A computational toolkit to boost precision oncology with pharmacogenomics

TESI DOCTORAL UPF / 2019

Supervised by Dr. Patrick Aloy
Tutored by Dr. Baldo Oliva
Structural Bioinformatics and Network Biology Group
Joint IRB-BSC-CRG Program in Computational Biology

IRB BARCELONA INSTITUTE FOR RESEARCH IN BIOMEDICINE

upf. Universitat Pompeu Fabra Barcelona

Thesis advisor: Professor Patrick Aloy Lidia Mateo

A computational toolkit to boost precision oncology with pharmacogenomics

Abstract

During the last decade much progress has been made in understanding the genomic basis of human cancer thanks to the massive molecular profiling of cancer patients. Although this is a first and essential step, the most imperative challenge ahead is to find the way to exploit this wealth of data to fulfill the promise of personalized medicine and improve patient treatment and prognosis.

In this thesis, we present a computational toolkit that could be used to expand the applicability domain of precision oncology, contributing to close the bench to bedside gap. First, we developed two tools for contextualizing individual mutations (PanorOmics) and individual patients (OncoGenomic Landscapes) within the body of clinical and scientific evidence currently available. Then, we went a step further and developed Targeted Cancer Therapy For You (TCT4U), a methodology that exploits driver gene co-occurrences to predict drug response. Finally, we extracted transcriptional signatures that allow us to monitor the oncogenic activity of 10 mitogenic pathways in each patient, and to identify drugs that are likely to revert this oncogenic endophenotype.
Keywords: precision oncology, drug response biomarkers, pharmacogenomics

Resum

Durant la darrera dècada la seqüenciació massiva de pacients de cancer ha permés avançar de forma significativa el coneixement sobre les bases moleculars del càncer. Un cop assolit aquest primer pass, el repte que tenim per davant és explotar aquestes dades en benefici dels pacients oncològics, en el marc de la medicina personalitzada.

En aquesta tesi, presentem una sèrie d'eines computacionals que pretenen expandir el domini d'aplicació de les teràpies dirigides, contribuint així a escurçar la bretxa entre la recerca i la pràctica clínica. Primerament, presentem dues eines que permeten visualitzar les mutacions (PanorOmics) i els pacients de càncer (OncoGenomic Landscapes) i contextualitzar-los en el corpus de dades i coneixement actuals. Anant un pas més enllà, presentem 'Targeted Cancer Therapy For You' (TCT4U), una metodologia que permet predir resposta a tractaments en base a la co-ocurrència d'alteracions oncogèniques. Finalment, hem extret signatures transcripcionals que ens permeten monitoritzar en cada pacient l'activitat oncogènica de 10 vies de senyalització mitogènica, i identificar compostos químics que tenen el potencial de revertir-la.
Paraules clau: medicina personalitzada, biomarcadors de resposta farmacològica, pharmacogenómica

Contents

A les meves filles, Laia i Júlia, perquè desitjo de tot cor que creixeu en un món millor, envoltades d'una societat que valori la recerca i se'n nodreixi. Gràcies pel temps que m'heu cedit per poder-me desenvolupar a nivell personal i acadèmic. Com a contrapartida, espero poder-vos contagiar l'amor per la vida i la ciència. Cada dia que passa us estimo una mica més.

Agraïments

En primer lloc, m'agradaria donar les gràcies al director de la meva tesi, en Patrick Aloy, per haver-me donat l'oportunitat i els mitjans per fer el doctorat a l'SBNB. M'has animat a explorar i aprendre coses noves cada dia i, alhora, m'has ajudat a no perdre el nord. La teva visió i tenacitat ens han portat a tots plegats molt lluny i estic convençuda que seguirem avançant i superant nous reptes.

Diuen que no hi ha patró sense mariners. O potser era al revés? Bé, tan hi fa... perquè en Miquel no és un mariner, és un contramestre! A més de tenir una ment brillant, la teva capacitat de treball i generositat fan que crèixer al teu costat sigu un autèntic privilegi. Si bé hem estat físicament separats, jo treballant des de Macondo i tu des dels racons més remots del món, sempre t'he sentit molt aprop. Gràcies, de veritat. En Patrick i tu heu lluitat molt per aixecar un projecte molt potent i ha arribat el moment d'unir forces i recollir-ne els fruits que tant us mereixeu. Martino i Oriol, 'there we go!!', sou uns cracks. Amb tu Adrià, formem el "bio*" team, i som els consumidors principals de gingebre confitat al lab. No me n'oblido de tu, Carles, ni del teu amor pels llevats i les catànies. Tampoc del Csaba, que és el nostre metge de guárdia! i per suposat, algú ha de treballar 'de debò'- Edu, Sergi, Víctor, Isabelle- formem un gran equip.

Com no podia ser d'altra manera, se m'esboça un somriure als llavis només de pensar en les *Sexies* de l'SBNB. Com us enyoro, noies! Samira, Tere, Laura i Richa, no podria haver tingut una millor acollida al grup. Gràcies pel vostre caliu i sororitat. Roberto, Eva, Emre, Francesco, gràcies per nodrir aquest ecosistema d'idees, i gràcies també als estudiants de màster que han anat aportant aire fresc i entusiasme.

Més enllà de l'IRB, he d'agrair el suport indispensable que he rebut per part de la família i amics. He de dir que arribar fins aquí ha estat un treball en equip. De fet, la meva dona, la Sandra, es mereix tant com jo qualsevol reconeixement que es derivi de la meva feina. Em sento molt afortunada de comptar amb el suport dels meus pares i els meus sogres que sempre estan dispostos a ajudar en el que calgui. Per descomptat, voldria acabar per donar les gràcies més sentides a les meves germanes Diana i Joana, i a la meva mare Macu, al meu pare Isidre, i a la Laura, la seva dóna. Vosaltres heu fet possible tot el que sóc i tot el que he fet.

Preface

I have always been passionate about science but I became really addicted to it when I felt for the first time the excitement of making a discovery. As speleologist, I had the opportunity to explore deep sinkholes and, many times, I had to struggle through narrow passages to discover remote places that nobody else had ever seen.

Later, as an undergraduate student, I realized that I felt a similar excitement when doing my research in population genomics and environmental adaptation. During my stay at the Evolutionary and Functional Genomics Lab I started to develop scientific reasoning skills, I designed and performed experiments to test my hypotheses, and I acquired a strong background in genomics. The time that has passed has allowed me to appreciate more the importance of basic research but, at that time, I did not feel completely satisfied.

My dream had always been to investigate about human health and disease; I wanted to ”cure people”. This was my vocation when I joined the Structural Bioinformatics and Network Biology Group at the beginning of 2015. I was very fortunate to join Patrick’s lab and we were both very excited to start a new translational research project in the field of personalized cancer medicine.

Introduction

The oncogenic processes underlying cancer

Cancer is a genomic disease that is driven by evolutionary forces. On the timescale of our lifetime, somatic cells accumulate genomic alterations due to the exposure to endogenous and exogenous mutational processes. This generates an intrinsic genomic diversity that is the substrate for natural selection.

Out of the thousands of genomic changes that are typically identified in a single cancer sample, a very small subset of 'driver' alterations might confer a selective advantage to the cell and lead to a clonal expansion (Merlo et al. 2006). On the other hand, most alterations are just 'passengers' and, as such, they do not confer oncogenic properties. However, their sequence base composition and distribution along the genome can still be very informative about the underlying mutational processes (Alexandrov et al. 2013) and even about the tissue of origin of the tumor (Salvadores, Mas-Ponte, and Supek 2019).

Moreover, some passenger alterations might acquire a driver role after there has been a lesion in another gene, triggering a second round of clonal expansion. Such genetic dependencies determine the multistep acquisition of driver alterations, a process that has been particularly well studied in colorectal (Fearon and Vogelstein 1990) and pancreatic carcinogenesis (Fearon and Vogelstein 1990).

Cancer driver genes

From individual cancer cells and clones to the entire organism and the whole population, biological systems evolve dynamically in space and time at different scales. The action of natural selection over thousands of years on humans has constrained somatic evolution to try to maintain genomic stability and homeostasis. This has lead to the emergence of tumor suppressor mechanisms but has also left some oncogenic vulnerabilities in our genomes (Merlo et al. 2006). For this reason, germline mutations tend to converge into cancer susceptibility genes and somatic oncogenic alterations do so into driver genes.

This is of crucial importance when it comes to the identification of the handful of driver alterations that are typically buried among thousands of passenger alterations in a single tumor. Current methods exploit the signatures of positive selection as if it was a powerful magnet helping to find a needle in a haystack. Several statistical approaches have been developed to detect statistically significant patterns of recurrence (Lawrence et al. 2014), functional impact bias (Mularoni et al. 2016), and/or clustering patterns both at sequence level (Arnedo-Pac et al. 2019) and in the three-dimensional structure of proteins (Porta-Pardo, Hrabe, and Godzik 2015).

The most comprehensive characterization of mutational driver genes published to date identified a total of 299 consensus genes after analyzing over 9,000 tumor samples, across 33 tissues of origin, with seven complementary algorithms (Bailey et al. 2018). Despite the large number of drivers identified per tumor type, every patient has a unique combination of mutations and copy number variants: ninety percent of patients show at least one putative driver alteration, but each sample only contains a median of three putatively altered drivers (Rubio-Perez et al. 2015).

Large-scale cancer sequencing analysis performed to date seem to be adequately-powered to identify driver events at gene-level resolution and, indeed, it seems that we might be nearing the saturation of cancer driver gene discovery in most tumor types (Hsiehchen and Hsieh 2018). However, the discovery of specific driver mutations and mutational hotspots within driver genes is still far from complete and will probably be expanded as prospective clinical tumor sequencing becomes increasingly adopted (Chang et al. 2018).

Cancer as a disease of pathways

Over the past two decades, the advent of Next-Generation Sequencing has revolutionized the field of cancer genomics and large-scale analyses have revealed a genomic architecture that is more complex than perhaps anticipated (Weinstein et al. 2013; Kandoth et al. 2013; Garraway and Lander 2013). For many cancer types there are relatively few driver genes that are recurrently mutated across tumors and, beyond these, most driver alterations lie on the 'long tail' of genes that are only rarely mutated (Senft et al. 2017; Garraway and Lander 2013; Krogan et al. 2015).

Many of those rare drivers converge into key signaling pathways linked to proliferation and survival and it is widely believed that the observed molecular heterogeneity is, at least in part, due to natural selection acting at pathway-level (S. Wang et al. 2018; Senft et al. 2017; Leiserson et al. 2015; Garraway and Lander 2013); (Krogan et al. 2015). Hanahan and Weinberg formulated the 'hallmark' processes that alter cell signaling and confer cancer cells the ability to divide and grow uncontrollably (Hanahan and Weinberg 2011; Hanahan and Weinberg 2000). Those processes include the enhancement of some biological capabilities such as proliferation, invasion or metastasis, and the evasion of homeostatic mechanisms such as cycle arrest, apoptosis or immune surveillance.

This small set of organizing principles represents a solid conceptual framework for the understanding of cancer biology. However, only a subset of driver genes can be assigned to one or a few of those hallmarks based on prior knowledge. Following this rational, recurrent driver alterations in the TCGA PanCancer Atlas project were aggregated at pathway-level, yielding as a result the systematic characterization and manual curation of a set of 10 well-known mitogenic signaling pathways (Sanchez-Vega et al. 2018).

Preclinical Cancer Models

The lack of reporting of longitudinal clinical outcomes, and associated treatment history, greatly undermines the potential utility of large genomic studies for the development of novel predictive algorithms and biomarkers. This precious information is buried inside electronic health records and dispersed worldwide (GENIE Consortium 2017). Advancements in the field of natural language processing will probably automate and reduce the prohibitive costs of manual clinical curation, which remains the gold standard today.

Due to the lack of data together with other ethical and logistic barriers, the research community largely relies on drug response data gathered from pre-clinical cancer models.

Cancer cell lines

Cancer cell lines are the most widely used *in vitro* model system and have been fundamental tools to set the grounds of our understanding of cancer biology. Hundreds of cell lines have been established and propagated since HeLa, the first cultured cell line, was derived from a cervical cancer patient back in 1951 (Shoemaker et al. 1983). Cancer cell line panels were initially conceived as drug screening platforms for early drug development that tried to recapitulate the observed inter-patient variability in response to chemotherapy (i.e NCI-60 (Shoemaker 2006)).

With the advent of the "-omics" revolution, the research community realized that the specific molecular alterations of a patient's tumor were critical determinants of drug response, pointing to the need for much larger cell line panels (Gillet, Varma, and Gottesman 2013). The the Cancer Cell Line Encyclopedia (CCLE, (Barretina et al. 2012)) and the GDSC (formerly referred to as the Cancer Genome Project or CGP, (Garnett et al. 2012), comprising almost 1,500 cell lines derived from all kind of tumor types, are the contemporary attempts to obtain a better representation of the overall genomic diversity of cancer. Importantly, they also include representative cell lines of specific subpopulations of patients that could potentially benefit from targeted therapies.

Despite the broader molecular representativity of current panels, the caveats of cancer cell lines are well known and raise concerns about their relevance to predict clinical response. Most cancer cell lines have been cultured as monolayers on plastic surfaces, and in growth-promoting conditions, for thousands of generations. As a consequence, most of them have suffered substantial transcriptomic drift and, moreover, most likely represent a cell subpopulation from the original primary tumor (Domcke et al. 2013; Gillet, Varma, and Gottesman 2013). A functional evidence of this drift is that the efficacy of drugs indicated to treat a specific cancer type is not associated with the tissue from which the cell line was derived, and not even with the level of expression of the intended drug target (Jaeger, Duran-Frigola, and Aloy 2015).

Another important limitation of the aforementioned pharmacogenomic screenings is the lack of functional insights. While we can identify statistical associations between certain genomic alterations and drug efficacy, we cannot know how those alterations interfere with the mechanism of action of the drug. Cellular perturbation experiments offer a completely orthogonal approach that can help elucidating those mechanistic connections.

Subramanian and colleagues have recently published a large-scale compendium of gene expression signatures induced by genetic and pharmacologic perturbations. This resource, comprising >1.3M gene expression signatures induced by >20.000 small molecules and >5.000 genetic perturbations profiled across several cell types, has proven to be useful for a wide variety of important tasks (Subramanian et al. 2017). Beyond the direct annotation of chemical compound bioactivity, this resource can inform about the mechanism of action of unannotated compounds, and even predict the functional impact of mutations in allelic series of cancer driver genes (Berger et al. 2016; Subramanian et al. 2017).

In summary, although cancer cell lines might no longer reflect the tumors from which they were established, they are still an effective means to assess the functional impact of chemical and genetic perturbations. Cancer cell lines are therefore one of the most powerful tools available for experimental research. Nevertheless, measurements of drug efficacy should be interpreted with caution and validated in alternative models that better mimic the *in vivo* cancer microenvironment before they can be translated to the clinical setting.

Patient-derived xenografts

The use of a mouse to perform experiments on human tumors that have been surgically excised and subcutaneously implanted into the animal (also known as Patient-Derived Xenografts or PDXs), has been successfully exploited for more than 50 years. This approach has gained traction during the last two decades and has become a popular alternative to overcome several of the limitations of conventional cancer cell lines (Villacorta-Martin, Craig, and Villanueva 2017; Willyard 2018).

Once engrafted, the tumors grow in an environment that, although not human, better mimics their native environment than a plastic petri dish. As a result, PDXs faithfully preserve genetic, transcriptomic and histopathological features of the parental tumors, as well as their metastatic potential (Bruna et al. 2016; Byrne et al. 2017; Izumchenko et al. 2017). More importantly, the tumors can be propagated during serial passages, allowing researchers to determine the sensitivity of a single tumor to multiple treatments.

Indeed, PDXs also generally retain similar drug sensitivity profiles to those of the corresponding patient tumors (Byrne et al. 2017). Not only they recapitulate treatment response rates reported by previous clinical trials (H. Gao et al. 2015), but also parallel the clinical responses achieved by their donor patients in co-clinical trials. Manuel Hidalgo and his colleagues reported that PDXs treated with the same drugs that had been administered to the donor patients could predict both positive and negative clinical outcomes with a very high accuracy (87%; 112 out of 129). Those encouraging results are in line with the results reported elsewhere (Bruna et al. 2016; Byrne et al. 2017; Hidalgo et al. 2014; Krepler et al. 2017; Pompili et al. 2016).

Unfortunately, no cancer model is perfect and there are some technical and logistic barriers limiting their application for clinical decision support. Perhaps one of the most relevant shortcomings for patients and oncologists is that the process of PDX generation and drug testing is often too slow to benefit the donor and, many times, tumors simply fail to engraft in mice. The tumor engraftment rate varies widely across cancer types, being relatively easy to establish PDXs from colorectal, skin or high-grade ovarian and uterine sarcoma primary tumors, with >75% take rate (Byrne et al. 2017). However, other malignancies such as ER+ breast primary tumors seem to be particularly difficult to engraft, with 4-7% take rate (Byrne et al. 2017). In general, it has been observed that higher engraftment rate correlates with cancer aggressiveness and poor patient survival (Bruna et al. 2016; Pergolini et al. 2017).

It is also apparent that the use of PDXs as a cancer model has some intrinsic limitations. One of the main concerns is that PDXs are usually generated in highly immunodeficient mice (i.e NOD/SCID/Il2rg-/- or NOD/Rag1-/-/Il2rg-/- among others). An important milestone to overcome this limitation has been the generation of 'humanized' PDXs that produce human immune cells. To achieve this, stem cells from human umbilical cord are injected into mice very early in their development. As the mice grow, those cells differentiate into components of the human immune system, such as T cells (Willyard 2018). Those next-generation PDX models are now feasible to generate and are even suitable to investigate immunotherapies (M. Wang et al. 2018).

