ABSTRACT

Title of Dissertation: GENOME WIDE DISCOVERY OF DISEASE MODIFIERS

Noam Auslander, Doctor of Philosophy, 2018

Dissertation directed by: Professor Eytan Ruppin, Department of Computer Science

Disease modifiers are genes that when activated can alter the expression of a phenotype associated with a disease. This can be done directly through affecting the expression of another gene that is causing the disease, or indirectly by affecting other factors that contribute to the phenotype's variability. Identification of disease modifiers is of great interest from both treatment and genetic counseling perspectives. We set here to develop computational approaches to identify and study disease modifiers. We focus on two research avenues for studying disease modifiers: (1) One aimed at identifying and investigating modifiers of cancer, a complex disease influenced by multiple genetic and environmental factors, and (2) the other focuses on the identification of disease modifiers for monogenetic disorders which involve a single disease causing gene.

Towards the first aim of studying cancer modifiers we take four complimentary approaches. (a) First, we developed a computational approach to

identify metabolic drivers of cancer that when applied to colorectal cancer, successfully identified FUT9 as a gene that strongly modifies tumors aggressiveness. (b) Second, to study metabolic *pathway-level modifications* in cancer, we developed an algorithm that summarizes cancer modifications to generate pathway compositions that best capture cancer associated alterations, which, as we show, enhances cancer classification and survival prediction. (c) Third, to identify *modifiers of cancer immunotherapy treatment,* we developed a new computational approach that robustly predicts the response to immune checkpoint blockage therapy. (d) Fourth, to identify *modifiers of cancer radiotherapy treatment* we built a robust predictor of rectal cancer patients' response to chemo-radiation-therapy (CRT), identifying a signature of genes that may serve a potential targets for modifying patients' response to CRT.

Towards the second aim of studying genetic modifiers of Mendelian diseases, we developed a computational approach for identifying a specific expression pattern associated with genes that are modifying disease severity. We show that we can successfully prioritize genes that are modifying disease severity in cystic fibrosis and spinal muscular atrophy, where we have identified a new modifier and validated it experimentally.

As will become evident from reading my dissertation, my work has naturally focused on developing a variety of computational approaches to analyze research questions that were of interest to me. Obviously, my work has greatly benefited and has been significantly enriched by close collaboration with many experimental labs

that have kindly embarked on testing the predictions made, and to whom I am indebted. In sum, we developed methods to identify and study disease modifiers for both cancer and Mendelian diseases. The applications of these methods generates a few promising leads for advancing the treatment for these diseases and improving clinical decision-making.

GENOME WIDE DISCOVERY OF DISEASE MODIFIERS

By

Noam Auslander

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2018

Advisory Committee:

Professor Eytan Ruppin, Chair
Hector Corrada Bravo
Max Leiserson
Stephan Mount
James A. Reggia

Preface

My long term research interests are in the development of computational approaches to identify and investigate disease modifiers that may improve the treatment for these diseases. I aim to develop approaches that can be robustly validated (using existing data or experimentally) and that are easily interpretable for our experimental collaborators and potential users from the biological fields.

In this dissertation I present five computational approaches, designed to answer five research questions with one common objective to identify factors that can modify the expression, severity and prognosis of a disease. My main research focus has been studying modifiers for cancer, with one exceptional project in which I investigated disease modifiers for monogenetic disorders. Each computational approach presented here utilizes computational tools via specific data representation, designed to particularly answer each research question considering the relevant data in availability.

Initially, I was interested in cancer driver genes identification, a well-studied and important problem, for which the main obstacle is distinguishing the genes that are driving cancer from these that are just associated with it (termed ‘passenger’ alterations). To this end, I utilized Genome Scale Metabolic Modeling that enables perturbation simulations and can be used to prioritize (metabolic) genes alterations that results with the transcriptional profile observed in tumors (and may hence be truly causal for the cancerous state).

As I became interested in the field of machine learning, a major part of my work was focused on developing machine learning solutions for biological problems that were interesting to me. For this, I investigated machine learning algorithms and data representation to solve different research questions; Studying pathway-level aggregation, I learned that aggregating gene expression via canonical pathway prohibits simple cancer classification. I hence developed a data-driven algorithm that aggregates gene expression for pathway composition that optimally differentiate between healthy and tumor tissues, which also enables cancer survival prediction. Later on, I was interested in predicting response to cancer treatment. Studying checkpoint blockage therapy (ICB) response prediction, I learned that while a few central features (immune checkpoint genes) play a key role in these treatments, the expression of these genes is a poor predictor of response. I hence developed a predictor that compares the expression of pairs of immune activators and inhibitors genes that can robustly predict ICB response and can be easily transferred across datasets. Then, Working with Thomas Ried at the NCI, I learned that approximately one third of rectal cancer patients (currently treated with neoadjuvant chemo-radiation treatment (CRT) followed by surgery) are tumor-free after CRT and might be equally well treated by a "watch and wait" strategy instead of surgery. I hence developed a predictor of response to CRT that specifically spots those complete responders and can be used to identify patients that may be spared from unnecessary surgery.

Finally, I became interested in disease modifiers for monogenetic diseases. I found that existing approaches require the utilization of large sequencing data, which is scares or completely absent for most of these diseases. I hence developed an approach

that can prioritize potential modifiers using healthy tissues gene expression by characterized genetic interactions patterns associated with such modifiers.

In sum, I have been studying different disease aspects and developed computational approaches adjusted to each research question considering the data availability. Working in close collaboration with different experimental labs on each of these projects provided me a better understanding of the research questions, which motivated the computational approaches I developed to answer each question. I believe that much of this work can be used in future studies to advance the development of treatments and improve clinical decision-making for these diseases.

Acknowledgements

This thesis recaps most of my work during the past four years, which has been the best and most significant period in my life so far. I wish to thank several people for their guidance, collaborations and support during these years.

First and foremost, I would like to express my profound gratitude and appreciation to my advisor Dr. Eytan Ruppin, for his continuous guidance and his boundless support. Thank you for your patience and kindness in mentoring me, for teaching me how to ask the right questions and how to answer them, and for being a true role model as a scientist and as a person.

Second, I would like to thank Dr. Thomas Ried for a wonderful mentorship and collaboration in the past two years, I feel fortunate for the privilege to work with you and to learn from you.

Third, I wish to thank my former lab members Keren Yizhak and Allon Wagner for their indispensable guidance early on.

I would also like to thank the current lab member with whom I was fortunate to work over the last few years: Joo Sang Lee, Sushant Patkar, Erez Persi, Rotem Katzir, Matthew Oberhardt, Hiren Karathia, Kuoyuan Cheng and Welles Robinson. Thank you for your support and collaboration.

I wish to also thank the excellent experimental collaborators with whom I was fortunate to work, Franco Vizeacoumar, Andrew Freywald, Gao Zhang, Meenhard Herlyn, Keith Flaherty, Genevieve Boland, Georg Emons, Ruediger meyer, Charlotte Sumner, Daniel Ramos and Sylvia LeDevedec.

Many thanks for the NCI-UMD fellowship for supporting me for the past two years.

Finally, a million thanks my husband, my parents and my sister for their unconditional support, love and care.

Table of Contents

Preface ii

Acknowledgements v

Table of Contents vi

Background viii

List of Tables xiii

List of Figures xiv

Chapter 1: Identification tumor modifying metabolic genes. 1

Introduction 1

Results 2

Pipeline: 2

FUT9 tumorigenic properties 4

Experimental work 17

Methods 24

Discussion 37

Chapter 2: Cancer pathway modifiers. 41

Introduction 41

Results 41

MCF pipeline 41

MCF predictive performance 45

MCF prediction of patients' survival 51

Methods 55

Discussion 63

Chapter 3: Cancer immunotherapy treatment modifiers 65
Introduction 66
Results 66
NB Spontaneous regression and ICB response in melanoma 66
IMPRES predictor 70
Methods 80
Discussion 84
Chapter 4: Cancer chemoradiotherapy treatment modifiers 85
Introduction 85
Results 86
Identification and cross validation 86
Validating with an independent datasets 88
Colorectal cancer patient survival prediction 90
Chapter 5: Monogenetic disorders modifiers 92
Introduction 92
Results 93
GENDULF pipeline 93
Cystic Fibrosis 95
Spinal Muscular Atrophy 99
Discussion 103
Bibliography 105

Background

Identification of cancer driver genes

Major tumor sequencing projects have been conducted and initiated in the past few years to identify genes that are frequently mutated and thereby are expected to have primary roles in the development of tumor [1–3]. Most common methods identify genes that are mutated more frequently than expected from the background mutation rate [4,5]. Other methods attempt to identify genes that exhibit other signals of positive selection across tumor samples, such as a high rate of non-silent mutations compared to silent mutations [5,6]. Nevertheless, driver genes mutated at low frequency are still difficult to detect with this approaches. Other methods hence attempt to identify genes that exhibit other signals of positive selection across tumor samples, such as a high rate of non-silent mutations compared to silent mutations [2,7] or a bias towards the accumulation of functional mutations (FM bias) [8].

Genome-scale metabolic modeling (GSMM) approaches to study human metabolism and cancer

A Genome Scale Metabolic Model (GSMM) is a computer program built around a set of reactions that comprise a metabolic network, accompanied by a mapping of genes and proteins to the reactions they catalyze within the network [9]. GSMM of human metabolism has become feasible in recent years thanks to the publication of the first full-fledged genome-scale human metabolic models (Recon1[10,11]). In addition to a network of more than 3000 metabolic reactions, Recon1 contains Boolean mappings

of approximately 1500 metabolic genes through their encoded enzymes to these reactions, sub-cellular compartmentalization of processes and pathways, and manually curated reaction stoichiometry and membrane transporters. A key critical merit of GSMM modeling is that it does not require the explication of detailed enzymatic kinetic information (which is yet unknown on a network scale) as it describes the metabolic state of cells at steady state. GSMM enables the integration of omics data collected at specific conditions to provide a genome wide view of their corresponding metabolism; that is, the prediction of the likely metabolic fluxes across the network, including uptake and secretion rates, cell proliferation and more. GSMMs can also be used to predict the phenotypic effects of genetic and environmental perturbations on the cell's flux distribution and viability. Such modeling studies have been employed in recent years to describe human metabolism [10] in general and in cancer [12–16].

Aggregating metabolic pathway information for cancer classification

Metabolism is universally conceptualized through the abstraction of *pathways*, which are groups of enzymatic reactions thought to operate coherently [17]. Undoubtedly, this abstraction is very useful and underlies many studies [18]. Hu et al. [19] showed that changes in the aggregate expression of canonical metabolic pathways that occur in individual tumors are reproducible in independent samples of the same tumor. On the other hand, it has also been observed that the canonical pathways abstraction does not capture the complexity of the metabolic network in full; Bordbar et al. [20] recently presented an algorithm for deriving metabolic pathways based on the principle of

parsimonious use of cellular components. They showed that it produces pathways that are more biologically plausible than the human defined 'canonical' pathways present in databases such as KEGG, EcoCyc, YeastCyc, and Gene Ontology.

There has been a considerable interest in cancer classifiers that utilize network- and pathway-based meta-features [21–24]. However, recent studies reported that many of these classifiers do not outperform models trained over single gene features [25–27].

Checkpoint blockade immunotherapy in cancer

Cancer immunotherapy using immune-checkpoint blockade (ICB) has created a paradigm shift in the treatment of advanced-stage cancers. The promising antitumour activity of monoclonal antibodies targeting the immune-checkpoint proteins CTLA-4, PD-1, and PD-L1 led to regulatory approvals of these agents for the treatment of a variety of malignancies. Patients might experience clinical benefits from treatment with these agents, despite unconventional patterns of tumour response that can be misinterpreted as disease progression, warranting a new, specific approach to evaluate responses to immunotherapy. However, only a subset of patients benefit from these treatments, while others may incur considerable side-effects and costs. Hence, predicting the patients' responsiveness to ICB is being extensively investigated in recent years.

Predicting clinical outcome of cancer and identification of prognostic biomarkers

It has been previously established that gene expression profiling can be used to predict the clinical outcome in different cancers; e.g., predicting patients survival in

breast cancer [28], predicting recurrence of treated patients [29] and predicting distant metastasis [30].

In rectal cancer, many studies have attempted to identify a clinically useful and reproducible gene expression signature capable of predicting response to neoadjuvant chemo radiation treatment (nCRT) using microarrays [31–36]. Most studies have focused on the identification of predictive signatures to distinguish "good" responders from "bad" responders and were primarily interested in the identification of patients who would benefit the most from nCRT and spare others from the potential toxicity of CRT. However, definition of "good" response to nCRT may not be straightforward; significant variations in definitions of responders and non-responders, in addition to the intrinsic subjectivity of these definitions, may be critical in this setting. Moreover, most of these studies include only few dozen of patients, perform the feature selection alongside with the training procedure and have very small test set. As a result, many of the identified signatures were found not reproducible and none these has been integrated into the clinic for prognostic use to this day [37].

Identification of genetic modifiers for monogenetic disorders

Strategies used to show the role of genetic factors in phenotypic expression are often classified into three categories depending on the type of data available [38]: (1) Association studies of case-control data, which is the most widely used strategy in the search for modifier genes, probably as it requires sampling patients only, rather than collecting familiar data. In association studies, the distribution of marker genotypes is compared in patients with different levels of the phenotype [39,40]. (2) Linkage studies,

which require available data from affected siblings. Linkage analysis compares the number of alleles shared identical by descent by affected siblings between phenotypically-concordant and discordant sibling pairs [41,42]. (3) Blind search - Systematic genome-wide screens, which consists in searching for the genetic factors involved in the phenotype of interest over the whole genome, to identify individuals that are resilient to mutations causing the phenotype of interest [43].

List of Tables

Table 1. Predicted tumor suppressors properties.

Table 2. Target metabolites selected for MCF.

Table 3. Summary of the datasets utilized for five cancer types.

Table 4. Response annotations for each melanoma dataset

List of Figures

Figure 1. Two-step pipeline for predicting metabolic tumor suppressors.

Figure 2. Tumorigenic attributes of FUT9.

Figure 3. Knockdown of FUT9 expression increases aggressiveness of colon cancer cells.

Figure 4. Expression of FUT9 supports tumor development.

Figure 5. Overview of the MCF algorithm.

Figure 6. Comparing the performance of MCF to MGE-SVM across integrated cancer-type datasets.

Figure 7. MCF survival prediction.

Figure 8. NB regression association with melanoma immune response.

Figure 9. IMPRES performance.

Figure 10. IMPRES features.

Figure 11. Cross validation performance.

Figure 12. Performance for two independent datasets

Figure 13. Patients score of response to CRT predicts survival in two independent datasets of colorectal cancer.

Figure 14. An overview of GENDULF computational pipeline.

Figure 15. CF identified modifiers.

Figure 16. U2AF1 gene.

Chapter 1: Identification tumor modifying metabolic genes.

Published as "An integrated computational and experimental study uncovers FUT9 as a metabolic driver of colorectal cancer", Molecular Systems Biology 2017[44]

Introduction

Altered metabolism is a core hallmark of cancer and yet, surprisingly, very few metabolic cancer genes are known to play a causal role in tumorigenesis. Here we present an integrated computational approach that combines a large-scale genomic analysis with a genome-scale metabolic modeling (GSMM) approach to identify new metabolic tumor suppressor genes. At the first step, our computational pipeline uses standard genomic approaches to identify potential candidates presenting tumorigenic molecular properties in patients' tumors. In a second step, we present a new GSMM method that identifies a subset of these genes that are likely to play a causal role in transforming the metabolic state of healthy colon tissue to a cancerous one. Our analysis predicts FUT9, as a causal metabolic driver of advanced stage colon cancer whose inhibition is predicted to modify the tumorigenic metabolic state from that of early colorectal tumors to that of late ones. The experimental testing of FUT9 inhibition reveals its complex dual role in this malignancy; while the knockdown of FUT9 enhances proliferation and migration of the bulk of colon cancer cells in monolayers, it suppresses colon cancer cells expansion in tumorspheres and inhibits tumor development in a mouse xenograft models, testifying on its context dependent modifying role in this malignancy.

Results

Pipeline:

An integrated genomic-modeling analysis predicts a modifying causal role of FUT9 in colorectal cancer [44]

We developed a two-step computational approach to predict metabolic tumor suppressors, that is, genes whose downregulation promotes cancer. Applied to study colon cancer, the first step employs a straightforward genomic analysis of the Cancer Genome Atlas (TCGA) database[45,46] to identify metabolic genes that are downregulated in colorectal cancer (Figure 1A). Subsequently, we performed a novel metabolic modeling analysis to identify, among the genes identified as *associated* with tumorogenesis in the first step, those whose downregulation is indeed most likely to result in the metabolic alterations observed in colorectal tumors and thus are more likely to play an actual *causal* role in the transformation of normal to cancerous tissues (Figure 1B). A detailed overview of each step follows.

Genomic identification of 34 candidate metabolic tumor suppressor genes in colorectal cancer: This step consists of three sub-steps that are applied sequentially, analyzing gene expression, Copy Number Variation (CNV), and survival data from 272 colorectal cancer samples and 42 matching healthy colon tissues samples in the TCGA[45,46]: (1) First, analyzing the transcriptomics data of these samples we identified 4593 genes that are significantly downregulated in colon cancer (one-sided Wilcoxon Rank-sum

test with multiple hypothesis correction (alpha=0.001)). (2) Second, 328 of these downregulated genes have significantly lower copy number in the tumors compared to the healthy samples (Q-values < 0.25). (3) Finally, a Kaplan Meier survival analysis further narrowed down this list to 177 candidate tumor suppressors whose downregulation is negatively correlated with patient survival (and thus, likely to enhance tumor progression; see Methods, Figure 1A). Reassuringly, the resulting list includes several known colon tumor suppressors such as APC[47,48], TCF7L2[49,50], MCC[51], PTEN [52,53], and SMAD4 [54,55]. It also includes 34 metabolic genes that are present in the human metabolic model, and which we further studied in the next modeling step.

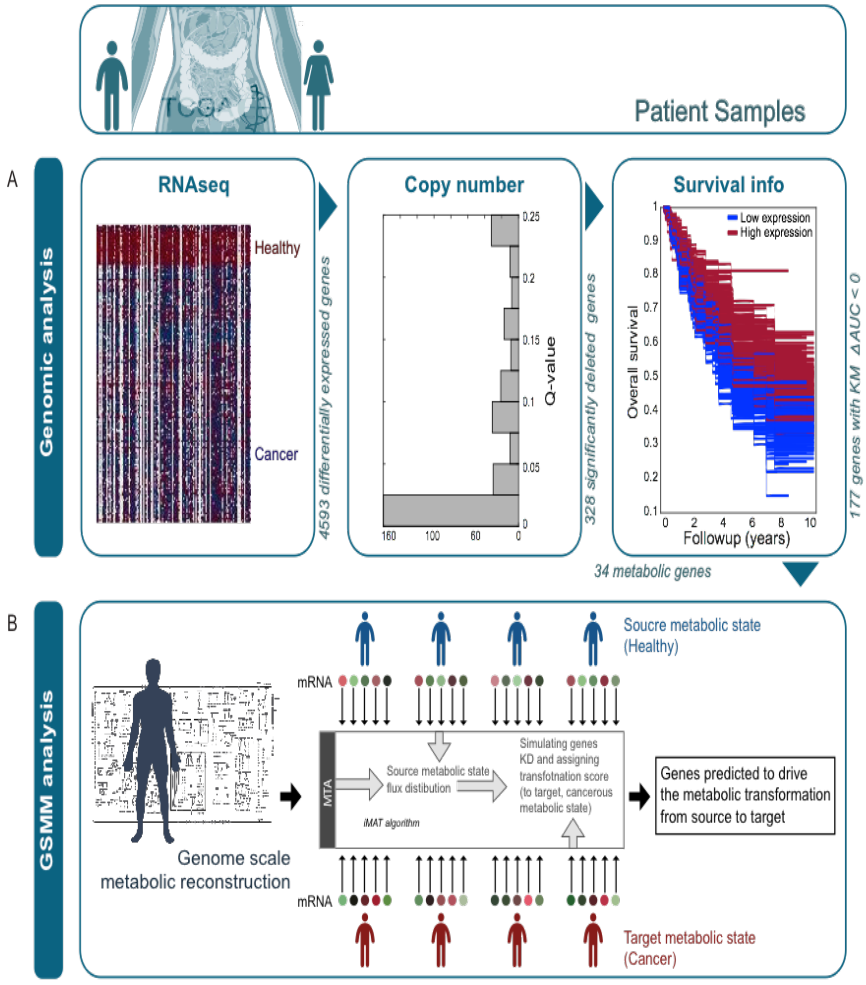

Figure 1. Two-step pipeline for predicting metabolic tumor suppressors. (A) Genomic analysis of three types of data yields an initial list of potential tumor suppressors. (B) GSMM-based approach of the potential tumor suppressors identifies metabolic genes whose knockdown may play a causal role in tumorogenesis.

FUT9 tumorigenic properties

To predict metabolic genes whose downregulation may play a causal role in colorectal cancer, we utilized a GSMM analysis approach termed the

Metabolic Transformation Algorithm (MTA)[56]. This algorithm was previously developed and used to successfully identify life-extending metabolic genes in yeast [56] and is employed here for the first time to search for metabolic tumor suppressors in cancer. MTA is a generic algorithm that aims to identify metabolic gene knockouts that are capable of driving a transformation from a given metabolic state to another, defined target state. The inputs to MTA are the pertaining transcriptomic measurements of these two given and targets states. Its output is a ranked list of metabolic genes whose inactivation has the potential to induce the transformation from the given to the target states (Methods)[56]. In our case, the given metabolic state is the healthy, non-malignant state, and the target state is the cancerous one, and correspondingly, the inputs to the algorithm are a set of gene expression data from matched healthy and tumor colon samples.

While the original publication of MTA has mainly focused on its testing and validation in a known collection of gene knockouts in microorganisms, it already showed that MTA correctly identifies fumarate hydratase as a gene whose knockdown may cause the metabolic transformations observed in HLRCC[57,58]. We now tested and validated that MTA successfully identifies the knockdown of succinate dehydrogenase (SDH) as a likely cause of the metabolic alterations observed in hereditary paraganglioma[59]. To further test the ability of MTA to identify the genes that where knocked down in mammalian screens from the pertaining transcriptomics data, we further

mined the literature to assemble a collection of 19 datasets of metabolic genes for which we found mouse or human gene expression data before and after the knockdown of each of these genes. For each of these knockdowns, we gave MTA these transcriptomics data as inputs and applied it to predict the most likely genes whose knockdowns may account for the transcriptomic changes observed in these experiments. MTA correctly predicted the experimentally knocked down genes in 13 of the 19 cases studied in the top 20% of the predictions (binomial P-value = 5.8266e-06, and its performance remains robust at multiple threshold setting, Appendix), validating MTA's predictive ability in mammalian tissues.

We then turned to apply MTA to identify metabolic genes that, when downregulated, can transform a healthy tissue to a cancerous one. We analyzed three independent transcriptomic datasets including 27 paired healthy/tumor samples from TCGA, 17 paired healthy/tumor samples from Khamas et al.[60] and 32 paired healthy/adenoma samples from Sabates-Bellver et al.[61]. In the first step, we ran an MTA analysis on each pair of matched healthy and tumor gene expression samples, yielding a ranked list of genes according to their *oncogenic transformation scores (OTS)* (Methods). OTS scores denote the likelihood that a gene knockout in the healthy cells can transform their metabolic state to a cancerous one. Following that, in a second step, an aggregate OTS was assigned to each metabolic gene by considering its scores across all samples and then, in a third step we aggregated the OTS

scores of each gene across all three datasets analyzed. We additionally analyzed colon polyp data from Sabates-Bellver *et al*[61], which includes 32 matched healthy and polyp samples. This data enabled us to perform two complementary MTA analyses, one predicting metabolic genes whose knockdown may cause the transformation to the polyp state, and one predicting metabolic genes whose inactivation may cause a further malignant transformation into colon cancer. (Methods).

The distribution of the resulting OTS scores of the 34 metabolic genes examined via these MTA analyses is presented in Table 1. While all 34 genes present genomic patterns that associate them with a tumorigenic state (using expression, copy number and survival data), only few are predicted by MTA to causally transform the metabolic healthy state to that of a cancerous one. As evident, only the knockdown of PTEN and FUT9 is predicted to transform the metabolic state of healthy cells as well as that of adenoma cells to that of colorectal tumors with high OTS scores (Methods). FUT9 is the most highly scored gene and is also strongly supported by the earlier genomic analysis: Its expression is strongly downregulated in colon cancer (Rank-sum P-value = 1e-22, Figure 2A), it is significantly deleted in colon cancer while not in other cancer types (Q-value = 0.0356, Figure 2B), its low expression is associated with poor survival in colon cancer (Kaplan-Meier (KM) ΔAUC = -0.1206, Figure 2C) (Table 1) (The resulting KM log-rank P-value is 0.1942, likely due to the small sample size of patients expressing FUT9 (only ~15% of

patients)). Interestingly though, while MTA highly scores FUT9 for all three transformations, FUT9 is not significantly downregulated at early stage colon adenomas using paired gene expression of healthy/adenoma samples from Sabates-Bellver et al.[61] (Paired student t-test, P-value = 0.47, Appendix Figure S1). This suggests that its inactivation may play a significant role only at later stages of colon cancer progression. Bearing this observation in mind, we set to study the role of FUT9 further, first computationally and then experimentally.

