (2014).
23. D. Hanahan, R. A. Weinberg, The hallmarks of cancer. *Cell* **100**, 57-70 (2000).
24. D. Hanahan, R. A. Weinberg, Hallmarks of cancer: the next generation. *Cell* **144**, 646-674 (2011).
25. G. I. Evan, K. H. Vousden, Proliferation, cell cycle and apoptosis in cancer. *Nature* **411**, 342-348 (2001).
26. B. Vogelstein, K. W. Kinzler, Cancer genes and the pathways they control. *Nature medicine* **10**, 789-799 (2004).
27. B. Vogelstein *et al.*, Cancer genome landscapes. *Science* **339**, 1546-1558 (2013).
28. D. L. Kasper, T. R. Harrison, *Harrison's principles of internal medicine*. (McGraw-Hill, Medical Pub. Division, New York, ed. 16th, 2005).
29. M. Imielinski *et al.*, Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing. *Cell* **150**, 1107-1120 (2012).

30. D. W. Parsons *et al.*, An integrated genomic analysis of human glioblastoma multiforme. *Science* **321**, 1807-1812 (2008).
31. E. R. Mardis *et al.*, Recurring mutations found by sequencing an acute myeloid leukemia genome. *The New England journal of medicine* **361**, 1058-1066 (2009).
32. S. Gross *et al.*, Cancer-associated metabolite 2-hydroxyglutarate accumulates in acute myelogenous leukemia with isocitrate dehydrogenase 1 and 2 mutations. *The Journal of experimental medicine* **207**, 339-344 (2010).
33. P. S. Ward, C. B. Thompson, Metabolic reprogramming: a cancer hallmark even warburg did not anticipate. *Cancer cell* **21**, 297-308 (2012).
34. O. Warburg, On the origin of cancer cells. *Science* **123**, 309-314 (1956).
35. M. Agathocleous *et al.*, Metabolic differentiation in the embryonic retina. *Nature cell biology* **14**, 859-864 (2012).
36. E. Gottlieb, I. P. Tomlinson, Mitochondrial tumour suppressors: a genetic and biochemical update. *Nature reviews. Cancer* **5**, 857-866 (2005).
37. H. R. Christofk *et al.*, The M2 splice isoform of pyruvate kinase is important for cancer metabolism and tumour growth. *Nature* **452**, 230-U274 (2008).
38. M. G. Vander Heiden, L. C. Cantley, C. B. Thompson, Understanding the Warburg effect: the metabolic requirements of cell proliferation. *Science* **324**, 1029-1033 (2009).
39. R. B. Hamanaka, N. S. Chandel, Warburg Effect and Redox Balance. *Science* **334**, 1219-1220 (2011).
40. R. A. Cairns, I. S. Harris, T. W. Mak, Regulation of cancer cell metabolism. *Nature reviews. Cancer* **11**, 85-95 (2011).
41. L. K. Boroughs, R. J. DeBerardinis, Metabolic pathways promoting cancer cell survival and growth. *Nature cell biology* **17**, 351-359 (2015).
42. C. R. Berkers, O. D. Maddocks, E. C. Cheung, I. Mor, K. H. Vousden, Metabolic regulation by p53 family members. *Cell metabolism* **18**, 617-633 (2013).
43. T. Li *et al.*, Tumor suppression in the absence of p53-mediated cell-cycle arrest, apoptosis, and senescence. *Cell* **149**, 1269-1283 (2012).
44. A. R. Choudhury *et al.*, Cdkn1a deletion improves stem cell function and lifespan of mice with dysfunctional telomeres without accelerating cancer formation. *Nature genetics* **39**, 99-105 (2007).
45. H. Ying *et al.*, Oncogenic Kras maintains pancreatic tumors through regulation of anabolic glucose metabolism. *Cell* **149**, 656-670 (2012).
46. G. M. DeNicola *et al.*, Oncogene-induced Nrf2 transcription promotes ROS detoxification and tumorigenesis. *Nature* **475**, 106-109 (2011).
47. F. Vazquez *et al.*, PGC1alpha expression defines a subset of human melanoma tumors with increased mitochondrial capacity and resistance to oxidative stress. *Cancer cell* **23**, 287-301 (2013).
48. P. Caro *et al.*, Metabolic signatures uncover distinct targets in molecular subsets of diffuse large B cell lymphoma. *Cancer cell* **22**, 547-560 (2012).
49. R. A. Gatenby, R. J. Gillies, Why do cancers have high aerobic glycolysis? *Nature reviews. Cancer* **4**, 891-899 (2004).
50. C. V. Dang, Glutaminolysis: supplying carbon or nitrogen or both for cancer cells? *Cell Cycle* **9**, 3884-3886 (2010).
51. J. Son *et al.*, Glutamine supports pancreatic cancer growth through a KRAS-regulated metabolic pathway. *Nature* **496**, 101-+ (2013).
52. F. Baenke, B. Peck, H. Miess, A. Schulze, Hooked on fat: the role of lipid synthesis in cancer metabolism and tumour development. *Disease models & mechanisms* **6**, 1353-1363 (2013).
53. P. M. Tedeschi *et al.*, Contribution of serine, folate and glycine metabolism to the ATP, NADPH and purine requirements of cancer cells. *Cell death & disease* **4**, e877 (2013).
54. H. Cheong, C. Lu, T. Lindsten, C. B. Thompson, Therapeutic targets in cancer cell metabolism and autophagy. *Nature biotechnology* **30**, 671-678 (2012).
55. J. Y. Guo, B. Xia, E. White, Autophagy-mediated tumor promotion. *Cell* **155**, 1216-1219 (2013).
56. D. Trachootham, J. Alexandre, P. Huang, Targeting cancer cells by ROS-mediated mechanisms: a radical therapeutic approach? *Nature Reviews Drug Discovery* **8**, 579-591 (2009).
57. W. W. Wheaton *et al.*, Metformin inhibits mitochondrial complex I of cancer cells to reduce tumorigenesis. *eLife* **3**, e02242 (2014).
58. K. Birsoy *et al.*, Metabolic determinants of cancer cell sensitivity to glucose limitation and biguanides. *Nature* **508**, 108-112 (2014).

59. A. Viale *et al.*, Oncogene ablation-resistant pancreatic cancer cells depend on mitochondrial function. *Nature* **514**, 628-632 (2014).
60. I. Elia, R. Schmieder, S. Christen, S. M. Fendt, Organ-Specific Cancer Metabolism and Its Potential for Therapy. *Handbook of experimental pharmacology*, (2015).
61. R. Possemato *et al.*, Functional genomics reveal that the serine synthesis pathway is essential in breast cancer. *Nature* **476**, 346-350 (2011).
62. O. D. Maddocks *et al.*, Serine starvation induces stress and p53-dependent metabolic remodelling in cancer cells. *Nature* **493**, 542-546 (2013).
63. W. C. Zhang *et al.*, Glycine decarboxylase activity drives non-small cell lung cancer tumor-initiating cells and tumorigenesis. *Cell* **148**, 259-272 (2012).
64. D. Kim *et al.*, SHMT2 drives glioma cell survival in ischaemia but imposes a dependence on glycine clearance. *Nature* **520**, 363-367 (2015).
65. J. D. Regan, N. Vodopick, S. Takeda, W. H. Lee, F. M. Faulcon, Serine requirement in leukemic and normal blood cells. *Science* **163**, 1452-1453 (1969).
66. M. G. Vander Heiden, Targeting cancer metabolism: a therapeutic window opens. *Nature reviews. Drug discovery* **10**, 671-684 (2011).
67. A. Schulze, A. L. Harris, How cancer metabolism is tuned for proliferation and vulnerable to disruption (vol 491, pg 364, 2012). *Nature* **494**, 130-130 (2013).
68. A. L. Barabasi, http://barabasilab.neu.edu/networksciencebook/, Ed. (2012).
69. M. Mitchell, *Complexity : a guided tour*. (Oxford University Press, Oxford England ; New York, 2009), pp. xvi, 349 p.
70. P. Erdos, A. Renyi, On the Evolution of Random Graphs. *B Int Statist Inst* **38**, 343-347 (1960).
71. D. J. Watts, S. H. Strogatz, Collective dynamics of 'small-world' networks. *Nature* **393**, 440-442 (1998).
72. A. L. Barabasi, R. Albert, Emergence of scaling in random networks. *Science* **286**, 509-512 (1999).
73. G. Lima-Mendez, J. van Helden, The powerful law of the power law and other myths in network biology. *Molecular bioSystems* **5**, 1482-1493 (2009).
74. M. P. Tordera, Complexity in Asthma: Inflammation and Scale-Free Networks. *Arch Bronconeumol* **45**, 459-465 (2009).
75. A. L. Barabasi, Z. N. Oltvai, Network biology: understanding the cell's functional organization. *Nature reviews. Genetics* **5**, 101-113 (2004).
76. M. Cassman, World Technology Evaluation Center., *Systems biology : international research and development*. (Springer, Dordrecht, The Netherlands, 2007), pp. xvii, 262 p.
77. I. Thiele, B. O. Palsson, A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nature protocols* **5**, 93-121 (2010).
78. E. J. O'Brien, J. M. Monk, B. O. Palsson, Using Genome-scale Models to Predict Biological Capabilities. *Cell* **161**, 971-987 (2015).
79. L. Varemo, I. Nookaew, J. Nielsen, Novel insights into obesity and diabetes through genome-scale metabolic modeling. *Frontiers in physiology* **4**, 92 (2013).
80. J. D. Orth, I. Thiele, B. O. Palsson, What is flux balance analysis? *Nature biotechnology* **28**, 245-248 (2010).
81. N. E. Lewis, H. Nagarajan, B. O. Palsson, Constraining the metabolic genotype-phenotype relationship using a phylogeny of in silico methods. *Nature Reviews Microbiology* **10**, 291-305 (2012).
82. A. Bordbar, J. M. Monk, Z. A. King, B. O. Palsson, Constraint-based models predict metabolic and associated cellular functions. *Nature reviews. Genetics* **15**, 107-120 (2014).
83. H. Ma *et al.*, The Edinburgh human metabolic network reconstruction and its functional analysis. *Molecular systems biology* **3**, 135 (2007).
84. N. C. Duarte *et al.*, Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proceedings of the National Academy of Sciences of the United States of America* **104**, 1777-1782 (2007).
85. A. Mardinoglu *et al.*, Genome-scale metabolic modelling of hepatocytes reveals serine deficiency in patients with non-alcoholic fatty liver disease. *Nature communications* **5**, 3083 (2014).
86. A. Mardinoglu *et al.*, Integration of clinical data with a genome-scale metabolic model of the human adipocyte. *Molecular systems biology* **9**, 649 (2013).
87. I. Thiele *et al.*, A community-driven global reconstruction of human metabolism. *Nature biotechnology* **31**, 419-425 (2013).

88. K. Yizhak, B. Chaneton, E. Gottlieb, E. Ruppin, Modeling cancer metabolism on a genome scale. *Molecular systems biology* **11**, (2015).
89. O. Folger *et al.*, Predicting selective drug targets in cancer through metabolic networks. *Molecular systems biology* **7**, 501 (2011).
90. C. Frezza *et al.*, Haem oxygenase is synthetically lethal with the tumour suppressor fumarate hydratase. *Nature* **477**, 225-228 (2011).
91. R. Agren *et al.*, Identification of anticancer drugs for hepatocellular carcinoma through personalized genome-scale metabolic modeling. *Molecular systems biology* **10**, 721 (2014).
92. K. Yizhak *et al.*, Phenotype-based cell-specific metabolic modeling reveals metabolic liabilities of cancer. *eLife* **3**, (2014).
93. K. Yizhak *et al.*, A computational study of the Warburg effect identifies metabolic targets inhibiting cancer migration. *Molecular systems biology* **10**, 744 (2014).
94. P. Ghaffari *et al.*, Identifying anti-growth factors for human cancer cell lines through genome-scale metabolic modeling. *Scientific reports* **5**, (2015).
95. L. Jerby *et al.*, Metabolic associations of reduced proliferation and oxidative stress in advanced breast cancer. *Cancer Res* **72**, 5712-5720 (2012).
96. A. Feizi, S. Bordel, Metabolic and protein interaction sub-networks controlling the proliferation rate of cancer cells and their impact on patient survival. *Scientific reports* **3**, 3041 (2013).
97. T. Shlomi, T. Benyamini, E. Gottlieb, R. Sharan, E. Ruppin, Genome-scale metabolic modeling elucidates the role of proliferative adaptation in causing the Warburg effect. *PLoS computational biology* **7**, e1002018 (2011).
98. S. Bordel, R. Agren, J. Nielsen, Sampling the solution space in genome-scale metabolic networks reveals transcriptional regulation in key enzymes. *PLoS computational biology* **6**, e1000859 (2010).
99. M. Ashburner *et al.*, Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics* **25**, 25-29 (2000).
100. M. Kanehisa *et al.*, Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic acids research* **42**, D199-205 (2014).
101. J. Hu *et al.*, Heterogeneity of tumor-induced gene expression changes in the human metabolic network. *Nature biotechnology*, (2013).
102. L. Varemo, J. Nielsen, I. Nookaew, Enriching the gene set analysis of genome-wide data by incorporating directionality of gene expression and combining statistical hypotheses and methods. *Nucleic acids research* **41**, 4378-4391 (2013).
103. R. Nilsson *et al.*, Metabolic enzyme expression highlights a key role for MTHFD2 and the mitochondrial folate pathway in cancer. *Nature communications* **5**, 3128 (2014).
104. B. Alberts, *Molecular biology of the cell*. (Garland Science, New York, ed. 4th, 2002), pp. xxxiv, 1548 p.
105. F. C. Neidhardt, R. Curtiss, *Escherichia coli and Salmonella : cellular and molecular biology*. (ASM Press, Washington, D.C., ed. 2nd, 1996).
106. M. Ehrenberg, C. G. Kurland, Costs of accuracy determined by a maximal growth rate constraint. *Quarterly reviews of biophysics* **17**, 45-82 (1984).
107. T. D. Halazonetis, V. G. Gorgoulis, J. Bartek, An oncogene-induced DNA damage model for cancer development. *Science* **319**, 1352-1355 (2008).
108. H. M. Temin, H. Rubin, Characteristics of an Assay for Rous Sarcoma Virus and Rous Sarcoma Cells in Tissue Culture. *Virology* **6**, 669-688 (1958).
109. G. K. Smyth, Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical applications in genetics and molecular biology* **3**, Article3 (2004).
110. F. Rapaport *et al.*, Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. *Genome biology* **14**, R95 (2013).
111. X. Yan, X. Su, *Linear regression analysis : theory and computing*. (World Scientific, Singapore ; Hackensack, NJ, 2009), pp. xix, 328 p.
112. E.-J. Wagenmakers, S. Farrell, AIC model selection using Akaike weights. *Psychonomic Bulletin & Review* **11**, 192-196 (2004).
113. D. Croft *et al.*, The Reactome pathway knowledgebase. *Nucleic acids research* **42**, D472-477 (2014).
114. D. W. Nebert, T. P. Dalton, The role of cytochrome P450 enzymes in endogenous signalling pathways and environmental carcinogenesis. *Nature Reviews Cancer* **6**, 947-960 (2006).

115. D. Z. Wang, R. N. Dubois, Eicosanoids and cancer. *Nature Reviews Cancer* **10**, 181-193 (2010).
116. B. I. Rini, S. C. Campbell, B. Escudier, Renal cell carcinoma. *Lancet* **373**, 1119-1132 (2009).
117. C. J. Creighton *et al.*, Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature*, (2013).
118. K. J. Kauffman, P. Prakash, J. S. Edwards, Advances in flux balance analysis. *Current opinion in biotechnology* **14**, 491-496 (2003).
119. M. Uhlen *et al.*, Towards a knowledge-based Human Protein Atlas. *Nature biotechnology* **28**, 1248-1250 (2010).
120. F. Gatto, I. Nookaew, J. Nielsen, Chromosome 3p loss of heterozygosity is associated with a unique metabolic network in clear cell renal carcinoma. *Proceedings of the National Academy of Sciences of the United States of America* **111**, E866-875 (2014).
121. K. R. Patil, J. Nielsen, Uncovering transcriptional regulation of metabolism by using metabolic network topology. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 2685-2689 (2005).
122. L. Varemo, F. Gatto, J. Nielsen, Kiwi: a tool for integration and visualization of network topology and gene-set analysis. *Bmc Bioinformatics* **15**, 408 (2014).
123. S. Pena-Llopis *et al.*, BAP1 loss defines a new class of renal cell carcinoma. *Nature genetics* **44**, 751-759 (2012).
124. Y. Wang *et al.*, Regulation of endocytosis via the oxygen-sensing pathway. *Nature medicine* **15**, 319-324 (2009).
125. N. Afratis *et al.*, Glycosaminoglycans: key players in cancer cell biology and treatment. *Febs J* **279**, 1177-1197 (2012).
126. R. L. Jackson, S. J. Busch, A. D. Cardin, Glycosaminoglycans: molecular properties, protein interactions, and role in physiological processes. *Physiological reviews* **71**, 481-539 (1991).
127. N. Volpi, R. J. Linhardt, High-performance liquid chromatography-mass spectrometry for mapping and sequencing glycosaminoglycan-derived oligosaccharides. *Nature protocols* **5**, 993-1004 (2010).
128. N. Volpi, F. Galeotti, B. Yang, R. J. Linhardt, Analysis of glycosaminoglycan-derived, precolumn, 2-aminoacridone-labeled disaccharides with LC-fluorescence and LC-MS detection. *Nature protocols* **9**, 541-558 (2014).
129. H. Moch *et al.*, Biomarkers in renal cancer. *Virchows Archiv : an international journal of pathology* **464**, 359-365 (2014).
130. R. Tibshirani, Regression shrinkage and selection via the Lasso. *J Roy Stat Soc B Met* **58**, 267-288 (1996).
131. B. E. Thacker, D. Xu, R. Lawrence, J. D. Esko, Heparan sulfate 3-O-sulfation: a rare modification in search of a function. *Matrix biology : journal of the International Society for Matrix Biology* **35**, 60-72 (2014).
132. C. Shao *et al.*, Comparative glycomics of leukocyte glycosaminoglycans. *Febs J* **280**, 2447-2461 (2013).
133. R. Moreno-Sanchez, S. Rodriguez-Enriquez, A. Marin-Hernandez, E. Saavedra, Energy metabolism in tumor cells. *Febs J* **274**, 1393-1418 (2007).
134. L. K. Boroughs, R. J. DeBerardinis, Metabolic pathways promoting cancer cell survival and growth. *Nature cell biology* **17**, 351-359 (2015).
135. G. Ciriello *et al.*, Emerging landscape of oncogenic signatures across human cancers. *Nature genetics* **45**, 1127-1133 (2013).
136. R. Agren *et al.*, Reconstruction of genome-scale active metabolic networks for 69 human cell types and 16 cancer types using INIT. *PLoS computational biology* **8**, e1002518 (2012).
137. Y. Wang, J. A. Eddy, N. D. Price, Reconstruction of genome-scale metabolic models for 126 human tissues using mCADRE. *BMC systems biology* **6**, 153 (2012).
138. S. Gupta, K. Chawla, Oncometabolomics in cancer research. *Expert review of proteomics* **10**, 325-336 (2013).
139. S. Patel, S. Ahmed, Emerging field of metabolomics: big promise for cancer biomarker identification and drug discovery. *Journal of pharmaceutical and biomedical analysis* **107**, 63-74 (2015).
140. M. O. Yuneva *et al.*, The metabolic profile of tumors depends on both the responsible genetic lesion and tissue type. *Cell metabolism* **15**, 157-170 (2012).
141. J. Fan *et al.*, Glutamine-driven oxidative phosphorylation is a major ATP source in transformed mammalian cells in both normoxia and hypoxia. *Molecular systems biology* **9**, 712 (2013).

142. X. F. Ye, J. Wang, W. T. Shi, J. He, Relationship between aspirin use after diagnosis of colorectal cancer and patient survival: a meta-analysis of observational studies. *British journal of cancer* **111**, 2172-2179 (2014).
143. A. T. Chan, S. Ogino, C. S. Fuchs, Aspirin Use and Survival After Diagnosis of Colorectal Cancer. *Jama-J Am Med Assoc* **302**, 649-659 (2009).
144. M. J. Thun, M. M. Namboodiri, C. W. Heath, Aspirin Use and Reduced Risk of Fatal Colon Cancer. *New Engl J Med* **325**, 1593-1596 (1991).
145. J. J. Li, M. D. Biggin, Gene expression. Statistics requantitates the central dogma. *Science* **347**, 1066-1067 (2015).
146. M. Vidal, M. E. Cusick, A. L. Barabasi, Interactome networks and human disease. *Cell* **144**, 986-998 (2011).
147. Y. H. Feng *et al.*, Global analysis of protein structural changes in complex proteomes. *Nature biotechnology* **32**, 1036-+ (2014).
148. S. C. Mack *et al.*, Epigenomic alterations define lethal CIMP-positive ependymomas of infancy. *Nature* **506**, 445-450 (2014).
149. I. Martincorena *et al.*, Tumor evolution. High burden and pervasive positive selection of somatic mutations in normal human skin. *Science* **348**, 880-886 (2015).
150. J. G. Paez *et al.*, EGFR mutations in lung cancer: Correlation with clinical response to gefitinib therapy. *Science* **304**, 1497-1500 (2004).
151. P. B. Chapman *et al.*, Improved Survival with Vemurafenib in Melanoma with BRAF V600E Mutation. *New Engl J Med* **364**, 2507-2516 (2011).
152. H. Joensuu *et al.*, Effect of the tyrosine kinase inhibitor STI571 in a patient with a metastatic gastrointestinal stromal tumor. *New Engl J Med* **344**, 1052-1056 (2001).
153. T. Wahlgren *et al.*, Treatment and overall survival in renal cell carcinoma: a Swedish population-based study (2000-2008). *British journal of cancer* **108**, 1541-1549 (2013).
154. R. Sikorski, B. Yao, Visualizing the Landscape of Selection Biomarkers in Current Phase III Oncology Clinical Trials. *Science translational medicine* **2**, (2010).
155. J. Brahmer *et al.*, Nivolumab versus Docetaxel in Advanced Squamous-Cell Non-Small-Cell Lung Cancer. *The New England journal of medicine* **373**, 123-135 (2015).
156. H. Gad *et al.*, MTH1 inhibition eradicates cancer by preventing sanitation of the dNTP pool. *Nature* **508**, 215-221 (2014).
157. Y. P. Yang *et al.*, Clinical Whole-Exome Sequencing for the Diagnosis of Mendelian Disorders. *New Engl J Med* **369**, 1502-1511 (2013).
158. M. R. Wilson *et al.*, Actionable Diagnosis of Neuroleptospirosis by Next-Generation Sequencing. *New Engl J Med* **370**, 2408-2417 (2014).
159. N. Rahman, Realizing the promise of cancer predisposition genes. *Nature* **505**, 302-308 (2014).
160. D. Tripathy, K. Harnden, K. Blackwell, M. Robson, Next generation sequencing and tumor mutation profiling: are we ready for routine use in the oncology clinic? *Bmc Med* **12**, (2014).
161. L. G. Biesecker, W. Burke, I. Kohane, S. E. Plon, R. Zimmern, Next-generation sequencing in the clinic: are we ready? *Nature Reviews Genetics* **13**, 818-824 (2012).
162. D. Peer *et al.*, Nanocarriers as an emerging platform for cancer therapy. *Nat Nanotechnol* **2**, 751-760 (2007).
163. O. Jonas *et al.*, An implantable microdevice to perform high-throughput in vivo drug sensitivity testing in tumors. *Science translational medicine* **7**, (2015).
164. P. S. Mitchell *et al.*, Circulating microRNAs as stable blood-based markers for cancer detection. *Proceedings of the National Academy of Sciences of the United States of America* **105**, 10513-10518 (2008).
165. J. Montero *et al.*, Drug-Induced Death Signaling Strategy Rapidly Predicts Cancer Response to Chemotherapy. *Cell* **160**, (2015).
166. N. A. Graham *et al.*, Glucose deprivation activates a metabolic and signaling amplification loop leading to cell death. *Molecular systems biology* **8**, (2012).
167. K. Komurov *et al.*, The glucose-deprivation network counteracts lapatinib-induced toxicity in resistant ErbB2-positive breast cancer cells. *Molecular systems biology* **8**, (2012).
168. L. Hood, N. D. Price, Demystifying Disease, Democratizing Health Care. *Science translational medicine* **6**, (2014).

PAPER I

Chromosome 3p loss of heterozygosity is associated with a unique metabolic network in clear cell renal carcinoma

F. Gatto, I. Nookaew, J. Nielsen

Proceedings of the National Academy of Sciences of the United States of America **111**, E866-875 (2014)

Chromosome 3p loss of heterozygosity is associated with a unique metabolic network in clear cell renal carcinoma

Francesco Gatto, Intawat Nookaew, and Jens Nielsen[1]

Department of Chemical and Biological Engineering, Chalmers University of Technology, 41296 Göteborg, Sweden

Edited by Robert Langer, Massachusetts Institute of Technology, Cambridge, MA, and approved January 24, 2014 (received for review October 11, 2013)

Several common oncogenic pathways have been implicated in the emergence of renowned metabolic features in cancer, which in turn are deemed essential for cancer proliferation and survival. However, the extent to which different cancers coordinate their metabolism to meet these requirements is largely unexplored. Here we show that even in the heterogeneity of metabolic regulation a distinct signature encompassed most cancers. On the other hand, clear cell renal cell carcinoma (ccRCC) strongly deviated in terms of metabolic gene expression changes, showing widespread down-regulation. We observed a metabolic shift that associates differential regulation of enzymes in one-carbon metabolism with high tumor stage and poor clinical outcome. A significant yet limited set of metabolic genes that explained the partial divergence of ccRCC metabolism correlated with loss of von Hippel-Lindau tumor suppressor (*VHL*) and a potential activation of signal transducer and activator of transcription 1. Further network-dependent analyses revealed unique defects in nucleotide, one-carbon, and glycerophospholipid metabolism at the transcript and protein level, which contrasts findings in other tumors. Notably, this behavior is recapitulated by recurrent loss of heterozygosity in multiple metabolic genes adjacent to *VHL*. This study therefore shows how loss of heterozygosity, hallmarked by *VHL* deletion in ccRCC, may uniquely shape tumor metabolism.

cancer metabolism | systems biology | genome-scale metabolic modeling | renal cancer

There is now widespread consensus that diversion of metabolism is among the most distinguished cancer phenotypes, and it is often postulated to characterize virtually all forms of cancer (1, 2). Indeed, many common oncogenic signaling pathways have been implicated in the emergence of specific metabolic features in cancer cells that have been associated with both survival and sustained abnormal proliferation rate (2–5). However, only a fraction of the metabolic reactions potentially occurring in a generic human cell are typically involved in such processes. Only recently a systemic study using transcriptional regulation has attempted to rule out the possibility that other metabolic processes in the network may achieve equal importance in cancer cells (6), and the idea that all cancer cells display a unique metabolic phenotype has spurred disputes that mainly highlighted a lack of comprehensive evidence (7). Taken together, we contend that only a systems perspective may help to elucidate the extent to which different cancer cells coordinate their metabolic activity.

In this context, systems biology approaches have been demonstrated to lead to the identification of altered metabolic processes in disease development with regard to those disorders that are driven or accompanied by metabolic reprogramming, including cancer (8–11). To this end, the reconstruction of genome-scale metabolic models (GEMs) is instrumental to knit high-throughput data into the metabolic network topology. Such integrative and network-dependent analysis enables prediction of how systems-level perturbations are translated into alterations in distinct and biologically meaningful modules and, at the same time, elucidation of genotype–phenotype relationships (12).

Results

Distinct Changes in Metabolic Gene and Protein Expression in Tumors. Until recently (6, 13, 14) it has been largely overlooked (*i*) the extent to which the metabolic phenotype is dissimilar with respect to healthy cells, and (*ii*) the extent to which it affects the complete metabolic network. We therefore used a GEM of the human cell and integrated high-dimension datasets of omics data, from both tumor-adjacent normal and cancer tissues. GEMs are models that account for all known reactions and matched metabolites in a cell and include the current knowledge for gene–protein reaction associations for each reaction. Here we used the human metabolic reaction (HMR) model, which comprises 7,943 reactions, 3,158 unique metabolites across eight compartments, and 3,674 genes and represents the most comprehensive compilation of human metabolic reactions (15). As for the omics data, we focused on RNAseq gene expression profiles and immunohistochemical proteomics. For cancer samples, we retrieved 539 transcriptomes and 25 proteomes, whereas for tumor-adjacent normal samples we retrieved 257 transcriptomes and 74 proteomes (*SI Appendix*, Table S1 and Dataset S1). We focused to include gene products that overlapped with the list of 3,674 genes in HMR. The HMR coverage was 97% for the transcript profiles in all cancers and tumor-adjacent normal samples. As for the protein profiles, because the protein coverage was heterogeneous across the samples, the HMR coverage was either 18% or 45%, depending on whether both tumor-adjacent normal and cancer samples or only cancer samples were pooled, respectively (*SI Appendix*, *SI Materials and*

Significance

It is suggested that regulation of metabolism is a point of convergence of many different cancer-associated pathways. Here we challenged the validity of this assertion and verified that a transversal metabolic signature in cancer emerges chiefly in the regulation of nucleotide metabolism. However, the most common form of renal cancer deviates from this behavior and presents some defects in its metabolic network not present in the normal kidney and unseen in other tumors. Notably, reduced copy number in key metabolic genes located adjacent to *VHL* (a tumor suppressor gene frequently deleted in this cancer) recapitulates these defects. These results are suggestive that recurrent chromosomal loss of heterozygosity in cancer may uniquely shape the metabolic network.

Author contributions: F.G., I.N., and J.N. designed research; F.G. performed research; F.G. and I.N. analyzed data; and F.G. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

Data deposition: All cancer genome-scale metabolic models are available through www.metabolicatlas.com.

[1]To whom correspondence should be addressed. E-mail: nielsenj@chalmers.se.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1319196111/-/DCSupplemental.

Methods). Even if metabolic-related proteins had lesser coverage, they are fairly representative for most canonical metabolic pathways (*SI Appendix*, Fig. S1).

The degree of similarity in metabolic gene expression between cancer and tumor-adjacent normal samples was assessed using principal component analysis (PCA) and mutual correlation-based hierarchical clustering. First, the similarity in the abundance of all metabolic transcripts across cancer and tumor-adjacent normal samples was evaluated (*SI Appendix*, Figs. S2 and S3). Both PCA and hierarchical clustering show that a group of cancer samples displays a substantial deviation from the general transcriptional pattern of most cancers. However, if these cancer samples are neglected, PCA reveals a consistent transcriptional response in different cancers opposed to tumor-adjacent normal samples, which is independent of cancer type and remarkable because the control samples were obtained adjacent to the tumor (note that the first principal component was neglected because it seems to account for few outliers; *SI Appendix*, Fig. S4). Conversely, hierarchical clustering shows a higher similarity in the gene expression of tissue-specific samples rather than across all cancer-labeled samples (*SI Appendix*, Fig. S3). This shows that cancers undergo a considerable alteration in metabolic gene expression profiles, but they also retain substantial similarity with the regulation of metabolic gene products of their matched tumor-adjacent normal tissue, in accordance with a recent study (6). This led us to speculate that although cancer samples regulate only a subset of metabolic genes upon transformation and preserve the expression of the remaining as in the tissue of origin, this regulation may be consistent and orchestrated across different cancer types. To verify this, PCA and hierarchical clustering were performed for those metabolic genes (~20%) that changed expression at statistical significance and across at least four histological cancer types, thereby subtracting the effect of tissue of origin (Fig. 1*A* and *SI Appendix*, Fig. S5 and Table S2). In both analyses, the distinction between most cancer and tumor-adjacent normal samples becomes apparent. In particular, hierarchical clustering provides clear evidence for the fact that most cancer samples modulate the expression of a distinct group of metabolic transcripts in a similar fashion, regardless of their histological classification. Next, multiple correspondence analysis (MCA) was performed to check whether the conclusions above also hold at the level of protein expression. Accordingly, proteomics data confirm that the expression of metabolic gene products is more similar between cancer samples than to normal tissues, which are distinctly separated (*SI Appendix*, Fig. S6*A*). However, within cancer samples no obvious cluster emerged, perhaps owing to less coverage (*SI Appendix*, Fig. S6*B*). Taken together, these analyses suggest that the transformation entails a partial yet significant remodeling of metabolic regulation, both at the transcript and protein level, which is transversal and to some extent coordinated within the disease phenotype and does not overlap with that of the tumor-adjacent normal tissue.

Deviation of the Transcriptional Program in Clear Cell Renal Carcinoma Metabolism. The above conclusion only holds provided that the cluster of deviating samples is not taken into account. We explored the nature of this cluster by correlating samples with available clinical data, and strikingly, all samples belonging to this cluster share a common histological type [i.e., clear cell renal cell carcinoma (ccRCC)] (*SI Appendix*, Figs. S2 and S3). Furthermore, such an anomalous profile is not attributable to an inherent elevated metabolic activity of the tissue of origin: when PCA was performed on the reduced pool of metabolic genes that significantly changed expression in most cancer types, ccRCC samples still separated clearly (Fig. 1*A* and *SI Appendix*, Fig. S5). Additionally, we noticed that papillary cell renal cell carcinoma samples present in the previous analyses did not overlap in terms of transcript abundance with ccRCC (*SI Appendix*, Fig. S7) and did not correlate with the previously neglected first principal

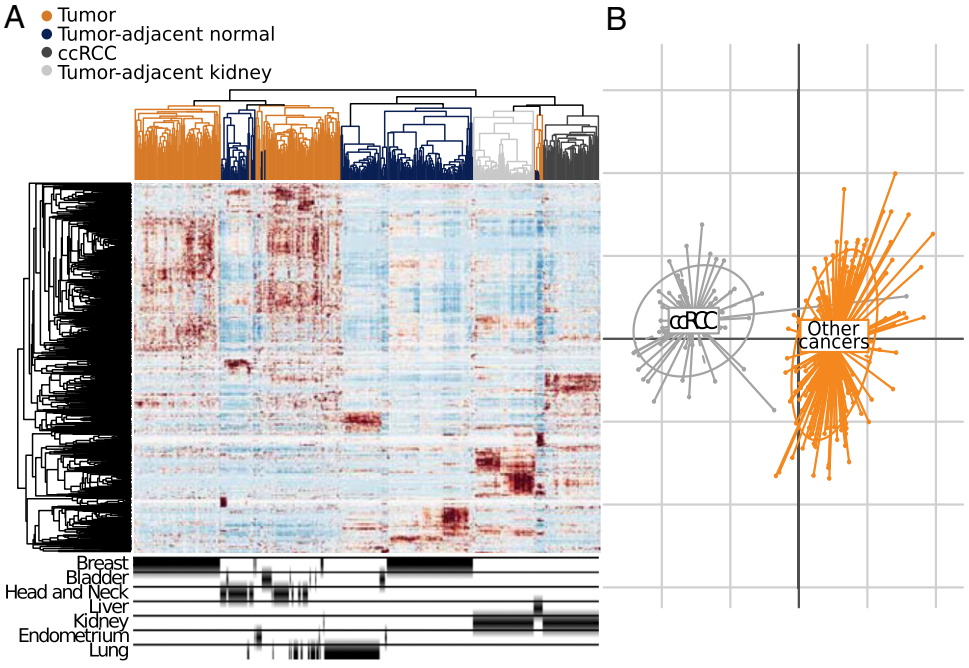


Fig. 1. Clustering analysis of metabolic gene expression profiles for cancer and tumor-adjacent normal samples. (*A*) Hierarchical clustering of absolute metabolic gene expression levels (RPKM) for cancer and tumor-adjacent normal samples, featuring only those genes that significantly changed expression across most cancer types upon transformation, thereby subtracting the effect of the tissue of origin. (*Lower*) Corresponding tissue of origin for each sample in the heatmap above. (*B*) PCA of $\log_2$ metabolic gene expression fold-change vs. matched tumor-adjacent normal samples for ccRCC (gray) and other cancer type samples (orange).

component (*SI Appendix*, Fig. S8). The deviation of ccRCC becomes even more evident when the metabolic gene expression fold-changes (FC) replace the transcript abundance in the PCA, which is suggestive of an opposite direction of regulation compared with other cancer types (Fig. 1*B*). Given that ccRCC samples present a lower and more variable tumor cellularity than the other cancer types in this study (16, 17), we tested the hypothesis that a higher stromal content may be responsible for the apparent outstanding regulation of metabolic gene expression in ccRCC. We inferred scores for tumor purity from gene expression profiles using ESTIMATE (17). These scores showed a moderated increase of infiltrating cells in ccRCC samples compared with others (*SI Appendix*, Fig. S9). However, when each metabolic gene expression fold-change value was adjusted using a simple linear regression on the stromal score for a given sample, we still observed a distinct separation of ccRCC from the other cancer types (*SI Appendix*, Fig. S10). These results suggest that unique patterns of regulation are rewiring ccRCC metabolism, which are markedly distinct from any other cancer in this study and that are not utterly ascribable to the tissue of origin or to the purity of the tumor. Indeed, the comprehensive molecular characterization of ccRCC that produced the data used in this study highlighted a considerable relationship between glycolytic metabolism and overall survival (16). However, the extent to which a shift toward a "Warburg effect"-like state may inherently explain the here-reported deviation of ccRCC metabolic regulation is questionable. Indeed, aerobic glycolysis is a hallmark reported in other different cancer types. Moreover, the number of regulated genes involved in this metabolic process (~40 as reported in ref. 16) is relatively low compared with the number of metabolic genes that clustered ccRCC distant from the other cancer types. These considerations hinted to us that other pathways may be strongly and uniquely regulated in ccRCC. Therefore, we computed for each cancer type the differential gene expression fold-changes compared with tumor-adjacent matched normal samples. We found that 2,539 metabolic genes are significantly regulated in ccRCC vs. tumor-adjacent kidney ($P < 0.05$), of which 329 genes are substantially up-regulated ($\log_2$FC ≥ 1) and 551 down-regulated ($\log_2$FC ≤ -1). This shows that there is a major disproportion toward down-regulation of metabolic genes in ccRCC. To test whether metabolic down-regulation is a common feature across different cancer types upon transformation, the actual discrete adjusted fold-change distribution in the population of ccRCC samples was compared with the population of the remaining cancer samples, and the former tend to have lower values ($P < 10^{-15}$, Mann-Whitney *U* test; *SI Appendix*, Fig. S11). We also observed that the adjustment for the stromal score corrects for an overestimation in terms of down-regulation. Additionally, the empirical cumulative distribution of adjusted fold-changes in the population of ccRCC samples was compared with the population of the remaining cancer samples. Again, we confirmed that there is a significant shift toward down-regulation in ccRCC with respect to the other cancer samples ($P < 10^{-15}$, Kolmogorov-Smirnov test; *SI Appendix*, Fig. S12). On the other hand, when the former test was repeated binning the remaining cancer samples according to their cancer type, endometrial cancer samples also showed a similar tendency (*SI Appendix*, Fig. S13). Taken together, these results are suggestive of widespread repression of metabolic gene expression in ccRCC, which in part explains the deviation observed above.

ccRCC Uniquely Regulates Nucleotide, Glycerolipid, and One-Carbon Metabolism Compared with Any Other Cancer Type, the Latter Being Implicated in Poor Prognosis. Next we sought to characterize the impact of a ccRCC divergent transcriptional program on metabolism as opposed to other cancer types. First, we checked in each PCA whether we could identify a set of relevant loadings responsible for the separation of ccRCC samples from the rest (i.e., the metabolic transcripts with the highest eigenvalues in each principal component—that is, highly associated with a separation on that component). However, neither in the PCA clustering on transcript abundance (*SI Appendix*, Fig. S14) nor in the PCA clustering on direction of gene expression regulation (*SI Appendix*, Fig. S15) could a well-defined set of genes be found. Therefore, we used network-dependent analyses to identify how such a unique program of transcriptional regulation diversely affected metabolism of ccRCC samples. For each of the cancer types we identified reporter metabolites and pathways (18) using our multiple gene-set analysis method (19) (Fig. 2 and *SI Appendix*, Fig. S16). As expected, in ccRCC diverse areas of the metabolic network were either uniquely regulated or not regulated compared with other cancer types, although an ostensible heterogeneity can be viewed across all cancer types (*SI Appendix*, *SI Text*). Among these, nucleotide metabolism and alanine, aspartate, and glutamate metabolism, which were generally found up-regulated in most cancer types, were not significantly altered in ccRCC. On the other hand, the metabolism of other amino acids (namely valine, leucine, isoleucine, cysteine, methionine, glycine, serine, and threonine) was significantly down-regulated only in ccRCC, as much as was the tricarboxylic acid (TCA) cycle and enzymes that participate in the metabolism of ubiquinone and ubiquinol (*SI Appendix*, Fig. S16), intermediates in the electron transport chain (ETC). Finally, we report unique changes in the metabolism of long-chain fatty acids and lactate (*SI Appendix*, Figs. S17 and S18).

Moreover, we noticed that 1,504 genes that showed statistical significance in patientwise expression fold-change across all cancer vs. matched normal samples ($P < 0.05$, rank–product test, Bonferroni correction) did not display any remarkable change in expression level when averaging in the pool of ccRCC samples. Unsupervised hierarchical clustering of patient-specific metabolic gene expression profiles featuring this set of genes revealed two different clusters with opposite regulatory directions (Fig. 3*A*). Interestingly, these clusters correlate with patients' tumor stage ($P = 0.041$, Pearson χ^2 test; Fig. 3*B*). Kaplan-Meier survival plots and log–rank tests were used to assess the differences in overall survival, and accordingly, the high tumor stage cluster is a predictor for poor prognosis ($P = 0.012$, log–rank test; Fig. 3*C*). Therefore, we sought to verify whether an advanced tumor stage drives per se a different transcriptional regulation of metabolism, as recently suggested (16). To test this, 170 metabolic genes that have significantly changed expression between high tumor stage (stage III to IV) and low tumor stage (stage I to II) samples ($P < 0.05$, Wilcoxon rank-sum test) were featured to cluster samples in a supervised fashion. Contrary to the premises, the two clusters that emerged from the analysis had a weaker association in relation to the tumor stage ($P = 0.1554$; *SI Appendix*, Fig. S19) but a comparable power to predict poor prognosis ($P = 0.025$) (Fig. 3*C*). In both scenarios, the curves strikingly superimpose with the survival plots based on the sole tumor stage information (high vs. low tumor stage; Fig. 3*C*), therefore suggesting a metabolic gene expression profile that is shaped after disease progression. To identify novel metabolic functions affected by the differential program of metabolic regulation between the two clusters, we used the reporter metabolite algorithm (18) (*SI Appendix*, Fig. S20). The analysis unveiled some unreported changes (*SI Appendix*, *SI Text*). Among these, we focused on dimethylglycine, a metabolite that belongs to one-carbon metabolism. Dimethylglycine is synthesized from betaine and subsequently converted into glycine (Fig. 3*D*). Most enzyme-coding genes that are uniquely attributable to this pathway display a significant difference in expression regulation between the low and high tumor stage cluster, especially betaine–homocysteine S-methyltransferase 1 (*BHMT*) and 2 (*BHMT2*) whose expression reverse direction completely (*SI Appendix*,

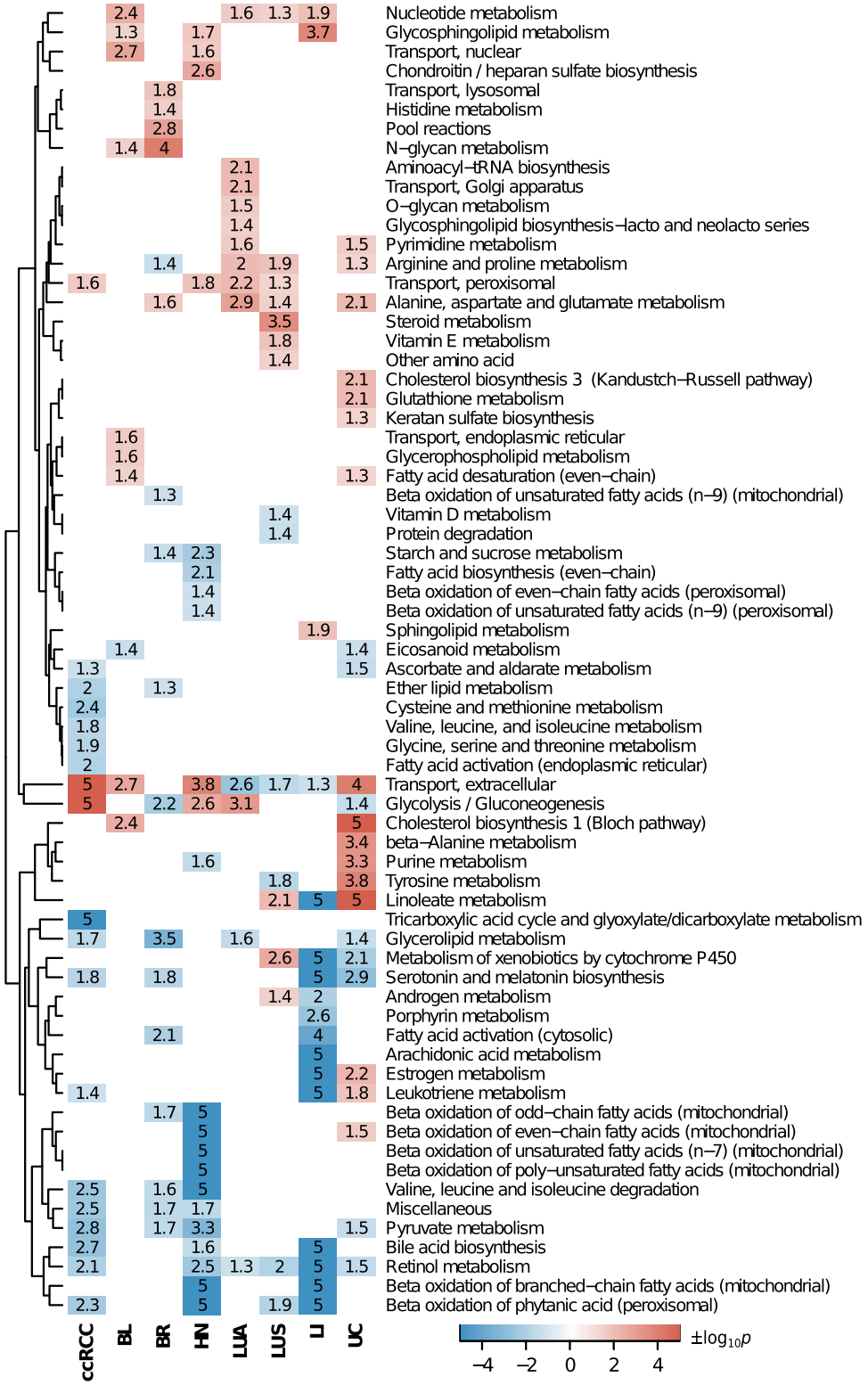


Fig. 2. Reporter canonical metabolic pathways in each cancer type according to significant changes in metabolic gene expression vs. matched tumor-adjacent normal tissues. Each box shows the $\log_{10}$ *P* value of the gene set representing a pathway in a certain cancer type, and the color indicates the overall direction of gene expression regulation for the gene set (red, up; blue, down). BL, bladder urothelial carcinoma; BR, breast invasive carcinoma; HN, head and neck squamous cell carcinoma; LUA, lung adenocarcinoma; LUS, lung squamous cell carcinoma; LI, liver hepatocellular Ccarcinoma; UC, uterine corpus endometrioid carcinoma.

Fig. S21). These results suggest an unanticipated role for betaine in ccRCC malignancy, which may parallel DNA hypermethylation frequency recently associated with an advanced stage and grade of this tumor (16), given the established role of betaine as a modulator of one-carbon metabolism and homocysteine levels through *BHMT* (20).