Another important concern in the generation of PDXs is the loss of the intrinsic heterogeneity that characterizes human cancers. Xenotransplantation affects the clonal architecture of the tumor in two important ways: (i) serial passaging causes stochastic genetic drift by the sequential bottle-necks and expansions that take place and (ii) adaptive evolutionary forces favor the expansion of clones with increased fitness in the mouse environment. It has been observed that clonal selection tends to be stronger at the initial establishment passage and remains quite low afterwards (Bruna et al. 2016; Eirew et al. 2015). However, sometimes a minor clone presents a gain of fitness in the mouse-specific environment and expands over passaging to become dominant (Eirew et al. 2015), leading to a divergent evolutionary trajectory between the same tumor grown in human and mice (Villacorta-Martin, Craig, and Villanueva 2017).

Despite the aforementioned changes in polyclonal dynamics, PDXs still preserve a substantial degree of the intratumoral heterogeneity seen in primary lesions (Bruna et al. 2016; Izumchenko et al. 2017). Moreover, it is very likely that with future refinements in their generation (i.e humanized mice), next-generation PDXs better preserve the human microenvironment and the selective forces that maintain clonal architecture.

In summary, PDXs are a reliable pre-clinical tool for pharmacogenomic studies and represent a more accurate approach than the use of cell lines to predict clinical response to anticancer therapies (Byrne et al. 2017). Indeed, in 2016 the US National Cancer Institute decided to stop screening anti-cancer drugs using the NCI-60 panel to focus on newer PDX models (Ledford 2016), and it is possible that other cancer research institutions follow the same trend.

Collaborative efforts engaging both public and private institutions are already flourishing (Conte et al. 2019) to overcome the logistic and financial barriers that the generation of large panels of PDXs could represent for most individual laboratories. Altogether, I believe that PDXs hold promise for improving cancer drug development success rates.

The Thesis into context

At the time I started my thesis, the TCGA consortium had recently released the molecular profiles of over 8,000 patients representing 27 tumor types (Weinstein et al. 2013). This tsunami of data changed the understanding of the genomic basis of human cancer, revealing extensive molecular heterogeneity within and across cancer types.

The tip of the iceberg was perhaps the observed inter-individual variability in drug response, even within subgroups of patients that had been selected on the basis of known biomarkers. A paradigmatic example of this was the disappointing result of the SHIVA clinical trial, which failed to demonstrate the superiority of targeted therapies with respect to the standard of care (Le Tourneau et al. 2015). Most biomarker-driven clinical trials are stratifying patients on the basis single-gene biomarkers, usually affecting the direct target of the test regimen. We realized that this approaches is too simple to account for the complexity of cancer and motivated us to explore the predictive power of driver alteration combinations.

Objectives

The general objective of this thesis was to develop computational methods to analyze, visualize, and contextualize the findings of prospective molecular profiling of cancer patients. In the following chapters we describe how we applied our toolkit to extract genotype-phenotype associations related to drug response and oncogenic pathway signaling:

(1) Starting from the ’precision oncology’ paradigm, we developed Cancer PanorOmics to contextualize each driver alteration identified in a personal cancer genome.

(2) We then explored the OncoGenomic Landscape of human cancer to unveil the structure of clinical cohorts of patients on the basis of combinations of driver alterations.

(3) With Targeted Cancer Therapy For You (TCT4U), we brought cancer therapies to the picture, and exploited driver co-occurrences to predict treatment outcomes in PDXs and also in patients.

(4) Finally, I present the preliminary results of a ’pharmacogenomics’ approach that we developed to determine: (i) which pathways are driving cancer progression in a given patient based on gene expression, and (ii) which compounds are likely to revert this endophenotype.

1

A PanorOmic View of Personal Cancer Genomes

AUTHORS: Lidia Mateo, Oriol Guitart, Carles Pons, Miquel Duran-Frigola, Roberto Mosca and Patrick Aloy.

PUBLICATION: Nucleic Acids Res. 2017 Jul 3; 45(Web Server issue): W195–W200.

Abstract

The massive molecular profiling of thousands of cancer patients has led to the identification of many tumor type specific driver genes. However, only a few (or none) of them are present in each individual tumor and, to enable precision oncology, we need to interpret the alterations found in a single patient. Cancer PanorOmics (http://panoromics.irbbarce- lona.org) is a web-based resource to contextualize genomic variations detected in a personal cancer genome within the body of clinical and scientific evidence available for 26 tumor types, offering complementary cohort- and patient-centric views. Additionally, it explores the cellular environment of mutations by mapping them on the human interactome and providing quasi-atomic structural details, whenever available. This 'PanorOmic' molecular view of individual tumors, together with the appropriate genetic counselling and medical advice, should contribute to the identification of actionable alterations ultimately guiding the clinical decision-making process.

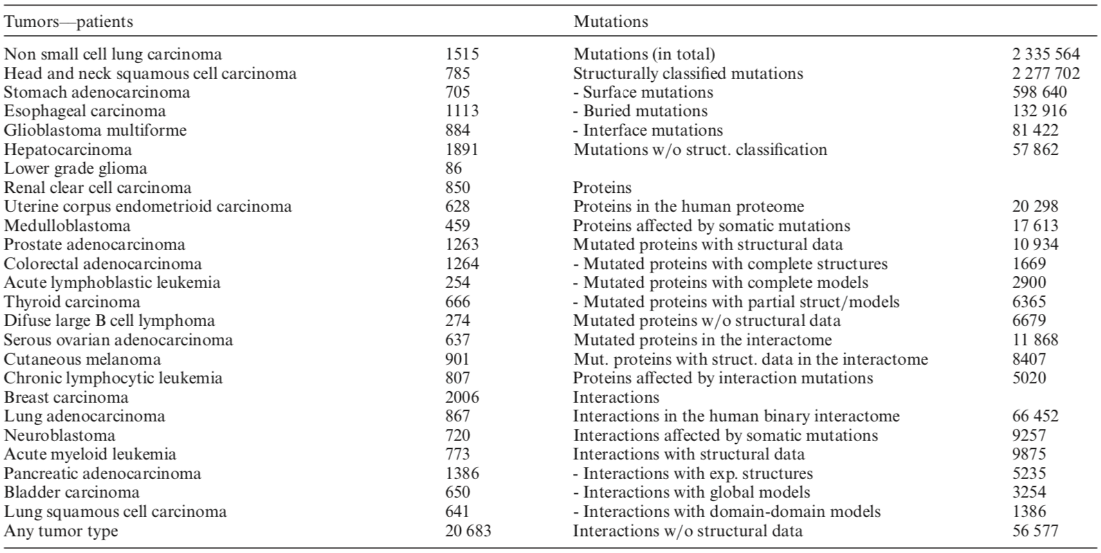
Introduction

Large-scale cancer genomics studies (Weinstein et al. 2013; International Cancer Genome et al. 2010) have shown that every personal cancer genome harbours thousands of genomic alterations –including somatic mutations, copy number alterations, gene expression changes and epigenetic modifications – that are not present in the patient's germline. Of these, a very small subset might be "driver" mutations, which confer a selective growth advantage through the enhancement of some biological capabilities like proliferation, angiogenesis, invasion or metastasis, or through the evasion of homeostatic mechanisms such as growth suppression, cell cycle arrest, cell death or immune surveillance (Hanahan and Weinberg 2011). However, the vast majority of genomic changes detected in cancer cells are just "passenger" and do not confer oncogenic properties (Vogelstein and Kinzler 2015). Distinguishing between driver and passenger mutations is a challenging task that has now become a matter of major interest, especially if we envision that the analysis of personal cancer genomes will become a common clinical practice. Current methods are based on the identification of signatures of positive selection at gene level that have been observed in large cohorts (i.e identification of genes showing high somatic mutation rate, mutation functional impact bias, and/or mutational clustering patterns that are more often observed in cancer than expected by chance) (Tokheim et al. 2016). These statistical approaches have permitted the identification of, for instance, over 200 driver genes for cutaneous melanoma or breast cancer (Rubio-Perez et al. 2015). However, when focusing on individual tumors, it is difficult to find more than 2-4 driver genes mutated and, in some patients, none of the usual suspects seems to be altered (Rubio-Perez et al. 2015). Moreover, the effect of distinct mutations in the same gene might be radically different (Mosca, Tenorio-Laranga, et al. 2015; Porta-Pardo, Hrabe, and Godzik 2015; Zhong et al. 2009), being it necessary to consider their exact location and molecular environment. Cancer PanorOmics integrates and displays high-throughput cancer sequencing analyses, together with data obtained from individual patients. The user can upload a list of somatic mutations, copy number alterations and/or gene expression changes detected in a personal cancer genome, and choose a reference cohort among the 26 tumor types available. Then, these genomic alterations can be explored in the light of what is known about the reference cohort. Additionally, the server maps mutations on the high-resolution 3D structure of proteins and protein-protein interactions to provide a molecular context to the genomic alterations analysed. Cancer PanorOmics is available at http://panoromics.irbbarcelona.org.

Data

The current version of Cancer PanorOmics catalogues 2,335,564 mutations found in 17,613 gene products from 20,683 cancer patients, representing 26 tumor types. These data have been compiled from IntOGen (Gonzalez-Perez et al. 2013) and COSMIC (Forbes et al. 2017) databases, and complemented with interactions and structural information from Interactome3D (Mosca, Ceol, and Aloy 2013) and dSysMap (Mosca, Tenorio-Laranga, et al. 2015). Cancer PanorOmics will be updated every 6 months. Please, see Table 1, or the Stats page at the server, for more detailed information.

Table 1. Cancer PanorOmics database content



Tumors—patients		Mutations	
Non small cell lung carcinoma	1515	Mutations (in total)	2 335 564
Head and neck squamous cell carcinoma	785	Structurally classified mutations	2 277 702
Stomach adenocarcinoma	705	- Surface mutations	598 640
Esophageal carcinoma	1113	- Buried mutations	132 916
Glioblastoma multiforme	884	- Interface mutations	81 422
Hepatocarcinoma	1891	Mutations w/o struct. classification	57 862
Lower grade glioma	86		
Renal clear cell carcinoma	850	Proteins	
Uterine corpus endometrioid carcinoma	628	Proteins in the human proteome	20 298
Medulloblastoma	459	Proteins affected by somatic mutations	17 613
Prostate adenocarcinoma	1263	Mutated proteins with structural data	10 934
Colorectal adenocarcinoma	1264	- Mutated proteins with complete structures	1669
Acute lymphoblastic leukemia	254	- Mutated proteins with complete models	2900
Thyroid carcinoma	666	- Mutated proteins with partial struct/models	6365
Difuse large B cell lymphoma	274	Mutated proteins w/o structural data	6679
Serous ovarian adenocarcinoma	637	Mutated proteins in the interactome	11 868
Cutaneous melanoma	901	Mut. proteins with struct. data in the interactome	8407
Chronic lymphocytic leukemia	807	Proteins affected by interaction mutations	5020
Breast carcinoma	2006	Interactions	
Lung adenocarcinoma	867	Interactions in the human binary interactome	66 452
Neuroblastoma	720	Interactions affected by somatic mutations	9257
Acute myeloid leukemia	773	Interactions with structural data	9875
Pancreatic adenocarcinoma	1386	- Interactions with exp. structures	5235
Bladder carcinoma	650	- Interactions with global models	3254
Lung squamous cell carcinoma	641	- Interactions with domain-domain models	1386
Any tumor type	20 683	Interactions w/o structural data	56 577

Workflow

User Input

The user can upload a list of somatic mutations, copy number and/or gene expression alterations detected in a personal cancer genome. Before submitting the query, the user should also choose one of the 26 tumor types available as a reference cohort. The server provides several examples to show the accepted formats of the input data.

Clinical Context

Cancer PanorOmics offers a contextualized view of the genomic alterations uploaded by the user within the available knowledge for the selected tumor type. More specifically, the user can easily see which altered genes in a given patient are known drivers, and how often each protein is mutated or copy number altered in the reference cohort. The results are displayed in two complementary Patient and Cohort Centric Views, as shown in Figure 1.

Patient Centric View. Here, all proteins altered in the patient are sorted by chromosomal coordinates and displayed in an interactive circular layout. The outermost ring, in gray, indicates the chromosomal location. Somatic mutations are represented as green cells in the inner data track, copy number and gene expression alterations are represented as red or blue cells in the second and third data tracks, respectively. Black circles are used to map know driver genes on the patient's alterations (filled circles indicate drivers specific for the selected tumor type, whereas empty circles indicate drivers in any other tumor type).

Cohort Centric View. This layout shows all known drivers for the reference tumor type, according to IntOGen (Gonzalez-Perez et al. 2013). The color scales indicate how often a protein is affected by each type of genomic alteration in the cohort. Somatic mutation frequency is represented in green in the inner data track, whereas copy number and gene expression alteration frequencies are represented in a red-white-blue scale in the second and third data tracks. Black circles are used to map the patient's genomic alterations on know driver genes.

Molecular Context

Cancer PanorOmics maps mutations on the high-resolution 3D structure of proteins and interactions. When the user selects a protein in the Patient Centric View or the Cohort Centric View, a new page is loaded to display the molecular context of the corresponding genomic alterations. *Protein-protein*

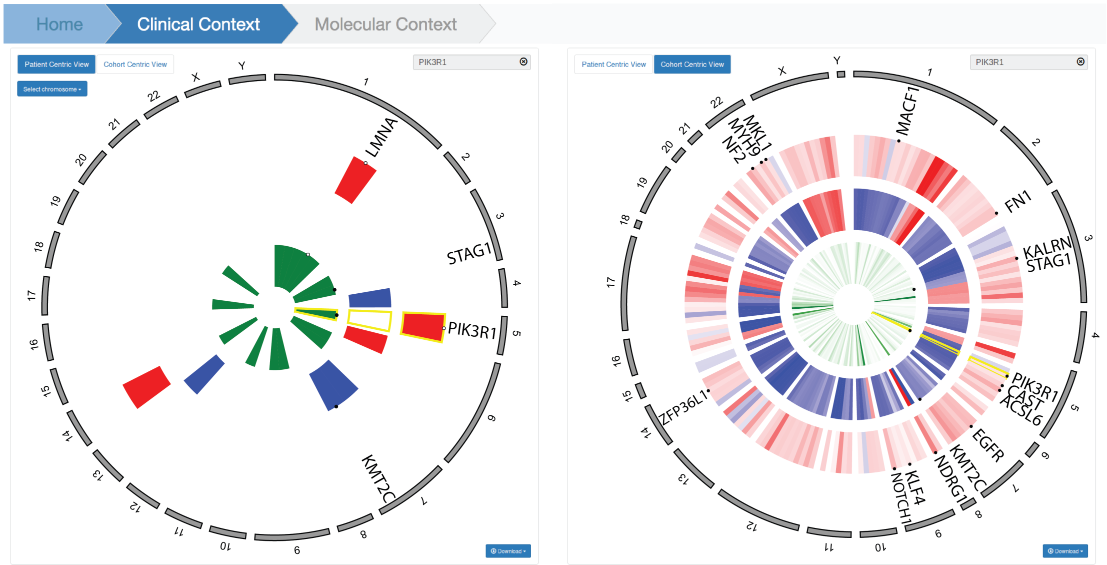
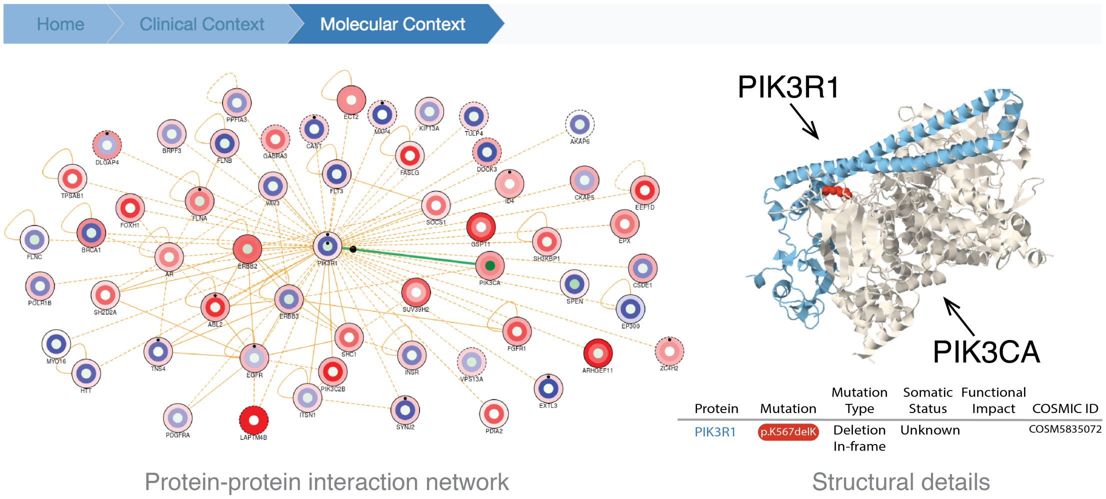


Figure 1.1: Clinical Context of the alterations detected in a ductolobular breast carcinoma patient (TCGA-A2-A4RX-01). The outermost rings represent the chromosomal location of genes and the next three data tracks show information about gene expression, copy number variation or somatic mutation in each gene. In the Patient Centric View (left panel), all somatic mutations and/or copy number alterations detected in this patient are displayed. Additionally, known drivers for the selected tumor type (filled circles), together with other cancer drivers (empty circles), are mapped onto patient alterations. In the Cohort Centric View (right panel), all known tumor type specific drivers are displayed and the frequency of each driver alteration is represented using a color gradient (green represents somatic mutation frequency, whereas red/blue represent copy number gain/loss or gene over/underexpression frequency). The alterations detected in this patient (filled circles) are mapped onto the known drivers. PIK3R1 gene is highlighted in yellow.