Gene	$\sum$ healthy → cancer (	Healthy→ adenoma OTS score	Adenoma →cancer OTS score	Differential expression P-value	CN Q-value	KM ΔAUC
FUT9	8.54	3.02	2.99	5.06E-24	0.0356	-0.120669976
AKR7A2	6.91	4.55	0.06	2.15E-14	3.46E-05	-0.198482955
CAT	5.78	0	0	5.76E-19	0.215	-0.124211074
PTEN	4.91	0.09	2.67	2.08E-19	0.00494	-0.009581467
PIK3CD	4.3	0	0.2	1.79E-11	0.00205	-0.048812

						134
FUCA1	4.07	0	0	1.27E-23	3.46E-05	-0.114652506
PLCE1	3.47	1.3	0	3.33E-23	0.0458	-0.104694133
STS	2.86	0	0.1	1.49E-08	0.0136	-0.080255382
SDHB	2.81	2.4	0	3.72E-16	3.46E-05	-0.106186688
MAN1C1	2.6	0	0.2	9.61E-12	3.78E-05	-0.023167238
MTHFR	2.14	0.21	0	2.06E-14	0.00205	-0.162335156
PIGN	2.1	0	0	1.41E-09	0.187	-0.029340173
FH	2.03	1.73	1.2	3.27E-11	1.24E-68	-0.033206093
PLA2G2D	1.66	0.9	0	3.48E-09	0.000147	-0.010922122
SLC18A2	1.48	0	0.73	3.27E-15	0.19	-0.095652366

LIPC	1.22	0.9	0	1.26E-19	0.215	-0.005134395
CYP2C18	1.2	1.64	0	1.43E-12	0.09	-0.048710276
HMGCL	1.2	1.09	0	1.40E-20	3.46E-05	-0.118954367
ACADS	1.11	2.4	0.12	2.73E-24	0.102	-0.065005732
PANK4	1.02	2.11	1.2	8.29E-13	0.0341	-0.044198756
COX6B2	0.82	0.76	0.1	1.88E-11	0.0397	-0.024199841
PDE4D	0.8	2	0	1.01E-17	0.00629	-0.076789143
ECHS1	0.71	1.2	0	5.59E-12	0.172	-0.054802279
INPP5A	0.32	3	0	2.59E-22	0.0599	-0.031649119
ITPKA	0.2	1.01	0.8	2.28E-18	0.172	-0.077477859
SLC25A4	0.2	0.2	0	4.23E-12	0.00872	-

						0.227357666
HS3ST5	0.11	0	0	1.28E-09	0.0676	-0.03426555
FECH	0	1.53	0	4.80E-17	0.227	-0.102888235
ME2	0	1.88	0	6.93E-17	0.0273	-0.083078298
NADK	0	0.96	0.01	2.27E-14	0.0815	-0.172031059
NDUFB8	0	1.5	0	2.10E-11	0.234	-0.023126419
NMNAT1	0	0.98	0	8.38E-19	0.0062	-0.124354125
PAFAH2	0	0.1	0	3.67E-21	3.47E-05	-0.021012791
PC	0	2.76	0	1.44E-16	0.0341	-0.214743894

Table 1. Predicted tumor suppressors properties. For each metabolic predicted tumor suppressor, the table displays: (1) the OTS scores for the three transformations, and genomic properties (2) differential expression P-value,

(3) Copy Number (CN) deletion Q-value (P-value that has been adjusted for the False Discovery Rate), and (4) Kaplan-Meier survival ΔAUC.

GSMM analysis of the metabolic implications of FUT9 inactivation: FUT9 belongs to the glycosyltransferase family and catalyzes the last step in the biosynthesis of Ley glycolipids in the carbohydrate antigen Lex[62,63]. This reaction takes place in the Golgi compartment, and the product is transported to the cytosol and secreted out from the cell[64]. The Ley glycolipid was previously reported to inhibit the procoagulant activity and metastasis of human adenocarcinoma[65–67]. The loss of FUT9 in the metabolic model prevents Ley glycolipid formation and secretion. To chart the network-wide metabolic alterations induced by FUT9 inactivation, we performed a Minimization Of Metabolic Adjustment (MOMA)[68] analysis to predict the metabolic state after FUT9 KD in late stage colorectal cancers, simulated by the Gene Inactivity Moderated by Metabolism and Expression (GIMME) algorithm [69] (Methods). This pinpoints reactions whose flux is predicted to be most afflicted by FUT9 inactivation in advanced stage cancer. We found that the loss of FUT9 in late stage colorectal cancers is predicted to cause an increase in the flux of 25 reactions, and a decrease in the flux of 6 reactions. The flux is predicted to increase in reactions associated with Glucose metabolism, and particularly TCA cycle (hyper-geometric P-value = 1.3676e-09, Figure 2D). We find that the expression of metabolic genes associated with reactions predicted to increase following FUT9 loss is significantly up-

regulated in stage 4 vs. stage 3 colon tumors when compared by their expression in TCGA data (hyper-geometric P-value = 0.0046). Experimental evaluation of these predictions using the Human Glucose Metabolism, RT² Profiler™ PCR Array revealed a good correlation with our computational prediction (Fig. 2D). In particular, 12 genes, including FH and SDHD proved to be upregulated in FUT9 silenced cells as expected from our computational analyses.

To evaluate the effect of FUT9 knockdown (KD) and overexpression (OE) on biomass production, Glucose consumption, Lactate production and Oxygen consumption in the benign colon adenoma state, we (1) simulated the wild-type metabolic state associated with colon adenoma. This was done by incorporating adenoma gene expression data from Sabates-Bellver et al.[61] using the GIMME algorithm. (2) We then sampled 100 flux distributions in the resulting predicted adenoma wild-type state. In each such sample we applied the MOMA[68] algorithm to predict the metabolic state after FUT9 KD and OE in adenoma, summing up the results overall 100 samples (Methods). We find that the biomass production predicted is significantly higher under FUT9 OE than its KD, as well as Lactate secretion rate (Wilcoxon rank-sum P-value = 0.0081 and 0.0173, respectively, Figure 2E). While Oxygen consumption rate is significantly higher under FUT9 KD (Wilcoxon rank-sum P-value = 6.79e-8, Figure 2E). These predictions imply that FUT9 activity is required for supporting cancer proliferation in the adenoma state, which are

consistent with the genomic findings we reported above that while FUT9 expression is strongly downregulated in colon cancer is not significantly downregulated at early stage colon adenomas.

We next evaluated the metabolic effects of FUT KD and OE in the colon tumor state. To this end we performed a similar analysis as described above for adenoma, while first inferring the likely metabolic state of colon tumors (Methods). Strikingly, we find that the predicted biomass production in the cancerous state is significantly higher under FUT9 KD than its OE (Wilcoxon rank-sum P-value = 0.0245, Figure 2F), and that lactate production rate is also increased under FUT9 KD (Wilcoxon rank-sum P-value = 0.0859, Figure 2F), opposite to the observed in simulated colon adenoma state. These predictions imply that the loss of FUT9, while hampering the growth of adenomas, is required for the proliferation of colon tumors, while its overexpression significantly reduces proliferation in that state.

Given the opposite predicted effects of KD perturbation in colon adenomas vs. tumors, we performed an additional GSMM analysis to study whether FUT9 inactivation at early colorectal cancer stages can induce the metabolic state observed at advanced tumors, or only its inactivation at late stages can induce this transformation. To this end we first inferred the likely metabolic state of advanced colorectal tumors using the GIMME algorithm[69], as done above in the adenoma analysis. We then predicted the likely metabolic states after the

loss if FUT9 in each of the four different stages of colorectal cancer progression, asking how similar is the metabolic state induced after the loss of FUT9 in each of these stages to the advanced, late cancerous state. The metabolic state after the KD of FUT9 in each stage-specific context was predicted using the MOMA algorithm[68] (Methods). This analysis revealed that the loss of FUT9 at early stages does not bring the metabolic state close to that observed in advanced cancer. Rather, for the FUT9 loss to cause such an effect, it has to occur in later stages of the disease (Figure 2G). This indicates that FUT9 downregulation is a tumor-transformative event only if occurs at later stages of tumor progression. To study this further from a genomic perspective, we analyzed the correlation between FUT9 copy number and the copy number levels of known early and late genetic markers of colorectal cancer. We find that FUT9 expression levels negatively correlate with the loss of the early markers APC and MCC (Spearman rho = -0.1726 and -0.1707, P-value < 0.05, respectively), while it is positively correlated with the loss of TP53, a marker of the advanced stage, [70,71] (Spearman rho = 0.1759, P-value <0.05, Figure 2H).

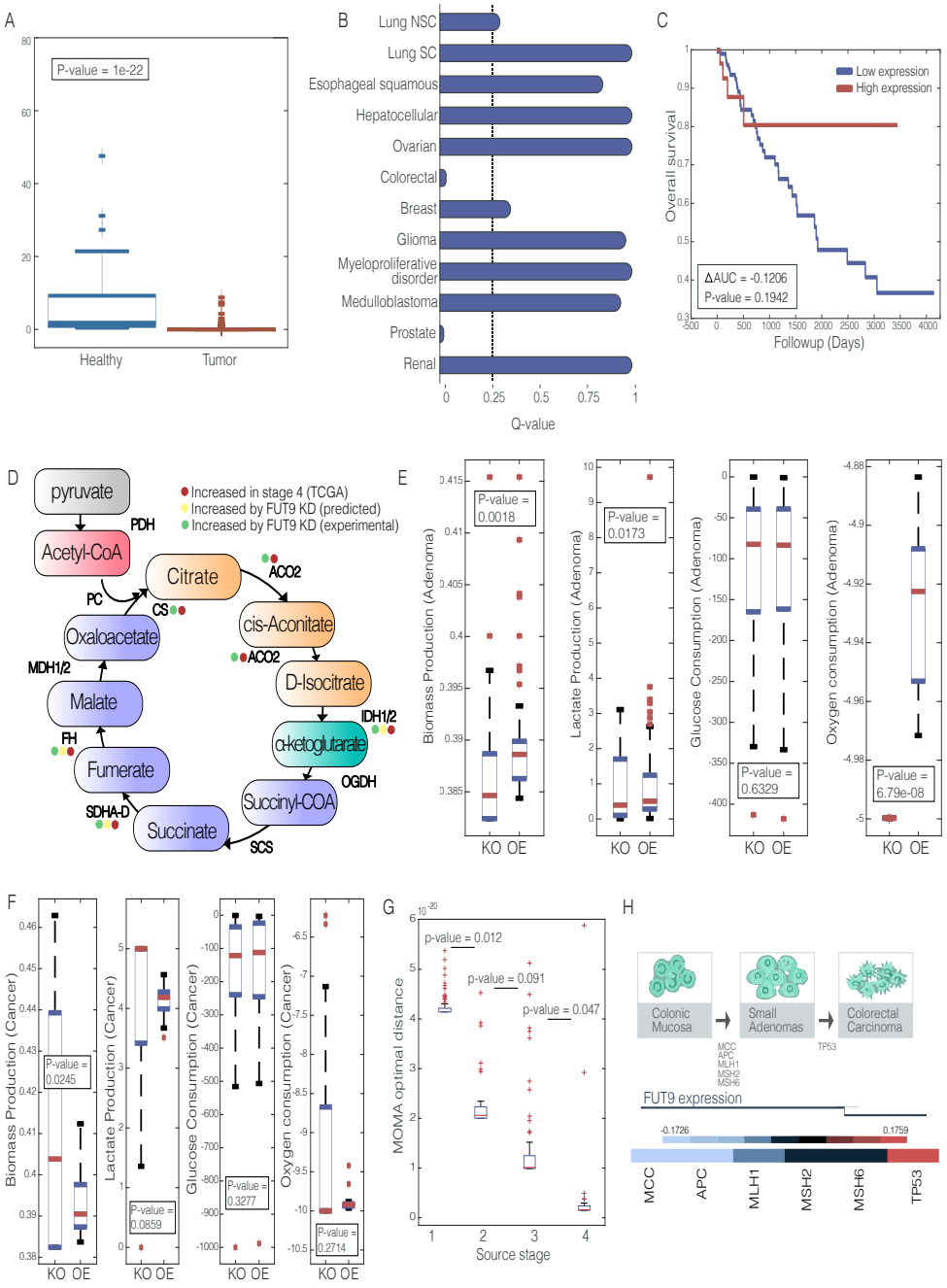

Figure 2. Tumorigenic attributes of FUT9. (A) A boxplot describing the expression of FUT9 in tumor vs. healthy colon tissues. (B) Q-value for CN of FUT9 in 12 different cancer types, the dashed line represents a significance threshold of 0.25. (C) Kaplan-Meier survival curve for FUT9 expression (top and bottom 0.5 quartiles). (D) The TCA cycle and its associated enzymes that are increased in stage 4 colorectal cancer (red), predicted to increase following

FUT9 KD (yellow) and increase following FUT9 KD experimentally (green). (E) Boxplot showing the distribution of biomass production, Glucose consumption, Lactate production and Oxygen consumption in adenoma state when FUT9 is knocked-down (KD) and overexpressed (OE). (F) Boxplot showing the distribution of biomass production, Glucose consumption, Lactate production and Oxygen consumption in cancer state when FUT9 is knocked-down (KD) and overexpressed (OE). (G) Boxplots sowing the MOMA scores obtained by the knock-down of FUT9 in stages 1-4. (H) Upper panel: Colorectal Adenoma-carcinoma sequence. Middle panel: the emerging role of FUT9 in colorectal tumor progression. Lower panel: Correlation heat-map of FUT9 copy number (CN) and early and late stage prognostic markers of colorectal cancer.

Experimental work

The experimental testing of these predictions shows that FUT9 plays a complex dual role in this malignancy. On one hand, the knockdown of FUT9 enhances proliferation and migration of the bulk of colon cancer cells in monolayers, pointing to a suppressive role (Figure 3). On the other hand, its knockdown suppresses colon cancer cells expansion in tumorspheres and inhibits tumor development in a mouse xenograft models, testifying to a tumor promoting role (Figure 4). These results suggest that FUT9's inhibition may have a differential effect on different types of tumor cells: its knockdown attenuates tumor initiating cells (TICs), which are known to dominate tumorspheres and early tumor seeding and growth, but promotes bulk tumor cells. In agreement, we find that FUT9 silencing decreases the expression of the colorectal cancer TIC marker, CD44 (Figure 4). Taken together, these

computational and experimental results testify that FUT9 acts first as an oncogene in TICs and enhances early stages of tumor formation, but later it acts as a tumor suppressor in bulk tumor cells and is hence lost at later tumor stages.

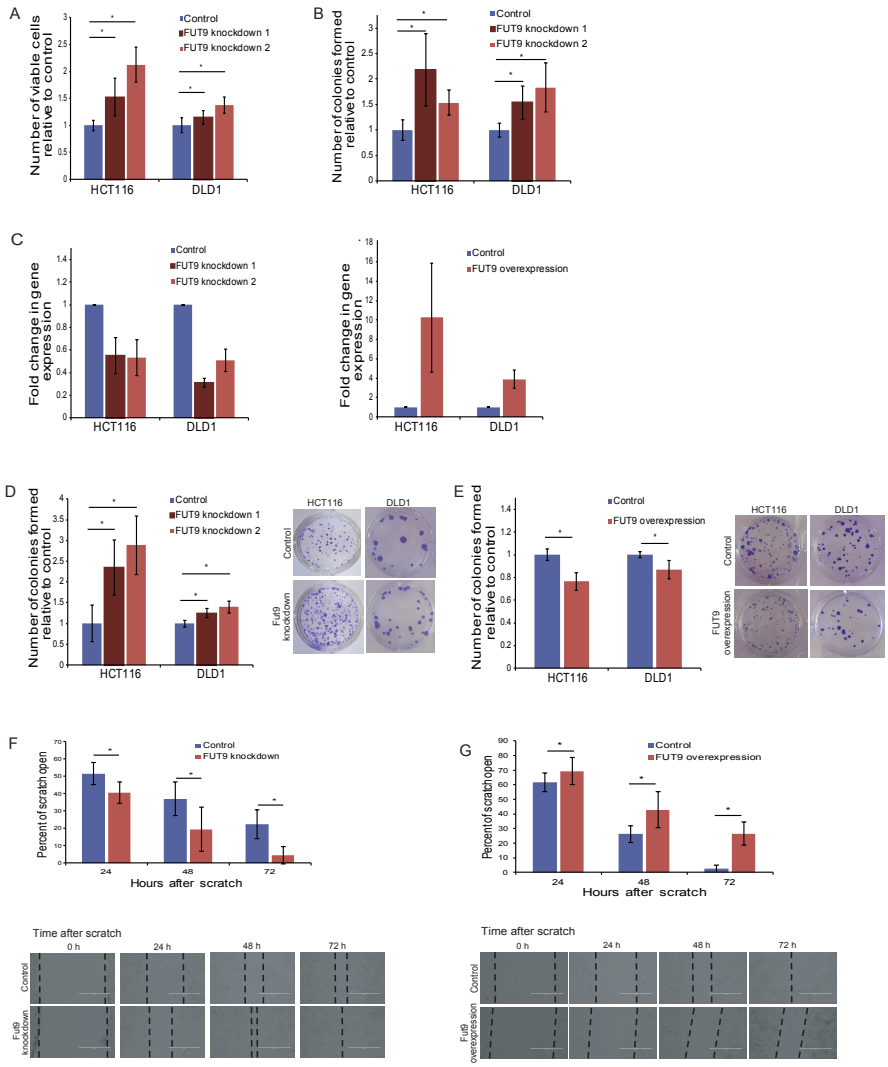

Figure 3. Knockdown of FUT9 expression increases aggressiveness of colon cancer cells. (A) HCT116 and DLD1 control and FUT9 knockdown cells were

seeded evenly in 96 well plates and the number of viable cells after 72 hours was analyzed using Resazurin absorbance reading. The graph represents the mean ± s.e. from three independent replicates normalized to the control cells. Six wells per replicate were analyzed. (B) The same cells from (A) were seeded in soft agar and cultured for 28 days. The number of colonies formed were quantified relative to the control cells. The mean ± s.e. from two independent replicates are represented. (C) The fold change in gene expression for the FUT9 knockdown and FUT9 overexpressing cells were analyzed using RT-qPCR. The graphs represent the mean ± s.e. fold change from three independent replicates. (D) HCT116 and DLD1 FUT9 knockdown cells were seeded at very low densities in a 24 well dish and cultured for 10 days. The number of colonies formed in each well were counted. The graph represents the mean ± s.e. of two independent replicates. Three wells were analyzed per replicate. Representative images of one well for each condition are shown. (E) The same colony formation assay as in (D) was performed and analyzed using FUT9 overexpressing cells. Representative images for each condition are shown. (F) HCT116 control and FUT9 knockdown cells were each seeded to form a confluent monolayer. A scratch was made in each monolayer and the width of the scratch monitored by imaging the same areas of each scratch (2 per scratch) at the time of scratching (0h) and 24, 48 and 72 hours later. The graph depicts the mean ± s.d. of two independent experiments and represents the percentage of scratch open at each time point relative to the 0h point. For optimal presentation, individual scratch images

are shown at different brightness and contrast settings. (G) The wound-healing assay was performed with HCT116 control and FUT9 overexpressing cells and analyzed as in (F). The graph summarizes the mean ± s.d. of two independent experiments and represents a percentage of scratch open at each time point relative to the 0h point. For optimal presentation, individual scratch images are shown at different brightness and contrast settings. * $P < 0.05$, Student's t-test

****Figure 3 and the work presented in it is generated by the Franco J. Vizeacoumar and his lab members* [44]

Our genomic analysis revealed that, while FUT9 is strongly downregulated at later stages of colon cancer development, it is still present in colon polyps and early adenoma**,** indicating that FUT9 activity may be required at the initial stages of tumor initiation. Thus, while FUT9 downregulation benefits the bulk of tumor cells as shown above, its activity may support the subpopulation of cancer stem cells or tumor initiating cells (TICs) that play a central role in tumor development. To study this hypothesis, HCT116 with FUT9 knockdown and matching control cells were cultured as tumorspheres, which are predominantly formed by TICs [72–75]. Consistent with our expectations, FUT9 knockdown reduced expansion of HCT116 cells in tumorspheres, while FUT9 overexpression produced enhanced proliferation of tumorsphere-forming cells **(**Figure 4A-B). On a molecular level, this was accompanied with the reduced expression of OCT4 transcription factor in FUT9 silenced cells.

Since OCT4 has been shown to support TIC formation [76,77], this observation provides a mechanistic explanation for FUT9 effect in supporting TIC activity. These results show that, in contrast to the anti-proliferative effects of FUT9 activity in the bulk of colon cancer cells (Figure 3A-D**)**, FUT9 activity may actually be required for the efficient expansion of TIC populations. This was further confirmed by flow cytometry analysis, showing that FUT9 silencing decreases the expression of a prominent colorectal cancer TIC marker CD44 [72,78] in HCT116 cells (Figure 4C).

Since TIC cells are essential for tumor initiation, tumor maintenance and tumor growth[79–84], increased TIC activity is expected to accelerate tumor growth *in vivo*[79,83,84]. To test the effect of FUT9 on this process, we generated a xenograft model of colorectal cancer in immune-deficient NOD/SCID gamma mice. HCT116 cells with silenced FUT9 expression or control cells transduced with non-targeting *sh*RNA were injected subcutaneously in equal numbers into the flank of the immuno-deficient mice and the growth of the resulting tumors was monitored. In agreement with its inhibitory effect in tumorspheres, FUT9 silencing also significantly reduced growth of xenograft tumors (Figure 4D). This may reflect the dual functionality of FUT9 where it supports tumor development by enhancing TIC activity (Figure 4E), while inhibiting the expansion of the bulk of tumor cells (Figure 3).

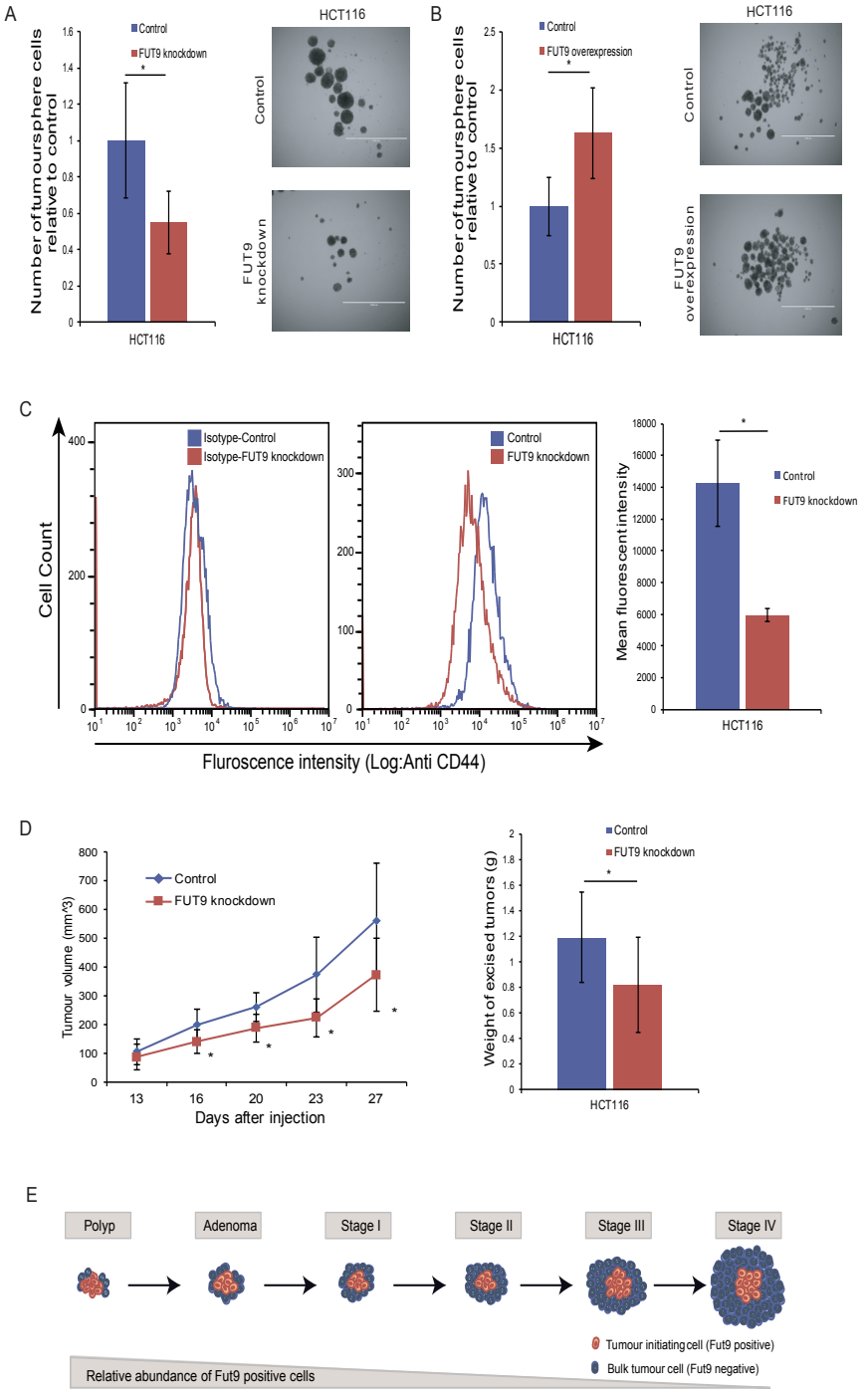

Figure 4. Expression of FUT9 supports tumor development. (A) HCT116 FUT9 knockdown and matching control cells were seeded in ultra-low attachment plates and cultured for one week. The resulting tumorspheres were

collected, dissociated, and the total number of cells counted. The graph represents the mean ± s.e. of two independent replicates normalized to the number of control cells. Each replicate represented tumorsphere cells collected from 24 independent wells. Representative images are shown. Scale bar, 1000 μm. **(B)** FUT9 overexpressing and control cells were cultured as tumorspheres and analyzed as in (A). Two independent replicates and representative pictures are depicted. **(C)** CD44 expression in FUT9 knockdowns (in red) and *sh*RFP control (in blue) in HCT116 cells were assessed using anti-CD44 and flowcytometry and representative histograms were overlayed (second panel). Isotype controls were also plotted and overlayed (first panel). Median fluorescent intensity (MFI) values derived from the software are plotted as bar chart. The graph represents the mean ± s.e. of two independent replicates. (D) HCT116 FUT9 knockdown or control cells were injected subcutaneously into the right flank of NOD/SCID mice and monitored for tumor formation. Each tumor was measured using calipers and the mean volume for the FUT9 knockdown and control tumors were graphed (first panel). The graph represents two independent experiments with a minimum of 11 mice analyzed per experimental condition. Mean tumor volumes ± s.d. are shown. Upon experiment termination, tumors were extracted, weighed, and the mean tumor weights ± s.d. are shown in the second panel. **(E)** A schematic showing the abundance of FUT9 positive cells over the course of colon cancer development. * $P < 0.05$, Student's t-test

****Figure 4 and the work presented in it is generated by the Franco J. Vizeacoumar, Andrew Freywald and their lab members* [44]

Methods

Kaplan-Meyer survival analysis of potential tumor suppressors.