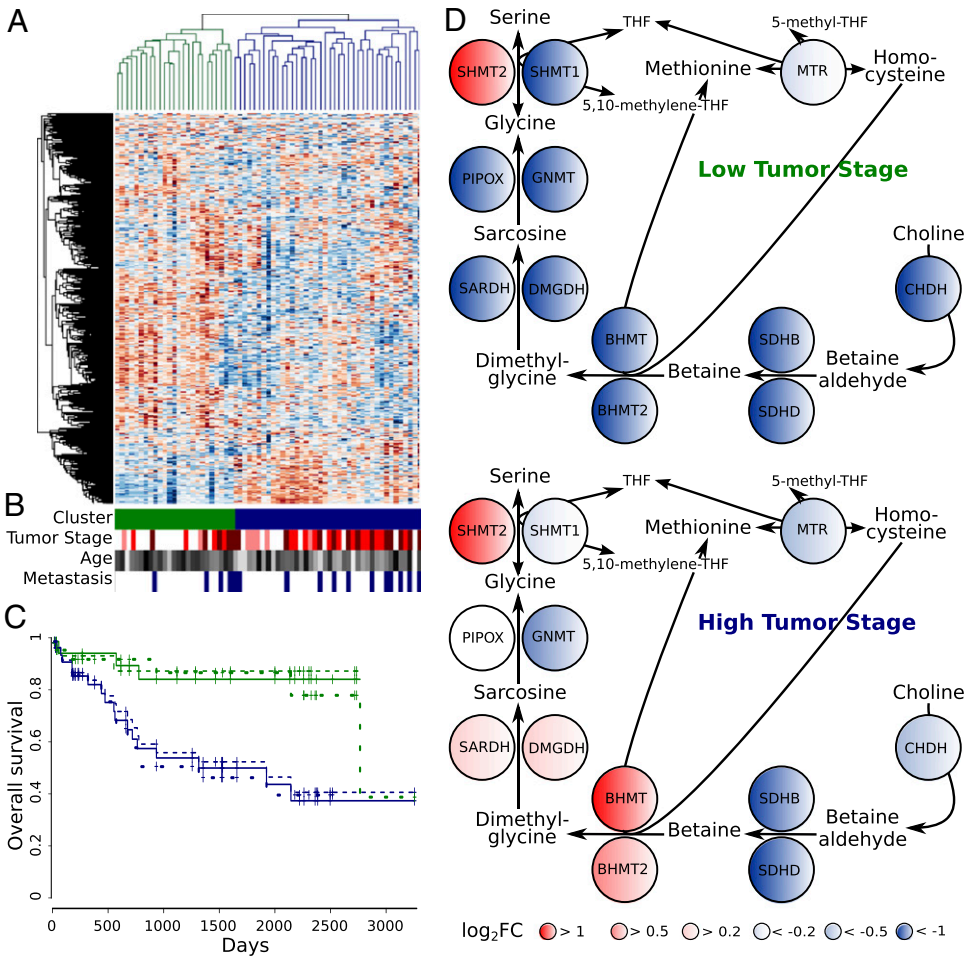
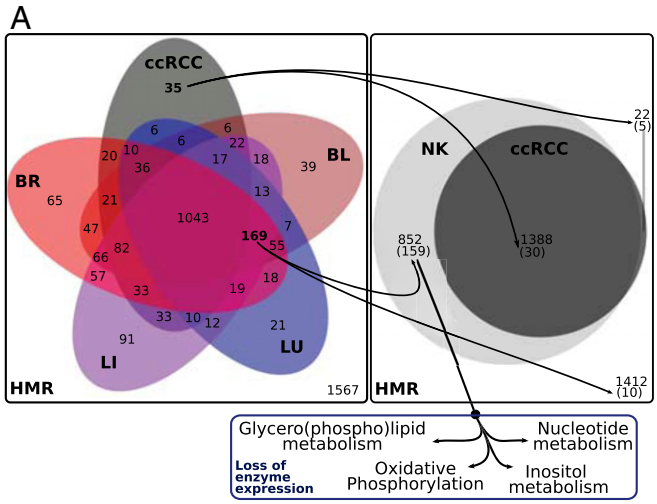


Fig. 3. Metabolic gene expression profiles distinguish two cluster of differential regulation in ccRCC that implicate a role of one-carbon metabolism in the malignancy. (*A*) Unsupervised hierarchical clustering of ccRCC $\log_2$ gene expression fold-changes vs. matched tumor-adjacent normal samples featuring the set of 1,504 metabolic genes significantly different across all $\log_2$ gene expression fold-changes in cancer vs. matched tumor-adjacent normal samples. (*B*) Clinical data for each ccRCC sample as ordered by the hierarchical clustering in *A*. Tumor stage ranges from stage 1 (white) to stage 4 (dark red); age is represented by a gray-scale in which the lowest value is 38 y and highest 90 y; metastatic tumors are depicted in blue. (*C*) Kaplan-Meier survival curves for the two clusters in *A* (solid line), for the two clusters in *SI Appendix*, Fig. S19 defined when featuring significantly changed genes between the pool of high tumor stage and low tumor stage samples (short dashed line), and for the two groups of samples solely identified by a high or low tumor stage (long dashed line). Blue and green lines refer to high and low tumor stage, respectively. (*D*) Gene expression regulation of enzymes involved in choline degradation to glycine in the two clusters in *A*.

Upon Transformation, the ccRCC Metabolic Network Features Widespread Loss of Function in Nucleotide, Glycerolipid, and One-Carbon Metabolism, in Contrast to Any Other Tumors. Because transcriptional regulation in ccRCC is suggestive of diffuse down-regulation of key metabolic functions, we evaluated whether the topology of its metabolic network reflected these premises. Therefore, we generated a cancer cell type-specific GEM based on protein evidence, using the INIT algorithm (21). Using the generic human model HMR as a template, INIT reduces the network according to proteomics evidence covering 15,156 proteins across at least 10 renal carcinomas (>70% are clear cell). Of 15,156 protein-coding genes, 2,482 could be mapped to HMR, and the reconstructed GEM, *iRenalCancer1410*, accounts for 3,913 reactions, 2,053 metabolites, and 1,410 genes, which means a reduction of 4,318 reactions (−53%) and 2,264 genes (−62%) from the generic human metabolic model, mostly due to no or little evidence for the encoded protein in most renal cancer samples (*SI Appendix*, Fig. S22*A*). To assess which losses of metabolic function are attributable to the transformation, we compared *iRenalCancer1410* with a published metabolic model of the kidney cell in tubules (21), accounting for 4,812 reactions, 2,268 metabolites, and 2,240 genes (*SI Appendix*, Fig. S22*B*). The comparison evidenced two interesting points. First, upon transformation, the renal cancer metabolic network shrinks with a reduction by approximately 20% of the number of reactions and a reduction of more than 35% of the number of associated genes, which clearly demonstrates a significant loss of metabolic functions. Second, 852 metabolic genes that are present at the protein level in the normal kidney cell are lost in the transformed cell. These genes mainly belong to the metabolism of nucleotides, glycerolipids, glycerophospholipids, glycosphingolipids, oxidative phosphorylation, and inositol metabolism, among others (Fig. 4*A*). Conversely, 22 metabolic genes present only the renal cancer network are associated with oxidoreductases in the endoplasmic reticulum. To validate whether these metabolic perturbations occurring during the transformation can be uniquely attributed to renal cancer and therefore provide insights on the deviating pattern of ccRCC metabolic regulation, we reconstructed cancer cell type-specific metabolic models for four additional cancer types (breast, lung, liver, and bladder) using the same procedure described above. In line with the previous results that highlighted a pronounced shift toward metabolic down-regulation in ccRCC



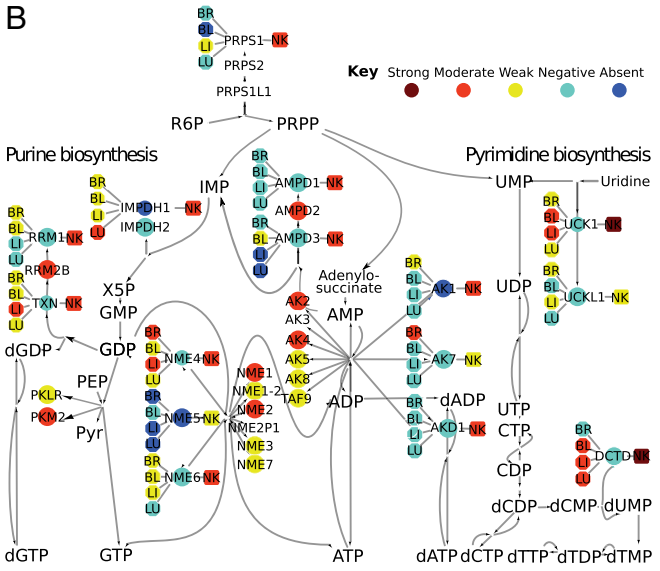


Fig. 4. Reconstruction and comparison of ccRCC-specific GEM (*iRenalCancer1410*) against other cancer-type GEMs and the normal kidney cell in tubules GEM. (*A*) Venn diagram for the metabolic genes present in *iRenalCancer1410* compared with other reconstructed GEMs (cancers, *Left*, kidney cell in tubules, *Right*). Metabolic genes absent in *iRenalCancer1410* but present both in the kidney cell in tubules GEM and other cancer type GEMs were used to enrich canonical pathways. (*B*) Nucleotide metabolism featuring the staining level of proteins taken from HPA for different tissues (*Center*: ccRCC; *Left*: BR, breast cancer; BL, bladder cancer; LI, liver cancer; LU, lung cancer; *Right*: NK, kidney cell in tubules).

compared with any other tumor, the metabolic network reconstruction of *iRenalCancer1410* resulted in a substantially smaller model than in any of the other cancers. Moreover, the comparison of each of the four metabolic networks against *iRenalCancer1410* revealed that 169 metabolic genes were present in all cancer models except for the renal cancer model, more than for any other cancer (on average, 41 ± 28 metabolic genes are lost in a model compared with the rest; Fig. 4*A*). Strikingly, 159 of these genes are also present in the normal kidney cell metabolic network, therefore confirming that the above-observed metabolic perturbations in ccRCC are unique. As a consequence, key enzymes are missing in the metabolism of glycerophospholipids, inositol, and one-carbon, as well as oxidative phosphorylation (Fig. 4*A* and *SI Appendix*, Fig. S23). Additionally, nucleotide metabolism is severely compromised, spanning both de novo biosynthesis, which is generally up-regulated in cancer (6), and successive degradation of nucleosides (Fig. 4*B*). In particular, we observe that absence of inosine 5′-monophosphate dehydrogenase 1 (IMPDH1) and 2 (IMPDH2), which commit IMP to initiate guanosine synthesis; RRM1 and TXN, which catalyze the synthesis of deoxyribonucleotides from the corresponding ribonucleotides; NME/NM23 family member 5 (NME5), NME/NM23 nucleoside diphosphate kinase 4 (NME4) and 6 (NME6), which are involved in nucleoside triphosphates biosynthesis from nucleoside diphosphates; and several enzymes in the 5′(3′)-deoxyribonucleotidase family, which dephosphorylates different deoxyribonucleotides. Taken together, a widespread repression of protein expression clearly limits the redundancy of genes in key metabolic pathways (most importantly nucleotide metabolism), a fact undetected in any other cancer type.

Divergence of ccRCC Metabolic Regulation Can Be Ascribed to Loss of Heterozygosity in Key Metabolic Genes Adjacent to von Hippel-Lindau Tumor Suppressor After Its Deletion. The divergence of ccRCC metabolic regulation compared with any other tumor and normal kidney is apparent both in the transcriptional regulation and its metabolic network. On one hand, the analysis of transcriptional regulation revealed that even though each type of cancer in this study displayed a complex and heterogeneous pattern of metabolic regulation in different pathways, ccRCC features a unique tendency to generally down-regulate large parts of metabolism. Only ccRCC strongly represses expression of cysteine and methionine metabolism (part of one-carbon metabolism), branched-amino acids metabolism, the metabolism of glycine, serine, and threonine, the TCA cycle and the ETC, as well as glycerolipid metabolism (with emphasis on fatty acid elongation). Additionally, whereas most tumors up-regulate nucleotide metabolism and the metabolism of alanine, aspartate, and glutamate, ccRCC shows a mixed pattern of regulation that results in both overexpressed and repressed genes in these pathways. On the other hand, the analysis of the metabolic network uncovered that ccRCC relies on a relatively small network, which features loss of gene redundancy in key metabolic pathways (e.g., oxidative phosphorylation and nucleotide, inositol, one-carbon, and glycerolipid metabolism) otherwise unaffected in any other cancer or in the normal kidney. Therefore, we sought to characterize the possible causes underlying the above-described metabolic features that render ccRCC different from all of the other cancer types upon transformation (Fig. 5).

First we considered the genetic background of ccRCC. Sporadic ccRCC (which represents 75% of all renal carcinomas) is generally characterized by mutations in the *VHL* tumor suppressor gene (16, 22). Deletion of Von Hippel-Lindau tumor suppressor (*VHL*) results in the stabilization of the hypoxia-inducible transcription factor (HIF) under normoxic conditions and thus entails a profound rewiring of mammalian oxygen-sensing pathways (23). Therefore, we tested the hypothesis that loss of *VHL* is one of the possible causes for the unique metabolic reprogramming of ccRCC. First, copy number variants (CNVs) for 62 ccRCC samples used in the PCA were checked in the region of *VHL*, and in 90% (56 of 62) of the samples the *VHL* gene locus was found consistently deleted, as opposed to matched tumor-adjacent normal samples (*SI Appendix*, Fig. S24). In addition, 52% of these samples (32 of 62) harbored a *VHL* mutation that mostly results in the deactivation of the corresponding protein (*SI Appendix*, Fig. S24). Second, the metabolic gene expression changes for each examined cancer type vs. matched normal tissues were correlated with an analogous comparison between two isogenic ccRCC cell lines (786-O), in which one of the cell lines is a *VHL* mutant with a frameshift deletion, whereas the second had *VHL* reintroduced (24). By comparing the *VHL*-deficient cell line against the *VHL*-reintroduced cell line, 1,339 genes were

found significantly regulated, of which 290 are metabolic ($P < 0.05$, 119 up-regulated, 171 down-regulated). Among all of the cancer types analyzed, only the direction changes in metabolic gene expression coming from ccRCC samples correlated with the above pattern of transcriptional regulation ($P < 0.05$; *SI Appendix*, Fig. S25). The correlation is even stronger when a more stringent significance level is used ($P < 0.01$; *SI Appendix*, Fig. S26). Despite the limited number of genes differentially regulated in the *VHL*-deficient cell line, the amount of metabolic genes coregulated with ccRCC (228) was found to be overrepresented ($P < 10^{-4}$, Fisher's exact test), in particular the up-regulated ones ($P < 10^{-4}$; *SI Appendix*, Fig. S27). Such genes are not exclusively ascribable to kidney or renal carcinoma, as revealed by functional clustering based on tissue expression data (*SI Appendix*, Table S3). These results were successfully replicated using an analogous yet independent dataset (25) (*SI Appendix*, Fig. S25). The analysis of these data indicates that in ccRCC *VHL* loss is indeed associated with the regulation of a significant portion of metabolic genes, whereas such association could not be recapitulated in any other cancer type. We therefore explored the metabolic functions affected by expression changes in the 228 genes regulated by loss of *VHL* that are associated with ccRCC. To this end, reporter pathways were computed as described above. Only three pathways were shown to be significantly regulated in a consistent direction by this set of genes, namely alanine, aspartate, and glutamate metabolism, valine, leucine, and isoleucine metabolism, and fatty acid elongation (*SI Appendix*, Fig. S28). In all three cases, gene expression was shifted toward down-regulation. Therefore, loss of *VHL* alone may explain why these pathways are repressed or mixed regulated only in ccRCC and not in the other tumors (Fig. 2 and *SI Appendix*, Fig. S16). Given *VHL* involvement in oxygen sensing, we also tested whether pseudohypoxia induced by *VHL* inactivation may drive per se a divergent transcriptional response with respect to environmental hypoxia seen in most tumors and the normal kidney (26, 27), but no such correlation could be found (*SI Appendix*, *SI Text* and Fig. S29).

Apart from *VHL* loss-mediated stabilization of HIF, other transcription factors may be triggered only in ccRCC, thus shedding light on the other differentially regulated metabolic functions. Hence, metabolic gene expression changes were used in our multiple gene-set analysis method to identify reporter transcription factors in each cancer type (*SI Appendix*, *SI Text* and Fig. S30). Notably, signal transducer and activator of transcription 1 (STAT1), an anticarcinogenic transcription factor (28), is deemed associated only to ccRCC as it regulates many of the up-regulated metabolic genes. Indeed, the STAT1 gene set comprises 4,381 genes, and 264 of them are metabolic genes that were significantly overexpressed in ccRCC ($P < 0.01$). Surprisingly, these genes were found to relate to inositol and nucleotide metabolism, among others (*SI Appendix*, Table S4), even though the above analysis of the ccRCC metabolic network rather suggested that these pathways were compromised. However, a detailed review unveiled that these metabolic genes are complementary to the ones found not expressed at the protein level in ccRCC. For instance, the nucleotide metabolism-related genes, namely *NME1*, *NME1-NME2*, *NME2*, and *PKM2*, are all part of ccRCC metabolic network and compensate the lack of expression of *NME4*, *NME5*, and *NME6* at the protein level (Fig. 4*B*). The same can be concluded for inositol metabolism, in which induction of *PIK3C2B*, *PIK3R3*, and *PIK3CD* contrasts with low to null expression of *PI4KB*, *PI4K2A*, and *PI4K2B* (*SI Appendix*, Fig. S23). Therefore, if STAT1 was indeed activated in ccRCC, then together with *VHL* loss it can explain most of the metabolic features that distinguished ccRCC from any other cancer in this study.

The mechanisms for other features unique to ccRCC, such as loss of gene redundancy in nucleotide and glycerolipid metabolism as well as down-regulation of one-carbon metabolism, still remained unsettled. Although this may be seen as part of a general shift toward down-regulation that has related to multistep cancer transformation and suggestive of dedifferentiation (29), we had previously ruled out a compelling role of the tissue of origin in ccRCC metabolic reprogramming (Fig. 1*A* and *SI Appendix*, Fig. S5). We therefore sought to identify whether other genetic alterations may be implicated. Thus we analyzed 488 ccRCC samples and as many matched normal samples for which CNVs were scanned using Affymetrix Genome-Wide SNP Array 6.0 (*SI Appendix*, Dataset S2). We restricted our analysis to those gene loci that overlap with the metabolic genes in HMR and that displayed appreciable mean segment amplitude with respect to the baseline ($>\pm 0.15$) across at least 50% of the samples. Furthermore, all mean segments amplitudes that were not found to be statistically different in the pool of ccRCC samples against tumor-adjacent normal samples were discarded ($P < 0.01$, Wilcoxon rank–sum test). In total, 108 metabolic genes were deemed to be recurrently deleted (107) or amplified (1) in ccRCC (*SI Appendix*, Fig. S31). Transcript and protein abundance for each of these genes were checked in ccRCC against tumor-adjacent normal samples, and 14 genes displayed a consistent trend with the presumptive CNV (*SI Appendix*, Fig. S32 and Table S5). Among these, abhydrolase domain containing 5 (*ABDH5*), choline dehydrogenase (*CHDH*), glycerol-3-phosphate dehydrogenase 1-like (*GPD1L*), *IMPDH2*, and pyruvate dehydrogenase beta (*PDHB*) are located within 3p14.3 and 3p22.3, a region that display significant decrease in gene copy number in the range of 75–81% of samples (Table 1). Reduced copy number for all these metabolic genes may share the same mechanism that induces early loss of *VHL* in ccRCC, being *VHL* located at 3p25.3. Only *PDHB* is known to be indirectly inhibited after *VHL* loss, via HIF-dependent expression of PDHK1, a PDH complex inhibitor (30). Remarkably, these deletions explain many defects previously unveiled in ccRCC metabolic regulation: ABHD5 and GPD1L are involved in glycerophospholipid metabolism; CHDH is implicated in one-carbon metabolism; PDHB commits pyruvate in the TCA cycle; and IMPDH2 is a key step in purine biosynthesis. Taken together, these results are suggestive of a multistep model for ccRCC metabolic reprogramming (Fig. 5): first, *VHL* loss in ccRCC ini-

Table 1. Potentially deleted genes according to copy number (CNV), transcript level (abundance [reads per kilobase per million reads (RPKM)] and regulation [$\log_2$FC]), and median protein staining level in malignant and healthy renal tissue

Gene	Gene locus	Mean CNV amplitude	CNV frequency (%)	RPKM	$\log_2$FC	Median staining renal cancers	Staining kidney cells in tubules	Enzymatic activity
ABHD5	3p21.31	−0.236	81.15	2.69	−0.94	Negative	Moderate	1-acylglycerol-3-phosphate O-acyltransferase
CHDH	3p21.1	−0.227	78.28	9.54	−1.05	Negative	Strong	Choline dehydrogenase
GPD1L	3p22.3	−0.234	80.94	4.96	−0.83	Negative	Moderate	Glycerol-3-phosphate dehydrogenase
IMPDH2	3p21.31	−0.236	81.15	8.19	−0.68	Negative	Moderate	IMP dehydrogenase
PDHB	3p14.3	−0.216	74.80	4.35	−1.30	Negative	Strong	Pyruvate dehydrogenase

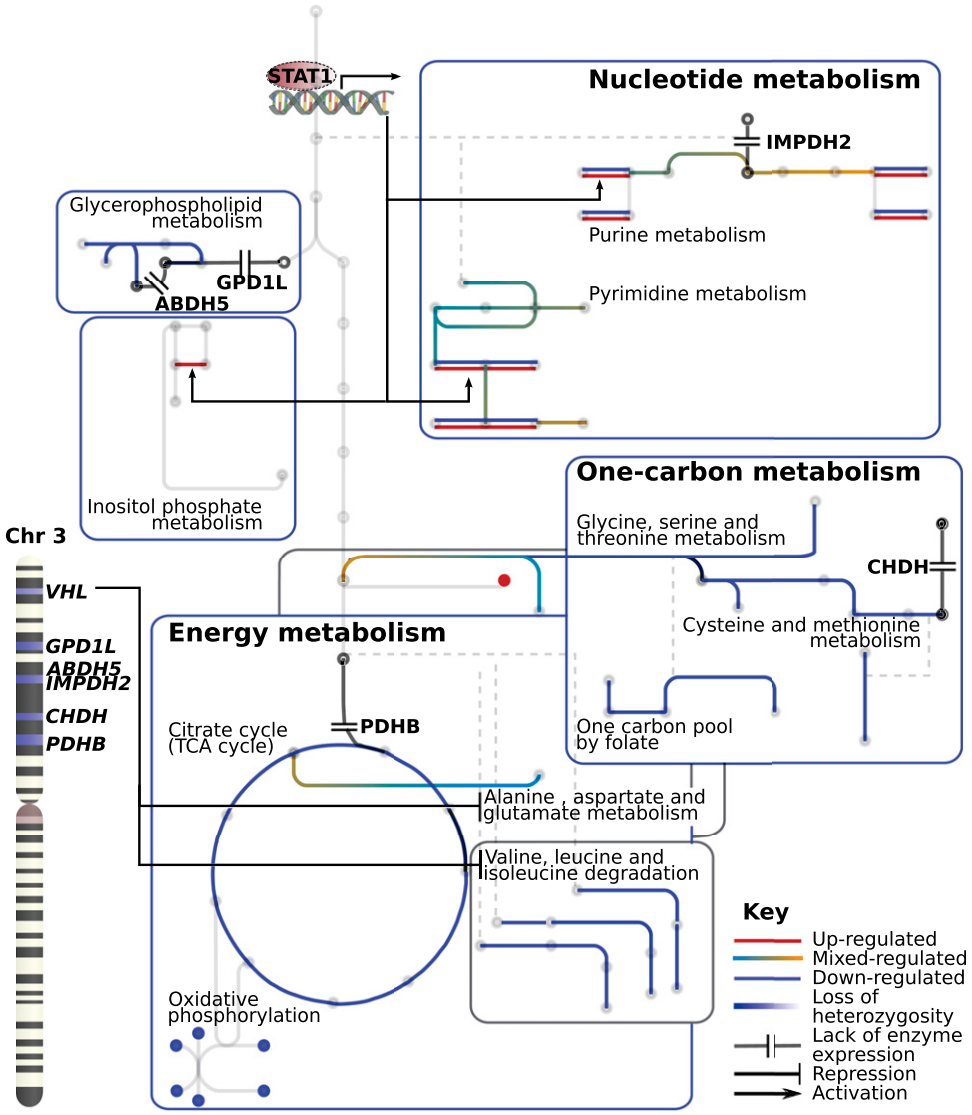


Fig. 5. An overview of the metabolic features unique to ccRCC in the landscape of cancer metabolic regulation. The figure shows reporter pathways (represented by edges; refer to Fig. 2) and metabolites (represented by nodes; refer to *SI Appendix*, Fig. S16) transcriptionally regulated only in ccRCC vs. matched tumor-adjacent normal tissue; and subnetworks (represented by rectangles) that feature lack of gene redundancy only in ccRCC metabolic network (refer to Fig. 4*A*). The mechanisms that contribute to this metabolic phenotype are summarized. First, loss of *VHL* represses expression of metabolic genes in alanine, aspartate, glutamate, and branched-chain amino acids metabolism. Second, potential activation of STAT1 up-regulates redundant genes in nucleotide biosynthesis and inositol metabolism. Third, loss of heterozygosity in metabolic genes adjacent to *VHL* affects several pathways previously identified as down-regulated or deficient only in ccRCC (represented by double bar).

tiates an extensive transcriptional program that also represses peripheral metabolism (e.g., branched-chain amino acids metabolism); then, recurrent loss of heterozygosity in 3p affects several adjacent metabolic genes that implicate reduced redundancy in the ccRCC metabolic network (e.g., impaired purine biosynthesis); finally, transition to malignancy and a possible activation of STAT1 may contribute to trigger adaptive mechanisms (e.g., regulation of one-carbon and nucleotide metabolism).

Discussion

The increasing body of evidence that the same deregulated signaling pathways that lead to the typical malignancy of cancer cells converge in the regulation of cell metabolism has lately gained attention for the possible implications in cancer therapy (31). Moreover, this fact propelled the idea that oncogene-directed metabolic reprogramming is a strict condition to support anabolic growth and meet the metabolic requirements for proliferation in any cancer cell (2, 5). However, the degree to which regulation of such reprogramming is equated across different cancer cells at the systems level has been largely overlooked. In this study, and in remarkable concordance with the work by Hu et al. (6), a systems analysis of the metabolic network revealed that cancer cells orchestrate the expression of metabolic genes in a similar fashion only when it comes to nucleotide, glutamate, and retinol metabolism, while retaining the expression of a substantial portion of metabolic genes unaltered with respect to their tissue of origin. As a proof of concept, it has been recently appreciated that sustained growth signaling via mTORC1, a pathway constitutively active in most human cancers, directly controls de novo pyrimidine biosynthetic flux (32, 33).

The fact that ccRCC has a radically different metabolic regulatory program at the systems level may therefore be important not only for the rational design of therapeutic targets against this particular neoplasm, but also to understand cancer metabolism

in general. Recently a comprehensive characterization of ccRCC unveiled an exceptional regulation of central carbon metabolism, associated with altered promoter methylation patterns and mutations in the PI(3)K/AKT pathway (16). Here we report that peripheral regions of metabolism (e.g., nucleotide, one-carbon, amino acid, and glycerolipid metabolism) are also uniquely affected in ccRCC, and we provide evidence that this divergence of ccRCC metabolic regulation can be ascribed to recurrent loss of heterozygosity. Indeed, in line with the idea that oncogenic pathways are implicated in the regulation of cancer metabolism, we show that loss of *VHL* in the 3p chromosome drives an initial reprogramming that matched exclusively in the case of ccRCC. Accordingly, a recent study that reconstructed a single tumor genetic phylogeny confirmed that *VHL* deletion is the earliest event in ccRCC cancerogenesis (34). Such reprogramming entails down-regulation of branched-chain amino acid metabolism, fatty acid elongation, and alanine, aspartate, and glutamate metabolism, the latter otherwise generally up-regulated in most cancers. Thereafter, loss of heterozygosity in key metabolic genes adjacent to *VHL* compromise nucleotide, one-carbon, and inositol metabolism and glycerolipid biosynthesis. In support of our conclusion, a study specifically aiming to map deletions on 3p besides *VHL* in ccRCC tissue samples revealed extensive loss of heterozygosity in gene loci within the chromosome, with remarkably higher frequency at a late tumor stage (35). Such events therefore imply in ccRCC loss of gene redundancy in a metabolic pathway, namely nucleotide metabolism, which is the most frequently overexpressed in cancer (6). As such, considering its essential role for cancer proliferation, such a finding paves the way for potential synthetic lethality strategies. Intriguingly, ccRCC cells seem to adapt to its defective network by up-regulating alternative pathways, as in the case of betaine, where we find that this adaptation may turn critical for the aggressiveness of the disease. In addition, a potential activation of STAT1 transcription factor was associated only with ccRCC. Such activation would trigger up-regulation of complementary genes in nucleotide and inositol metabolism, explaining the mixed regulation of the former in ccRCC. Although STAT1 activation requires experimental validation, this event may be linked to immune response (28) or to treatment with interferon-α (36). Finally, we considered environmental hypoxia and kidney cell dedifferentation as other potential factors that may contribute to a different metabolic reprogramming in ccRCC, but no compelling role could be demonstrated in either case. In conclusion, there is evidence that ccRCC metabolic regulation is uniquely shaped upon loss of heterozygosity in the 3p chromosome, where VHL, the tumor suppressor gene most commonly associated to ccRCC, is located.

Finally, Hu et al. (6) reported that in the few cases in which nucleotide biosynthesis was not up-regulated in cancer, the overall metabolic gene expression was most down-regulated or not changed. Indeed, here we found that ccRCC displays a significant shift toward metabolic down-regulation, which translated in the smallest metabolic network, and features a compromised nucleotide metabolism. We therefore addressed this discrepancy in ccRCC and propose that loss of heterozygosity may be instrumental in shaping the metabolic topology of these cancer cells. As such, we also believe that these results reinforce the idea that among all other cancers nucleotide biosynthesis is a crucially altered pathway marked with increased activity.

Materials and Methods

Data. RNAseq profiles for primary tumor and matched tumor-adjacent normal tissues were obtained at The Cancer Genome Atlas (TCGA, tcga-data.nci.nih.gov). Immunohistochemical protein profiles were retrieved at the Human Protein Atlas (HPA version 11, www.proteinatlas.org). GEMs for a generic human cell (HMR3674, shortly HMR) and the kidney cell in tubules were downloaded from the Human Metabolic Atlas (www.metabolicatlas.com). Detailed information is given in *SI Appendix, SI Materials and Methods*.

Cluster Analysis. PCA and hierarchical clustering (Pearson correlation metric, average linkage), were performed on the basis of metabolic transcript abundance profiles [measured in reads per kilobase per million reads (RPKM)] or $\log_2$ metabolic gene expression fold-change against matched tumor-adjacent normal samples focusing only on those genes included in HMR. MCA was based on four categorical staining levels (strong, moderate, weak, and negative) for metabolic gene encoded proteins. Detailed information is given in *SI Appendix, SI Materials and Methods*.

Gene-Set Analysis. Multiple gene-set analyses were implemented using PIANO R-package (19), and each gene set was defined as either the set of genes constituting a pathway in HMR (reporter pathway), or as the set of genes that encode for all reactions involving a certain metabolite in HMR (reporter metabolites), or as the set of genes for which a peak was detected in any ChIP-seq experiment, as collected in Cscan (37), targeting a certain transcription factor (reporter transcription factors). For each directionality class (up-, down-, or mixed regulated), the statistical significance returned is the median significance reported by eight gene-set analysis methods. Then, the most significant directionality class is reported. Detailed information is given in *SI Appendix, SI Materials and Methods*.

Statistical Analysis. Details on the statistical tests reported in the text are available in *SI Appendix, SI Materials and Methods*. For gene expression differential analysis, cancer type-wise statistics were computed from empirical Bayes estimation and generalized linear models to fit a negative binomial distribution on the read counts; patient-wise statistics were computed using the rank–product test adjusted using the Bonferroni correction. Detailed information is given in *SI Appendix, SI Materials and Methods*.

Cancer GEMs Reconstruction. The reconstruction of cancer type-specific GEMs was performed for breast, bladder, liver, lung, and renal cancer using the INIT algorithm (21) within the RAVEN Toolbox (38). Scoring for evidence of a reaction to be occurring was based on HPA protein profiles for each cancer type. INIT reconstructs a GEM by maximizing the reaction score while preserving network connectivity and functionality (i.e., the resulting GEM must be able to perform a list of metabolic tasks, including biomass growth). Detailed information is given in *SI Appendix, SI Materials and Methods*. All cancer models are available through www.metabolicatlas.com.

CNV Analysis. SNP arrays for CNV analyses were obtained at TCGA for ccRCC and matched tumor-adjacent normal samples, and segment amplitude across each chromosome was calculated using the GADA R-package (39). Detailed information is given in *SI Appendix, SI Materials and Methods*.

ACKNOWLEDGMENTS. We thank Amir Feizi, Rahul Kumar, Adil Mardinoglu, Natapol Pornputtapong, Kaisa Thorell, Sergio Velasco, Leif Väremo, Rasmus Ågren, and Tobias Österlund for discussion and computational support, and Verena Siewers and Josè Luis Martinez for editing the article. The Cancer Genome Atlas provided access and diffusion of restricted data. The computations were performed on resources provided by the Swedish National Infrastructure for Computing at C3SE. This work was sponsored by the Knut and Alice Wallenberg Foundation and the Chalmers Foundation.

1. Hanahan D, Weinberg RA (2011) Hallmarks of cancer: The next generation. *Cell* 144(5):646–674.
2. Ward PS, Thompson CB (2012) Metabolic reprogramming: A cancer hallmark even Warburg did not anticipate. *Cancer Cell* 21(3):297–308.
3. Schulze A, Harris AL (2012) How cancer metabolism is tuned for proliferation and vulnerable to disruption. *Nature* 491(7424):364–373, and erratum (2012) 494(7435):130.
4. Cairns RA, Harris IS, Mak TW (2011) Regulation of cancer cell metabolism. *Nat Rev Cancer* 11(2):85–95.
5. Vander Heiden MG, Cantley LC, Thompson CB (2009) Understanding the Warburg effect: The metabolic requirements of cell proliferation. *Science* 324(5930):1029–1033.
6. Hu J, et al. (2013) Heterogeneity of tumor-induced gene expression changes in the human metabolic network. *Nat Biotechnol* 31(6):522–529.
7. Moreno-Sánchez R, Rodríguez-Enríquez S, Marín-Hernández A, Saavedra E (2007) Energy metabolism in tumor cells. *FEBS J* 274(6):1393–1418.
8. Mardinoglu A, Gatto F, Nielsen J (2013) Genome-scale modeling of human metabolism—a systems biology approach. *Biotechnol J* 8(9):985–996.
9. Jerby L, Ruppin E (2012) Predicting drug targets and biomarkers of cancer via genome-scale metabolic modeling. *Clin Cancer Res* 18(20):5572–5584.
10. Väremo L, Nookaew I, Nielsen J (2013) Novel insights into obesity and diabetes through genome-scale metabolic modeling. *Front Physiol* 4:92.

11. Lewis NE, Abdel-Haleem AM (2013) The evolution of genome-scale models of cancer metabolism. *Front Physiol* 4:237.
12. Lewis NE, Nagarajan H, Palsson BO (2012) Constraining the metabolic genotype-phenotype relationship using a phylogeny of in silico methods. *Nat Rev Microbiol* 10(4):291–305.
13. Frezza C, et al. (2011) Haem oxygenase is synthetically lethal with the tumour suppressor fumarate hydratase. *Nature* 477(7363):225–228.
14. Folger O, et al. (2011) Predicting selective drug targets in cancer through metabolic networks. *Mol Syst Biol* 7:501.
15. Mardinoglu A, et al. (2013) Integration of clinical data with a genome-scale metabolic model of the human adipocyte. *Mol Syst Biol* 9:649.
16. Creighton CJ, et al. (2013) Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature* 499(7456):43–49.
17. Yoshihara K, et al. (2013) Inferring tumour purity and stromal and immune cell admixture from expression data. *Nat Commun* 4:2612.
18. Patil KR, Nielsen J (2005) Uncovering transcriptional regulation of metabolism by using metabolic network topology. *Proc Natl Acad Sci USA* 102(8):2685–2689.
19. Väremo L, Nielsen J, Nookaew I (2013) Enriching the gene set analysis of genome-wide data by incorporating directionality of gene expression and combining statistical hypotheses and methods. *Nucleic Acids Res* 41(8):4378–4391.
20. Ueland PM, Holm PI, Hustad S (2005) Betaine: A key modulator of one-carbon metabolism and homocysteine status. *Clin Chem Lab Med* 43(10):1069–1075.
21. Agren R, et al. (2012) Reconstruction of genome-scale active metabolic networks for 69 human cell types and 16 cancer types using INIT. *PLOS Comput Biol* 8(5):e1002518.
22. Cohen HT, McGovern FJ (2005) Renal-cell carcinoma. *N Engl J Med* 353(23):2477–2490.
23. Kim WY, Kaelin WG (2004) Role of VHL gene mutation in human cancer. *J Clin Oncol* 22(24):4991–5004.
24. Vanharanta S, et al. (2013) Epigenetic expansion of VHL-HIF signal output drives multiorgan metastasis in renal cancer. *Nat Med* 19(1):50–56.
25. Papandreou I, Cairns RA, Fontana L, Lim AL, Denko NC (2006) HIF-1 mediates adaptation to hypoxia by actively downregulating mitochondrial oxygen consumption. *Cell Metab* 3(3):187–197.
26. Harris AL (2002) Hypoxia—a key regulatory factor in tumour growth. *Nat Rev Cancer* 2(1):38–47.
27. Jiang Y, et al. (2003) Gene expression profiling in a renal cell carcinoma cell line: Dissecting VHL and hypoxia-dependent pathways. *Mol Cancer Res* 1(6):453–462.
28. Dranoff G (2004) Cytokines in cancer pathogenesis and cancer therapy. *Nat Rev Cancer* 4(1):11–22.
29. Danielsson F, et al. (2013) Majority of differentially expressed genes are down-regulated during malignant transformation in a four-stage model. *Proc Natl Acad Sci USA* 110(17):6853–6858.
30. Jonasch E, et al. (2012) State of the science: an update on renal cell carcinoma. *Mol Cancer Res* 10(7):859–880.
31. Cheong H, Lu C, Lindsten T, Thompson CB (2012) Therapeutic targets in cancer cell metabolism and autophagy. *Nat Biotechnol* 30(7):671–678.
32. Robitaille AM, et al. (2013) Quantitative phosphoproteomics reveal mTORC1 activates de novo pyrimidine synthesis. *Science* 339(6125):1320–1323.
33. Ben-Sahra I, Howell JJ, Asara JM, Manning BD (2013) Stimulation of de novo pyrimidine synthesis by growth signaling through mTOR and S6K1. *Science* 339(6125):1323–1328.
34. Gerlinger M, et al. (2012) Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N Engl J Med* 366(10):883–892.
35. Singh RB, Amare Kadam PS (2011) Investigation of tumor suppressor genes apart from VHL on 3p by deletion mapping in sporadic clear cell renal cell carcinoma (cRCC). *Urol Oncol* 31(7):1333–1342.
36. Brinckmann A, et al. (2002) Interferon-alpha resistance in renal carcinoma cells is associated with defective induction of signal transducer and activator of transcription 1 which can be restored by a supernatant of phorbol 12-myristate 13-acetate stimulated peripheral blood mononuclear cells. *Br J Cancer* 86(3):449–455.
37. Zambelli F, Prazzoli GM, Pesole G, Pavesi G (2012) Cscan: Finding common regulators of a set of genes by using a collection of genome-wide ChIP-seq datasets. *Nucleic Acids Res* 40(Web Server issue):W510–W515.
38. Agren R, et al. (2013) The RAVEN toolbox and its use for generating a genome-scale metabolic model for *Penicillium chrysogenum*. *PLOS Comput Biol* 9(3): e1002980.
39. Pique-Regi R, Cáceres A, González JR (2010) R-Gada: A fast and flexible pipeline for copy number analysis in association studies. *BMC Bioinformatics* 11:380.

PAPER II

Metabolic reprogramming resulting from oncogenic mutations converges in the deregulation of arachidonate and xenobiotics metabolism

F. Gatto, A. Schulze, J. Nielsen

Submitted for publication

Metabolic reprogramming resulting from oncogenic mutations converges in the deregulation of arachidonate and xenobiotics metabolism

F. Gatto[1], A. Schulze[2,3], J. Nielsen[1,*]

Affiliations:

[1]Department of Biology and Biological Engineering, Chalmers University of Technology, Göteborg, Sweden.

[2]Theodor-Boveri-Institute, Biocenter, Würzburg, Germany.

[3]Comprehensive Cancer Center Mainfranken, Würzburg, Germany

*Correspondence to: nielsenj@chalmers.se

SUMMARY

Mutations stand at the basis of the clonal evolution of most cancers. Nevertheless, it is still elusive whether mutations induce a reprogramming of gene expression that results in the emergence of the hallmarks of cancer, regardless of the cancer type. Here we analysed the genome and transcriptome of 1,082 primary tumors and show that 12 common cancer mutations independently lead to the deregulation of genes with metabolic functions. From our analysis, we derived a network of reactions, termed AraX, that involve the glutathione- and oxygen-mediated metabolism of arachidonic acid and xenobiotics. Deregulation of AraX significantly correlated with all 12 mutations. We further show that, among all metabolic pathways, AraX deregulation represents the strongest predictor for patient survival. These findings suggest that oncogenic mutations drive a selection process that converges on deregulation of the AraX network to gain growth advantage during cancer evolution.

Introduction

Sequencing of an increasing number of cancer genomes has revealed the extent of genomic heterogeneity of the disease, which stems from a complex interplay of mutations and the natural selection of clones (Yates and Campbell, 2012). The complexity of the cancer genome is a daunting challenge for the rational treatment of the disease. While progress has been made in the attempt to tailor treatments to the defined molecular features of individual tumors, the need for ever more precise patient stratification provides a rational limit for these strategies (Chin et al., 2011). Moreover, the concept of convergent evolution in cancer could explain the acquisition of the cancer phenotype through multiple routes (Gerlinger et al., 2014; Hanahan and Weinberg, 2011; Weinberg, 2014).

Mutations are central in the evolution of most cancers and, once acquired, they are liabilities that cancers carry throughout their progression. In addition to direct effects on cellular signaling networks and the reprograming of gene expression, cancer mutations also initiate a process of natural selection, which results in the emergence of cell lineages exhibiting the transformed characteristic of cancer (Vogelstein et al., 2013). In the light of this, it is likely that the aggregate of molecular features of a given tumor, including the presence of a given mutation, is represented in its gene expression profile. In other words, it is conceivable to factorize the expression level of each gene as the contribution of different tumor features, and extract the contribution due to occurrence of a cancer mutation. In turn, common transcriptional changes attributable to different mutations, i.e. convergence towards a common set of deregulated genes, should correspond to the deregulation of key biological processes. These key processes are then selected for via mutagenesis and natural selection, and define the phenotype of cancer.

Many studies have characterized the gene expression changes occurring due to prominent cancer-associated mutations in cell line and animal models (DeNicola et al., 2011; Fodde et al., 1994; Johnson et al., 2001; Podsypanina et al., 1999; Sasaki et al., 2012). However, these mechanistic studies are technologically limited to focus on one or few cancer mutations in one or few cancer types. On the contrary, a systematic analysis can identify meaningful correlations, but it requires simultaneous knowledge of the presence of a cancer mutation and the levels of all transcripts in the same sample and in a sufficient large number of samples that span distinct cancer types. Examples of such pan-cancer studies have so far concentrated in the identification of biological processes putatively affected by cancer mutations and/or epigenetic alterations, without taking in account the underlying changes in gene expression (Ciriello et al., 2013; Hofree et al., 2013; Kandoth et al., 2013). Here we have used genomic and transcriptomic data from 1,082 human tumor samples across 13 cancer types to derive genome-wide correlations between cancer mutations and transcript levels in human primary tumors. These associations were used to investigate whether

different mutations converge in the transcriptional regulation of defined biological processes. These processes are likely to represent cellular functions that are critical for positive selection during cancer evolution.

Results

Definition of the factors that contribute to gene expression changes in cancer using generalized linear models

We first sought to test the existence of a statistical association between gene expression changes and the presence of a cancer-associated mutation in the tumor, i.e. if occurrence of a mutation correlates with an increase or decrease in the mRNA abundance of a gene. RNA-seq profiles for 1,082 primary tumor samples were retrieved from 13 distinct cancer types (range of 21-199 samples per type, Fig. S1) for which a validated mutation spectrum was available (Cerami et al., 2012) (Fig. 1A). In this cohort, we focused on the 158 genes mutated at moderate frequency (>2% samples), of which 12 are mutated at high frequency (>10% samples, Fig. S2). We hypothesized that the level of gene expression could be factorized as the contribution of four sample features: the histopathological cancer type; the expression level of transcription factors; the presence or absence of a mutation; and the synergy induced by occurrence of a mutation in a particular cancer type. We therefore employed the established statistical framework of generalized linear models (GLM) to perform a linear regression of gene expression on the following factors: the 13 cancer types (*CT*); the activation status of 119 well-characterized transcription factors (*TFs*) (Zambelli et al., 2012); the presence or absence of a mutation in one of the 158 genes for which mutations were found at moderate frequency (*Muts*); and the interaction terms between the presence of a high frequency mutation and the cancer type where it occurred (*Ints*) (Fig. 1B). This generated an initial GLM (*All*), which comprised 416 factors.

Likely, many of these factors do not contribute significantly to explain the expression level of a gene. Hence, we employed different methods for model selection, including backward selection and regularized regression via the Lasso algorithm (Tibshirani, 1996). These methods identify a minimal number of relevant factors while maintaining an acceptable accuracy of prediction of the observed gene expression levels. Each method thus returned a set of relevant factors that constitute an alternative GLM to the initial *All* model (Fig. 1B). In total, we generated 17 GLMs: a backward selection (*BS*) model (yielding 84 factors); three Lasso models (*lasso1, lasso5, lasso10*), depending on the number of factors with a non-null regression coefficient in 1%, 5%, or 10% of all genes (yielding 328, 101, and 59 factors respectively); and 13 models solely based on a subset of the four sample features (i.e. only *CT,* or only *TFs,* or only *Muts,* or only *Ints* factors, or any other combination of these). The goodness-of-fit between observed and predicted expression level for each gene is dependent on the GLM used. The best GLM was selected by counting for each GLM the number of genes whose expression was accurately predicted (i.e. an acceptable goodness-of-fit), while relying on the least number of factors. A quality measure of this trade-off is the Akaike information criterion, AIC, where low AIC values are indicative of good quality. Using each GLM, we calculated the AIC values for each gene (Fig. 2A). The best distribution of AIC values was achieved by applying the *BS* model compared to any of the other GLMs (Fig. 2A). The conditional probability that a particular GLM performs better in the prediction of the expression level of a given gene can also be derived by directly comparing the AIC values of the alternative GLMs in the form of AIC weights (Wagenmakers and Farrell, 2004). The number of genes for which the *BS* model generated the highest probability of predicting the expression more accurately than the other GLMs was 7320, followed by the *lasso10* model (5999) and the GLM in which only cancer type factors were used (*onlyCT*, 4295) (Fig. 2B). Overall, the goodness-of-fit between observed vs. predicted gene expression levels across all 1,082 samples using the *BS* model generated a Pearson correlation coefficient $R = 0.963$ (Fig. 2C). Considering these results, we adopted the *BS* model to test for associations between gene expression and cancer mutations.

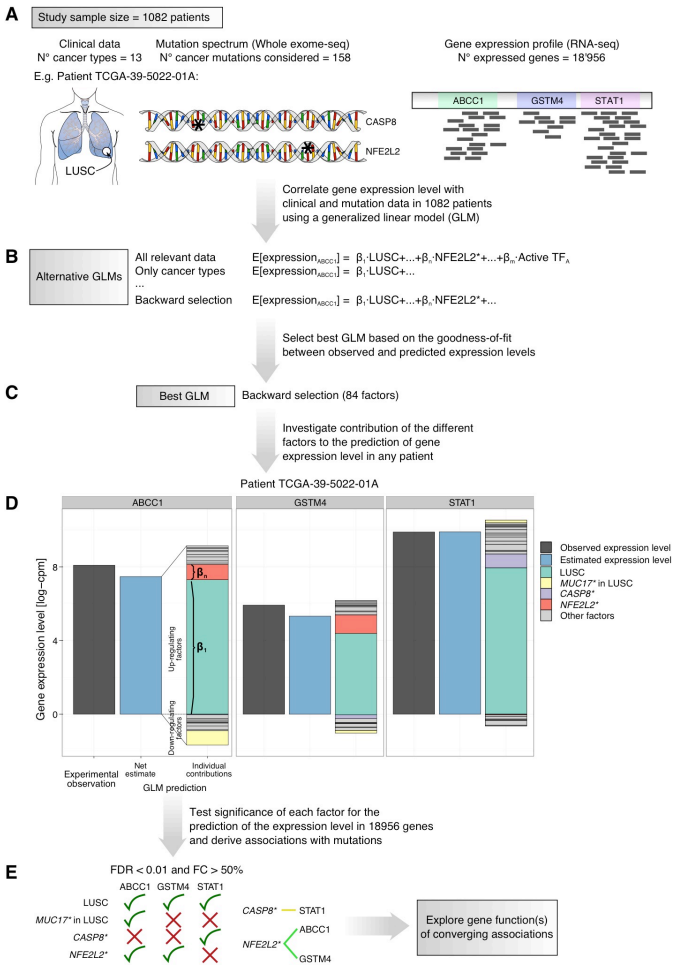


Fig. 1. Workflow used to derive statistical associations between gene expression changes and cancer mutations. (**A**) Input data for the study were collected from 1,082 patients for which clinical, mutation, and gene expression level data were simultaneously generated. LUSC: Lung squamous cell carcinoma. (**B**) The observed level of gene expression was correlated to clinical and mutation data by deriving alternative generalized linear models (GLMs). Each GLM factorizes the contribution of predefined factors to the expression level of a given gene (e.g. *ABCC1*) as a linear regression, where coefficients are estimated by fitting the observed gene expression level in the 1,082 samples. Each GLM predicts an expected value for the expression level of a gene in a sample given the factor values for that sample (e.g. if the sample is LUSC, the GLM adds a contribution equal to its estimated coefficient, β_1). (**C**) Model selection is performed to decide which GLM returns the best predictions while using a minimal number of factors. (**D**) The predicted expression is net sum of positive and negative factors as determined by the model. As example, expression of *ABCC1* is positively affected by a cancer type factor (LUSC, green bar) and a mutation in *NFE2L2* (red bar) but negatively affected by an interaction term, the context specific mutation of *MUC17* in LUSC (yellow bar). (**E**) The significance of each factor can be tested using a threshold for the moderated *t*-statistics and for the minimum expression fold-change. The factors representing mutations can hereby be associated with gene expression changes. For example, a mutation in *NFE2L2* showed a significant statistical association with expression changes in *ABCC1* (green line). Associations identified in this manner were used to derive networks of deregulated biological processes that are independently associated with cancer mutations.