Protein	Mutation	Mutation Type	Somatic Status	Functional Impact	COSMIC ID
PIK3R1	p.K567delK	Deletion In-frame	Unknown		COSM5835072

Protein-protein interaction network

Structural details

Figure 1.2: Molecular Context of PIK3R1$^{\mathrm{K567delK}}$ and PIK3R1$^{\mathrm{overexpr.}}$ alterations detected in a ductolobular breast carcinoma patient (TCGA-A2-A4RX-01). The protein-protein interaction network shows the 50 most frequently altered interactors of PIK3R1. Each node describes a protein with three concentric circles informing about the frequency of each type of alteration: the inner green circle represents somatic mutation frequency, the middle red/blue circle represents copy number gain/loss frequency and the outer red/blue circle represents over/underexpression frequency in the reference cohort. The edges represent physical interactions between proteins. Genomic alterations detected in this patient are represented by small black dots. Genomic alterations can affect one or more molecular layers of a protein. Somatic mutations can also affect the protein-protein interaction interface between two proteins. Nodes or edges represented with a solid line in the network have structural information available. The structural details show the two proteins involved in the selected interaction: PIK3R1 (blue) and PIK3CA (gray). The precise location of the aminoacid affected by the somatic mutation PIK3R1$^{\mathrm{K567delK}}$ detected in this patient is highlighted in red.

interaction network. The current human interactome contains 66,452 interactions (Mosca, Ceol, and Aloy 2013). We zoom in to the protein of interest and its neighbors, and represent on each of them the recurrence of genomic alterations in the reference cohort. Small black circles indicate the presence of a genomic alteration either on a protein or at the interaction interface between two proteins. Every node contains three concentric circles displaying the information available for three molecular data types: somatic mutations, copy number and gene expression variations. Additional details about the known or potential role in cancer of each of the proteins in the network are shown in a table.

*Structural details.*Whenever available, structural 3D information is shown in an interactive panel. The mutations selected by the user are highlighted with a sphere representation and classified as 'buried', 'surface' or 'interface', based on their location in the 3D structure. We provide 3D structural details for over 2 million point mutations mapped onto 10,934 proteins, representing more than 95% of the cancer point mutations currently catalogued. Of these, 81,422 mutations lay at protein-protein interaction interfaces (i.e. edgetic) and are thus likely to affect the connectivity of the human interactome (Mosca, Tenorio-Laranga, et al. 2015; Zhong et al. 2009). Below the structural representation, we report information about the genomic alterations detected in the selected proteins, such as the mutation type, somatic status and FATHMM (Shihab et al. 2015) predicted functional impact for those mutations contained in COSMIC, as well as a brief description of the protein function or the method used to identify the selected interaction. We also list the mutations detected in the reference cohort that map onto the same protein or protein-protein interaction interface, so that the user can identify highly mutated protein domains or regions.

Use Case

To illustrate the applicability of Cancer PanorOmics, we studied the genomic alterations detected in a 67 years old woman diagnosed with ductolobular breast cancinoma (TCGA-A2-A4RX-01). The molecular profile of this patient consisted of 21 somatic mutations, 5 copy number alterations and 1,120 gene expression changes.

The Patient Centric View (left panel in Figure 1) showed that six genomic alterations affect either known breast cancer driver genes (PIK3R1^{K567delK}, STAG1^{E449Q}, KMT2CCN^{loss}, PIK3R1$^{overexpr.}$) or a driver gene that is not annotated to any specific tumor type (LMNA^{G567V}, LMNA$^{overexpr.}$). The Cohort Centric View (right panel in Figure 1) indicates that STAG1 and PIK3R1 are only rarely mutated in this reference cohort (mutated in less than 2% of patients). It also shows that KMT2C is more frequently altered (deleted in 12% of patients). Many other gene expression alterations appear in this view, being the upregulation of NDRG1 the most recurrent one (more than 12% of patients). We next explored the molecular context of the PIK3R1^{K567delK} mutation. The network view (left panel in Figure 2) shows that 11 of the direct interactors of PIK3R1 exhibit gene expression changes (e.g EGFR and ABL2 upregulation, which occur in more than 5% of breast cancer patients). Moreover, PIK3R1^{K567delK} lays at the interaction interface with PIK3CA (right panel in Figure 2), one of the most frequently mutated breast cancer drivers, likely triggering and edgetic perturbation. Indeed, we observed that patients with putative PIK3R1 edgetic mutations have an overall survival curve that is more similar to that of patients with edgetic mutations in the PIK3CA side of the interaction than patients with non-edgetic PIK3R1 mutations (Figure 3). Although experimental validation is needed to know whether this mutation is really disrupting the PIK3R1-PIK3CA interaction, and whether this might have an impact on the oncogenic potential, the survival analysis indicates that the distinction between edgetic and non-edgetic mutations might have a prognostic value and deserves further investigation.

Concluding Remarks

In the era of personalized medicine, we can now generate comprehensive and accurate portraits of individual tumors by putting together multiple layers of molecular profiles, such as somatic mutations,

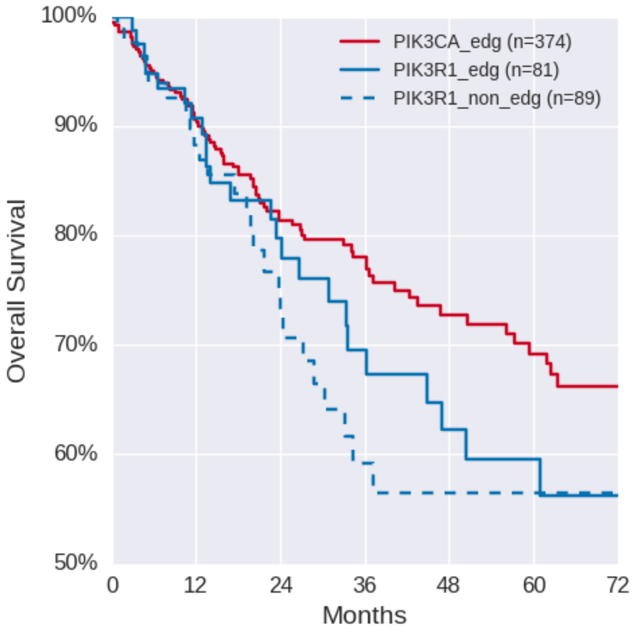


Figure 1.3: Kaplan-Meyer estimate of probability of overall survival (OS) in patients with non-edgetic PIK3R1 mutations and patients with putative PIK3R1-PIK3CA edgetic mutations at both sides of the interaction interface. The estimated 3-year survival probability of patients with putative PIK3R1-PIK3CA edgetic mutations is 69.58% (PIK3R1 interface, n=81) and 78.02% (PIK3CA interface, n=374), which are considerably higher than the 59.21% (n=89) observed in patients with non-edgetic PIK3R1 mutations. The survival analysis has been generated with somatic mutations and survival data of 4,795 cancer patients obtained from cBioPortal (Cerami et al. 2012; J. Gao et al. 2013)

DNA copy-number alterations or mRNA expression. Computational solutions to large-scale data integration and visualization are needed to facilitate the interpretation of such complex data. Cancer PanorOmics has a significant added value with respect to other resources separately, since it switches between patient- and cohort-centric perspectives, contextualizing the alterations of interest, and offering a seamless systemic (network) view annotated with 3D structures. In addition, no programming skills are needed to submit an integrated query and visualize the results with an enhanced and interactive display. We believe that servers like this one will help bridging the gap between cancer genomics and clinical oncologists, and play a central role in future personalized cancer management.

Server Information

Cancer PanorOmics runs on Apache HTTP Server, and it is written in PHP. After the automated validation by the server, the submitted mutations are stored into a PostgreSQL database and a user id is generated to identify access to the data. The server frontend is designed using the Bootstrap css framework. The interactive display of the mutation data relies on BioCircos.js (Cui et al. 2016) and Cytoscape.js (Franz et al. 2016) javascript libraries, which were modified to fulfill the visualization needs. D3.js (https://d3js.org/) javascript library is also used to display legend information, and Jsmol (http://www.jmol.org/) to provide interactive visualisation of protein 3D structures.

2

Exploring the OncoGenomic Landscape of cancer

AUTHORS: Lidia Mateo, Oriol Guitart, Miquel Duran-Frigola and Patrick Aloy.

PUBLICATION: Genome Med. 2018 Aug 3; 10: 61. doi: 10.1186/s13073-018-0571-0

Abstract

The widespread incorporation of next-generation sequencing into clinical oncology has yielded an unprecedented amount of molecular data from thousands of patients. A main current challenge is to find out reliable ways to extrapolate results from one group of patients to another, and to bring rationale to individual cases in the light of what is known from the cohorts.

OncoGenomic Landscapes displays thousands of cancer genomic profiles in a 2D space. Our tool allows to rapidly assess the heterogeneity of large cohorts, enabling the comparison to other groups of patients, and using driver genes as landmarks to aid in the interpretation of the landscapes. We also offer the possibility of mapping new samples and cohorts onto 22 predefined landscapes related to cancer cell line panels, organoids, patient-derived xenografts and clinical tumor samples.

Contextualizing individual subjects in a more general landscape of human cancer is, we believe, a valuable aid for basic researchers and clinical oncologists trying to identify treatment opportunities.

Introduction

The widespread incorporation of next-generation sequencing into clinical oncology has yielded an unprecedented amount of molecular data from thousands of patients, holding promise for a healthcare revolution (Biankin 2017; Gerstung et al. 2017). One of the current challenges is to find out reliable ways to extrapolate results from one group of patients to another, and to bring rationale to individual patients in the light of what is known from the cohorts. In this context, visualization tools that enable the exploration and analysis of large genomic datasets become essential for efficient interpretation and effective communication. Conventional strategies often represent dysregulated genes in a cohort as a matrix, with samples as columns and genes as rows, sorted according to the frequency of the genomic alterations (Kandoth et al. 2013; Hoadley, Yau, Wolf, et al. 2014; GENIE Consortium 2017; Bailey et al. 2018). Although this representation is useful to identify the main driver genes and to find recurrent patterns, they miss the capability of capturing the global structure of a cohort of patients or the comparison to other cohorts. Other approaches are more focused on exploiting population structure patterns based genomic profile similarities computed considering the whole genome or transcriptome (Newton et al. 2017; Nicolau, Levine, and Carlsson 2011; Prokopenko et al. 2016). The representations generated by those methods are difficult to interpret from a biological point of view, since most of the genomic alterations considered are of unknown functional impact. Furthermore, except for the TumorMap (Newton et al. 2017), the available tools do not offer a means to locate individual patient data within the cohort as a whole. In this context, as a complementary approach, we have developed a visualization tool that is mainly focused on the global characterization of cohorts but that only considers driver alterations with known or predicted functional impact on oncogenesis, yielding a global picture of a cohort that is biologically interpretable.

Implementation

Dataset summary

We collected 16,508 genomic profiles (coding somatic mutations and copy number variants) that are representatives of several cohorts of patients and cancer models (patient-derived xenografts, organoids and cell lines). We considered the 92.15% of samples having a putatively oncogenic alteration in one or more genes covered by the IMPACT410 gene panel (see Table 1). If solicited, future updates can easily incorporate larger patient cohorts, such as the complete TCGA (Weinstein et al. 2013) and ICGC (International Cancer Genome et al. 2010) sets, and whole exome sequencing data, to complement the 410 genes included in the IMPACT panel (Cheng et al. 2015).

Variant filtering

In order to filter out as many passenger alterations as possible, we applied a strict filtering pipeline described below, which was slightly tailored to each dataset:

TCGA patients. We downloaded the Catalog of Driver Mutations 2016.5, a curated dataset of known and predicted oncogenic coding mutations identified after analyzing 6,792 exomes of a pan-cancer cohort of 28 tumor types (Rubio-Perez et al. 2015). We could complement this information with copy number variation data (Cerami et al. 2012; J. Gao et al. 2013) for 4,058 patients, representing 16 tumor types. In addition to the known and predicted oncogenic coding mutations, we also considered as oncogenic the deletion (GISTIC score $\leq$ -2) of tumor suppressor genes and the amplification (GISTIC score $\geq$ 2) of oncogenes. The role of driver genes was established by inspecting the Catalog of Cancer Genes (Tamborero et al. 2018).

MSKCC patients. We obtained both protein coding mutations and copy number variants from the MSK_IMPACT Clinical Sequencing Cohort (Zehir et al. 2017) through cBioPortal (Cerami et al. 2012; J. Gao et al. 2013). Genes with a copy number alteration score $\leq$ -2 or $\geq$ 2 were considered as putative deletions or amplifications, respectively.

Novartis PDXs. We collected the 375 PDXs for which both mutations and copy number alterations were available (H. Gao et al. 2015). After analyzing the probability distribution of the estimated absolute copy number per gene, we considered absolute copy numbers below 1 or above 4 as gene deletions or amplifications, respectively. Using these criteria, we observed significant differences in gene expression between deleted tumor suppressors and amplified oncogenes (Supplementary Figure 1A), confirming that those thresholds are biologically relevant.

GDSC cell lines. We used gene level copy number data reported in the GDSC (W. Yang et al. 2013), which is based on PICNIC analysis of Affymetrix SNP6.0 arrays. We considered genes with a minimum copy number of any genomic segment mapping to that gene below 1 or above 6 as gene

Table 2.1: Summary of sample size and provenance.

	Biological Source	No. of samples with SNV and CNV	No. of samples with driver alt. in whole exome	No. of samples with driver alt. in IMPACT410
TCGA	patients	4,058	3,935	3,850
MSK-IMPACT	patients	10,945	-	9,869
Novartis PDXs	PDXs	375	375	375
OncoTrack	patients	117	109	109
	PDXs	59	59	59
	organoids	46	46	46
GDSC Cell Lines	cell lines	908	904	904
TOTAL	-	16,508	5,428	15,212

deletions or amplifications, respectively. Using those thresholds, we observed significant differences in gene expression between deleted tumor suppressors and amplified oncogenes (Supplementary Figure 1B), as described above for the analysis of copy number variants in PDXs.

OncoTrack (Schutte et al. 2017). We downloaded the genomic profiles of a biobank of 106 tumors, 35 organoids and 59 xenografts. Copy number alterations were already annotated as ”Amplification” or ”Deletion”.

For MSK-IMPACT, Novartis PDXs, GDSC cell lines and OncoTrack datasets, protein coding somatic mutations (following HGVS nomenclature recommendations) and copy number variants were classified into predicted passenger or known/predicted oncogenic alterations using the cancer genome interpreter resource (Tamborero et al. 2018).

After filtering out putative passenger alterations, we subsampled the dataset to consider only oncogenic alterations covered by the IMPACT410 gene panel (Cheng et al. 2015), which provided a much larger reference cohort ($>$ 10,000 patients MSKCC (Zehir et al. 2017)) while retaining enough signal to build meaningful OncoGenomic Landscapes.

2D projections

We built a Boolean matrix encoding the oncogenic alterations identified in each sample (in rows) and driver gene (in columns). We then calculated the Jaccard distance between all pairs of unique samples and used the resulting distance matrix as input for a metric Multi-Dimensional Scaling (MDS), carried out using the scikit-learn implementation of MDS (Pedregosa et al. 2011) with default parameters (2 components, 4 SMACOF initializations and a maximum of 300 iterations per run). As a result, we obtained (x,y)-coordinates for each of the samples (i.e. a 2D projection). The corresponding level plots were generated by the 2D kernel density estimate function of the seaborn library, using 20 levels and a grayscale color-map as background. The PanCancer and more specific landscapes are the result of applying this procedure to the whole dataset and sample subsets, respectively.

To benchmark both the distance metric and the dimensionality reduction strategy used to generate the landscapes, we examined whether the organization of samples in the Pan-Cancer Landscape reflects the tissue-of-origin of the tumor. We observed a significant clustering of samples based on tissue-of-origin when examining both the Jaccard similarity coefficient in the multidimensional space and the Euclidean proximity in the MDS space. To evaluate the robustness of the current strategy, we also assessed the clustering of samples when using a Kernel PCA projection, an approach previously used in the field (Prokopenko et al. 2016). Moreover, we observed that the MDS projection yields greater spatial resolution compared to Kernel PCA and that the proximity in the MDS space has a stronger correlation with the proximity in the multidimensional space (Supplementary Figure 2).

When new samples are to be mapped onto a given landscape, we approximate their location by a nearest neighbor search in the original multidimensional space of genomic alterations (i.e. Jaccard distance). A new sample is assigned the (x,y)-coordinate of its nearest neighbor, and the distance between them serves as a confidence score of the mapping. We found this simple strategy to be sufficient, as it yields an error comparable to the intrinsic one of SMACOF MDS (Supplementary Figure 3).

Cohort overlays

In order to highlight the territory occupied by a subset of samples, we obtained the (x,y)-coordinates of the selected samples in a given landscape and generated a 2D kernel density estimate with the kdeplot function using 20 levels, a transparent background, and contours colored using a color-map that represents probability density as heat.

Driver landmark overlays

Similarly, to highlight the territory occupied by samples that have an oncogenic alteration in a given driver gene, we obtained the coordinates of those samples and generated a 2D kernel density estimate using 4 levels. We modified the resulting plots by removing the level with lowest density and setting the same color and transparency to the rest of levels.

Survival Analysis

We used the median distance to the 22 nearest PDXs, which correspond to 5% of the total 434 PDXs, as a measure of how far a patient is to the PDXs. Patients in the upper and lower quartiles of the median distance distribution were considered to be distal or proximal to PDXs, respectively. We compared the lifespans of patients that are proximal or distal to PDXs using the Kaplan-Meyer estimate of the survival function, and performed a log-rank test to assess the statistical significance of the observed difference using the lifelines library. Additionally, we investigated the effect of distance to PDXs on survival using Cox's proportional hazards regression model, adjusting for tumor type and patient provenance covariates.