For each gene found to be significantly lowly expressed and deleted through gene expression and copy number data, we applied Kaplan-Meyer survival analysis to examine the association of its downregulation with poor patient survival. We use TCGA COAD survival and gene expression data, and separate the expression of each gene to 'high' and 'low' bins by its median level. We calculate the ΔAUC resulting from the two Kaplan-Meyer curves and select only genes with $\Delta AUC < 0$ indicating that their low expression is associated with poor survival.

A constraint-based model of metabolism.

A metabolic network consisting of m metabolites and n reactions can be represented by a *stoichiometric matrix* S, where the entry S_{ij} represents the stoichiometric coefficient of metabolite i in reaction j. A Constraint-Based Model (CBM) imposes mass balance, directionality and flux capacity constraints on the space of possible fluxes in the metabolic network's reactions through a set of linear equations

$$S \cdot v = 0 \qquad (1)$$

$$v_{min} \leq v \leq v_{max} \qquad (2)$$

Where v is the flux vector for all reactions in the model (i.e. the *flux distribution*). The exchange of metabolites with the environment is represented as a set of *exchange (transport) reactions*, enabling a pre-defined set of metabolites to be either taken up or secreted from the growth media. The steady-state assumption represented in Equation (1) constrains the production rate of each metabolite to be equal to its consumption rate. Enzymatic directionality and flux capacity constraints define lower and upper bounds on the fluxes and are embedded in Equation (2). In the following, flux vectors satisfying these conditions will be referred to as feasible steady-state flux distributions. Gene knockdowns are simulated by constraining the flux through the corresponding metabolic reaction to zero. Similarly, environmental perturbations are simulated by constraining the flux through the associated exchange reaction to zero.

For each of the dataset analyzed here, we simulated the same media that was used in the experiment (DMEM). For modeling human metabolism we have used Recon1. [64]

Metabolic Transformation algorithm (MTA).

MTA receives as input the gene expression measurement of two distinct metabolic states, termed source and target states. Next the algorithm executes

the following steps: (1) determine the flux distribution that corresponds to the source state using integration Metabolic Analysis Tool (iMAT) ; (2) identify the set of genes whose expression have significantly elevated or reduced between the source and targets states, and the set of genes whose expression remained relatively constant between the states. Next, the algorithm searches for perturbations that can alter the fluxes of the changed reactions in the observed direction, while keeping the fluxes of the unchanged reaction as close as possible to their predicted source state. Finally, MTA outputs a ranked list of candidate perturbations according to their ability to transform from the source to the target metabolic state.

The Transformation Score

Relying on the optimization value obtained by MTA to rank the transformations induced by different perturbations is suboptimal, since the integer-based scoring of the changed reactions is coarse-grained and does not distinguish between solutions achieving large flux alterations and those obtaining flux changes barely crossing the ε threshold. Therefore, we chose to quantify the success of a transformation by a scoring function based on the resulting flux distributions rather than on the optimization objective values themselves. First, we denote the resulting flux distribution obtained in a given MIQP solution (for a given reaction knock-out) as v^{res}. Second, reactions found in R_F and R_B are classified into two groups $R_{success}$ and $R_{unsuccess}$, denoting whether they achieved a change in flux rate in the required direction

(forward or backward) or not. The following scoring function is then used to assess the global change achieved by the employed perturbation:

$$\frac{\sum_{i\epsilon R_{success}} abs\left[\left(v_i^{ref} - v_i^{res}\right)\right] - \sum_{i\epsilon R_{unsuccess}} abs\left[\left(v_i^{ref} - v_i^{res}\right)\right]}{\sum_{i\epsilon R_S} abs(v_i^{ref} - v_i^{res})} \qquad (*)$$

The numerator of this function is the sum over the absolute change in flux rate for all reactions in $R_{success}$, minus a similar sum for reactions in $R_{unsuccess}$. The denominator is then the corresponding sum over reactions in R_S (the reactions which should stay untransformed). Following, perturbations achieving the highest scores under this definition are the ones most likely to perform a successful transformation by both maximizing the change in flux rate for significantly changed reactions, and minimizing the corresponding change in flux of unchanged reactions. Using an alternative scoring function based on the Euclidean distance instead of absolute values yielded similar results.

While we believe that the TS score (Equation (*)) is the right one to pursue from a biological point of view, optimizing it directly is a very difficult mathematical task. To accomplish that one would need to develop a novel optimization algorithm for solving a mixed *non-linear* programming problem, whose objective function is non-smooth and non-differentiable, requiring non-smooth optimization tools. Attempting such a solution directly would greatly complicate the problem as one would need to add many variables and constraints. Furthermore, the specific form of this ratio is actually dependent on the solution itself (as it evaluates $R_{success}$ and $R_{unsuccess}$ separately) making the entire task infeasible. In light of these evident difficulties we have

chosen to take a two-step approach in this study that is sub-optimal but yet tractable. While the wild-type solution always achieves maximal values in terms of the original proxy objective function used in step 3 (by definition), it does not necessarily achieve high transformation scores (step 4). This is because the wild type solution is the least constrained, and hence most of the solutions found in step 3 can be satisfied by achieving only a minimal epsilon change; Those are obviously non-optimal from a biological standpoint as they do not really come close to the desired objective, and hence their TS score (in step 4) is sub-optimal in many of the cases, correctly ruling them out as biologically viable solutions. MTA analysis is established upon learning the regulatory effects of the knockdown of metabolic genes via the direct stoichiometric flux coupling of the reactions they encode to other reactions in the human metabolic network (which are inherently embedded in the reactions stoichiometric matrix it includes).

Aggregated oncogenic transformation scores (OTS).

MTA scores each reaction according to the extent of which its knockout is predicted to cause the observed transformation from normal to cancer. For each reaction i (RXN_i) we define the aggregated OTS score by:

$$OTS(RXN_i) = \sum_{j \in matched\ paires} I_{ij} \times (1 - P(I_{ij} = 1))$$

Where I_{ij} is one when reaction i was scores higher than random (MTA score when no perturbation is simulated) and zero otherwise. $P(I_{ij} = 1)$ is a reaction's probability to be scored higher than random in matched pair j (which is the number of perturbation that are scored higher then no perturbation in pair j). Thus, paired samples in which fewer reactions received a significant score are more heavily weighted.

Reaction-to-gene mapping of OTS.

OTS is assigned to each reaction in the metabolic model. Each metabolic gene is assigned the highest score assigned to one of its associated reactions, using the reaction-to-gene mapping defined by the Recon1 metabolic model.

Colon polyp and colon tumor gene expression normalization.

To apply MTA from polyp to tumor, we applied quantile normalization to the 1496 metabolic genes present in Recon1 metabolic model. We used 27 colon samples from TCGA that were used for the paired-MTA analysis and 32 colon adenoma sample, when the reference distribution is the mean expression of these 1496 metabolic genes across all 272 colon tumors in TCGA.

Utilizing MOMA and GIMME algorithms to predict the pathway-level effect of FUT9 inactivation in late stage colon cancer.

To investigate FUT9 role in tumorigenesis in the metabolic model, we set to discover which metabolic flux alterations are induced by the loss of FUT9 in

late stage colon cancer. To this end, we utilized the GIMME algorithm to simulate metabolic flux of stage 3 colon tumors. To evaluate FUT9 effect on metabolic fluxes at that stage, we then utilize the MOMA algorithm and sample 100 flux distributions with and without FUT9 knockdown. For each reaction, we compare the MOMA sampled flux distributions with and without FUT9 KD using one-sided Wilcoxon rank-sum test. We define the set of reactions that are increased following FUT9 knockdown as reactions whose sampled flux is increased when FUT9 knockdown is simulated vs. WT (Wilcoxon rank-sum P-value<0.05) and the set of reactions that are decreased following FUT9 knockdown as reactions whose sampled flux is decreased when FUT9 knockdown is simulated vs. WT (Wilcoxon rank-sum P-value<0.05).

Utilizing the MOMA algorithm to evaluate the effect of FUT9 knockdown and over-expression on biomass production, Glucose consumption, Lactate production and Oxygen consumption.

To predict the effect of FUT9 levels on Biomass production, Glucose consumption, Lactate production and Oxygen consumption we utilized the GIMME algorithm to simulate metabolic flux of (1) colon adenoma state using the 32 adenoma samples from Sabates-Bellver et al.[61] (2) Colon cancer state using 268 cancerous samples from the TCGA. For each of the adenoma and cancer predicted flux distributions, we sampled 100 flux distributions for FUT9 KD and another 100 for FUT9 OE (defined by setting the lower bound

of FUT9 associated reactions to 80% of their maximum), using MOMA algorithm, aiming to minimize the metabolic adjustments after FUT9 perturbations, from the initial adenoma or cancerous metabolic state. In both cases we set the lower bound of the biomass reaction to be at least 80% of its optimal rate to simulate proliferating cells and restrict variability in the resulting fluxes.

Utilizing MOMA algorithm to predict stage specific context in which the loss of FUT9 is tumorigenic.

To predict the context in which the loss of FUT9 drives the oncogenic transformation, we used colorectal cancer gene expression measurements from the TCGA database. For each sample, we predict a flux distribution using the GIMME algorithm[69] (the mean flux distribution over 100 sample points was used) and the metabolic model in which FUT9 is knocked down. We then predict a flux distribution typical for stage 4 samples (using the GIMME algorithm[69], genes are considered downregulated with FDR corrected P-value <0.05, compared to all other stages). Then, we compute the MOMA score obtained when aiming to minimize the metabolic adjustment from each sample to the metabolic state predicted for stage 4 samples. Finally, we compare the MOMA score distributions obtained for samples in each of the stages (1-4), describing for each such sample the extent to which the KO of FUT9 is predicted to bring the metabolic flux distribution closer to that of

stage 4. A similar analysis was repeated when using iMAT instead of GIMME to predict flux distributions, yielding similar results.

Cell lines and transfections.

HCT116 and DLD1 colon cancer cell lines were selected based on expression data for FUT9. Both cell lines were cultured in McCoy's 5A medium supplemented with (Fisher Scientific, SH3020001) supplemented with 10% (v/v) FBS (Life Technologies, 12483020), 100 units/mL penicillin-streptomycin solution (Thermo Scientific, SV30010) at 37°C with 5% CO_2. HEK293T cells were used to generate lentivirus and cultured in DMEM (Fisher Scientific, SH3024301) containing 10% (v/v) FBS and 100 units/mL penicillin-streptomycin at 37°C with 5% CO_2. Cells were passaged using 0.25% trypsin-EDTA at 70% confluency.

Transfections were done using X-tremeGENE 9 (Roche, 6365809001) as per the manufacturer's instructions. Lentivirus was generated by transfecting HEK293T cells cultured in 100 mm dishes with psPAX2, pMD2.6, and pLKO.1-*sh*RNA or pLX304 expression plasmids. Media was replaced after 24 hours with DMEM containing 2% (w/v) bovine serum albumin (BSA) (Fisher Scientific, BP9703100) and lentivirus was harvested after 24 and 48 hours and pooled.

To generate the FUT9 knockdown cells, HCT116 and DLD1 cells were transduced with lentivirus containing *sh*RNA sequences specific to FUT9. Two *sh*RNA sequences for FUT9 were used, which were transduced

separately or, in subsequent experiments, pooled and transduced together. An *sh*RNA sequence specific to RFP (Sigma) was used as a non-targeting control. For each transduction, 0.5 mL of each *sh*RNA lentivirus was added to $2x10^5$ cells in a 35 mm dish in a final volume of 3 mL with 8 μg/mL of polybrene (Sigma, 107689). Twenty-four hours after transduction, the media was removed and replaced with media containing 2 μg/mL puromycin (Fisher Scientific, BP2956100) for selection. Cells were selected for a minimum of 48 hours before use in experiments. Knockdown cells were passaged a maximum of five times. The FUT9 overexpressing cells were generated by transducing HCT116 cells with lentivirus containing pLX304-FUT9 (DNA SU, HsCD00444887) using the same transduction method as above. After transduction, cells were selected using 4 μg/mL of blasticidin (VWR, 89149-988) for 14 days. Cells were maintained with 1 μg/mL of blasticidin.

Quantitative real-time PCR (RT-qPCR) analysis.

RNA was isolated from cell pellets using RNeasy mini kit (Qiagen, 74104) according to the manufacturer's instructions including DNase treatment (Qiagen, 79254). RNA quantification was performed using a NanoDrop 2000c spectrophotometer (Thermo Scientific) and RNA integrity was verified spectrophotometrically by A260/A280 ratios between 1.8 to 2.0 and A260/A230 ratios greater than 1.7. Equal quantities of RNA were used to generate cDNA using the RT^2 First strand kit (Qiagen, 330401) according to the manufacturer's instructions.

FUT9 expression levels were evaluated using TaqMan real-time PCR gene expression assay (Life Technologies, 4369016 and 4331182, assay ID: Hs00276003_m1). The fold change in gene expression was analyzed using the ΔΔCT method. Human Glycosylation-related gene expression was evaluated using RT2 Profiler human glycosylation PCR array (Qiagen, 330231 PAHS-046ZA) according to the manufacturer's instructions. Data analysis was performed using the ΔΔCT method as described in the manufacturer's web portal (SABiosciences).

Cell viability assay.

Equal numbers of Fut9 knockdown and control cells were seeded in 96 well plates ($5x10^3$ cells per well). After 72 hours, the abundance of viable cells was analyzed using Resazurin (Fisher Scientific, AR002). Resazurin was added to each well at a concentration of 10% (v/v) and the plates were incubated at 37°C and read using SpectraMax M5 microplate reader (VWR) after one, two, three, and four hours. An increased number of viable cells reflects increased cell expansion.

Growth on soft agar.

The ability of FUT9 knockdown and control cells to grow in low-anchorage conditions was determined by seeding cells in a soft agar medium. Cells were trypsinized and $2.5x10^4$ cells suspended in 0.35% agar-media supplemented with 10% (v/v) FBS and 4% (v/v) minimum essential medium vitamin

solution (Life Technologies, 11120052) and layered on a 0.6% agar-media bottom layer in 6 well plates. Cells were allowed to grow for 28 days and colonies were imaged using an EVOS FL Cell Imaging System microscope at 40x magnification (Life Technologies) and the density of colonies was quantified using ImageJ software.

Colony formation assay.

The ability of individual cells to form colonies was shown by seeding a low density of cells (50 to 200 cells per well) in a 24-well culture plate. After ten days, the colonies were fixed with 100% cold methanol for 10 minutes and stained using 1% crystal violet. The numbers of visible colonies were counted.

Wound-healing assay.

Cells were cultured in 6-well plates and allowed to grow to a confluent monolayer. A scratch was made in each well by scraping with 100 μL pipette tip across the cell monolayer (time point zero of the experiment). Wells were rinsed with PBS three times to remove floating cells. The same areas of each scratch (2 per scratch) were imaged at the time of scratch (0 hours), 24, 48, and 72 hours using an EVOS FL Cell Imaging System microscope at 100 x magnification. The width of scratch in each image was measured using PowerPoint software.

Tumorsphere culture and tumorsphere-forming cell counts.

For tumorsphere culture, 2 x 10^3 cells from monolayer cultures were seeded into 96-well Ultra-Low attachment plates (Corning, 07-200-603) in complete Mammocult medium (Stemcell Technologies, 05620), prepared according to the manufacturer's instruction. Cells were cultured for seven days, tumorspheres in each well were imaged with an EVOS FL Cell Imaging System microscope. Tumorspheres were then collected, dissociated, and cells were counted using a hemocytometer. For each replicate in this experiment, tumorspheres from 24 independent wells were collected into a 15 mL tube and centrifuged at 300 x g for 5 minutes. Collected tumorspheres were dissociated into a single cell suspension in 500 µL of pre-warmed Trypsin-EDTA. Cells were washed with tumorsphere culture medium containing 2% FBS and resuspended in serum-free tumorsphere culture medium for cell counting.

Xenograft models.

All animal experimental procedures were reviewed and approved by the University of Saskatchewan Animal Research Ethics Board. Mice used in the present study were from our established colony of NOD SCID gamma mice at the Laboratory Animal Services Unit (LASU), University of Saskatchewan. Mice were maintained at the LASU during the course of the experiments. Control *sh*RFP and *sh*FUT9 knockdown HCT116 cells were trypsinized and resuspended in ice cold PBS. Cells were mixed 1:1 with Matrigel (Corning, CB-40234) and 3 x 10^6 cells in a total volume of 100 µL and injected subcutaneously into the left flank of 6 to 8 week old immunodeficient

NOD/SCID gamma mice. At least five mice that developed tumors were used in our analysis for each experimental condition in each biological replicate. One of the mice in the control group was excluded from the analysis of the last two time points due to lethality. Tumors were measured every 3 to 4 days using a digital caliper, and the tumor volume was calculated using the tumor ellipsoid formula $A/2*B^2$ where A and B represent the long and the short diameter of the tumor respectively. Upon experiment termination, tumors were extracted, fixed in 10% formalin, and weighed.

FACS analysis.

Cells were harvested and washed 3 times with ice-cold PBS containing 0.25% FBS. Cells were incubated with FITC-conjugated mouse-anti-human CD44 antibody (BD, 555478) or FITC-conjugated mouse IgG2b antibody (BD, 555742) for 30 min at 4ºC in the dark. Cells were then washed thrice with PBS, run through a Beckman Coulter CytoFLEX flow cytometer at 488 nm, and analyzed using CytExpert V1.2 software.

Discussion

We present a novel approach for identifying metabolic tumor suppressors that leads to the discovery of the complex, multi-faceted role of FUT9 in colon cancer. On the methodological side, we show here that a metabolic modeling MTA analysis can successfully identify metabolic genes that play a causal role in cancer initiation and progression from an initial list of genes that are

formed via a standard genome-wide analysis. Such an analysis may be thus performed to further identify causal metabolic cancer genes given any list of candidate cancer drivers emerging from a genomic analysis, in other cancer types.

The role of FUT9 in colorectal cancer appears to be rather complex. Our results indicate that FUT9 activity promotes the expansion of TICs, while its downregulation supports expansion and aggressiveness of bulk of tumor cells. TICs represent a higher proportion of the overall cell population in a tumor at earlier stages of tumor development. At later stages however, TICs are gradually outgrown by the rest of the tumor cells (Figure 4E), but they are still required for efficient tumor growth and maintenance[79–84]. Since our experimental data suggests that FUT9 provides an advantage for TIC populations, while its reduced activity benefits other tumor cells, its relative abundance should be expected to gradually drop with tumor progression, mirroring a decrease in the proportional representation of TICs. Notably, in accordance with that, we found that FUT9 expression is maintained in earlier tumors: colorectal polyps and colorectal adenoma at the levels observed in healthy colon tissue (studied in paired, matched samples; Appendix Figure S1), while FUT9 levels progressively decrease from the M0 to M1 stages (Appendix Figure S5). Reduced FUT9 expression at the M1 metastatic stage also matches our observations, suggesting that FUT9 downregulation enhances migration of colorectal cancer cells. This further supports a notion

that as tumors develop, FUT9 activity is switched off in the bulk of tumor cells to enhance their aggressiveness, which should negatively affect patient survival. In agreement, our computational analysis showed a positive correlation between FUT9 expression and survival of colorectal cancer patients.

This study is focused on the identification of tumor suppressor genes, as simulating a gene's knockdown in the metabolic model is very well defined, while simulating the over-expression of genes is more complex and challenging. Thus, developing an MTA approach to identify causal metabolic oncogenes whose overexpression is transforming the metabolic state remains an open challenge. Cancer evolution usually involves a sequence of genetic and environmental events; indeed, while our computational analysis points to the central role that FUT9 plays in generating a tumorigenic metabolic state in colon cancer, we find that its role depends on the overall genomic context, such as the cell types in which it occurs and the staging of the tumors. In agreement, our experimental data reveal that, while FUT9 activity enhances OCT4 expression, and is essential for the formation of tumor initiating cells, it also show that FUT9 downregulation enhances the invasive behavior of bulk colon cancer cells, which hence contributes at later stages following tumor initiation. Hence, our results should be viewed bearing this reservation in mind.

Overall, our findings support a dual role for FUT9 in colorectal cancer. They suggest that it may act in this malignancy in a manner similar to the reported actions of the EphB2 receptor, a known hallmark of colorectal cancer TICs[85] that is also downregulated to allow colorectal cancer tumor progression[86]. Our description of this complex action of FUT9 identifies an entirely new player in colorectal cancer and adds another intriguing member to the rather short list of metabolic genes that have been shown to play a critical role in tumor biology.

Chapter 2: Cancer pathway modifiers.

Published as "Data-driven metabolic pathway compositions enhance cancer survival prediction" PLOS Computational Biology, 2016 [87]

Introduction

Altered cellular metabolism is an important characteristic and driver of cancer. Surprisingly however, we find here that aggregating individual gene expression using canonical metabolic pathways fails to enhance the classification of noncancerous vs. cancerous tissues and the prediction of cancer patient survival. This supports the notion that metabolic alterations in cancer rewire cellular metabolism through unconventional pathways. Here we present MCF (Metabolic classifier and feature generator), which incorporates gene expression measurements into a human metabolic network to infer new cancer-mediated pathway compositions that enhance cancer vs. adjacent noncancerous tissue classification across five different cancer types. These data-driven pathways, in contrast to the canonical literature-based pathways, successfully generate clinically relevant features that are predictive of breast cancer patients' survival in an independent dataset.

Results

MCF pipeline

We first tested if the use of canonical pathways enhances the accuracy of cancer classification. We overlaid gene expression data derived from 3611 samples across ten datasets of five cancer types (including breast, lung, colon, prostate and head and neck squamous cell carcinoma) onto canonical metabolic pathways

defined in the RECON1 human metabolic model [10] and quantified the expression of every metabolic pathway based on the sum of the expression of all genes associated with this pathway (Methods, which in this case yields better performance than using the mean expression). We then trained SVM classifiers of cancer vs. adjacent noncancerous tissue samples using either the expression of individual metabolic genes (henceforth, MGE-SVMs) or human canonical metabolic pathways' expression (Methods). Testing the classifiers in five-fold cross validation we found that using the canonical pathway expression leads to inferior performance in these classification tasks compared to using the individual metabolic gene expression. These findings motivated us to identify pathways whose activity may better reflect the altered rewiring of metabolism in cancer and enhance cancer prediction.