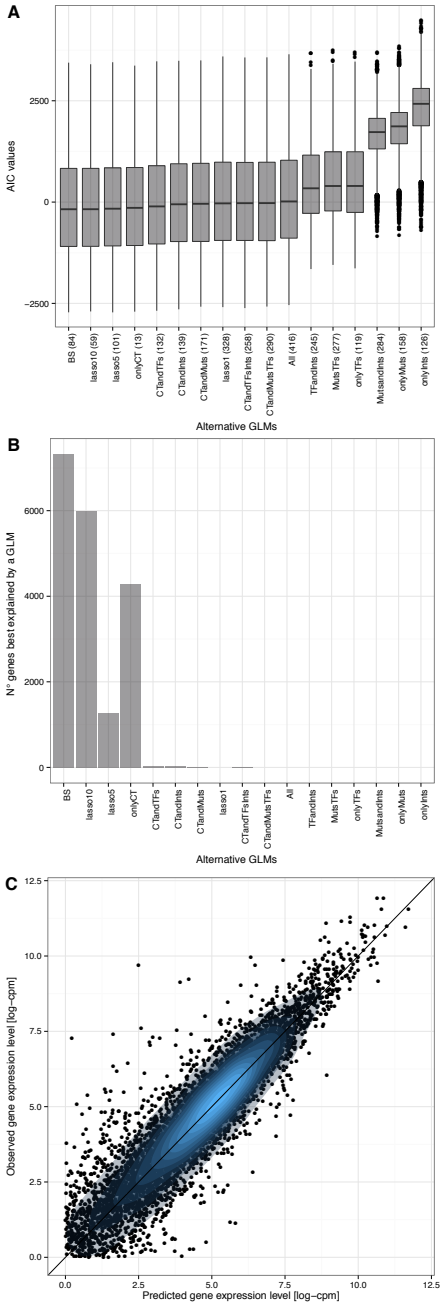


Fig. 2. Model selection according to the minimum Akaike information criterion (AIC) reveals that the backward selection (*BS*) model is better at fitting gene expression across samples than the alternative GLMs. (**A**) Boxplot of AIC values (one for each gene) using alternative GLMs. Key: *BS* – Backward selection model (84 factors); *lasso1* – Lasso non null factors in >1% of all genes (328 factors); *lasso5* – Lasso non null factors in at least >5% of all genes (101 factors); *lasso10* – Lasso non null factors in >10% of all genes (59 factors); *CT* – Cancer type factors (13 factors); *TFs* – Transcription factor expression level factors (120 factors); *Muts* – Presence of a mutation factors (158 factors); *Ints* – Interaction term between presence of a mutation and cancer type (126 factors); *All* – All factors (416 factors). (**B**) Number of genes whose expression is best explained by one of the alternative GLMs based on AIC weights. (**C**) Correlation between observed and predicted gene expression levels using the *BS* model. Bluer contours define areas with increasing density of points.

We also sought to validate whether the genes found here to be associated with one of the 12 mutations actually change their expression when the mutation is present. To this end we used 189 experimentally derived gene-sets, each representing genes whose expression is altered in response to perturbation in a key cancer-associated gene (Subramanian et al., 2005). We then performed a gene-set analysis for each mutation in order to evaluate if the genes found to be associated with it are enriched in any of these 189 gene-sets. We observed an overall high consistency between the direction of regulation of the genes found here to be associated with a given mutation and the corresponding experimentally derived gene-set (Fig. S5). For example, genes here found to be down-regulated when *TP53* is mutated significantly enriched the P53_DN.V1_DN gene-set, which features genes down-regulated in cell lines from the NCI-60 panel with mutated TP53. Taken together, these results suggest that the 12 cancer mutations identified as factors in the BS model are each linked to defined gene expression changes that encompass all tumors bearing the mutation, and are not attributable to a specific cancer type.

Convergence of mutation-associated gene expression changes in the regulation of metabolism

Next, we were interested in elucidating if the genes associated with each mutation are involved in specific biological processes. In particular, we expected that the 12 mutations associate *independently* with processes linked to important cancer-relevant phenotypes, known as the hallmarks of cancer (Hanahan and Weinberg, 2011). Convergence on any of these processes would provide strong evidence that cancer mutations drive the selection of clones that feature properties reflecting these hallmarks. Hence, we checked if the genes associated with the 12 mutations are enriched in any particular biological process, each represented by a distinct Gene Ontology (GO) term. We employed consensus gene-set analysis using Piano (Varemo et al., 2013), which revealed a diverse number of GO biological processes that are significantly associated with each of the examined mutations (FDR < 0.01, Fig. S6). However, contrary to the premises, only a small number of GO biological processes simultaneously associated with more than one mutation (Fig. 3). We further classified these processes with highly significant convergence ($p < 0.01$) according to the 24 ancestor categories they are assigned to within the GO hierarchy. Hereby we observed an over-representation of the GO categories metabolism, immune system processes, response to stimulus and multi-organism process. Intriguingly, metabolism is the GO category with the most robust convergence measured in terms of stable overrepresentation when more stringent criteria for convergence are enforced, followed by immune system processes (Fig. S7). Taken together, these results suggest that the presence of each of these 12 mutations entails a diverse spectrum of gene expression changes in terms of biological processes that are affected, but that the reprogramming induced by these mutations converges in the deregulation of metabolism and immune system processes

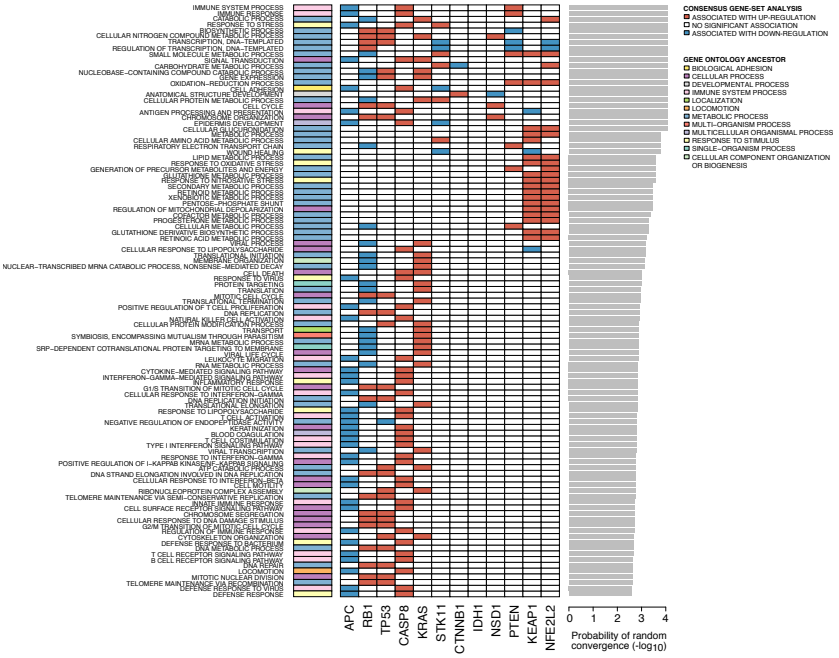


Fig. 3. Mutations converge on the regulation of GO biological processes that relate primarily to metabolism and immune system processes. Each row indicates a GO term that is enriched in the up- (red) or down- (blue) regulated genes associated with each mutation (column) in the consensus gene-set analysis. GO terms are classified according to the ancestor GO category and sorted by the significance of the convergence (barplot on the right).

Mutation-associated gene expression changes converge on a sub-network of metabolic reactions

Metabolism appeared to be the biological process that displayed the largest extent of regulation associated with the 12 mutations. Indeed, mutations in cancer genes have been recognized to deregulate metabolism to meet the metabolic requirements of rapid proliferation and allow cancer cells to adapt to the microenvironment (Cairns et al., 2011; Schulze and Harris, 2013). Others and we have previously found that distinct cancer types featured some common gene expression changes in metabolism compared to their matched normal tissues, and these common changes were primarily ascribed to altered nucleotide biosynthesis (Gatto et al., 2014; Hu et al., 2013; Nilsson et al., 2014). However, these studies could not distinguish whether the observed changes are attributable to a common adaptation process during cancer progression or are rather the consequence of a specific mutation event. To interrogate this, we selected among the genes here associated with 12 mutations those that overlapped with the 3765 genes that participate in the human metabolic network (Mardinoglu et al., 2014). This set corresponds to 525 metabolic genes, each associated with the presence of at least one of the 12 mutations.

The network of associations between a mutation and deregulated metabolic genes revealed a number of genes on which multiple mutations converge (Fig. 4A and S8). However, no metabolic gene showed convergent association with all mutations, nor was there a canonical metabolic process to which all mutations are associated (Fig. 3). We therefore tested the hypothesis that mutations collectively associate with metabolic genes encoding for a common yet non-canonical sub-network of reactions. We first mapped for each reaction in the human metabolic reaction network the number of mutations that converge on it, through the association with the underlying reaction-coding gene(s) (Fig. 4B). This highlighted distinct clusters of reactions within the metabolic network. To extract the largest functional cluster, we searched for a connected sub-network of reactions in which the number of converging mutations is maximized by using the jActiveNetworks algorithm (Ideker et al., 2002). This approach returned a single high-convergence reaction sub-network (Fig. 4C). We characterized this sub-network by determining whether its nodes significantly enrich any pathway and/or metabolite compared to the background human metabolic network. We uncovered that the sub-network featured an over-representation of the metabolism of xenobiotics and polyunsaturated very long-chain fatty acids (Fig. 4D). In addition, individual metabolites, such as glutathione, oxygen, and arachidonic acid, were also over-represented within the sub-network (Fig. 4D). Collectively, these findings suggest that a sub-network of reactions that connects arachidonic acid and xenobiotics via glutathione and oxygen is associated independently with 12 frequent mutations in cancer.

Curation of the high-convergence sub-network of metabolic reactions: AraX

Starting from the high-convergence reaction sub-network, we manually curated a representation of the candidate pathway that best represents these reactions according to the literature. We termed this pathway AraX (Fig. 5), for arachidonic acid and xenobiotic metabolism. The AraX pathway is encoded by 101 metabolic genes. It contains 23% of all mutation-metabolic gene associations. One branch of the AraX pathway comprises reactions that control the availability of arachidonic acid and catalyze its conversion to eicosanoids (34 genes). The second branch facilitates the detoxification of xenobiotics (41 genes). Importantly, nine enzymes encoded by the genes associated with this pathway are involved in both branches (e.g. CYP2C8). In addition, there are 5 transporters that can secrete the end products of the pathway (Fig. 5). The main co-substrates for arachidonic acid and xenobiotic metabolism are oxygen and glutathione, whose levels are controlled by the remaining 21 genes. The overrepresentation of xenobiotics metabolism with cancer mutations was unexpected, considering that the samples used for this study were derived from untreated tumors. The importance of AraX in cancer may reside in its individual components, some of which have established roles in cancer initiation and progression. Aberrant arachidonic acid metabolism regulates processes critical for cancer progression, mainly by establishing a tumor-supporting microenvironment where immune cells and endothelial cells are recruited to produce mitogens, pro-inflammatory cytokines, and angiogenic factors (Wang and Dubois, 2010). Enzymes within the xenobiotics metabolism form reactive intermediates from exogenous and endogenous substrates that can cause cancer initiation, potentially by promoting genotoxicity (Nebert and Dalton, 2006). Both pathways are a primary source of cytosolic reactive oxygen species, which exhibit a characteristically abnormal concentration in many types of cancer cells (Trachootham et al., 2009). Finally, a number of xenobiotic-metabolizing enzymes and transporters in AraX confer cancer cells with mechanisms of detoxification and drug-resistance (Fletcher et al., 2010b). Taken together, this suggests that AraX is implicated in a number of host-cancer interactions that result in pro-tumorigenic functions. We also confirmed that compared to all 186 KEGG metabolic pathways AraX is, on average, the most significantly enriched pathway by the genes associated with a mutation (odds ratio, 12.07; 95% bootstrap confidence interval [CI], 4.75 to 17.66); Fig. S9), followed by xenobiotics metabolism by cytochrome P450 (odds ratio, 5.72; 95%

bootstrap CI, 1.04 to 8.90). Remarkably, similar results were obtained when AraX was compared to the 674 Reactome pathways, which also include signaling pathways that should be highly deregulated in human cancer and that include non metabolic genes, upon which AraX could not be constructed (Fig. S10).

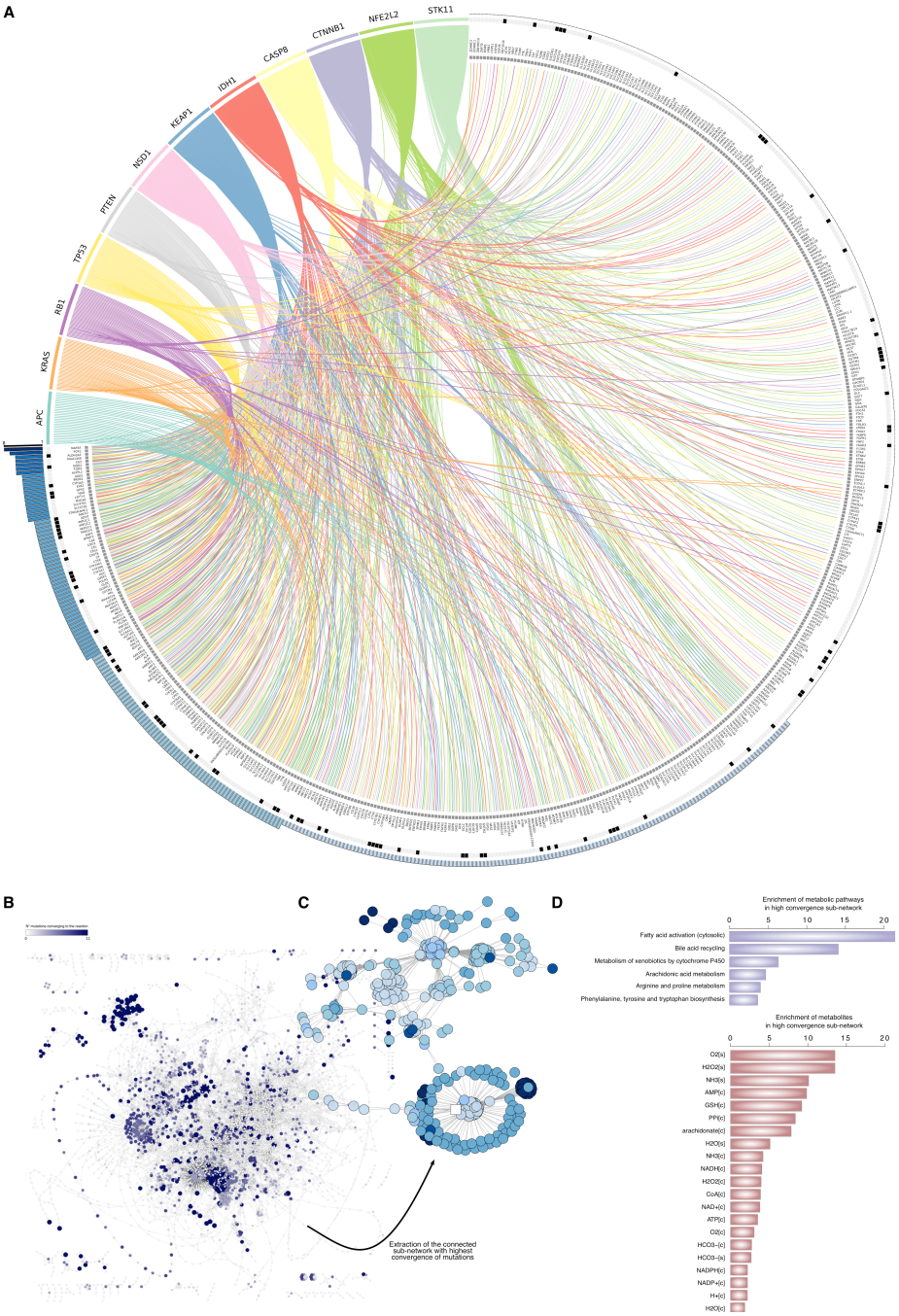


Fig. 4. The network of associations between cancer mutations and metabolic genes reveals a region of high convergence in which genes encode for a metabolic sub-network revolving around arachidonic acid and xenobiotics. (**A**) Circos plot where cancer mutations are connected to metabolic genes if a statistical association was found (high resolution in Fig. S8). Metabolic genes are sorted counter-clockwise according to the number of links (i.e. the number of mutation-metabolic gene associations). Bars indicate the number of mutations that are converging to a particular gene. Black entries in the outer circle indicate genes belonging to AraX (introduced in Fig. 5). (**B**) The human metabolic reaction network where each node is a reaction and the blue gradient indicates the number of mutations converging to it via association with any reaction-encoding gene. (**C**) Extraction of the sub-network where the number of converging mutation-driven transcriptional changes are maximized. (**D**) Characterization of the sub-network in terms of over-represented pathways (top) and metabolites (bottom) compared to the background human metabolic network.

Overall, this finding suggests that deregulation of a network of metabolic reactions connected to arachidonic acid and xenobiotics metabolism and mediated by glutathione and oxygen is advantageous in cancer, since 12 frequent mutations independently entail transcriptional changes that converge on this pathway.

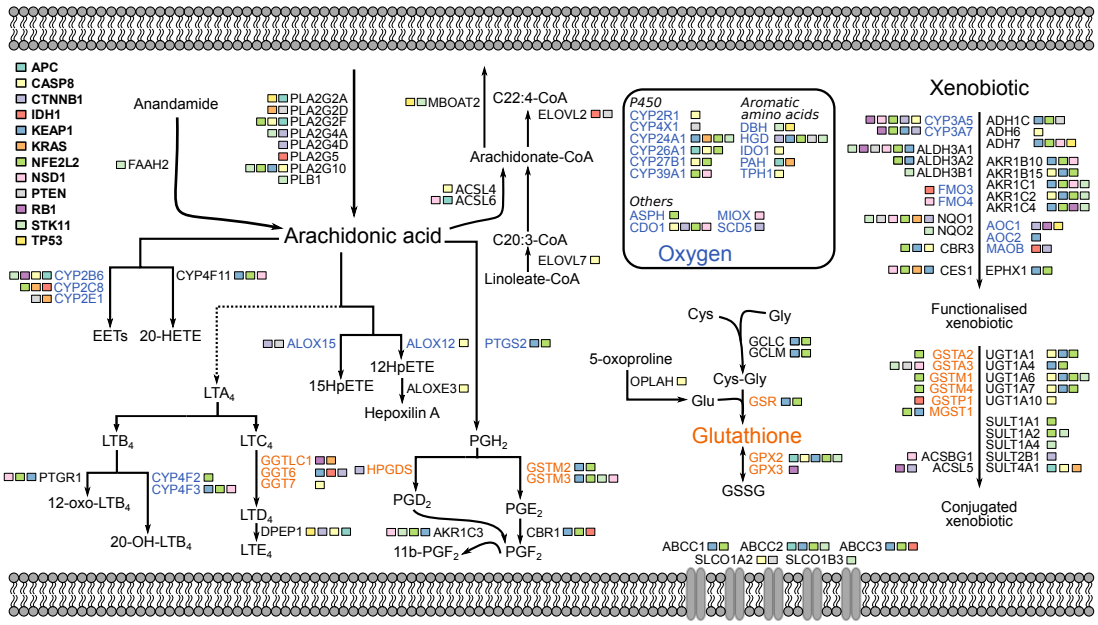


Fig. 5. A literature curated sub-network of reactions that revolves around arachidonic acid and xenobiotic metabolism (AraX) shows convergence by multiple cancer mutations. The boxes next to each gene indicate which mutations are associated with it.

Deregulation of AraX in cancer is the strongest predictor of survival among metabolic pathways

Next, we sought to investigate the implication of the convergence on AraX in cancer. We observed no obvious pattern in the direction of the regulation of AraX by the different mutations. Instead, we found a mutation-specific modulation in the expression of AraX genes, with varying degrees of overlap (Fig. S11). This poses a challenge when devising an intervention strategy to normalize the expression or activity of the AraX pathway aimed at halting cancer progression. On the other hand, this also suggests that a generic deviance (i.e. deregulation) in the expression of AraX is likely to confer a selective advantage in cancer, regardless of the disease type. Hence, we estimated a deregulation score for the AraX pathway in each tumor sample using Pathifier (Drier et al., 2013). This score captures the extent to which the expression of a pathway in a tumor sample deviates from its expression in the normal tissue of origin (Fig. S12). Then we sought to verify whether the extent of AraX deregulation in the tumor is predictive of an independent measure of selective advantage, as determined by patient's survival. Survival analysis for a subset of 783 samples, for which reference normal samples were available, revealed that tumors with a "low" deregulation score, i.e. with an expression level of the AraX pathway similar to that of normal tissue from which the cancers originate, have significantly better prognosis than tumors with a "high" deregulation score. Tumors with a very high deregulation score showed the worst prognosis ($p = 8e^{-6}$, Fig. 6). Compared to the 186 KEGG metabolic pathways, the deregulation of AraX ranks as the best predictor for survival, as estimated by a Lasso penalized Cox proportional hazard model (Fig. S13). At the optimal penalty value (log-λ = -2.6), only four KEGG pathways are predictive of survival, with AraX providing the most robust result (log-hazard ratio per unit of deregulation score, 0.39). Taken together, the strong association of AraX deregulation with poor prognosis suggests that aberrant expression of this pathway confers a stronger selective advantage for cancer progression compared to other metabolic processes.

Discussion

Cancer cells exhibit heterogeneous combinations of genetic alterations that are the result of a process of natural selection. Through this process, cancer cells deregulate critical biological functions to establish the hallmarks of the transformed phenotype (Vogelstein et al., 2013). The concept of convergent evolution in cancer

implies that different genetic alterations can result in functionally similar outputs, which are likely to reflect an evolutionary advantage for the cancer cells with respect to their microenvironment (Gerlinger et al., 2014).

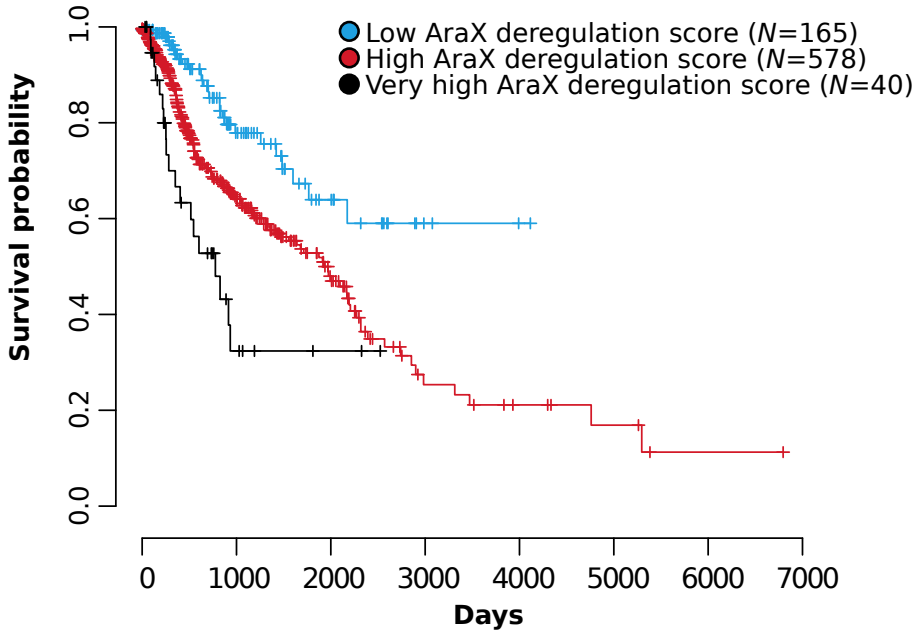


Fig. 6. Kaplan-Meier survival plots for the subset of 783 tumor samples, for which reference normal samples were available, classified with very high (black line), high (dark blue) or low (light blue) deregulation score in the AraX pathway.

In this study, we uncovered one such node of convergence, a metabolic pathway that we termed AraX. Encoded by 101 genes, AraX is a network of metabolic reactions that revolve around the metabolism of arachidonic acid and xenobiotics mediated by oxygen and glutathione. We demonstrate that 12 frequent cancer mutations converge in a significant association with transcriptional deregulation of AraX, more than with any other metabolic or biological pathway. This convergence is striking in that it occurs regardless of the cancer type and independent of the expression of a number of transcription factors.

Intriguingly, the fact that AraX is a transcriptionally regulated pathway of oxygen-consuming reactions could reflect a strategy by which cancer cells adapt to tumor hypoxia by regulating oxygen-dependent enzymes to compensate for reduced oxygen availability. Cancer mutations select independently for the deregulation of this pathway, potentially under the selective pressure of hypoxia. Hence, we speculate that an effective strategy to arrest cancer evolution could be represented by modulating the activity of the AraX pathway, potentially using a multi-targeted approach rather than selective inhibition of individual components, a strategy also advocated by network pharmacology (Hopkins, 2008).

Experimental procedures

Data and analyses used in this study are deposited in Synapse (ID: syn3163200).

Data retrieval. RNAseq gene expression profiles and clinical data for 1,082 primary tumor samples encompassing 13 cancer types (BLCA – Bladder adenocarcinoma, BRCA – Breast carcinoma, COAD – Colon adenocarcinoma, GBM – Glioblastoma multiforme, HNSC – Head and neck squamous cell carcinoma, KIRC – Clear cell renal cell carcinoma, LGG – Low grade glioma, LUAD – Lung adenocarcinoma, LUSC – Lung squamous cell carcinoma, OV – Ovarian carcinoma, READ - Rectum adenocarcinoma, PAAD – Pancreatic adenocarcinoma, UCEC – Uterine corpus endometrial carcinoma) were downloaded from the Cancer Genome Atlas (TCGA) in November 2013. At least 20 samples per cancer type were required to be included in this cohort. Mutation profiles for the same samples were obtained from the cBioPortal (Gao et al., 2013).

Differential gene expression analysis. RNAseq-generated read count tables were used to estimate gene expression in each sample in the pan-cancer cohort. To this end, we adopted voom, an approach that extends the generalized linear model (GLM) for microarray gene expression signals to analyze count-based expression data

(Law et al., 2013). The gene-wise count variance is calculated from the linear regression of gene-wise observed log-counts across all samples in the cohort according to a number of factors as the gene-wise residual standard deviation of the regression. If a lowess curve is fitted to square-root residual standard deviation as a function of mean log-counts, it is possible to predict the square-root standard deviation of each observation (i.e. log-counts for a given gene in a given sample) from the fitted log-count of that observation according to the linear regression (i.e. the mean-variance trend). Differential gene expression analysis for each factor is then performed using the standard linear modeling procedure proposed by limma (Smyth, 2004), with the addition that the log-counts-per-million of each observation are corrected using the predicted variance as an inverse weight. Even if voom assumes that each observation is normally distributed, this method proved to outperform count-based approaches in differential expression analysis comparison studies (Rapaport et al., 2013). The significance of each factor in the regression of the expression of each gene is then tested using moderated *t*-statistics. So generated *p*-values were corrected for multiple testing by controlling the false discovery rate (FDR) across genes using the Benjamini and Hochberg correction and by adopting the nestedF correction across contrasts. A factor is deemed significant in the regression of the expression of a gene if it is associated to at least 50% fold change (|log-FC| > 1.5) with a FDR < 0.01.

Model selection. In order to perform the differential gene expression analysis above, it is required to define the factors for the regression. These factors are devoted to explain the biological variability of gene-wise counts across the samples in the pan-cancer cohort. They should capture the main contributions and some smaller contributions interesting to our investigation. Hence we tentatively selected the following factors for an initial design (*All*):

- The cancer types, i.e. the belonging to a histopathologically defined cancer type among the 13 types in the cohort;
- The mutation status of 158 cancer-associated genes. An initial list of 260 genes was generated by merging the Cancer5000 and Cancer5000-S lists in (Lawrence et al., 2014). We excluded *HIST1H3B, HIST1H4E,* and *MLL4*, which could not be uniquely mapped using the Ensembl v.73 annotation. Furthermore, 102 genes that were not mutated at moderate frequency in the cohort (>2%) were also excluded.
- The activation status of 119 well-characterized transcription factors (Zambelli et al., 2012), defined by the belonging to a certain quintile of expression in the pan-cancer cohort.
- The interaction terms between a cancer type and a cancer-associated gene mutated at high frequency. These are defined as the 12 mutations with a frequency >10% across the pan-cancer cohort. There are 126 such interaction terms, excluding those linearly dependent on the other factors. These factors take in account cancer type-dependent contributions of mutations.

Using the same notation (where appropriate) as in voom, the GLM (1) is:

$$(1) \qquad E(y_{g,i}) = \mu_{g,i} = I_g + x_i^T \beta_g$$

where $y_{g,i}$ is the log-counts per million (log-cpm) value for gene g in sample i, $\mu_{g,i}$ is the expected value, x_i is the vector of covariate values in sample i, β_g is the (unknown) vector of coefficients representing the contribution of each covariate on the expected value, and I_g is the explicitly formulated intercept of the GLM. In our formulation, the *All* model (2) becomes:

$$(2) \qquad \mu_{g,i} = I_g + \left(\sum_{m=1}^{nCancerMutations} x_m + \sum_{t=1}^{nCancerTypes} x_t + \sum_{f=1}^{nTranscriptionFactors} x_f + \sum_{I=1}^{nInteractions} x_I\right)^T \beta_g$$

where x_m is a binary value {0,1} indicating the absence or presence of mutation m in the sample i; x_t is a binary value {0,1} indicating the belonging of sample i to the cancer type t; x_f is a ternary value {-1,0,1} indicating whether the expression of transcription factor f in sample i is in the bottom quintile, 2nd to 4th quintile, or top quintile with respect to the distribution of its expression values in the pan-cancer cohort; and x_I is a binary value {0,1} indicating whether there is the interaction I between the cancer type to which sample i belongs and a frequently mutated gene.

We excluded the following observations from this study:

1. All genes that have ambiguous annotation in Ensembl v73. This set corresponds to 565 genes.
2. All genes that were not detected in any sample. A gene is detected if at least 10 counts were reported in 10% of the samples. Although the opposite may occur due to an actual repression of the gene, this signal cannot be distinguished from genes that are misannotated or, more likely, from genes whose transcripts cannot be detected due to technical limitation in the sensitivity of the sequencing instrument. These observations do not add any information on the expression status of the (presumptive) gene and thus their removal will not alter the result of downstream analyses. This set corresponds to 1075 genes.

Overall, 1575 genes were excluded from the initial set of 20,531 genes (65 overlapped between the above mentioned filtered sets), yielding a total of 18,956 genes analyzed.

Many factors in the *All* model are unlikely to contribute in explaining the expression of most genes, thereby increasing the risk of over-fitting. We adopted two different model selection methods to derive the most relevant factors while using a minimal number of factors. First, backward selection (Yan and Su, 2009) was used to exclude, at each iteration, the factor that is associated with the least number of differentially expressed genes. The procedure was stopped once the number of differentially expressed genes was greater than 1% of all genes (i.e. 190 genes). The resulting GLM contains 84 factors (*BS* design). Second, we used L1-constrained regression shrinkage using the Lasso algorithm (Tibshirani, 1996) to compute, for each gene, the factors in the *All* model with a non-null coefficient. The penalty value used for the Lasso regression was calculated such that the mean 10-fold cross-validated error is minimum. The Lasso method was implented using the R-package glmnet (Friedman et al., 2010). We constructed three GLMs based on the factors with a non-null coefficient in at least 1%, 5%, or 10% of all genes (*lasso1, lasso5,* and *lasso10*). Finally, we constructed alternative GLMs that feature either only the cancer type (*CT*) or the transcription factor levels (*TF*) or the mutation statuses (*Muts*) or the interactions (*Ints)* factors, or any other combination of these classes. The best GLM was evaluated by first calculating the Akaike information criterion (AIC) values for the goodness-of-fit of all genes by each GLM. This criterion was chosen for its ability to capture the trade-off between the goodness of fit and the number of factors utilized in the regression of the expression of a gene (for each GLM, there is an AIC value per gene). We next computed, for each gene, the difference between the AIC value returned by the current GLM and the minimum AIC value observed using any GLM. From this, we calculated the AIC weight of the alternative GLMs in the regression of each gene. The AIC weights were transformed into probabilities that a certain GLM is the most likely to explain the expression of that gene. Finally, we counted for each GLM the number of genes whose expression is best explained by that GLM. The model selection was implemented in R 3.1.2.

Gene-set analyses. The gene-set analyses were performed using the R-package Piano (Varemo et al., 2013). In all analyses, we evaluated the significance of a gene-set using the genes found here to be associated with a mutation (here on mutation-associated genes). For each mutation, the list of mutation-associated genes is generated using the differentially gene expression analysis based on the *BS* model (see Differential gene expression analysis). In the case of enrichment of the 189 gene-sets representing each a genetic perturbation in a key cancer-associated gene [retrieved from the Molecular Signatures Database (MSigDB) (Subramanian et al., 2005)], the significance of a gene-set was tested using the Fisher's test, and the *p*-values were controlled for multiple testing by transformation to FDR using the Benjamini and Hochberg correction. To check for consistency between the genetic perturbation represented by a gene-set and the expected effect on gene expression by a mutation, we compared separately the top five ranked gene-sets (if significant, i.e. gene-set FDR < 0.01) mostly associated with up-regulated or down-regulated genes (in Piano, so called "mixed directional" classes). For example, genes here found up-regulated when *APC* is mutated are significantly associated with the BCAT_BILD_ET_AL_UP gene-set, in which β-catenin (BCAT), a direct target of APC, was over-expressed in primary epithelial breast cancer cell.
In the case of enrichment of GO biological processes, 8255 gene-sets were retrieved using the R-package biomaRt (Durinck et al., 2009). The significance of a gene-set was tested using the consensus between six tests (Fisher's test, Stouffer's test, Reporter test, Tail strength test, mean, and median), and the *p*-values were controlled for multiple testing by transformation to FDR using the Benjamini and Hochberg correction. If gene-set FDR < 0.01, the underlying biological process is deemed significantly associated with the mutation. To compute the probability that multiple mutations are simultaneously associated with a gene-set, we designed a permutation test in which the gene-sets significantly associated with a mutation are randomly permuted 10'000 times. Then, we calculated a *p*-value as the frequency at which a gene-set is randomly associated with a number of mutations greater or equal to that observed prior randomization. Next, we computed using the Fisher' Exact Test which ancestor GO category (defined as the children of the GO term *biological process*) were overrepresented by the GO terms that showed significant convergence. Finally, we estimated the robustness of the supposed overrepresentation of a primary GO biological processes repeating this above operation using only those GO terms that showed convergence by an increasing number of mutations.

Extraction of the high-convergence reaction sub-network. The human genome-scale metabolic model HMR2 was downloaded from http://www.metabolicatlas.com/. We generated a reaction network from the model where reactions are nodes, and an edge links two nodes if there is at least one metabolite shared by the two reactions. We excluded 18 metabolites with exceptionally high degree (>200) to prevent a combinatorial explosion of reaction-reaction edges. Then, we used the jActiveNetwork algorithm (Ideker et al., 2002) to extract from this reaction network a connected sub-network that maximizes the number of mutations converging to it. To this end, we counted for each reaction the number of times that any mutation is found associated with a gene encoding that reaction. Each

reaction of the network was then scored using this count. We subtracted a penalty equal to 3 to the score to ensure that the extracted sub-network was reasonably small yet comprised as many reactions with at least 2 mutations converging to them. Artificial reactions introduced in HMR2 for modeling purposes (defined by the HMR2 sub-systems *Isolated*, *Artificial reactions*, *Exchange reactions*, *Pool reactions*) were further penalized with a score of -100. The search was implemented using the R-package BioNet (Beisser et al., 2010). The returned high-convergence reaction sub-network contained 389 reactions (nodes) out of the 8184 reactions that were present in the reaction network.

Analysis of the high-convergence reaction sub-network. We characterized the high-convergence reaction sub-network by comparing the frequency of metabolites and pathways represented by the reactions in the sub-network to the background frequency in HMR2. The overrepresentation of metabolites and pathways was calculated using the Fisher's Exact Test. To further aid the interpretation of the reactions part of the high-convergence reaction sub-network, this was broken down in reaction clusters, defined as sets of reactions that share the same gene-reaction association. These are returned by applying unsupervised hierarchical clustering to the gene-reaction association matrix in HMR2 limited to include the reactions in the high-convergence reaction sub-network and the genes associated with at least one mutation. This operation reduced the complexity of the high-convergence reaction sub-network to 57 reaction clusters.

Curation of the high-convergence reaction sub-network. Starting from the above analysis, we consulted the literature to frame the high-convergence reaction sub-network in the context of well-defined metabolic functions and reconstruct a comprehensive pathway. Also, we manually reviewed every metabolic gene associated with at least one mutation and verified if there exist a relation with the emerging pathway. We initially focused on arachidonic acid and its metabolism given its prominent enrichment in the high-convergence reaction sub-network compared to HMR2. Reaction cluster #14 revealed a significant number of enzymes responsible for the cleavage of arachidonic acid from cellular lipids among the mutation-associated genes. PLA2G2A, PLA2G2D, PLA2G2F, PLA2G4A, PLA2G4D, PLA2G5, PLA2G10, and PLB1 all belong to the class of phospholipases A_2 and function to release free fatty acids from the *sn*-2 position of phospholipids (Astudillo et al., 2012). Noteworthy, PLA2G2A shows an exquisite preference towards phospholipids containing arachidonic acid at the *sn*-2 position (Murphy and Gijon, 2007). In this context, reaction cluster #2 implicates three fatty acid binding proteins (FABP1, FABP4, and FABP6) in the trafficking of long chain fatty acids. However, the exact role and affinity towards arachidonic acid is still debated (Anderson and Stahl, 2013; Furuhashi and Hotamisligil, 2008). We thus decided to exclude these genes from the candidate pathway. In contrast, manual review revealed that another mutation-associated gene, FAAH2, affects arachidonic acid availability. Specifically, FAAH2 degrades endogenous cannabinoid anandamide to release arachidonic acid (Wei et al., 2006). The reaction cluster #8 indicates inclusion of reactions belonging to the cytochrome P450-pathways of arachidonic acid. These include reactions in epoxygenase pathway, catalyzed by CYP2B6, CYP2C8, CYP2E1, and in the hydroxylase pathway, catalyzed by CYP4F11 (Arnold et al., 2010; Kroetz and Zeldin, 2002). CYP4X1 is also a likely member to the epoxygenase pathway, but evidence for specificity to arachidonic acid is still unconclusive (Kumar, 2015; Stark et al., 2008). Reaction clusters #7, #11, #12 and #41 and manual review of mutation-associated genes implicate the lipoxygenase (LOX) pathway of arachidonic acid and the metabolism of a class of LOX products, leukotrienes. CYP4F2, CYP4F3, and PTGR1 catalyze the inactivation of leukotriene B_4, a product of arachidonic acid metabolism, either by w-oxidation or via the 12HDH/15oPGR pathway (Murphy and Gijon, 2007). Four other mutation-associated genes are involved in the conversion of leukotriene C_4 to leukotriene D_4 and then leukotriene E_4, on one hand GGTLC1, GGT6, GGT7 (Murphy and Gijon, 2007), and the other DPEP1 (Croft et al., 2014). In addition three other mutation-associated genes belong to the LOX pathway, on one hand ALOX12B and ALOXE3, on the other ALOX15. ALOX12B and ALOXE3 are responsible for the synthesis of another class of LOX products, hepoxilins, and in particular hepoxilin A (Munoz-Garcia et al., 2014). ALOX15 catalyzes the first step in the synthesis of yet another class of LOX products, lipoxilins (Schneider and Pozzi, 2011). Reaction cluster #1 implicates activation of very long-chain fatty acid, such as arachidonic acid, by acyl-CoA synthetases ACSL4, ACSL5, ACSL6, and ACSBG1. In particular, ACSL4 and ACSL6 show selectivity towards arachidonic acid, in that they constitute the first step for its incorporation into cellular lipids (Astudillo et al., 2012). Intriguingly, reaction cluster #49 connects fatty acid elongases to such activation of very long-chain fatty acid. In particular, ELOVL7 and ELOVL2 participate in the elongation of w-6 fatty acids, respectively upstream and downstream of arachidonate (Ohno et al., 2010). MBOAT2 is instead involved in the Land's cycle to reincorporate activated arachidonic acid in the membrane lipids (Astudillo et al., 2012). Besides the LOX and cytochrome P450 pathway, another major route of arachidonic acid is the cyclooxygenase (COX) pathway to produce prostaglandins. Manual review implicates six mutation-associated genes in the metabolism of prostaglandin H_2,

the first product of arachidonic acid conversion in the COX pathway. Prominently, PTGS2 (also known as COX-2) catalyzes the first common step in the COX pathway from arachidonic acid to prostaglandin H_2 (Schneider and Pozzi, 2011). HPGDS converts prostaglandin H_2 to prostaglandin D_2 (Schneider and Pozzi, 2011). GSTM2 and GSTM3 can convert prostaglandin H_2 to prostaglandin E_2 (Hayes et al., 2005), which in turn can be converted to prostaglandin F_{2a} by CBR1 (2015; Malatkova et al., 2010). AKR1C3 can reduce prostaglandin H_2 and D_2 to prostaglandin F_{2a} and 11b-prostaglandin F_{2a}, respectively (Penning, 2014). Next we focused on xenobiotics metabolism, among the most enriched pathways in the high-convergence reaction sub-network. We first noticed that nine genes overlap with the metabolism of arachidonic acid. ACSL5 and ACSBG1 (Wermuth, 2003), AKR1C3 (Penning, 2014), CBR1 (Malatkova et al., 2010), CYP2B6, CYP2C8, CYP2E1 (Wermuth, 2003), GSTM2 and GSTM3 (Hayes et al., 2005) have also reported activity in the detoxification of electrophilic xenobiotics. Reaction clusters #4, #7, #16 and #21 implicate phase I of xenobiotics metabolism (also called functionalization). After manual review, we gathered a total of 23 genes involved in the functionalization phase. The great majority (21) are oxidoreductases in the family of cytochrome P450 (CYP3A5, CYP3A7), alcohol dehydrogenases (ADH1C, ADH6, ADH7), flavin-containing monoxygenases (FMO3, FMO4), aldo-keto reductases (AKR1B10, AKR1B15, AKR1C1, AKR1C2, AKR1C4), quinone reductases (NQO1, NQO2), carbonyl reductases (CBR3), aldehyde dehydrogenases (ALDH3A1, ALDH3A2, ALDH3B1), and amine oxidases (AOC1, AOC2, MAOB) (Brozic et al., 2011; Quinn et al., 2008; Wermuth, 2003). The two remaining genes, CES1 and EPHX1, belong instead to the class of hydrolases (Wermuth, 2003). Reaction cluster #3 implicates phase II of xenobiotics metabolism, also known as conjugation. Collectively, we found 14 genes that can catalyze conjugation reactions among the mutation-associated genes, besides the above-mentioned ACSL5 and ACSBG1. UGT1A1, UGT1A4, UGT1A6, UGT1A7, and UGT1A10 are UDPGA transferases that carry glucuronidation reactions on xenobiotics (Wermuth, 2003). GSTA2, GSTA3, GSTM1, GSTM4, GSTP1, and MGST1 catalyze the conjugation of glutathione (Hayes et al., 2005). SULT1A1, SULT1A2, SULT1A4, SULT2B1, and SULT4A1 belong to the family of sulfotransferases and are responsible for sulfonation reactions on xenobiotics using PAPS as cofactor (Wermuth, 2003). Finally, reaction clusters #22 and #56 include transporters for both arachidonic acid-derived products and solubilized xenobiotics. The organic anion transporters SLCO1A2 and SLCO1B3 show affinity for prostaglandin E_2 and leukotriene C_4, respectively (Thiriet, 2012). The ABC transporters ABCC1 and ABCC3 are renowned for their ability to move a variety of xenobiotics, but other substrates include prostaglandin A_1, A_2, D_2, E_2, 15d J_2 and leukotriene C_4 (Fletcher et al., 2010a). Manual review revealed an additional ABC transporter with related activity among the mutation-associated genes, ABCC2 (Fletcher et al., 2010a). The enrichment for the occurrence of oxygen- and glutathione-consuming reactions in the high-convergence reaction sub-network persuaded us to investigate which other genes support their metabolism. There are six enzymes among the mutation-ssociated genes that are involved in glutathione biosynthesis, GCLC, GCLM, GPX2, GPX3, GSR, and OPLAH (Pompella et al., 2002). These expand the list of glutathione-utilizing enzymes in the candidate pathway to a total of 17 members. Also, 15 additional mutation-associated genes encode for reactions that use oxygen: six belong to the cytochrome P450 (CYP2R1, CYP4X1, CYP24A1, CYP26A1, CYP27B1, CYP39A1); five participate in the metabolism of aromatic amino acids (DBH, HGD, IDO1, PAH, TPH1); while the remaining genes have disparate metabolic activities (ASPH, CDO1, MIOX, SCD5). We neglected the result on the enrichment for the pathways bile acid recycling and phenylalanine, tyrosine, and tryptophan biosynthesis because the associated genes that drove the enrichment are best explained by xenobiotics metabolism. The so-reconstructed candidate pathway features 34 genes attributable to arachidonic acid metabolism, 41 genes attributable to xenobiotics metabolism, 21 genes that mediate glutathione and oxygen metabolism, and 5 genes in the transport system. We reviewed each protein in this pathway in UniProt and/or Reactome to validate the gene annotation provided by literature (2015; Croft et al., 2014). In total, 101 out of 525 of all mutation-associated metabolic genes are represented in this pathway. We termed this pathway AraX.

Enrichment of pathways by mutation-associated genes. We calculated the overrepresentation of AraX by each group of mutation-associated genes compared to any other KEGG metabolic pathways (189) or Reactome pathways (674), as retrieved in MSigDB, using the Fisher's Exact Test. The mean enrichment of a pathway across all mutations was subject to bootstrapping (1'000 replicates) in order to calculate the 95% confidence interval for the mean enrichment. This operation allows evaluating the robustness of a pathway mean enrichment to outliers (i.e. mutations strongly associated with a pathway).

Survival analysis. The deregulation at the level of gene expression for a metabolic pathway in a sample was estimated using Pathifier (Drier et al., 2013). This algorithm returns a score between 0 and 1 that represents the extent to which the expression of a pathway in a sample is deviating from the centroid pathway expression in normal samples. Hence, we calculated the score for all

tumor samples in this study belonging to six cancer types for which matched normal samples were available in TCGA. These normal samples were used to provide the reference expression level of the pathway in a tissue. Next, we binned the tumor samples according to their deregulation score into "low" (if the score is below the 95th percentile of the scores across all normal samples), "high" (if above), or "very high" (if above the 95th percentile of the scores across all tumor samples). Thus, if a tumor sample has a "low" score, then its deregulation is similar to the most deregulated of the normal samples of reference. Kaplan-Meier curves were generated for each group, and the significance of survival difference was estimated using the log-rank test. In order to calculate which metabolic pathway deregulation has the foremost effect in the prediction of survival, we used a lasso penalized Cox regression model. Patient survival was regressed using a Cox proportional hazards model that uses as variables the deregulation score of 186 KEGG metabolic pathways and AraX. The selection of variables relevant to predict survival was performed using increasing values for the lasso penalty (log-λ) used in the regression. The optimal penalty value was calculated such that the mean 10-fold cross-validated error was minimum. Out of 187 initial variables, only 4 variables are predictive of survival at the optimal penalty.

Author contributions

FG designed the study and performed the analyses; FG, AS, and JN discussed and interpreted the results; FG and JN conceived the study.

Acknowledgements

The authors would like to acknowledge the following people for their contributions to this study: Erik Kristiansson's group in Chalmers University of Technology for support in statistics; Pontus Hjortskog, David Jensen, Kenny Nilsson, and Luuk van Egeraat for support in the data retrieval and storage; Martin Eilers in University of Würzburg for discussion; Adil Mardinoglu, Ivan Mijakovic, and Leif Väremo for a critical review of the manuscript. The computations were performed on resources provided by the Swedish National Infrastructure for Computing (SNIC) at C3SE. F.G. and J.N. acknowledge Knut and Alice Wallenberg Foundation for financing this work.

References

(2015). UniProt: a hub for protein information. Nucleic acids research 43, D204-212.

Anderson, C.M., and Stahl, A. (2013). SLC27 fatty acid transport proteins. Molecular aspects of medicine 34, 516-528.

Arnold, C., Konkel, A., Fischer, R., and Schunck, W.H. (2010). Cytochrome P450-dependent metabolism of omega-6 and omega-3 long-chain polyunsaturated fatty acids. Pharmacological reports : PR 62, 536-547.

Astudillo, A.M., Balgoma, D., Balboa, M.A., and Balsinde, J. (2012). Dynamics of arachidonic acid mobilization by inflammatory cells. Biochimica et biophysica acta 1821, 249-256.

Beisser, D., Klau, G.W., Dandekar, T., Muller, T., and Dittrich, M.T. (2010). BioNet: an R-Package for the functional analysis of biological networks. Bioinformatics 26, 1129-1130.

Brozic, P., Turk, S., Rizner, T.L., and Gobec, S. (2011). Inhibitors of aldo-keto reductases AKR1C1-AKR1C4. Current medicinal chemistry 18, 2554-2565.

Cairns, R.A., Harris, I.S., and Mak, T.W. (2011). Regulation of cancer cell metabolism. Nature reviews. Cancer 11, 85-95.

Cerami, E., Gao, J., Dogrusoz, U., Gross, B.E., Sumer, S.O., Aksoy, B.A., et al. (2012). The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. Cancer discovery 2, 401-404.