Results

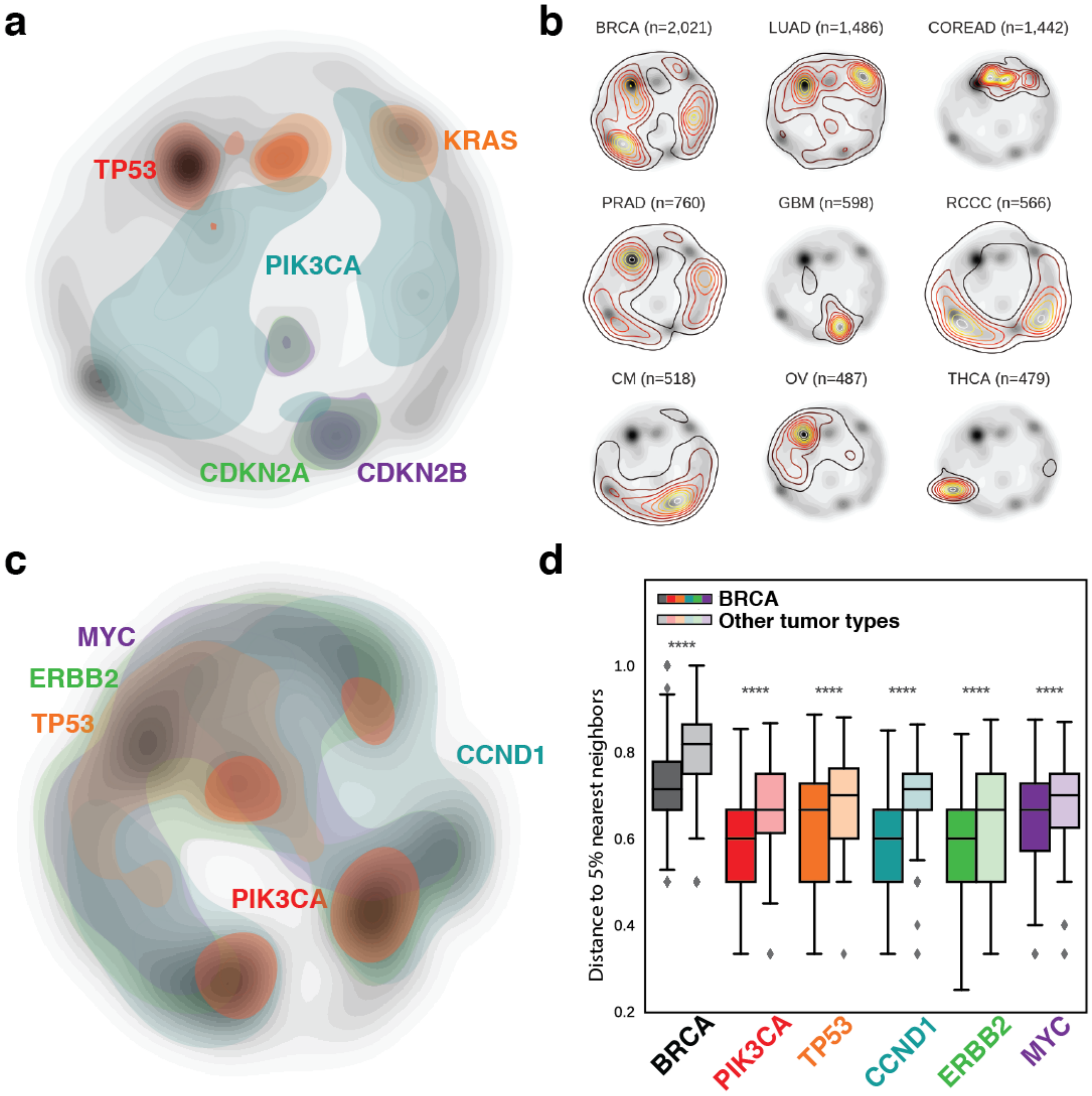
We have developed a visualization tool that is mainly focused on the global characterization of cancer cohorts. Our computational pipeline mines and integrates genomic profiles from 13,827 cancer patients and 1,385 cancer models (434 patient-derived xenografts, 46 organoids and 905 cell lines), compares pairs of samples based on shared oncogenic alterations, and plots the results in a 2D space that we called OncoGenomic Landscape. We offer our tool as a web-based interface that enables the comparison of the main cohorts published to date, as well as the possibility of mapping new samples or cohorts on any of the available landscapes. Below, we describe some test cases to illustrate the utility of our tool and we also provide a step-by-step tutorial on how to perform basic downstream analyses (available at https://oglandscapes.irbbarcelona.org/tutorial).

Figure 1 displays the distribution of samples across the PanCancer landscape, including 15,212 genomic profiles from different tissues (see Table 1). As expected, territories corresponding to recurrent drivers such as TP53 or KRAS are well populated (Figure 1a). Perhaps more interesting is the relatively large amount of patients that occupy a territory shared by TP53 and KRAS alterations, consistent with a significant co-occurrence observed in the MSK-IMPACT Clinical Sequencing Cohort (Zehir et al. 2017; J. Gao et al. 2013), and suggesting a synergistic effect between these alterations. It is also apparent that samples with alterations in CDKN2A and CDKN2B occupy almost identical regions, which agrees with the finding that these two tumor suppressors are usually co-deleted as they are encoded next to each other in a very small locus (Tu et al. 2018).

Beyond key gene alterations, the PanCancer landscape retains the tissue of origin of the tumors (Figure 1b). We can observe how certain tumor types (e.g. glioblastoma or colorectal adenocarcinoma) often present a limited set of driver mutations and are thus restricted to very specific areas in the map, while other types (e.g. breast cancer or prostate adenocarcinoma) show a much more diverse pattern of oncogenic alterations, and are widely spread. In both cases, it is possible to cluster cancer patients based on the tissue of origin of their tumor, and to identify dominant groups representing each tumor type (Supplementary Figure 2), as previously suggested for the 12 major cancer types (Hoadley, Yau, Wolf, et al. 2014; Kandoth et al. 2013) and, more recently, for the 33 cancer types that comprise the complete TCGA PanCancer Analysis (Bailey et al. 2018). Moreover, we can zoom in on a region that is specific for a certain tumor type and capture patterns that might otherwise be hidden in the broader PanCancer landscape (Figure 1c). For instance, despite their considerable heterogeneity, we see that breast cancer samples are closer to each other than to other tumor types (Figure 1d). The observed proximity cannot be only attributed to the presence of common driver genes, since we observe that tumor samples in different tissues sharing the most frequent driver alterations in breast cancer are significantly more distal. These results strongly suggest that our tumor type specific territories capture complex mutational signatures that cannot be attained by analyzing driver genes individually.

Figure 2.1 *(following page)*: Visual display of the Oncogenomic Landscape of cancer.(a) PanCancer landscape populated by 15,212 samples of 19 major tumor types of different biological origin (13,827 patients, 434 PDXs, 46 organoids, 905 cell lines). The territories occupied by samples that have at least one of the five most recurrent oncogenic alterations are shaded in different colors and serve as landmarks for molecular interpretation. (b) Distinct territories occupied by the nine most comprehensively characterized tumor types are depicted as transparent level plots overlaid on the PanCancer landscape background. BRCA: breast carcinoma, LUAD: lung adenocarcinoma, COREAD: colorectal adenocarcinoma, PRAD: prostate cancer, GBM: glioblastoma multiforme, RCCC: renal clear cell carcinoma, CM: cutaneous melanoma, OV: ovarian cancer and THCA: thyroid cancer. (c) The OncoGenomic Landscape of breast invasive carcinoma (BRCA) patients is shown to illustrate how each of the 19 tumor type specific landscapes are displayed in our web-server. Colors represent the territories occupied by samples having oncogenic alterations in five breast cancer specific landmark driver genes. (d) Boxplot showing the median distance of breast cancer samples to the 5% nearest neighbors in each comparison. The first two boxes compare the median distance of all breast cancer patients among themselves and to patients with other tumor types. The remaining pairs of boxes focus on patients that have an oncogenic alteration in each of the main five BRCA driver genes. Panels a, b and c are screenshots directly obtained from the webserver, although the color scheme and transparency of the landmark driver genes have been adapted to enhance their appearance in the printed figure. Panel d was generated after performing the statistical analysis outside of the app.

Figure 2.1: (continued)



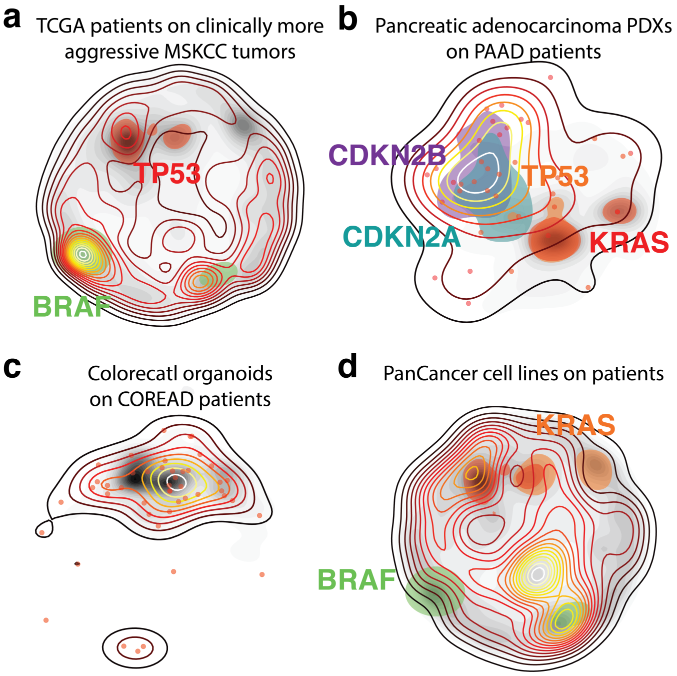


Figure 2.2: Overlay of different OncoGenomic Landscapes. (a) The cohort of primary tumors from TCGA (n=3,850) is displayed as a transparent level plot overlaid on a largest cohort of clinically aggressive tumors from MSKCC (n=9,869), represented as a background landscape in gray scale. In a similar way, (b) pancreatic adenocarcinoma PDXs are overlaid on a cohort of PAAD patients (n=377), (c) OncoTrack colorectal organoids (n=46) are overlaid on colorectal adenocarcinoma patients (n=1,141), and (d) a panel of 905 cell lines are overlaid on 13,827 PanCancer patients. Panels b and d are screenshots directly obtained from the webserver, although the color scheme and transparency of the landmark driver genes have been adapted to enhance their appearance in printed the figure. Panels a and d are comparisons of two PanCancer background overlays in which one of the two overlays in each comparison was represented as gray scale landscape to enable a more direct comparison in the static view offered by the printed version of the figure.

The accurate comparison of patient or cancer model cohorts is fundamental to evaluate their molecular diversity and, more importantly, to assess whether information such as treatment benefits or prognostic factors learned from a reference group can be safely transferred to a new cohort. For instance, by comparing primary resections of treatment naïve tumors (3,850 patients from The Cancer Genome Atlas (TCGA)) to 9,869 clinically aggressive tumors from the Memorial Sloan Kettering Cancer Center (MSKCC), we can readily see than alterations in TP53 are much more common in the MSKCC cohort than in TCGA, as recently reported (Zehir et al. 2017), while BRAF alterations show the opposite trend (Figure 2a). We believe that portrayals like this might also guide the design of clinical basket trials, where patients are selected based on their oncogenomic profiles regardless of their specific tumor type (Hyman et al. 2015).

We can also use OncoGenomic Landscapes to assess the molecular representativity of different model systems (cell lines, organoids or patient derived xenografts (PDXs)) with respect to a reference clinical cohort. For example, even though alterations in TP53, KRAS and CDKN2A are the most prevalent in pancreatic ductal adenocarcinoma patients (Tu et al. 2018), when we look at the tumors that successfully engrafted in mice (i.e. PDXs), we clearly see that CDKN2A-CDKN2B co-alterations are much more frequent in PDXs than it would be expected from clinical data (Figure 2b), supporting the idea that the simultaneous inactivation of CDKN2A and CDKN2B is required for the induction of pancreatic cancer in adult mice with overexpressed KRAS^{G12D} and loss of TP53 (Tu et al. 2018). Conversely, we observe that the small collection of 69 OncoTrack colorectal organoids (Schutte et al. 2017) spans the molecular diversity seen in a much larger cohort of COREAD patients (188 from TCGA and 953 from MSKCC) (Figure 2c). Finally, the overlay of 905 cancer cell lines (W. Yang et al. 2013) on top of patient samples reveals a lack of cell models to study the effects of KRAS and BRAF mutations alone (Figure 2d).

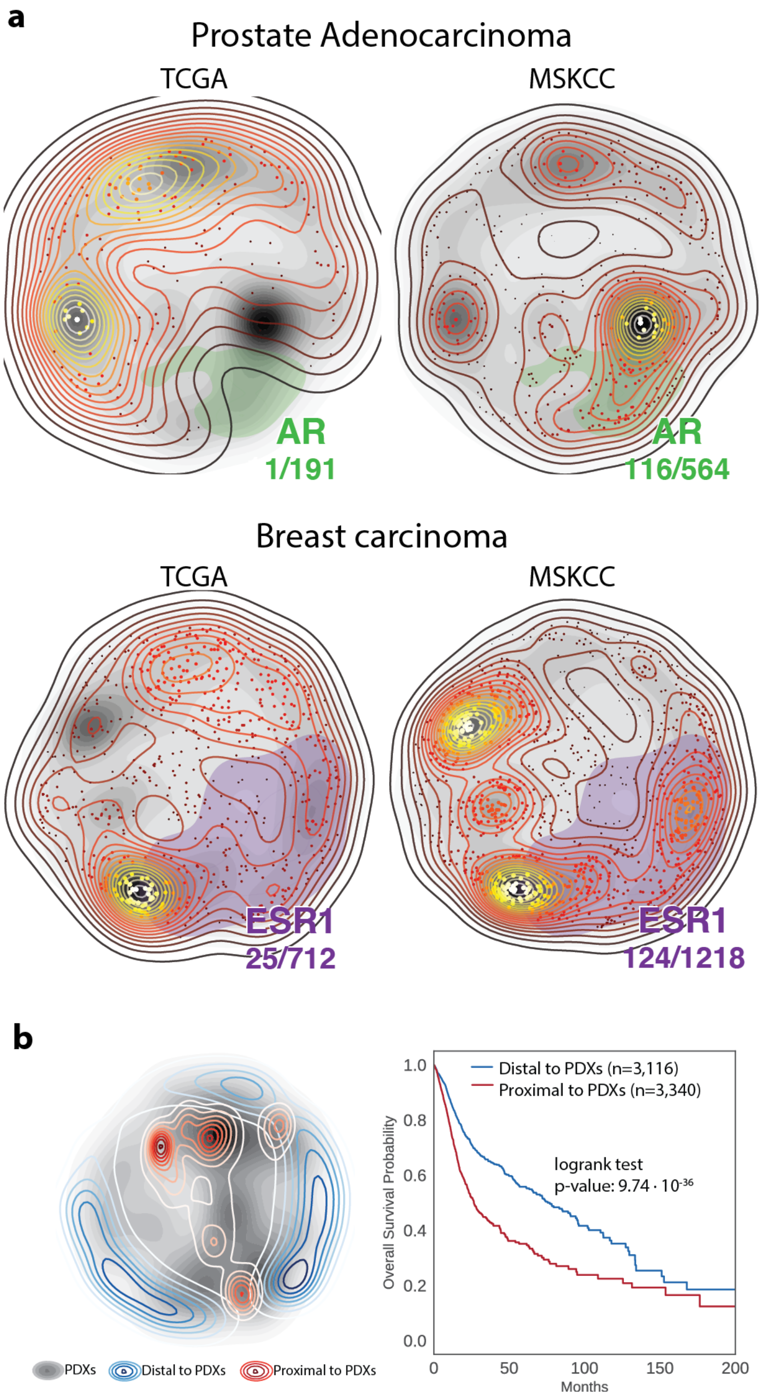
Interestingly, we also find that distances in OncoGenomic Landscapes correlate with relevant clinical features. Mutations in the Androgen Receptor (AR) in prostate and in Estrogen Receptor (ESR1) in breast cancer are related to acquired resistance to hormonal therapies. The density of patients with mutations in those genes is notoriously higher in MSKCC than in TCGA, consistent with the known clinico-pathological differences of those two cohorts (Figure 3a). We can also relate territories in the landscape to overall survival probabilities (Figure 3b). It is well documented that during the establishment of PDXs there is an engraftment bias towards more aggressive tumors (Pergolini et al. 2017; Whittle et al. 2015). Accordingly, we see that patients that are proximal to successfully-engrafted tumors show a significantly worse prognosis than patients that are distal to PDXs (p-value $9.74 \cdot 10^{-36}$), and the trend remains significant (Cox regression p-value $2.23 \cdot 10^{-12}$) after adjusting for possible confounding factors such as tumor type and patient provenance (TCGA or MSKCC). This observation is in line with the recent finding that pancreatic ductal adenocarcinoma patients whose tumors did engraft in mice had significantly shorter recurrence-free and overall survivals than patients whose tumors failed to engraft (Pergolini et al. 2017).

Concluding Remarks

In summary, OncoGenomic Landscapes is a web-based visualization tool that organizes tumor samples, and other cancer models, in a 2D space, enabling the comparison of large cohorts and capturing their molecular heterogeneity. We offer the possibility of mapping new samples and cohorts onto a set of 22 predefined landscapes, providing an intuitive means to visualize user's data and enrich it with knowledge transferred from the large corpus of cancer samples available today. Contextualizing individual patients in a more general landscape of human cancer is, we believe, a valuable aid for clinical oncologists trying to identify treatment opportunities, maybe in a compassionate use basis, for patients that ran out of standard therapeutic options.

Figure 2.3 *(following page)*: Clinical relevance of Oncogenomic Landscapes. (a) Differences between TCGA and MSKCC cohorts related to resistance to endocrine therapy in PRAD and BRCA. The fraction of patients in each cohort presenting alterations in the androgen receptor (AR) and the estrogen receptor (ESR1) are shown in green and magenta, respectively. (b) Patient distance to PDXs correlates with overall survival probability. The territories occupied by PDXs are shown as a background landscape in gray scale whereas the location of patients that are proximal (red) or distal (blue) to PDXs are shown as transparent level plots. Kaplan-Meyer analysis comparing the overall survival rate of patients that are proximal (red) or distal (blue) to PDXs. Panel a is composed of screenshots directly obtained from the webserver, although the color scheme and transparency of the landmark driver genes have been adapted to enhance their appearance in printed the figure. Panel b was generated outside the app following the steps described in the tutorial available at https://oglandscapes.irbbarcelona.org/tutorial.