To this end we developed a new data-driven algorithm, called the Metabolic classifier and feature generator (MCF): (1) We first define a differentially expressed reaction as a reaction whose ranked expression level within a sample is significantly different in noncancerous vs. cancerous samples (using a Wilcoxon rank-sum p-value with $\alpha = 0.05$, Methods). (2) The next step of MCF follows the concept of *reporter metabolites* [88] - it identifies metabolites that participate in differentially expressed reactions between the noncancerous and cancerous samples. (3-4) The key novelty of MCF is to use these reporter metabolites as centerpieces for building novel *composite pathways* leading from each reporter metabolite s to a group of target metabolites T_s that show consistent differential expression between the cancerous and noncancerous states. These pathways are

(by construction) predictive of the cancer vs. non-cancer states. (5) We then build a support vector machine (SVM-MCF) ensemble classifier of cancer vs. noncancerous tissue based on the gene expression of the new composite pathways as classification features. We apply a five-fold cross validation procedure to test the classification rate (accuracy) and area under the cover (AUC) for each dataset studied (Methods). The main steps of MCF are outlined below and in Figure 5 (see Methods for a formal description):

(1) Rank-transform the gene expression data: We first rank-transform the gene expression data and convert it biochemical reaction expression values using the human model's genes-to-reactions mapping. This results in patient specific weighted metabolic networks in which the weights of each reaction edge correspond to the rank assigned to this reaction for a certain patient.

(2) Identify seed reporter metabolites: For computational tractability, we limited the search to simple paths in which the first reaction is differentially expressed between the two states. To this end, we identify metabolites that are substrates in a large number of reactions that are differentially expressed between cancerous and noncancerous samples.

(3) Assigning 'expression weights' from each seed reporter metabolite on the paths to all other metabolites in the network: We calculate the heaviest distances (i.e. the weight of a simple path with the largest sum of reactions' expression values) from each seed metabolite to all other metabolites in the network. For the purpose of identifying the new composite paths, the metabolic network hypergraph is transformed to a regular graph

representation having metabolite nodes and (directed) edge connecting any two metabolites that participate in a given reaction as a substrate and a product, respectively (if the reaction is directed).

(4) Identify the most differentially expressed ('heaviest') pathways: For each source metabolite *s* we find the L=10 target metabolites T_s such that the heaviest distance from *s* leading to each of the targets in T_s differs most between the noncancerous and cancer training sets.

(5) Building an SVM classifier: For each of the N source metabolites *s* we train an independent SVM classifier to distinguish cancerous from noncancerous samples using the weight of the L selected paths from *s* to T_s as features. This results in an ensemble of N SVMs. A test sample is then classified by a majority vote over the N classifiers.

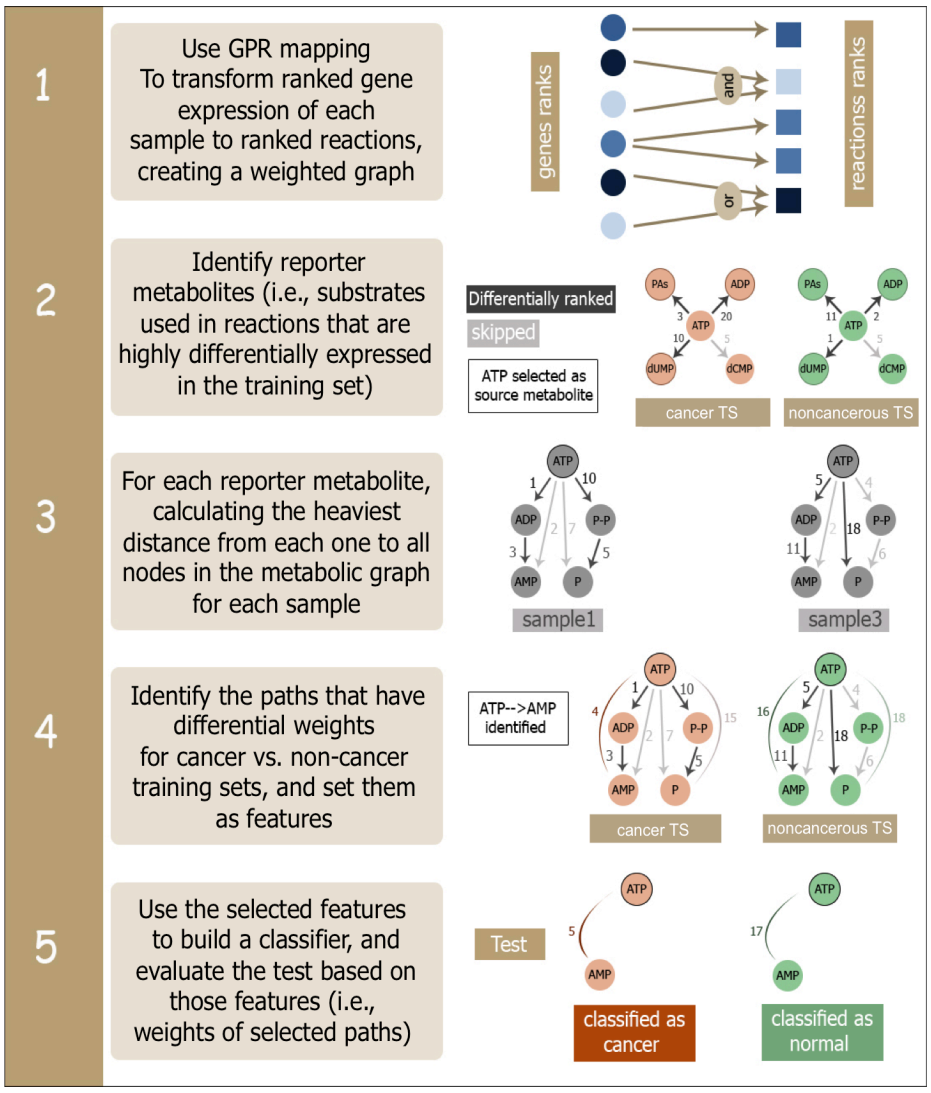

Figure 5. Overview of the MCF algorithm.

MCF predictive performance

We compared the accuracy of the MCF to MGE-SVMs classifiers that are based on individual metabolic gene expression by comparing their AUC and mean accuracy scores in a five-fold cross validation on various cancerous vs noncancerous classification tasks. We find that MCF performs as well as

MGE-SVM in all 10 datasets studied spanning five different cancer types, and significantly outperforms MGE-SVM in five of these datasets.

As MCF aggregates transcriptional information in network-based manner, we hypothesized that it will be more robust than MGE-SVM when trained on data of the same cancer type but aggregated from multiple studies. To test this we merged the available tumor/tissue samples expression (rank-transformed, Methods) data from both GEO and TCGA, producing a combined dataset for each of the five different cancer types studied. We compared the performance (AUC and accuracy) of MCF and MGE-SVM on each of the five combined datasets using a standard five-fold cross-validation procedure. Combining datasets in this manner accentuated the higher predictive performance of MCF vs. MGE-SVM across all cancer types studied (Figure 6), including colon cancer where no significant performance difference was observed previously.

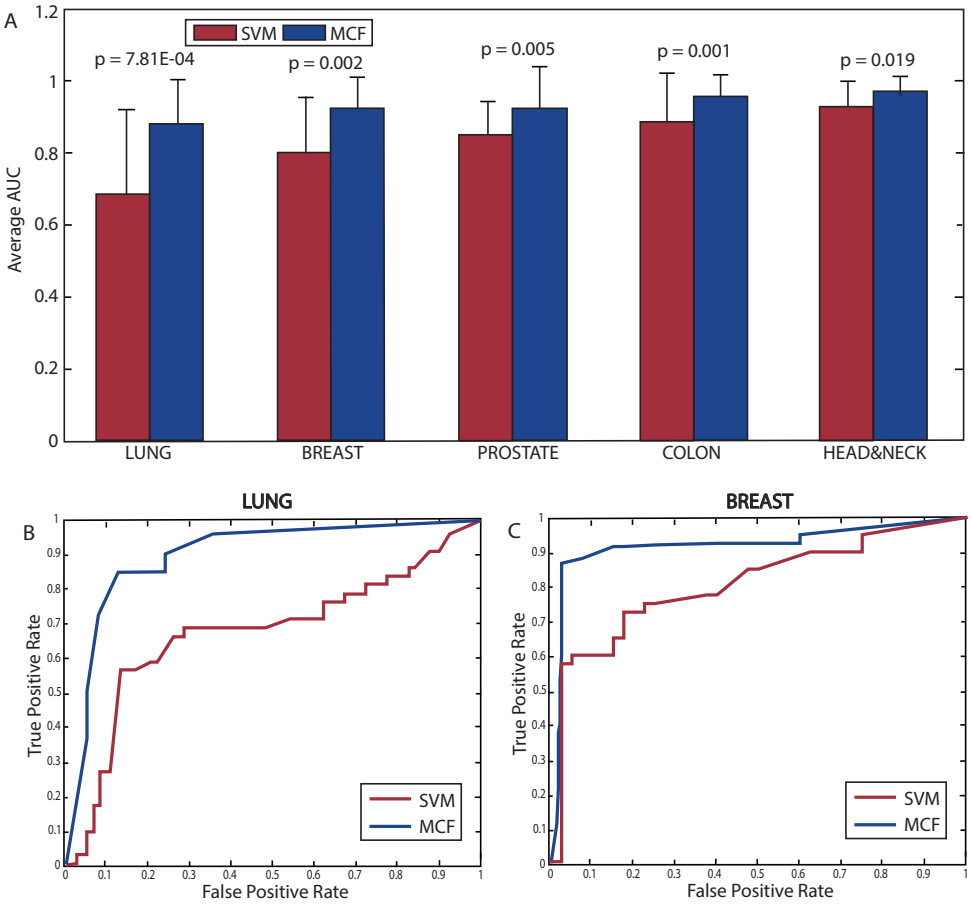

Figure 6. Comparing the performance of MCF to MGE-SVM across integrated cancer-type datasets. (A) A bar plot describing the predicted AUC obtained over the combined datasets of the same cancer type using a five-fold cross validation procedure for MGE-SVM (red bars) and MCF (blue bars) classifiers. AUC denotes the area under the curve. Error bars represent one standard deviation, and p-values are for a one-sided, paired-sample t-test for the AUC of each of the five folds. (B), (C) present the receiver operating characteristic (ROC) curves obtained in the classification of the lung and breast cancer combined datasets, respectively.

Notably, source metabolites that strongly differ in usage between noncancerous and cancerous tissues may constitute interesting cancer biomarker candidates. We find that there is a small set of such source metabolites that recur in multiple cancer types, and they vanish in randomly shuffled data. These include currency energy metabolites (e.g., NAD+ and ATP), a finding consistent with the large alterations seen in energy metabolism in cancer. We examined the target metabolites T_s that contribute most to ATP being differentially utilized. As the paths leading to them from ATP are most differentially expressed, this may testify that the consumption of ATP to produce each of these metabolites is altered in cancer (and may possibly serve as correlate to their overall production levels). These target metabolites are specific for cancer type (Table 1, a pattern that remained robust to the introduction of noise to the data (See Methods). This suggests that while ATP is differentially utilized between tumors and their noncancerous tissues counterparts in all cancer types, there exists considerable variance in the ways it is utilized.

prostate	Breast	Colon	head&neck	lung
↑ 3alpha,7alpha,12alpha -Trihydroxy-5beta -cholestanoyl-CoA(S)	↑ dADP	↓ O-Acetylcarnitine	↑ CTP	↑ Hydroxy-methylglutaryl -CoA

↑ 3alpha,7alpha-Dihydroxy-5beta-cholest-24-enoyl-CoA	↑ Oxidized thioredoxin	↑ 5-Phospho-beta-D-ribosylamine	↑ dATP	↑ Spermine
↓ 3alpha,7alpha,26-Trihydroxy-5beta-cholestane	↓ Hydrogen peroxide	↑ Spermine	↑ dCTP	↑ D-Mannose 1-phosphate
↓ 3alpha,7alpha,12alpha-Trihydroxy-5beta-cholestan-26-al	↓ L-Threonate	↑ Fumarate	↑ dGTP	↑ Deoxycytidine
↓ 7alpha-Dihydroxy-5beta-cholestan-26-al	↓ Hydrogen peroxide	↑ GMP	↑ dITP	↑ Diphosphate
↓ 3alpha,7alpha,12alpha,26-Tetrahydroxy-5beta-cholestane	↓ Iodine	↓ retinoyl glucuronide	↑ dTTP	↑ UDP-D-glucuronate

↑ 5-Amino-1 -(5-Phospho-D-ribosyl) imidazole-4-carboxamide		↓ UDP		↑ Phospho enolpyruvate
		↑ Leukotriene B4		↓ Oxalate

Table 2. Target metabolites selected for MCF. The target T_s metabolites that MCF selected when it choses ATP as a seed (↑ denotes increased formation from ATP in cancer and ↓ denotes decreased formation from ATP in cancer compared to noncancerous tissue counterpart, Methods). The table shows one instance of each selected target although in some cases the same target metabolite was identified in multiple compartments (e.g. UDP in the cytosol and in the mitochondria).

Several of the target metabolites are known to be associated with their respective cancers: Oxalate has been studied as a survival marker in lung cancer [89]; spermine has been observed to be differentially expressed in lung and colon cancer [90–92]; Carnitine was shown to slow down tumor development in colon cancer [93]; and blockage of Leukotriene B4 was reported to suppress cell proliferation in colon cancer patients [94]. Thus, MCF identifies key

metabolites that take part in metabolic processes that are altered in the specific cancers they occur.

MCF prediction of patients' survival

As we have shown, MCF generates new composite pathways that show more power than traditional pathways in classifying normal versus cancer samples. To evaluate the clinical significance of these new features we examined whether they are also predictive of a different objective, the prediction of survival of breast cancer patients. Furthermore, to test whether the clinical utility of MCF pathways carried between datasets, we trained and tested the pathways on *independent* datasets. For training we used the combined GEO and TCGA breast cancer data. For testing, we used an independent dataset (METABRIC, [95]) that includes gene expression measurements from 1,981 cancer patients and their corresponding survival information. Remarkably, we find that out of the 80 pathways that MCF identified as differentially expressed in the original classification task on the combined TCGA and GEO data (L=10 targets from 8 identified source metabolites), 58 pathways are predictive for survival in the METABRIC data using Kaplan-Meier estimator [96] (FDR corrected Kaplan-Meier log-rank p-value < 0.05; methods). In marked contrast, the expression levels of *none* of the canonical metabolic pathways defined by Recon1 are predictive of survival in this dataset. This is in line with our previous observation that the activity of the canonical

metabolic pathways is not helpful in distinguishing between cancerous vs. noncancerous samples.

To evaluate the aggregate predictive power of the set of pathways selected by MCF as a whole, we compared patients predicted by MCF to have the best and worst prognosis (top and bottom 10%, respectively; Methods) and found that they indeed have a marked difference in their survival as predicted (Figure 7A, delta-AUC = 0.2436, and Kaplan-Meier log-rank P-value < 1.0e-30). In contrast, when aggregating information across the canonical human metabolic model pathways in a similar manner we find that pathways predicted to have best and worst prognosis show no difference in survival (Figure 7B, delta-AUC = 0.0176, and Kaplan-Meier log-rank P-value = 0.4282). We then examined whether the aggregated pathway score can be used as a survival model for the METABRIC dataset, using the conventional concordance index (C-index) [97]. We find that while the pathways selected by MCF are predictive of patients survival, the canonical human metabolic model pathways do not show such predictive power (C-index = 0.69 vs. 0.52, respectively). Interestingly we find that the predictive power of individual MCF selected pathways in the original task of predicting cancer vs. noncancerous samples (i.e. the AUC obtained from the cross validation procedure on the combined datasets from TCGA and GEO) markedly correlates with their predictive power for survival in the METABRIC dataset (Spearman ρ = 0.58, p-value <1.4e-09). This finding explains their predictive

power across these different tasks and datasets, and further testifies to their clinical significance.

Finally, we performed a canonical pathway enrichment analysis over the reactions participating in the MCF composite pathways identified in breast cancer that are predictive of survival. We find that the most enriched canonical pathways emerging in this analysis are already known to be associated with cancer initiation and progression, such as fatty acid related metabolic pathways [98–100], the citric acid cycle [101,102] and cholesterol and steroid metabolism [103] (Figure 7D). Hence, even though aggregated gene expression through canonical pathways does not show survival predictive power, the composite alterations in cancer do rewire its metabolism using components of these traditional pathways, albeit via different composition.

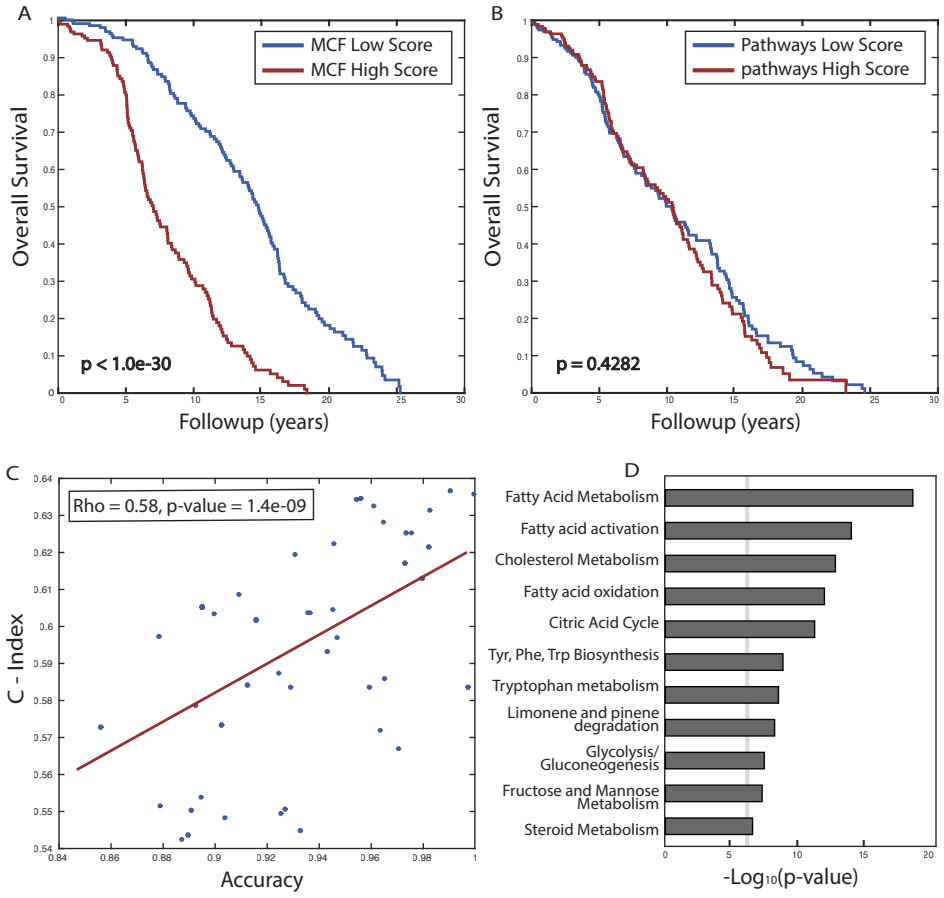

Figure 7. MCF survival prediction. MCF pathway utilization predicts the survival of breast cancer patients, while canonical pathways show no such signal. Shown in (A) and (B) are the Kaplan-Meier survival curves for patients predicted by MCF and canonical pathways respectively to have the best and worst prognosis (top and bottom 10% of patients scores, respectively; Methods). (C) A scatter plot showing the correlation between the prediction classification accuracy achieved using each individual MCF pathway in the combined breast cancer data from TCGA and GEO (where they are identified) (X-label) and the C-index obtained using each such pathway in predicting

patients' survival on the (unseen) METABRIC data. (D) The canonical pathway enrichment of the reactions participating in the MCF composite pathways predictive of survival. The dashed line represents a significance threshold of 0.05 (corrected for multiple hypotheses testing).

Methods

Gene expression datasets

We focused on five cancer types, and for each one utilized datasets from TCGA [45] and GEO [104], as summarized in Table 3.

	TCGA data		GEO data	
Cancer type	TCGA designation	sample count (N/C)	GEO accession	sample count (N/C)
Prostate	PRAD	487/52	GSE32448 [105]	40/40
Lung adeno-carcinoma	LUAD	58/490	GSE19804 [91]	60/60
Colon	COAD	41/273	GSE32323 [60]	17/17
Head&neck	HNSC	43/498	GSE6631 [106]	22/22
Breast	BRCA	111/1098	GSE10780 [107]	140/42

Table3. Summary of the datasets utilized for five cancer types. N and C stand for number of normal and cancerous samples in the data, respectively.

In addition, we used the METABRIC breast cancer database by Curtis et al. [95] to test the predictive power of MCF pathways with respect to patient survival.

Evaluation of classifiers

Throughout this study, we evaluate classifier performance by computing the AUC and average accuracy in a five-fold cross-validation procedure. We repeated 100 times the following:

- Down-sample either the cancerous or normal groups: Assume that the data has N normal samples and C cancerous samples and |N|>|C|. We randomly chose |C| samples out of the normal group and excluded the rest. Similarly, if the data had more cancerous samples than normal ones, we down-sampled the cancerous group to the size of the normal group. This ensures that the accuracy statistic is not biased due to an over-representation of one of the groups, which occurs in many of the datasets studied here.
- 5-fold cross validation: We split the chosen samples into 5 folds, each time training on 4/5 of them and testing by computing the AUC or accuracy on the remaining 1/5.

The AUC and accuracy shown here is the average of the 100 repetitions, and the paired t-test p-values are from the resulting vector of 100 AUC or accuracy values for each such random selection.

Metabolic gene expression SVMs (MGE-SVMs)

To classify cancer vs. normal samples according to metabolic gene expression, we trained a support vector machine (SVM) using the expression of 1,496 metabolic genes as features. We denote these machines MGE-SVMs. Metabolic genes are defined in this study as the set of 1,496 genes annotated in Recon1 [10] a well-curated reconstruction of the global human metabolic network.

We observed that SVMs trained on this reduced set of gene expression features consistently outperformed SVMs trained on the expression of all genes. This is not surprising seeing that the metabolic subset has roughly one-order of magnitude smaller dimensionality, and yet remains highly informative because of the key role of metabolic adaptations in cancer [108–110]. Applying further dimensionality reduction on the set of 1,496 metabolic genes (e.g., through PCA) had little effect on the results. In addition, we observed that training MGE-SVMs with ranked expression values (that we use for MCF) achieves similar, but slightly inferior, results to the ones obtained using the expression values themselves.

Converting gene expression into biochemical reaction expression

Recon1 defines a mechanistic genotype-phenotype relationship through Boolean rules that encode gene-protein-reaction (GPR) associations. To convert ranked gene expression to biochemical reaction expression, we evaluated the Boolean GPR rule of that reaction while replacing the "AND" and "OR" operators with "min" and "max", respectively as described in [111]. Differential expression between biochemical reaction is determined by a Wilcoxon rank sum test with a significance threshold of 0.05, Bonferroni-adjusted for multiple hypotheses where appropriate.

Computing metabolic pathway expression

Classification based on metabolic pathways relied on the pathway definitions embedded in Recon1, which associates every reaction with a single pathway out of a total of 99 pathways defined based on the Kyoto Encyclopedia of Genes and Genomes (KEGG) LIGAND database. To compute a pathway expression, we first converted the ranked gene expression to ranked reaction expression as described above, and then summed the ranked expression of all the reactions associated with the pathway. An alternative methods of computing pathway in which for each pathway we use the sum the ranks of all its associated genes showed inferior performance comparing to the method presented here, as well as using the mean of ranked reaction expression instead of the sum.

Identifying seed reporter metabolites

MCF builds metabolic pathways that have highly differential expression between the two target states (i.e., cancerous and non-cancerous). However, identifying the most differentially expressed pathways between two groups of weighted networks is a NP-hard problem by reduction from the problem of finding the longest-path [112] (Given a directed weighted graph G, let w be the smallest weight in G. Create a copy G' of G with all edge weights set to w-c for some constant c>0. The most differentiating path between G and G' is the heaviest (i.e., longest) path in G). For computational tractability, we limited the search for simple paths in which the first reaction is differentially expressed between the two states. We chose source metabolites that are substrates in at least k>=5 differentially expressed reactions with Wilcoxon rank-sum p-value corrected for multiple hypothesis.

Building the classifier

To build a classifier based on the differential expression of the pathway from source metabolite *s* to L=10 target metabolites, we do the following: we compute the heaviest distances (i.e. the weight of a simple path with largest sum of reactions expression values) from *s* to the all other metabolites in the network in all of the train samples. For the purpose of computing paths, we followed the common approach [113,114] transforming the hypergraph into a digraph and limiting ourselves to pathways that are simple directed paths in the digraph. The metabolic hypergraph is viewed as a standard graph with

metabolite nodes and a directed edge (u,v) connecting any two metabolites such that u and v participate in some reaction as a substrate and a product, respectively. We then select a set T_s of L target metabolites for which the paths from *s* were most differentially expressed. I.e., for every target metabolite *t* we compute the Wilcoxon rank sum p-value when comparing the heaviest distance from *s* to *t* in the normal vs. the cancer samples, and we finally choose the T_s with L metabolites that obtained the smallest p-values out of all possible targets. The distances from *s* to the chosen L metabolites (denoted T_s) are used as features for an SVM.