Chin, L., Andersen, J.N., and Futreal, P.A. (2011). Cancer genomics: from discovery science to personalized medicine. Nature medicine 17, 297-303.

Ciriello, G., Miller, M.L., Aksoy, B.A., Senbabaoglu, Y., Schultz, N., and Sander, C. (2013). Emerging landscape of oncogenic signatures across human cancers. Nature genetics 45, 1127-1133.

Croft, D., Mundo, A.F., Haw, R., Milacic, M., Weiser, J., Wu, G., et al. (2014). The Reactome pathway knowledgebase. Nucleic acids research 42, D472-477.

DeNicola, G.M., Karreth, F.A., Humpton, T.J., Gopinathan, A., Wei, C., Frese, K., et al. (2011). Oncogene-induced Nrf2 transcription promotes ROS detoxification and tumorigenesis. Nature 475, 106-109.

Drier, Y., Sheffer, M., and Domany, E. (2013). Pathway-based personalized analysis of cancer. Proceedings of the National Academy of Sciences of the United States of America 110, 6388-6393.

Durinck, S., Spellman, P.T., Birney, E., and Huber, W. (2009). Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. Nature protocols 4, 1184-1191.

Fletcher, J.I., Haber, M., Henderson, M.J., and Norris, M.D. (2010a). ABC transporters in cancer: more than just drug efflux pumps. Nature reviews. Cancer 10, 147-156.

Fletcher, J.I., Haber, M., Henderson, M.J., and Norris, M.D. (2010b). ABC transporters in cancer: more than just drug efflux pumps. Nature Reviews Cancer 10, 147-156.

Fodde, R., Edelmann, W., Yang, K., van Leeuwen, C., Carlson, C., Renault, B., et al. (1994). A targeted chain-termination mutation in the mouse Apc gene results in multiple intestinal tumors. Proceedings of the National Academy of Sciences of the United States of America 91, 8969-8973.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. J Stat Softw 33, 1-22.
Furuhashi, M., and Hotamisligil, G.S. (2008). Fatty acid-binding proteins: role in metabolic diseases and potential as drug targets. Nature reviews. Drug discovery 7, 489-503.
Gao, J., Aksoy, B.A., Dogrusoz, U., Dresdner, G., Gross, B., Sumer, S.O., et al. (2013). Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Science signaling 6, pl1.
Gatto, F., Nookaew, I., and Nielsen, J. (2014). Chromosome 3p loss of heterozygosity is associated with a unique metabolic network in clear cell renal carcinoma. Proceedings of the National Academy of Sciences of the United States of America 111, E866-875.
Gerlinger, M., McGranahan, N., Dewhurst, S.M., Burrell, R.A., Tomlinson, I., and Swanton, C. (2014). Cancer: Evolution Within a Lifetime. Annu Rev Genet 48, 215-236.
Hanahan, D., and Weinberg, R.A. (2011). Hallmarks of cancer: the next generation. Cell 144, 646-674.
Hayes, J.D., Flanagan, J.U., and Jowsey, I.R. (2005). Glutathione transferases. Annual review of pharmacology and toxicology 45, 51-88.
Hofree, M., Shen, J.P., Carter, H., Gross, A., and Ideker, T. (2013). Network-based stratification of tumor mutations. Nature methods 10, 1108-1115.
Hopkins, A.L. (2008). Network pharmacology: the next paradigm in drug discovery. Nat Chem Biol 4, 682-690.
Hu, J., Locasale, J.W., Bielas, J.H., O'Sullivan, J., Sheahan, K., Cantley, L.C., et al. (2013). Heterogeneity of tumor-induced gene expression changes in the human metabolic network. Nature biotechnology.
Ideker, T., Ozier, O., Schwikowski, B., and Siegel, A.F. (2002). Discovering regulatory and signalling circuits in molecular interaction networks. Bioinformatics 18 Suppl 1, S233-240.
Johnson, L., Mercer, K., Greenbaum, D., Bronson, R.T., Crowley, D., Tuveson, D.A., et al. (2001). Somatic activation of the K-ras oncogene causes early onset lung cancer in mice. Nature 410, 1111-1116.
Kandoth, C., McLellan, M.D., Vandin, F., Ye, K., Niu, B., Lu, C., et al. (2013). Mutational landscape and significance across 12 major cancer types. Nature 502, 333-339.
Kroetz, D.L., and Zeldin, D.C. (2002). Cytochrome P450 pathways of arachidonic acid metabolism. Current opinion in lipidology 13, 273-283.
Kumar, S. (2015). Computational identification and binding analysis of orphan human cytochrome P450 4X1 enzyme with substrates. BMC research notes 8, 9.
Law, C.W., Chen, C., Shi, W., and Smyth, G.K. (2013). Voom! precision weights unlock linear model analysis tools for RNA-seq read counts. In http://www.statsci.org/smyth/pubs/VoomPreprint.pdf.
Lawrence, M.S., Stojanov, P., Mermel, C.H., Robinson, J.T., Garraway, L.A., Golub, T.R., et al. (2014). Discovery and saturation analysis of cancer genes across 21 tumour types. Nature 505, 495-501.
Malatkova, P., Maser, E., and Wsol, V. (2010). Human carbonyl reductases. Current drug metabolism 11, 639-658.
Mardinoglu, A., Agren, R., Kampf, C., Asplund, A., Uhlen, M., and Nielsen, J. (2014). Genome-scale metabolic modelling of hepatocytes reveals serine deficiency in patients with non-alcoholic fatty liver disease. Nature communications 5, 3083.
Munoz-Garcia, A., Thomas, C.P., Keeney, D.S., Zheng, Y., and Brash, A.R. (2014). The importance of the lipoxygenase-hepoxilin pathway in the mammalian epidermal barrier. Biochimica et biophysica acta 1841, 401-408.
Murphy, R.C., and Gijon, M.A. (2007). Biosynthesis and metabolism of leukotrienes. The Biochemical journal 405, 379-395.
Nebert, D.W., and Dalton, T.P. (2006). The role of cytochrome P450 enzymes in endogenous signalling pathways and environmental carcinogenesis. Nature Reviews Cancer 6, 947-960.
Nilsson, R., Jain, M., Madhusudhan, N., Sheppard, N.G., Strittmatter, L., Kampf, C., et al. (2014). Metabolic enzyme expression highlights a key role for MTHFD2 and the mitochondrial folate pathway in cancer. Nature communications 5, 3128.
Ohno, Y., Suto, S., Yamanaka, M., Mizutani, Y., Mitsutake, S., Igarashi, Y., et al. (2010). ELOVL1 production of C24 acyl-CoAs is linked to C24 sphingolipid synthesis. Proceedings of the National Academy of Sciences of the United States of America 107, 18439-18444.
Penning, T.M. (2014). The aldo-keto reductases (AKRs): Overview. Chemico-biological interactions.
Podsypanina, K., Ellenson, L.H., Nemes, A., Gu, J., Tamura, M., Yamada, K.M., et al. (1999). Mutation of Pten/Mmac1 in mice causes neoplasia in multiple organ systems. Proceedings of the National Academy of Sciences of the United States of America 96, 1563-1568.
Pompella, A., Bánhegyi, G.b., and Wellman-Rousseau, M. (2002). Thiol metabolism and redox regulation of cellular functions. (Amsterdam ; Washington, DC: IOS Press).
Quinn, A.M., Harvey, R.G., and Penning, T.M. (2008). Oxidation of PAH trans-dihydrodiols by human aldo-keto reductase AKR1B10. Chemical research in toxicology 21, 2207-2215.
Rapaport, F., Khanin, R., Liang, Y., Pirun, M., Krek, A., Zumbo, P., et al. (2013). Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. Genome biology 14, R95.
Sasaki, M., Knobbe, C.B., Munger, J.C., Lind, E.F., Brenner, D., Brustle, A., et al. (2012). IDH1(R132H)

mutation increases murine haematopoietic progenitors and alters epigenetics. Nature 488, 656-659.
Schneider, C., and Pozzi, A. (2011). Cyclooxygenases and lipoxygenases in cancer. Cancer metastasis reviews 30, 277-294.
Schulze, A., and Harris, A.L. (2013). How cancer metabolism is tuned for proliferation and vulnerable to disruption (vol 491, pg 364, 2012). Nature 494, 130-130.
Smyth, G.K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Statistical applications in genetics and molecular biology 3, Article3.
Stark, K., Dostalek, M., and Guengerich, F.P. (2008). Expression and purification of orphan cytochrome P450 4X1 and oxidation of anandamide. Febs J 275, 3706-3717.
Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., et al. (2005). 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 102, 15545-15550.
Thiriet, M. (2012). Signaling at the cell surface in the circulatory and ventilatory systems. (New York: Springer).
Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. J Roy Stat Soc B Met 58, 267-288.
Trachootham, D., Alexandre, J., and Huang, P. (2009). Targeting cancer cells by ROS-mediated mechanisms: a radical therapeutic approach? Nature Reviews Drug Discovery 8, 579-591.
Varemo, L., Nielsen, J., and Nookaew, I. (2013). Enriching the gene set analysis of genome-wide data by incorporating directionality of gene expression and combining statistical hypotheses and methods. Nucleic acids research 41, 4378-4391.
Vogelstein, B., Papadopoulos, N., Velculescu, V.E., Zhou, S., Diaz, L.A., Jr., and Kinzler, K.W. (2013). Cancer genome landscapes. Science 339, 1546-1558.
Wagenmakers, E.-J., and Farrell, S. (2004). AIC model selection using Akaike weights. Psychonomic Bulletin & Review 11, 192-196.
Wang, D.Z., and Dubois, R.N. (2010). Eicosanoids and cancer. Nature Reviews Cancer 10, 181-193.
Wei, B.Q., Mikkelsen, T.S., McKinney, M.K., Lander, E.S., and Cravatt, B.F. (2006). A second fatty acid amide hydrolase with variable distribution among placental mammals. The Journal of biological chemistry 281, 36569-36578.
Weinberg, R.A. (2014). Coming full circle-from endless complexity to simplicity and back again. Cell 157, 267-271.
Wermuth, C.G. (2003). The practice of medicinal chemistry. (Amsterdam ; London: Academic).
Yan, X., and Su, X. (2009). Linear regression analysis : theory and computing. (Singapore ; Hackensack, NJ: World Scientific).
Yates, L.R., and Campbell, P.J. (2012). Evolution of the cancer genome. Nature reviews. Genetics 13, 795-806.
Zambelli, F., Prazzoli, G.M., Pesole, G., and Pavesi, G. (2012). Cscan: finding common regulators of a set of genes by using a collection of genome-wide ChIP-seq datasets. Nucleic acids research 40, W510-515.

PAPER III

Flux balance analysis predicts essential genes in clear cell renal cell carcinoma metabolism

F. Gatto*, H. Miess*, A. Schulze, J. Nielsen

Scientific Reports ***5****, 10738* (2015)

*Authors contributed equally to this work

SCIENTIFIC REPORTS

OPEN

Flux balance analysis predicts essential genes in clear cell renal cell carcinoma metabolism

Received: 20 January 2015
Accepted: 27 April 2015
Published: 04 June 2015

Francesco Gatto[1,*], Heike Miess[2,*], Almut Schulze[2,3,4] & Jens Nielsen[1]

Flux balance analysis is the only modelling approach that is capable of producing genome-wide predictions of gene essentiality that may aid to unveil metabolic liabilities in cancer. Nevertheless, a systemic validation of gene essentiality predictions by flux balance analysis is currently missing. Here, we critically evaluated the accuracy of flux balance analysis in two cancer types, clear cell renal cell carcinoma (ccRCC) and prostate adenocarcinoma, by comparison with large-scale experiments of gene essentiality *in vitro*. We found that in ccRCC, but not in prostate adenocarcinoma, flux balance analysis could predict essential metabolic genes beyond random expectation. Five of the identified metabolic genes, *AGPAT6*, *GALT*, *GCLC*, *GSS*, and *RRM2B*, were predicted to be dispensable in normal cell metabolism. Hence, targeting these genes may selectively prevent ccRCC growth. Based on our analysis, we discuss the benefits and limitations of flux balance analysis for gene essentiality predictions in cancer metabolism, and its use for exposing metabolic liabilities in ccRCC, whose emergent metabolic network enforces outstanding anabolic requirements for cellular proliferation.

The regulation of metabolism has been recognised to be of central importance in cancer[1–3]. Several studies have collectively suggested that cancer selects for cell clones that have reprogrammed their metabolism, resulting in distinct cancer type-dependent metabolic phenotypes[4–11]. These programs enforce cancer cell dependence on specific flux distributions, and disruption of the underlying pathways mostly results in cell death[12–17].

Under these premises, metabolic modelling using flux balance analysis (FBA)[18] is the only approach that can predict the effect of genetic and environmental perturbations in the disruption of such metabolic phenotypes at the genome scale[19,20], and applications of these models for studying cancer or metabolic diseases have been advocated[21–24]. Contrary to other systems biology approaches, FBA typically involves only limited fundamental assumptions (e.g., mass and charge balance in all reactions, and thermodynamically constrained reaction directionality) and little to no parameter fine-tuning (e.g., non-growth and growth-associated ATP maintenance), yet still allows for meaningful genome-wide predictions of gene essentiality in a variety of model organisms[25,26], provided that a genome-scale metabolic model for the organism is available. Nevertheless, a number of algorithms are now available to infer the active metabolic network in human cells[27–34], and the constraints required to formulate a plausible FBA can now be more readily obtained due the increased availability of high-throughput data. Despite these promising conditions, use of FBA to predict gene essentiality in cancer metabolism is still at its infancy, and besides the extensive theoretical formulations reported in the literature, few practical studies have so far benefited from the systematic analyses enabled by FBA-based studies[35–39].

[1]Department of Biology and Biological Engineering, Chalmers University of Technology, Göteborg 41296, Sweden. [2]Gene Expression Analysis Laboratory, Cancer Research UK London Research Institute, London WC2A 3LY, United Kingdom. [3]Theodor-Boveri-Institute, Biocenter, Am Hubland, 97074 Würzburg, Germany. [4]Comprehensive Cancer Center Mainfranken, Josef-Schneider-Str.6, 97080 Würzburg, Germany. *These authors contributed equally to this work. Correspondence and requests for materials should be addressed to J.N. (email: nielsenj@chalmers.se) or A.S. (email: almut.schulze@uni-wuerzburg.de)

In this study, we critically and systematically assessed the benefits and limitations of FBA for performing genome-scale predictions of gene essentiality in cancer metabolism. In particular, we were interested in the use of FBA to expose metabolic liabilities in clear cell renal cell carcinoma (ccRCC), the most common form of kidney cancer[40]. This cancer type was chosen as because we have recently uncovered that it features a compromised metabolic network[41]. We also verified whether the accuracy of FBA extends to a second cancer type, prostate adenocarcinoma (PC) and analysed the essentiality of selected genes in metabolic models of non-malignant tissues. Our findings suggest that FBA is suitable to uncover essential genes in cancers whose emergent metabolic network enforces outstanding anabolic requirements for cellular proliferation. Hereby we demonstrate that ccRCC depends on the expression of *AGPAT6, GALT*, *GCLC, GSS*, and *RRM2B*, which, although essential for cancer cells, are potentially nonessential in normal cells.

Results

Strategy used to benchmark predictions of gene essentiality in cancer metabolism. Flux balance analysis (FBA) is possibly the only modelling approach that has the potential to predict gene essentiality in cancer metabolism at the genome scale[39]. In this study, we sought to systematically validate whether FBA can be used to determine gene essentiality in cancer cell metabolism by comparing predictions with large-scale experimental datasets (Fig. 1). Therefore, the FBA problem was formulated to scan for a feasible flux distribution that enables the simultaneous biosynthesis of all human biomass components, the so-called biomass equation, in cancers growing in defined serum-containing medium[42]. In these conditions, the metabolic network is free to absorb any medium or serum metabolites (at any rate), which include sugars, amino acids, several metabolic intermediates and short chain fatty acids. In FBA, the emergence of a feasible flux distribution that can support biomass formation is generally limited by the introduction of constraints[43,44] that can represent molecular or environmental limitations (e.g., the absence of a given enzyme in a cancer type or the unavailability of a nutrient in the microenvironment).

Here, we considered two typical sets of constraints: A) the topology of the cancer specific-metabolic network; and B) a profile of experimentally measured fluxes for a number of exchange metabolites (i.e., exchange fluxes) in a panel of cancer-specific cell lines (generally more than one cell line for each type of cancer). Using either of these two constraints we predicted gene essentiality using FBA by introducing a constraint that disables flux in the univocally encoded reaction(s). This constraint is commonly referred to as *in silico* single-gene knockout, and the gene is essential if the *in silico* single-gene knockout ablates biomass production. A gene knockout ablates biomass production if there is no flux distribution that allows the biomass equation to carry a flux, or if the knockout results in a substantial flux reduction. However, a gene knockout consents biomass production if there is no change in the flux through the biomass equation. Single-gene knockout resulting in no change in biomass production is mostly explained due to one of the following reasons: 1) gene redundancy, i.e., more than one gene encodes for the reaction(s) associated with the knockout; 2) pathway redundancy, i.e., there is an alternative pathway with the same overall stoichiometry that can compensate for the knockout; or 3) the reaction(s) encoded by the knocked-out gene are not active (dead end) at the studied condition. Depending on this outcome, a gene is declared essential or nonessential *in silico* for a certain cancer. If constraint B) is implemented, an *in silico* single-gene knockout may ablate or consent biomass production, depending on which profile of exchange fluxes is used as a constraint. In this case, the corresponding gene is declared essential *in silico* for the cancer type only if biomass production is ablated using exchange flux profiles from at least 70% of its corresponding cancer cell lines.

In principle, the proposed approach should capture all metabolic liabilities related to biomass formation induced by the network topology and to the activation of metabolic pathways induced by the exchange flux profile of a certain cancer. At the same time, it is noteworthy that the FBA problem formulated herein will not uncover other metabolic liabilities known to be associated with cancer survival, for example, maintenance of anti-oxidant pools[45]. To evaluate the gene essentiality predictions, we compared these to large-scale experimental data *in vitro*: in this case, a panel of cancer-specific cell lines derived from prostate adenocarcinoma (PC) or clear cell renal cell carcinoma (ccRCC), both cultured in defined serum-containing medium. The cells were transfected with a library of siRNA oligonucleotides that target approximately 230 metabolic genes. In the PC screen, induction of caspase activity was quantified after 96 h following transfection, whereas in the ccRCC screen, reduction in cell number was monitored. If at least 70% of the cancer cell lines passed a given threshold for caspase activity or cell number reduction, then the gene was declared essential *in vitro* for this cancer type (or nonessential *in vitro* if *vice versa*). The accuracy of the predictions was calculated using the Matthews correlation coefficient (MCC) and the related Fisher's exact test statistics.

Accuracy of flux balance analysis for gene essentiality in clear cell renal cell carcinoma metabolism. We decided to assess *in vitro* gene essentiality in the metabolism of ccRCC, as this is the most common form of kidney cancer[40] and it exhibits a strong regulation and dependence on a reprogrammed metabolism following transformation[46–48]. Additionally, we have recently shown that it features a characteristically compromised metabolic network[41]. The reliance on specific metabolic reactions for survival suggests that this cancer may be particularly susceptible to disruptions in the metabolic network. A panel of 5 ccRCC cell lines (786-O, A498, 769-P, RCC4, and UMRC2) was transfected with a custom library of

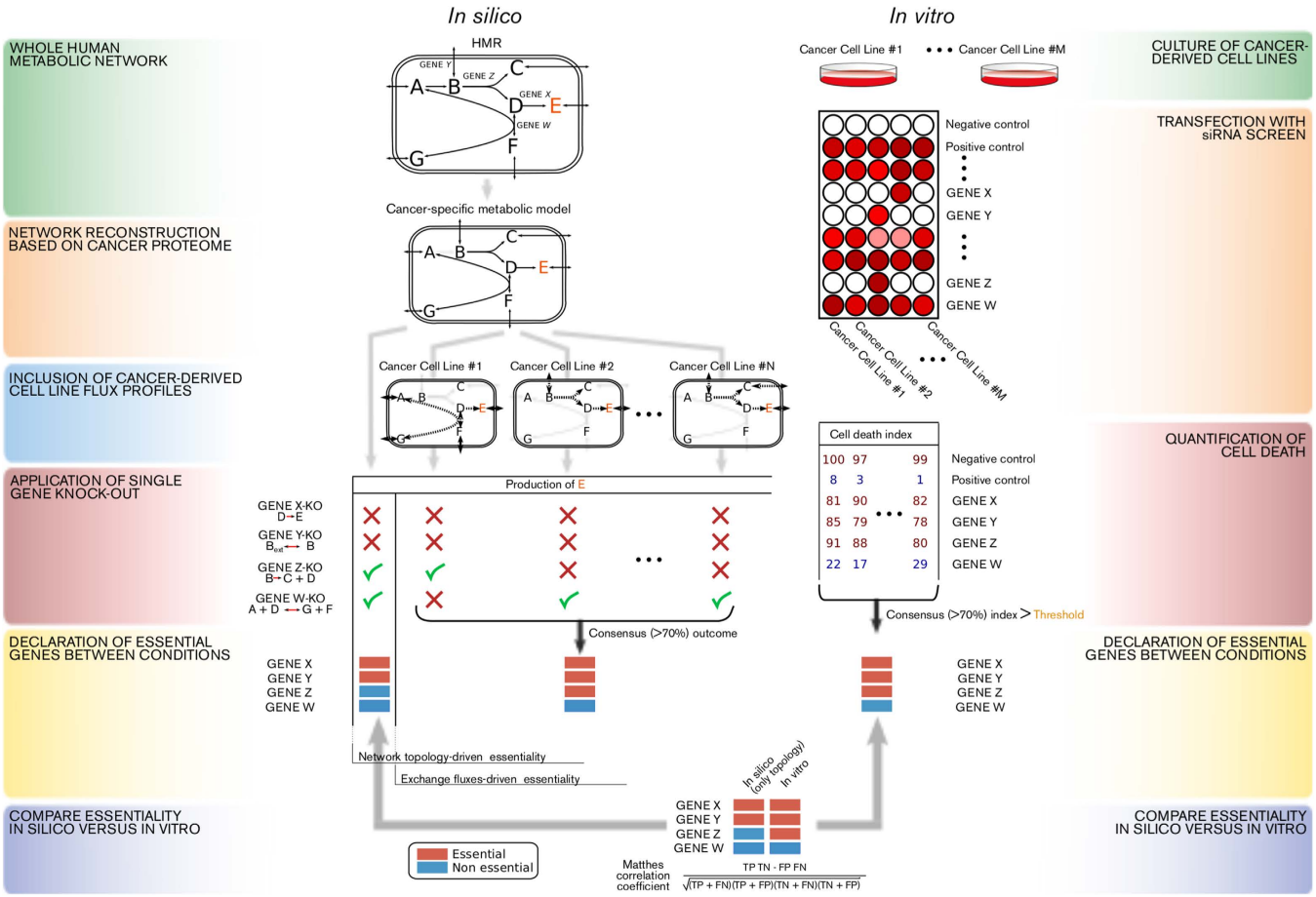


Figure 1. Strategy to measure the accuracy of flux balance analysis predictions of gene essentiality in cancer metabolism . Left part: the Human Metabolic Reaction (HMR) database was used as a generic genome-scale metabolic network to reconstruct a cancer-type specific network based on proteome data obtained from cancer specimen (in the example, the reaction $A \Rightarrow B$ is absent in the cancer-specific model due to lack of the matched enzyme at the protein level for that cancer type). Successively, flux balance analysis is used to simulate whether a flux towards production of biomass (metabolite *E*) was feasible after every single gene-knockout, using as constraints either the topology of the cancer type-specific metabolic network or the measured fluxes for a number of exchange metabolites in different cancer type-derived cell lines. In the latter case a gene is deemed essential if it disables biomass production in $\geq$70% of the cell lines. Grey arrows indicate reactions not occurring in the network. Dashed arrows indicate measured fluxes in a cell line. Right part: cancer-derived cell lines were cultured and transfected with a library of siRNAs that target ~230 metabolic genes and cell number was determined after 4 days. If $\geq$70% of the cell lines passed a given threshold of cell death, the corresponding gene was deemed essential. Bottom: gene essentiality for the ~230 genes targeted by the siRNA library was compared *in silico* vs. *in vitro* and the accuracy of the predictions was calculated by several statistical measures (e.g. the Matthews Correlation Coefficients).

siRNA oligonucleotides targeting 230 different metabolic enzymes, transporters, and regulators involved in central carbon metabolism. For each siRNA, loss of viability was quantified by determining the mean cell number reduction relative to a negative control (non-targeting RISC-free) and a positive control (siRNA targeting ubiquitin B). The number of genes declared essential *in vitro* depends on the threshold chosen for the mean cell number reduction. We selected a 30% reduction for benchmarking purposes because the quantity of essential genes appears to reach a plateau at this value; note that no siRNA caused a cell number reduction greater than 50% (Supplementary Fig. 1). With this threshold, of the 217 tested siRNAs that overlap with the human metabolic network[49], 20 gene knockdowns caused death in at least 70% (4 of 5) of the ccRCC cell lines and were thus deemed essential *in vitro* (Supplementary Fig. 2). In contrast, 136 tested siRNAs did not significantly affect cell number in at least 70% of the ccRCC cell lines and were conversely deemed nonessential *in vitro* (Supplementary Data 1). The remaining 61 genes were not classified, as their knockdowns had mixed effects across cell lines and therefore were not directly attributable to the ccRCC phenotype.

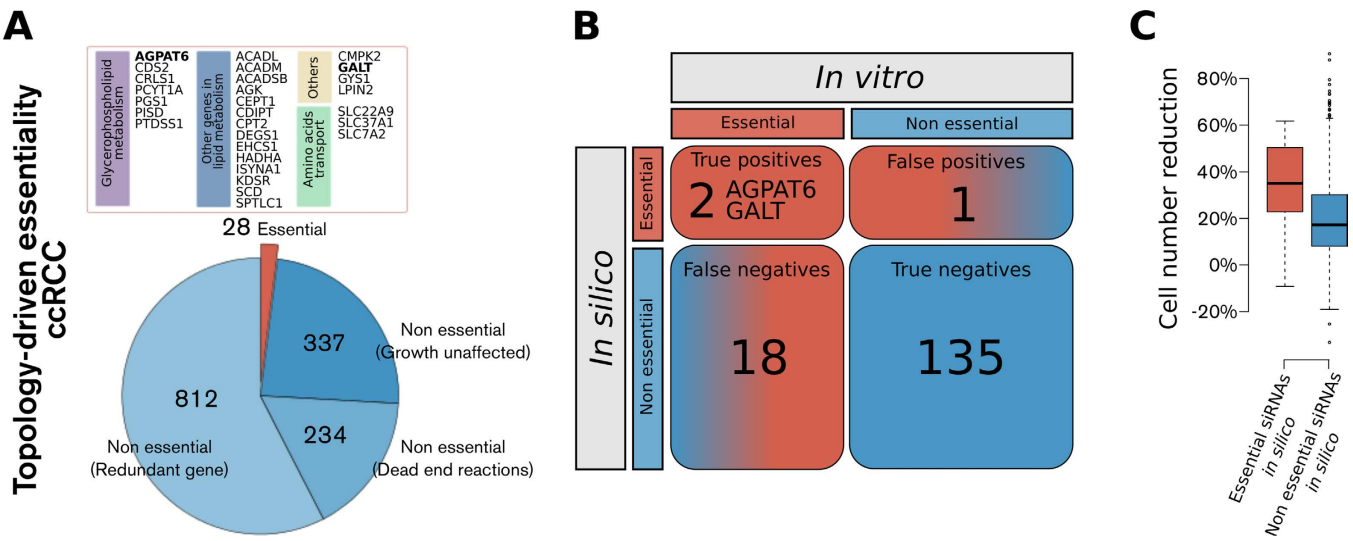


Figure 2. Gene essentiality in ccRCC metabolism as predicted by flux balance analysis using the metabolic network topology as a sole constraint for biomass formation positively compares to a functional RNAi screen targeting ~230 metabolic genes in a panel of ccRCC cell lines . **A**) Gene essentiality in ccRCC according to flux balance analysis using the metabolic network topology as only constraint for biomass formation. **B**) Contingency table for the comparison between the declaration of gene essentiality *in silico* vs. *in vitro* for those siRNAs in the library that had consensus effect in terms of cell number reduction in ≥70% of the cell lines. AGPAT6 and GALT are considered true positives ($p = 0.04$) because their ablation results in cell death *in silico* and *in vitro*. **C**) Boxplots of total cell number reduction for the groups of siRNAs predicted to be either essential (red) or non essential (blue) *in silico*.

Next, we predicted *in silico* gene essentiality using as the sole constraint the topology of the ccRCC metabolic network, as defined by a ccRCC genome-scale metabolic network[41]. We identified 28 essential genes and 1,383 nonessential genes (Fig. 2A). Topology-driven gene essentiality was found to be accurate at a statistically significant level (MCC = 0.226, $p = 0.043$, Fig. 2B). This approach detected two true positives (i.e., candidates essential both *in silico* and *in vitro*), namely *AGPAT6* and *GALT* (Fig. 2A); the expected number of true positives by chance is close to approximately zero ([TP] = 0.174). In this sense, we can assume that *AGPAT6* and *GALT* represent *bona fide* pivotal metabolic nodes in ccRCC, regardless of the exchange fluxes, which suggests that their essentiality is due to a loss of alternative redundant metabolic pathways or genes in ccRCC. Interestingly, siRNAs corresponding to genes predicted to be essential *in silico* result overall in a mean cell number reduction significantly higher than that for siRNAs corresponding to genes predicted not to be essential ($p < 0.001$, Wilcoxon rank-sum test, Fig. 2C).

Next, we also implemented exchange fluxes from a panel of seven ccRCC cell lines (786-O, A498, ACHN, CAK1-1, TK-10, RXF-393, and UO-31) as constraints[50,51]. Using this approach, eighty-seven genes were predicted to be essential in at least 70% (5 of 7) of the cell lines (Fig. 3A). When exchange fluxes were considered, the gene essentiality prediction was found to have an increased accuracy, when compared to the *in vitro* data (MCC = 0.235, $p = 0.010$, Fig. 3B). Additionally, in this case we observed a substantial mean cell number reduction for the group of siRNAs targeting genes predicted to be essential *in silico* compared to those predicted to be nonessential ($p < 0.001$, Fig. 3C). In particular, four additional genes were identified as true positives using this approach, namely *CAD*, *DHCR24*, *FDFT1*, and *ODC1* (Fig. 3A). It is likely that the essentiality of these genes is attributable to common metabolic requirements among ccRCC cell lines (e.g., a high lactate secretion to glucose uptake ratio or secretion of secondary metabolites), which induces dependence on the expression of enzymes that activate the related metabolic pathways. Interestingly, the accuracy of these predictions was not preserved if only exchange fluxes were considered, but the topology of the ccRCC metabolic network was neglected: we observed no significant predictive ability when the generic human metabolic network was used (MCC = 0.086, $p = 0.339$, Supplementary Fig. 3). The results of the accuracy achieved by FBA in these scenarios are reported in Table 1.

Taken together, these results suggest that in ccRCC metabolism, FBA is able to predict gene essentiality, although to a limited degree. Gene essentiality as exposed by FBA is in turn attributable to a rewiring of the metabolic network and exchange fluxes that contribute to biomass production. Conversely, it is conceivable that the 14 genes that were found to be essential *in vitro* but were not captured by FBA are essential because the gene products carry out metabolic tasks that are not ascribable to the biomass production simulated here. Alternatively, it also possible that redundant pathways available in the metabolic

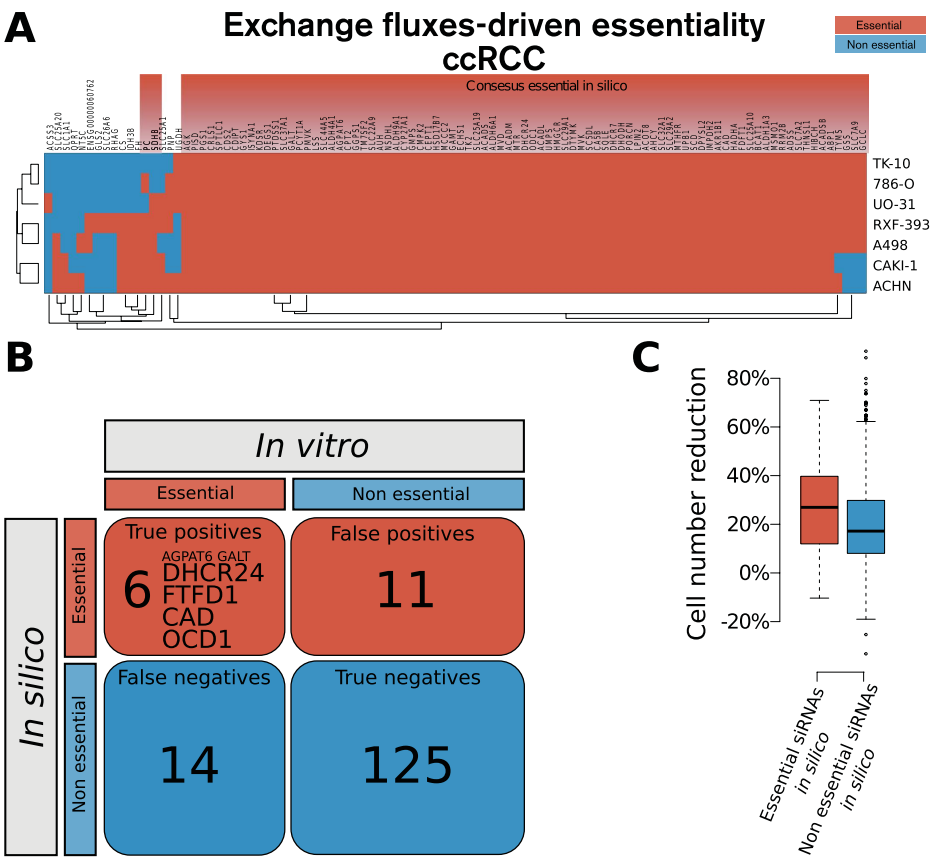


Figure 3. Gene essentiality in ccRCC metabolism as predicted by flux balance analysis using the profile of exchange fluxes from seven ccRCC cell lines in addition to the ccRCC network topology shows increased accuracy when compared with the RNAi screen . **A**) Gene essentiality in ccRCC according to flux balance analysis using the profile of exchange fluxes from seven ccRCC cell lines on top of ccRCC network topology. Each profile of exchange fluxes representing a ccRCC cell line entails a set of genes essential when using that profile. The heatmap features only genes that are essential using at least one flux profile. Finally, genes that are essential using at least 70% of the cell line flux profiles are deemed essential *in silico* in ccRCC. **B**) Contingency table for comparison between the declaration of gene essentiality *in silico* vs. *in vitro* for those siRNAs in the library that had consensus effect in terms of cell number reduction in $\geq$ 70% of the cell lines. Other than AGPAT6 and GALT, DHCR24, FTFD1, CAD, and OCD1 are true positives ($p = 0.007$). **C**) Boxplots of total cell number reduction for the groups of siRNAs predicted to be either essential (red) or non essential (blue) *in silico*.

network are not active due to the presence of regulation *in vitro* or *in vivo* that is not considered in the FBA simulations, as suggested by studies of gene deletion in yeast[52].

Accuracy of flux balance analysis for gene essentiality in prostate adenocarcinoma metabolism. We next sought to define whether the accuracy of FBA predictions is cancer type-dependent. To this end, we used a published dataset that applied the same custom siRNA library in a panel of three prostate adenocarcinoma (PC) cell lines (LNcaP, PC3, DU145)[53]. Cell death was defined by induction of caspase activity, and we declared a gene essential *in vitro* if the corresponding siRNA caused apoptosis with a caspase activity z-score ≥ 2.5 (i.e., number of standard deviations from control) in at least 2 of the 3 cell lines, as adopted in the original study. Using these criteria, 14 metabolic genes were found to be essential in the PC cell lines (Supplementary Fig. 4). The topology of a PC specific metabolic network was reconstructed using the same pipeline followed to generate the previously employed ccRCC genome-scale metabolic model[41] and was used as the sole constraint to perform FBA to predict *in silico* gene essentiality. We identified 37 essential genes, whereas 1,638 genes were classified as nonessential (Supplementary Fig. 5A). We also implemented exchange fluxes from a panel of two PC cell lines, PC3 and DU145, as constraints[50,51], which resulted in the classification of 35 additional genes as essential in both these cell lines (Supplementary Fig. 5B).

Contrary to the results obtained for ccRCC, the accuracy of FBA predictions in PC was considerably lower when using metabolic network topology as the sole constraint (MCC = 0.082, $p = 0.233$,

Cancer type	FBA constraints	Medium	TP	FN	FP	TN	Fisher exact test *p*-value	MCC
Clear cell renal cell carcinoma	Topology	FBS	2	18	1	135	0.043	0.226
		HAM	5	15	12	124	0.046	0.174
	Topology + Exchange fluxes	FBS	6	14	11	125	0.010	0.235
		HAM	6	14	15	121	0.032	0.186
	Exchange fluxes	FBS	1	19	2	134	0.339	0.086
Prostate adenocarcinoma	Topology	FBS	2	12	12	186	0.233	0.082
		HAM	2	12	14	184	0.285	0.068
	Topology + Exchange fluxes	FBS	2	12	19	179	0.635	0.039
		HAM	2	12	27	171	1	0.005

Table 1. Statistical measure of accuracy of the flux balance analysis predictions on gene essentiality compared to *in vitro* results for different set of constraints, media, and cancer types. Key: TP – true positive (essential *in silico* and *in vitroin vitro*); FN – false negative (non essential *in silico*, essential *in vitro*); FP – false positive (essential *in silico*, non essential *in vitro*); TN – true negative (non essential *in silico*, non essential *in vitro*); MCC – Matthews Correlation Coefficient.

Supplementary Fig. 6A), and even worsened with the implementation of exchange fluxes (MCC = 0.039, $p = 0.635$, Supplementary Fig. 6C). However, when only topology was used, we observed a slightly higher mean caspase activity for the group of siRNAs targeting genes predicted to be essential *in silico* ($p = 0.011$, Supplementary Fig. 6B); this did not hold when exchange fluxes were also used as constraints ($p = 0.152$, Supplementary Fig. 6D). The results of the accuracy achieved by FBA in these scenarios are reported in Table 1. Interestingly, most genes predicted to be essential *in silico* participate in the biosynthesis of steroids. In particular, the two true positive genes detected by FBA, *MVD* and *NSDHL*, belong to this pathway. The inefficacy of exchange fluxes to unveil additional liabilities may be due to the low number of flux profiles available as constraints for PC (only 2 cell lines), as opposed to ccRCC (7 cell lines). However, it is also possible that the altered exchange fluxes in PC cells fuel pathways other than those required for biomass production and are therefore not captured by the FBA model used here. One of these pathways could indeed involve the synthesis of cholesterol for the production of steroid hormones, which play a major role in the development of PC[54].

These results suggest that contrary to ccRCC, FBA fails to accomplish acceptable predictions of gene essentiality in PC metabolism. This may reflect the fact that PC cells are more robust in the task of synthesising biomass components. In support of this, Ros and colleagues identified a metabolic liability in PC that does not relate to biomass formation, but is involved in detoxification of reactive oxygen species (ROS)[53]. In addition, ccRCC metabolism could represent an ideal situation for the identification of metabolic liabilities using FBA because of its highly compromised metabolic network.

Effect of the medium metabolites on the accuracy of flux balance analysis predictions. In FBA, the definition of metabolites available for uptake is a decisive constraint for the prediction of gene essentiality[43]. In simulations with microorganisms, the list of metabolites available for uptake mirrors the medium composition used in the controlled experimental setup. Because human cancer cell lines are normally cultured in serum-containing medium, the list of 150 metabolites adopted so far may potentially contain a large number of compounds that can be utilised *in silico* even though they are not utilised in any metabolic reactions by cells *in vitro* (e.g., bilirubin). To explore the extent to which the medium composition affects the accuracy of FBA predictions, we repeated all simulations using Ham's medium, a nutrient poor medium adopted in previous studies to predict *in silico* gene essentiality of cancer cells[33,39]. This less permissive medium decreases the availability of alternative pathways. Thus, the number of essential genes predicted *in silico* increases for both ccRCC (Supplementary Fig. 7) and PC (Supplementary Fig. 8), when only the topology is used as a constraint and when exchange fluxes are also considered. However, these genes were mostly not found to be essential *in vitro*, and therefore the accuracy of the FBA predictions was lower for all four scenarios (Supplementary Fig. 9). We conclude that a broader definition of the medium improves FBA simulations in human systems and reduces the number of false negatives (i.e., genes essential *in silico* but not *in vitro*) induced by incorrect assumptions regarding the unavailability of certain metabolites to the cells. The results of the accuracy achieved by FBA in these scenarios are reported in Table 1.

Effect of the choice for the cell death threshold *in vitro* on the accuracy of flux balance analysis predictions. Given that the definition of gene essentiality *in vitro* depends on the threshold selected for cell death (namely the mean cell number reduction in the ccRCC screen and the caspase activity *z*-score in the PC screen), we performed a sensitivity analysis on these thresholds for all tested scenarios (implemented constraints, cancer types, and medium definition). In the case of ccRCC, we observed a positive relationship between the accuracy of FBA predictions and the strictness of the definition of the threshold for cell death, at least up to the point where the number of essential genes *in vitro* is less than 10, which occurs for mean cell number reduction >40% (Supplementary Fig. 10). This trend was conserved in all scenarios, with the highest accuracy being achieved when using the topology of the ccRCC metabolic network in a serum-containing medium as the sole constraint to perform FBA; the lowest accuracy was observed when constraining the exchange fluxes in Ham's medium. In the case of PC, the above trend was not observed for any scenarios (Supplementary Fig. 11). In particular, the accuracy of predictions was not noticeably different from a random predictor. Taken together, this indicates that the accuracy of *in silico* predictions is significant in ccRCC (but not in PC) for a reasonable range of thresholds upon which a gene is declared essential *in vitro*.

Effect of cancer cell line exchange fluxes on the inference of gene essentiality in a certain cancer. Because FBA proved powerful in exposing the metabolic liabilities of ccRCC, we decided to validate some of the *in silico* predictions of gene essentiality. In particular, we tested the extent to which exchange fluxes from ccRCC cell lines can be used to infer gene essentiality attributable to the ccRCC phenotype. To this end, we selected some genes that were differentially classified as essential depending on the cell line flux profile, but still classified as essential in ccRCC according to a consensus outcome, i.e., essential in >70% of cell lines. We chose to test the predictions for *GCLC*, *GSS*, *SLC7A9* (considered essential *in silico* for ccRCC because they were classified as such in 5 of the 7 cell lines), and *PNP* (considered nonessential *in silico* for ccRCC because it was classified as such in 3 of the 7 cell lines). In addition, we tested *UMPS* and *RRM2B*, which were deemed essential *in silico* for all ccRCC cell lines upon implementation of every exchange flux profile. Next, the corresponding genes were silenced in five of the seven cell lines whose exchange fluxes were used to constrain the FBA predictions (786-O, A498, CAKI-1, TK10, and UO31). In accordance with the threshold for cell death adopted above, a gene was declared essential *in vitro* for ccRCC if more than 70% of cell lines tested (e.g., at least 4 of 5) exhibited at least 30% mean cell number reduction compared to control (Fig. 4).

At the level of gene essentiality in ccRCC, the consensus predictions for *RRM2B, GCLC*, *UMPS*, and *GSS* were confirmed *in vitro.* However, *PNP* and *SLC7A9* knockouts showed mixed effects across cell lines *in vitro*. Hence, the essentiality of these genes in ccRCC could not be inferred from this experiment. Overall, this result suggests that ccRCC cell line exchange fluxes can entail some common metabolic requirements associated with the ccRCC phenotype, which can be exploited to predict gene essentiality in ccRCC metabolism. However, the exchange flux measurements appear to be insufficient *per se* to achieve reliable predictions for a specific cell line. Indeed, we observe that only 17 of the 30 individual predictions were replicated *in vitro* if the cell-line-specific exchange fluxes were used for the prediction of essentiality for the corresponding cell line.

Characterisation of gene essentiality in ccRCC metabolism. FBA exposed some metabolic liabilities in ccRCC that are unlikely to have been predicted by chance. Therefore, we sought to characterise those genes that were classified in this study as essential *in silico* and validated *in vitro.* This list includes ten metabolic genes: *AGPAT6, CAD*, *DHCR24*, *FDFT1*, *GALT, GCLC, GSS, ODC1, RRM2B,* and *UMPS.* First, we predicted whether these gene knockouts would be toxic for the execution of essential metabolic functions, i.e., whether the *in silico* gene knockouts compromise the metabolism of normal cell types. As previously described[33], we simulated the essentiality of these genes in 83 normal cell types by checking whether 56 primary metabolic tasks (e.g., synthesis of cholesterol or oxidative phosphorylation) could be carried out *in silico* upon application of the corresponding *in silico* gene knockout. In all normal cell types, the simulation revealed that knockout of *CAD* or *UMPS* ablates the *de novo* biosynthesis of pyrimidines, while *FDFT1* and *DHCR24* knockouts impede the production of cholesterol in normal human cell types (Fig. 5A). However, the remaining 6 genes had only minor toxic effects (in < 50% of cell types), and can thus be regarded as nontoxic to normal cells.

Next, we specifically checked the toxicity of these gene knockouts in tubular kidney cells, where ccRCC is thought to originate from[55]. In this case, the *in silico* knockout of *ODC1* was found to be toxic because it impaired seven essential metabolic tasks in normal kidney cells. On the contrary, *AGPAT6, GALT, GCLC, GSS*, and *RRM2B* knockouts did not compromise any metabolic task and can thus be considered as selectively essential in ccRCC (Fig. 5B). To test the quality of these predictions, we ablated *GCLC, GSS, RRM2B*, and *UMPS* in an immortalised, non-tumourigenic kidney epithelial cell line (HK-2) using RNAi. These four genes were not part of the siRNA screening library but were predicted by FBA to be essential both *in silico* and *in vitro*. In accordance with the *in silico* predictions of toxicity, we observed cell death when *UMPS* was knocked out in HK-2 cells, while *GCLC, GSS*, and *RRM2B* knockouts caused a minor cell number reduction, above the adopted threshold for cell death (Fig. 5C).

Subsequently, we attempted to elucidate the putative mechanisms at the flux level underlying the essentiality of the *AGPAT6, GALT, GCLC, GSS*, and *RRM2B* genes, which were predicted to be toxic to

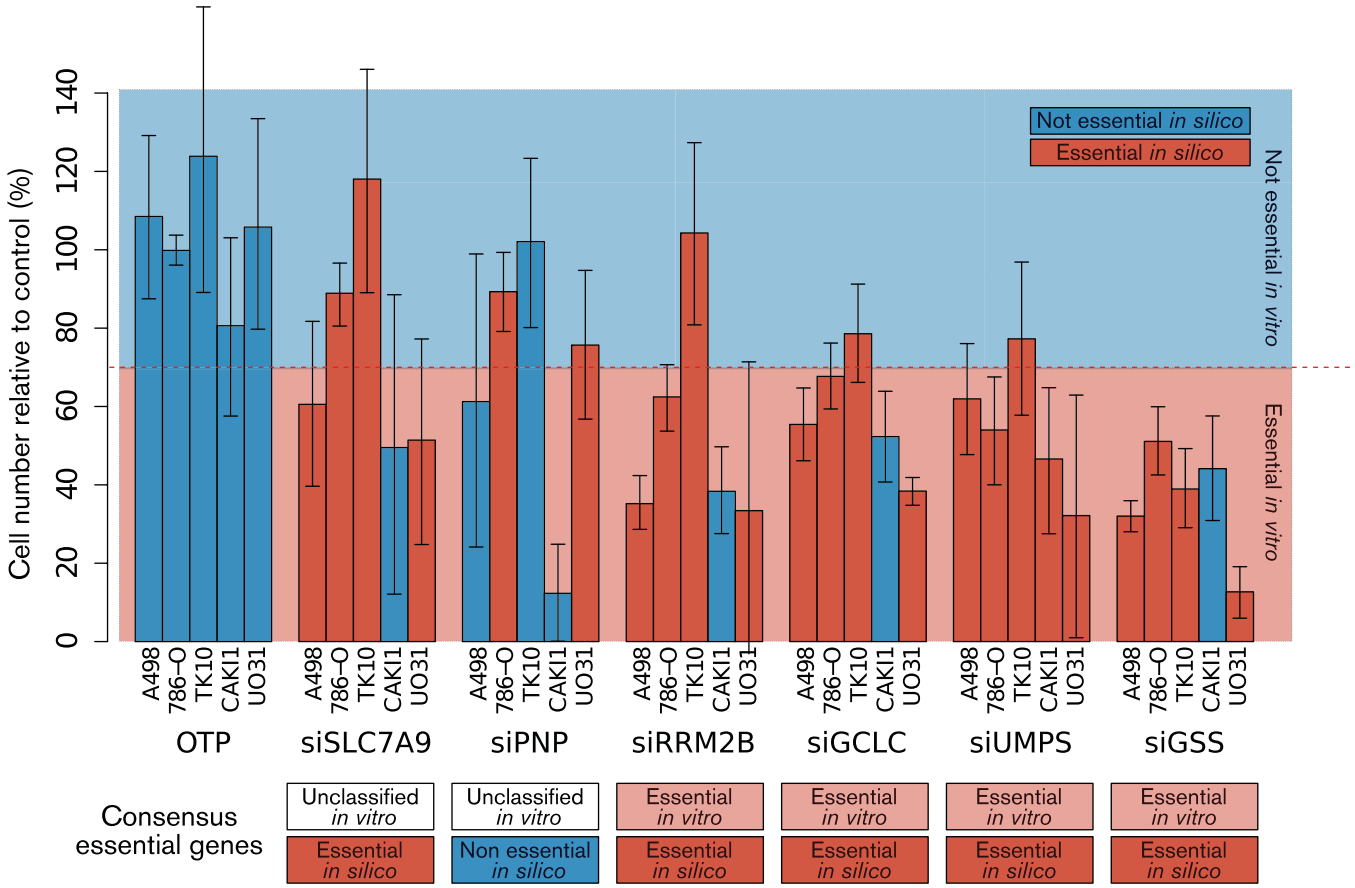


Figure 4. Validation of predicted gene essentiality in ccRCC . Five ccRCC cell lines that match the flux profile constraints implemented to predict gene essentiality in ccRCC were transfected with siRNA targeting SLC7A9, PNP, RRM2B, GCLC, UMPS and GSS. A non-targeting oligonucleotide, OTP (scrambled siRNA), was used as negative control. . Each bar represents the mean cell number reduction relative to control together with the 95% highest density interval of two experiments performed in triplicate. The consensus outcome across cell lines in terms of gene essentiality is shown below each set of bars corresponding to a certain silenced gene. Genes that, if silenced, cause a $\geq$ 30% reduction in cell number relative to the non-targeting RISC-free siRNA in $\geq$ 70% of ccRCC cell lines are deemed essential *in vitro*. The consensus outcome for each silenced gene is compared to the prediction of essentiality *in silico* for the corresponding gene in ccRCC (compare with Fig. 3A).

none or only to a few of the normal human cell types, and in particular were predicted to be nontoxic to tubular kidney cells. *AGPAT6* and *GALT* were found to be essential when using the topology of the ccRCC metabolic network as the sole constraint for FBA, which is indicative of a loss of pathway redundancy in key steps involved in biomass synthesis. In the human metabolic network, the *AGPAT6*-encoded reaction, i.e., the conversion of glycerol-3-phosphate to 1-acyl-glycerol-3-phosphate, is associated with additional isoenzymes, *AGPAT9, GPAT2,* and *GPAM*. However, according to the Human Protein Atlas, which supported the reconstruction of the ccRCC metabolic network, *AGPAT6* is the only member of the family of lysophosphatidic acid acyltransferase genes appreciably expressed in ccRCC[56]. Therefore, when *AGPAT6* is knocked out, the production of glycerolipids, which is required for biomass production, becomes unfeasible (Fig. 6A), making *AGPAT6* an essential gene in ccRCC.