Figure 2.3: (continued)



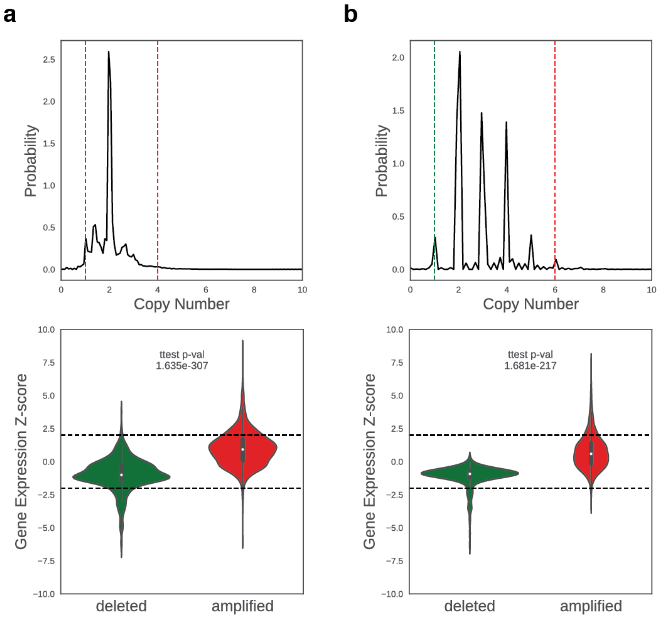


Figure S1: Copy number alterations in Novartis PDXs (a) and CDSC Cell Line panel (b). The top plots show the probability distribution of the estimated absolute copy number per gene in all the samples. Dashed lines indicate the thresholds that were chosen to call deletions (green) or amplifications (red). The bottom plots show the difference in gene expression between deleted tumor suppressors and amplified oncogenes in each dataset.

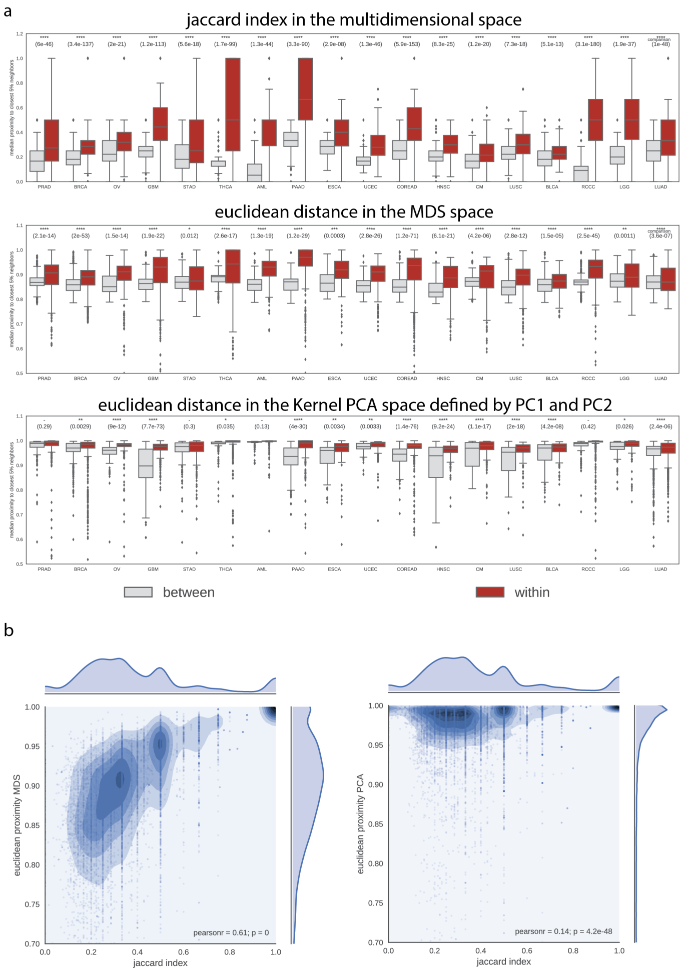


Figure S2: The organization of samples across the PanCancer landscape reflects the tissue-of-origin of the tumor. (a) Samples of a given tumor type (within) are closer to each other in the PanCancer landscape than to samples of other tumor types (between) when measuring the median Jaccard similarity coefficient with respect to the closest 5% neighbors in the multidimensional space. The clustering of samples based on tissue-of-origin is also significant when measuring the median Euclidean proximity to the closest 5% neighbors in the two-dimensional MDS space and in the space defined by the first two components of a Kernel Principal Component Analysis (Kernel PCA). (b) Correlation between the proximity of each sample to the closest 5% samples of the same tumor type in the MDS space (left) or in the Kernel PCA (right) and the proximity to the closest 5% samples of the same tumor type in the multidimensional space (Jaccard index).

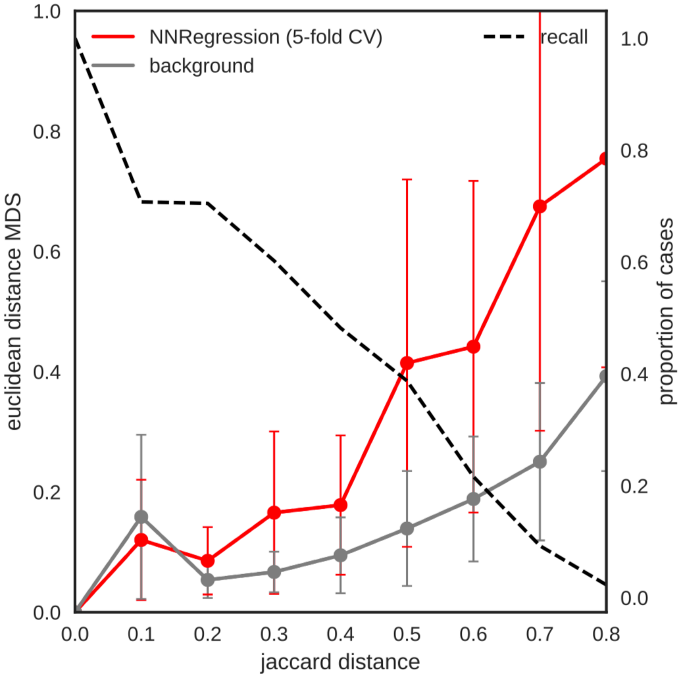


Figure S3: Comparison between the mapping errors associated to the nearest neighbor search to the intrinsic error of SMACOF MDS. The red line represent the performance of the Nearest Neighbor Regressor in a five-fold corss-validation exercise, in terms of Euclidean distance between the predicted and the observed coordinates in the MDS space with respect to the Jaccard distance between the test sample and its closest neighbor. The gray line represents the variablitiy in the SMACOF MDS projections of pairs of samples, measured as Euclidean distance in the MDS space, with respect to their Jaccard distance in the n-dimensional space. Solid lines represent the median distance, with error bars indicating its median absolute deviation. The dashed line represents the cumulative proportion of cases in the test set that are at a given distance or greater from its nearest neighbor.

3

Cancer Therapy Prioritization Based on Driver Alteration Co-occurrences

AUTHORS: Lidia Mateo, Miquel Duran-Frigola, Albert Gris-Oliver, Marta Palafox, Maurizio Scaltriti, Pedram Razavi, Sarat Chandarlapaty, Joaquin Arribas, Meritxell Bellet, Violeta Serra and Patrick Aloy

PUBLICATION: Preprint 10.1101/772673 posted on bioRxiv. Submitted to Genome Medicine.

Abstract

Molecular profiling of personal cancer genomes, and the identification of actionable vulnerabilities and drug-response biomarkers, are the basis of precision oncology. Tumors often present several driver alterations that might be connected by cross-talk and feedback mechanisms, making it difficult to mark single oncogenic variations as reliable predictors of therapeutic outcome.

In the current work, we uncover and exploit driver alteration co-occurrence patterns from a recently published in vivo screening in patient-derived xenografts (PDXs), including 187 tumors and 53 drugs. For each treatment, we compare the mutational profiles of sensitive and resistant PDXs to statistically define Driver Co-Occurrence (DCO) networks, which capture both genomic structure and putative oncogenic synergy. We then use the DCO networks to train classifiers that can prioritize, among the available options, the best possible treatment for each tumor based on its oncogenomic profile. In a cross-validation setting, our drug-response models are able to correctly predict 66% of sensitive and 77% of resistant drug-tumor pairs, being applicable to several tumor types and drug classes for which no biomarker has yet been described. Additionally, we experimentally validated the performance of our models on 15 new tumor samples engrafted in mice, achieving an overall accuracy of 75%.

Finally, we adapted our strategy to derive drug-response models from continuous clinical outcome measures, such as progression free survival, which better represent the data acquired during routine clinical practice and in clinical trials. We believe that the computational framework presented here could be incorporated into the design of adaptive clinical trials, revealing unexpected connections between oncogenic alterations and increasing the clinical impact of genomic profiling.

Introduction

In light of the complexity and molecular heterogeneity of tumors, clinical and histopathological evaluation of cancer patients is nowadays complemented with genomic information. Genome-guided therapy has been shown to improve patient outcome (Stockley et al. 2016; Schwaederle et al. 2015) and clinical trial success rate (Jardim et al. 2015) and, despite some controversy (Prasad 2016), prospective molecular profiling of personal cancer genomes has enabled the identification of an increasing number of actionable vulnerabilities (Chang et al. 2018).

Cancer genome sequencing initiatives have found that any given tumor contains from tens to thousands of mutations. However, only a few of them confer a growth advantage to cancer cells, driving thus the tumorigenic process. The most comprehensive study of 'driver' genes published to date has analyzed over 9,000 tumor samples, across 33 tissues of origin, and has systematically identified driver mutations in 258 genes (Bailey et al. 2018). Approximately half (142) of those driver genes were associated with a single tumor type, whereas 87 genes seem to provide a growth advantage in several tumor types. The number of drivers detected per tumor type varies widely, ranging from 2 in kidney chromophobe cancer to 55 in uterine cancer. Despite the large number of drivers identified per tumor type, every patient has a unique combination of mutations and copy number variants: ninety percent of patients show at least one putative driver alteration, but each sample only contains a median of three putative altered drivers(Rubio-Perez et al. 2015).

On top of identifying key alterations in tumor development, it is fundamental to pinpoint those that can shed light on the most appropriate therapy to treat each tumor (i.e. biomarkers). Often, patients with similar clinicopathological characteristics might be molecularly different (Bailey et al. 2018), this inter-patient heterogeneity is one of the reasons why only a subset of them will actually respond to a given targeted treatment. Computational studies suggest that up to 90% of patients may benefit from molecularly-guided therapy when biomarkers of uncertain clinical significance, as well as off-label and experimental drugs, are used to guide treatment selectionRubio- (Rubio-Perez et al. 2015; Senft et al. 2017).

Although randomized controlled trials are still considered the gold standard in the clinics, they cannot address all possible patient clinicopathologic and molecular subtypes (Das and Lo 2017). Precision medicine has prompted the reconsideration of clinical drug development pipelines, with the implementation of more sophisticated clinical trial designs, such as umbrella, basket, and platform trials to account for inter-patient heterogeneity (Simon 2017). In particular, the implementation of adaptive enrichment strategies allows for continual learning and modification of the eligibility criteria as data accumulate, with the objective of recruiting those patients that are most likely to benefit from treatment (Das and Lo 2017; Pallmann et al. 2018; Thorlund et al. 2018; Simon 2017).

However, despite the implementation of these novel experimental designs, currently only alterations in 25 genes have accumulated enough clinical evidence to be approved as biomarkers by the FDA (Chakravarty et al. 2017). Indeed, a recent comprehensive analysis of 6,729 pan-cancer tumors could only identify actionable mutations with therapeutic options available in clinical practice (FDA-approved or international guidelines) or reported in late phase (III−IV) clinical trials in 5.2% and 3.5% of the samples, respectively (Tamborero et al. 2018). These figures coincide with clinical trial enrolment rates (Stockley et al. 2016), where only 89 out of 1,640 of patients could enter genotype-matched treatment trials, the vast majority of which involved mutations in four genes, namely PIK3CA, KRAS, BRAF and EGFR. This highlights an acute need to expand the current repertoire of response biomarkers.

The eligibility criteria of most genomically-matched basket clinical trials are based on the single-gene biomarkers. However, most tumors do not present a single actionable mutation but have co-occurring driver alterations that might simultaneously alter key players of signaling pathways connected by cross-talk and feedback mechanisms (Jaeger, Duran-Frigola, and Aloy 2015; Sanchez-Vega et al. 2018). There are many documented cases of functionally relevant co-occurring oncogenic mutations, such as the concomitant inactivation of TP53 and RB1 (Huun, Lonning, and Knappskog 2017), co-deletion of CDKN2A and CDKN2B (Tu et al. 2018), co-amplification of MDM2 and CDK4 (Dembla et al. 2018; Laroche-Clary et al. 2017), 1p/19q co-deletion in glioma (Lauber, Klink, and Seifert 2018), MYC amplification and TP53 mutations (Ulz, Heitzer, and Speicher 2016) or activating alterations in KRAS and BRAF (Chang et al. 2018). At pathway level, the concomitant activation PI3K signaling pathway with FGF signaling (FGFR2 and FGFR3), or with NRF2 mediated oxidative response have also been identified in several tumor types(Sanchez-Vega et al. 2018). In this context, a single-gene based stratification of patients into subtypes and treatment arms might be over-simplistic, and novel frameworks that exploit co-mutational patterns might prove more effective.

As in the identification of driver mutations, the discovery of drug response biomarkers requires large numbers of patient molecular profiles matched to treatment outcomes. Unfortunately, treatment history information of large-scale genomics endeavors has not been systematically collected (e.g. TCGA (J. Liu et al. 2018)) or is not yet publicly available (e.g. GENIE Consortium (GENIE Consortium 2017)). Even though better data sharing policies are needed, many concerns are raised regarding privacy, property and the preliminary nature of confidential biomedical data. Safer alternative ways of sharing biomedical data are already on the table (Guinney and Saez-Rodriguez 2018) but, until the access to systematically annotated clinical records becomes a reality, the research community largely relies on drug response data gathered from pre-clinical models.

Cancer cell lines are the most widely used in vitro model system, and have been fundamental tools to set the grounds of our understanding of cancer biology and to assess the efficacy of a broad spectrum of cancer drugs (Iorio et al. 2016). Unfortunately, cancer cell lines have been cultured as monolayers on plastic surfaces, and in growth-promoting conditions, for decades. As a consequence, most of them have suffered a substantial transcriptional drift, and they likely represent a cell subpopulation from the original primary tumor (Gillet, Varma, and Gottesman 2013). Those facts have fueled the debate regarding how well cancer cell lines resemble the tumors from which they were established and to which extent they are clinically relevant (Jaeger, Duran-Frigola, and Aloy 2015; Gillet, Varma, and Gottesman 2013).

A more realistic model to bridge the bench-to-bedside gap is the patient-derived mouse xenograft (PDX) (H. Gao et al. 2015). To some extent, PDXs preserve inter- and intra-tumoral heterogeneity, and mimic the clinical course of the disease and response to targeted therapy, at least in certain tumor types (Einarsdottir et al. 2014; Bruna et al. 2016; Krepler et al. 2017). Indeed, a recent review reported a 91% (153 out of 167) correspondence between the clinical responses of patients and their cognate PDXs (Pompili et al. 2016). Although this data are more time-intensive and expensive to generate, it is still feasible to establish large in vivo screenings, covering a wide diversity of tumor types and drugs. PDXs are thus a clinically relevant platform for pre-clinical pharmacogenomic studies, and represent a more accurate approach to identify predictive biomarkers compared with the use of cancer cell lines (Byrne et al. 2017).

Here, we present a computational strategy to uncover and exploit driver alteration co-occurrence patterns in PDXs. By comparing the molecular profiles of sensitive and resistant PDXs to a given drug, we identify driver co-occurrence networks and use them as a new type of drug-response indicators, applicable much beyond known biomarkers. We apply our strategy to the largest panel of PDXs and drugs available to date (H. Gao et al. 2015), and prospectively validate our findings in vivo. Finally, we adapt our strategy to derive response predictive models directly from continuous clinical outcome measures, such as progression free survival, and evaluate them on a cohort of breast cancer patients.

Results and Discussion

Driver co-occurrence networks of drug response

Although thousands of genomic profiles of patient tumors are available, accurate information about pharmacological interventions and treatment outcome has not been systematically collected, or has not been disclosed yet. Thus, to bypass these limitations, we compiled drug response data obtained in PDXs, since they preserve the overall molecular profile of the original tumor, and maintain its cellular and histological structure (Pompili et al. 2016). In particular, we based our study on 375 PDXs for which somatic mutations and copy number alterations have been characterized, together with their response to 62 treatments across six indications, using the 'one animal per model per treatment (1x1x1)' experimental design (H. Gao et al. 2015).