Let N be the number of source metabolites detected. MCF repeats the procedure described above for each of the source metabolites *s*, and for each *s* a distinct SVM is trained. This results in an ensemble of N SVMs. A test sample is then classified by the majority vote of the N individual classifiers (no ties ever occurred in the present study).

MCF classification score

The MCF classifier is an ensemble of N SVMs (for each detected source metabolite). The MCF classification score for classifying observation x is the sum of N scores assigned to x by the N SVMs. Therefore:

$$MCF_{score(x)} = \sum_{i=1}^{N} f_i(x)$$

Where $f_i(x)$ is the predicted response of x for the trained classification function f_i (trained on the features selected for source metabolite i)

$$f_i(x) = \sum_{j=1}^{n} \alpha_{i,j}\, y_{i,j} G(X_i, X) + b_i$$

Where $(\alpha_{i,1\ldots}\alpha_{i,n}, b_i)$ are the estimated parameters, $G(X_i, X)$ is the dot product in the predictor space between X and the support vectors and the sum indicates training set observations.

Predicting patient survival by canonical or MCF pathways

To train the model and select the features we use the combined GEO and TCGA breast cancer datasets and train it on the original classification task of separating noncancerous from cancer tissues (when all samples are used). This results in 80 composite pathways that are generated and selected by MCF (for comparison, the human metabolic network defines 99 different pathways). We then use the METABRIC dataset and calculate the weights of the 80 selected pathways for this dataset (by generating a weighted metabolic graph for each sample in the MTABRIC dataset and calculating the heaviest distance between each seed metabolite and the target metabolites selected for it for the combined dataset from GEO and TCGA) as well as the weight of the 99 human metabolic network pathways. In the two pathways sets, we define the weight of each patient for every pathway by the sum of ranks of the reactions associated with the pathway. For every pathway we evaluated the KM log-rank p-value taking top 10% and bottom 10% weighted pathways.

To calculate an aggregated pathway score using either the 80 MCF selected pathways or the 99 canonical model pathways we calculate the weights of these pathways using the METABRIC gene expression data. We compute for each patient's tumor two aggregate scores (one over the MCF pathways and over the model pathways) as follows:

$$score(patient_i) = \frac{\sum_{p \in P_c} weight_i\,(p)}{\sum_{p \in P_n} weight_i\,(p)}$$

When $weight_i(p)$ is the weight of pathway p for patient i. P_c is the set of pathways (either MCF selected pathways of canonical pathways) in which high expression levels were associated with cancer state, and P_n is the set of pathways in which high expression levels were associated with noncancerous healthy state. Both P_c and P_n are determined by analyzing the two breast cancer datasets from TCGA and GEO (the mean of each pathway was evaluated for noncancerous and cancer samples to decide whether a pathway is in P_c or in P_n). These P_c and P_n set of pathways were then used to predict the patients survival an independent METABRIC breast cancer dataset, by assessing $weight_i(p)$ for every sample based on its transcriptomics and computing $score(patient_i)$ accordingly. A KM analysis is then employed to examine the survival difference of high score versus low score patients' samples.

MCF Robustness to gene expression noise

To test MCF's robustness, we introduced noise into every sample's gene expression vector by adding random Gaussian noise with distributions N(0,1), N(0,2) and N(0,3). We then trained MCF classifiers based on the perturbed data and evaluated the source and target metabolites MCF selected.

Discussion

We present a novel method termed MCF that identifies data-driven pathway compositions that best differentiate the metabolic alterations occurring in cancerous vs. noncancerous tissues. MCF leverages a priori knowledge on the structure of the human metabolic network (ignoring its conventional decomposition to canonical pathways) to inform the analysis of cancer vs. noncancerous gene expression. It detects key hubs of metabolic alterations and infers the composition of non-standard pathways altered in a specific cancer type. Applied across five different cancer types, we find that MCF outperforms standard methods in the basic task of cancer vs. noncancerous classification. Remarkably, MCF derived pathways successfully predict patients' survival in an independent dataset while standard metabolic pathways fail to do so, testifying on the robustness and utility of the metabolic features learned by MCF.

Meta-learning is of great relevance to cancer classification as it can potentially exploit one of the hallmarks of cancer, deregulation of pathways and cellular

processes, by taking knowledge on relations between genes and pathways into account in the classifier [24,115,116]. However, recent studies have reported that many of these methods do not outperform a model trained over single gene features [25–27,117]. MCF offers a solution to some of the main issues that hampered previous methods. First, some previous studies are based on pre-defined gene sets [118] or networks [119] characterizing healthy cells while cancer may rewire many functions, and in particular its metabolism. To this end, MCF performs unsupervised pathway generation and selection that captures key metabolic alterations occurring in cancer. Second, some studies relied on the topology of a pre-defined biological network such as a co-expression network [119], cellular pathway map [120] or protein–protein interaction (PPI) network [121] that have been inferred from high-throughput studies. In difference, MCF relies on a manually curated metabolic network that is extensively supported by experimental evidence [10]. The metabolic network is thus less noisy, while still highly informative due to metabolism's role in cancer growth and development. Third, it has been shown that structural and directional information improves the predictive power of meta-features over single genes [117]; In accord, the metabolic network is directional and highly structured which allows MCF to infer pathways of biological relevance.

While metabolic reprogramming is a substantial part of cancer biology, the methodological insights obtained from developing MCF are general, and could potentially be built into path-centric approaches that would involve

other cellular networks. This could lead to stronger predictors based on reliable models of signaling and regulatory networks on a genome scale. Second, finding the most separating paths in differently weighted graphs is an NP-complete problem. Here, we only offer a heuristic solution that is obviously sub-optimal. This could be improved upon by employing more exhaustive and/or efficient weighted path searching methods. We can expect that follow-up work will advance the identification of top separating pathways in differentially weighted metabolic graphs, potentially improving the power of MCF further.

Chapter 3: Cancer immunotherapy treatment modifiers

Accepted to *Nature Medicine* as :"Robust prediction of therapeutic response to immune checkpoint blockage therapy in metastatic melanoma" [122]

Introduction

Immune checkpoint blockade (ICB) therapy provides remarkable clinical gains, where melanoma is at the forefront of its success. However, only a subset of patients with advanced tumors currently benefit from these therapies, while incurring considerable side-effects and costs. Hence, constructing predictors of patient's response is of crucial value, and such accurate predictors are yet absent. This is a serious challenge due to the complexity of the immune response and the lack of large ICB-treated patient cohorts with omics and response data, which handicaps the construction of robust predictors that are transferable across different datasets. Here we build an immune-centered predictor of ICB-response that utilizes immune checkpoints' transcriptomic relations mediating spontaneous tumor regression. It robustly predicts melanoma patients ICB-response and can capture almost all true responders while sparing treatment for more than half of the non-responders. It achieves an overall accuracy of 0.83 over 11 datasets spanning 297 samples including unpublished data, outperforming existing predictors.

Results

NB Spontaneous regression and ICB response in melanoma

We hypothesized that an immune-based predictor of spontaneous regression may capture the immune activity and could thus be used more generally to predict response to ICB for patients with melanoma. To test this hypothesis, we built a predictor of spontaneous regression in NB, analyzing the

transcriptomics data of 108 patients, who include both spontaneously regressing (patients considered as low risk NB and with no tumor progression) and high risk progressing patients (i.e., without spontaneous regression, Methods)[123]. As we are interested in predicting the response to ICB, we focused on 28 immune checkpoint genes collected from literature reports that were included in all RNA-sequencing (RNA-seq) datasets available to us. To capture the predictive relations based on these immune checkpoint genes, we based the NB predictor on pairwise relations between the (normalized) expression levels of these genes. Each predictive feature compares the expression of two checkpoint genes A and B, capturing a logical relation between their transcriptional levels (e.g., A > B). We performed a feature selection procedure searching for a subset of these features that best separates patients with the spontaneously regressing NB from patients with non-regressing NB, resulting in 15 most predictive features (Methods). Based on these features, the prediction of spontaneous regression of a tumor sample from its expression data is simply made by counting the number of predictive feature pairs that are fulfilled (true) in that sample. This number, ranging from 0-15, denotes its IMmuno-PREdictive Score (IMPRES), with higher scores predicting spontaneous regression. The resulting predictor obtains an accuracy of 0.9 (in terms of the Area Under the Receiver Operator Curve (AUC)) in the NB dataset. Reassuringly, examining tumors derived from patients with melanoma who were not treated with ICB[124], we find that the IMPRES scores for patients denoted as 'high immune response' are significantly higher than

that of other subtypes (Rank-sum p-value = 9.6E-5 and 0.05 for the test and validation datasets, Figure 8A). Additionally, we find that IMPRES is significantly and positively associated with higher overall survival in these datasets[124] (Figure 8B), testifying to its ability to capture immune activity that is associated with improved melanoma prognosis in the absence of ICB treatment.

We next turned to investigate whether there are similarities between the cellular processes mediating the immune response in melanoma and those mediating spontaneous tumor regression in NB. To this end, we collected 9 gene expression and ICB response datasets including patients treated with anti-CTLA-4, anti-PD-1 or their combination[125–130]. First, we identified immune related, Consistently Differentially expressed Pathways (termed CDPs) in ICB responders versus non-responders (evaluated separately for patients treated with anti-PD-1 or anti-CTLA-4 treatments, Methods). We find seven CDPs across all anti-PD-1 datasets and four CDPs across all anti-CTLA-4 datasets (an overlap which is significantly higher than expected, permutation P-value=0.001 for anti-PD1 and P-value=0.03 for anti-CTLA-4 datasets respectively). Second, we find that the CDPs are also differentially expressed in a similar manner in the 'high immune response' melanomas compared with other subtypes (Binomial P=0.003 and 0.0623 for the test and validation sets)[124] and in spontaneously regressing vs high risk progressing NB tumors (Binomial P=0.009, Figure 8C) (Methods). To test the relation

between the features identified in NB and the activity of the CDPs, we computed the correlations between the expression ratios of each of these features with each CDP expression. As evident from Figure 8D, these associations are consistently maintained across the four groups studied (All pairwise comparisons hypergeometric P-values < 0.01).

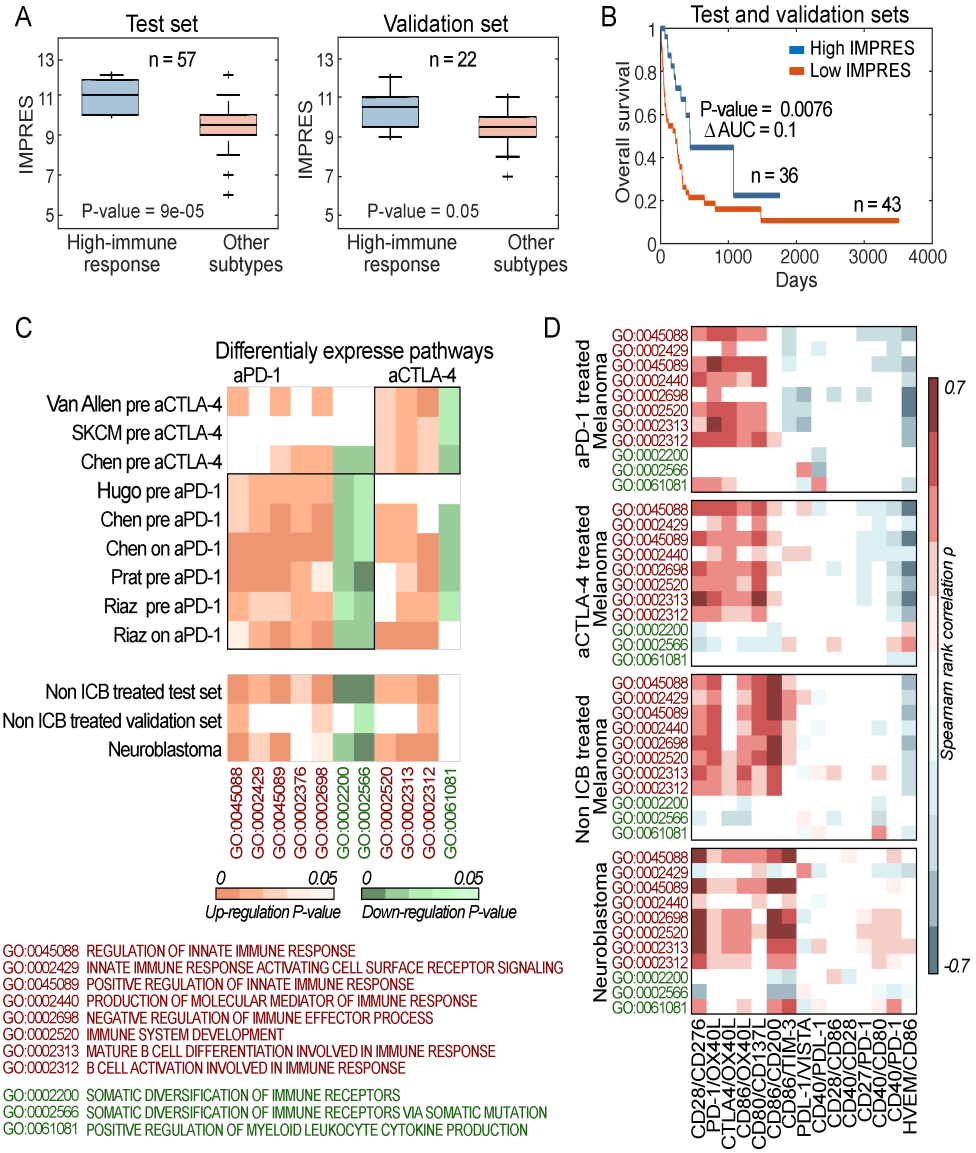

Figure 8. NB regression association with melanoma immune response. (A) Boxplots showing IMPRES of high vs low immune response in test and validation datasets of non-ICB treated melanoma patients[124]; p-values are computed via a Rank-sum test. **(B)** Kaplan-Meier survival curves of patients with high versus low IMPRES (computed over the combined test and validation datasets[124]). The median IMPRES is used to define the "Low IMPRES" and "High IMPRES" subgroups. **(C)** Upper Panel**:** Heatmaps showing the enrichment P-values for CDPs that are up (orange) or down (green) regulated in responders versus non-responders across the anti-PD-1 (encapsulated in the left rectangle) and the anti-CTLA-4 melanoma datasets[125,127,128,130] (right rectangle). The lower Panel displays the enrichment P-values for these CDPs in high immune response vs other subtypes in non-ICB treated melanoma, and in spontaneous regression vs non-spontaneous regression in the NB dataset. **(D)** Heatmaps showing the rank correlation ρ between expression levels of each CDP and each of the IMPRES features ratios, computed separately over the anti-PD-1 datasets, the anti-CTLA-4 datasets, the non-ICB treated melanoma datasets and the neuroblastoma dataset. White-colored entries denote non-statistically significant associations.

IMPRES predictor

We turned to apply IMPRES to predict the responses of melanoma patients to ICB treatments, without any further training. To this end, we analyzed 256 samples from 6 studies including patients treated with anti-CTLA-4, anti-PD-

1 or their combination[125–130]. We first computed the IMPRES of each melanoma sample based on its expression data and used that for the respective Receiver Operator Characteristic (ROC) classification curves. IMPRES achieves an AUC= 0.77 for van Allen et al.[125] (anti-CTLA-4); AUC = 0.83 for Hugo et al.[127] (anti-PD-1); AUC = 0.8 for TCGA SKCM[128] (anti-CTLA-4); AUC = 0.96, 0.77 and 0.80 for Chen et al[126] (on-treatment with anti-PD-1; pre-treatment with anti-PD-1 (post-CTLA-4 treatment) and pre-treatment with anti-CTLA-4, respectively) and AUC = 0.78 and 0.85 for Riaz et al. (pre- and on-treatment with anti-PD-1, respectively)[130]. A lower performance of AUC = 0.73 is obtained for Prat et al[129] (anti-PD-1), a nanostring dataset with low coverage of the IMPRES checkpoint molecules[126,129] (Figure 9A).

We further tested the predictive ability of IMPRES in a new unpublished dataset in which we carried out RNA-seq of tumor biopsies derived from 41 samples of patients with metastatic melanoma who were treated with different ICB therapies at the Massachusetts General Hospital (Methods). IMPRES achieves an AUC of 0.81 and 0.97 on the anti-PD-1 and anti-CTLA-4 samples respectively (Figure 9B). Evaluating the predictive accuracy of IMPRES on the aggregate collection of all the datasets studied above (a total of 297 samples), IMPRES obtains an AUC of 0.83, significantly superior to all other existing published predictive signatures, as shown in detail further below. Its aggregate performance is AUC=0.84 on all anti-PD-1 treated samples and AUC=0.8 for all samples treated with anti-CTLA-4 (Figure 9B). To appreciate

the potential translational impact of IMPRES, Figure 9C shows the number of true/false positives (responders) and true/false negatives (non-responders) obtained on this aggregated data at different IMPRES score thresholds. In total, there are 89 samples labeled as responders (considered 'positive' in the classification) and 208 non-responder samples (considered 'negative') across all datasets. If one adopts a very conservative approach and predicts responders only if their IMPRES score is greater-than/equal-to 12, few such predictions arise but all of them are true (top row pair). At a more relaxed threshold of 8, IMPRES correctly captures almost all true responders, while misclassifying less than half of the non-responders (that is, sparing unnecessary treatments for the majority of non-responding patients). When further decreasing the classification decision threshold more samples are predicted as responders, manifesting the known tradeoff between precision and recall (Figure 9D). A qualitatively similar picture emerges when considering the anti-PD-1 and anti-CTLA-4 treated patients separately. Higher IMPRES scores are also associated with improved overall survival and progression-free survival (PFS) in ICB treated melanoma patients (Methods, Figure 9E-H).

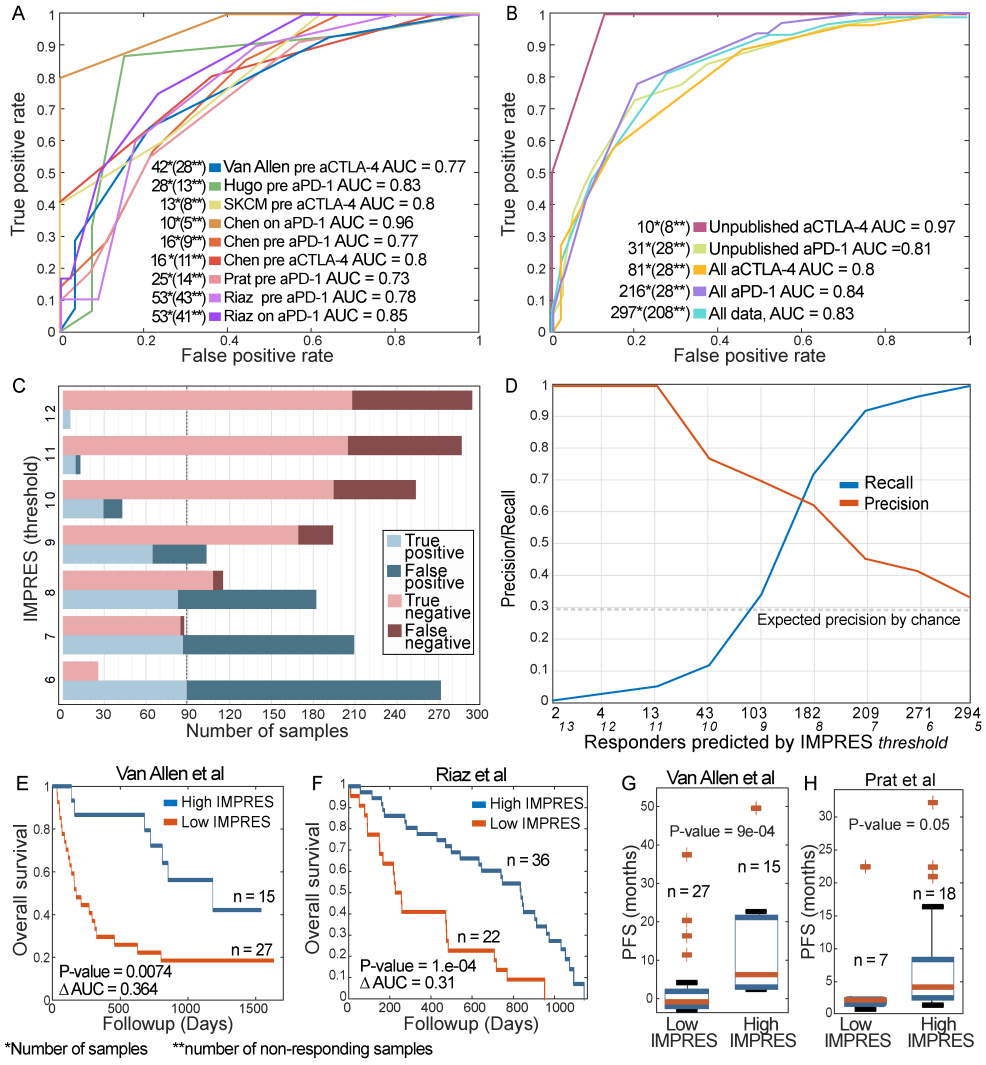

Figure 9. IMPRES performance. (A) Receiver Operating Characteristic (ROC) curves quantifying IMPRES prediction AUC across numerous publically available ICB response datasets[125–130]. (B) ROC curves for an independent dataset of ICB response (with 10 patients treated with anti-CTLA-4 and 31 patients treated with anti-PD-1) and for the aggregate datasets including all 297 samples, the 216 samples of patients treated with anti-PD-1 and 81 with anti-CTLA-4. (C) Bar plots showing the prediction accuracy and error types for different IMPRES thresholds (where a positive label corresponds to a

'responder' prediction) on the aggregate compendium of 297 patients included in all 11 datasets studied. The dashed line represents the total number of responders. (D) Precision/recall evaluation of IMPRES on the same aggregate compendium. The Y-axis displays the precision/recall as a function of the number of 'responder' predictions made (shown on the X-axis, obtained by decreasing the classification threshold, whose value is also displayed in italic font). (E)-(F) Kaplan Meier survival curves for the Van Allen et al and Riaz et al. ICB treatment datasets, respectively, with high vs. low IMPRES scores (using the median IMPRES as a threshold differentiating between the high and low groups). (G)-(H) Boxplots showing progression free survival for low vs. high IMPRES in the Van Allen and Prat et al. ICB datasets (using the median IMPRES as a differentiating threshold).

We next turned to compare the predictive accuracy of IMPRES with that of current state-of-the-art predictors. Even though there is a clear association between the tumor's mutational load and patient's response to ICB, the resulting predictive power is fairly moderate, with AUCs in the range of 0.6-0.7[127,131–133], and similarly for predictors based on the neoantigen landscape[125,131,134,135]. Studies based on transcriptomic signatures have reported AUCs in the range of 0.6-0.8[127,136], but these performance levels are mainly limited to the single dataset that was used for their construction. To perform the comparison, we built predictors of response to ICB based on each of the published transcriptomic signatures. The overall performance of

IMPRES is significantly superior to each of the other predictors (Paired Rank-sum test P-value<0.004) (Figure 10A). This observation holds true when we compared the performances for each ICB-treatment group separately (Figure 10B). Compared to IMPRES, the second best predictor, cytolytic-activity estimation[137], has an overall AUC of 0.68, and that of each of the other methods 0.6 or lower. Overall, the predictors built on biologically motivated scores (cytolytic-activity[137] and PDL-1 expression) generalize better than the machine learning based predictors constructed on transcriptomic signatures identified in isolated, specific cohorts. Of note, while we find a significant correlation between IMPRES and abundances of $CD8^+$ and $CD4^+$ T cells inferred via CIBERSORT, the inferred abundances of immune cells themselves are poor predictors of response to ICB. IMPRES superiority is particularly notable because for most existing signature-based predictors (all but cytolytic-activity[137] and PDL-1 expression) we had to re-train the latter separately for each dataset, otherwise their overall performance was dismal, testifying to their poor generalizability between different datasets (Methods). In contrast, IMPRES is constructed only once from the NB data and never trained on any melanoma dataset; thus, it is markedly less prone to over-fitting, a paramount concern regarding standard cancer transcriptomics predictors[138–140]. To further study the importance of training on the independent NB data, we trained ICB response predictors based on melanoma data instead of NB, following exactly the same representation and training procedure as used in IMPRES. In this case we obtain markedly lower

prediction performances on the melanoma datasets that were not used for training compared with those achieved with the original IMPRES procedure. For completeness, we additionally compared the performance of IMPRES to all other predictors but excluded patients annotated with 'stable disease' from the analysis. This results in an overall similar picture of superior performance of IMPRES versus the other classifiers with a slightly improved performance in both.