Regarding GALT, this enzyme is pivotal in the ccRCC metabolic network because it catalyses the second step of the Leloir pathway of galactose metabolism (conversion of UDP-galactose to UDP-glucose). Examination of the flux space in ccRCC revealed that this reaction fuels the production of UDP-glucose, which is needed for the biosynthesis of glycogen. Knockout of *GALT* thus results in growth ablation due to the inability to produce glycogen, here considered to be an essential biomass component. This pathway can be bypassed via UGP2, which condenses glucose-1-phosphate with UTP to yield UDP-glucose, but *UGP2* is not expressed in ccRCC (Fig. 6B). The essentiality of GALT in ccRCC is determined by the inactivity of this parallel pathway; this represents an example of loss of redundancy within the topology of the metabolic network.

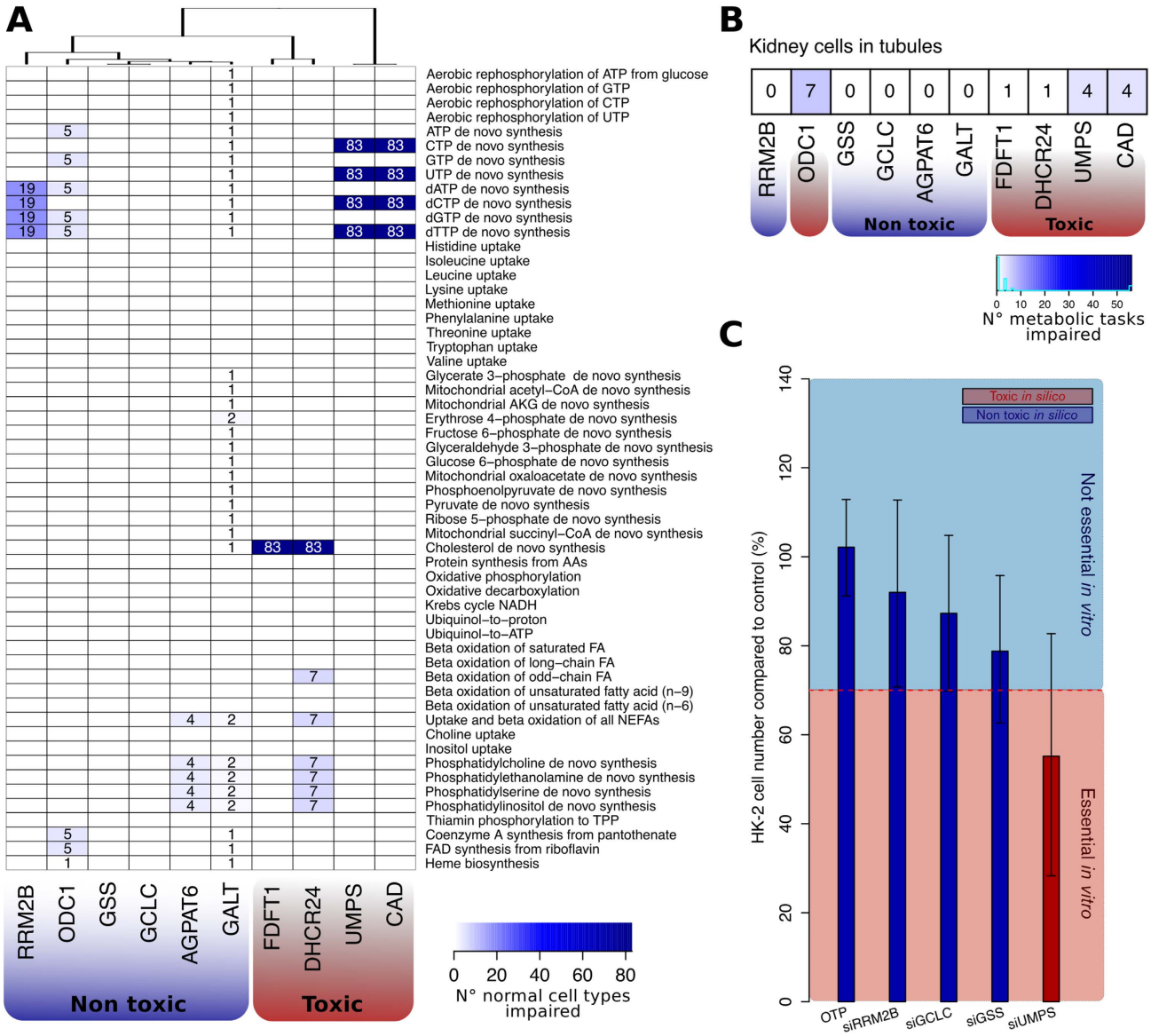


Figure 5. In silico toxicity for the knockouts in normal cells of genes essential in ccRCC . **A**) Number of normal cell types where a certain metabolic task is impaired upon *in silico* knockout. For each of the ten genes found essential in ccRCC according to this study (*RRM2B, ODC1, GSS, GCLC, AGPAT6, GALT, FDFT1, DHCR24, UMPS,* and *CAD*; columns), it was tested if the corresponding *in silico* gene knockout affects the feasibility of 56 different metabolic tasks (rows) in 83 genome-scale metabolic models representing normal, non-tumourigenic cell types. The numbers within the heatmap indicate how many of normal cell types (out of the 83) showed a certain metabolic task that was no more feasible upon the knockout (white cells indicate that none of the cell lines showed an effect). The knockout of *AGPAT6, GALT, GCLC, GSS, ODC1*, and *RRM2B* did not impair more than 50% of normal cell types and are hence considered non-toxic to normal cells. On the contrary, *FDFT1, DHCR24, UMPS* and *CAD* knockouts affected some essential metabolic tasks in all normal cell types and are thereby considered toxic to normal cells. **B**) Number of metabolic tasks impaired upon *in silico* knockouts in a kidney cell in tubule model. In addition to *FDFT1, DHCR24, UMPS* and *CAD, OCD1* knockout is predicted to be toxic because it disables seven metabolic tasks. On the other hand, knockouts of *AGPAT6, GALT, GCLC, GSS,* and *RRM2B* are predicted to be non-toxic and these genes are thus considered selectively essential in ccRCC. **C**) Validation of toxicity for *GCLC, GSS, RRM2B* and *UMPS* knockouts in a normal kidney epithelial cell line, HK-2. Cells were transfected with siRNA targeting RRM2B, GCLC, UMPS and GSS and a non-targeting scrambled siRNA, OTP, was used as negative control. Each bar represents the mean cell number reduction relative to control together with the 95% highest density interval of two experiments performed in triplicate. In line with the predictions, *UMPS* knockout caused a substantial cell number reduction in HK-2 cells compared to knockouts of *GCLC, GSS,* and *RRM2B*, thereby indicating a substantially superior toxicity in normal kidney epithelial cells.

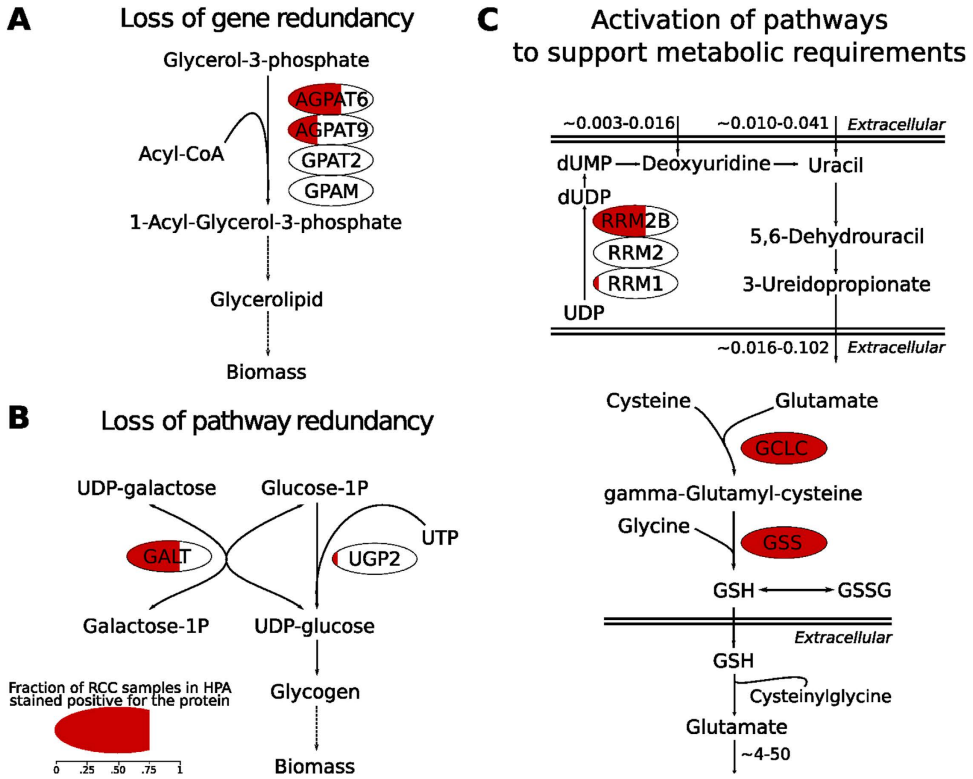


Figure 6. In silico elucidation of the mechanisms of essentiality for the five genes selectively essential in ccRCC . **A**) *AGPAT6* is essential only in ccRCC because of loss of gene redundancy. In ccRCC, the repression of *AGPAT9*, *GPAT2*, and *GPAM* in glycerolipid metabolism renders the pathway solely dependent on AGPAT6 to produce essential lipids for biomass. **B**) *GALT* is selectively essential because of loss of pathway redundancy in ccRCC. Low or no expression of UGP2 forces the flux through GALT to produce glycogen in ccRCC. **C**) *RRM2B, GCLC* and *GSS* are essential only in ccRCC because of specific metabolic requirements of ccRCC cells that activate the corresponding pathway (flux rates are shown in fmol cell^{-1} h^{-1}). Top: the measured secretion rate of 3-ureidopropionate in ccRCC cell lines is not matched by the observed uptake rate of its direct precursors, uracil and deoxyuridine. This forces a flux active in the catabolism of UDP (part of the pyrimidine degradation pathway) to compensate for the observed 3-ureidopropionate secretion rate. One of the pathway steps is uniquely catalyzed by RRM2B, given that the other genes associated to this reaction (RRM1 and RRM2) are not expressed in ccRCC. Bottom: ccRCC cell lines secrete glutamate at a high rate and the only flux distribution that fits glutamate secretion in the ccRCC metabolic network requires the cleavage of extracellular glutathione (GSH). Extracellular GSH is in turn derived from *de novo* GSH intracellular synthesis that is catalyzed by GSS and GCLC. Noteworthy, the reduction of reactive oxygen species like H_2O_2 by GSH is a metabolic function preserved in the predicted flux distribution. For each protein, the red shading represents the fraction of ccRCC samples in which the protein is expressed according to the Human Protein Atlas.

The three additional genes, *GCLC, GSS*, and *RRM2B*, were classified as essential only when constraints on exchange fluxes were implemented. Their essentiality is likely to be due to a loss of redundancy in the ccRCC network when metabolite fluxes are constrained by measured uptake and secretion rates. To explore how the fluxes were distributed before the implementation of the exchange fluxes, we relieved each of these constraints one at the time until biomass production was restored, thereby allowing us to associate gene essentiality with a particular exchange flux. We found that *RRM2B* is associated with the flux of 3-ureidopropionate, a product of the uracil degradation pathway, which is secreted by all ccRCC cell lines (range: 0.016 to 0.102 fmol cell^{-1} h^{-1}) (Fig. 6C, left). This secretion rate is not matched by the uptake rate of either of its two precursors, uracil and deoxyuridine (range: 0.003 to 0.016 and 0.010 to 0.041 fmol cell^{-1} h^{-1}, respectively). Thus, it is necessary for cells to activate a flux to degrade UDP to sustain the given 3-ureidopropionate secretion rate, and one of these steps is catalysed by RRM2B. According to the Human Protein Atlas, the two other genes associated with this step (namely *RRM1* and *RRM2*) are not expressed in ccRCC, and thus they cannot compensate for this flux if *RRM2B* is knocked out, making *RRM2B* essential.

In the case of *GCLC* and *GSS*, the essentiality is associated with the secretion of glutamate, which occurs at remarkably high rates (approximately 4 to 50 fmol cell^{-1} h^{-1}) in ccRCC cell lines. The analysis

of the ccRCC flux space unveiled that elevated rates of extracellular glutamate accumulation derive from the catabolism of extracellular glutathione (GSH) carried out by different gamma-glutamyl transferases. Indeed, the reconstructed ccRCC metabolic network does not include alternative pathways that support the secretion of glutamate, such as the x_C^- system. Despite the evidence that the x_C^- system, an antiporter responsible for cystine uptake via a 1:1 exchange with glutamate, is expressed in the kidney[57], no evidence for the encoding gene, *SLC7A11*, is reported at the protein level by the Human Protein Atlas; it was therefore not included in the reconstructed network. In the absence of alternative pathways, the only flux distribution returned by FBA that fits glutamate secretion requires the cleavage of extracellular GSH. This is in turn dependent on the secretion of *de novo* synthesised intracellular GSH, which is catalysed by GSS and GCLC (Fig. 6D, right). The genes are therefore classified as essential to support this flux distribution. At the same time, GSH is also utilised to reduce peroxides/reactive oxygen species (ROS). In this process, GSH is oxidised and dimerises with another moiety to form GSSG, which can be catalytically recycled to GSH. Therefore, this flux distribution also includes the expected role of GSH in ROS detoxification. Overall, FBA was able to predict a model that associates the essentiality of *GSS* and *GCLC* to the observed secretion of glutamate. Nevertheless, we acknowledge that model incompleteness (attributable to a lack of functional gene annotation in the metabolic network, as in the case of *SLC7A11* in ccRCC) may be a factor that affects the reliability of this prediction, as recognised in earlier works on FBA predicted gene essentiality[58,59].

Discussion

In the last decade, increasing evidence supports the notion that cancer cells reprogram their metabolism and are therefore susceptible to disruption of the metabolic network[3]. Despite the promise that flux balance analysis (FBA) enables prediction of gene essentiality in cancer metabolism at the genome scale[21], we have observed a scarcity of methodical studies that assess these potential benefits critically. Considering the widespread use of FBA in the systems biology community[60], we applied fundamental principles of FBA to measure the accuracy of the predictions against large-scale gene essentiality experiments performed *in vitro*. We evaluated the efficacy of this method for two cancer types, clear cell renal cell carcinoma (ccRCC) and prostate adenocarcinoma (PC), for a variety of parameters: types of FBA constraints used, complexity of the *in silico* medium (i.e., the spectrum of metabolites available for uptake), and the numerical threshold for cell death applied to the *in vitro* experiments. A summary of the accuracy for all tested scenarios is given in Table 1.

Our findings suggest that FBA is sufficiently accurate to expose metabolic liabilities in ccRCC. The highest accuracy was achieved by FBA using a ccRCC-specific metabolic network that was further constrained by exchange fluxes determined experimentally across a panel of seven ccRCC cell lines and using an *in silico* serum-containing medium, which allows the flexible uptake of 150 different metabolites (Matthews Correlation Coefficient = 0.235). The accuracy improved with stricter definitions of the threshold for cell death *in vitro,* whereas it worsened when a more restricted medium metabolite composition, as found in Ham's medium, was applied *in silico.* This medium composition is not representative of the actual culture conditions used for the *in vitro* experiments, but had previously been used in similar studies[39,61]. However, FBA was found not to be predictive for gene essentialities in PC, where the accuracy was not better than a random predictor (MCC = 0.039, for the same scenario described above).

In general, poor predictions may be ascribed to different assumptions beyond FBA. First, we considered gene essentialities in metabolism based on the ability to carry flux towards biomass formation. Although this is an undisputable requirement for cancer cell proliferation, there is evidence that survival of cancer cells also depends on other metabolic functions, most notably NADPH production and anti-oxidant synthesis[45,62]. Second, we also classify a gene as essential if the *in silico* knockout cannot satisfy certain cancer type-specific metabolic requirements, here represented by the profiles of exchange fluxes in several cancer type-specific cell lines (one such requirement could be lactate secretion in a specified range of rates). The number of profiles available for a cancer type may affect the specificity of certain metabolic requirements (for PC, only two profiles of exchange fluxes were available), therefore biasing the quality of the *in silico* predictions when these are enforced as constraints in FBA. Third, we disregarded enzyme complexes because current human genome-scale metabolic models do not report such annotation systematically[63,64]. Hence, if more than one enzyme is associated to a reaction in the model, all genes encoding for these enzymes are automatically excluded from the *in silico* single-gene knockout and classified as redundant. Furthermore, genome-scale metabolic models such as those used for this study represent the best models to our knowledge in terms of metabolic reactions occurring in a cell, but model incompleteness is known to affect *in silico* predictions of essential metabolic genes[59]. Finally, the evaluation of the accuracy of the *in silico* prediction is also affected by the accuracy of the *in vitro* experiments. Indeed, metabolic screens using siRNA libraries may produce false negatives due to insufficient silencing and are liable to significant off-target effects that can both positively and negatively affect the viability of transfected cells.

As part of this evaluation, FBA unveiled the inherent fragility of ccRCC metabolic processes that contribute to biomass growth or support certain metabolic requirements. A number of recent studies have evidenced the centrality of metabolism in ccRCC[41,46,47], and these findings further support the notion that ccRCC is dependent on specific metabolic genes to sustain proliferation. At the same time, our work leverages on a metabolic model for ccRCC and cultured ccRCC cell lines to assess vulnerabilities for this

disease. Hence, this approach cannot likely capture the genomic diversity and complexity of ccRCC[65,66]. However, it should prove useful to expose metabolic liabilities that are transversal to the ccRCC phenotype. In this context, we describe five genes, *AGPAT6*, *GALT, GCLC, GSS*, and *RRM2B,* which are essential to ccRCC but are potentially dispensable in normal cell types. In addition, FBA can also be used to explore the mechanisms that render a gene essential *in silico*.

One of the mechanisms by which this essentiality arises is loss of gene redundancy. AGPAT6 is the only expressed enzyme isoform that can commit glycerol-3-phosphate into glycerolipid biosynthesis. Thus, glycerolipid synthesis is clearly a sensitive pathway in ccRCC, perhaps exacerbated by the lack of expression of enzymes within alternative routes due to a loss of heterozygosity in the corresponding gene loci, as recently suggested[41]. Interestingly, some members of the AGPAT family may have a causal role in cancer development[67], and further work is required to elucidate the roles of these genes in ccRCC. A second possible mechanism is loss of pathway redundancy, resulting in enhanced dependence on the remaining reactions. This mechanism of gene essentiality revealed by FBA in ccRCC is exemplified by GALT, a component of the Leloir pathway. Downregulation of the enzymes of an alternative pathway in ccRCC induced dependence on *GALT* expression. In accordance with the findings of our simulations, there is evidence that *UGP2* and *GALT* homologs provide redundancy for this pathway in yeast[68]. Moreover, it is has been shown that *GALT*-deficient mice can sustain glycogen synthesis through the pathway branch catalysed by the murine homolog of *UGP2*[69], thus supporting the predicted mechanism that confers essentiality to *GALT* in ccRCC. The pivotal role of GALT in glycogen biosynthesis may in addition underscore the typical phenotype of ccRCC cells, which are characterised by high levels of glycogen accumulation. Finally, FBA using flux rates as additional constraints allowed us to identify gene essentialities associated with specific exchange fluxes in ccRCC cell lines. The essentiality of *GCLC* and *GSS* was linked to the secretion of glutamate, whereas the essentiality of *RRM2B*, an enzyme in deoxyribonucleotide metabolism, was linked to the secretion of 3-ureidopropionate. *GCLC* and *GSS* play a fundamental role in the intracellular detoxification of ROS by catalysing two successive steps in the biosynthesis of glutathione[70]. This biological process plays a prominent role in carcinogenesis[71] and has been postulated to be of central importance in the rewiring of cancer metabolism[45]. Therefore, it is likely that the essentiality of *GCLC* and *GSS* in the *in vitro* experiments stems from their functions in the control of intracellular ROS levels.

Here, we find that *de novo* synthesis of GSH is also associated with glutamate secretion in the absence of other systems that can fulfil this function in ccRCC. Although this may be an artefact due to model incompleteness, we show that GCLC-GSS can sustain a flux distribution in which extracellular GSH is catabolised into cysteinylglycine and glutamate, therefore explaining the observed glutamate secretion in ccRCC cell lines. In a similar fashion, the consistent secretion of 3-ureidopropionate observed in ccRCC cell lines combined with an unmatched uptake rate of its direct precursors implies that *RRM2B* is active in supporting uridine-derived 3-ureidopropionate. *RRM2B* exerts its function in deoxyribonucleoside biosynthesis and in DNA damage repair, and in this role it appears to hinder cancer progression[72–75]. Nevertheless, *RRM2B* function in ccRCC may be different given the lack of expression of *RRM1* and *RRM2* for supporting nucleotide biosynthesis. Indeed, this pathway was found to be compromised uniquely in ccRCC compared to four other cancer types[41].

In conclusion, in this study we show the strength and limitations of FBA for the prediction of gene essentiality at a genome scale in cancer metabolism. In addition, we report five metabolic genes selectively essential in a particular cancer type, i.e., ccRCC. Importantly, FBA can be used to identify potential mechanisms by which these gene essentialities arise and thereby provide testable hypotheses. We argue that accounting for metabolic liabilities other than biomass generation and the integration of additional layers of high-throughput data may lead to an even more complete description of the essentiality landscape in cancer metabolism.

Materials and Methods

Cell culture and reagents. 786-O, 769-P, A498, CAKI-1, RCC4, TK10, UMRC2, and UO31 clear cell renal cell carcinoma (ccRCC) cell lines were maintained (and transfected) in DMEM supplemented with 4.5 g/l D-Glucose, 0.11 g/l Sodium Pyruvate (Gibco), 4 mM Glutamine (CRUK, Clare Hall, Cell Services), 100 Units/ml Penicillin / 100 ug/ml Streptomycin (Gibco) and 10% Fetal Bovine Serum (Gibco). For the RNAi screen, RCC4, UMRC2, A498, 786-O and 769-P cells were transfected in triplicates in a 96-well format with 37.5 nM siRNA SMARTpools (Dharmacon siGENOME) targeting the genes of interest using Dharmafect2 as transfection reagent. For the validation experiment for *GCLC, GSS, PNP, RRM2B, SLC7A9,* and *UMPS*, 786-O, A498, CAKI-1, TK10, and UO31 cells were transfected in two independent experiments, triplicates each in a 96-well format with 37.5 nM siRNA SMARTpools (Dharmacon siGENOME) targeting the genes of interest using Dharmafect2 as transfection reagent. For the validation experiment for *GCLC, GSS, RRM2B,* and *UMPS*, HK-2 cells were transfected in two independent experiments, triplicates each in a 96-well format with 37.5 nM siRNA SMARTpools (Dharmacon siGENOME) targeting the genes of interest using Dharmafect2 as transfection reagent. In all cases, after 96 h (with a media top up after 24 h), cells were fixed in 80% ethanol over night and subsequently stained with DAPI (Sigma). Cell number was determined using an ACUMEN ^{e}X3 laser-scanning fluorescent micro plate cytometer. For the purpose of data normalization, the non-targeting RISC-free transfection was used

as negative control, while ubiquitin B (*UBB*) and polo-like kinase 1 (PLK1) served as positive killing control.

Quantification of cell death. In the case of ccRCC, cell death was quantified in terms of reduction of cell number upon siRNA transfection in cells scaled to the effect of the negative and positive controls. For each replicate, 9 positive killing controls (*UBB*) and 12 negative controls (RISC-free) were transfected. In a given cell line c for a given replicate r, the cell number reduction caused by siRNA s was linearly interpolated as in equation (1):

$$CellNumberReduction_{s,c,r} = \frac{\text{median}\left(N^{\circ}cells_{RISCfree}\right)_{c,r} - N^{\circ}cells_{s,c,r}}{\text{median}\left(N^{\circ}cells_{RISCfree}\right)_{c,r} - \text{median}\left(N^{\circ}cells_{UBB}\right)_{c,r}} \quad (1)$$

Then for each cell line, the mean cell number reduction is computed as the average across replicates. We declare a gene essential *in vitro* in ccRCC if the mean cell number reduction upon transfection of the corresponding siRNA for at least 70% of the tested cell lines is above 30%. For all these genes, we verified that the associated mean cell number reduction is statistically significantly greater than 0 (one sided t-test, $p < 0.05$). This data is collected in Supplementary Data 1. In the case of PC, processed data containing caspase activity z-score was retrieved from the study[53]. We declare a gene essential *in vitro* in PC if the caspase activity z-score upon transfection of the corresponding siRNA for at least 2 of the 3 tested cell lines is above 2.5, as adopted in the original study.

Statistical tests. The Fisher's Exact Test was carried out using the total number of tested siRNA in the library that could be compared to the *in silico* single gene-knockout test as the universe and it was performed in R. The 95% highest density interval (HDI) for cell number variation relative to RISC-free control was calculated by Bayesian estimation under the following assumptions: data are sampled from a t-distribution of unknown and to be estimated normality (i.e. degrees of freedom); high uncertainty on the prior distributions; the marginal distribution is well approximated by a Markov chain Monte Carlo sampling with no thinning and chain length equal to 100'000. The estimation was performed using the BEST R-package[76] (the above assumptions are reflected by the default parameters).

Flux balance analysis. The ccRCC genome-scale metabolic model (iRenalCancer1410) was downloaded at www.metabolicatlas.org. The PC genome-scale metabolic model (iProstateCancer1675) was reconstructed using the same pipeline as for iRenalCancer1410[41]. The models are inherently mass and charge balanced, and reaction directionalities to reflect thermodynamic constraints were not modified. Apart from changing a misannotated reaction incorrectly associated with *OGDH* to its correct associated gene, *ODC1,* no other modifications to the models were operated. The lists of metabolites available for uptake or secretion in the serum-containing medium (FBS) or in Ham's medium are given in Supplementary Data 2. For FBS, this list was compiled by merging the list of metabolites exchanged in cell line cultures growing in fetal bovine serum medium according to Jain and coworkers[51] and other compounds known to be present in this medium[42]. In the case of ccRCC, 94 metabolites could be matched with a preexisting uptake reaction in the network, 6 metabolites are present in the extracellular compartment but did not have a preexisting uptake reaction, and 51 metabolites are only present in the cytosolic compartment. We added an exchange reaction for each of the latter 57 metabolites, modeled as energy-free diffusion (i.e. => metA[s/c]). In the case of PC, 92 metabolites could be matched with a preexisting uptake reaction in the network, 7 metabolites are present in the extracellular compartment but did not have a preexisting uptake reaction, and 46 metabolites are only present in the cytosolic compartment. We added an exchange reaction for each of the latter 53 metabolites, modeled as energy-free diffusion (i.e. => metA[s/c]). For Ham's medium, the list was retrieved from[33]. In both the case of ccRCC and PC, 38 metabolites could be matched with a preexisting uptake reaction in the network, 1 metabolite is present in the extracellular compartment but did not have a preexisting uptake reaction, and 4 metabolites are only present in the cytosolic compartment. We added an exchange reaction for each of the latter 5 metabolites, modeled as energy-free diffusion (i.e. => metA[s/c]). Unless cell line specific exchange fluxes were used, all the above exchange reaction can span any real value from −1000 to +1000, while any other exchange reaction was bounded to 0 for uptake. It should be noticed that this is a critical step for the simulations which follow: the possibility to freely exchange these metabolites allows to take in account all the extremely different metabolic states that a cell may adopt in response to the availability of these metabolites (e.g. lactate may be either secreted as a by-product of glycolysis or absorbed and catabolized as a carbon source). For both ccRCC and PC, the used biomass equation is the built-in reaction in iRenalCancer1410, which was in turn adapted from[35]. This reaction accounts for all major macromolecular components in the biomass (e.g., membrane lipids, proteins, etc.), and the respective stoichiometric coefficient is reflective of the contribution of each component in 1g of cancer biomass. The simulation of a single gene knockout using FBA was performed by formulating the linear program problem (2) for each gene g in the model:

$$\max \ \boldsymbol{v}_{obj} \tag{2}$$

$$0 \ < \boldsymbol{v}_{obj} \leq \boldsymbol{\mu} \tag{3}$$

$$\boldsymbol{S} \cdot \boldsymbol{v} = 0 \tag{4}$$

$$1000 \ \leq \boldsymbol{v}_{\boldsymbol{i}} \leq \ + 1000 \ \ \forall \ \boldsymbol{i} \ \in \ \{\text{Exchange reaction indexes for FBS metabolites}\} \tag{5}$$

$$\boldsymbol{\alpha}_{\boldsymbol{j}} \leq \boldsymbol{v}_{\boldsymbol{j}} \leq \boldsymbol{\beta}_{\boldsymbol{j}} \ \ \forall \ \boldsymbol{j} \ \in \ \{\text{Exchange reaction indexes for measured metabolites}\} \tag{6}$$

$$\boldsymbol{v}_{\boldsymbol{r}} = 0 \text{ where r } \in \ \{\text{Reaction indexes univocally encoded by gene g}\} \tag{7}$$

where v_{obj} is the flux through the biomass equation, $\boldsymbol{\mu}$ is the experimental growth rate (for simulation using cell line specific fluxes) or an arbitrary number (for simulation without specific constraints on the exchange reaction), $\boldsymbol{S}$ is the stoichiometric matrix of the model (that is a m x n matrix where m is the number of metabolites and n is the number of reactions and each (i,j) entry is the stoichiometric coefficient of the metabolite corresponding to row i in the reaction corresponding to column j), $\boldsymbol{v}$ is the vector containing the values of the fluxes through each reaction in the model, and α (resp. β) are the lower (resp. upper) bound for the exchange flux corresponding to each metabolite measured in[51] and adjusted by[50]. These bounds were implemented only in the simulations using cell line specific fluxes. For a given cell line, they were calculated for each measured metabolite j as $\overline{\boldsymbol{v}}_{\boldsymbol{j}} \ \pm \ 2 \cdot \sigma_{v_j}$, where $\overline{\boldsymbol{v}}$ and $\sigma_{\boldsymbol{v}}$ are the mean and the standard deviation of the corresponding exchange flux in the two replicate measurements. In the simulations using cell line specific exchange fluxes, the biomass equation coefficients were multiplied by a conversion factor equal to 550 pg_{DW} $cell^{-1}$ to accommodate the fact that exchange fluxes were measured in fmol $cell^{-1}$ h^{-1} instead of mmol $g_{DW}{}^{-1}$ h^{-1} as normally assumed in genome-scale metabolic modeling[77].

The problem was formulated using native functions in the RAVEN Toolbox[78] and solved using MOSEK v.7. Simulation results are reported in Supplementary Data 3. All constrained simulation-ready models are available through the website http://www.metabolicatlas.com/.

Importantly, for the purpose of the study, the optimization part is not relevant. Indeed, we are interested on whether a feasible region exists upon the constraint imposed by the gene knockout (7), i.e. whether the fact that the encoded reactions cannot carry flux implies no flux in the biomass equation. However, if also exchange fluxes were implemented to perform FBA (6), we took in account a significant reduction (min. 50%) of the optimum (which is upper bounded by the experimental growth rate) to classify a gene as essential *in silico*. Besides this, a gene is deemed essential *in silico* when there is no solution to (2), i.e. there cannot be found a flux distribution such that the biomass equation carries flux. This is indeed valuable in light of the consideration above: among all the available metabolic states permitted by the availability of serum metabolites, no scenario allows for a flux towards all biomass precursors simultaneously.

It should be noticed that, when implementing cell line specific exchange fluxes, a set of numerical constraints had to be neglected and converted to ± 1000. This operation is obligated by the fact that some measured fluxes are not consistent with the network topology. Thus, either the measured compounds are not used by the cellular reaction network or further experimental validation is required. It may also be that the model is not complete and it should be updated in an iterative fashion, a process normally encountered in genome-scale models[79]. Examples of these inconsistencies are the conjugated bile acids (such as glycochenodeoxycholate) or anthranilate, that were measured to be absorbed by the cancer cell lines in Jain *et al.* study[51], yet there is no evidence that these compounds can be degraded in any metabolic reaction accounted in the model. In other cases, coupled measured fluxes are stoichiometrically unbalanced. To get the minimum set of constraints (6) that had to be lifted to get a feasible solution to the problem (2) (neglecting the constraint 7), we formulated a linear program that iteratively searches for the minimum sum of the fluxes that must be supplemented to each exchange flux in (6) such that all other constraints are satisfied while $v_{obj} > 0$. In the end, all exchange fluxes in (6) that required a supplementary flux (i.e. the imposed bounds in the original problem would be infeasible) were instead bounded to ± 1000. This procedure had been repeated for each cell line specific to a certain cancer type, and for all successive simulations, only the exchange fluxes that were feasible in all cell lines for a given cancer were retained. In ccRCC, 61 and 38 exchange fluxes were coordinately bounded for all seven ccRCC cell lines in FBS and Ham's medium respectively (Supplementary Fig. 12–15). In PC, 60 and 57 exchange fluxes were coordinately bounded for all two PC cell lines in FBS and Ham's medium respectively (Supplementary Fig. 16–19).

Characterization of in silico essentiality. The test for toxicity in normal cell types was performed as previously described[33]. Briefly, 83 normal cell type genome-scale metabolic models were downloaded at www.metabolicatlas.org. For each model, a list of 56 metabolic tasks was simulated under the constraint (7) for each gene classified as essential *in silico* and validated as such *in vitro*. If no solution were found, the gene knockout is deemed toxic for a certain normal cell type. If more than 50% of the 83 normal cell types show no toxicity in any metabolic task upon knockout of a gene essential in a cancer, than the gene is regarded non toxic to normal cells. Furthermore, if knockout of a gene essential in a cancer shows no toxicity in any metabolic task in the supposed cell type of origin, than the gene is regarded selectively essential to a cancer.

To elucidate the mechanism through which a gene is selectively essential *in silico*, different simulations were carried out according to the constraint that first resulted in an unfeasible solution:

1. In the case of genes essential using as a sole constraint to perform FBA the topology of a cancer metabolic network, there are two possible explanations. In the first scenario, a reaction essential to carry flux towards the biomass equation is encoded by a single gene in the cancer-specific metabolic network because all other isoenzymes are not expressed in the cancer (loss of gene redundancy). This was the case of *AGPAT6*, and it was found by constraining the flux of the encoded reaction to zero in the generic human metabolic network from which the ccRCC metabolic network topology is derived. This constraint results in an unfeasible solution using the generic network, suggesting that at least one of the isoenzymes must be expressed to sustain biomass formation. In the second scenario, there are two alternative routes to support a flux towards the biomass equation, each encoded by a single gene, but just one is expressed in the cancer-specific metabolic network (loss of pathway redundancy). This was the case of *GALT*. To verify this, the reaction encoded by *GALT* was constrained to zero in the generic human metabolic network, and then a second round of single gene-knockouts was performed using the *GALT*-KO generic human metabolic network. In the end, the double *GALT-UGP2* knockout results in an unfeasible solution in the generic human metabolic network, indicative that *UGP2* encodes for a potential alternative pathway to *GALT* that is not expressed in ccRCC.
2. In the case of genes essential when using also the exchange fluxes to perform FBA, but non essential when the sole topology was used as a constraint, an unfeasible solution is trivially attributable to the implementation of one (or more) of these additional constraints. Therefore, for each essential gene, all constraints in (6) are released (set the lower and upper bound to −1000 and +1000 respectively) one at the time, until the metabolite whose constraint on the exchange flux caused unfeasibility is spotted. For *GSS* and *GCLC,* this metabolite is glutamate, while for *RRM2B* 3-ureidopropionate. Further interpretation of the individual mechanisms was achieved by following the fluxes around each key metabolite in the simulation, either when no knockout was applied or when the knockout was applied neglecting the constraint in the key metabolite exchange flux.

The fraction of ccRCC samples where a protein involved in any of the mechanisms above is stained with at least a weak signal was retrieved from the Human Protein Atlas v. 11[56].

References

1. Vander Heiden, M. G., Cantley, L. C. & Thompson, C. B. Understanding the Warburg effect: the metabolic requirements of cell proliferation. *Science* **324,** 1029–1033, doi:10.1126/science.1160809 (2009).
2. Ward, P. S. & Thompson, C. B. Metabolic reprogramming: a cancer hallmark even warburg did not anticipate. *Cancer Cell* **21,** 297–308, doi:10.1016/j.ccr.2012.02.014 (2012).
3. Schulze, A. & Harris, A. L. How cancer metabolism is tuned for proliferation and vulnerable to disruption (vol 491, pg 364, 2012). *Nature* **494,** 130–130, doi:Doi 10.1038/Nature11827 (2013).
4. Ben-Sahra, I., Howell, J. J., Asara, J. M. & Manning, B. D. Stimulation of de Novo Pyrimidine Synthesis by Growth Signaling Through mTOR and S6K1. *Science*, doi:10.1126/science.1228792 (2013).
5. Robitaille, A. M. *et al.* Quantitative Phosphoproteomics Reveal mTORC1 Activates de Novo Pyrimidine Synthesis. *Science*, doi:10.1126/science.1228771 (2013).
6. Jeong, S. M. *et al.* SIRT4 has tumor-suppressive activity and regulates the cellular metabolic response to DNA damage by inhibiting mitochondrial glutamine metabolism. *Cancer Cell* **23,** 450–463, doi:10.1016/j.ccr.2013.02.024 (2013).
7. Letouze, E. *et al.* SDH mutations establish a hypermethylator phenotype in paraganglioma. *Cancer Cell* **23,** 739–752, doi:10.1016/j.ccr.2013.04.018 (2013).
8. Vazquez, F. *et al.* PGC1alpha expression defines a subset of human melanoma tumors with increased mitochondrial capacity and resistance to oxidative stress. *Cancer Cell* **23,** 287–301, doi:10.1016/j.ccr.2012.11.020 (2013).
9. Faubert, B. *et al.* AMPK is a negative regulator of the Warburg effect and suppresses tumor growth in vivo. *Cell Metab* **17,** 113–124, doi:10.1016/j.cmet.2012.12.001 (2013).
10. Yang, L. *et al.* Metabolic shifts toward glutamine regulate tumor growth, invasion and bioenergetics in ovarian cancer. *Mol Syst Biol* **10,** 728, doi:10.1002/msb.20134892 (2014).
11. Birsoy, K. *et al.* Metabolic determinants of cancer cell sensitivity to glucose limitation and biguanides. *Nature* **508,** 108–112, doi:10.1038/nature13110 (2014).
12. Patra, K. C. *et al.* Hexokinase 2 is required for tumor initiation and maintenance and its systemic deletion is therapeutic in mouse models of cancer. *Cancer Cell* **24,** 213–228, doi:10.1016/j.ccr.2013.06.014 (2013).
13. Chen, L. *et al.* SYK inhibition modulates distinct PI3K/AKT- dependent survival pathways and cholesterol biosynthesis in diffuse large B cell lymphomas. *Cancer Cell* **23,** 826–838, doi:10.1016/j.ccr.2013.05.002 (2013).

14. Possemato, R. *et al.* Functional genomics reveal that the serine synthesis pathway is essential in breast cancer. *Nature* **476,** 346–350, doi:10.1038/nature10350 (2011).
15. Wheaton, W. W. *et al.* Metformin inhibits mitochondrial complex I of cancer cells to reduce tumorigenesis. *Elife* **3,** e02242, doi:10.7554/eLife.02242 (2014).
16. Cunningham, J. T., Moreno, M. V., Lodi, A., Ronen, S. M. & Ruggero, D. Protein and nucleotide biosynthesis are coupled by a single rate-limiting enzyme, PRPS2, to drive cancer. *Cell* **157,** 1088–1103, doi:10.1016/j.cell.2014.03.052 (2014).
17. Ding, J. *et al.* The histone H3 methyltransferase G9A epigenetically activates the serine-glycine synthesis pathway to sustain cancer cell survival and proliferation. *Cell Metab* **18,** 896–907, doi:10.1016/j.cmet.2013.11.004 (2013).
18. Orth, J. D., Thiele, I. & Palsson, B. O. What is flux balance analysis? *Nat Biotechnol* **28,** 245–248, doi:10.1038/nbt.1614 (2010).
19. Joyce, A. R. & Palsson, B. O. Predicting gene essentiality using genome-scale in silico models. *Methods Mol Biol* **416,** 433–457, doi:10.1007/978-1-59745-321-9_30 (2008).
20. Suthers, P. F., Zomorrodi, A. & Maranas, C. D. Genome-scale gene/reaction essentiality and synthetic lethality analysis. *Mol Syst Biol* **5,** 301, doi:10.1038/msb.2009.56 (2009).
21. Jerby, L. & Ruppin, E. Predicting Drug Targets and Biomarkers of Cancer via Genome-Scale Metabolic Modeling. *Clinical Cancer Research* **18,** 5572–5584, doi:Doi 10.1158/1078-0432.Ccr-12-1856 (2012).
22. Mardinoglu, A., Gatto, F. & Nielsen, J. Genome-scale modeling of human metabolism - a systems biology approach. *Biotechnology Journal*, doi:10.1002/biot.201200275 (2013).
23. Lewis, N. E. *et al.* Large-scale in silico modeling of metabolic interactions between cell types in the human brain. *Nat Biotechnol* **28,** 1279–1285, doi:10.1038/nbt.1711 (2010).
24. Varemo, L., Nookaew, I. & Nielsen, J. Novel insights into obesity and diabetes through genome-scale metabolic modeling. *Front Physiol* **4,** 92, doi:10.3389/fphys.2013.00092 (2013).
25. Lewis, N. E., Nagarajan, H. & Palsson, B. O. Constraining the metabolic genotype-phenotype relationship using a phylogeny of in silico methods. *Nature Reviews Microbiology* **10,** 291–305, doi:Doi 10.1038/Nrmicro2737 (2012).
26. Bordbar, A., Monk, J. M., King, Z. A. & Palsson, B. O. Constraint-based models predict metabolic and associated cellular functions. *Nat Rev Genet* **15,** 107–120, doi:10.1038/nrg3643 (2014).
27. Becker, S. A. & Palsson, B. O. Context-specific metabolic networks are consistent with experiments. *PLoS Comput Biol* **4,** e1000082, doi:10.1371/journal.pcbi.1000082 (2008).
28. Schmidt, B. J. *et al.* GIM3E: condition-specific models of cellular metabolism developed from metabolomics and expression data. *Bioinformatics* **29,** 2900–2908, doi:10.1093/bioinformatics/btt493 (2013).
29. Shlomi, T., Cabili, M. N., Herrgard, M. J., Palsson, B. O. & Ruppin, E. Network-based prediction of human tissue-specific metabolism. *Nat Biotechnol* **26,** 1003–1010, doi:10.1038/nbt.1487 (2008).
30. Jerby, L., Shlomi, T. & Ruppin, E. Computational reconstruction of tissue-specific metabolic models: application to human liver metabolism. *Mol Syst Biol* **6,** 401, doi:10.1038/msb.2010.56 (2010).
31. Wang, Y., Eddy, J. A. & Price, N. D. Reconstruction of genome-scale metabolic models for 126 human tissues using mCADRE. *BMC Syst Biol* **6,** 153, doi:10.1186/1752-0509-6-153 (2012).
32. Vlassis, N., Pacheco, M. P. & Sauter, T. Fast reconstruction of compact context-specific metabolic network models. *PLoS Comput Biol* **10,** e1003424, doi:10.1371/journal.pcbi.1003424 (2014).
33. Agren, R. *et al.* Identification of anticancer drugs for hepatocellular carcinoma through personalized genome-scale metabolic modeling. *Mol Syst Biol* **10,** 721, doi:10.1002/msb.145122 (2014).
34. Agren, R. *et al.* Reconstruction of genome-scale active metabolic networks for 69 human cell types and 16 cancer types using INIT. *PLoS Comput Biol* **8,** e1002518, doi:10.1371/journal.pcbi.1002518 (2012).
35. Frezza, C. *et al.* Haem oxygenase is synthetically lethal with the tumour suppressor fumarate hydratase. *Nature* **477,** 225–228, doi:10.1038/nature10363 (2011).
36. Agren, R. *et al.* Identification of anticancer drugs for hepatocellular carcinoma through personalized genome-scale metabolic modeling. *Mol Syst Biol* **10,** 721, doi:10.1002/msb.145122 (2014).
37. Yizhak, K. *et al.* A computational study of the Warburg effect identifies metabolic targets inhibiting cancer migration. *Mol Syst Biol* **10,** 744, doi:10.15252/msb.20134993 (2014).
38. Jerby-Arnon, L. *et al.* Predicting Cancer-Specific Vulnerability via Data-Driven Detection of Synthetic Lethality. *Cell* **158,** 1199–1209, doi:10.1016/j.cell.2014.07.027 (2014).
39. Folger, O. *et al.* Predicting selective drug targets in cancer through metabolic networks. *Mol Syst Biol* **7,** 501, doi:10.1038/msb.2011.35 (2011).
40. Rini, B. I., Campbell, S. C. & Escudier, B. Renal cell carcinoma. *Lancet* **373,** 1119–1132, doi:10.1016/S0140-6736(09)60229-4 (2009).
41. Gatto, F., Nookaew, I. & Nielsen, J. Chromosome 3p loss of heterozygosity is associated with a unique metabolic network in clear cell renal carcinoma. *Proc Natl Acad Sci U S A* **111,** E866–875, doi:10.1073/pnas.1319196111 (2014).
42. Freshney, R. I. *Culture of animal cells : a manual of basic technique and specialized applications*. 6th edn (Wiley-Blackwell, 2010).
43. Price, N. D., Reed, J. L. & Palsson, B. O. Genome-scale models of microbial cells: evaluating the consequences of constraints. *Nat Rev Microbiol* **2,** 886–897, doi:10.1038/nrmicro1023 (2004).
44. Palsson, B. *Systems biology : properties of reconstructed networks*. (Cambridge University Press, 2006).
45. Cairns, R. A., Harris, I. S. & Mak, T. W. Regulation of cancer cell metabolism. *Nat Rev Cancer* **11,** 85–95, doi:10.1038/nrc2981 (2011).
46. Creighton, C. J. *et al.* Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature*, doi:10.1038/nature12222 (2013).
47. Li, B. *et al.* Fructose-1,6-bisphosphatase opposes renal carcinoma progression. *Nature*, doi:10.1038/nature13557 (2014).
48. Nilsson, H. *et al.* Primary clear cell renal carcinoma cells display minimal mitochondrial respiratory capacity resulting in pronounced sensitivity to glycolytic inhibition by 3-Bromopyruvate. *Cell Death Dis* **6,** e1585, doi:10.1038/cddis.2014.545 (2015).
49. Mardinoglu, A. *et al.* Integration of clinical data with a genome-scale metabolic model of the human adipocyte. *Molecular Systems Biology* **9,** doi:10.1038/msb.2013.5 (2013).
50. Dolfi, S. C. *et al.* The metabolic demands of cancer cells are coupled to their size and protein synthesis rates. *Cancer Metab* **1,** 20, doi:10.1186/2049-3002-1-20 (2013).
51. Jain, M. *et al.* Metabolite profiling identifies a key role for glycine in rapid cancer cell proliferation. *Science* **336,** 1040–1044, doi:10.1126/science.1218595 (2012).
52. Forster, J., Famili, I., Palsson, B. O. & Nielsen, J. Large-scale evaluation of in silico gene deletions in Saccharomyces cerevisiae. *OMICS* **7,** 193–202 (2003).
53. Ros, S. *et al.* Functional metabolic screen identifies 6-phosphofructo-2-kinase/fructose-2,6-biphosphatase 4 as an important regulator of prostate cancer cell survival. *Cancer Discov* **2,** 328–343, doi:10.1158/2159-8290.CD-11-0234 (2012).
54. Cai, C. *et al.* Intratumoral de novo steroid synthesis activates androgen receptor in castration-resistant prostate cancer and is upregulated by treatment with CYP17A1 inhibitors. *Cancer Research* **71,** 6503–6513, doi:10.1158/0008-5472.CAN-11-0532 (2011).