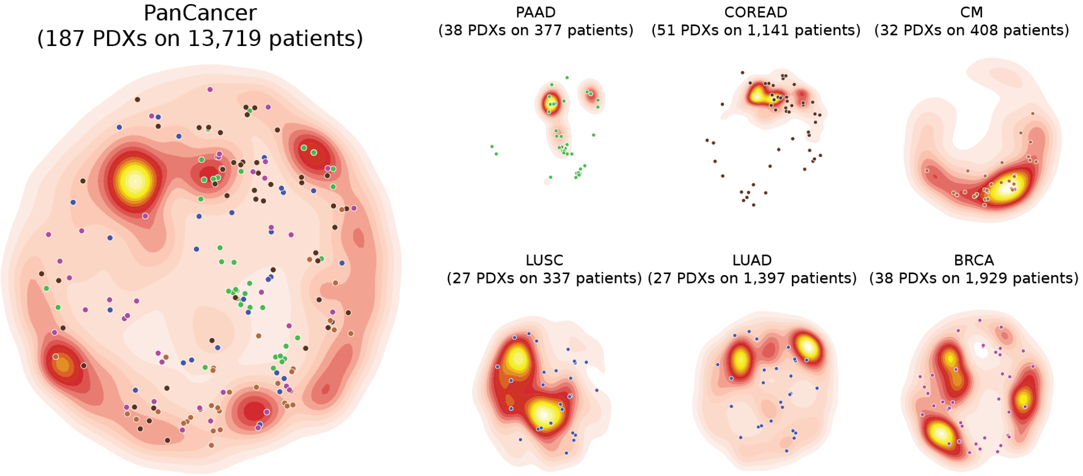


Figure 3.1: Molecular representativity of PDXs. OncoGenomic Landscape 2D representations of the molecular heterogeneity of the 187 PDXs annotated with both drug response data and oncogenic alterations, compared with that of their corresponding reference cohorts of cancer patients from TCGA and MSKCC. The points represent the location of each individual PDX, colored by tumor type. The distribution of the 187 PDXs can be compared to the distribution of patient samples, represented as density color-scale map in the background. PanCancer (n=15,212), BRCA (breast cancer, n=2,021), CM (cutaneous melanoma, n=492), COREAD (colorectal carcinoma, n=1,442), LUAD (lung adenocarcinoma, n=1,486), LUSC (lung squamous cell carcinoma, n=352), PAAD (pancreatic adenocarcinoma, n=442). Non-small cell lung cancer PDXs were mapped on top of both LUSC and LUAD reference populations.

Of the 62 drugs and drug combinations tested, we selected 53 treatment arms that showed significant inter-individual heterogeneity (i.e. a sufficient number of 'sensitive' and 'resistant' tumors) to model drug response. In total, these data comprised 3,127 experiments performed on 187 PDXs (H. Gao et al. 2015) for which we had, at least, 5 sensitive and 5 resistant PDXs. First, we assessed whether this set of PDXs is representative of the genomic diversity observed in human tumors by comparing their alterations to the oncogenomic profiles extracted from 13,719 cancer patients (Mateo et al. 2018). We found that the 187 PDXs considered broadly covered the whole oncogenomic landscape represented by the full cohort ('PanCancer' cohort in Figure 1). When analyzing tumor types individually, we observed that, while the mutational diversity of some of them is perfectly reflected in the PDX samples (e.g. colorectal and cutaneous melanoma tumors), the distribution of mutated genes showed clear differences in others (e.g. pancreatic cancer). Overall, there are PDXs representing the most populated areas of the PanCancer cohort, suggesting that the full collection of PDXs may be used in downstream analyses.

We adopted the Modified Response Evaluation Criteria in Solid Tumors (mRECIST) (H. Gao et al. 2015; Therasse et al. 2000) to assess the change in tumor volume in response to treatment. We considered to be 'sensitive' those PDXs that showed a Complete Response (CR), Partial Response (PR) or Stable Disease (SD), and 'resistant' those with a Progressive Disease (PD) status. We used the Cancer Genome Interpreter (Tamborero et al. 2018) to filter out passenger mutations from PDX profiles, and only worked with driver somatic mutations and copy number alterations.

For each treatment, we grouped sensitive and resistant PDXs, irrespective of the origin of their tumors. We then identified driver alterations that were overrepresented in sensitive or resistant PDXs, as well as pairs of driver alterations showing statistically significant patterns of co-occurrence in each subpopulation (see Materials and Methods for details). Finally, for each treatment, we built a Driver Co-Occurrence (DCO) network in sensitive PDXs consisting of overrepresented drivers (nodes) and pairs of co-occurring drivers (edges), another DCO network for resistant PDXs, and a third general one consisting of all drivers and co-occurrences associated with both treatment responses (Figure 2A).

DCO networks for each of the 53 drugs are detailed in Table S1 and can be visualized using Cytoscape (Shannon et al. 2003) (Supplementary Data S1). The total number of drivers and driver co-occurrences captured in the DCO networks varied substantially among treatments, ranging from 28 to 196 driver genes (median of 109 nodes, IQR: 82-136) and 20 to 1,499 pairs of drivers (median of 300 edges, IQR: 220-471) overrepresented in PDXs treated with Ruxolitinib+Biminetinib and the tankyrase inhibitor LJC049, respectively. However, when considering individual animals, the number of altered drivers and pairs of drivers was small and remained quite stable across treatments, with a median of only 9 genes (IQR: 5-15) and 7 driver co-occurrences (IQR: 2-29) per PDX (Figure 2B).

We next sought to assess the novelty of our DCO networks by comparing the overrepresented driver genes, and the co-occurring pairs, to the set of annotated response and resistance biomarkers for each treatment (Tamborero et al. 2018). Figure 2C shows that, although there is some overlap, our approach vastly expands the set of genes to be considered in downstream treatment prioritization applications. More specifically, 47 of the 58 genes annotated as approved or experimental response biomarkers are present in at least one DCO network, and 28 of them are related to the same drug or drug class. Additionally, our DCO networks include 331 novel genes that might be associated to treatment sensitivity or resistance.

Exploring the functional relevance of DCO networks

Even in targeted therapies, where the drugs are rationally designed to modulate well-characterized oncogenic alterations (e.g. HER2 amplification or the BRAFV600 mutation), it is known that alterations in other proteins do also influence drug response. For example, activating alterations in the MAPK or in PI3K/AKT pathways have been related to resistance to BRAF inhibition (Haarberg and Smalley 2014). We thus explored the functional relationship between the inferred DCO networks and the suggested mechanisms of action of each treatment through the analysis of the ten main oncogenic signaling pathways(Sanchez-Vega et al. 2018).

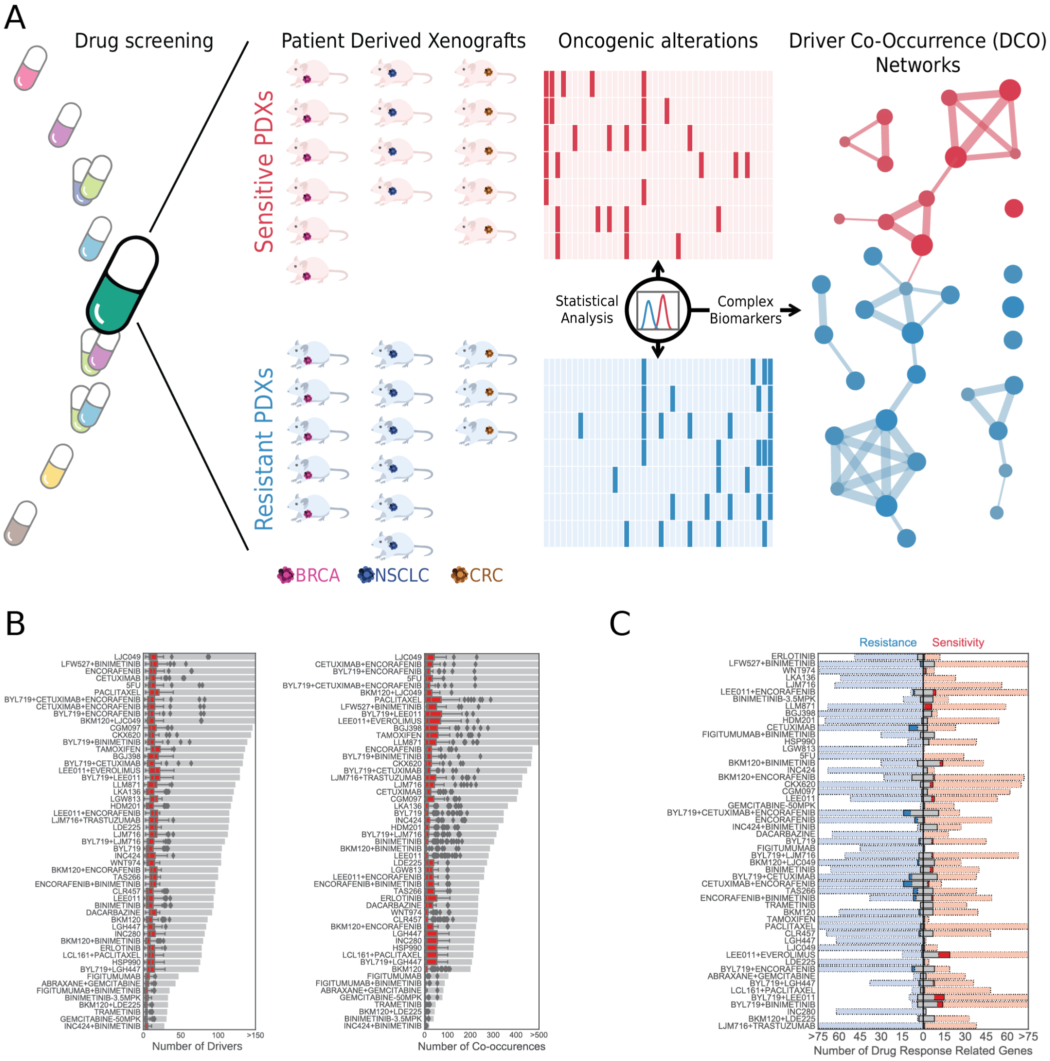
As expected, we find that approved or experimental biomarkers of drug response are directly related to the pathways where their intended targets belong (Fisher's exact test odds-ratio (OR) 5.14, p-value $1.4 \cdot 10^{-8}$). On the other hand, the genes in the DCO networks inferred for each treatment, while keeping certain functional coherence with the therapeutic targets (OR 1.59, p-value 0.090), show a larger functional diversity (Figure 3A, Table S2).

For instance, we found that the DCO networks derived in response to Tamoxifen are enriched in RTK/RAS/MAPK signaling proteins (OR 2.70, p-value 0.006; see Figure 3A and Table S2). When we analyzed separately sensitive and resistant DCO networks we found that the RTK/RAS/MAPK pathway, which is known to mediate resistance to this therapy (McGlynn et al. 2009), is indeed only overrepresented in the resistance DCO network (OR 3.08, p-value 0.004). Interestingly, other DCO networks are enriched in pathways not directly related to the known mechanism of action of the drug (e.g. 5FU, Cetuximab or LJC049, see Figure 3A and Table S2). However, the most striking observation is that cell cycle related proteins seem to play a central role in the inferred DCO networks for more than half of the treatments (35 of 53), irrespectively of the mechanism of action of the drug administered. This trend is not apparent when considering differentially altered drivers alone (Figure 3A).

Beyond the main oncogenic pathways, topological analysis of DCO networks revealed several

Figure 3.2 *(following page)*: Computational strategy and description of Driver Co-Occurrence (DCO) networks. (A) We inferred DCO networks from the analysis of 3,127 in vivo experiments that screened the efficacy of 53 treatments against a panel of 187 molecularly characterized PDXs of several tumor types. We first compared the patterns of oncogenic mutations and CNVs. in sensitive and resistant PDXs, regardless of the tissue of origin of the tumors. Next, we identified sets of driver genes showing differential alteration rates between responders and non-responders (DiffD), which are represented as red or blue nodes in DCO networks, respectively. Additionally, we identified pairs of genes whose alteration co-occurred more often than expected given the alteration rate of each driver (Ps), and that did so more often in one of the two response groups. We represented each pair of co-altered drivers as two nodes connected by an edge. We derived a sensitivity, resistance and global DCO network for each treatment. (B) Gray bars show the number of drivers and pairs of co-occurring drivers included in each DCO network derived from whole exome sequencing data. Red boxplots show the distribution of the number of drivers or driver co-occurrences identified in each individual PDX. (C) Blue and red boxes represent the overlap between DCO drivers and genes with annotated biomarkers of resistance or sensitivity, respectively. We show in light blue and light red the number of drivers in the resistance and sensitivity DCO networks that were not previously associated to drug response. Likewise, gray bars indicate the number of drug response associated genes that were not included in our DCO networks. In this analysis, we only considered as drug response associated those genes with biomarkers identified in two or more PDXs, which is a requirement that any driver needs to satisfy in order to be incorporated to a DCO network.

Figure 3.2: (continued)



large, strongly connected modules, composed of driver genes that had been co-amplified or co-deleted as part of the same genomic segment. To account for this effect, we clustered driver genes that are close in the genome and show similar alteration patterns (Materials and Methods; Supplementary Figure S2). After the filtering, we could still recapitulate known cases of co-amplification and simultaneous overexpression of adjacent oncogenes shown to provide a cellular cross-talk between tumorigenic pathways. For instance, in the LEE011(ribociclib)+Encorafenib, Dacarbazine, LDE225 (sonidegib), LGW813 (an IAP inhibitor), and TAS266 (a DR5 agopinst) DCO networks, we find links between MDM2 and CDK4, which are frequently co-altered as part of the same amplicon in the 12q chromosomal region (Dembla et al. 2018; Laroche-Clary et al. 2017).

Indeed, their hypothetical cooperation has triggered the use of CDK4/6 inhibitors as potentiators of MDM2 antagonists (Laroche-Clary et al. 2017), which are currently being tested to treat liposarcoma in clinical trials (NCT02343172 and NCT01692496). Another example is the concomitant amplification of ERBB2 and TOP2A as part of the 7q12 amplicon, which occurs in 40–50% of breast cancers (Arriola et al. 2008) and provides a rational basis for the addition of anthracyclines targeting TOP2A as adjuvant chemotherapy in the treatment of HER2-positive breast cancer (Gennari et al. 2008). However, the addition of doxorubicin to the standard regimen can potentially increase cardiotoxicity and failed to demonstrate a significant clinical improvement with respect to trastuzumap+paclitaxel in a phase III trial (Baselga et al. 2014), suggesting that further clinical evaluation is still needed.

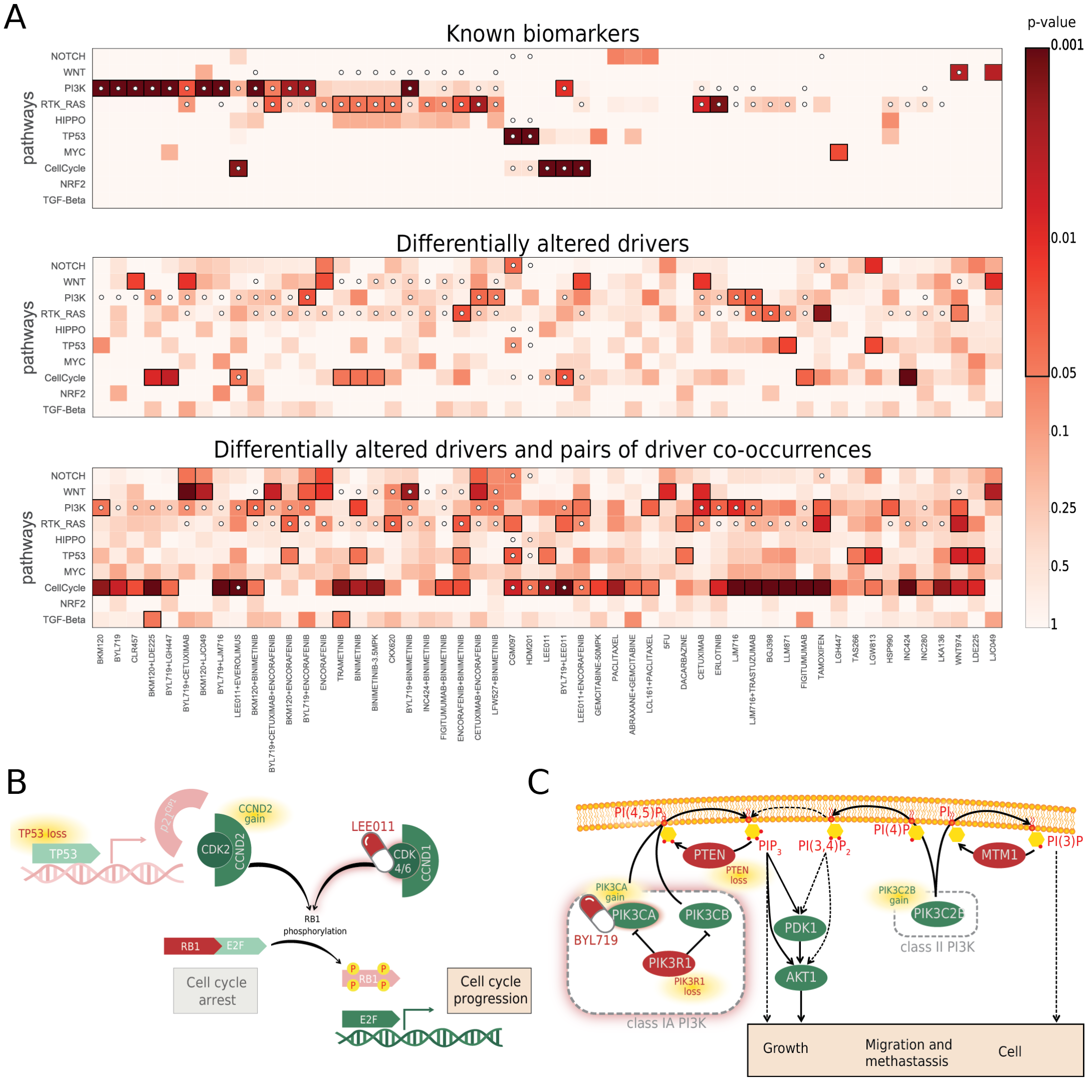
DCO networks do not only capture genome structure, but also functional relationships between oncogenic alterations found far apart in the genome. For instance, we observed that PDXs treated with the CDK4/6 inhibitor LEE011 show a markedly reduced response rate when they have co-occurring alterations in CCND2 and TP53 (27%, 3 out of 11) compared to PDXs with alterations in TP53 alone (41%, 29 out of 70) or wild-type TP53 and CCND2 (43%, 39 out of 90). Interestingly, alterations in TP53-CCND2 also tend to co-occur in a large collection of 74,247 pan-cancer samples compiled from 240 cancer studies (J. Gao et al. 2013), with TP53-CCND2 being co-altered in approximately 1% of the samples (751 patients, OR 2.57, p-value < 0.001). In particular, TP53-CCND2 alterations significantly co-occur in three out of five tumor-type specific cohorts: in 2,173 breast cancer patients (METABRIC, OR > 8, p-value < 0.001), in 479 patients with skin cutaneous melanoma (TCGA, OR 6.47, p-value 0.01), and in 507 patients with lung adenocarcinoma (TCGA, OR 5.24, p-value 0.015). However, we found no evidence of TP53-CCND2 co-occurrence in patient cohorts of other tumor types, such as colorectal adenocarcinoma (220 samples from TCGA), or pancreatic adenocarcinoma (175 samples from TCGA).