The features composing IMPRES uncover a few insights that are biologically interesting. Reassuringly, the relatively higher expression of known immune stimulatory genes (such as HVEM, CD27 and CD40) is associated with a better response, while the higher expression of known immune inhibitory genes (such as CD276, TIM-3, CD200 and VISTA) is associated with a worse response to ICB, as expected (Figure 10C). Higher expression of CD40 compared to that of PD-1, PDL-1, CD80 and CD28 is associated with a better response to ICB, in line with the recent findings that agonists of CD40 reverse resistance to anti-PD-1 therapy, and that induced PD-1 expression mediates acquired resistance to antagonist CD40 treatment[141]. Additionally, the higher expression of the immune stimulator CD27 compared to that of PD1 (but not compared to CTLA-4) is associated with improved response. This is in line with recent findings that the combination of a CD27 agonist plus anti-PD-1 recapitulates the effects of CD4+ T helper cells on tumor control, while the

combination of a CD27 agonist plus anti-CTLA-4 did not improve tumor control[142].

We further studied the individual predictive power of each of the 15 IMPRES features, by considering the expression ratio of each predictive pair (Methods). We find that some features are specifically more predictive for anti-PD-1 pre-treatment (CD28/CD86, Rank-sum P-value = 0.05) or on-treatment (PD1/OX40L, CD86/OX40L and CD86/CD200, Rank-sum P-value = 0.018 for all). Notably, no feature emerges as being strongly predictive of response to anti-CTLA-4 specifically (Figure 10D). Next, we examined all possible associations between these 15 features (using their expression ratio) and the abundance of all 22 types of immune cells inferred by CIBERSORT in the datasets of melanoma treated with ICB. Notably, we find two significant associations (Bonferroni-Corrected for multiple comparisons, $\alpha = 0.05$) between $CD8^+$ T cells abundance and IMPRES features that hold across melanoma datasets: the first involves a significant and consistent negative correlation with the CD40/PD-1 expression ratio and the second involves a positive correlation with the PD1/OX40L expression ratio (Figure 10E). Finally, a feature reduction analysis shows that the overall predictive performance of IMPRES can be maintained with a subset of 11 of the 15 original features, but beyond that it markedly decreases.

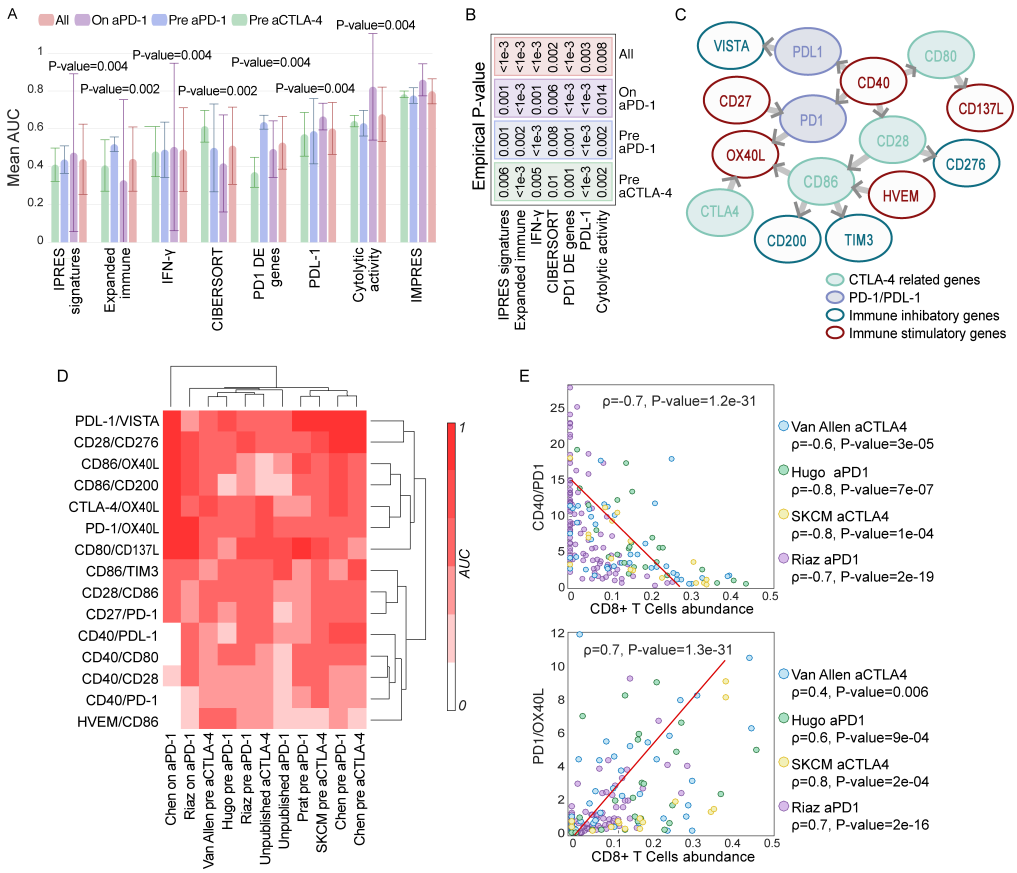

Figure 10. IMPRES features. (A) Bar plot comparing IMPRES performance to that of other published approaches across 9 publicly available ICB treatment datasets. The performance obtained by each approach is displayed via four bars, each representing the AUC for a specific treatment group, with the rightmost bar denoting the overall performance across all samples. The Rank-sum P-values comparing the performance of each classifier evaluated to that of IMPRES over all samples are presented (P-value of 0.002 is achieved when IMPRES AUC is larger than that obtained by the other predictor for all 9 datasets, and 0.004 when it is larger for 8/9 datasets). **(B)** A table showing the empirical P-values comparing IMPRES performance to that of each of the other predictors in the three different ICB treatment classes and for the

aggregate of all datasets (the value of '<1e-3' denotes that IMPRES' prediction performance was superior to that of the predictor with which it was compared in all 1,000 sampled test repetitions). **(C)** A network representation of the 15 pairwise features comprising IMPRES. Each node represents an immune checkpoint gene and each edge describes a pairwise relation (an IMPRES feature). The direction of edge A -> B denotes that the higher expression of A vs. that of B is associated with better patients' response. The color of the outline of each node denotes if it is inhibitory or activating and its fill color denotes whether it belongs to the PD1 or CTLA-4 pathways. **(D)** Clustogram of the individual predictive power of the 15 IMPRES features (based on their expression ratios) in each of the melanoma treatment datasets studied (the color scaling denotes the AUC obtained using each individual ratio as a response predictor, ranging from 0 to 1). **(E)** Scatter plots showing the correlation between CIBERSORT-inferred CD8+ T cells abundance (X-axis) and the gene expression ratios of two IMPRES features that are significantly associated with it (Y-axis); CD40/PD1 (upper panel) and PD1/OX40L lower panel). The Spearman ρ and associated P-values are shown for each ICB response data[125,127,128,130] individually and for all four datasets together (excluding nanostring datasets in which low coverage severely degrades CIBERSORT performance and were hence not included in this analysis).

Methods

Collection of immune checkpoint molecules

To build a predictor based on pair-wise relations between checkpoint genes' expression, we formed a list of 45 immune checkpoint genes with known co-stimulatory or co inhibitory effects, collected from literature reports [143–146]. From these, we focus on 28 genes that were measured in all RNA-sequencing datasets analyzed in this paper.

Feature selection and IMPRES construction on the NB data

For feature selection, we use the quantile-normalized expression of the 28 immune checkpoint genes selected above in the 108 NB tumor samples studied, using the following expression function of pairs of checkpoint genes as features:

$$F_{i,j}(\mathrm{x}) = \begin{cases} 1, & exp_i(x) < exp_j(x) \\ 0, & otherwise, \end{cases}$$

Where $exp_i(x)$ and $exp_j(x)$ denote the expression of genes $i\ and\ j$ in sample x.

We focus on pairs where at least one of the genes is among the six genes that are directly associated with anti-CTLA-4 and anti-PD1 blockade therapy, including CTLA-4, CD28, CD80/CD86, PD-1 and PD-L1¹⁴⁷ (Buchbinder & Desai, 2016)(Buchbinder & Desai, 2016)(Buchbinder & Desai, 2016)(Buchbinder & Desai, 2016), which together form 294 potential

gene pairs. To select features that best separate positive from negative samples in the NB data, we performed a hill climbing aggregative feature selection involving 500 iterations of a five-fold cross validation procedure, where the features that highly scored consistently across folds were selected for IMPRES.

Immune pathway enrichment analysis

To identify CDPs (consistently differentially expressed immune pathways in melanoma ICB responders), we first identified the genes that are up and down regulated in ICB responders vs non-responders for each of the datasets[125,127,128,130] (using one sided Rank-sum P-value<0.05). Then, we performed a GO pathway[148] enrichment analysis for immune related pathways via a hyper-geometric test, to identify (1) pathways that are consistently up or down regulated (hyper-geometric P-value<0.05) in responders for all anti-PD-1 melanoma datasets, and (2) pathways that are consistently up or down regulated in responders for all anti-CTLA-4 melanoma datasets (Figure 8C).

To correlate CDPs with the IMPRES features, we then evaluated the Spearman rank correlation coefficients (ρ) and corresponding P-values between the median pathway expression level of each CDP (using the median expression of all genes in a pathway) and each of the IMPRES expression ratios. This is done across all samples in each of the following datasets: (1) the anti-PD-1 treated melanoma datasets (2) the anti-CTLA-4 treated melanoma

datasets (3) the non ICB-treated melanoma datasets and (4) the neuroblastoma dataset.

Computing IMPRES features' expression ratio

To evaluate the predictive performance and functional associations of individual IMPRES features in a more refined manner we used the expression ratio instead of the binary indicators in each sample (i.e. for each feature A>B we used A/B instead). The resulting AUCs obtained with each ratio feature for each ICB response data are presented in Figure 10D.

Applying IMPRES to predict ICB response of melanoma patients

To apply IMPRES, we calculate for each sample x, the $F_{i,j}(x)$ over the 15 IMPRES checkpoint pairs (features). This leads to a binary vector of length 15 for each sample. The total number of '1's in this vector denotes the sample's IMPRES score (ranging between 0 and 15). High scores predict good response. By varying the classification threshold over the different possible IMPRES score values we generate the ROC curves and the resulting AUCs presented in the main text for each melanoma dataset.

Unpublished data collection and preparation

RNA-sequencing of 31 anti-PD-1 pre- and on-treatment tumor specimens, and 10 anti-CTLA-4 pre- and on- treatment tumor specimens derived from patients with metastatic melanoma was conducted as previously described in

Jenkins et al[149]. These patients were enrolled in clinical trials at Massachusetts General Hospital. Clinical trial registration numbers at ClinicalTrials.gov are NCT01714739; NCT02083484; NCT01543698; NCT01072175; NCT00949702; NCT01783938; NCT01006980.

Clinical response classification

Table 4 enclosed by summarizes the response annotations and criteria used for establishing them in the original study.

	Van Allen et al.	Hugo et al.	TCGA SKCM	Chen et al.	Prat et al.	Riaz et al.	Unpublished data
classified as "response"	'response',	'Complete Response',	'Complete Response',	'R' = freedom from disease/ decreased tumor > 6 months	'CR', 'PR'	'CR', 'PR'	'CR', 'PR'
		'Partial Response'	'Partial Response'				
classified as "non-response"	'nonresponse','long-survival'	'Progressive Disease'	'Clinical Progressive Disease', 'Stable Disease'	'NR' = tumor growth on serial CT scans or a clinical benefit lasting 6 months or less	'PD'	'PD', 'SD'	'PD'
Protocol	irRECIST[150]	RECIST[151]	RECIST[151]	Nan	RECIST[151]	RECIST[151]	RECIST[151]

Table 4. Response annotations for each melanoma dataset

Kaplan Meier survival analysis

Kaplan Meier analysis is performed by comparing the survival of patients with high IMPRES scores (> median(IMPRES)) to those with low IMPRES scores (< median(IMPRES)) using a log-rank test. The patients with median IMPRES score (= median(IMPRES)) are grouped with the smaller-size group among the two groups mentioned above.

Discussion

In summary, IMPRES' high predictive performance is mainly due to two key conjectures: (a) key immune mechanisms underlining spontaneous regression in NB can predict response to ICB, and (b) specific pairwise relations of immune checkpoint genes' expression can be predictive of spontaneous regression of NB and response to ICB in melanoma. Our results demonstrate that building on these assumptions leads to a predictor of response to checkpoint therapy that is significantly superior to the state-of-the-art and displays robust performance across many different melanoma datasets. From a translational standpoint, we show that IMPRES can correctly capture almost all true responders while misclassifying less than half of the non-responders, sparing unnecessary treatments for non-responding patients. Future studies are warranted to further study the predictive performance of the approach presented here in other cancer types where ICB is approved.

Chapter 4: Cancer chemoradiotherapy treatment modifiers

Prediction of patients with complete pathological response to chemo-radiation therapy (CRT) and identification of targets that modulate patients' response and mitigate resistance to CRT

Introduction

Rectal carcinomas account for approximately 20% of all colorectal cancers. Patients with stage II and III rectal carcinoma are treated with chemo-radiotherapy (CRT) before surgery to reduce the rate of local recurrences. However, not all patients respond equally well to CRT, with response ranging from complete response, i.e., no tumor cells left (Approximately 15-27% of patients (Sanghera, Wong et al. 2008, Maas, Nelemans et al. 2010)), to primary resistance. Clearly, accurately predicting the response to CRT before treatment commences would be immensely useful: patients with a predicted complete response and having other comorbidities might be spared surgery. Alternatively, patients whose tumors are resistant could be treated to increase sensitively to CRT. Our goal is to build a predictor that will predict complete response to CRT based on gene expression data characterizing the tumor of each individual patient. This analysis will also reveal the set of discriminating genes and accompanying molecular features that play a key role in modulating the response to CRT in rectal cancer and will enable identification of targets that play a role in modifying resistance to CRT in rectal cancer, whose inhibition may mitigate this resistance.

Results

Identification and cross validation

To avoid over-fitting we first performed a feature selection procedure using only small number of randomly selected samples (20% of the data). To this end, we randomly select eight positive (5% of the patients, **Tumor Regression Grade (TRG)** = 100%) and eight negative (5% of the patients, **Tumor Regression Grade (TRG)** < 45%) cases and search for differentially expressed transcripts between these samples, from which to selected the features for the classifier. To reliably identify complete responders, without misclassifying partial responders, we aimed to select features that will maximize the sensitivity of the generated classifier. Therefore, we randomly select a second set of 8 positive and 8 negative samples and performed a hill-climbing procedure [152], gradually adding transcripts from the group of differently expressed transcripts, that improve the sensitivity of a resulting SVM classifier, when applied to the second group of samples. A signature of 42 transcripts resulted in a classifier with maximal sensitivity on the 16 randomly selected test samples. The resulting classifier generated from 42-transcript signature was than assessed through a five-fold cross validation procedure on training set comprising 32 positive (TRG = 100%) and 32 negative (TRG < 45%) cases. The cross validation procedure resulted in a sensitivity of 0.46 (when allowing zero false positives), AUC of 0.86 and an Accuracy of 0.8 (Figure 11A). Based on these 64 cases, an SVM machine was established and applied to all 161 cases with TRG ranging from 10% to 100%.

Encouragingly, we found that when applying our classifier to the entire data, we get a sensitivity of 0.31, AUC of 0.97 and an Accuracy of 0.86 (Figure 11B-C), indicating that even when considering the full range of TRG (10-100%), our classifier can correctly identify more than 30% of the complete responders with no errors, using a group of 42 transcripts (36 unique genes).

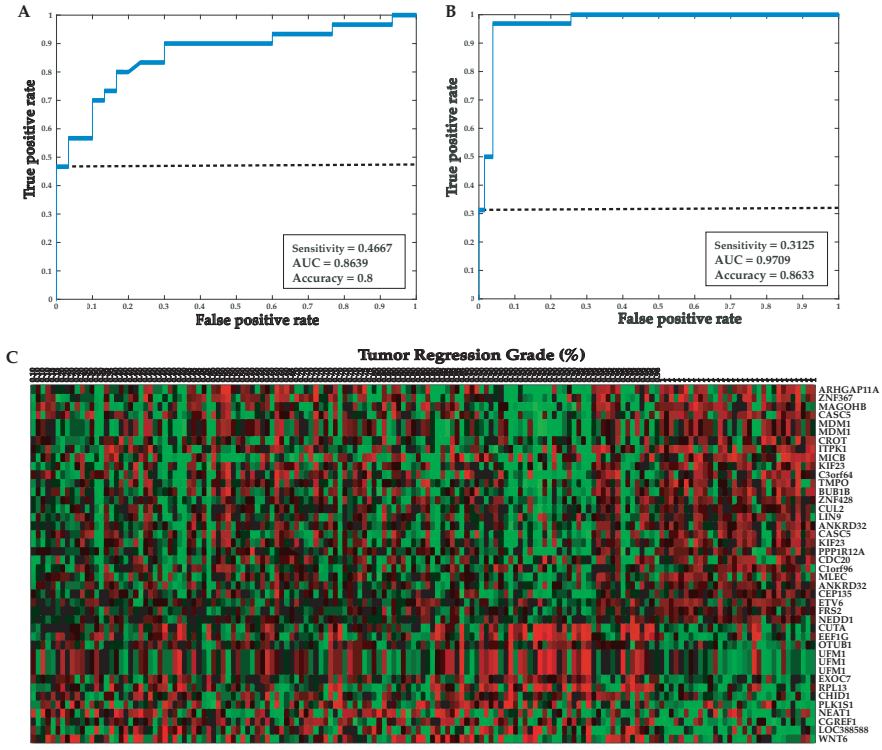

Figure 11. Cross validation performance. (A) and (B) Receiver operating characteristic (ROC) curve for the cross validation procedure and for the full set of 161 patients, respectively. The dashed line represents the objective of the rate of true positive sample when the false positive is zero. (C) A heatmap showing the 42-gene signature for all 161 patients, sorted by TRG of these patients.

Validating with an independent datasets

A major issue impeding the emergence of local advanced rectal cancer prognostic or diagnostic tools from molecular signatures is that these signatures are often found not reproducible when applied on independent datasets (Conde-Muino, Cuadros et al. 2015). To prove this point, we first applied our classifier on an independent test that was completely left out to this point, including 14 samples from which 4 are complete responders. Our classifier results in a sensitivity of 0.25 (when allowing zero false positives), AUC of 0.6250 and an Accuracy of 0.785 (Figure 12A-B)

Next, we tested several recently published classifiers on our data set consisting of 161 patients and an independent data set from (Millino, Maretto et al. 2017). We find that none of these signatures show a predictive signal when applied to our or the data from Millino et al.

In contrast, when applied to the same independent dataset from of 0.82 and an Accuracy of 0.84 (Figure 12C-D), indicating that even when applied on a completely independent dataset, our predictor can accurately foresee 25% of the complete responders while never misclassifying an incomplete responder.

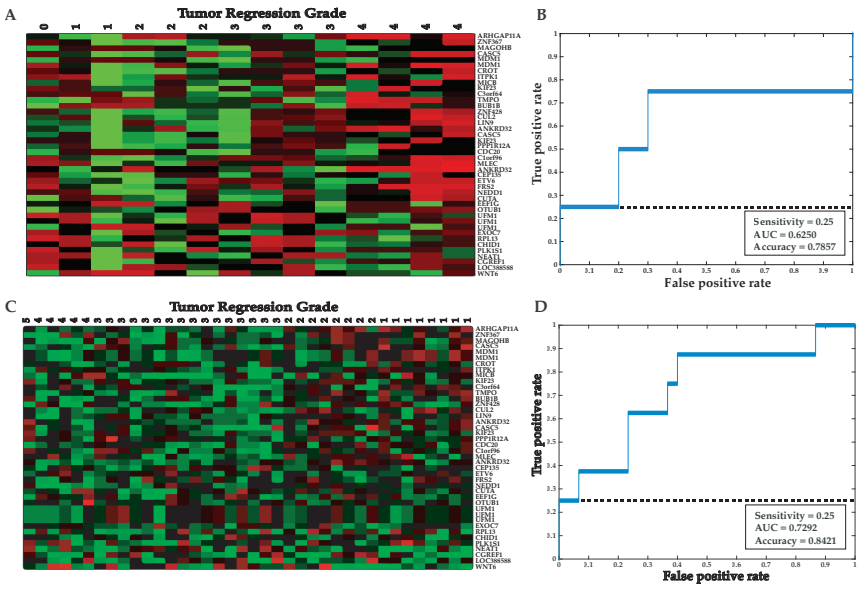

Figure 12. Performance for two independent datasets. (A) A heatmap showing the 42-gene signature for the 14 patients in our test set (B) Receiver operating characteristic (ROC) curve resulting from applying our classifier to the left-out test set. (C) A heatmap showing the 42-gene signature for the 38 patients in the dataset from (Millino, Maretto et al. 2017), sorted by TRG of these patients. (D) Receiver operating characteristic (ROC) curve obtained by applying our classifier on independent dataset with 38 patients. The dashed line represents the objective of the rate of true positive sample when the false positive is zero.

Colorectal cancer patient survival prediction

Patient complete response (pCR) is associated with an improved survival and a more favorably oncological outcome. We therefore hypothesized that due to their biological function; the expression patterns of genes in the classifier predicting pCR in rectal cancer patients should be also associated with good prognosis. To this end, we calculated a patient specific score using the 42 genes in the selected signature. For each gene, $g \in G_{rsp}$ is defined if its expression is increased in responder samples in the training set, and $g \in G_{rst}$ is defined if its expression is increased in resistant samples in the training set. The score of each patient is then defined by,

$$score(patient_i) = \frac{\sum_{i \in G_{rsp}} exp_i(g)}{\sum_{i \in G_{rst}} exp_i(g)}$$

As expected, we find significant correlation between these scores and Tumor regression grade (rho = 0.3398, Spearman p = 1.03e-05, Figure 13A). We then calculated this score for each patient in the TCGA (Cancer Genome Atlas 2012) colorectal cancer database (n=276) and in a second dataset from (Jorissen, Gibbs et al. 2009) colorectal cancer (n=290). Strikingly, we find that high score is significantly associated with improved survival (Log-rank p=0.0051 and 0.021, respectively. Figure 13B-C). This testifies further on the robustness of our signature and its prognostic value.

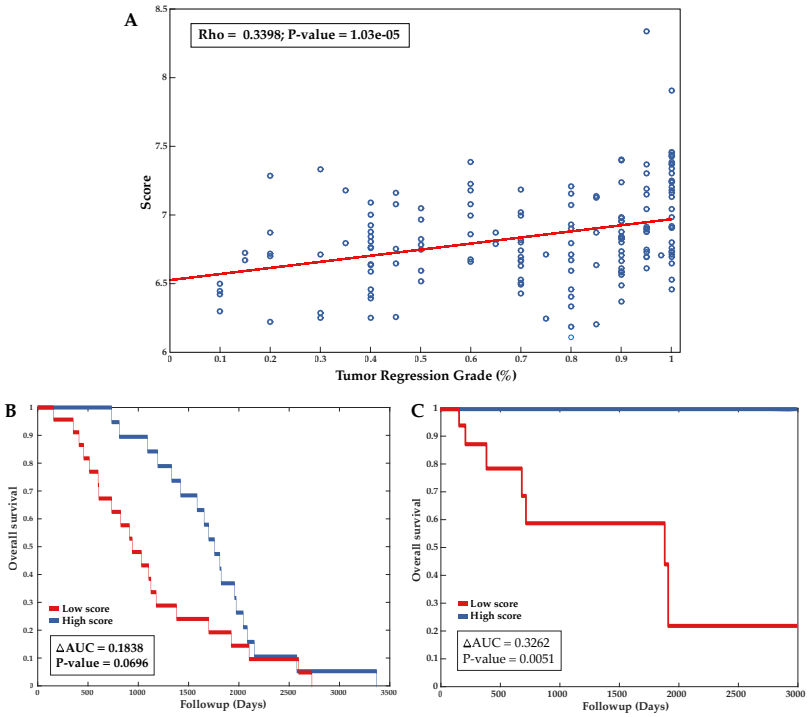

Figure 13. Patients score of response to CRT predicts survival in two independent datasets of colorectal cancer. (A) A scatter plot showing the correlation between the tumor regression grade in percent (X-label) and the gene signature score assigned to each patients representing predicted response to CRT (Y-label). (B) and (C) are the Kaplan-Meier survival curves for colorectal cancer patients predicted by using this score to have the best and worst prognosis (top and bottom 10% of patients scores, respectively; Methods) in the TCGA database and data from [153]

Chapter 5: Monogenetic disorders modifiers

In preparation as "System-wide identification of genomic modifiers of monogenetic disorders"

Introduction

Substantial clinical variability is observed in many Mendelian diseases, so that patients with the same mutation may develop a very severe form of disease, a mild form or show no symptoms at all. Among the factors that may explain these differences in disease manifestations are *modifier genes* [154]. Identifying these genetic modifiers may be of great interest from both treatment and genetic counseling perspectives [155]. However, very few modifier genes have been identified so far and the mechanisms underlying clinical variability of Mendelian disorders remain poorly understood, mainly due to the low frequency of the mutations causing these disorders and the scarcity of available data.