55. Kum, J. B. *et al.* Mixed epithelial and stromal tumors of the kidney: evidence for a single cell of origin with capacity for epithelial and stromal differentiation. *Am J Surg Pathol* **35,** 1114–1122, doi:10.1097/PAS.0b013e3182233fb6 (2011).
56. Fagerberg, L. *et al.* Contribution of Antibody-based Protein Profiling to the Human Chromosome-centric Proteome Project (C-HPP). *J Proteome Res*, doi:10.1021/pr300924j (2012).
57. Burdo, J., Dargusch, R. & Schubert, D. Distribution of the cystine/glutamate antiporter system xc- in the brain, kidney, and duodenum. *J Histochem Cytochem* **54,** 549–557, doi:10.1369/jhc.5A6840.2006 (2006).
58. Monk, J., Nogales, J. & Palsson, B. O. Optimizing genome-scale network reconstructions. *Nat Biotechnol* **32,** 447–452, doi:10.1038/nbt.2870 (2014).
59. Becker, S. A. & Palsson, B. O. Three factors underlying incorrect in silico predictions of essential metabolic genes. *BMC Syst Biol* **2,** 14, doi:10.1186/1752-0509-2-14 (2008).
60. Hubner, K., Sahle, S. & Kummer, U. Applications and trends in systems biology in biochemistry. *Febs Journal* **278,** 2767–2857, doi:10.1111/j.1742-4658.2011.08217.x (2011).
61. Agren, R. *et al.* Identification of anticancer drugs for hepatocellular carcinoma through personalized genome-scale metabolic modeling. *Mol Syst Biol* **10** (2014).
62. Jeon, S. M., Chandel, N. S. & Hay, N. AMPK regulates NADPH homeostasis to promote tumour cell survival during energy stress. *Nature* **485,** 661–665, doi:10.1038/nature11066 (2012).
63. Mardinoglu, A. *et al.* Genome-scale metabolic modelling of hepatocytes reveals serine deficiency in patients with non-alcoholic fatty liver disease. *Nat Commun* **5,** 3083, doi:10.1038/ncomms4083 (2014).
64. Thiele, I. *et al.* A community-driven global reconstruction of human metabolism. *Nat Biotechnol* **31,** 419–425, doi:10.1038/nbt.2488 (2013).
65. Ciriello, G. *et al.* Emerging landscape of oncogenic signatures across human cancers. *Nat Genet* **45,** 1127–1133, doi:10.1038/ng.2762 (2013).
66. Gerlinger, M. *et al.* Genomic architecture and evolution of clear cell renal cell carcinomas defined by multiregion sequencing. *Nat Genet* **46,** 225–233, doi:10.1038/ng.2891 (2014).
67. Agarwal, A. K. Lysophospholipid acyltransferases: 1-acylglycerol-3-phosphate O-acyltransferases. From discovery to disease. *Curr Opin Lipidol* **23,** 290–302, doi:10.1097/MOL.0b013e328354fcf4 (2012).
68. Lai, K. & Elsas, L. J. Overexpression of human UDP-glucose pyrophosphorylase rescues galactose-1-phosphate uridyltransferase-deficient yeast. *Biochem Biophys Res Commun* **271,** 392–400, doi:10.1006/bbrc.2000.2629 (2000).
69. Leslie, N., Yager, C., Reynolds, R. & Segal, S. UDP-galactose pyrophosphorylase in mice with galactose-1-phosphate uridyltransferase deficiency. *Mol Genet Metab* **85,** 21–27, doi:10.1016/j.ymgme.2005.01.004 (2005).
70. Pompella, A., Bánhegyi, G. b. & Wellman-Rousseau, M. *Thiol metabolism and redox regulation of cellular functions*. (IOS Press, 2002).
71. Toyokuni, S. Iron and thiols as two major players in carcinogenesis: friends or foes? *Front Pharmacol* **5,** 200, doi:10.3389/fphar.2014.00200 (2014).
72. Liu, X. *et al.* Ribonucleotide reductase small subunit M2B prognoses better survival in colorectal cancer. *Cancer Research* **71,** 3202–3213, doi:10.1158/0008-5472.CAN-11-0054 (2011).
73. Zhang, K. *et al.* p53R2 inhibits the proliferation of human cancer cells in association with cell-cycle arrest. *Mol Cancer Ther* **10,** 269–278, doi:10.1158/1535-7163.MCT-10-0728 (2011).
74. Cho, E. C. *et al.* Tumor suppressor FOXO3 regulates ribonucleotide reductase subunit RRM2B and impacts on survival of cancer patients. *Oncotarget* **5,** 4834–4844 (2014).
75. Chang, L., Guo, R., Huang, Q. & Yen, Y. Chromosomal instability triggered by Rrm2b loss leads to IL-6 secretion and plasmacytic neoplasms. *Cell Rep* **3,** 1389–1397, doi:10.1016/j.celrep.2013.03.040 (2013).
76. Kruschke, J. K. Bayesian estimation supersedes the t test. *J Exp Psychol Gen* **142,** 573–603, doi:10.1037/a0029146 (2013).
77. Feist, A. M. & Palsson, B. O. The biomass objective function. *Curr Opin Microbiol* **13,** 344–349, doi:10.1016/j.mib.2010.03.003 (2010).
78. Agren, R. *et al.* The RAVEN toolbox and its use for generating a genome-scale metabolic model for Penicillium chrysogenum. *PLoS Comput Biol* **9,** e1002980, doi:10.1371/journal.pcbi.1002980 (2013).
79. Thiele, I. & Palsson, B. O. A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nat Protoc* **5,** 93–121, doi:10.1038/nprot.2009.203 (2010).

Acknowledgements

The authors wish to thank Dr. M. Jiang, Dr. R. Saunders and Dr. M. Howell (LRI High Throughput Screening Facility) for help with performing the RNAi screen. The computations were performed on resources provided by the Swedish National Infrastructure for Computing (SNIC) at C3SE. F.G. and J.N. acknowledge Knut and Alice Wallenberg Foundation and Chalmers Foundation for sponsoring this work.

Author Contributions

F.G. performed the computational analyses and drafted the manuscript. H.M. performed the experiments. F.G., H.M., A.S., and J.N. designed the study and edited the manuscript in the final form.

Additional Information

Supplementary information accompanies this paper at http://www.nature.com/srep

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Gatto, F. *et al.* Flux balance analysis predicts essential genes in clear cell renal cell carcinoma metabolism. *Sci. Rep.* **5**, 10738; doi: 10.1038/srep10738 (2015).

PAPER IV

Changes in glycosaminoglycan biosynthesis regulation in clear cell renal cell carcinoma are detectable in patients' plasma and urine and can diagnose metastasis

F. Gatto, N. Volpi, H. Nilsson, I. Nookaew, M. Maruzzo, A. Roma, M. E. Johansson, U. Stierner, S. Lundstam, U. Basso, J. Nielsen

Submitted for publication

Changes in glycosaminoglycan biosynthesis regulation in clear cell renal cell carcinoma are detectable in patients' plasma and urine and can diagnose metastasis

Francesco Gatto[1], Nicola Volpi[2], Helén Nilsson[3], Intawat Nookaew[1,*], Marco Maruzzo[4], Anna Roma[4], Martin E. Johansson[3], Ulrika Stierner[5], Sven Lundstam[6], Umberto Basso[4], and Jens Nielsen[1].

Affiliations:

[1] Department of Biology and Biological Engineering, Chalmers University of Technology, Göteborg, Sweden.

[2] Department of Life Sciences, University of Modena and Reggio Emilia, Modena, Italy.
[3] Department of Translational Medicine Malmö, Center for Molecular Pathology, Lund University, Skåne University Hospital, Malmö, Sweden.
[4] Medical Oncology Unit 1, Istituto Oncologico Veneto IOV - IRCCS, Padova, Italy.
[5] Department of Oncology, Institute of Clinical Sciences, Sahlgrenska Academy at the University of Gothenburg, Sahlgrenska University Hospital, Göteborg, Sweden
[6] Department of Urology, Sahlgrenska University Hospital and Sahlgrenska Academy, University of Gothenburg, Göteborg, Sweden
[*] Present address: Comparative Genomics Group, Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831, United States

Contacts: To whom correspondence should be addressed: J. Nielsen nielsenj@chalmers.se

SUMMARY

Clear cell renal cell carcinoma (ccRCC) is the most common form of kidney cancer. Currently, no diagnostic biomarkers entered the clinical routine. Here, we used genome-scale metabolic modeling to pinpoint unique metabolic reprogramming in ccRCC, given recent findings of its involvement in ccRCC progression. Contrary to other six cancers, we discovered a strong coordinated regulation of the glycosaminoglycan (GAG) biosynthesis only in ccRCC. Extracellular GAGs are macromolecules previously implicated in tumor metastasis. We speculated that such regulation could be translated to develop an accessible diagnostic marker for metastatic ccRCC (mccRCC). We measured 18 independent GAG properties in plasma and urine of 34 mccRCC patients and 16 healthy individuals. The GAG profile was distinctively altered in mccRCC. Based on the data we designed three GAG markers that distinguished mccRCC from healthy individuals with accuracy ranging 82.7% to 100%. A negative predictive value equal to 100% was validated in an independent cohort of 18 mccRCC patients and 9 healthy individuals. In addition, these markers were predictive independent of age, gender, BMI, or dietary intake. These results demonstrate that a coordinated regulation of GAG biosynthesis takes place in ccRCC and that GAG profiling in accessible fluids is suitable for diagnosis of mccRCC.

Introduction

Clear cell renal cell carcinoma (ccRCC) is the most common form of kidney cancer (Rini et al., 2009) and it is responsible for 100'000 deaths worldwide (Ferlay et al., 2010). Roughly 50% of ccRCC are expected to develop metastatic disease, which is usually incurable. In sharp contrast to early diagnosed ccRCC, the median survival of metastatic patients is significantly worse (Gupta et al., 2008), even though the prognosis has improved after the introduction of modern targeted therapies (Wahlgren et al., 2013). Recent trials emplyoing sequential use of the tyrosine-kinase inhibitor sunitinib and mTOR inhibitor everolimus reported a median survival that exceeded 30 months (Motzer et al., 2015). However, no biomarker is currently approved for diagnosis and monitoring of metastatic ccRCC (Jonasch et al., 2012; Moch et al., 2014).

The search for molecular biomarkers has focused on ccRCC genetics and angiogenesis, but none of these biomarkers have entered clinical routine, nor are easily accessible or indicative of metastasis (Finley et al., 2011; Moch et al., 2014). On the other hand, other molecular processes prominent in ccRCC may fulfill this lack. In this sense, accumulating evidence suggests that the proliferation and survival of cancer cells rely upon a shift in their metabolism (Schulze and Harris, 2013; Vander Heiden et al., 2009; Ward and Thompson, 2012). In particular, ccRCC has been recently proved to feature a strong regulation and dependence on a distinctive metabolic reprogramming, which is pivotal to its progression (Creighton et al., 2013; Gatto et al., 2015; Gatto et al., 2014; Li et al., 2014). These outstanding metabolic changes may be of clinical interest since they have the potential to be translated as ccRCC biomarkers.

Under these premises, we here follow up on our recent study that revealed a deviating regulation of metabolism in ccRCC in contrast to seven common epithelial tumors (Gatto et al., 2014) and further computationally characterized metabolic regulation in ccRCC leveraging on a larger number of samples and using state-of-the-art genome-scale metabolic modeling (Bordbar et al., 2014; Jerby and Ruppin, 2012; Mardinoglu et al., 2013; Mardinoglu and Nielsen, 2015). As a result, we uncovered a previously unreported coordinated regulation of glycosaminoglycan (GAG) biosynthesis, exacerbated in metastasis. This led us to speculate that such regulation may be detectable in metastatic ccRCC. Hence, we designed an observational study to measure GAG profiles in accessible fluids of metastatic ccRCC patients and sought to characterize the suitability of GAG profiles as a diagnostic marker for the disease.

Results

Metabolic modeling reveals a coordinated regulation of glycosaminoglycan biosynthesis unique to clear cell renal cell carcinoma

Our recent study suggests that metabolic reprogramming in clear cell renal cell carcinoma (ccRCC) is unique likely due to genetic alterations in the tumor progression (Gatto et al., 2014). The exceptionality of metabolic regulation in ccRCC may have important clinical implications as a potential molecular biomarker. Thus, we sought to fully characterize metabolic regulation in ccRCC computationally. We retrieved a larger number of gene expression profiles from The Cancer Genome Atlas (TCGA) than in our previous study (481 tumor samples vs. 71 tumor-adjacent normal samples, here simply referred to as normal, Table S1) and employed two recent methods in genome-scale metabolic modeling named Piano (Varemo et al., 2013) and Kiwi (Varemo et al., 2014) to pinpoint respectively deregulated metabolic pathways or connected components in the metabolic network of ccRCC. When we analyzed differential gene expression in ccRCC vs. normal samples using these methods, we observed, in line with previous studies (Creighton et al., 2013; Gatto et al., 2014), widespread down-regulation of many interconnected components in the tricarboxylic acid cycle and branched-chain amino acid metabolism and up-regulation of components in the pentose phosphate pathway (Fig. S1-2). However, the analyses also returned a previously unreported sub-network of metabolites that comprises precursors of chondroitin and heparan sulfates (Fig. 1A). Furthermore, the corresponding biosynthetic pathways display a distinct and opposite regulation in ccRCC vs. normal samples (Fig. S1). Chondroitin (CS) and heparan (HS) sulfates are glycosaminoglycans (GAGs) that share a common biosynthetic route in the linkage to the core protein, but thereafter they differ in the polymerization: CS repeating disaccharide is constituted by *N*-acetylgalactosamine and glucuronic acid residues, while HS repeating disaccharide is constituted by *N*-acetylglucosamine and glucuronic acid residues (Kreuger and Kjellen, 2012; Mikami and Kitagawa, 2013). In ccRCC, we observed a coordinated regulation of GAG biosynthesis, defined by a substantial up-regulation of most genes specific to CS biosynthesis (11/13) and a concurrent down-regulation of genes specific to HS biosynthesis (8/13), pointing to a relative change in GAG disaccharide composition, sulfation, and chain length in ccRCC (Fig. 1B, Table S2). We confirmed such coordinated regulation of GAG biosynthesis in two independent datasets that compared gene expression in ccRCC vs. normal samples (Pena-Llopis et al., 2012; Wang et al., 2009), with high and significant correlations between expression fold-changes in these studies and the TCGA samples ($r = 0.87$-0.89, Fig. 1C). To verify to which extent this regulatory pattern is ccRCC-specific, we repeated an analogous analysis for six other epithelial cancer types for which at least 20 normal samples were found in The Cancer Genome Atlas (breast invasive carcinoma, colon adenocarcinoma, head and neck squamous cell carcinoma, lung adenocarcinoma, lung squamous cell carcinoma, and uterine corpus endometrial carcinoma). None of these cancers displayed the same coordinated pattern as in ccRCC, which is a clear outlier according to unsupervised hierarchical clustering, even though we found cancer type-dependent regulation of individual enzymes involved in GAG biosynthesis (Fig. 1D, Table S2). In addition, we never observed both the CS and the HS biosynthesis pathway among the top ranked regulated pathways in any of these cancers types (Fig. S3).

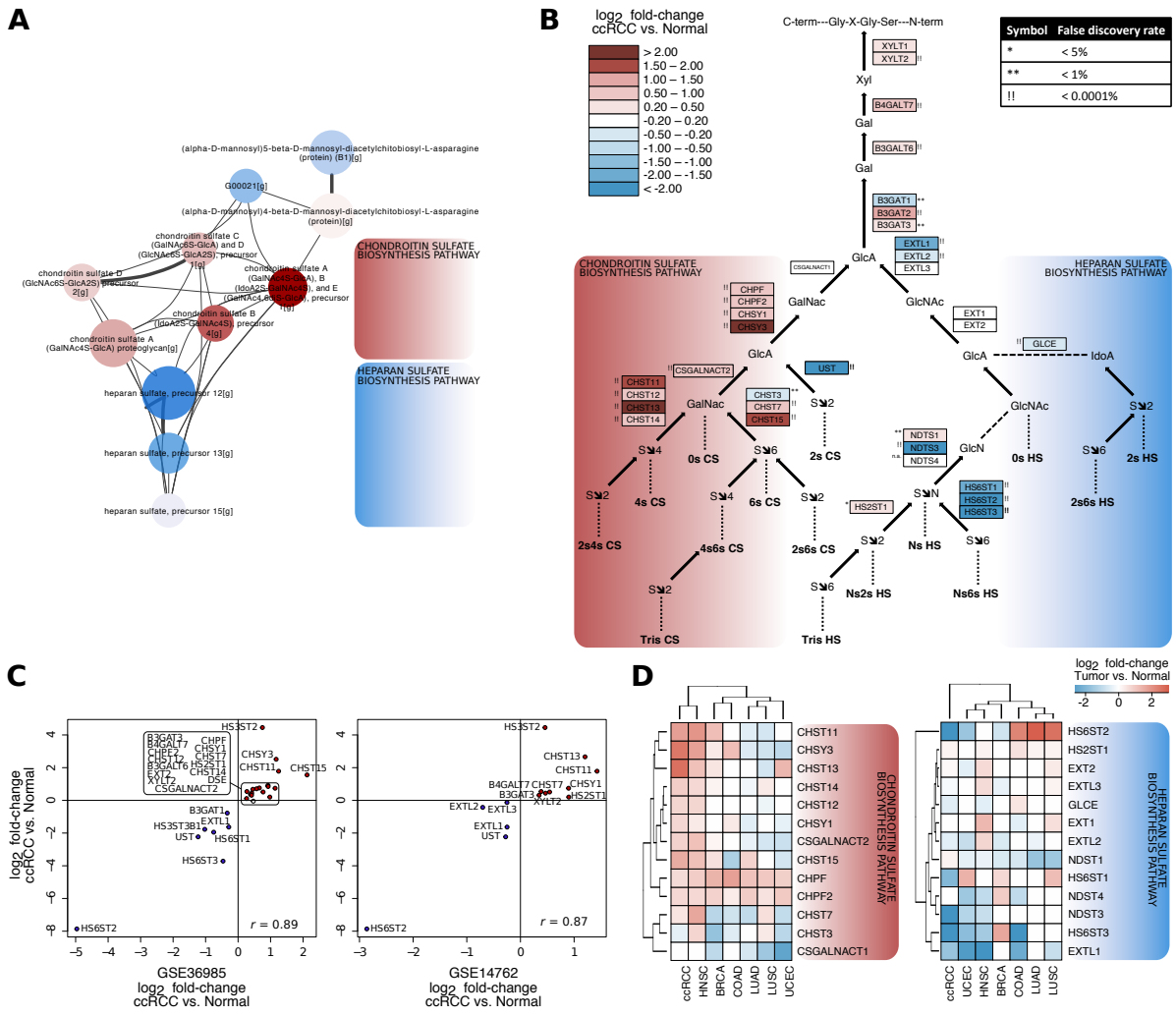


Figure 1 – Coordinated regulation of glycosaminoglycan biosynthesis in ccRCC vs. normal kidney. A) Genome-scale metabolic modeling using Kiwi reveals a coordinated transcriptional regulation in a subnetwork of metabolites belonging to the chondroitin sulfate and heparin sulfate biosynthetic pathway. The node color indicates the general direction of regulation of the genes associated with the metabolite (red – up-regulation; blue – down-regulation). See also Fig. S1-2 for further gene expression analysis in ccRCC at the metabolite and pathway level respectively. B) Pathway-view of glycosaminoglycan biosynthesis in ccRCC. Each box shows the enzyme(s) carrying out a given reaction in the pathway. The color represents the $\log_{10}$ fold-change in ccRCC vs. normal for the enzyme-coding gene, while the symbol next to each box reports the significance for the corresponding gene regulation (in terms of false discovery rate). The pathway has been drawn according to KEGG gene associations (Note that genes related to dermatan sulfate biosynthesis or sulfation at C3 in heparan sulfate are not shown, the latter event being rarely observed (Thacker et al., 2014)). Solid arrows indicate addition of a molecule, dashed lines indicate conversion of a molecule, and dotted lines indicate the final disaccharide composition up to that point. C) Correlation of gene expression $\log_2$ fold-changes in the glycosaminoglycan biosynthesis pathway between TCGA samples (y-axis) and two independent studies (GSE36986 and GSE14762, (Pena-Llopis et al., 2012; Wang et al., 2009)). D) Gene expression $\log_2$ fold-changes in the glycosaminoglycan biosynthesis pathway in ccRCC as opposed to other cancers vs. matched normal tissues. Key: HNSC – Head and neck squamous cell carcinoma; BRCA – Breast invasive carcinoma; COAD – Colon adenocarcinoma; LUAD – Lung adenocarcinoma; LUSC – Lung squamous cell carcinoma; UCEC – Uterine corpus endometrial carcinoma. See also Fig. S3 for gene expression analysis in other cancers at the pathway level.

In order to evaluate if the coordinated regulation of GAG biosynthesis is represented also at the level of protein expression, we used immunohistochemistry of a ccRCC tissue microarray to detect the presence of three representative proteins characteristic for the pathway (CHPF2 in CS biosynthesis and HS6ST2 and EXTL1 in HS biosynthesis) in ccRCC vs. normal kidney samples (Fig. 2A). In accordance with gene expression changes, CHP2 displayed strong staining in all tested tumor samples (positive in 21 of 21 samples) and only weak and likely unspecific staining in the kidney proximal tubule cells (0/2); HS6ST2 showed no or weak staining in all

tested tumor samples (positive in 0 of 32) while it was detected in both the podocytes in the kidney glomeruli and the endothelial cells of larger vessels (2/2); and EXTL1 was undetected in 96% of the tested tumor samples (positive in 1 of 27), but it stained strongly in the kidney collecting duct cells (2/2) (representative samples in Fig. 2B). Taken together, these results suggest that a coordinated regulation of GAG biosynthesis is a prominent metabolic event occurring exquisitely in the kidney during ccRCC transformation.

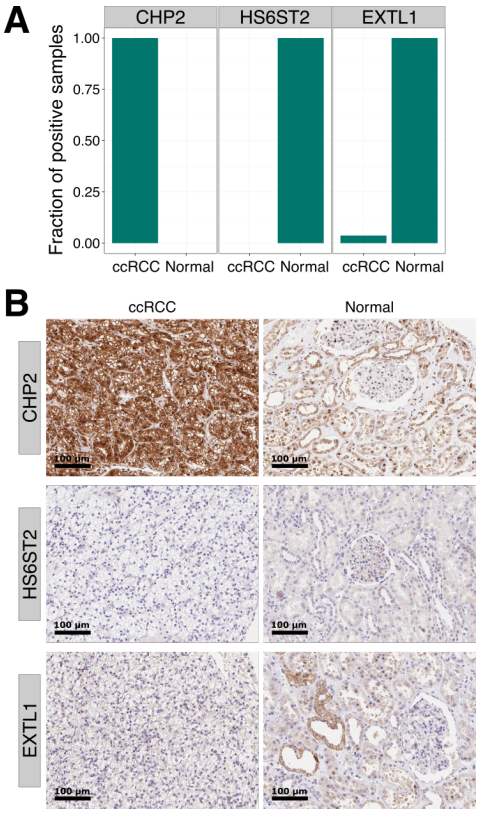


Figure 2 – Immunohistochemical staining of three proteins in glycosaminoglycan biosynthesis in ccRCC vs. normal kidney. A) Fraction of samples positive for CHP2, HS6ST2, and EXTL1 in ccRCC (21 to 27 tissue samples) vs. normal kidney (2 samples). Results are presented as the consensus of staining performed in duplicates. B) Staining for CHP2, HS6ST2, and EXTL1 in representative ccRCC and normal samples.

Altered regulation of glycosaminoglycan biosynthesis is exacerbated in metastasis and it is detectable in patients' urine and plasma

CS and HS have been long implicated in the regulation of angiogenesis, adhesion, invasion, and migration, key steps in the metastatic cascade (Afratis et al., 2012; Jackson et al., 1991). We extended our differential gene expression analysis to verify whether genes in GAG biosynthesis showed further regulation in ccRCC patients with metastasis. We found that 11 genes in GAG biosynthesis were differentially regulated in metastasis, exacerbating the overexpression of CS-associated genes and the repression of HS-associated genes (Fig. S4). This is suggestive that a coordinated regulation of GAG biosynthesis is an event accentuated with metastasis. While the assembly of GAGs chains takes place intracellularly, the completed proteoglycan is secreted in the extracellular matrix (Silbert and Sugumaran, 2002). Hence, all considered, we speculated that not only eventual changes in GAGs due to ccRCC progression might be reflected in kidney-proximal fluids, but also that these changes should be easier to detect in metastatic ccRCC (mccRCC) patients. This speculation leverages on the fact that variations in GAG concentration and composition have been observed in the proximal fluids of other diseases in which GAGs were implicated (Anower et al., 2013; Mannello et al., 2014; Mannello et al., 2015; Schmidt et al., 2014; Volpi et al., 2015).

In order to verify whether changes in the GAG profile occur in mccRCC and can be measured in accessible body fluids, we recruited a discovery cohort of 50 subjects, consisting of 34 patients with mccRCC and 16 healthy individuals (Table 1, Table S3). Plasma and urine samples were taken from all subjects, except in 21 mccRCC patients for whom only plasma samples were available. CS and HS concentration and their disaccharide composition were quantified in the samples using liquid chromatography with on-line electrospray ionization mass spectrometry (ESI-MS). In total, 18 independent GAG properties were measured in every fluid sample (note that the GAG charge is the sum of all sulfated disaccharide fractions). The collection of all these data points defines a GAG profile. We observed remarkable differences between the GAG profile of mccRCC patients compared to healthy individuals, both in the plasma and in the urine (Fig. 3A). Principal component analysis (PCA) of GAG profiles that combine plasma and urine measurements revealed that mccRCC patients clearly separate from healthy individuals (71% of variance is explained along the first component, Fig. 3B). Similar separations were achieved by using solely measurements in the plasma (81% variance) or in the urine (63% variance). These results indicate that mccRCC entails alterations in systemic GAG composition that are markedly distinctive compared to healthy individuals.

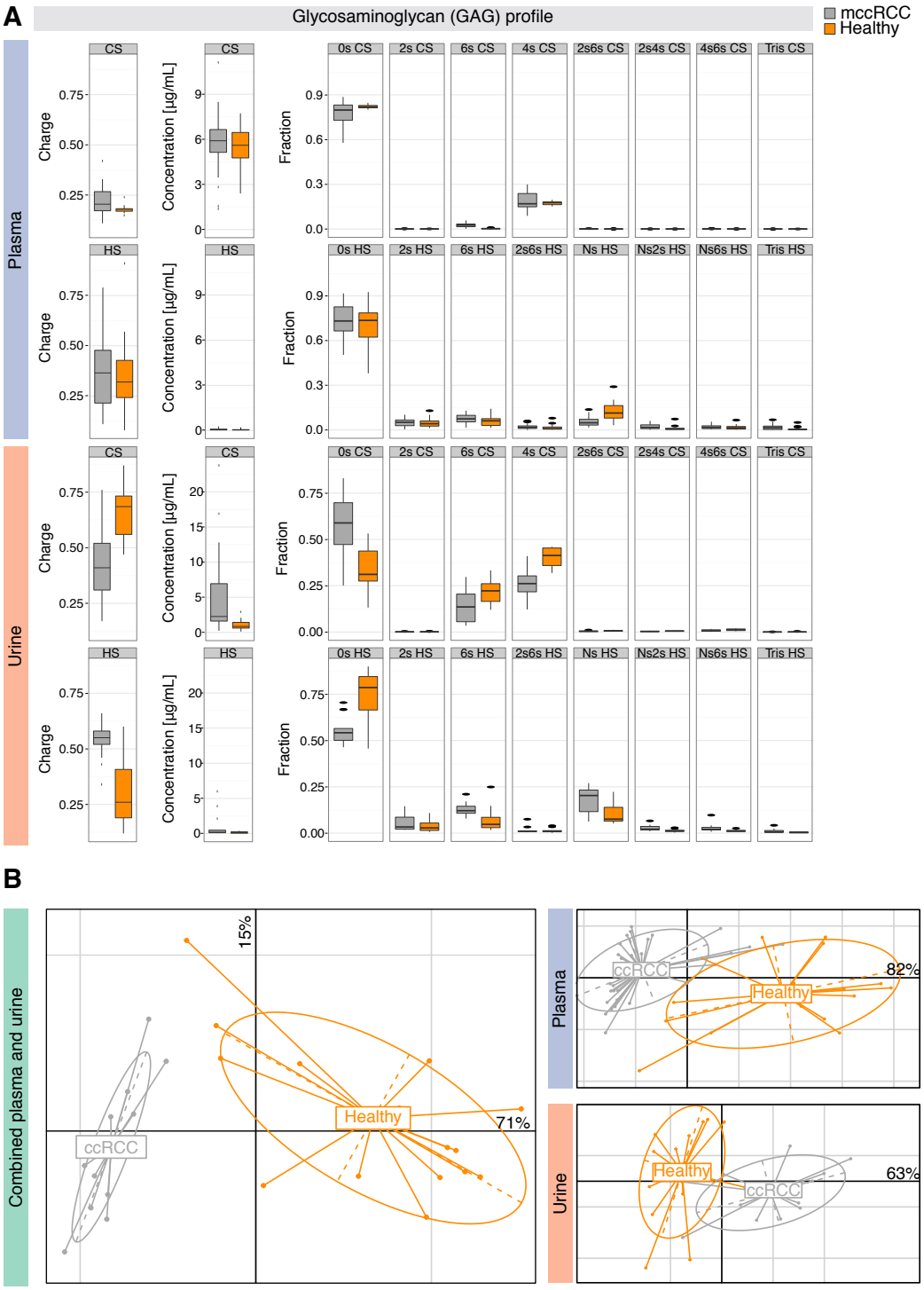


Figure 3 – The glycosaminoglycan plasma and urine profile of mccRCC patients is markedly distinct than healthy individuals. A) The glycosaminoglycan profile of mccRCC patients (grey boxplots) and healthy individuals (orange boxplots) in the plasma (top) and urine (bottom). Each profile comprises 18 independent measurements of GAGs (9 related to CS and 9 related to HS), which refer to the total concentration and the disaccharide composition. B) Principal component analysis of sample GAG profiles using measurements from plasma, urine, or both.

The changes in the plasma and urine GAG profile here largely attributed to the occurrence of mccRCC opens the opportunity to design accessible markers based on the GAG properties that best distingu

ish the disease from an healthy state. We utilized Lasso penalized logistic regression (Tibshirani, 1996) with leave-one-out cross-validation to select robust GAG properties that are most predictive of the clinical outcome (i.e. mccRCC vs. healthy). A marker was subsequently designed as a ratio, where the numerator is the sum of the properties associated with mccRCC and the denominator is the sum of the properties associated with the healthy state. Each term was normalized using the regression coefficients. We derived three potential disease markers, based on either plasma or urine or combined measurements:

$$\text{Plasma score} = \frac{[6s\ CS] + CS_{tot}}{\frac{3}{10}\frac{[4s\ CS]}{[6s\ CS]} + [Ns\ HS]}$$

$$\text{Urine score} = \frac{[Ns6s\ HS] + 60 \cdot \text{Charge}\ HS}{[4s\ CS]}$$

$$\text{Combined score} = \text{mean}(\text{Plasma score}, \text{Urine score})$$

where terms in brackets represent the fraction of the disaccharide for the corresponding GAG (the abbreviations describe different sulfation patterns for CS and HS as per Fig. 1B), CS_{tot} is the total concentration of CS (in μg/mL) and $\text{Charge}\ HS$ is the total fraction of sulfated disaccharides of HS. We then calculated the three scores for each sample and observed that mccRCC samples have recurrently elevated scores with respect to healthy samples (Fig. 4A). We computed a significant non-null mean difference in all three scores between the two groups using robust Bayesian estimation. The mean difference is equal to 2.15 for the combined score (95% High Density Interval [HDI] 1.72 to 2.60), 2.49 for the plasma score (95% HDI 1.94 to 3.05), and 0.79 for the urine score (95% HDI 0.52 to 1.06). The performance of the three markers was evaluated using the receiver-operating-characteristic (ROC) curves, and the area under the curve (AUC) was found to be 1 (perfect classifier) in the case of the combined and plasma score, and 0.966 for the urine score (Fig. 4B, Table 2). A straightforward clinical implementation of these markers would be to monitor mccRCC patients after surgery and diagnose a recurrence using a simple accessible test in adjunct to or in substitution of standard radiological tests. Thus, from each ROC curve, we computed a score cut-off that maximizes the negative predictive value (NPV) of the marker (Lopez-Raton et al., 2014) (Table 2). Taken together, these findings demonstrate that alterations in plasma and urine GAG composition occurring in mccRCC can be summarized into scores. In turn, these scores accurately distinguished diseased from healthy individuals.

In order to validate whether these scores have a reproducible accuracy in an independent cohort, we recruited 27 subjects, consisting of 18 patients with mccRCC and 9 healthy individuals (Table 1, Table S4). Plasma and urine samples were taken from all subjects, except in 11 mccRCC patients for whom only plasma samples were available. We analyzed the three markers for each individual and computed the corresponding scores. The scores were remarkably higher in mccRCC compared to healthy also in the validation cohort (Fig. 4C). We computed an AUC value equal to 1 for all three markers (Fig. 4D). Also, the NPV at the previously determined cut-off score was 100% for all markers (Table 2). This evidence strongly suggests that the three markers have the potential to indicate the occurrence of mccRCC by means of an accessible analytical test. Nevertheless, a rigorous validation of this test as a tool for mccRCC follow-up requires the calculation of the scores in subjects previously diagnosed with mccRCC but with no evidence of disease. Indeed, we cannot rule out that previous exposure to the disease may have prolonged effects on the systemic GAG composition, thus altering the scores and weakening their clinical utility. Therefore, we analyzed the markers and calculated the corresponding scores in a cohort of 8 individuals diagnosed with mccRCC but with no evidence of disease at the time of sampling. We observed a remarkable decrease in the scores from the expected value in mccRCC, even though the accuracy of the classification differed among scores. The highest accuracy was achieved for the plasma score, where the computed scores lied below the cut-off in 7 of 8 cases and hence 87.5% of the subjects were correctly classified as healthy (Fig. 5). The accuracy was lowest in the case of the urine scores, with only 2 of 8 subjects (25%) were classified as healthy. Nevertheless, it is noteworthy that, with regards to the scope for which the cut-off scores were derived, at least one subject was correctly identified as healthy (the NPV for the test was therefore 100% for all three scores, which is trivial given the cohort only includes negative controls). Even though the sample size might be too small for meaningful statistics, these results argue that plasma and urine GAG composition can be used as a robust and accurate diagnostic biomarker for the occurrence of mccRCC.

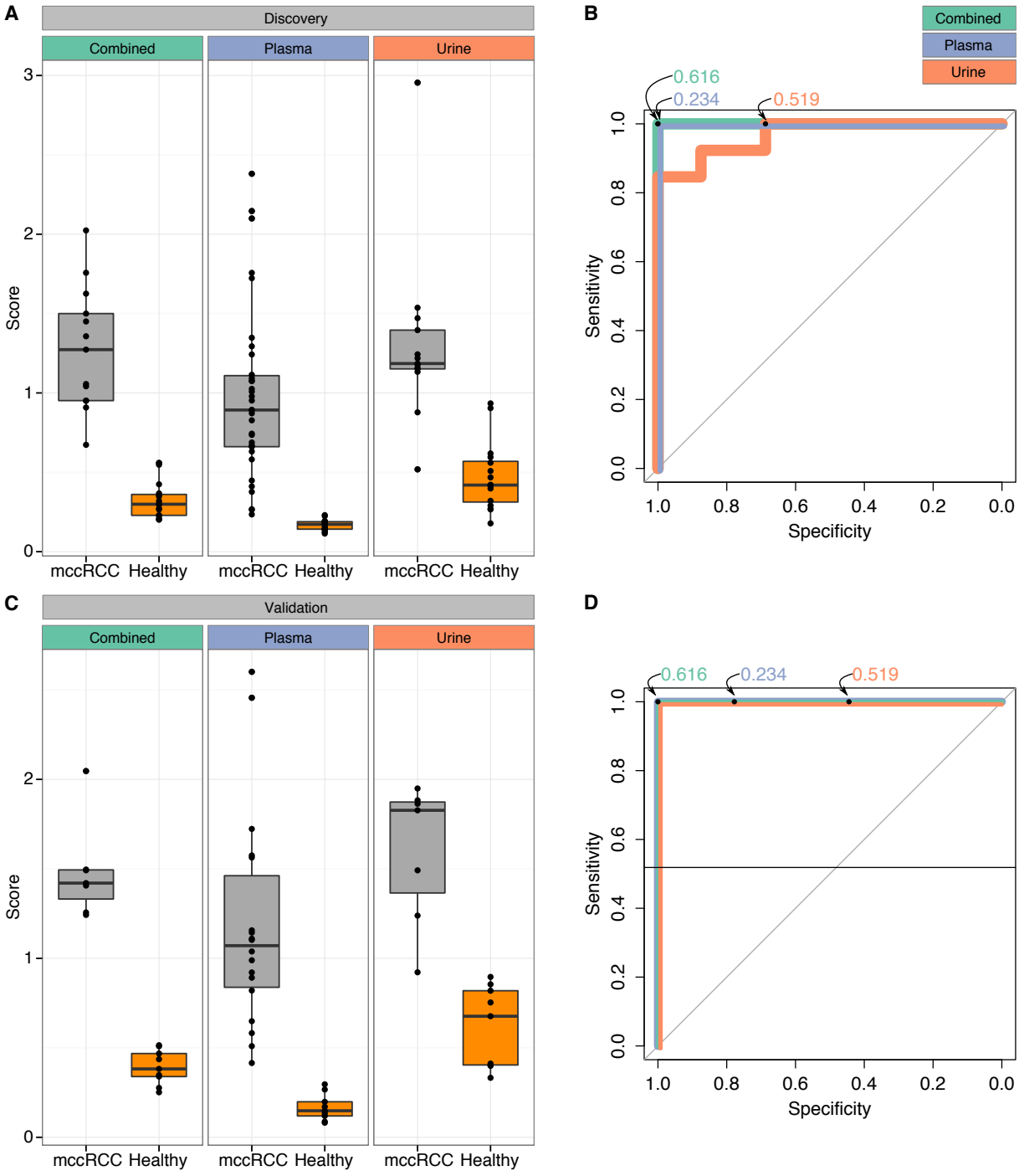


Figure 4 – The glycosaminoglycan profile can be summarized in three scores (based on measurements in the plasma, urine, or both) that can accurately predict occurrence of mccRCC. A) Plasma, urine, and combined scores in mccRCC patients (grey boxplots) and healthy individuals (orange boxplots) belonging to the discovery cohort (34 samples vs. 17, respectively). B) Receiver-operating-characteristic (ROC) curves in the classification of samples of the discovery cohort as either mccRCC or healthy based on the combined, plasma, and urine scores. For each marker, an optimal cut-off scores that maximizes the negative predictive value is indicated. C) Plasma, urine, and combined scores in mccRCC patients (grey boxplots) and healthy individuals (orange boxplots) belonging to the validation cohort (18 samples vs. 9, respectively). D) ROC curves in the classification of samples of the validation cohort as either mccRCC or healthy based on the combined, plasma, and urine scores. See also Fig. S5 for score correlations with confounding factors.

Glycosaminoglycan composition in plasma and urine is distinctively different and predictive of metastatic clear cell renal cell carcinoma regardless of confounding factors

We sought to identify the extent to which the measured systemic GAG alterations are purely attributable to ccRCC progression, as suggested by the underlying transcriptional regulation, or are also dependent on other confounding factors. Therefore, we gathered clinical and dietary information, which may confound the association of the scores with the clinical outcome, for 33 individuals (17 mccRCC and 16 healthy, Table S3). As reported in Table 1, we observed an uneven distribution of some baseline characteristics, for example gender, pasta consumption and alcohol consumption. Therefore, we tested whether the clinical outcome could be purely inferred by some of the confounding factors rather than the marker scores. First, we determined which are the most biased factors between the mccRCC vs. healthy groups. To this end, we regressed the clinical outcome based on the confounding factors and the combined score, using Lasso penalized logistic regression. This analysis selected four potentially relevant confounding factors: age, weekly consumption of pasta and rice, and use of alcohol. Then, we performed analysis of covariance using logistic regression to test the strength of the association between the clinical outcome and the combined score using the four confounding factors as covariates. Notably, none of the covariates have a significant contribution in the regression of the clinical outcome (p = 0.27 to and 0.44, Fig S5). In addition, we calculated that the logistic regression model based solely on the combined score is the most likely (p = 99.2%) according to the minimum Kullback–Leibler divergence criterion: the Akaike information criterion for the regression based on the sole combined score is significantly lower than for the regression based also on the four covariates (7.8 vs. 17.5, respectively). A similar conclusion was reached for the plasma score (6.0 vs. 17.9), but not for the urine score (23.0 vs. 17.5), where pasta consumption displayed a significant effect in the regression of the clinical outcome (p = 0.03). Taken together, these results indicate that the combined and plasma score alone (but not the urine score) have a strong association with the clinical outcome regardless of any here-considered confounding factor, prompting use of GAG measurements as unbiased predictors for occurrence of mccRCC.

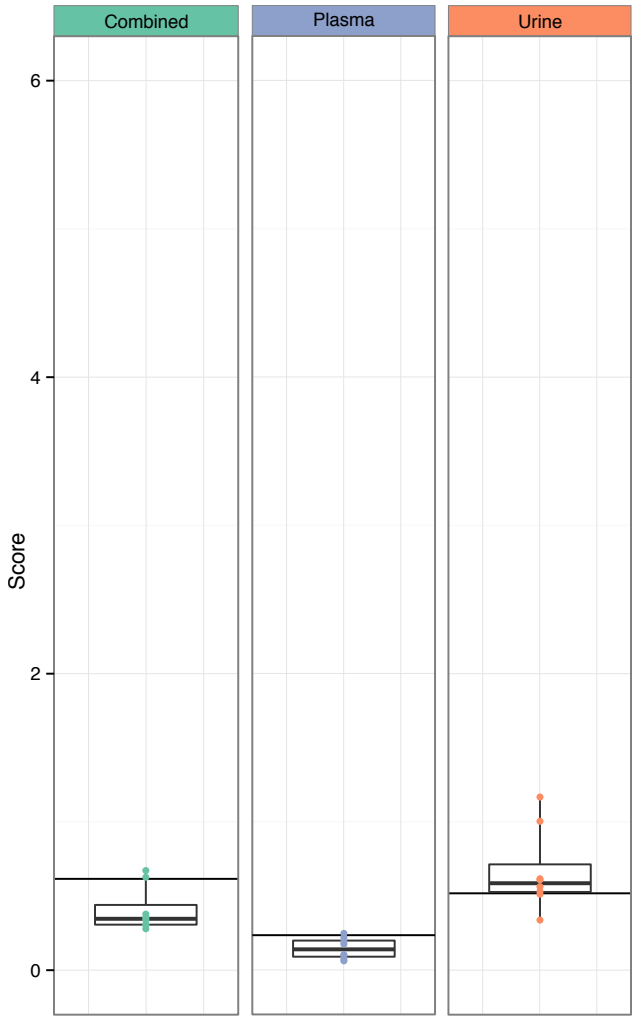


Figure 5 – Combined, plasma, and urine scores in subjects previously diagnosed with mccRCC but with no evidence of disease at the time of sampling. The horizontal lines represent the optimal cut-off scores at which a subject is classified as either mccRCC or healthy.

Finally, we explored if systemic therapy has an effect on the marker scores given that these are calculated by profiling body fluids. We limited our analysis to the patients for which solely plasma samples were collected (and hence we checked solely the effect on the plasma scores), because for this group we noticed a comparable number of treated (n = 19) and untreated (n = 33) patients. We did not observe any significant correlation between the plasma score and the use of systemic therapy, based on a linear regression of the score on the treatment status of the sample (p = 0.518) and the type of treatment (sunitinib vs. other regiments, p = 0.508). Overall, these analyses of covariance show that GAG measurements in the form of the proposed scores can robustly predict the occurrence of mccRCC despite baseline and treatment

differences across patients. This robustness is likely due to a coordinated regulation of GAG biosynthesis intrinsic to ccRCC progression, which is mirrored at the level of kidney-adjacent fluids.

Discussion

This study reveals that a coordinated regulation of GAG biosynthesis that features a concurrent up-regulation of the branch leading to CS formation and down-regulation of the branch leading to HS formation is a prominent transcriptional event in ccRCC. Also, many pathway-associated genes are further up- or down-regulated in metastasis. This discovery was enabled by leveraging on an increased number of samples and recent advances in metabolic network analysis: indeed, traditional gene-set enrichment analysis likely misses the distinctive regulation of the two branches within the gene-set, because the opposite fold-changes would cancel each other out. Altered expression of CS and HS, particularly in the glycan composition and sulfation, has been implicated in the promotion of migration, metastasis, and angiogenesis in a number of tumor models, including skin(Smetsers et al., 2004), lung(Mizumoto et al., 2012), brain(Wade et al., 2013), and breast(Fernandez-Vega et al., 2013); however, to our knowledge, this is the first report of an extensive, consistent, and coordinated regulation of the whole biological process of GAG biosynthesis in a cancer type. The relevance of such precise regulation of GAGs in ccRCC may be attributed to their role in the remodeling of the extracellular matrix that strongly depends on their composition and abundance(Afratis et al., 2012). For example, a chondroitin sulfate-rich matrix is linked to the development of self-contained and defined lesions in lower grade glioma (as opposed to microscopic infiltrations typical of glioblastomas)(Silver et al., 2013), a tumor growth model closely resembled by ccRCC(Rini et al., 2009). However, it remains to be explored how this regulatory program is mechanistically linked to metastasis rather than representing a coordinated metabolic event attributable to the remodeling of the kidney caused by the disease.

Since GAGs localize and act in the extracellular matrix, we assumed that changes in their regulation reflect changes in their profile in body fluids proximal to the kidney, e.g. blood and urine, as seen in other pathologies (Anower et al., 2013; Mannello et al., 2014; Mannello et al., 2015; Schmidt et al., 2014; Volpi et al., 2015). In particular, this behavior should be exacerbated in metastasis. Currently, there is no diagnostic biomarker that has entered routine practice for metastatic ccRCC (Jonasch et al., 2012; Moch et al., 2014). At the same time, the metastatic disease is invariably incurable, although rare complete responses were reported in association with oncological targeted therapies with or without metastasectomy (Albiges et al., 2012). Therefore it would undoubtedly represent an important clinical advancement if changes in the GAG profile could constitute an indicator of occurrence of the disease. The availability of such test would be valuable for a number of medical decisions: to monitor ccRCC before and after surgery or systemic treatment; to rule out the relapse of the disease also during a longer period of time after which a patient is typically declared cured; to assess the occurrence of ccRCC in a population at risk, such as genetically predisposed individuals; to ascertain whether a metastasis is due to ccRCC or other neoplasms; and to follow treatment response in mccRCC. In consideration of this, we designed three markers that are distinctive of occurrence of mccRCC, that are calculated based on measurements in accessible fluids, that are predictive of the clinical outcome independently of the here-considered confounding factors, and that, most importantly, are accurate and robust predictor of the disease.

The plasma and urine GAG profiles loosely resemble the expected pattern from the underlying transcriptional regulation, i.e. an increased output of CS with respect to HS in ccRCC. Noteworthy, a previous study examined CS/HS concentration in low tumor stage ccRCC tissue samples (Batista et al., 2012) and re-elaboration of this data to highlight the CS/HS ratio delineated a similar trend (Fig. S6). At the same time, the GAG profiles reveal some novel biological insights attributable to the occurrence of this cancer type. The GAG composition in the plasma of healthy individuals is typically not affected by any tissue. Here, we observed a systemic alteration of GAG composition concomitant to metastatic ccRCC. The enrichment of 4s- and 6s CS and 6s HS in mccRCC samples is strikingly reminiscent of the GAG composition of lymphocytes (Shao et al., 2013). It is therefore tempting to speculate that infiltration of the immune system in mccRCC could lie behind the observed transcriptional regulation in the tumor. In the urine, the GAG composition in healthy individuals is not as well characterized. The alterations here reported in the GAG profile of mccRCC samples might reflect a progressive

damaging of cells in the kidney glorumeli (McCarthy and Wassenhove-McCarthy, 2012; Miner et al., 2011). Collectively, these evidences seem to underscore the importance of alterations in GAGs in the progression of ccRCC.

So far, among the major difficulties that have impaired biomarker discovery and its translation in the clinical practice are the detection of targets in accessible samples and the reproducibility of results (Sawyers, 2008). Here, we provide evidence for a plasmatic and/or urinary marker of metastatic ccRCC that is supported by an intensely and consistently regulated biological process in ccRCC samples. We envision that future longitudinal studies based on our findings may establish these markers for a diverse range of diagnostic tools in the clinical management of ccRCC.