Although the role of TP53 status in response to CDK4/6 inhibition is controversial (Patnaik et al. 2016; Knudsen and Witkiewicz 2017), TP53 loss is thought to reduce the expression of its target p21CIP1 (CDKN1A) and consequently relieve CDK2 from its inhibition. On the other hand, cyclin D2 (CCND2) preferentially activates CDK2, although it can also activate CDK4 (Sweeney et al. 1997). Thus, based on our observations and the available literature, we hypothesize that concomitant oncogenic alterations in TP53 and CCND2 could shift the CDK4/6 dependency towards an alternative CDK2-dependent activation of G1/S transition, rendering those tumors insensitive to CDK4/6 inhibition (Figure 3B).

PIK3CA-mutant tumors are sensitive to isoform-selective PI3K inhibitors such as BYL719 (alpelisib) (Juric, Janku, et al. 2019; Juric, Rodon, et al. 2018; Andre et al. 2019). However, PIK3CA-independent mechanisms of PI3K activation (e.g. activating alterations in PIK3CB or PTEN loss) often confer re-

Figure 3.3 *(following page)*: Functional analysis of Driver Co-Occurrence (DCO) networks. (A) The three heatmaps show the enrichment of 10 main oncogenic signaling pathways across the set of genes with FDA-approved biomarkers, the set of drivers with differential alteration rate between responders and non-responders (DiffD), and the whole set of drivers and pairs of drivers in the DCO networks (DiffD_DiP). Associations with a one-sided Fisher's Exact test p-value < 0.05 are squared in black. White circles denote the presence of at least one drug target in a pathway, which is informative of the mechanism of action of each treatment. This representation shows that reported biomarkers tend to be enriched in the same pathways they are directly targeting, whereas DCO networks expand beyond the drug target, with the potential to uncover more distant functional relationships. Cell cycle related proteins seem to play a central role in the DCO networks inferred for almost half of the treatments (35 of 53), irrespectively of the mechanism of action of the drug. (B) The DCO network of Ribociclib, a CDK4/6 inhibitor, is enriched in cell cycle related proteins, such as CCND2, CCND3, CDKN2A, CDKN2B, CDK6 or RB1 (OR 6.54, p-value 0.0028; see Table S2). Based on the observed driver alteration co-occurrence patterns, we propose that the co-alteration of TP53 and Cyclin D2 (CCND2) might abrogate CDK4/6 dependency, rendering tumors insensitive to Ribociclib. More specifically, we hypothesize that TP53 loss would relieve CDK2 from the inhibitory activity of one of its major transcriptional targets, p21CIP1. This would synergize with the gain of function of CCND2, which preferentially binds to and activates CDK2, facilitating an alternative CDK4/6-independent activation of G1/S transition. (C) The DCO network of BYL719, an isoform-selective PI3K inhibitor, includes four proteins that are involved in PI3K signaling: PIK3CA, PIK3R1, PIK3C2B and PTEN. Tumors that depend exclusively on PIK3CA for the activation of PI3K signaling respond well to this treatment (65.2% response rate), whereas tumors in which PIK3CA alteration co-occurs with either PIK3R1, PIK3C2B or PTEN alterations show a response rate very similar tot that of wild-type PIK3CA tumors (45.45% and 44%, respectively). PIK3C2B, a member of class II PI3K family, tends to be co-altered with PIK3CA more often than expected, and more frequently in resistant than in sensitive PDXs. PIK3C2B contributes to phosphatydil inositol signaling by phosphorylating the third position of the inositol ring, taking as substrates both phosphatidyl inositol and phosphatidyl-4-phosphate inositol. The resulting products might directly or indirectly contribute to cell survival, growth or metastasis in a PIK3CA-independent manner, which would represent a novel mechanism of resistance to PIK3CA inhibition.

Figure 3.3: (continued)



sistance to this treatment (Nakanishi et al. 2016; Juric, Castel, et al. 2015). The DCO network of the BYL719 drug contains four proteins involved in PI3K signaling, namely PIK3CA, PIK3R1, PIK3C2B and PTEN.

Indeed, we observed a higher response rate (65%, 15 out of 23) among PDXs with oncogenic PIK3CA alterations compared to PDXs with wild-type PIK3CA (44%, 52 out of 117), which agrees with the mechanism of action of BYL719. More interestingly, we found that PIK3CA-altered PDXs having no co-occurring oncogenic alterations in the PI3K pathway (n=23) showed an even higher response rate (83 %, 10 out of 12) than those with co-occurring alterations in PIK3R1, PIK3C2B or PTEN (45%, 5 out of 11). These co-occurring alterations likely activate PI3K signaling in a PIK3CA-independent manner, hence the limited response to BYL719 treatment (Juric, Castel, et al. 2015) (Figure 3C).

Out of the three genes co-altered with PIK3CA, only PIK3C2B is found co-altered more often than expected in PDXs treated with BYL719 (4.06% inferred co-occurrence rate vs. 1.50% expected; expected value of the difference (e-value) 0.006), and we indeed observed that PDXs with PIK3CA-PIK3C2B co-alteration showed a lower response rate (33%, 2 out of 6) than those with PIK3CA alteration alone (76%, 13 out of 17). Finally, PIK3CA and PIK3C2B alterations also co-occur in approximately 1% of the 74,247 pan-cancer samples (758 patients, OR 2.99, p-value < 0.001), being particularly co-altered in breast cancer patients (METABRIC, OR 1.48, p-value < 0.001) and pancreatic adenocarcinoma patients (TCGA, OR > 8, p-value 0.002).

Overall, DCO networks capture co-occurring alterations associated to drug resistance in PDXs, as illustrated by the concomitant alteration of CCND2-TP53 in relation to CDK4/6 inhibition and that of PIK3CA-PIK3C2B in relation to PI3K inhibition. Moreover, many of these co-occurrence patterns are also found in patient cohorts, indicating a potential clinical translation of these findings.

TCT4U: A collection of 53 drug response classifiers for genome-driven treatment prioritization

We then explored whether the sets of differentially (co-)altered genes in sensitive and resistant PDXs can be used to predict drug response. For each treatment, we used the DCO networks to statistically classify PDXs as resistant or sensitive. The goal of this exercise is to identify, among the available treatments, the best possible option for each individual based on its oncogenomic profile. We thus named the set of developed drug response classifiers Targeted Cancer Therapy for You (TCT4U).

In brief, for each treatment arm, we combined the probabilities assigned by three Naïve Bayes (NB) classifiers, trained with sensitivity, resistance and general DCO networks, into a single prediction score per drug-PDX pair. Figure 4A shows the performance of the NB classifiers in a leave-one-out cross-validation setting, whereby the oncogenomic profile of PDXs is used to predict response to each treatment.

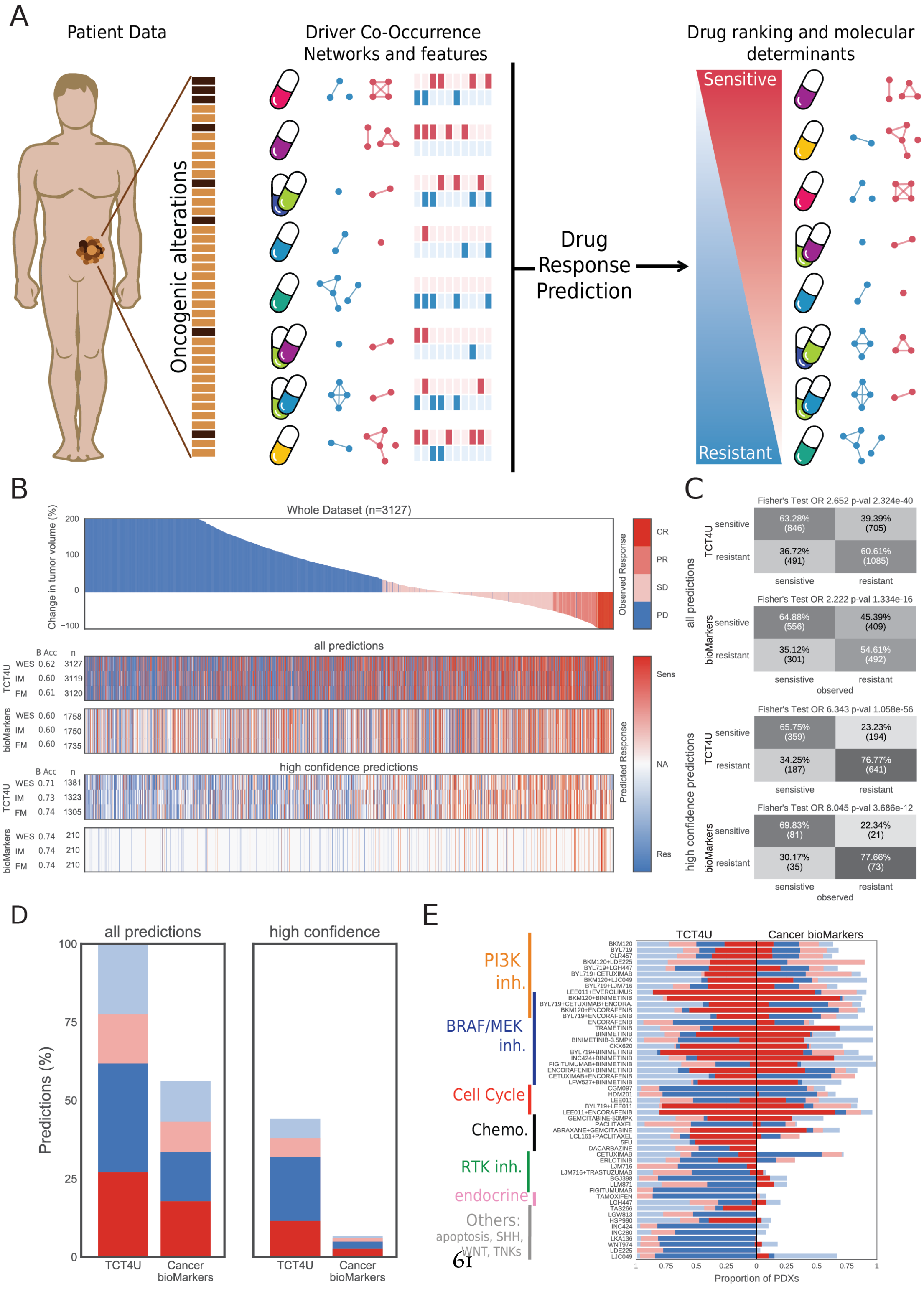
Additionally, to increase the clinical translatability of our approach, we repeated the calculations considering only those alterations detectable by the Memorial Sloan Kettering-Integrated Mutation Profiling of Actionable Cancer Targets (IM) (Cheng et al. 2015; Zehir et al. 2017) and the Foundation Medicine (FM) gene panels (Frampton et al. 2013), which contain probes to detect 410 and 287 mutated genes, respectively, and are widely used in clinical settings. Finally, we assessed the effectiveness of TCT4U by comparing its predictive power to that of FDA-approved and experimental biomarkers (see Materials and Methods for details).

We collected the change in tumor volume and the mRECIST classification for a total of 3,127 experiments with reported treatment outcome, comprising 187 PDXs tested for response to 53 treatments. Figure 4B shows that TCT4U models are applicable to all drug-PDX pairs (3,127), while alterations in approved and experimental biomarkers can only be found in about half of them (1,758). However, wherever applicable, the accuracy attained by both methods is almost identical: TCT4U correctly classified 64% of sensitive and 56% of resistant drug-PDX pairs, while approved or experimental biomarkers attained sensitivity and resistance accuracies of 65% and 55%, respectively. Overall, TCT4U models yielded correct predictions for 1,866 (60%) drug-PDX pairs, while the figure achieved by known biomarkers is 1,048 (33%). It is also remarkable that, even if they consider a much lower number of genes, both IM and FM derived models achieved comparable prediction accuracies (Figure 4B, Supplementary Figure S3).

In a treatment decision setting, we would not need to predict the effects of every possible drug on each patient, but only those drugs that might work best or, also importantly, those drugs that would

Figure 3.4 *(following page)*: Targeted Cancer Therapy for You (TCT4U), a collection of Naïve Bayes drug response classifiers based on DCO networks. (A) Given a new tumor sample, we compare it to the patterns of driver alterations and co-alterations associated to sensitivity or resistance to any of the treatments in TCT4U, and rank the drugs accordingly, predicting whether a drug will or not be effective. Since the number of driver alterations that tumors typically have is relatively small, we can know which are the molecular determinants used by the classifier and use this information for functional interpretation of the predictions. (B) Waterfall plot representation of the outcome of the in vivo pharmacogenomic screening used to infer our collection of DCO networks and TCT4U drug response classifiers. Each bar represents the best average response of one of the 3,127 in vivo experiments, sorted (left to right) from the worst to the best response to treatment, and colored according to the mRECIST classification as PD (progressive disease), SD (stable disease), PR (partial response) and CR (complete response), as proposed by [REF Gao]. The heatmaps below show the predictions of TCT4U in a leave-one-out cross validation setting and the predictions made on the basis of known biomarkers. Each heatmap has three rows, which correspond to the predictions obtained when examining the whole exome (WES) or a subsampled molecular profile containing the genes covered by IMPACT410 (IM) or Foundation Medicine (FM) targeted gene panels. The number of predictions and their balanced accuracy are annotated along the y-axis. The set of 'high-confidence' predictions refers, on the one hand, to the subset of 10 highest scoring sensitivity and resistance predictions per PDX, and to the subset of clinically approved biomarkers on the other hand. (C) Contingency tables showing the association between the observed and the predicted drug responses based on WES profiles. (D) The precision of each set of predictions is illustrated by the red and blue sections of the stacked bar plots, which represent the proportion of correct sensitivity and resistance predictions. Analogously, incorrect predictions are represented in faint colors. Missing predictions (NA) are represented in white to offer a comparative overview of the recall. (E) Stacked bar plots representing the precision and recall of all TCT4U predictions and all reported biomarkers, covered by WES profiles split by treatment arm.

Figure 3.4: (continued)



not work. Thus, when we considered only the top-10 highest scoring sensitivity and resistance drugs for each of the 187 PDXs (i.e. high-confidence predictions), the precision of TCT4U significantly improved to 66% and 77%, respectively. We found very similar numbers for approved biomarkers (70% and 78%) although, in this case, they could only predict drug response in 210 of the drug-PDX pairs, spanning 59 PDXs (Figure 4D).

Overall, we obtained a strong association between predicted and observed drug responses when analyzing all TCT4U models (OR 2.65, p-value $2.4 \cdot 10^{-40}$) that was even stronger when we focused on the high-confidence ones (OR 6.34, p-value $1.1 \cdot 10^{-56}$). In both cases, these associations are two-fold stronger than the ones achieved by approved and experimental biomarkers (p-value $1.3 \cdot 10^{-16}$).

Finally, if we only focus on the drug with the highest probability of response per PDX (i.e. the most realistic scenario), TCT4U correctly predicted 56 effective drugs in 70 PDXs (80% accuracy) and 49 inefficacious drugs in 74 PDXs (66% accuracy), while the corresponding figures achieved by approved and experimental biomarkers are 16 out of 18 correct sensitive (89% accuracy) and 18 out of 23 resistance predictions (83% accuracy). Please, note that for the remaining PDXs (117 sensitive and 113 resistant), our top sensitivity and resistance predictions had no experimental data available and, thus, we cannot assess their accuracy.

Finally, while the coverage of approved or experimental biomarkers is mostly limited to BRAF/MEK inhibitors, PI3K/mTOR inhibitors or cell cycle related treatments, the predictions made by TCT4U also cover other drug families including chemotherapies, RTK inhibitors, endocrine therapies, and more experimental treatments targeting WNT (WNT974), SHH (LDE225) or apoptosis related pathways (TAS266, LGW813), among others (Figure 4E).

Experimental validation of TCT4U drug response predictions on a prospective PDX dataset

Additionally, we sought to prospectively evaluate the performance of the TCT4U models in new tumors. To this aim, we selected, from our VHIO collection of molecularly-characterized breast cancer PDXs, a subset of 15 tumors for which TCT4U prediction were of high confidence (i.e. in the top 10). Moreover, to assess the added value of our drug response predictors, we selected drug-PDX pairs for which the anticipated outcome did not agree with approved or experimental biomarkers, either because the individual oncogenomic profiles did not have any biomarker altered (n=9), or the TCT4U predictions were opposed to those suggested by known biomarkers (n=7).

The final validation set consisted of 16 drug-PDX pairs, with 10 tumors predicted to be sensitive and 6 to be resistant, comprising an isoform-selective PI3Kα inhibitor (BYL719, n=5), a CDK4/6 inhibitor (LEE011, n=2), the combination of both (BYL719+LEE011, n=3), a MEK inhibitor (selumetinib, n=2), an estrogen receptor antagonist (tamoxifen, n=2), and a taxane (paclitaxel, n=2). We subcutaneously implanted the tumors in immunocompromised mice and let the tumors grow until they reached a volume of 120-150 mm3. We then treated the PDXs for 15-57 days and measured their response to the administered drugs following the mRECIST guidelines (see Materials and Methods for details). The complete results of our study, including treatment setting (drug dose, duration, etc.) and tumor response (tumor growth, mRECIST classification, etc.) for every PDX can be found in Table S3, and are summarized in Figure 5.