Strategies used to show the role of genetic factors in phenotypic expression are often classified into three categories depending on the type of data available [38]: (1) Association studies of case-control data, which is the most widely used strategy in the search for modifier genes, probably as it requires sampling patients only, rather than collecting familiar data. In association studies, the distribution of marker genotypes is compared in patients with different levels of the phenotype. (2) Linkage studies, which require available data from affected siblings. Linkage analysis compares the number of alleles shared identical by descent by affected siblings between phenotypically-concordant and discordant sibling pairs. (3) Blind search - Systematic

genome-wide screens, which consists in searching for the genetic factors involved in the phenotype of interest over the whole genome, to identify individuals that are resilient to mutations causing the phenotype of interest.

Here we present an approach for genome-wide GENtic moDULators identiFication (GENDULF) that is applicable in the lack of large DNA sequencing data. GENDULF operates by mining tissue gene expression of healthy and disease bearing individuals to identify expression patterns of genes that may modify disease severity. We first apply our approach to identify tissue specific modifiers of Cyctic Fibrosis (CF), for which considerable efforts has already been invested to find genetic modifiers. GENDULF prioritize most of the modifiers previously identified for CF in both lung and colon tissues (via linkage and association studies), and points to a few new candidates that may potentially bear a modifying role. To experimentally validate our approach, we then apply it to Spinal Muscular Atrophy (SMA), for which fewer modifiers have been previously discovered. We find one gene, U2AF1, out of four candidates arising initially, that consistently increases the ratio between full length SMN2 to Δ7SMN2, thus potentially increasing the levels of the SMN protein and improving the phenotype for SMA patients.

Results

GENDULF pipeline

We set to search for genes that modify the phenotype associated with a monogenetic disorder, and specifically, we aim to find genes whose down

regulation would result with a healthier phenotype, as these may have therapeutic value if targeted by drugs. GENDULF is based on the notion that modifier genes may be active in *healthy individuals* where a gene that is causing a monogenetic disorder (termed GCD) is lowly expressed, but that these are probably inactive in *disease bearing individuals*. Based on this rational, **GENDULF** proceeds in two main steps (Figure 14): (1) First, for a given monogenetic disorder studied, we identify Potential Modifiers (PMs) which are genes that are particularly lowly expressed when the GCD is lowly expressed in the relevant tissue for the disease. This association may potentially underlie the rescued phenotype observed when the GCD is inactive in healthy individuals. (2) Then, we identify Disease associated PMs (DPMs). To this end, we examine the expression of the PMs in studies containing both diseased and control samples. We hypothesize that if a gene is a genetic modifier whose low expression confers a healthy phenotype (as indicated by the first step) then we should expect to find it highly expressed in *disease* samples, which evidently are not rescued (i.e. DPM). This is in contradistinction to genes that are co-expressed with a gene causing a monogenetic disorder in healthy tissues but are not genetic modifiers, whose expression should remain correlated with that of the GCD also in the disease samples (non DPMs).

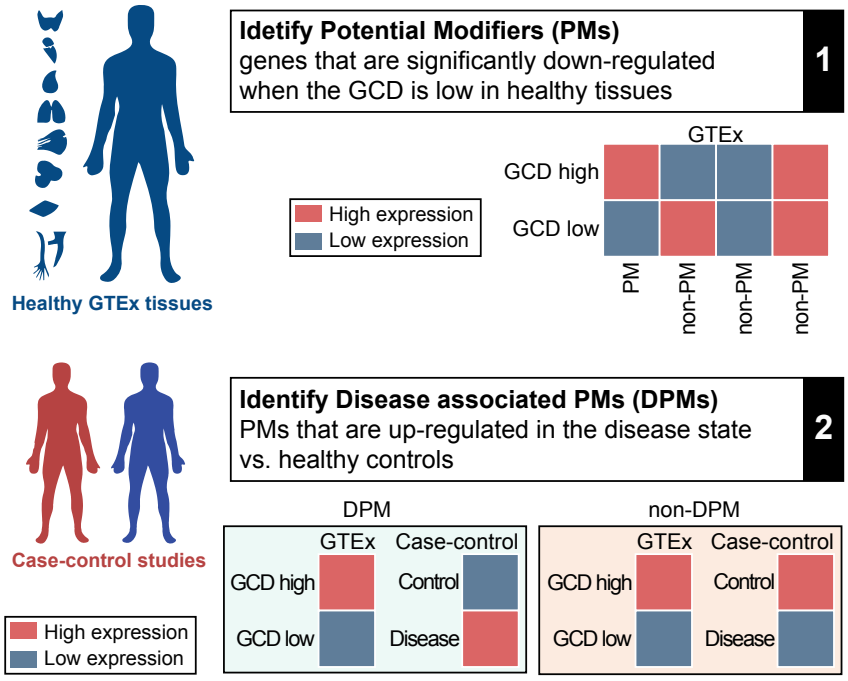

Figure 14. An overview of GENDULF computational pipeline. (1) Mining transcriptomics of healthy tissues afflicted by a disorder to identify PMs. (2) Evaluate the expression of the PMs in studies containing both diseased and control samples to find DPMs genes that are not co-expressed with the GCD in disease tissues.

Cystic Fibrosis

Cystic fibrosis (CF) is an inherited disorder that causes thick, sticky mucus to form in the lungs, pancreas, colon and other organs. In the lungs, thick mucus can damage tissue and block airways, making it difficult for patients to breathe and promoting lung infections [156]. CF is caused by mutations in the *CFTR* gene result in defective cystic fibrosis transmembrane conductance regulator (CFTR) proteins [157,158]. Normally, CFTR proteins located on the surface of the epithelial membrane act as chloride channels that in turn

regulate the epithelial sodium channel and other anion channels at the cell surface. The complex interplay of these channels regulates the electrochemical gradient that allows appropriate airway surface liquid depth and mucus viscosity [159,160]. CF affects 60,000 individuals worldwide and is a good model for the identification and characterization of factors that influence disease variation as its high prevalence provides a large number of accessible patients to perform detailed phenotypic analyses [161,162]. So far a few modifiers has been identified for CF lung disease severity through association studies [162,163]. We hence set to identify tissue specific CFTR modifier and compare our findings against these recently identified genetic modifiers.

We First applied GENDULF to analyze 320 healthy lung samples from the GTEx database, which yielded 55 PMs. We then examine the expression of these genes in nasal brushings of the inferior turbinates of mild and sever CF patients and healthy controls [164]. We find that the while the CFTR gene expression generally decreases in mild and severe CF patients vs. controls, the expression of the 14 of the PMs increases in severe CF (Figure 15A), testifying to their potential modifying role.

Reassuringly, some of the CFTR modifiers identified in lung tissues are known modifiers of CF manifestations in the lung, including the EHF gene, a known modifier gene of lung disease severity in CF [162] and SLC6A14, which has been recently identified as potential modifier of lung disease severity in CF [163].

Second, we applied GENDULF to analyze 345 healthy colon samples from the GTEx database, yielding 11 candidate genes. Examining the expression of these genes in rectal mucosal epithelia from CF patients and healthy controls [165], we find that the expression levels of four of these candidates is higher in CF than in normal control samples, thus all predicted as CF genomic modifiers in the colon (Figure 15B). Interestingly, the knockdown of one of the four identified modifiers, FABP1, rescues a lethal intestine defect in mouse model of CF [166]

A recently published study [163] has pointed to five loci that display significant association with variation in CF lung disease. The identification of a gene that causally affects disease variation is challenging as such association loci may typically include many genes [38]. We applied **GENDULF** to evaluate the genes within these five loci to identify candidate modifiers of CF lung disease severity.

To this end, in a given loci, for each gene we evaluated the level by which its expression is significantly down regulated when the expression of CFTR is extremely low in healthy lung tissues, quantified via a Wilcoxon rank-sum P-value. We find that in four of the five loci we studied (all but chr5p15.3) at least one gene is showing such a significant *functional association (FA)* and that the level of these FAs varies monotonically with the genomic location of the gene (it is gradually increasing up to a maximum and then gradually decreasing). Robustness analysis shows that this pattern of monotonically

varying FAs does not occur in random, as it is never obtained when analyzing randomly shuffled CFTR expression values. This, it is likely that these ordered monotonic expression patterns truly capture FAs with the CFTR gene. Furthermore, notably the modifier genes of CF pointed by the authors of [162,163] mostly fall at maxima points, as shown in Figure 15C, including EHF, MUC4, HLA-DRA and SLC614.

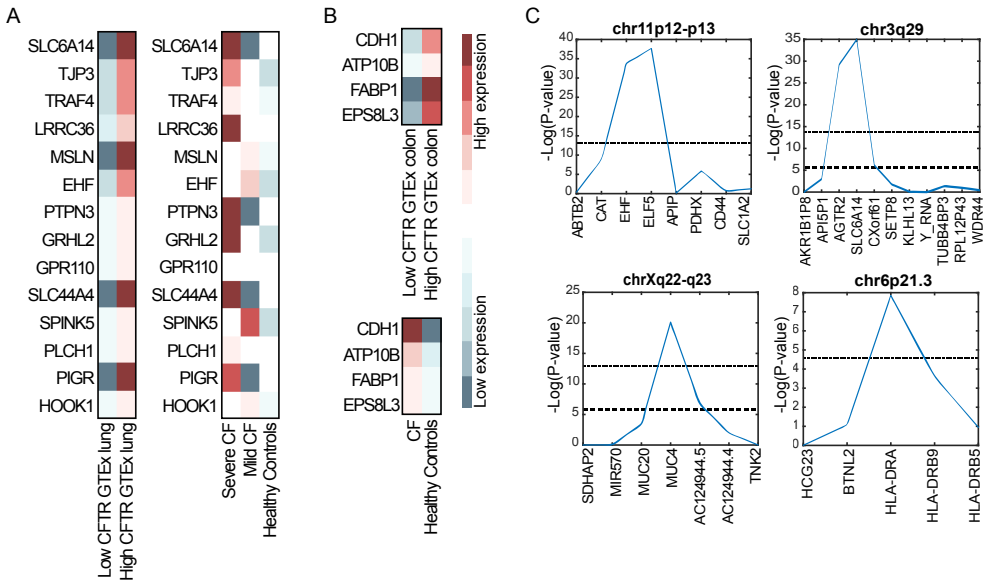

Figure 15. CF identified modifiers. (A) CF identified modifiers expression in healthy lung for high vs. low CFTR expression (left panel) and in severe CF, mild CF and healthy controls. (B) CF identified modifiers expression in healthy colon for high vs. low CFTR expression (top panel) and for CF and vs controls (bottom panel). (C) The p-values assigned to genes within chr11p12-p13, chr6p21.3, chr3q29 and chrXq22-q23 ordered by their location. The lower dashed line represent a significant threshold corrected for the number of

evaluated genes and the upper dashed line represent a significant threshold corrected for all genes and transcripts in GTEx with alpha = 0.05.

Spinal Muscular Atrophy

Spinal muscular atrophy is an autosomal recessive neurodegenerative disease characterized by degeneration of spinal cord motor neurons, atrophy of skeletal muscles, and generalized weakness. SMA is caused by deletion or mutation of the *survival of motor neuron 1* (*SMN1*) gene, which encodes the SMN protein [167,168]. Humans carry two paralogous *SMN1* and *SMN2* genes that are ubiquitously expressed. The SMN protein is principally produced from *SMN1* full-length mRNA, as *SMN2* transcripts often goes through alternative splicing and exclusion of exon 7, resulting in mRNA that encodes an unstable SMN protein that is rapidly degraded [169,170]. However, the low levels of *SMN2* transcripts still produce small amounts of the fully functional SMN protein.

Here, We use healthy Muscle and Spinal cord tissues, in which SMN1 expression is considerably variable; to evaluate the expression of different SMN2 isoforms when SMN1 is especially low. In both tissues we find increased full-length SMN2 transcript levels when SMN1 expression levels become very low (Spearman Rho = -0.1427, -0.1432 and Rank-sum p-value = 1.3657e-05 and 0.084, for healthy muscle and spinal chord tissues, respectively). This support the assumption that rescuing mechanisms are

indeed in play in healthy tissues when SMN1 expression is very low, thus investigating the trascriptomic changes occurring in such healthy samples may reveal other GM of SMA, whose alterations may modify SMN levels in diseased tissues as well.

Hence, we applied GENDULF to find genetic modifiers for SMA whose knockdown (KD) will result in improved phenotype. To this end, we search for genes that fulfill all the four following criteria:

(1) Genes whose expression is significantly lower than expected when SMN1 levels are extremely low (bottom 10th percentile) in healthy muscle and spinal cord tissues, suggesting that their low expression may have a compensating affect for the loss of SMN1.

(2) Genes who's under activation is associated with higher ratio between full-length SMN2 levels and exon7 skipped SMN2 levels, testifying for their potential compensating affect through reducing the exon7 skipping of SMN2.

(3) Genes whose expression is not reduced by the KD of SMN1 (to simulate the SMA disease state, in either in iPSC-derived motor neurons or human SH-SY5Y cells). This indicated that the observed association between these genes and SMN1 in healthy tissues is not due to mere co-expression but may signify a true functional rescue effect in the maintaining these tissues as healthy, while the lack of this rescue effect results in an SMA disease like phenotype.

(4) Genes whose expression is higher in SMA vs. controls in Muscle or Spinal cord tissues, further testifying on a true modifying effect rather than a co-expression pattern.

This analysis points to two potential targets that withstand all three criteria with sufficient statistical significance, U2AF1 (Figure 16A-E) and HNRNPA0.

Experimentally testing these targets, we find that the top predicted target, U2AF1, indeed increases the ratio between the levels of full-length SMN2 levels and exon7 skipped SMN2 levels, as predicted (Figure 16F).

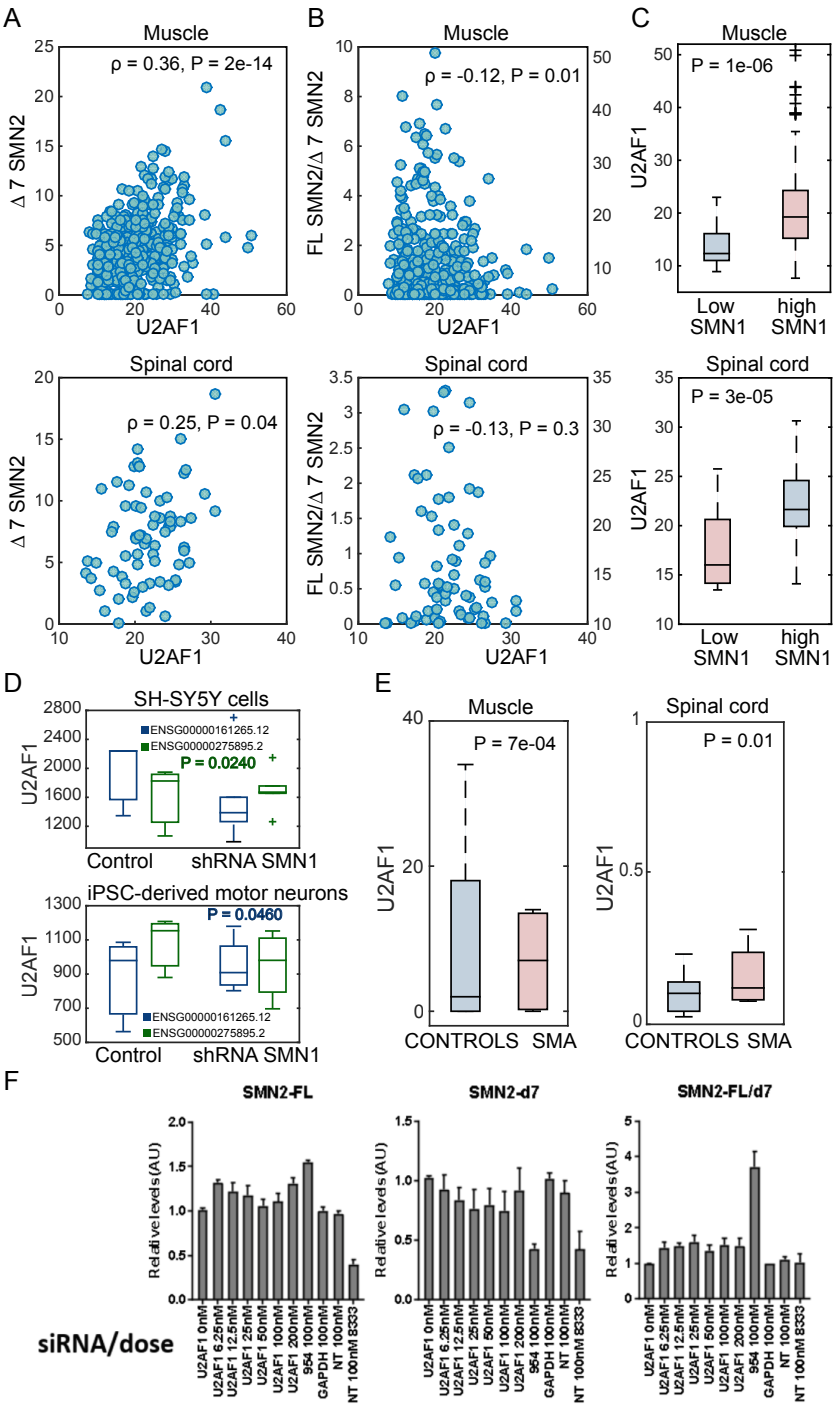

Figure 16. U2AF1 gene. (A) Scatter plots showing the correlation between U2AF1 and Δ7$SMN2$ levels in healthy muscle (upper panel) and healthy spinal cord (lower panel) (B) Scatter plots showing the correlation between U2AF1 and the ratio between full length SMN2 and Δ7$SMN2$ in healthy

muscle (upper panel) and healthy spinal cord (lower panel) (C) boxplot showing U2AF1 expression in healthy muscle (upper panel) and healthy spinal cord (lower panel) for samples with high vs. low SMN1 expression (D) Boxplot showing the expression of the two U2AF1 transcripts in controls vs. SMN1 shRNA in human SH-SY5Y cells (upper panel) and iPSC-derived motor neurons (lower panel). (E) Boxplots showing U2AF1 expression in controls vs. SMA in muscle (left panel) and spinal cord (right panel) (F) Experimental testing the affect of U2AF1 KD on full length SMN2, Δ7 SMN2 and the ratio between full length SMN2 and Δ7 SMN2.

****Figure 16F and the work presented in it is generated by Charlotte Sumner and Daniel Ramos, Johns Hopkins University School of Medicine*

Discussion

We present GENDULF, a novel systematic approach to identify generic genetic modulators for monogenetic diseases. GENDULF identifies expression patterns of genes that may modify disease severity using gene expression of healthy and disease bearing individuals, and it is the first approach for modifiers identification that is applicable in the lack of large DNA sequencing data. We first validate GENDULF in Cyctic Fibrosis (CF), where we show that it prioritizes most of the modifiers previously identified for CF in both lung and colon tissues (via linkage and association studies).

Additionally, we validate GENDULF in Spinal Muscular Atrophy (SMA), where we identify U2AF1 gene, which, as predicted, consistently increases the ratio between full length SMN2 to Δ7SMN2.

GENDULF may be applied to a spectrum of monogenetic diseases to prioritize potential modifiers and enable treatment possibilities for these orphan diseases.

Bibliography

1. Ding, L. *et al.* Somatic mutations affect key pathways in lung adenocarcinoma. *Nature* **455,** 1069–75 (2008).
2. Greenman, C. *et al.* Patterns of somatic mutation in human cancer genomes. *Nature* **446,** 153–158 (2007).
3. Wood, L. D. *et al.* The genomic landscapes of human breast and colorectal cancers. *Science* **318,** 1108–13 (2007).
4. Dees, N. D. *et al.* MuSiC: Identifying mutational significance in cancer genomes. *Genome Res.* **22,** 1589–1598 (2012).
5. Lawrence, M. S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499,** 214–8 (2013).
6. Gonzalez-Perez, A. *et al.* Computational approaches to identify functional genetic variants in cancer genomes. *Nat. Methods* **10,** 723–729 (2013).
7. Hodis, E. *et al.* A landscape of driver mutations in melanoma. *Cell* **150,** 251–263 (2012).
8. Gonzalez-Perez, A. & Lopez-Bigas, N. Functional impact bias reveals cancer drivers. *Nucleic Acids Res.* **40,** (2012).
9. Orth, J. D., Thiele, I. & Palsson, B. Ø. What is flux balance analysis? *Nat Biotechnol* **28,** 245–248 (2010).
10. Duarte, N. & Becker, S. a. Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proc. Natl. Acad. Sci. U. S. A.*

104, 1777–1782 (2007).

11. Ma, H. *et al.* The Edinburgh human metabolic network reconstruction and its functional analysis. *Mol. Syst. Biol.* **3,** 135 (2007).
12. Folger, O. *et al.* Predicting selective drug targets in cancer through metabolic networks. *Mol. Syst. Biol.* **7,** 501 (2011).
13. Yizhak, K. *et al.* A computational study of the Warburg effect identifies metabolic targets inhibiting cancer migration. *Mol. Syst. Biol.* **10,** 744 (2014).
14. 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,** (2012).
15. Agren, R. *et al.* Identification of anticancer drugs for hepatocellular carcinoma through personalized genome-scale metabolic modeling. *Mol. Syst. Biol.* **10,** (2014).
16. Nam, H. *et al.* A Systems Approach to Predict Oncometabolites via Context-Specific Genome-Scale Metabolic Networks. *PLoS Comput. Biol.* **10,** (2014).
17. Berg, J. M., Tymoczko, J. L. & Stryer, L. *Stryer Biochemie*. *Biochemistry textbook* (2007). doi:10.1007/978-3-8274-2989-6
18. Khatri, P., Sirota, M. & Butte, A. J. Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS Comput. Biol.* **8,** e1002375 (2012).
19. Hu, J. *et al.* Heterogeneity of tumor-induced gene expression changes in the human metabolic network. *Nat. Biotechnol.* **31,** 522–9 (2013).
20. Bordbar, A. *et al.* Minimal metabolic pathway structure is consistent with

associated biomolecular interactions. *Mol. Syst. Biol.* **10,** (2014).

21. Gatza, M. L. *et al.* A pathway-based classification of human breast cancer. *Proc. Natl. Acad. Sci. U. S. A.* **107,** 6994–6999 (2010).
22. Huang, S., Yee, C., Ching, T., Yu, H. & Garmire, L. X. A Novel Model to Combine Clinical and Pathway-Based Transcriptomic Information for the Prognosis Prediction of Breast Cancer. *PLoS Comput. Biol.* **10,** (2014).
23. Huang, S. *et al.* Novel personalized pathway-based metabolomics models reveal key metabolic pathways for breast cancer diagnosis. *Genome Med.* **8,** 34 (2016).
24. Taylor, I. W. *et al.* Dynamic modularity in protein interaction networks predicts breast cancer outcome. *Nat. Biotechnol.* **27,** 199–204 (2009).
25. Staiger, C. *et al.* A critical evaluation of network and pathway-based classifiers for outcome prediction in breast cancer. *PLoS One* **7,** e34796 (2012).
26. Cun, Y. & Fröhlich, H. F. H. Prognostic gene signatures for patient stratification in breast cancer - accuracy, stability and interpretability of gene selection approaches using prior knowledge on protein-protein interactions. *BMC Bioinformatics* **13,** 69 (2012).
27. Staiger, C., Cadot, S., Györffy, B., Wessels, L. F. a & Klau, G. W. Current composite-feature classification methods do not outperform simple single-genes classifiers in breast cancer prognosis. *Front. Genet.* **4,** 1–15 (2013).
28. van ’t Veer, L. J. *et al.* Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415,** 530–6 (2002).
29. Paik, S. *et al.* A Multigene Assay to Predict Recurrence of Tamoxifen-Treated,

Node-Negative Breast Cancer. *N. Engl. J. Med.* **351,** 2817 (2004).

30. Wang, Y. *et al.* Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet* **365,** 671–9 (2005).

31. Ghadimi, B. M. *et al.* Effectiveness of gene expression profiling for response prediction of rectal adenocarcinomas to preoperative chemoradiotherapy. *J. Clin. Oncol.* **23,** 1826–1838 (2005).

32. Watanabe, T. *et al.* Prediction of sensitivity of rectal cancer cells in response to preoperative radiotherapy by DNA microarray analysis of gene expression profiles. *Cancer Res.* **66,** 3370–3374 (2006).

33. Kim, I.-J. *et al.* Microarray gene expression profiling for predicting complete response to preoperative chemoradiotherapy in patients with advanced rectal cancer. *Dis. Colon Rectum* **50,** 1342–53 (2007).

34. Rimkus, C. *et al.* Microarray-based prediction of tumor response to neoadjuvant radiochemotherapy of patients with locally advanced rectal cancer. *Clin. Gastroenterol. Hepatol.* **6,** 53–61 (2008).

35. Brettingham-Moore, K. H. *et al.* Pretreatment transcriptional profiling for predicting response to neoadjuvant chemoradiotherapy in rectal adenocarcinoma. *Clin. Cancer Res.* **17,** 3039–3047 (2011).

36. Watanabe, T. *et al.* Prediction of Response to Preoperative Chemoradiotherapy in Rectal Cancer by Using Reverse Transcriptase Polymerase Chain Reaction Analysis of Four Genes. *Dis. Colon Rectum* **57,** (2014).