Experimental Procedures

Gene expression analysis. RNAseq gene expression profiles for 481 ccRCC primary tumor and 71 tumor-adjacent normal-like samples were retrieved at The Cancer Genome Atlas (TCGA) (Table S1). Differential expression analysis for ccRCC vs. normal was performed using voom(Law et al., 2014). 2090 genes with no annotation (3%) or no more than 10 counts in less than 10% of the samples (7%) were discarded. The effect of metastasis was accounted by adding the metastatic status of each sample as a covariate in the linear model used in voom. Two independent microarray-generated datasets where retrieved in GEO (GSE36895 (Pena-Llopis et al., 2012) and GSE14762 (Wang et al., 2009)) and the differential expression analysis for ccRCC vs. normal was performed using limma (Smyth, 2004). The significance for changes in gene expression using either RNAseq or microarray data was tested using empirical Bayes estimation on a linear model for a given comparison (in the case of RNAseq the count variance was moderated as proposed in voom (Law et al., 2014)). Consensus gene-set enrichment analysis (GSA) using piano (Varemo et al., 2013) was performed using as gene-sets either KEGG pathways or metabolites (i.e., a gene-set is the list of reaction-encoding genes that involve a given metabolite (Patil and Nielsen, 2005)), where the gene-set *p*-value is defined as the median *p*-value for the following GSA methods: Fisher's test, Stouffer's test, reporter test, tail-strength test, mean, and median. The significance of a gene-set for each GSA method was tested using a permutation test by shuffling gene labels 10'000 times. The gene-sets ranked among the top 30 by most GSA methods are shown in a heatmap that is hierarchically clustered. The differential gene expression analysis and multiple gene-set analysis (limited to KEGG pathways) was then repeated for six other cancer types (breast invasive carcinoma, colon adenocarcinoma, head and neck squamous cell carcinoma, lung adenocarcinoma, lung squamous cell carcinoma, and uterine corpus endometrial carcinoma) compared to matched tumor-adjacent normal samples (also retrieved at TCGA, Table S1). All analyzed cancer types were subsequently hierarchically clustered upon $\log_2$ fold-change in the expression of genes belonging to the KEGG glycosaminoglycan biosynthesis pathways (excluding genes belonging to dermatan sulfate biosynthesis and sulfotransferases on the C3 of heparan sulfate) compared to matched normal samples. Gene-set relatedness between gene-sets was computed in terms of the underlying network using Kiwi (Väremo et al., 2014), where gene-set were considered related if the mutual shortest path length is lesser than 2 in the network (to increase interpretability, gene-sets with more than 10 genes were neglected). In the case of metabolites, the gene-set network was extracted from the genome-scale metabolic model HMR2 (Agren et al., 2014). All the above methods were implemented using the respective R-packages, except Kiwi that is a Python module.

Immunohistochemical staining. A tissue microarray containing 32 ccRCC samples and 2 normal kidney samples in duplicates was prepared and used for immunohistochemistry. An experienced urological pathologist selected all cases. The ethical approval was granted by the ethical committee at Lund University (LU289-07). Tissue sections of 4μm were deparaffinized and rehydrated according to standard protocols. Antigen retrieval was performed using pressure cooking of the samples for 20 minutes in 10 mmol/L citrate buffer, pH 6.0. Immunohistochemical staining was performed using a Dako Techmate 500 unit, according to the manufacturer's instructions (Dako, Glostrup, Denmark). Antibodies and dilutions used were HPA020992 (CHP2 1:35), HPA034625 (HS6ST2 1:125) and HPA037749 (EXTL1 1:35), all from Atlas Antibodies AB, Stockholm, Sweden. Only tumor samples where both duplicates could be scored were included in the analysis (21 for CHP2, 32 for HS6ST2 and 27 for EXTL1).

Blood and urine analysis. In the discovery cohort, plasma and urine samples were obtained from 34 patients with metastatic clear cell renal carcinoma in two sites, IOV-IRCCS, Padova, Italy and Sahlgrenska University Hospital, Göteborg, Sweden. For 21 patients, only plasma samples were obtained. A control group was formed using 16 healthy individuals without any renal or liver malignancy, nor inflammatory pathologies. In the validation cohort, plasma and urine samples were obtained from 18 patients with metastatic clear cell renal carcinoma in two sites, IOV-IRCCS, Padova, Italy and Sahlgrenska University Hospital, Göteborg, Sweden. For 11 patients, only plasma samples were obtained. A control group was formed using 9 healthy individuals without any renal or liver malignancy, nor inflammatory pathologies. All subjects provided written informed consent. Clinical and dietary information is available in Table S3. The present observational study was notified to the Ethics Committee at IOV-IRCCS, Padova, Italy on January 2013. The approval to collect and analyze blood samples at the Sahlgrenska University Hospital, Göteborg, Sweden was obtained from the Regional Ethics Board of Västra Götaland, Sweden. Whole blood samples were collected in EDTA-coated tubes. The tubes were centrifuged (2,500*g* for 15 minutes at 4 °C) and the plasma extracted and collected in a separate tube. Urine were collected in polypropilene tubes. The samples were stored at -80°C until they were shipped for analysis in dry ice. The present study was carried out with compliance to the regulations of local Human Ethics Research Committee at the IOV-IRCCS, Padova, Italy. The samples were analyzed using HPLC with on-line electrospray ionization mass spectrometry (ESI-MS) as described in (Volpi et al., 2014; Volpi and Linhardt, 2010). Sixteen independent GAG properties were measured in each sample (either plasmatic or urinary): CS concentration, HS concentration, and fractions of disaccharide composition for both CS and HS. The charge is the sum over all sulfated disaccharide fractions. Principal component analysis was performed on available GAG properties for three cases: only plasmatic, only urinary, or both plasmatic and urinary (combined). Principal component analysis was implemented using R-package ade4 (Dray and Dufour, 2007) (centering was performed by the mean). All measurements are available in Table S3.

Marker design. To design the markers in the only plasmatic or in the only urinary case, we used Lasso penalized logistic regression (Tibshirani, 1996) with leave-one-out cross-validation to select those GAG properties that are most predictive of the clinical outcome (i.e. mccRCC vs. healthy) at the optimal Lasso penalty value. This was calculated using the glmnet R-package (Friedman et al., 2010) as the penalty value for which the cross-validation error was within 1 standard error of the minimum. The markers were built as the ratio between the sum of the GAG properties robustly predictive of mccRCC over the sum of the GAG properties robustly predictive of healthy state. Each property value was normalized using the respective regression coefficient (rounded to the nearest rational number). The marker for the combined case was taken as the mean of the so-designed plasmatic and urinary markers. The highest density interval (HDI) for the mean difference in marker scores between mccRCC vs. healthy was calculated using Bayesian estimation under the following assumptions: scores are sampled from a *t*-distribution of unknown and to be estimated normality (i.e. degrees of freedom); high uncertainty on the prior distributions; the marginal distribution is well approximated by a Markov chain Monte Carlo sampling with no thinning and chain length equal to 100'000. The estimation was performed using *BEST* (Kruschke, 2013) (the above assumptions are reflected by the default parameters). Bayesian estimation was preferred over the widely used *t*-test since it provides a robust and reliable estimation of mean difference even under uncertainty of the underlying score distribution for the two groups (that is the case when the number of samples is limited) (Nuzzo, 2014).

Accuracy metrics. For each marker (plasma, urine, or combined), we evaluated its performance in the binary classification of a sample as either mccRCC or healthy at varying threshold scores by deriving the receiver-operating-characteristic (ROC) curves. We measured the accuracy of each marker as the area under the curve (AUC) of its ROC curve (AUC is 1 for a perfect classifier and 0.5 for a random classifier). We selected as a potential cut-off value for a given marker the score for which the negative predictive value was maximum, i.e. a sample whose marker score is below this cut-off value has the maximum probability of not being mccRCC. We assumed a prevalence equal to the proportion of mccRCC samples in each cohort. The ROC curves were calculated using the pROC R-package (Robin et al., 2011), while the optimal cut-off using the OptimalCutpoints R-package (Lopez-Raton et al., 2014).

Analysis of covariance. The analysis of covariance was performed using logistic regression on the clinical outcome (mccRCC vs. control) on selected covariates among those reported in the clinical and dietary information in Table S3. These covariates were selected using Lasso penalized logistic regression with leave-one-out cross-validation as the most predictive of the clinical outcome at the optimal Lasso penalty value (chosen as described in Marker design). These covariates are age, weekly consumption of pasta and rice, and use of alcohol. Next, we performed logistic regression on the clinical outcome based on the combined score and the four selected covariates. The significance of each coefficient was tested using the Wald *z*-statistics for the hypothesis that the corresponding parameter is zero. The same procedure was followed to check the effect of systemic therapy as covariate, but using only plasma samples to regress the clinical outcome (since only for such sub-cohort there were enough patients that did not undergo any systemic therapy). In this case, either only one covariate was used to indicate the presence or absence of undergoing therapy, or a second covariate to account for the specific effect of sunitinib was added. Logistic regression was implemented adopting the Firth bias-reduction method using the brglm R-package. The performance of the two alternative models for logistic regression (either combined score + age + weekly consumption of pasta + weekly consumption of rice + use of alcohol; or combined score) was evaluated according to the minimum Kullback–Leibler divergence criterion by calculating the Akaike's information criterion (AIC) for the models and deriving the model probability in terms of AIC weights (Wagenmakers and Farrell, 2004).

Acknowledgments

The authors wish to thank M.A. Pinhal (Federal University of São Paulo) for valuable discussions, and S. Khoomrung (Chalmers University of Technology) for sample preparation. We thank the research nurses Cristina Magro and Orejeta Diamanti at the Medical Oncology Unit 1, Istituto Oncologico Veneto IOV and Kristina Welinder at the Sahlgrenska University Hospital for performing blood draws and storing samples. We also thank Dina Petranovic and Adil Mardinoglu for critically reviewing the article. This work was financially supported by the Kurt and Alice Wallenberg Foundation.

Author contributions

F.G. performed all computational analyses. M.J. constructed the ccRCC tissue microarray. H.N. and M.J. performed and evaluated the protein staining. U.B., A.R., M.M., U.S., and S.L. designed and supervised the blood and urine sampling. N.V. supervised the glycosaminoglycan analytical measurements. I.N., U.B., and J.N. supervised the study. F.G. and J.N. conceived and designed the study. F.G. wrote the manuscript. All authors edited and approved the manuscript in its final form.

References

Afratis, N., Gialeli, C., Nikitovic, D., Tsegenidis, T., Karousou, E., Theocharis, A.D., et al. (2012). Glycosaminoglycans: key players in cancer cell biology and treatment. Febs J 279, 1177-1197.

Agren, R., Mardinoglu, A., Asplund, A., Kampf, C., Uhlen, M., and Nielsen, J. (2014). Identification of anticancer drugs for hepatocellular carcinoma through personalized genome-scale metabolic modeling. Molecular systems biology 10, 721.

Albiges, L., Oudard, S., Negrier, S., Caty, A., Gravis, G., Joly, F., et al. (2012). Complete remission with tyrosine kinase inhibitors in renal cell carcinoma. Journal of clinical oncology : official journal of the American Society of Clinical Oncology 30, 482-487.

Anower, E.K.M.F., Matsumoto, K., Habuchi, H., Morita, H., Yokochi, T., Shimizu, K., et al. (2013). Glycosaminoglycans in the blood of hereditary multiple exostoses patients: Half reduction of heparan sulfate to chondroitin sulfate ratio and the possible diagnostic application. Glycobiology 23, 865-876.

Batista, L.T., Matos, L.L., Machado, L.R., Suarez, E.R., Theodoro, T.R., Martins, J.R., et al. (2012). Heparanase expression and glycosaminoglycans profile in renal cell carcinoma. International journal of urology : official journal of the Japanese Urological Association 19, 1036-1040.

Bordbar, A., Monk, J.M., King, Z.A., and Palsson, B.O. (2014). Constraint-based models predict metabolic and associated cellular functions. Nature reviews. Genetics 15, 107-120.

Creighton, C.J., Morgan, M., Gunaratne, P.H., Wheeler, D.A., Gibbs, R.A., Gordon Robertson, A., et al. (2013). Comprehensive molecular characterization of clear cell renal cell carcinoma. Nature.

Dray, S., and Dufour, A.B. (2007). The ade4 package: Implementing the duality diagram for ecologists. J Stat Softw 22, 1-20.

Ferlay, J., Shin, H.R., Bray, F., Forman, D., Mathers, C., and Parkin, D.M. (2010). Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008. International

journal of cancer. Journal international du cancer 127, 2893-2917.
Fernandez-Vega, I., Garcia, O., Crespo, A., Castanon, S., Menendez, P., Astudillo, A., et al. (2013). Specific genes involved in synthesis and editing of heparan sulfate proteoglycans show altered expression patterns in breast cancer. BMC cancer 13, 24.
Finley, D.S., Pantuck, A.J., and Belldegrun, A.S. (2011). Tumor biology and prognostic factors in renal cell carcinoma. The oncologist 16 Suppl 2, 4-13.
Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. J Stat Softw 33, 1-22.
Gatto, F., Miess, H., Schulze, A., and Nielsen, J. (2015). Flux balance analysis predicts essential genes in clear cell renal cell carcinoma metabolism. Scientific reports 5, 10738.
Gatto, F., Nookaew, I., and Nielsen, J. (2014). Chromosome 3p loss of heterozygosity is associated with a unique metabolic network in clear cell renal carcinoma. Proceedings of the National Academy of Sciences of the United States of America 111, E866-875.
Gupta, K., Miller, J.D., Li, J.Z., Russell, M.W., and Charbonneau, C. (2008). Epidemiologic and socioeconomic burden of metastatic renal cell carcinoma (mRCC): a literature review. Cancer treatment reviews 34, 193-205.
Jackson, R.L., Busch, S.J., and Cardin, A.D. (1991). Glycosaminoglycans: molecular properties, protein interactions, and role in physiological processes. Physiological reviews 71, 481-539.
Jerby, L., and Ruppin, E. (2012). Predicting Drug Targets and Biomarkers of Cancer via Genome-Scale Metabolic Modeling. Clinical Cancer Research 18, 5572-5584.
Jonasch, E., Futreal, P.A., Davis, I.J., Bailey, S.T., Kim, W.Y., Brugarolas, J., et al. (2012). State of the science: an update on renal cell carcinoma. Molecular cancer research : MCR 10, 859-880.
Kreuger, J., and Kjellen, L. (2012). Heparan sulfate biosynthesis: regulation and variability. The journal of histochemistry and cytochemistry : official journal of the Histochemistry Society 60, 898-907.
Kruschke, J.K. (2013). Bayesian estimation supersedes the t test. Journal of experimental psychology. General 142, 573-603.
Law, C.W., Chen, Y., Shi, W., and Smyth, G.K. (2014). Voom: precision weights unlock linear model analysis tools for RNA-seq read counts. Genome biology 15, R29.
Li, B., Qiu, B., Lee, D.S., Walton, Z.E., Ochocki, J.D., Mathew, L.K., et al. (2014). Fructose-1,6-bisphosphatase opposes renal carcinoma progression. Nature.
Lopez-Raton, M., Cadarso-Suarez, C., Rodriguez-Alvarez, M.X., and Gude-Sampedro, F. (2014). OptimalCutpoints: An R Package for Selecting Optimal Cutpoints in Diagnostic Tests. J Stat Softw 61, 1-36.
Mannello, F., Maccari, F., Ligi, D., Canale, M., Galeotti, F., and Volpi, N. (2014). Characterization of oversulfated chondroitin sulfate rich in 4,6-O-disulfated disaccharides in breast cyst fluids collected from human breast gross cysts. Cell biochemistry and function 32, 344-350.
Mannello, F., Maccari, F., Ligi, D., Santi, M., Gatto, F., Linhardt, R.J., et al. (2015). Breast cyst fluid heparan sulphate is distinctively N-sulphated depending on apocrine or flattened type. Cell biochemistry and function.
Mardinoglu, A., Gatto, F., and Nielsen, J. (2013). Genome-scale modeling of human metabolism - a systems biology approach. Biotechnology Journal.
Mardinoglu, A., and Nielsen, J. (2015). New paradigms for metabolic modeling of human cells. Current opinion in biotechnology 34C, 91-97.
McCarthy, K.J., and Wassenhove-McCarthy, D.J. (2012). The Glomerular Basement Membrane as a Model System to Study the Bioactivity of Heparan Sulfate Glycosaminoglycans. Microsc Microanal 18, 3-21.
Mikami, T., and Kitagawa, H. (2013). Biosynthesis and function of chondroitin sulfate. Biochimica et biophysica acta 1830, 4719-4733.
Miner, J.H., Towler, D., Chen, F., Kopan, R., and Hruska, K. (2011). Organogenesis of the kidney glomerulus Focus on the glomerular basement membrane. Organogenesis 7, 75-82.
Mizumoto, S., Takahashi, J., and Sugahara, K. (2012). Receptor for advanced glycation end products (RAGE) functions as receptor for specific sulfated glycosaminoglycans, and anti-RAGE antibody or sulfated glycosaminoglycans delivered in vivo inhibit pulmonary metastasis of tumor cells. The Journal of biological chemistry 287, 18985-18994.
Moch, H., Srigley, J., Delahunt, B., Montironi, R., Egevad, L., and Tan, P.H. (2014). Biomarkers in renal cancer. Virchows Archiv : an international journal of pathology 464, 359-365.
Motzer, R.J., Rini, B.I., McDermott, D.F., Redman, B.G., Kuzel, T.M., Harrison, M.R., et al. (2015). Nivolumab for Metastatic Renal Cell Carcinoma: Results of a Randomized Phase II Trial. Journal of clinical oncology : official journal of the American Society of Clinical Oncology 33, 1430-1437.
Nuzzo, R. (2014). Scientific method: statistical errors. Nature 506, 150-152.
Patil, K.R., and Nielsen, J. (2005). Uncovering transcriptional regulation of metabolism by using metabolic network topology. Proceedings of the National Academy of Sciences of the United States of America 102, 2685-2689.
Pena-Llopis, S., Vega-Rubin-de-Celis, S., Liao, A., Leng, N., Pavia-Jimenez, A., Wang, S., et al. (2012). BAP1 loss defines a new class of renal cell carcinoma. Nature genetics 44, 751-759.

Rini, B.I., Campbell, S.C., and Escudier, B. (2009). Renal cell carcinoma. Lancet 373, 1119-1132.
Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.C., et al. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. Bmc Bioinformatics 12, 77.
Sawyers, C.L. (2008). The cancer biomarker problem. Nature 452, 548-552.
Schmidt, E.P., Li, G.Y., Li, L.Y., Fu, L., Yang, Y.M., Overdier, K.H., et al. (2014). The Circulating Glycosaminoglycan Signature of Respiratory Failure in Critically Ill Adults. Journal of Biological Chemistry 289, 8194-8202.
Schulze, A., and Harris, A.L. (2013). How cancer metabolism is tuned for proliferation and vulnerable to disruption (vol 491, pg 364, 2012). Nature 494, 130-130.
Shao, C., Shi, X.F., White, M., Huang, Y., Hartshorn, K., and Zaia, J. (2013). Comparative glycomics of leukocyte glycosaminoglycans. Febs J 280, 2447-2461.
Silbert, J.E., and Sugumaran, G. (2002). Biosynthesis of chondroitin/dermatan sulfate. IUBMB life 54, 177-186.
Silver, D.J., Siebzehnrubl, F.A., Schildts, M.J., Yachnis, A.T., Smith, G.M., Smith, A.A., et al. (2013). Chondroitin sulfate proteoglycans potently inhibit invasion and serve as a central organizer of the brain tumor microenvironment. The Journal of neuroscience : the official journal of the Society for Neuroscience 33, 15603-15617.
Smetsers, T.F., van de Westerlo, E.M., ten Dam, G.B., Overes, I.M., Schalkwijk, J., van Muijen, G.N., et al. (2004). Human single-chain antibodies reactive with native chondroitin sulfate detect chondroitin sulfate alterations in melanoma and psoriasis. The Journal of investigative dermatology 122, 707-716.
Smyth, G.K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Statistical applications in genetics and molecular biology 3, Article3.
Thacker, B.E., Xu, D., Lawrence, R., and Esko, J.D. (2014). Heparan sulfate 3-O-sulfation: a rare modification in search of a function. Matrix biology : journal of the International Society for Matrix Biology 35, 60-72.
Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. J Roy Stat Soc B Met 58, 267-288.
Vander Heiden, M.G., Cantley, L.C., and Thompson, C.B. (2009). Understanding the Warburg effect: the metabolic requirements of cell proliferation. Science 324, 1029-1033.
Varemo, L., Gatto, F., and Nielsen, J. (2014). Kiwi: a tool for integration and visualization of network topology and gene-set analysis. Bmc Bioinformatics 15, 408.
Väremo, L., Gatto, F., and Nielsen, J. (2014). Kiwi: a tool for integration and visualization of network topology and gene set analysis. (Submitted).
Varemo, L., Nielsen, J., and Nookaew, I. (2013). Enriching the gene set analysis of genome-wide data by incorporating directionality of gene expression and combining statistical hypotheses and methods. Nucleic acids research 41, 4378-4391.
Volpi, N., Coppa, G.V., Zampini, L., Maccari, F., Galeotti, F., Garavelli, L., et al. (2015). Plasmatic and urinary glycosaminoglycan profile in a patient affected by multiple sulfatase deficiency. Clinical Chemistry and Laboratory Medicine 53, E157-E160.
Volpi, N., Galeotti, F., Yang, B., and Linhardt, R.J. (2014). Analysis of glycosaminoglycan-derived, precolumn, 2-aminoacridone-labeled disaccharides with LC-fluorescence and LC-MS detection. Nature protocols 9, 541-558.
Volpi, N., and Linhardt, R.J. (2010). High-performance liquid chromatography-mass spectrometry for mapping and sequencing glycosaminoglycan-derived oligosaccharides. Nature protocols 5, 993-1004.
Wade, A., Robinson, A.E., Engler, J.R., Petritsch, C., James, C.D., and Phillips, J.J. (2013). Proteoglycans and their roles in brain cancer. Febs J.
Wagenmakers, E.-J., and Farrell, S. (2004). AIC model selection using Akaike weights. Psychonomic Bulletin & Review 11, 192-196.
Wahlgren, T., Harmenberg, U., Sandstrom, P., Lundstam, S., Kowalski, J., Jakobsson, M., et al. (2013). Treatment and overall survival in renal cell carcinoma: a Swedish population-based study (2000-2008). British journal of cancer 108, 1541-1549.
Wang, Y., Roche, O., Yan, M.S., Finak, G., Evans, A.J., Metcalf, J.L., et al. (2009). Regulation of endocytosis via the oxygen-sensing pathway. Nature medicine 15, 319-324.
Ward, P.S., and Thompson, C.B. (2012). Metabolic reprogramming: a cancer hallmark even warburg did not anticipate. Cancer cell 21, 297-308.

PAPER V

Kiwi: a tool for integration and visualization of network topology and gene-set analysis

L. Varemo*, F. Gatto*, J. Nielsen

BMC Bioinformatics **15**, 408 (2014)

*Authors contributed equally to this work

Väremo et al. BMC Bioinformatics 2014, 15:408
http://www.biomedcentral.com/1471-2105/15/408

BMC Bioinformatics

SOFTWARE Open Access

Kiwi: a tool for integration and visualization of network topology and gene-set analysis

Leif Väremo†, Francesco Gatto† and Jens Nielsen*

Abstract

Background: The analysis of high-throughput data in biology is aided by integrative approaches such as gene-set analysis. Gene-sets can represent well-defined biological entities (e.g. metabolites) that interact in networks (e.g. metabolic networks), to exert their function within the cell. Data interpretation can benefit from incorporating the underlying network, but there are currently no optimal methods that link gene-set analysis and network structures.

Results: Here we present Kiwi, a new tool that processes output data from gene-set analysis and integrates them with a network structure such that the inherent connectivity between gene-sets, i.e. not simply the gene overlap, becomes apparent. In two case studies, we demonstrate that standard gene-set analysis points at metabolites regulated in the interrogated condition. Nevertheless, only the integration of the interactions between these metabolites provides an extra layer of information that highlights how they are tightly connected in the metabolic network.

Conclusions: Kiwi is a tool that enhances interpretability of high-throughput data. It allows the users not only to discover a list of significant entities or processes as in gene-set analysis, but also to visualize whether these entities or processes are isolated or connected by means of their biological interaction. Kiwi is available as a Python package at http://www.sysbio.se/kiwi and an online tool in the BioMet Toolbox at http://www.biomet-toolbox.org.

Keywords: Gene-set analysis, Transcriptomics, Network analysis, Visualization tool

Background

Gene-set analysis (GSA) is a widely used category of bioinformatics methods and there are many available tools that perform GSA [1,2]. In GSA, genes known to contribute to a certain function, or share a relevant biological feature, are collected into sets. If these gene-sets are enriched by transcriptome or other high-throughput data, GSA directly highlights the most prominent among these sets, and thereby the underlying functions that are implicated by the data [2]. Networks stand at the basis of complex biological systems [3] and in many cases gene-sets represent elements that are connected, not simply because of gene overlap, but rather to exert a coordinated function through their interactions (the gene-set interaction network). Examples of elements that can be used as gene-sets and where an interaction network can be defined include: transcription factors in a gene regulatory network [4]; the hierarchical network of Gene Ontology terms [5]; and metabolite gene-sets in a metabolic network [6]. In particular the last example provides a very useful case since metabolite gene-sets (genes that are associated to reactions in which the metabolite takes part in) are connected through reaction pathways, but will usually not share any common genes (unless they participate in the same reaction). Thus, when several metabolite gene-sets in a pathway are significant their important biological connection will be lost, unless the gene-set interaction network is taken into account.

With this in mind, interpretation and visualization of the results from a GSA currently suffers from several limitations. Typically, the results are presented as a list of the most significant gene-sets, or visualized in a heatmap where gene-sets are clustered according to either the pattern of significance across several conditions or their direction of regulation. In both cases, the biologically relevant connections between gene-sets, defined by their interaction network, are ignored. Multiple connected significant gene-sets will likely represent an important biological process, but with the current visualization

* Correspondence: nielsenj@chalmers.se
†Equal contributors
Department of Chemical and Biological Engineering, Chalmers University of Technology, Gothenburg 412 96, Sweden

approaches these connections are lost and are tedious to elucidate manually.

On the other hand, it is not unusual to see GSA results presented as networks, with nodes representing the most significant gene-sets [1,7-9]. However, in these cases edges between nodes simply represent gene overlap. This can help to reduce the bias from redundant gene-sets by clustering gene-sets with overlapping gene content together. Nevertheless, a network visualization approach where the edges represent gene-set interactions is advantageous in the context of biological interpretation. Indeed, different tools can be used to visualize data on gene-set interaction networks [10-14], although some of them are not specifically made for that purpose. Unfortunately, these tools suffer from one or several of the following drawbacks:

- The tool is not made specifically to handle GSA data, which requires the user to tweak the input (e.g. common identifiers and color-coding scheme) in the best way possible to fit the framework of that tool.
- The tool is only made for a specific type of network (e.g. KEGG pathways or GO-terms), constraining the user to only one single gene-set type.
- The tool is not effectively reducing the network to highlight the significant results, but instead simply overlaying the data on the original, and potentially huge, gene-set interaction network.

Here we address the current limitations by developing a new network-based visualization approach and implement it in the software tool Kiwi. Contrary to other available tools, Kiwi explicitly embraces the paradigm that gene-sets can be biological entities that interact and it therefore aims at visualizing GSA results in the context of the gene-set interaction network in such way that the biological connections between all significant gene-sets become apparent. This is done by taking into account both the directionality and significance of the gene-sets and by removing non-interesting gene-sets from the visualized network. Further on, Kiwi is made as general as possible, in the sense that it accepts input from any GSA tool and any gene-set interaction network defined by the user. Finally, since the biological measurements behind the data are made at the gene-level, Kiwi enables the user to go from the visualization network of significant gene-sets back to the gene-level data, in order to detect driver genes behind the regulated biological elements that the gene-sets represent.

Implementation

Input data

The input to Kiwi is at minimum the gene-set interaction network and a table of p-values for the gene-sets, which can be collected from the output of any GSA tool. Apart from this, it is recommended to also supply the gene members of the gene-sets as well as the gene-level statistics (e.g. p-values and fold-changes) that were used as input to the GSA. Full details and required format for the input files can be found in the online Kiwi reference manual.

Processing

An outline of the network visualization process performed by Kiwi is shown in Figure 1. First, non-significant gene-sets are filtered out according to a user-set cutoff. The remaining gene-sets are used as nodes in a new visualization network. In this visualization network the edges between gene-sets should reflect how closely they interact. The shortest path length (SPL) measures the shortest distance between two gene-sets and is a property of the network that indicates whether the two gene-sets are interacting directly or indirectly via a certain number of intermediates. Hence, the SPL between all pair of nodes in the gene-set interaction network is calculated. If the SPL between two gene-set nodes is below a user-set cutoff an edge is drawn between those nodes, with an edge thickness relative to the SPL. The SPL cutoff can be seen as a measure of the relatedness of two gene-sets in the gene-set interaction network, and it controls at what distance these gene-sets should not any longer be considered biologically connected. For each node, only the edge or edges with the lowest SPL are kept, so that each node is connected only to its closest nodes of those present in the visualization network. Finally, the visualization network is drawn using a force-based layout. Nodes are resized to reflect the gene-set significance and color-coded to capture the general direction of change of the genes in the set (refer to the online documentation for further details).

Output

Kiwi produces two figures: a network and a heatmap. The network presents an uncluttered view where the most important features are highlighted. The node sizes and color-codes are adjusted according to the gene-set significance and general direction of change. The heatmap serves as a complement to the network by displaying the gene-level statistics for each gene-set in the network. The rows (gene-sets) and columns (genes) are hierarchically clustered, which enables the identification of (i) gene-sets with similar gene content and (ii) the significant genes that are driving the observed changes. Both figures can be fine-tuned by the user through several parameters and the network can also be saved in graphML format and imported into Cytoscape for further customization.

Case studies

To illustrate the advantages of Kiwi, we use two case studies. The first one is based on a differential gene expression dataset from lung adenocarcinoma vs. normal

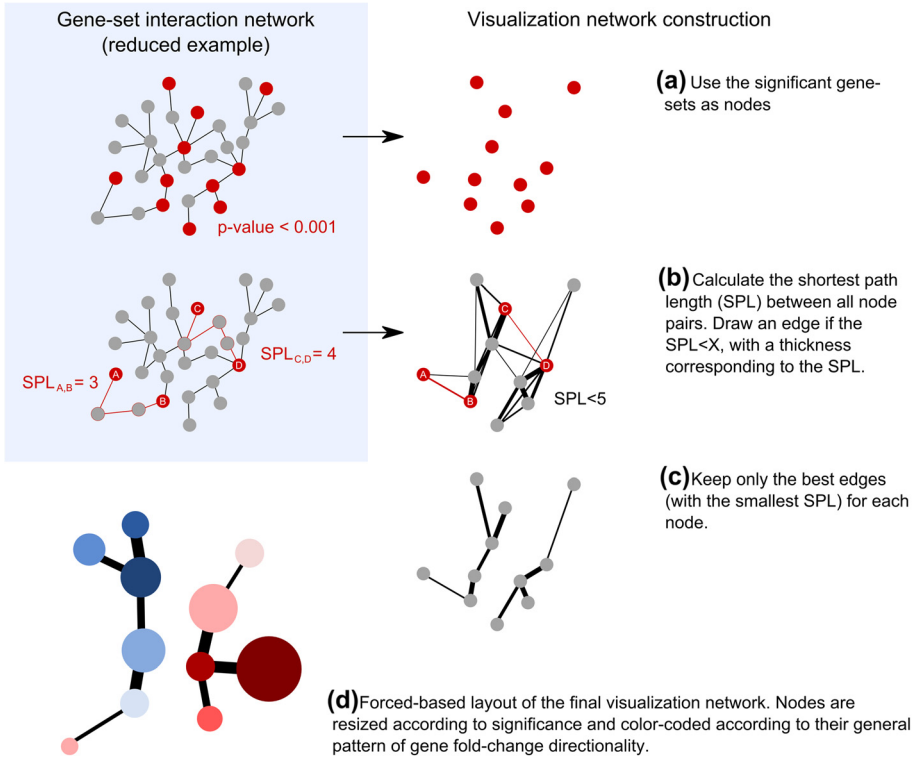


Figure 1 Overview of the Kiwi workflow. (a) Significant gene-sets are selected based on a user-set cutoff and used as nodes in the visualization network. **(b)** The shortest path length (SPL) between all node pairs in the gene-set interaction network is calculated. In the example, the SPL between node A and B is 3, and between node C and D is 4. If the SPL between two nodes is below a user-set cutoff (5 in the example), an edge is drawn between those nodes, with a thickness corresponding to the SPL (an SPL = 1 will generate the thickest edge). In the example, the edges between nodes A and B, and C and D, respectively, are marked in red, corresponding to the SPLs shown in the gene-set interaction network. **(c)** To avoid a cluttered network with too many edges, only the best edges (with lowest SPL) are kept for each node. Note that a node may still have multiple edges of different thickness if, for example, a thinner edge is the best one of a neighbouring node. (This step is optional.) **(d)** Finally, the visualization network is drawn using a forced-based layout. Nodes are resized according to the gene-set significance and color-coded in order to reflect the general direction of change of the gene-set.

lung tissue [15]. Metabolites from a human genome-scale metabolic model [16] were used as gene-sets and the GSA was carried out using the Bioconductor R-package piano [1].

For the second case study we used gene expression data from a study on Kras conditional activation in mouse xenograft tumors [17]. Metabolites from a mouse genome-scale metabolic model, derived from the human genome-scale metabolic model used in case study 1, using gene homology as described in [18], were used as gene-sets. The GSA was carried out using the Bioconductor R-package piano.

Kiwi version 0.2.8 was used for both case studies. The heatmaps and network plots shown in Figure 2a,d and Figure 3b,c are the direct output from Kiwi, however, to provide as clear of a figure as possible, the node labels in the networks have been manually shifted. The data and scripts for running these case studies are available as Additional file 1.

Results and discussion

In order to show the advantages, in terms of biological interpretation, of using Kiwi to visualize GSA results in the context of a gene-set interaction network, we performed two case studies. In both cases we used a genome-scale metabolic model to define a metabolite-metabolite network (connecting metabolites if they are substrates or products of the same reaction). A metabolite gene-set is defined by the group of genes that are associated with reactions in which the metabolite participates in.

Metabolic changes associated with lung adenocarcinoma transformation

To illustrate the benefits of exploiting the gene-set interaction network, compared to only considering the gene overlap, we re-analysed a differential gene expression dataset from lung adenocarcinoma vs. normal lung tissue [15]. Metabolites from the human genome-scale metabolic

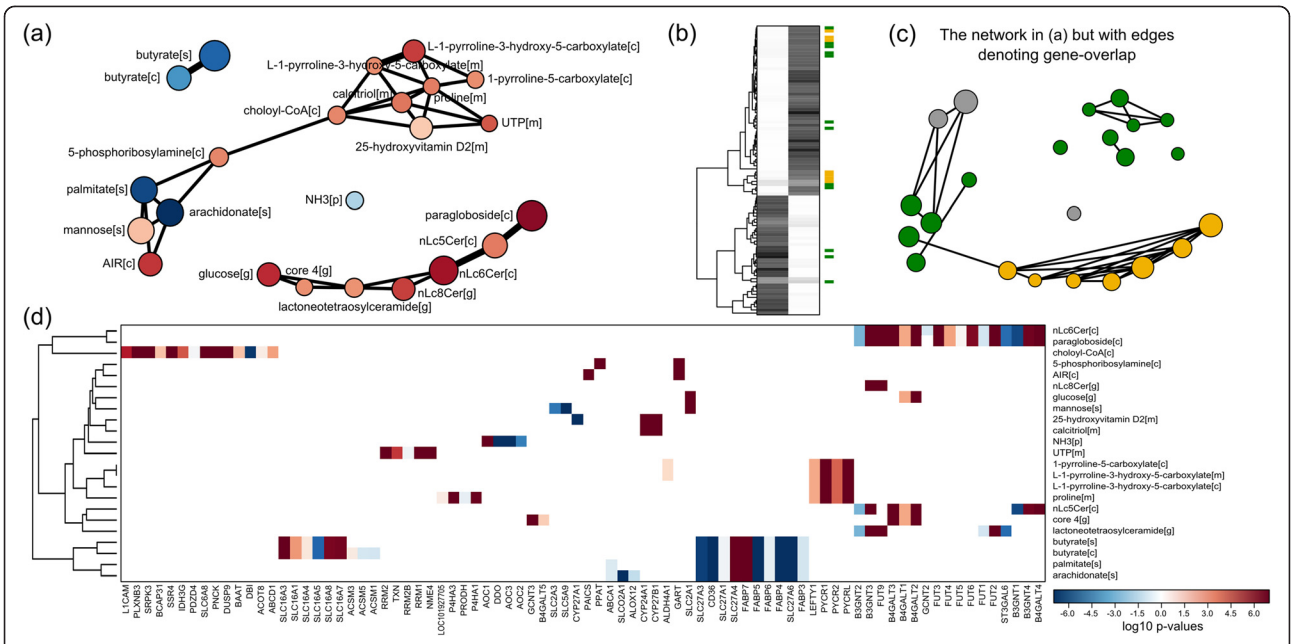


Figure 2 The output from Kiwi in case study 1, metabolic changes upon transformation to lung adenocarcinoma. (a) The Kiwi network shows the significant gene-sets and their general direction of change. Up-regulation: red, down-regulation: blue. Two metabolically connected pathways emerge, a pattern that is not easily detected in **(b)** a traditional heatmap or **(c)** a gene-overlap network. Green and orange indicate gene-sets of the two pathways, respectively. **(d)** The Kiwi heatmap displays the individual gene-level statistics for each significant gene-set. It also clusters similar gene-sets based on gene overlap in order to identify driver genes or potential bias.

model HMR2 [16] were used as gene-sets (i.e. genes associated with reactions in which a specific metabolite participates) and the GSA was carried out using the Bioconductor R-package piano [1], which produces files that can be directly imported by Kiwi. The Kiwi network (Figure 2a) clearly identifies significant gene-sets composing two metabolically connected pathways. For example, 5-phosphoribosylamine and 1-pyrroline-5-carboxylate both participate in pyrimidine biosynthesis, but their relatedness becomes apparent if the underlying metabolic network that measures the mutual distance is considered. These important connections are lost when the results are presented as a traditional heatmap (Figure 2b) or a network based on overlap of gene members of the different gene-sets (Figure 2c). The Kiwi heatmap (Figure 2d) shows the gene-level transcriptional changes for each gene-set enabling the identification of interacting gene-sets without gene overlap, and their driver-genes. For example, 5-phosphoribosylamine is a significant gene-set because of *GART* and *PPAT* up-regulation, while 1-pyrroline-5-carboxylate is significant due to *LEFTY1* and *PYCR* up-regulation. The heatmap also simplifies the detection of similar gene-sets, as e.g. nLc6Cer[c] and paragloboside[c].

Metabolic changes associated with activation of oncogenic Kras in mouse tumor xenografts

Using a second case study we sought to test if Kiwi is able to reproduce networks known to be informative in a certain condition. To this end, we re-analyzed gene expression data from a study where the oncoprotein Kras was conditionally activated in mouse xenograft tumors [17]. The authors showed that activation of oncogenic Kras entails extensive metabolic reprogramming, in particular up-regulation of steroid biosynthesis. We therefore performed GSA [1] in the context of a mouse genome-scale metabolic network (Figure 3a) and tested if Kiwi could capture the relevant network of gene-sets upon Kras activation. In line with the results in the aforementioned study, we observe the emergence of the steroid biosynthetic pathway, which is overexpressed in different steps (Figure 3b). Indeed, despite the fact that isopentenyl-pPP, 14-demethyllanosterol, squalene, and lanosterol are not overlapping gene-sets (as shown by the heatmap in Figure 3c), Kiwi relates the metabolites given their vicinity in the underlying mouse metabolic network. Notably, contrary to the gene-set enrichment analysis used by the authors, Kiwi also identifies which pathway among the different branches of steroid biosynthesis is truly up-regulated by Kras activation, namely lanosterol synthesis.

Conclusions

Kiwi is a new tool tailored for the visualization of GSA results in a gene-set interaction network context. As opposed to available tools, Kiwi starts from the premise that gene-sets can be precise biological entities that achieve a certain function by means of their interactions,

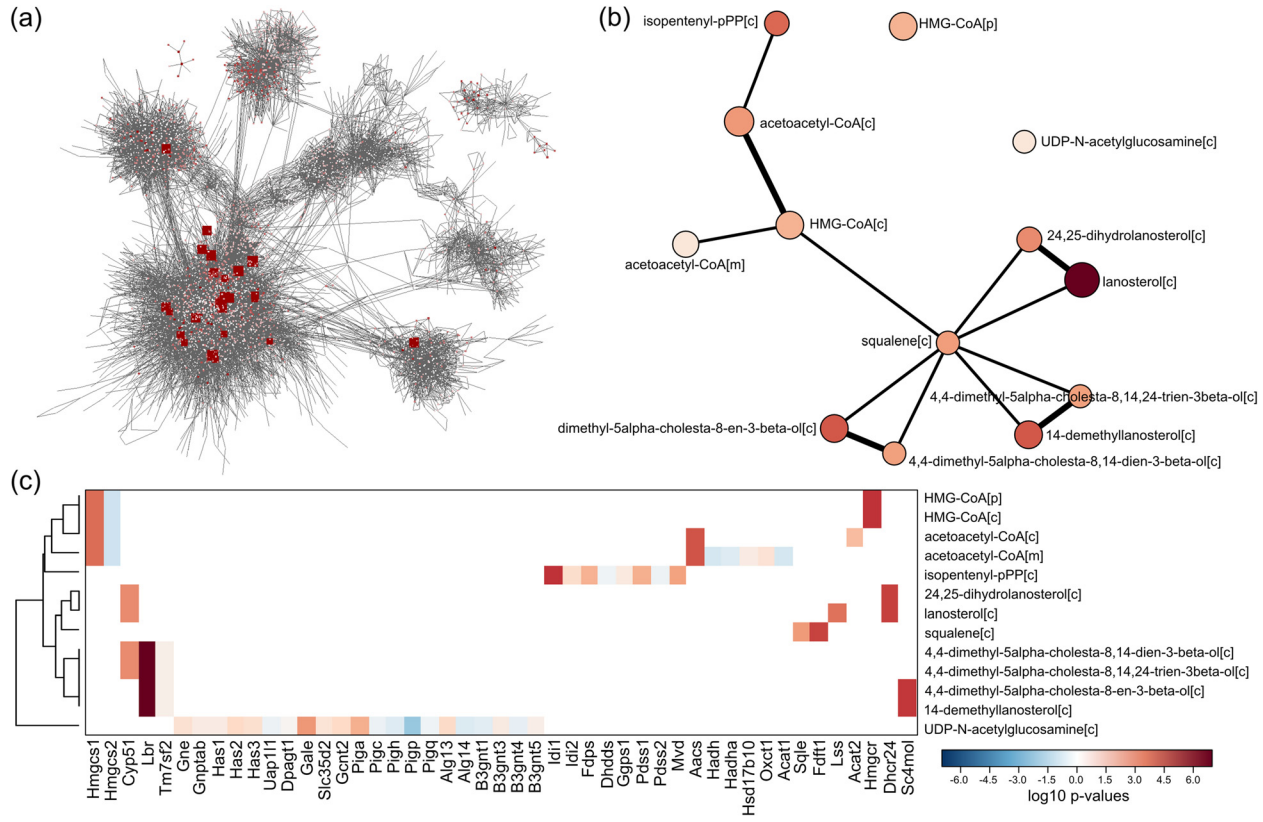


Figure 3 The output from Kiwi in case study 2, metabolic changes upon activation of Kras in mouse tumors. (a) The significant gene-sets overlayed on the metabolite-metabolite network (the full gene-set interaction network). As the network is too big and the gene-sets are too spread around the network it is not possible, in a simple way, to draw any biological conclusions about the data. **(b)** Kiwi effectively reduces the gene-set interaction network and pulls out the significant gene-sets, while maintaining their biological relatedness given in the original network. Here, Kiwi outputs a network showing the steroid biosynthetic pathway, in line with the original study. **(c)** The heatmap for the gene-sets in the Kiwi network shows that isopentenyl-pPP, 14-demethyllanosterol, squalene, and lanosterol are not overlapping in terms of gene members, yet they are connected in the Kiwi network.

such as metabolites in a pathway. This paradigm significantly improves the interpretation of the effect of transcriptional regulation in a certain context, such as metabolism, because it adds an extra layer of information to the GSA results. As exemplified in the two case studies, such addition is fundamental to capture certain transcriptionally regulated processes. In the case of the transformation to lung adenocarcinoma, we observe that the up-regulation of pyrimidine biosynthesis is mediated by the connection provided by choloyl-CoA. In the case of oncogenic Kras activation in mouse tumors, not only do we reproduce the up-regulation of the steroid biosynthetic process, but we also report that this is ascribed mainly to the synthesis of lanosterol. In neither case could such results be highlighted by connecting gene-sets using gene overlap (see Figure 2c) or by overlaying the GSA results on the corresponding gene-set interaction network (see Figure 3a). In favour of a clean layout for enhanced interpretation, Kiwi reduces the gene-set interaction network while maintaining and highlighting the important gene-set connections. It works with the output from any GSA tool and any collection of gene-sets that can be described as a network. For full usability, from raw data to final figure, it integrates seamlessly with the Bioconductor R-package piano (for GSA) and Cytoscape (for advanced layout and customization). Kiwi is available as a Python package at http://www.sysbio.se/kiwi and an online tool in the BioMet Toolbox at http://www.biomet-toolbox.org [19].

Availability and requirements

Project name: Kiwi
Project home page: www.sysbio.se/kiwi
Operating system(s): Platform independent
Programming language: Python
Other requirements: Kiwi depends on the following python packages: numpy > = 1.8.0; matplotlib > = 1.3.1; networkx > = 1.8.1; mygene > = 2.1.0; pandas > = 0.13.1; scipy > = 0.13.3.
License: MIT
Any restrictions to use by non-academics: None

Additional file

Additional file 1: Case study demo files.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
LV and FG wrote the code and developed the software. LV drafted the manuscript. FG carried out the case studies and wrote corresponding parts of the manuscript. JN supervised the project. All authors read, edited and approved the final manuscript.

Acknowledgements
The authors would like to thank Adil Mardinoglu for reconstructing the mouse genome-scale metabolic model and Subazini Thankaswamy for including Kiwi in the BioMet Toolbox. This work was funded by Knut and Alice Wallenberg foundation, and Chalmers foundation.

Received: 1 September 2014 Accepted: 3 December 2014
Published online: 11 December 2014

References
1. Väremo L, Nielsen J, Nookaew I: **Enriching the gene set analysis of genome-wide data by incorporating directionality of gene expression and combining statistical hypotheses and methods.** *Nucleic Acids Res* 2013, **41**(8):4378–4391.
2. Hung JH, Yang TH, Hu Z, Weng Z, Delisi C: **Gene set enrichment analysis: performance evaluation and usage guidelines.** *Briefings Bioinform* 2012, **13**(3):281–291.
3. Barabasi A-L, Oltvai ZN: **Network biology: understanding the cell's functional organization.** *Nat Rev Genet* 2004, **5**(2):101–113.
4. Oliveira AP, Patil KR, Nielsen J: **Architecture of transcriptional regulatory circuits is knitted over the topology of bio-molecular interaction networks.** *BMC Syst Biol* 2008, **2**:17.
5. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA: **Gene Ontology: tool for the unification of biology.** *Nat Genet* 2000, **25**(1):25.
6. Patil KR, Nielsen J: **Uncovering transcriptional regulation of metabolism by using metabolic network topology.** *Proc Natl Acad Sci U S A* 2005, **102**(8):2685–2689.
7. Chen E, Tan C, Kou Y, Duan Q, Wang Z, Meirelles G, Clark N, Ma'ayan A: **Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool.** *BMC Bioinf* 2013, **14**(1):128.
8. Merico D, Isserlin R, Stueker O, Emili A, Bader GD: **Enrichment map: a network-based method for gene-set enrichment visualization and interpretation.** *PLoS One* 2010, **5**(11):e13984.
9. Wang X, Terfve C, Rose JC, Markowetz F: **HTSanalyzeR: an R/Bioconductor package for integrated network analysis of high-throughput screens.** *Bioinformatics* 2011, **27**(6):879–880.
10. Eden E, Navon R, Steinfeld I, Lipson D, Yakhini Z: **GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists.** *BMC Bioinf* 2009, **10**(1):48.
11. Yamada T, Letunic I, Okuda S, Kanehisa M, Bork P: **iPath2.0: interactive pathway explorer.** *Nucleic Acids Res* 2011, **39**(suppl 2):W412–W415.
12. Luo W, Brouwer C: **Pathview: an R/Bioconductor package for pathway-based data integration and visualization.** *Bioinformatics* 2013, **29**(14):1830–1831.
13. Al-Shahrour F, Minguez P, Tárraga J, Montaner D, Alloza E, Vaquerizas JM, Conde L, Blaschke C, Vera J, Dopazo J: **BABELOMICS: a systems biology perspective in the functional annotation of genome-scale experiments.** *Nucleic Acids Res* 2006, **34**(suppl 2):W472–W476.
14. Bates JT, Chivian D, Arkin AP: **GLAMM: Genome-Linked Application for Metabolic Maps.** *Nucleic Acids Res* 2011, **39**(suppl 2):W400–W405.
15. Gatto F, Nookaew I, Nielsen J: **Chromosome 3p loss of heterozygosity is associated with a unique metabolic network in clear cell renal carcinoma.** *Proc Natl Acad Sci U S A* 2014, **111**(9):E866–E875.
16. Mardinoglu A, Agren R, Kampf C, Asplund A, Uhlen M, Nielsen J: **Genome-scale metabolic modelling of hepatocytes reveals serine deficiency in patients with non-alcoholic fatty liver disease.** *Nat Commun* 2014, **5**:3083.
17. Ying H, Kimmelman Alec C, Lyssiotis Costas A, Hua S, Chu Gerald C, Fletcher-Sananikone E, Locasale Jason W, Son J, Zhang H, Coloff Jonathan L, Yan H, Wang W, Chen S, Viale A, Zheng H, J-h P, Lim C, Guimaraes Alexander R, Martin Eric S, Chang J, Hezel Aram F, Perry Samuel R, Hu J, Gan B, Xiao Y, Asara John M, Weissleder R, Wang YA, Chin L, Cantley Lewis C, *et al*: **Oncogenic Kras Maintains Pancreatic Tumors through Regulation of Anabolic Glucose Metabolism.** *Cell* 2012, **149**(3):656–670.
18. Sigurdsson M, Jamshidi N, Steingrimsson E, Thiele I, Palsson B: **A detailed genome-wide reconstruction of mouse metabolism based on human Recon 1.** *BMC Syst Biol* 2010, **4**(1):140.
19. Garcia-Albornoz M, Thankaswamy-Kosalai S, Nilsson A, Väremo L, Nookaew I, Nielsen J: **BioMet Toolbox 2.0: genome-wide analysis of metabolism and omics data.** *Nucleic Acids Res* 2014, **42**(Web Server issue):W175–W181.

doi:10.1186/s12859-014-0408-9
Cite this article as: Väremo *et al.*: **Kiwi: a tool for integration and visualization of network topology and gene-set analysis.** *BMC Bioinformatics* 2014 **15**:408.

PAPER VI

In search for symmetries in cancer metabolism

F. Gatto, J. Nielsen

Submitted for publication

In search for symmetries in the metabolism of cancer

Francesco Gatto[1] and Jens Nielsen[1].

Affiliations:

[1] Department of Biology and Biological Engineering, Chalmers University of Technology, Göteborg, Sweden.

Contacts: To whom correspondence should be addressed: J. Nielsen nielsenj@chalmers.se

SUMMARY

Even though aerobic glycolysis, or the Warburg effect, is arguably the most common trait of metabolic reprogramming in cancer, it is unobserved in certain tumor types. Systems biology advocates a global view on metabolism to dissect which traits are consistently reprogrammed in cancer, and hence likely to constitute an obligate step for the evolution of cancer cells. We refer to such traits as symmetrical. Here, we review early systems biology studies that attempted to reveal symmetrical traits in the metabolic reprogramming of cancer, discuss the symmetry of reprogramming of nucleotide metabolism, and outline the current limitations that, if unlocked, could elucidate whether symmetries in cancer metabolism may be claimed.

Introduction

The modern theory on the origin of cancer prescribes that every tumor is a unique experiment of nature, in which mutations in key genes (driver mutations) together with accumulation of mutations in secondary genes (passenger mutations) define the fitness of a cancer clone in its strive to survive and proliferate in the host environment, the human body (Gerlinger et al., 2014; Vogelstein et al., 2013). At the time of writing, COSMIC (Catalogue Of Somatic Mutations In Cancer), a manually curated database for cancer mutations, lists 572 genes in which a mutation is causally implicated in some form of cancer (Forbes et al., 2015). Thanks to significant advances in the DNA-sequencing technology, in particular the advent of next-generation sequencing, it has been recently estimated that even though we have possibly discovered all the most recurrently mutated genes, the number of genes that are rarely mutated (yet potentially implicated in cancer) will likely continue to rise in the future (Lawrence et al., 2014). In addition, the heterogeneity of mutated genes can encompass both localized and distal tumors in the same patient (Gerlinger et al., 2012; Johnson et al., 2014; Nik-Zainal et al., 2012).

Despite this heterogeneity, some characteristics are clearly shared by all cancers. Phenotypic traits as aberrant proliferation and invasion are observed in virtually all cancer types (Evan and Vousden, 2001). In a landmark review in 2001, Hananah and Weinberg described six phenotypic traits that all cancers seem to acquire, which they called the hallmarks of cancer (Hanahan and Weinberg, 2000, 2011). The origin of cancer itself should explain the convergence on these phenotypic traits, in that the number of mutations that initiate cancer is higher than the number of pathways altered by the mutations, but all these mutations are ultimately responsible for the acquisition of the hallmarks (Vogelstein and Kinzler, 2004; Vogelstein et al., 2013). We will refer to a specific phenotypic trait as symmetric if convergent evolution has been observed. This term is, reminiscent of the fact that any two cancers as different as they may appear may always be repositioned to look symmetric under that specific phenotypic trait. In principle, a symmetric trait should be distinctive of the disease.

Different lines of evidence recognized a hallmark status to the reprogramming of metabolism in cancer. At the genetic level, two studies published in the early NGS era reported an unprecedented mutation in the cytosolic NADP+-dependent isocitrate dehydrogenase 1 gene (*IDH1*) in 12% of glioblastoma multiforme patients and in 8% of acute myeloid leukemia (Mardis et al., 2009; Parsons et al., 2008). *IDH1* encodes for an enzyme responsible for catalysis of a reaction in central carbon metabolism that converts isocitrate to 2-oxoglutarate. The discovery of *IDH1* mutations provided a direct connection between the origin of cancer and deregulation of metabolism. At the metabolic level, the first report that cancers may reprogram metabolic fluxes to match different requirements for proliferations is historically attributed to Otto Warburg (Warburg, 1956a). Warburg noted that, even in the presence of oxygen, cultured cancer cells prefer to metabolize glucose to lactic acid, rather than completely oxidize it through the tricarboxylic acid (TCA) cycle, which is more favorable in terms of ATP yield (a phenomenon dubbed aerobic glycolysis or the Warburg effect) (Warburg, 1956b). Accumulated evidence at both the genetic and metabolic level led researchers to conclude that cancer reprograms its metabolism as part of the transformation to facilitate cell proliferation and survival (Cairns et al., 2011; Schulze and Harris, 2013; Vander Heiden et al., 2009; Ward and Thompson, 2012).

Despite this, a consensus over the extent and prominence of metabolic reprogramming in cancer has not yet been established. In other words, the diversity of conclusions in molecular biology studies across different cancer models question the symmetry of metabolic reprogramming (Boroughs and DeBerardinis, 2015; Elia et al., 2015). Systems biology represents a global approach that aims at explaining biological behaviors of interest (like metabolism) by modeling the interactions of *all* components within the system. As such, it may aid in the establishment of consensus. In this review, we will summarize the arguments beyond the symmetry of metabolic reprogramming evidenced by molecular biology and attempt to reconcile them with early global studies of cancer metabolism in systems biology.

The symmetry of metabolic reprogramming in molecular biology

Metabolic reprogramming is defined as the rewiring of a metabolic flux distribution from a functional steady state to another functional steady state. Hence, when researchers talk about metabolic reprogramming in cancer, it must be detailed on which fluxes the rewiring occurs and to which new value. Recurrent phenomena observed in cancer that fall under this definition are: aerobic glycolysis (Gatenby and Gillies, 2004; Ying et al., 2012), addiction to glutamine (Dang, 2010; Son et al., 2013), *de novo* lipogenesis (Baenke et al., 2013), essentiality of one-carbon intermediates (Tedeschi et al., 2013), reliance on autophagy and macropinocytosis (Cheong et al., 2012; Guo et al., 2013), reactive oxygen species homeostasis (Trachootham et al., 2009) and, more recently, dependence on mitochondrial respiration (Birsoy et al., 2014; Viale et al., 2014; Wheaton et al., 2014) (Figure 1A).

Unfortunately, none of these phenotypic traits are universally shared among all cancer cells. Not even aerobic glycolysis is ubiquitous in cancer (Moreno-Sanchez et al., 2007), despite the fact that this phenomenon serves as the basis for fluorodeoxyglucose-mediated positron emission tomography (FDG-PET), among the most accurate diagnostic tool to detect cancer metastasis. For example, a subset of melanomas defined by overexpression of PPARGC1A (also known as PGC1a) displays a distinctive metabolic state characterized by elevated mitochondrial respiration, as opposed to PGC1a-negative melanomas that are highly glycolytic (Vazquez et al., 2013). A similar observation regards diffuse large B cell lymphoma, where a tumor subset insensitive to inhibition of B cell receptor signaling also featured a higher rate of mithochondrial respiration (Caro et al., 2012). A consensus model that accommodates the limits of the symmetry for these phenotypic traits has been constantly challenged by newer discoveries and remains therefore elusive (Boroughs and DeBerardinis, 2015).

Nevertheless, molecular studies in the last decade clearly highlighted that these diverse manifestations of metabolic reprogramming are direct targets of oncogenes and tumor suppressor genes at the origin of cancer. As such, metabolic reprogramming represented cancer vulnerabilities and reversal of the metabolic phenotype induced by cancer mutations often resulted in tumor regression. For instance, mechanistic links between cancer mutations and metabolic reprogramming have been demonstrated for oncogenic c-Myc (Gao et al., 2009), KRAS (Flier et al., 1987; Gaglio et al., 2011; Ying et al., 2012) and BRAF (Haq et al., 2013), loss of tumor suppressors SIRT6 (Sebastian et al., 2012), or oncogene-induced activation of the PI3K-AKT-mTOR pathway (Duvel et al., 2010; Elstrom et al., 2004; Masui et al., 2013), Nrf2 (DeNicola et al., 2011), and b-catenin in Wnt signaling (Cadoret et al., 2002). In the case of the most commonly mutated gene in cancer, *TP53*, it was noted that the profoundly studied and intuitive effects of its inactivation are yet less robust than the effects elicited by *TP53* mutations in metabolism (Berkers et al., 2013). *TP53* is essential to arrest cell cycle and induce apoptosis in the event of genotoxic stress, and loss of these functions is widely regarded to be the key mechanism of cancer initiation. Nevertheless, different *TP53* mutated mice failed to develop tumors even if the encoded protein effectively lost the above-mentioned functions (Choudhury et al., 2007; Li et al., 2012). At the same time, in these experiments the protein retained a similar metabolic control to the wild-type protein. This suggests that the metabolic reprogramming induced by *TP53* mutations is not only oncogenic, but also central to tumor formation (Jiang et al., 2013).

Collectively, these studies demonstrate that as much as it can be heterogeneous and context-dependent in its realization, the reprogramming of metabolism is still consequential to an oncogenic event. This is suggestive that clones with metabolic reprogramming are strongly selected for in the evolution of a tumor. In the same fashion as other hallmarks like sustained proliferative signaling, metabolic reprogramming provides cancer cells with a selective growth advantage. It is imperative to distinguish which traits unlock this growth advantage in spite of the genomic and micro-environmental heterogeneity; if and where metabolic reprogramming is symmetric. The challenge of integrating this diversity and provide a global view of cancer metabolism was recently undertaken by a number of systems biology studies.

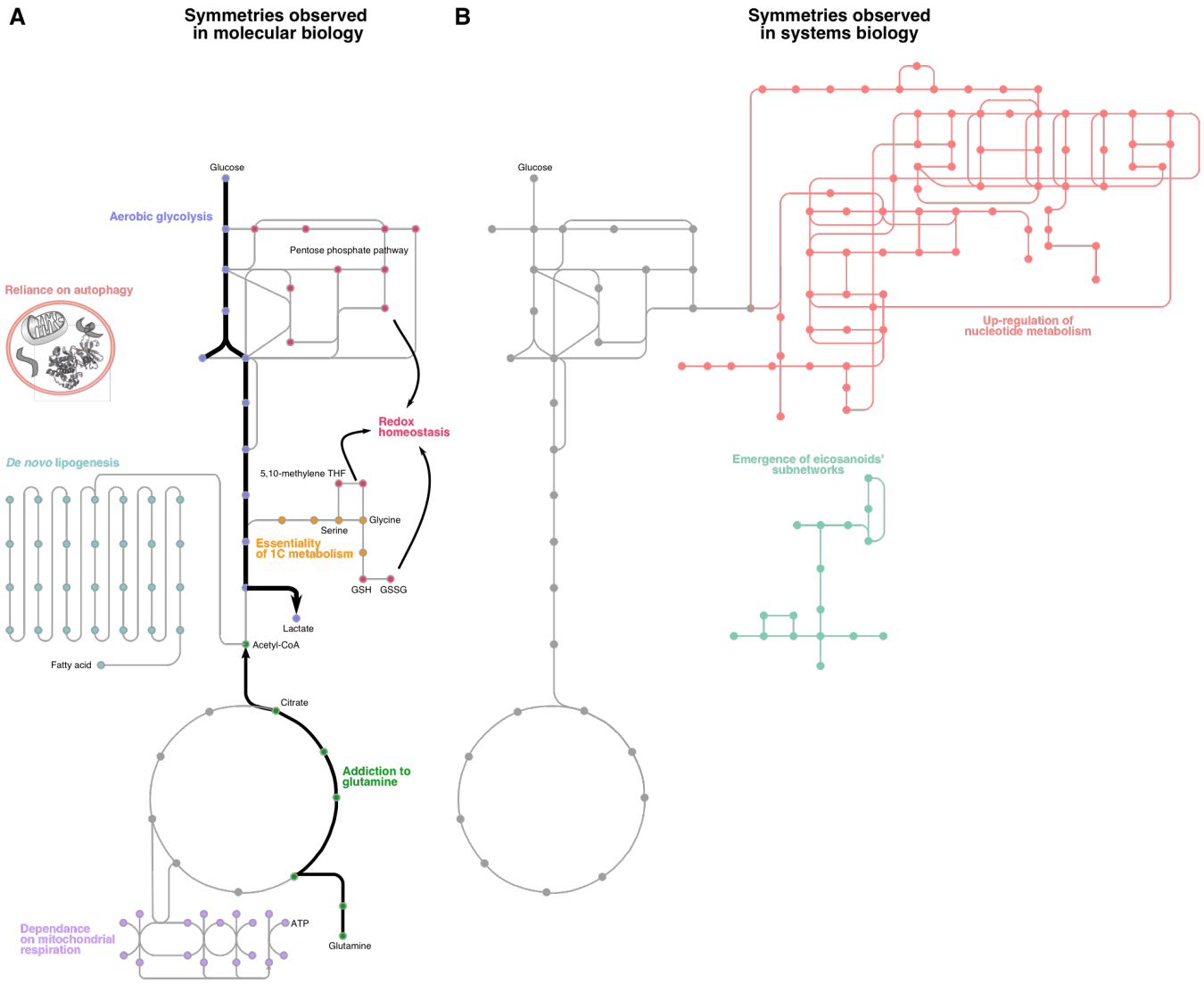


Figure 1 – Traits of metabolic reprogramming in cancer deemed symmetrical in molecular biology (A) and systems biology (B). Key: 1C – One-carbon; GSH – Glutathione; GSSG – Glutathione disulfide; THF – Tetrahydrofolate.

The symmetry of metabolic reprogramming in systems biology

Systems biology is a holistic approach that aims at explaining behaviors of interest by modeling the interactions of all components within a complex system(Cassman and World Technology Evaluation Center., 2007). Metabolism is a complex system. In humans, the metabolic network emerges from interactions of 3765 gene products, according to a recent genome-scale reconstruction(Mardinoglu et al., 2014), and the corresponding set of genes is commonly referred to as metabolic genes. To tackle the challenge of symmetry in cancer metabolism, it would be required to find which traits within metabolic reprogramming are recurrent in all possible cancer cells. In this section, we will delineate some systems biology studies on this subject that have attempted to maintain a global view on metabolism by accounting for most metabolic genes in pan-cancer cohorts. However, due to technological limitations each study relied on its own approximation for the definition of phenotypic traits. In general, the higher the level of abstraction adopted for a trait, the larger was the sample size available for the study. In light of this, these systems biology studies embodied early efforts to uncover symmetries within metabolic reprogramming, but technological progress will refine the approximations introduced to define a phenotypic trait.

A first level of approximation is to define a trait by the presence of genetic or epigenetic alterations in its corresponding metabolic pathway, as defined through a canonical classification by the Gene Ontology Consortium (GO)(Ashburner et al., 2000) or the Kyoto Encyclopedia of Genes and Genomes (KEGG)(Kanehisa et al., 2014). Metabolic reprogramming is thus the

putative effect of such alterations on cancer cells as opposed to genetically intact parental cells. This analysis is implicitly unlocked from pan-cancer studies that mapped genetic alterations to all known cellular processes and pathways, hence including metabolism. In one such studies, Ciriello et al. condensed these alterations across 3,299 tumor samples from 12 tumor types into 31 oncogenic signatures(Ciriello et al., 2013). None of these signatures was found to specifically enrich metabolic pathways or processes. Nevertheless, we note that all signatures collectively encompassed 5 cancer pathways previously mentioned to control metabolic reprogramming, namely signaling through the PI3K-AKT, RAS, Wnt, p53, and c-Myc.

A second level of approximation is to define a trait by the expression level of the genes belonging to its corresponding metabolic pathway, and metabolic reprogramming as a change in expression of the trait in the tumor compared to the matched normal tissue. In a first study from 2013, Hu et al. looked for symmetries in the gene expression profiles of 1,421 metabolic genes of 1,646 tumor samples spanning 22 types vs. 962 matched normal tissues, based on data generated from a human microarray platform(Hu et al., 2013). The authors found limited metabolic reprogramming in cancer, in that the metabolic gene expression profile is largely similar to the normal tissue of origin, and cancers of different types do to a great extent, *not* share a specific type metabolic reprogramming. They quantified this in terms of Euclidean distance in the metabolic gene expression profile. By assuming that a normal tissue (e.g. breast) is 100% dissimilar to another tissue (e.g. prostate), than the corresponding cancers (i.e. breast and prostate cancer, respectively) are only 63% dissimilar on average to the normal tissue of origin, but 83% dissimilar between each other. Despite this heterogeneity, they report a symmetric behavior in the biosynthesis of pyrimidines, where they observed that most genes are up-regulated, and, to a lesser extent, in purine and aminoacyl-tRNA biosynthesis, glycolysis, retinol and xenobiotics metabolism, and the degradation of essential amino acids and fatty acids. In a subsequent study in 2014, we searched analogous symmetries in 3,674 metabolic genes across 257 pairs of cancer vs. adjacent normal tissues encompassing 7 tumor types, based on data from a next-generation sequencing platform(Gatto et al., 2014). Here, we confirmed all the broad conclusions reached by the previous study. In our data set, we reported that cancer cells orchestrated the expression of metabolic genes in a similar fashion only when it comes to nucleotide, glutamate, and retinol metabolism. A last study by Nilsson et al. further corroborated these findings. Their data set consisted of 1,981 tumors of 19 different types vs. 931 matched normal tissues, where they interrogated the expression of 1,454 metabolic genes using data normalized from multiple human microarray platforms. Even though the authors did not look for symmetries by formally collecting these genes into pathway, they reported that consistently regulated genes in cancer belong to glycolysis, anti-oxidant metabolism, glycosylation pathways, nucleotide and deoxynucleotide metabolism, while the top regulated gene, methylenetetrahydrofolate dehydrogenase 2 (*MTHFD2*), is part of one-carbon metabolism. Instead of looking to deregulation of pathways, an alternative is to investigate deregulation of gene co-expression patterns. Reznik and Sander focused on pairs of metabolic genes that are differentially co-regulated in 1394 tumor samples with respect to 177 matched normal tissues in 2 tumor types(Reznik and Sander, 2015). Strikingly, almost no overlap was observed in the patterns of differential co-expression between the 2 tumor types. This finding was further documented when the authors included 1755 samples from 5 other tumor types and their corresponding normal tissues: no gene pair was simultaneously differentially co-regulated in all tumor types. This is suggestive that symmetries, if any, are to be searched at the pathway level rather that as genetic interactions. Taken together, the consensus between three independent studies is that any tumor undertakes an obligate step during the transformation: the metabolism of nucleotides is up-regulated. Considering the global scale of these systems biology studies, one could even argue that nucleotide metabolism might be the *only* coordinately regulated metabolic pathway at the transcriptional level in cancer.

A third level of approximation defines a trait by the expression of sufficiently connected components with the metabolic network, like metabolites and reactions, rather than relying on canonical and arbitrary definition of metabolic processes. Here, metabolic reprogramming consists in the change of expression of small networks, which are recurrently over-represented in tumors compared to the corresponding normal tissues. In 2012, Ågren et al. compared the metabolic networks of 16 cancer types vs. 24 matched normal cell types to identify metabolites involved in sub-networks observed more often in cancer at the protein level(Agren et al., 2012). Four groups of connected metabolites were reported: polyamines (e.g. spermidine), isoprenoids (e.g. geranylgeranyl diphosphate), eicosanoids (e.g. 5-hydroperoxyeicosatetraenoic acid, or 5-HPETE), and heme catabolites (e.g. bilirubin). Later that year, Wang et al. compared the metabolic networks of 17 cancer types vs. 18 matched normal tissues and ranked pathways according to over-representation of reactions in cancer at the gene expression level(Wang et al., 2012). The most significant enrichments were detected for eicosanoid, nucleotide and one-carbon metabolism and lysosomal transport. Collectively, the two studies show that cancers are symmetric in the formation of reaction sub-networks that revolve around the metabolism of eicosanoids.

A fourth level of approximation defines a trait by the abundance of related metabolites, which are informative of the metabolic state of that trait. Metabolic reprogramming consists in the change of metabolite abundance in tumors compared to normal tissues. A global view on this definition of metabolic reprogramming is provided by untargeted metabolomics (Gupta and Chawla, 2013). This technology holds the promise to profile the abundance of all metabolites in a tumor, even though at the current state the characterization of many metabolites remains elusive (Weljie and Jirik, 2011). Also, as for today, there is no systematic study in which untargeted metabolomics has been applied in a diverse group of cancer types to reveal differential metabolite abundance compared to matched normal tissues. Considering this, the picture offered by this technology is still too incomplete to infer symmetric traits of metabolic reprogramming. Nevertheless, Patel and Ahmed recently reviewed a progresses in individual cancer types spanning 12 tissues (Patel and Ahmed, 2015). Despite differences in the metabolomics tool used and thereby detection sensitivity, the authors reported some recurrent perturbations between cancers arising from various tissues. They ascribed these to glycolysis (reprogrammed in 6 of 12 tissues), TCA cycle (6 of 12), choline metabolism (9 of 12) and fatty acid metabolism (9 of 12). In terms of specific metabolites, the most recurrently perturbed are choline, phosphatidylcholine and lysophosphatidylcholine, each differentially abundant in five distinct cancer tissues. Reprogramming of choline metabolism was undetected only in tumors of the urinary and female reproductive system and in multiple myeloma. Whilst awaiting for a systematic approach to unveil symmetries in cancer metabolism using untargeted metabolomics, these studies collectively point to central carbon metabolism, choline and fatty acid metabolism as the metabolic traits mostly affected by malignant transformation.

A fifth level of approximation defines a trait by the flux through sufficiently connected components of the metabolic network, and a strict application for cancer of the definition of metabolic reprogramming is in this way readily possible if the change is measured before and after the transformation. Necessarily, the scale of these studies is dramatically reduced and so the reach of any claim about the symmetry. Nevertheless, they provide evidences on non-symmetric traits. In 2012, Yuneva et al. checked glucose and glutamine central carbon metabolism in mouse cancers arising from two different tissues via two distinct evolutionary trajectories(Yuneva et al., 2012). They found no conserved pattern of metabolic reprogramming. In 2013, Fan et al. followed the fluxes in a small network of central carbon metabolism in a parental cell line after induction of two different oncogenes(Fan et al., 2013). They observed a single symmetric trait attributable to the transformation, that is a substantial increase in the fermentative flux of glucose to lactate (i.e. aerobic glycolysis). However, this phenomenon seemed to fulfill distinct metabolic requirements according to which oncogene was activated. Overall, these studies seem to rule out the symmetry in the reprogramming of central carbon metabolism, in that no trait was purely ascribable to cancer regardless of genetic and/or micro-environmental heterogeneity.

The symmetries in cancer metabolism today

The symmetries unveiled by systems biology are so far circumstantiated to two events: transcriptional up-regulation of nucleotide metabolism and formation of sub-networks in the metabolism of eicosanoids (Figure 1B). Given the scale of the studies in which the former event was observed compared to the latter, we focus the following discussion on the symmetry of nucleotide metabolism.

In the first place, if true, why a symmetric regulation only in nucleotide metabolism? And within nucleotide metabolism, why preferentially pyrimidines? Why is it up-regulated? Finally, perhaps most compellingly, is this metabolic reprogramming an adaptive or an oncogenic process? In other words, considering that these studies compared transformed proliferating cells in an abnormal microenvironment to wild-type (mostly) quiescent cells in a physiologically normal tissue, is this metabolic reprogramming a feature of cancer due to adaptation to the transformation or is it driven by the mutations at the origin of cancer rendering the disease vulnerable to disruption in this process?

These questions require mechanistic molecular studies that, as noted before, are hard to tackle at the same scale as the systems biology studies mentioned above. However some speculations are due. In proliferating cells, nucleotides are continuously synthesized to meet the increased requirements of RNA and DNA due to growth. Nevertheless, this argument should apply to all macromolecular classes (proteins, lipids, etc.) and not only RNA or DNA. Hence, we should have observed up-regulation of the corresponding anabolic pathways as well. As a matter of fact, these other classes are more prominent in terms of cellular composition. In an average mammalian cell, DNA accounts for only 1% of dry weight, RNA slightly more, about 4%. For comparison, proteins accounts for 60%, and lipids for 18% (Alberts, 2002). On the other, contrary to lipids and proteins, nucleotides are unique in that they cannot be readily scavenged from the extracellular environment. In addition, the nucleotide-based macromolecular class of RNA is unique compared to other classes in terms of its relation with growth rate. Experiments in bacteria have shown that when cells grow at higher rates, RNA levels exhibits the greatest relative change, in sharp contrast with DNA and proteins (Neidhardt and Curtiss, 1996).

This reflects the increasing concentration of ribosomes needed for increased protein synthesis with increasing growth rate. Whereas there is convincing evidence that DNA concentration is not limiting, the ribosome concentration and the corresponding protein synthesis rate are. Thus, in 1983, Ehrenberg and Kurland proposed that an energy-efficient metabolic strategy for ribosomes at increasing growth rates is to proportionally increase both the substrate pool and the ribosome concentration (Ehrenberg and Kurland, 1984). Since the most prominent phenotypic trait of a cancer cell is the growth advantage over the neighboring normal cells, this growth rate dependent effects on the cell macromolecular composition should be observed during the transformation. In this fashion, the here-in reported convergence on nucleotide metabolism up-regulation seems to support the model of Ehrenberg and Kurland. In other words, this symmetry is probably not an oncogenic process, but an evolutionary conserved metabolic strategy that cells adopt when their growth rate increases. We acknowledge two alternative hypotheses that argue in favor of the oncogenic nature of nucleotide metabolism up-regulation. The first one is the process of DNA damage. More than (virtually) any other normal cell, cancer cells suffer significant DNA damage, which leads to unregulated cell division, and the reincorporation of nucleosides during DNA reparation requires a sanitized pool of these metabolites. Chemical interference with this process leads to tumor regression in mouse models (Gad et al., 2014). In this case, up-regulation of nucleotide metabolism may serve as a cancer-specific reprogramming and explain the observed symmetry. The second one stems from the observation that *de novo* pyrimidine biosynthesis is a process stimulated by mTOR signaling (Ben-Sahra et al., 2013; Robitaille et al., 2013). This is turn is constitutively activated as a result of oncogenic aberrations in upstream pathways, including three of the five above mentioned cancer pathways recurrently associated with the oncogenic signatures, PI3K-AKT, RAS, ans Wnt signaling (Shimobayashi and Hall, 2014). The fact that mTOR is also central to ribosome biogenesis (Gentilella et al., 2015) interlaces its oncogenic control of pyrimidine biosynthesis with the above discussed crucial need for RNA precursors to support a higher growth rate. A simple view of the symmetry in cancer metabolism would be an adaptive process in which proliferation is unlocked by oncogenic activation of mTOR that up-regulate nucleotide biosynthesis, whose products are in turn limiting for proliferation. However, as much as mTOR can be presumptively assumed active in the transformation of many disparate cancer types (Kandoth et al., 2013), this symmetry should also be validated.

The fact that up-regulation of nucleotide metabolism is symmetrical in cancer opens a therapeutic window, which, in fact, represents one of the most long standing and effectively exploited chemotherapeutic strategies. Indeed, among the oldest available treatments for cancers of various type is a chemotherapy consisting of antimetabolites (Tennant et al., 2010). These agents are analogs to human metabolites and they function by interfering with those reactions that use these metabolites as substrate. Gemcitabine, decitabine, and fluorouracil are examples of these drugs. They are all antimetabolites to nucleotides, and specifically pyrimidines. Their clinical use against a number of disparate cancers underscores the symmetry of nucleotide metabolism in cancer. However, it does not necessarily corroborate the above claim about its uniqueness in the landscape of metabolic pathways. Indeed, another prominent class of antimetabolite drugs is represented by anti-folates (e.g. methotrexate). This, in turn, can be reconciled to the finding reported by Nilsson and colleagues about *MTHFD2* and the related mitochondrial folate pathway. Taken together, the direct observations in human tumors of the efficacy of inhibition of such (and perhaps *only* such) metabolic pathways seem to support their symmetry in cancer metabolism. At the same time, the fact that cancer can relapse after antimetabolite treatment is also indicative that such symmetric regulation is yet circumventable, hence suggesting that the process is rather adaptive then oncogenic.

In our study (Gatto et al., 2014), we also noted that the symmetry is broken by the case of clear cell renal cell carcinoma (ccRCC), the most common form of kidney tumor (Rini et al., 2009). Here we reported, for example, that nucleotide metabolism is down-regulated at the transcript and protein level. We ascribed the uniqueness of ccRCC metabolic reprogramming to recurrent copy number alterations in the chromosome 3p. Here is also located the most commonly mutated tumor suppressor gene in ccRCC, the von Hippel-Lindau (*VHL*) tumor suppressor (Creighton et al., 2013). This proposed model of metabolic reprogramming, which remains to be validated, prescribes that exceptional genetic events following loss of *VHL* lead to the symmetry break while still promoting tumor progression.

Perspectives

A central question in the search for symmetries in systems biology was introduced above and regards the oncogenic nature of a symmetrical process. In contrast to molecular biology where mechanisms induced by an oncogenic mutation can be studied in great detail, such approaches have not yet been tackled by systems biology.

In addition, we identified five weaknesses that hamper any definitive claim about symmetries even for the outlined large-scale studies:

1. First and foremost, the phenotype of a tumor is only approximated by its gene expression profile. The central dogma commends that the proteins are the ultimate effectors for the phenotype of a cell.

Although recent estimates recognize to the process of transcription a prominent role in the control of protein levels (accounting for ~70% of the variance (Li and Biggin, 2015)), translation and degradation cannot be neglected. Even under this approximation, proteins exert their function and define a cell phenotype by means of interactions within each other and with the environment (Barabasi and Oltvai, 2004; Vidal et al., 2011). These interactions depend on the availability of certain compounds at a given time, potential modifications of protein active sites (or even distant sites) via post-translational modifications, and probably on processes that we do not fully understand or not know at all. Indeed, researchers rely and attempt to corroborate the paradigm introduced with the central dogma. This imposes tremendous limitations on our ability to describe and interpret a phenotype. Most of these results will undoubtedly collapse or necessitate to be revisited in light of a future paradigm shift.

2. The sample size of these studies is still limited, with only thousands of cancers across tens of cancer types. It is worth reminding that some 15 million new cases of cancer are diagnosed every year, classified in over 100 types.

3. Genetic and epigenetic alterations may not be the only drivers of cancer evolution. For example, in a recent publication, Martincorena et al. found that skin cells displayed a surprisingly high number of mutations in cancer driver genes, despite being physiologically normal. As many as 83 clones per square centimeter of skin positively selected for mutations in *NOTCH* genes, a family causally implicated in cancer due to their role in the regulation of stem cell biology. Even though these are aged and UV-exposed cells and even considering that the mutation burden is still at the lower end for most skin cancers, the fact that normal cells carry so many cancer-causing mutations is sufficient to question "what combinations of events are sufficient for transformation" (Martincorena et al., 2015).

4. The technology is limiting. This is a fact that will always impinge the spectrum of scientific questions that can be legitimately answered. In this case, the advancement in our understanding on which and to which extent a gene is expressed is outstanding compared to twenty years ago or so. Yet, these studies relied on pictures of the transcriptome that are static, estimated, and related to a variegated population of human cells.

5. In close relation with the previous point, these questions should be addressed experimentally. These experiments are in turn dependent on the presence of a scalable and practical technology. Only repeated experimental observation may corroborate the boundaries of the symmetry so far claimed for the metabolism of cancer.

Conclusion

We believe that the search for symmetries is essential to understand the role of metabolism in cancer evolution. These symmetries represent basic requirements for the existence of cancer. So far, systems biology studies have elucidated some instances where such symmetry can be claimed or should be ruled out. Even though the findings unlocked by systems biology, like the centrality of nucleotide metabolism, are intuitive for many cancer researchers, the global approach of systems biology helps to set the borders and circumstantiate what really matters. If anything, it draws our attention away from processes that are just a corollary to the origin of cancer and promotes our curiosity towards new hypotheses.

Acknowledgments

The authors would like to thank Leif Väremo and Costas Lyssiotis for critically reviewing the manuscript. This work was sponsored by a grant from the Knut and Alice Wallenberg Foundation.

References

Agren, R., Bordel, S., Mardinoglu, A., Pornputtapong, N., Nookaew, I., and Nielsen, J. (2012). Reconstruction of genome-scale active metabolic networks for 69 human cell types and 16 cancer types using INIT. PLoS computational biology 8, e1002518.

Alberts, B. (2002). Molecular biology of the cell. (New York: Garland Science).

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., et al. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nature genetics 25, 25-29.

Baenke, F., Peck, B., Miess, H., and Schulze, A. (2013). Hooked on fat: the role of lipid synthesis in cancer metabolism and tumour development. Disease models & mechanisms 6, 1353-1363.

Barabasi, A.L., and Oltvai, Z.N. (2004). Network biology: understanding the cell's functional organization. Nature reviews. Genetics 5, 101-113.

Ben-Sahra, I., Howell, J.J., Asara, J.M., and Manning, B.D. (2013). Stimulation of de Novo Pyrimidine Synthesis by Growth Signaling Through mTOR and S6K1. Science.

Berkers, C.R., Maddocks, O.D., Cheung, E.C., Mor, I., and Vousden, K.H. (2013). Metabolic regulation by p53 family members. Cell metabolism 18, 617-633.

Birsoy, K., Possemato, R., Lorbeer, F.K., Bayraktar, E.C., Thiru, P., Yucel, B., et al. (2014). Metabolic determinants

of cancer cell sensitivity to glucose limitation and biguanides. Nature 508, 108-112.
Boroughs, L.K., and DeBerardinis, R.J. (2015). Metabolic pathways promoting cancer cell survival and growth. Nature cell biology 17, 351-359.
Cadoret, A., Ovejero, C., Terris, B., Souil, E., Levy, L., Lamers, W.H., et al. (2002). New targets of beta-catenin signaling in the liver are involved in the glutamine metabolism. Oncogene 21, 8293-8301.
Cairns, R.A., Harris, I.S., and Mak, T.W. (2011). Regulation of cancer cell metabolism. Nature reviews. Cancer 11, 85-95.
Caro, P., Kishan, A.U., Norberg, E., Stanley, I.A., Chapuy, B., Ficarro, S.B., et al. (2012). Metabolic signatures uncover distinct targets in molecular subsets of diffuse large B cell lymphoma. Cancer cell 22, 547-560.
Cassman, M., and World Technology Evaluation Center. (2007). Systems biology : international research and development. (Dordrecht, The Netherlands: Springer).
Cheong, H., Lu, C., Lindsten, T., and Thompson, C.B. (2012). Therapeutic targets in cancer cell metabolism and autophagy. Nature biotechnology 30, 671-678.
Choudhury, A.R., Ju, Z., Djojosubroto, M.W., Schienke, A., Lechel, A., Schaetzlein, S., et al. (2007). Cdkn1a deletion improves stem cell function and lifespan of mice with dysfunctional telomeres without accelerating cancer formation. Nature genetics 39, 99-105.
Ciriello, G., Miller, M.L., Aksoy, B.A., Senbabaoglu, Y., Schultz, N., and Sander, C. (2013). Emerging landscape of oncogenic signatures across human cancers. Nature genetics 45, 1127-1133.
Creighton, C.J., Morgan, M., Gunaratne, P.H., Wheeler, D.A., Gibbs, R.A., Gordon Robertson, A., et al. (2013). Comprehensive molecular characterization of clear cell renal cell carcinoma. Nature.
Dang, C.V. (2010). Glutaminolysis: supplying carbon or nitrogen or both for cancer cells? Cell Cycle 9, 3884-3886.
DeNicola, G.M., Karreth, F.A., Humpton, T.J., Gopinathan, A., Wei, C., Frese, K., et al. (2011). Oncogene-induced Nrf2 transcription promotes ROS detoxification and tumorigenesis. Nature 475, 106-109.
Duvel, K., Yecies, J.L., Menon, S., Raman, P., Lipovsky, A.I., Souza, A.L., et al. (2010). Activation of a metabolic gene regulatory network downstream of mTOR complex 1. Molecular cell 39, 171-183.
Ehrenberg, M., and Kurland, C.G. (1984). Costs of accuracy determined by a maximal growth rate constraint. Quarterly reviews of biophysics 17, 45-82.
Elia, I., Schmieder, R., Christen, S., and Fendt, S.M. (2015). Organ-Specific Cancer Metabolism and Its Potential for Therapy. Handbook of experimental pharmacology.
Elstrom, R.L., Bauer, D.E., Buzzai, M., Karnauskas, R., Harris, M.H., Plas, D.R., et al. (2004). Akt stimulates aerobic glycolysis in cancer cells. Cancer Res 64, 3892-3899.
Evan, G.I., and Vousden, K.H. (2001). Proliferation, cell cycle and apoptosis in cancer. Nature 411, 342-348.
Fan, J., Kamphorst, J.J., Mathew, R., Chung, M.K., White, E., Shlomi, T., et al. (2013). Glutamine-driven oxidative phosphorylation is a major ATP source in transformed mammalian cells in both normoxia and hypoxia. Molecular systems biology 9, 712.
Flier, J.S., Mueckler, M.M., Usher, P., and Lodish, H.F. (1987). Elevated levels of glucose transport and transporter messenger RNA are induced by ras or src oncogenes. Science 235, 1492-1495.
Forbes, S.A., Beare, D., Gunasekaran, P., Leung, K., Bindal, N., Boutselakis, H., et al. (2015). COSMIC: exploring the world's knowledge of somatic mutations in human cancer. Nucleic acids research 43, D805-811.
Gad, H., Koolmeister, T., Jemth, A.S., Eshtad, S., Jacques, S.A., Strom, C.E., et al. (2014). MTH1 inhibition eradicates cancer by preventing sanitation of the dNTP pool. Nature 508, 215-221.
Gaglio, D., Metallo, C.M., Gameiro, P.A., Hiller, K., Danna, L.S., Balestrieri, C., et al. (2011). Oncogenic K-Ras decouples glucose and glutamine metabolism to support cancer cell growth. Molecular systems biology 7, 523.
Gao, P., Tchernyshyov, I., Chang, T.C., Lee, Y.S., Kita, K., Ochi, T., et al. (2009). c-Myc suppression of miR-23a/b enhances mitochondrial glutaminase expression and glutamine metabolism. Nature 458, 762-765.
Gatenby, R.A., and Gillies, R.J. (2004). Why do cancers have high aerobic glycolysis? Nature reviews. Cancer 4, 891-899.
Gatto, F., Nookaew, I., and Nielsen, J. (2014). Chromosome 3p loss of heterozygosity is associated with a unique metabolic network in clear cell renal carcinoma. Proceedings of the National Academy of Sciences of the United States of America 111, E866-875.
Gentilella, A., Kozma, S.C., and Thomas, G. (2015). A liaison between mTOR signaling, ribosome biogenesis and cancer. Biochimica et biophysica acta.
Gerlinger, M., McGranahan, N., Dewhurst, S.M., Burrell, R.A., Tomlinson, I., and Swanton, C. (2014). Cancer: Evolution Within a Lifetime. Annu Rev Genet 48, 215-236.
Gerlinger, M., Rowan, A.J., Horswell, S., Larkin, J., Endesfelder, D., Gronroos, E., et al. (2012). Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. The New England journal of medicine 366, 883-892.
Guo, J.Y., Xia, B., and White, E. (2013). Autophagy-mediated tumor promotion. Cell 155, 1216-1219.
Gupta, S., and Chawla, K. (2013). Oncometabolomics in cancer research. Expert review of proteomics 10, 325-336.

Hanahan, D., and Weinberg, R.A. (2000). The hallmarks of cancer. Cell 100, 57-70.
Hanahan, D., and Weinberg, R.A. (2011). Hallmarks of cancer: the next generation. Cell 144, 646-674.
Haq, R., Shoag, J., Andreu-Perez, P., Yokoyama, S., Edelman, H., Rowe, G.C., et al. (2013). Oncogenic BRAF regulates oxidative metabolism via PGC1alpha and MITF. Cancer cell 23, 302-315.
Hu, J., Locasale, J.W., Bielas, J.H., O'Sullivan, J., Sheahan, K., Cantley, L.C., et al. (2013). Heterogeneity of tumor-induced gene expression changes in the human metabolic network. Nature biotechnology.
Jiang, P., Du, W., Mancuso, A., Wellen, K.E., and Yang, X. (2013). Reciprocal regulation of p53 and malic enzymes modulates metabolism and senescence. Nature 493, 689-693.
Johnson, B.E., Mazor, T., Hong, C., Barnes, M., Aihara, K., McLean, C.Y., et al. (2014). Mutational analysis reveals the origin and therapy-driven evolution of recurrent glioma. Science 343, 189-193.
Kandoth, C., McLellan, M.D., Vandin, F., Ye, K., Niu, B., Lu, C., et al. (2013). Mutational landscape and significance across 12 major cancer types. Nature 502, 333-339.
Kanehisa, M., Goto, S., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M. (2014). Data, information, knowledge and principle: back to metabolism in KEGG. Nucleic acids research 42, D199-205.
Lawrence, M.S., Stojanov, P., Mermel, C.H., Robinson, J.T., Garraway, L.A., Golub, T.R., et al. (2014). Discovery and saturation analysis of cancer genes across 21 tumour types. Nature 505, 495-501.
Li, J.J., and Biggin, M.D. (2015). Gene expression. Statistics requantitates the central dogma. Science 347, 1066-1067.
Li, T., Kon, N., Jiang, L., Tan, M., Ludwig, T., Zhao, Y., et al. (2012). Tumor suppression in the absence of p53-mediated cell-cycle arrest, apoptosis, and senescence. Cell 149, 1269-1283.
Mardinoglu, A., Agren, R., Kampf, C., Asplund, A., Uhlen, M., and Nielsen, J. (2014). Genome-scale metabolic modelling of hepatocytes reveals serine deficiency in patients with non-alcoholic fatty liver disease. Nature communications 5, 3083.
Mardis, E.R., Ding, L., Dooling, D.J., Larson, D.E., McLellan, M.D., Chen, K., et al. (2009). Recurring mutations found by sequencing an acute myeloid leukemia genome. The New England journal of medicine 361, 1058-1066.
Martincorena, I., Roshan, A., Gerstung, M., Ellis, P., Van Loo, P., McLaren, S., et al. (2015). Tumor evolution. High burden and pervasive positive selection of somatic mutations in normal human skin. Science 348, 880-886.
Masui, K., Tanaka, K., Akhavan, D., Babic, I., Gini, B., Matsutani, T., et al. (2013). mTOR complex 2 controls glycolytic metabolism in glioblastoma through FoxO acetylation and upregulation of c-Myc. Cell metabolism 18, 726-739.
Moreno-Sanchez, R., Rodriguez-Enriquez, S., Marin-Hernandez, A., and Saavedra, E. (2007). Energy metabolism in tumor cells. Febs J 274, 1393-1418.
Neidhardt, F.C., and Curtiss, R. (1996). Escherichia coli and Salmonella : cellular and molecular biology. (Washington, D.C.: ASM Press).
Nik-Zainal, S., Van Loo, P., Wedge, D.C., Alexandrov, L.B., Greenman, C.D., Lau, K.W., et al. (2012). The life history of 21 breast cancers. Cell 149, 994-1007.
Parsons, D.W., Jones, S., Zhang, X., Lin, J.C., Leary, R.J., Angenendt, P., et al. (2008). An integrated genomic analysis of human glioblastoma multiforme. Science 321, 1807-1812.
Patel, S., and Ahmed, S. (2015). Emerging field of metabolomics: big promise for cancer biomarker identification and drug discovery. Journal of pharmaceutical and biomedical analysis 107, 63-74.
Reznik, E., and Sander, C. (2015). Extensive decoupling of metabolic genes in cancer. PLoS computational biology 11, e1004176.
Rini, B.I., Campbell, S.C., and Escudier, B. (2009). Renal cell carcinoma. Lancet 373, 1119-1132.
Robitaille, A.M., Christen, S., Shimobayashi, M., Cornu, M., Fava, L.L., Moes, S., et al. (2013). Quantitative Phosphoproteomics Reveal mTORC1 Activates de Novo Pyrimidine Synthesis. Science.
Schulze, A., and Harris, A.L. (2013). How cancer metabolism is tuned for proliferation and vulnerable to disruption (vol 491, pg 364, 2012). Nature 494, 130-130.
Sebastian, C., Zwaans, B.M., Silberman, D.M., Gymrek, M., Goren, A., Zhong, L., et al. (2012). The histone deacetylase SIRT6 is a tumor suppressor that controls cancer metabolism. Cell 151, 1185-1199.
Shimobayashi, M., and Hall, M.N. (2014). Making new contacts: the mTOR network in metabolism and signalling crosstalk. Nature reviews. Molecular cell biology 15, 155-162.
Son, J., Lyssiotis, C.A., Ying, H., Wang, X., Hua, S., Ligorio, M., et al. (2013). Glutamine supports pancreatic cancer growth through a KRAS-regulated metabolic pathway. Nature 496, 101-+.
Tedeschi, P.M., Markert, E.K., Gounder, M., Lin, H., Dvorzhinski, D., Dolfi, S.C., et al. (2013). Contribution of serine, folate and glycine metabolism to the ATP, NADPH and purine requirements of cancer cells. Cell death & disease 4, e877.
Tennant, D.A., Duran, R.V., and Gottlieb, E. (2010). Targeting metabolic transformation for cancer therapy. Nature reviews. Cancer 10, 267-277.
Trachootham, D., Alexandre, J., and Huang, P. (2009). Targeting cancer cells by ROS-mediated mechanisms: a radical therapeutic approach? Nature Reviews Drug Discovery 8, 579-591.

Vander Heiden, M.G., Cantley, L.C., and Thompson, C.B. (2009). Understanding the Warburg effect: the metabolic requirements of cell proliferation. Science 324, 1029-1033.

Vazquez, F., Lim, J.H., Chim, H., Bhalla, K., Girnun, G., Pierce, K., et al. (2013). PGC1alpha expression defines a subset of human melanoma tumors with increased mitochondrial capacity and resistance to oxidative stress. Cancer cell 23, 287-301.

Viale, A., Pettazzoni, P., Lyssiotis, C.A., Ying, H., Sanchez, N., Marchesini, M., et al. (2014). Oncogene ablation-resistant pancreatic cancer cells depend on mitochondrial function. Nature 514, 628-632.

Vidal, M., Cusick, M.E., and Barabasi, A.L. (2011). Interactome networks and human disease. Cell 144, 986-998.

Vogelstein, B., and Kinzler, K.W. (2004). Cancer genes and the pathways they control. Nature medicine 10, 789-799.

Vogelstein, B., Papadopoulos, N., Velculescu, V.E., Zhou, S., Diaz, L.A., Jr., and Kinzler, K.W. (2013). Cancer genome landscapes. Science 339, 1546-1558.

Wang, Y., Eddy, J.A., and Price, N.D. (2012). Reconstruction of genome-scale metabolic models for 126 human tissues using mCADRE. BMC systems biology 6, 153.

Warburg, O. (1956a). On respiratory impairment in cancer cells. Science 124, 269-270.

Warburg, O. (1956b). On the origin of cancer cells. Science 123, 309-314.

Ward, P.S., and Thompson, C.B. (2012). Metabolic reprogramming: a cancer hallmark even warburg did not anticipate. Cancer cell 21, 297-308.

Weljie, A.M., and Jirik, F.R. (2011). Hypoxia-induced metabolic shifts in cancer cells: moving beyond the Warburg effect. The international journal of biochemistry & cell biology 43, 981-989.

Wheaton, W.W., Weinberg, S.E., Hamanaka, R.B., Soberanes, S., Sullivan, L.B., Anso, E., et al. (2014). Metformin inhibits mitochondrial complex I of cancer cells to reduce tumorigenesis. eLife 3, e02242.

Ying, H., Kimmelman, A.C., Lyssiotis, C.A., Hua, S., Chu, G.C., Fletcher-Sananikone, E., et al. (2012). Oncogenic Kras maintains pancreatic tumors through regulation of anabolic glucose metabolism. Cell 149, 656-670.

Yuneva, M.O., Fan, T.W., Allen, T.D., Higashi, R.M., Ferraris, D.V., Tsukamoto, T., et al. (2012). The metabolic profile of tumors depends on both the responsible genetic lesion and tissue type. Cell metabolism 15, 157-170.