37. Lopes-Ramos, C. *et al.* Comprehensive evaluation of the effectiveness of gene expression signatures to predict complete response to neoadjuvant

chemoradiotherapy and guide surgical intervention in rectal cancer. *Cancer Genet.* **208,** (2015).

38. Génin, E., Feingold, J. & Clerget-Darpoux, F. Identifying modifier genes of monogenic disease: Strategies and difficulties. *Human Genetics* **124,** 357–368 (2008).

39. Feingold, E. Methods for linkage analysis of quantitative trait loci in humans. *Theor. Popul. Biol.* **60,** 167–180 (2001).

40. Macgregor, S., Craddock, N. & Holmans, P. A. Use of phenotypic covariates in association analysis by sequential addition of cases. *Eur. J. Hum. Genet.* **14,** 529–534 (2006).

41. Li, J.-L. *et al.* A genome scan for modifiers of age at onset in Huntington disease: The HD MAPS study. *Am. J. Hum. Genet.* **73,** 682–7 (2003).

42. Hauser, E. R. *et al.* Ordered subset analysis in genetic linkage mapping of complex traits. *Genet. Epidemiol.* **27,** 53–63 (2004).

43. Chen, R. *et al.* Analysis of 589,306 genomes identifies individuals resilient to severe Mendelian childhood diseases. *Nat. Biotechnol.* **34,** 531–538 (2016).

44. Auslander, N. *et al.* An integrated computational and experimental study uncovers FUT 9 as a metabolic driver of colorectal cancer. 1–16 (2017). doi:10.15252/msb.20177739

45. Barretina, J. *et al.* The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483,** 603–307 (2012).

46. Beroukhim, R. *et al.* Supp info: The landscape of somatic copy-number alteration across human cancers. *Nature* **463,** 899–905 (2010).

47. Fearnhead, N. S., Britton, M. P., Bodmer, W. F., Hospital, J. R. & Ox, O. The ABC of APC. *Hum. Mol. Genet.* **10,** 721–33 (2001).

48. Aoki, K. & Taketo, M. M. Adenomatous polyposis coli (APC): a multi-functional tumor suppressor gene. *J. Cell Sci.* **120,** 3327–35 (2007).

49. Hazra, A., Fuchs, C. S., Chan, A. T., Giovannucci, E. L. & Hunter, D. J. Association of the TCF7L2 polymorphism with colorectal cancer and adenoma risk. *Cancer Causes Control* **19,** 975–980 (2008).

50. Slattery, M. L. *et al.* Transcription factor 7-like 2 polymorphism and colon cancer. *Cancer Epidemiol. Biomarkers Prev.* **17,** 978–982 (2008).

51. Kinzler, K. W. *et al.* Identification of a Gene Located at Chromosome-5q21 That Is Mutated in Colorectal Cancers. *Science (80-.).* **251,** 1366–1370 (1991).

52. Song, M. S., Salmena, L. & Pandolfi, P. P. The functions and regulation of the PTEN tumour suppressor. *Nat Rev Mol Cell Biol* **13,** 283–296 (2012).

53. Nassif, N. T. *et al. PTEN mutations are common in sporadic microsatellite stable colorectal cancer. Oncogene* **23,** (2004).

54. Miyaki, M. *et al.* Higher frequency of Smad4 gene mutation in human colorectal cancer with distant metastasis. *Oncogene* **18,** 3098–3103 (1999).

55. Alazzouzi, H. *et al.* SMAD4 as a prognostic marker in colorectal cancer. *Clin. Cancer Res.* **11,** 2606–2611 (2005).

56. Yizhak, K., Gabay, O., Cohen, H. & Ruppin, E. Model-based identification of drug targets that revert disrupted metabolism and its application to ageing. *Nat. Commun.* **4,** 2632 (2013).

57. King, a, Selak, M. a & Gottlieb, E. Succinate dehydrogenase and fumarate

hydratase: linking mitochondrial dysfunction and cancer. *Oncogene* **25,** 4675–4682 (2006).

58. Kiuru, M. *et al.* Few FH mutations in sporadic counterparts of tumor types observed in hereditary leiomyomatosis and renal cell cancer families. *Cancer Res.* **62,** 4554–4557 (2002).

59. Frezza, C., Pollard, P. J. & Gottlieb, E. Inborn and acquired metabolic defects in cancer. *Journal of Molecular Medicine* **89,** 213–220 (2011).

60. Khamas, A. *et al.* Screening for epigenetically masked genes in colorectal cancer using 5-aza-2′-deoxycytidine, microarray and gene expression profile. *Cancer Genomics and Proteomics* **9,** 67–75 (2012).

61. Sabates-Bellver, J. *et al.* Transcriptome profile of human colorectal adenomas. *Mol. Cancer Res.* **5,** 1263–1275 (2007).

62. Gouveia, R. *et al.* Expression of glycogenes in differentiating human NT2N neurons. Downregulation of fucosyltransferase 9 leads to decreased Lewisx levels and impaired neurite outgrowth. *Biochim. Biophys. Acta - Gen. Subj.* **1820,** 2007–2019 (2012).

63. Nishihara, S. *et al.* ,3-Fucosyltransferase IX (Fut9) determines Lewis X expression in brain. *Glycobiology* **13,** 445–455 (2003).

64. Duarte, N. C. *et al.* Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proc. Natl. Acad. Sci. U. S. A.* **104,** 1777–1782 (2007).

65. Inufusa, H. *et al.* Ley glycolipid-recognizing monoclonal antibody inhibits procoagulant activity and metastasis of human adenocarcinoma. *Int. J. Oncol.*

19, 941–946 (2001).

66. Suzuki, M. *et al.* Le(y) glycolipid acts as a co-factor for tumor procoagulant activity. *Int. J. Cancer* **73,** 903–909 (1997).

67. Nudelman, E., Levery, S. B., Kaizu, T. & Hakomori, S. Novel fucolipids of human adenocarcinoma: Characterization of the major Le(y) antigen of human adenocarcinoma as trifucosylnonaosyl Le(y) glycolipid (III3FucV3FucVI2FucnLc6). *J. Biol. Chem.* **261,** 11247–11253 (1986).

68. Segrè, D., Vitkup, D. & Church, G. M. Analysis of optimality in natural and perturbed metabolic networks. *Proc. Natl. Acad. Sci. U. S. A.* **99,** 15112–15117 (2002).

69. Becker, S. A. & Palsson, B. O. Context-specific metabolic networks are consistent with experiments. *PLoS Comput. Biol.* **4,** (2008).

70. Lurje, G., Zhang, W. & Lenz, H.-J. Molecular prognostic markers in locally advanced colon cancer. *Clin. Colorectal Cancer* **6,** 683–690 (2007).

71. Fearon, E. R. Genetic alterations underlying colorectal tumorigenesis. *Cancer Surv* **12,** 119–136 (1992).

72. Qureshi-Baig, K., Ullmann, P., Haan, S. & Letellier, E. Tumor-Initiating Cells: a criTICal review of isolation approaches and new challenges in targeting strategies. *Mol. Cancer* **16,** 40 (2017).

73. Rybicka, A. & Król, M. Identification and characterization of cancer stem cells in canine mammary tumors. *Acta Vet. Scand.* **58,** 86 (2016).

74. Chan, T. S. *et al.* Metronomic chemotherapy prevents therapy-induced stromal activation and induction of tumor-initiating cells. *J. Exp. Med.* 1–22 (2016).

doi:10.1084/jem.20151665

75. Liu, C.-C. *et al.* Suspension survival mediated by PP2A-STAT3-Col XVII determines tumour initiation and metastasis in cancer stem cells. *Nat. Commun.* **7,** 11798 (2016).

76. Chiou, S. H. *et al.* Coexpression of Oct4 and Nanog enhances malignancy in lung adenocarcinoma by inducing cancer stem cell-like properties and epithelial-mesenchymal transdifferentiation. *Cancer Res.* **70,** 10433–10444 (2010).

77. Levings, P. P. *et al.* Expression of an exogenous human Oct-4 promoter identifies tumor-initiating cells in osteosarcoma. *Cancer Res.* **69,** 5648–5655 (2009).

78. Dalerba, P. *et al.* Phenotypic characterization of human colorectal cancer stem cells. *Proc. Natl. Acad. Sci. U. S. A.* **104,** 10158–10163 (2007).

79. Beck, B. & Blanpain, C. Unravelling cancer stem cell potential. *Nat. Rev. Cancer* **13,** 727–738 (2013).

80. Ricci-Vitiani, L., Fabrizi, E., Palio, E. & De Maria, R. Colon cancer stem cells. *Journal of Molecular Medicine* **87,** 1097–1104 (2009).

81. Cheng, X. & O'Neill, H. C. Oncogenesis and cancer stem cells: Current opinions and future directions. *J. Cell. Mol. Med.* **13,** 4377–4384 (2009).

82. Zhang, W. C. *et al.* Tumour-initiating cell-specific miR-1246 and miR-1290 expression converge to promote non-small cell lung cancer progression. *Nat. Commun.* **7,** 11702 (2016).

83. Bansal, N. *et al.* BMI-1 targeting interferes with patient-derived tumor-

initiating cell survival and tumor growth in prostate cancer. *Clin. Cancer Res.* (2016). doi:10.1158/1078-0432.CCR-15-3107

84. Grinshtein, N. *et al.* Small molecule kinase inhibitor screen identifies polo-like kinase 1 as a target for neuroblastoma tumor-initiating cells. *Cancer Res.* **71,** 1385–1395 (2011).

85. Merlos-Suárez, A. *et al.* The intestinal stem cell signature identifies colorectal cancer stem cells and predicts disease relapse. *Cell Stem Cell* **8,** 511–524 (2011).

86. Cortina, C. *et al.* EphB–ephrin-B interactions suppress colorectal cancer progression by compartmentalizing tumor cells. *Nat. Genet.* **39,** 1376–1383 (2007).

87. Auslander, N., Wagner, A., Oberhardt, M. & Ruppin, E. Data-Driven Metabolic Pathway Compositions Enhance Cancer Survival Prediction. *PLoS Comput. Biol.* **12,** (2016).

88. Patil, K. R. & Nielsen, J. Uncovering transcriptional regulation of metabolism by using metabolic network topology. *Proc. Natl. Acad. Sci. U. S. A.* **102,** 2685–2689 (2005).

89. Farah, I. O., Lewis, V. L., Ayensu, W. K. & Cameron, J. A. Role of fructose diphosphate (fdp) and glycerol on the differential survival of mrc-5 and a549 cell lines. *Biomed Sci Instrum* **48,** 112–8 (2012).

90. Casero Jr., R. A. *et al.* Differential Induction of Spermidine/Spermine N1-Acetyltransferase in Human Lung Cancer Cells by the Bis(ethyl)polyamine Analogues. *Cancer Res* **49,** 3829–3833 (1989).

91. Hong, S.-H. *et al.* Suppression of lung cancer progression by biocompatible glycerol triacrylate- spermine-mediated delivery of shAkt1. *Int. J. Nanomedicine* **7,** 2293–2306 (2012).

92. Allen, W. L. *et al.* The role of spermidine/spermine N1-acetyltransferase in determining response to chemotherapeutic agents in colorectal cancer cells. *Mol. Cancer Ther.* **6,** 128–37 (2007).

93. Roscilli, G. *et al.* Carnitines slow down tumor development of colon cancer in the DMH-chemical carcinogenesis mouse model. *J. Cell. Biochem.* **114,** 1665–1673 (2013).

94. Ihara, A. *et al.* Blockade of leukotriene B4 signaling pathway induces apoptosis and suppresses cell proliferation in colon cancer. *J. Pharmacol. Sci.* **103,** 24–32 (2007).

95. Curtis, C. *et al.* The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* **486,** 346–52 (2012).

96. Kaplan, E. L. & Meier, P. Nonparametric Estimation from Incomplete Observations. *J. Am. Stat. Assoc.* **53,** 457–481 (1958).

97. Harrell, F. E., Lee, K. L. & Mark, D. B. Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat. Med.* **15,** 361–387 (1996).

98. Carracedo, A., Cantley, L. C. & Pandolfi, P. P. Cancer metabolism: fatty acid oxidation in the limelight. *Nature reviews. Cancer* **13,** 227–32 (2013).

99. Menendez, J. a & Lupu, R. Fatty acid synthase and the lipogenic phenotype in cancer pathogenesis. *Nature reviews. Cancer* **7,** 763–777 (2007).

100. Kuhajda, F. P. Fatty acid synthase and cancer: New application of an old pathway. *Cancer Research* **66,** 5977–5980 (2006).

101. Possemato, R. *et al.* Functional genomics reveal that the serine synthesis pathway is essential in breast cancer. *Nature* **476,** 346–50 (2011).

102. DeBerardinis, R. J., Lum, J. J., Hatzivassiliou, G. & Thompson, C. B. The Biology of Cancer: Metabolic Reprogramming Fuels Cell Growth and Proliferation. *Cell Metabolism* **7,** 11–20 (2008).

103. Furberg, A.-S., Veierød, M. B., Wilsgaard, T., Bernstein, L. & Thune, I. Serum high-density lipoprotein cholesterol, metabolic profile, and breast cancer risk. *J. Natl. Cancer Inst.* **96,** 1152–60 (2004).

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

105. Derosa, C. a *et al.* Elevated osteonectin/SPARC expression in primary prostate cancer predicts metastatic progression. *Prostate Cancer Prostatic Dis.* **15,** 150–6 (2012).

106. Kuriakose, M. a *et al.* Selection and validation of differentially expressed genes in head and neck cancer. *Cell. Mol. Life Sci.* **61,** 1372–1383 (2004).

107. Chen, D. T. *et al.* Proliferative genes dominate malignancy-risk gene signature in histologically-normal breast tissue. *Breast Cancer Res. Treat.* **119,** 335–346 (2010).

108. Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: the next generation. *Cell* **144,** 646–74 (2011).

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

110. Ward, P. S. & Thompson, C. B. Metabolic Reprogramming: A Cancer Hallmark Even Warburg Did Not Anticipate. *Cancer Cell* **21,** 297–308 (2012).

111. Colijn, C. *et al.* Interpreting expression data with metabolic flux models: predicting Mycobacterium tuberculosis mycolic acid production. *PLoS Comput. Biol.* **5,** e1000489 (2009).

112. Cormen, T. H., Leiserson, C. E. & Rivest, R. L. *Introduction to Algorithms , Second Edition*. *Computer* **7,** (2001).

113. Alpert, C. J. & Kahng, A. B. Recent directions in netlist partitioning: a survey. *Integration, the VLSI Journal* **19,** 1–81 (1995).

114. Gallo, G., Longo, G., Pallottino, S. & Nguyen, S. Directed hypergraphs and applications. *Discret. Appl. Math.* **42,** 177–201 (1993).

115. Pujana, M. A. *et al.* Network modeling links breast cancer susceptibility and centrosome dysfunction. *Nat. Genet.* **39,** 1338–1349 (2007).

116. Dao, P. *et al.* Inferring cancer subnetwork markers using density-constrained biclustering. *Bioinformatics* **26,** 625–631 (2010).

117. Allahyar, A. & De Ridder, J. FERAL: Network-based classifier with application to breast cancer outcome prediction. in *Bioinformatics* **31,** i311–i319 (2015).

118. Venet, D., Dumont, J. E. & Detours, V. Most random gene expression signatures are significantly associated with breast cancer outcome. *PLoS Comput. Biol.* **7,** (2011).

119. Park, M. Y., Hastie, T. & Tibshirani, R. Averaged gene expressions for regression. *Biostatistics* **8,** 212–227 (2007).

120. Lee, E., Chuang, H.-Y., Kim, J.-W., Ideker, T. & Lee, D. Inferring pathway activity toward precise disease classification. *PLoS Comput. Biol.* **4,** e1000217 (2008).

121. Chuang, H.-Y., Lee, E., Liu, Y.-T., Lee, D. & Ideker, T. Network-based classification of breast cancer metastasis. *Mol. Syst. Biol.* **3,** 140 (2007).

122. Auslander, N. *et al.* 1,2, ,. 1–22

123. Su, Z. *et al.* An investigation of biomarkers derived from legacy microarray data for their utility in the RNA-seq era. *Genome Biol.* **15,** 523 (2014).

124. Jönsson, G. *et al.* Gene expression profiling-based identification of molecular subtypes in stage IV melanomas with different clinical outcome. *Clin. Cancer Res.* **16,** 3356–3367 (2010).

125. Van Allen, E. M. *et al.* Genomic correlates of response to CTLA-4 blockade in metastatic melanoma. *Science (80-.).* **350,** 207–211 (2015).

126. Chen, P. L. *et al.* Analysis of immune signatures in longitudinal tumor samples yields insight into biomarkers of response and mechanisms of resistance to immune checkpoint blockade. *Cancer Discov.* **6,** 827–837 (2016).

127. Hugo, W. *et al.* Genomic and Transcriptomic Features of Response to Anti-PD-1 Therapy in Metastatic Melanoma. *Cell* **165,** 35–44 (2016).

128. Weinstein, J. N. *et al.* The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.* **45,** 1113–20 (2013).

129. Prat, A. *et al.* Immune-related gene expression profiling after PD-1 blockade in

non–small cell lung carcinoma, head and neck squamous cell carcinoma, and melanoma. *Cancer Res.* **77,** 3540–3550 (2017).

130. Riaz, N. *et al.* Tumor and Microenvironment Evolution during Immunotherapy with Nivolumab. *Cell* (2017). doi:10.1016/j.cell.2017.09.028

131. Ejeta, G. *et al.* Genomic correlates of response to CTLA-4 blockade in metastatic melanoma. (2015).

132. Snyder, A. *et al.* Genetic Basis for Clinical Response to CTLA-4 Blockade in Melanoma. *N. Engl. J. Med.* 2189–2199 (2014). doi:10.1056/NEJMoa1406498

133. Colli, L. M. *et al.* Burden of nonsynonymous mutations among TCGA cancers and candidate immune checkpoint inhibitor responses. *Cancer Res.* **76,** 3767–3772 (2016).

134. Verdegaal, E. M. E. *et al.* Neoantigen landscape dynamics during human melanoma-T cell interactions. *Nature* **536,** 91–5 (2016).

135. Bhinder, B. & Elemento, O. Towards a better cancer precision medicine: systems biology meets immunotherapy. *Curr. Opin. Syst. Biol.* **2,** 67–73 (2017).

136. Ayers, M. *et al.* IFN- γ – related mRNA profile predicts clinical response to PD-1 blockade. *J. Clin. Invest.* **127,** 1–11 (2017).

137. Rooney, M. S., Shukla, S. A., Wu, C. J., Getz, G. & Hacohen, N. Molecular and genetic properties of tumors associated with local immune cytolytic activity. *Cell* **160,** 48–61 (2015).

138. Simon, R., Radmacher, M. D., Dobbin, K. & McShane, L. M. Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. *J.*

Natl. Cancer Inst. **95,** 14–18 (2003).

139. Tinker, A. V., Boussioutas, A. & Bowtell, D. D. L. The challenges of gene expression microarrays for the study of human cancer. *Cancer Cell* **9,** 333–339 (2006).

140. Ransohoff, D. F. Bias as a threat to the validity of cancer molecular-marker research. *Nat. Rev. Cancer* **5,** 142–149 (2005).

141. Zippelius, A., Schreiner, J., Herzig, P. & Muller, P. Induced PD-L1 Expression Mediates Acquired Resistance to Agonistic Anti-CD40 Treatment. *Cancer Immunol. Res.* **3,** 236–244 (2015).

142. Ahrends, T. *et al.* CD27 Agonism Plus PD-1 Blockade Recapitulates CD4+ T-cell Help in Therapeutic Anticancer Vaccination. *Cancer Res.* **76,** 2921–2931 (2016).

143. Chen, L. & Flies, D. B. Molecular mechanisms of T cell co-stimulation and co-inhibition. *Nature Reviews Immunology* **13,** 227–242 (2013).

144. Zhang, Q. & Vignali, D. A. A. Co-stimulatory and Co-inhibitory Pathways in Autoimmunity. *Immunity* **44,** 1034–1051 (2016).

145. Fuertes Marraco, S. A., Neubert, N. J., Verdeil, G. & Speiser, D. E. Inhibitory receptors beyond T cell exhaustion. *Frontiers in Immunology* **6,** (2015).

146. Ramsay, A. G. Immune checkpoint blockade immunotherapy to activate anti-tumour T-cell immunity. *British Journal of Haematology* **162,** 313–325 (2013).

147. Buchbinder, E. I. & Desai, A. CTLA-4 and PD-1 Pathways: Similarities, Differences, and Implications of Their Inhibition. *Am. J. Clin. Oncol.* **39,** 98–106 (2016).

148. Ashburner, M. *et al.* Gene ontology: Tool for the unification of biology. *Nature Genetics* **25,** 25–29 (2000).

149. Jenkins, R. W. *et al.* Ex Vivo Profiling of PD-1 Blockade Using Organotypic Tumor Spheroids. *Cancer Discov.* CD-17-0833 (2017). doi:10.1158/2159-8290.CD-17-0833

150. Eisenhauer, E. A. *et al.* New response evaluation criteria in solid tumours: Revised RECIST guideline (version 1.1). *Eur. J. Cancer* **45,** 228–247 (2009).

151. Hoos, A., Wolchok, J. D., Humphrey, R. W. & Hodi, F. S. CCR 20th anniversary commentary: Immune-related response criteria - Capturing clinical activity in immuno-oncology. *Clinical Cancer Research* **21,** 4989–4991 (2015).

152. Kohavi, R. & John, G. H. Wrappers for feature subset selection. *Artif. Intell.* **97,** 273–324 (1997).

153. Jorissen, R. N. *et al.* Metastasis-associated gene expression changes predict poor outcomes in patients with Dukes stage B and C colorectal cancer. *Clin. Cancer Res.* **15,** 7642–7651 (2009).

154. Nadeau, J. H. Modifier genes in mice and humans. *Nat. Rev. Genet.* **2,** 165–174 (2001).

155. Antonarakis, S. E. & Beckmann, J. S. Mendelian disorders deserve more attention. *Nat. Publ. Gr.* **7,** 277–282 (2006).

156. O'Sullivan, B. P. & Freedman, S. D. Cystic fibrosis. *Lancet* **373,** 1891–1904 (2009).

157. Kerem, E. *et al.* The Relation between Genotype and Phenotype in Cystic

Fibrosis — Analysis of the Most Common Mutation (ΔF 508). *N. Engl. J. Med.* **323,** 1517–1522 (1990).

158. Mohon, R. T., Wagener, J. S., Abman, S. H., Selfzer, W. K. & Accurso, F. J. Relationship of genotype to early pulmonary function in infants with cystic fibrosis identified through neonatal screening. *J. Pediatr.* **122,** 550–555 (1993).

159. Rich, D. P. *et al.* Expression of cystic fibrosis transmembrane conductance regulator corrects defective chloride channel regulation in cystic fibrosis airway epithelial cells. *Nature* **347,** 358–63 (1990).

160. Bear, C. E. *et al.* Purification and functional reconstitution of the cystic fibrosis transmembrane conductance regulator (CFTR). *Cell* **68,** 809–818 (1992).

161. Cutting, G. R. Modifier genes in Mendelian disorders: The example of cystic fibrosis. *Ann. N. Y. Acad. Sci.* **1214,** 57–69 (2010).

162. Wright, F. A. *et al.* Genome-wide association and linkage identify modifier loci of lung disease severity in cystic fibrosis at 11p13 and 20q13.2. *Nat. Genet.* **43,** 539–46 (2011).

163. Corvol, H. *et al.* Genome-wide association meta-analysis identifies five modifier loci of lung disease severity in cystic fibrosis. *Nat. Commun.* **6,** 8382 (2015).

164. Wright, J. M. *et al.* Respiratory epithelial gene expression in patients with mild and severe cystic fibrosis lung disease. *Am. J. Respir. Cell Mol. Biol.* **35,** 327–336 (2006).

165. Stanke, F. *et al.* The CF-modifying gene EHF promotes p.Phe508del-CFTR residual function by altering protein glycosylation and trafficking in epithelial

cells. *Eur. J. Hum. Genet.* **22,** 660–6 (2014).

166. Zhou, L. *et al.* Correction of lethal intestinal defect in a mouse model of cystic fibrosis by human CFTR. *Science* **266,** 1705–1708 (1994).

167. Brzustowicz, L. M. *et al.* Genetic mapping of chronic childhood-onset spinal muscular atrophy to chromosome 5q11.2-13.3. *Nature* **344,** 540–541 (1990).

168. Lefebvre, S. *et al.* Identification and characterization of a spinal muscular atrophy-determining gene. *Cell* **80,** 155–165 (1995).

169. Cho, S. & Dreyfuss, G. A degron created by SMN2 exon 7 skipping is a principal contributor to spinal muscular atrophy severity. *Genes Dev.* **24,** 438–442 (2010).

170. Burnett, B. G. *et al.* Regulation of SMN protein stability. *Mol. Cell. Biol.* **29,** 1107–15 (2009).