We treated five PDXs with BYL179, four of which (PDX131, PDX293, PDX156 and PDX191) were predicted to be sensitive to the drug by TCT4U models, and one (PDX153) to be resistant. The four PDXs predicted to be sensitive showed co-alterations of CCND1, FGF3 and FGF4. These genes are located in the 11q13.3 genomic segment, and DCO networks found this region to be amplified more often in sensitive than in resistant PDXs, with an alteration rate of 7.46% and 1.37%, respectively (e-value 0.05). It is worth noting that our model, which was derived from 140 PDXs of different tumor types (i.e. 38 BRCA, 42 COADREAD, 25 NSCLC and 35 PDAC), did not show a significant tendency towards co-occurrence of PIK3CA and the 11q amplicon (OR 2.69 p-value 0.26).

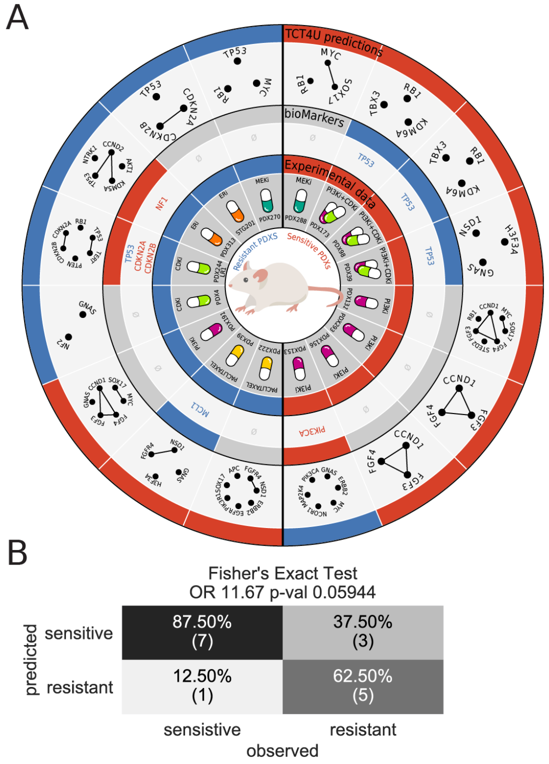
Dysregulation of FGFR signaling can lead to downstream activation of PI3K/AKT pathway and, indeed, a recent study reported that 73% of patients (8 of 11) with both an alteration in the PI3K/AKT/mTOR pathway and FGF/FGFR amplification experienced clinical benefit when treated with therapy targeting the PI3K/AKT/mTOR pathway, whereas only 34% of patients (12 of 35) with PI3K/AKT/mTOR alterations alone did so (Wheler et al. 2016).

However, the implication of FGF signaling with respect to the clinical benefit of PI3K/AKT/mTOR blockage remains controversial. The retrospective analysis of a large subset of patients enrolled in the BOLERO-2 trial (Hortobagyi et al. 2016) showed that alterations in FGF signaling had a negligible impact (FGFR1) or slightly decreased (FGFR2) the clinical benefit of everolimus treatment. In line with these findings, ER+/ERBB2- metastatic breast cancer patients with FGFR1 and FGFR2 amplification did not derive a clinical benefit from BYL719+letrozole (Mayer et al. 2017). Accumulating evidence suggests that FGF signaling by FGFR1/2 amplification attenuates the response to PI3K blockage in PIK3CA mutant breast cancer. However, the impact of FGF signaling in response to BYL719 in PIK3CA wild-type tumors originated from breast as well as from other tissues has yet to be determined.

In our dataset, three out of the four PDXs responded to the treatment. In particular, in PDX293 we observed a partial response (PR) after 18 days of treatment, with a reduction of 65% in the initial

Figure 3.5 *(following page)*: In vivo validation of 16 high-confidence TCT4U predictions based on IMPACT410 profiles with missing or conflicting reported biomarkers. (A) The circular plot summarizes the results of 16 in vivo experiments comprising 5 PDXs treated with the BYL719 isoform-selective PI3K inhibitor (PI3Ki), 2 PDXs treated with LEE011 CDK4/6 inhibitor (CDKi), 3 PDXs treated with the combination of both (PI3Ki+CDKi), 2 PDXs treated with binimetinib MEK inhibitor (MEKi), 2 PDXs treated with tamoxifen ER antagonist (ERi), and 2 PDXs treated with paclitaxel. The innermost track shows the experimentally determined treatment outcome, in which responder tumors showing disease stabilization or regression are represented in red, and non-responders are represented in blue. The middle track represents the predictions based on reported biomarkers and the genes to which they are annotated, when available. The outermost track represents the predictions based on TCT4U and their underlying molecular determinants. TCT4U predictions are sorted from correct to incorrect following clockwise and anticlockwise directions for sensitivity and resistance, respectively. (B) Contingency table showing the association between the observed and predicted responses to treatment.

Figure 3.5: (continued)



tumor volume. PDX131 and PDX156 showed a stable disease (SD) after 20 and 11 days of treatment, respectively. On the other hand, in PDX191 the tumor increased its volume by 80% after 13 days of treatment (PD), and we thus considered it a wrong prediction. However, after 43 days of treatment, we could observe a halving in tumor growth (235%) with respect to untreated animals (501%) (Table S3).

PDX153 was the only PDX with an oncogenic PIK3CA mutation (p.K111E) reported to confer sensitivity to the treatment (Tamborero et al. 2018) and, indeed, we observed a significant reduction of 83% in the tumor volume after 35 days of treatment (i.e. a PR outcome). Our model classified this PDX as resistant because it also had other alterations overrepresented among resistant PDXs, such as MAP2K4 (e-value 0.011) or NCOR1 (e-value 0.016). The DCO networks also considered PIK3CA status, which is more frequently altered in sensitive PDXs (22.29%) than in resistant PDXs (11.40%; e-value 0.071). However, it seems that the final prediction was driven by additional oncogenic alterations that showed stronger statistical association than PIK3CA status, although they proved to be less informative.

We administered LEE011, a CDK4/6 inhibitor, to PDX4 and PDX244_LR1, with the TCT4U prediction that the two tumors would be resistant to the drug. PDX4 did not present any known biomarker of drug response, but it showed a heterozygous loss of NF2. Oncogenic alterations in NF2 are overrepresented among resistant PDXs in the DCO network (e-value 0.037) and for this reason PDX4 was predicted to be resistant. Interestingly, loss of NF2 has been associated to increased CDK6 expression and was previously identified as mechanism of resistance to CDK4/6 inhibition in ER+ mestastatic breast cancer patients (Li et al. 2018).

On the other hand, we also treated PDX244_LR1, which is a model of acquired resistance to LEE011 derived from a sensitive parental tumor (PDX244). Accordingly, PDX244_LR1 simultaneously showed known biomarkers of sensitivity (CDKN2A-CDKN2B co-deletion) and resistance (TP53 p.C176R) to the treatment (Tamborero et al. 2018). Although both genomic events were also considered by TCT4U models and, in line with what has been reported, CDKN2A-CDKN2B co-deletion is slightly more common in sensitive than in resistant PDXs (32.78% vs. 25.34%; e-value 0.200), we did not find a significant association of TP53 with resistance (44.65% vs. 48.39%; e-value 0.480), and thus it is not included in the LEE011 DCO network.

Moreover, PDX244_LR1 presents an oncogenic mutation in RB1 (p.M695Nfs*26), which showed a strong association with resistance to CDK4/6 inhibition in the DCO networks (4.29% vs. 12.65%, DiffD e-val 0.037). RB1 is the primary target of CDK4/6 and its status is a key determinant of CDK4/6 inhibition efficacy (Shapiro 2017). Accordingly, RB1 overexpression is reported to confer sensitivity to CDK4/6 inhibition in prostate cancer, but its loss or deletion is not currently reported as a resistance biomarker (Tamborero et al. 2018). Our experiments showed that, and in agreement with TCT4U predictions, the tumors increased their volume between 45 and 215%, being thus catalogued as PD.

We also treated three PDXs (PDX173, PDX98 and PDX39) with the same PI3Kα and CDK4/6 inhibitors in combination (BYL719+LEE011). The three of them had oncogenic mutations in TP53 (p.R249S, p.R249S and p.V157I), which are associated with resistance to CDK4/6 inhibition (Tamborero et al. 2018). However, DCO networks found additional sensitivity-associated genomic features and thus TCT4U models predicted them as sensitive to this drug combination. More specifically, TBX3 disrupting mutation, present in PDX173 and PDX98 DCO networks, is significantly associated to sensitivity to this treatment (6.19% vs. 0.05%; e-value 0.006).

We found that, indeed, all three tumors responded to the combination treatment: PDX173 became completely tumor free (CR), PDX39 showed a reduction of 47% (SD), and PDX98 of 25% (SD). Interestingly, TBX3 is a transcriptional repressor of p21 and p14, which are directly upstream of cyclin-CDKs, and also of PTEN (Willmer et al. 2015). TBX3 has been shown to directly repress PTEN in neck squamous carcinoma cells (Burgucu et al. 2012), and thus TBX3 loss would result in PTEN up-regulation influencing the response to PI3K inhibition. Although this hypothesis seems plausible, TBX3 loss showed a low allelic fraction in PDX173 and its direct implication in response to this drug combination should be confirmed experimentally.

Two PDXs were treated with the MEK inhibitor Binimetinib, with TCT4U models predicting PDX270 to be resistant and PDX288 to respond to the drug. Both PDXs presented RB1 loss (a loss-of-function mutation p.Y321* and a deletion, respectively), which is significantly associated to resistance in the DCO network (5.51% vs. 16.89%; e-value 0.015). Additional alterations necessarily contributed to the divergent prediction of those PDXs.

The prediction of resistance in PDX270 was not likely to be driven by TP53 loss, since DCO networks did not find this alteration significantly associated to MEK inhibition response (51.53% vs. 41.03%; e-value 0.165). The two PDXs also shared MYC amplification, which in our DCO networks is also not significantly associated to differential response to MEK inhibition (21.85% vs. 24.19%; e-value 0.730). However, we found that MYC was significantly co-altered with SOX17 in PDX288. This co-alteration is distinctive of sensitive PDXs, with an observed co-occurrence rate of 13.11% with respect to an expected 3.09% (e-value $< 1 \cdot 10^{-4}$), and drove the TCT4U prediction. In this case, neither tumor presented known biomarkers of response to MEK inhibition.

When treated with Binimetinib, PDX270 was classified as non-responder (PD), as the tumor volume had increased by 144%, even more than in untreated animals (117%). On the contrary, and validating the TCT4U models, PDX288 responded well to treatment (SD), and tumors did not show any significant growth. Interestingly, an integrative genomics screen performed in 229 primary invasive breast carcinomas identified the co-amplification of MYC and the 8p11-12 genomic region, together with aberrant methylation and expression of several genes spanning the 8q12.1-q24.22 genomic region (Parris et al. 2014). This observation coincides with our DCO network derived from whole exome sequencing data, where we could detect the co-amplification of a large cluster of genes located in the 8p11-p12 (HOOK3, TCEA1) and 8q11.23-q24.22 genomic regions (SOX17, PLAG1, CHCHD7, NCOA2, COX6C, MYC, NDRG1) in sensitive PDXs, but not in resistant ones (Supplementary Figure S1C-D).

We selected two additional PDXs to be treated with an estrogen receptor (ER) antagonist (Tamoxifen) and, in agreement with TCT4U prediction,, we could confirm that both tumors were resistant to this treatment. PDX313 was ER+ but TCT4U models predicted it as resistant because it presents several alterations, namely AKT1 p.E17K (0.08% vs. 8.75%, e-value 0.007), NTRK1 amplification in chromosome 1 (0.20% vs. 23.29%, e-value 0.004), and the co-amplification of CCND2-KDM5A in chromosome 12 (0.06% vs. 23.15%, e-value 0.002; and 0.06% vs. 20.26%, e-value 0.002), that are associated to resistance in the DCO network. Moreover, this PDX had an oncogenic mutation in NF1 (p.L375V) that, despite being associated to Tamoxifen sensitivity in neuroblastoma (Tamborero et al. 2018; Byer et al. 2011), has been associated to endocrine resistance in HR+HER2− breast cancer patients (Razavi et al. 2018).

Likewise, our models predicted resistance in STG201, an ER- PDX. It is noteworthy that the current implementation of TCT4U does not consider ER status because this biomarker cannot be determined from somatic DNA alterations. However, our method might still be able to detect subtype specific dependencies that might influence response to endocrine therapy. In Tamoxifen DCO, the co-alteration of CDKN2A-CDKN2B, but not the alteration of CDKN2A alone (19.79% vs. 29.60%; e-value 0.515), is associated to resistance (17.89% vs. 5.30%, e-value 0.019) and is the genomic feature that more likely driving this prediction. Moreover, both tumors present oncogenic alterations in TP53, CDKN2A-CDKN2B, CCND2 and RB1 that might uncouple ER signaling and cell cycle progression (PDX313), which is a reported mechanism of resistance to endocrine therapy (Thangavel et al. 2011).

Finally, we explored the TCT4U prediction capacity in cytotoxic chemotherapy, where specific oncogenic characteristics should be less related to treatment efficacy. We selected PDX222 and PDX39 to be treated with Paclitaxel. While PDX222 did not present any known biomarker of response, PDX39 sowed an MCL1 amplification, which has been reported to promote resistance to antitubulin chemotherapeutics (Tamborero et al. 2018; Wertz et al. 2011). Although PDX222 showed alterations that are slightly more common in resistant than in sensitive PDXs (EGFR, SOX17 and APC, all with insignificant e-values), it also presented an ERBB2 amplification that in our model is strongly associated to sensitivity (14.70% vs. 0.05%; e-value $6 \cdot 10^{-4}$), and a co-amplification of FGFR4-NSD1 in chromosome 5, which also occurred more often than expected in sensitive than in resistant PDXs (7.33% vs. 0.54%, e-value 0.015).

Regarding PDX39, the genomic feature with the strongest association in TCT4U was the same co-amplification of FGFR4-NSD1 mentioned above, followed by the alteration of GNAS (7.49% vs. 2.81%; e-value 0.334), which is also slightly more frequent in sensitive than in resistant PDXs. Accordingly, we predicted that both tumors would respond to the drug. When treated with paclitaxel, both PDXs showed a progressive disease (PD), proving the TCT4U predictions wrong, although the growth of the tumors was 75% and 33% smaller in treated than in untreated mice.

Overall, TCT4U models correctly predicted the outcome of 12 of the 16 (75%) treatments tested, validating 70% of sensitivity (7 of 10) and 83% of resistance (5 of 6) predictions, which is in good agreement with the cross-validation results for high-confidence predictions (66-77% precision). However, in this challenging prospective validation, known biomarkers only predicted correctly 2 of the 16 (12%) treatment outcomes. In particular, two of the TCT4U misclassified responses were correctly predicted by known biomarkers, while the rest were either incorrect (5 of 7) or missing (9).

Bringing TCT4U from the workbench to the clinics

To explore the clinical potential of TCT4U methodology, we analyzed a cohort of 116 metastatic breast cancer patients being treated at the Memorial Sloan Kettering Cancer Center (Li et al. 2018), and for which we have recorded information of their oncogenomic profile and clinical outcome (Table S4). These metastatic patients had received between 1 and 17 rounds of treatments (median of 2) before being selected for a trial to test a combination of CDK4/6 and aromatase inhibitors. Each tumor was genetically profiled, using the MSK-IMPACT panel, and the clinical outcome of the treatment was recorded as progression free survival (PFS).

In this study, one third of the patients did not derive a clinical benefit and relapsed before 5 months. At the other extreme of the distribution, one third of the patients could be treated for more than 10 months and were considered to present a durable clinical benefit. We are aware that a threshold of 10 months might not be relevant in a first line treatment setting, where this drug combination has shown to achieve a median PFS of 24 months (Tanguy et al. 2018). However, the PFS decreases in subsequent lines of therapy and, in a metastatic setting where over half of patients have received prior therapies, a PFS of more than 10 months might still be good surrogate measure of the clinical benefit.

We did not have PDXs treated with a combination of CDK4/6 and aromatase inhibitors, and the best TCT4U model for it was derived in response to CDK4/6 inhibition (LEE011), based on 71 sensitive and 100 resistant PDXs. Using this model, only 6 out of 216 patients were predicted to be sensitive to treatment, and only one of them showed a clinically significant PFS (13.5 months). The majority of patients (78%) relapsed within the first year of treatment but, unfortunately, we have no data in this clinical series as to whether the tumors regressed, at least initially. It thus seems that the outcome measure used to train the TCT4U model (mRECIST), based on relative tumor growth, is not appropriate in most clinical settings.

Without a model for this specific drug combination, and with the aforementioned differences in outcome measures, we decided to adapt our methodology to classify patients based on the duration of the treatment before cancer relapsed. For this, we divided the cohort in three groups and considered the 40 patients for which the tumors relapsed before 4.2 months after the start of the treatment as resistant, and the 40 for which the time to progression was longer than 9.7 months as responsive to the treatment.

The resulting DCO networks for this treatment, which are relatively small compared to TCT4U DCO networks, contain a total of 18 drivers and 16 co-occurring pairs (see Figure 6A). The strongest associations captured by the DCO network are MYC, MAP3K1, ATR and ERBB2 alterations, which happen more frequently in sensitive than in resistant patients (e-values of 0.001, 0.002, 0.004, and 0.033, respectively). On the other hand, we find that MAP2K4, FGFR2, FAT1, ESR1 and BCL6 are more frequently altered in resistant than in sensitive patients (e-values of 0.005, 0.005, 0.014, 0.017 and 0.044, respectively; Table S1). Indeed, FAT1 loss has been recently associated to resistance to this treatment through a mechanism that involves the activation of Hippo pathway, leading to an increase in CDK6 expression (Li et al. 2018). Oncogenic mutations in ESR1 are also common in metastatic and pretreated breast cancer, emerging as a mechanism of acquired resistance to endocrine therapies that can ultimately result in resistance to the combinational tehrapy (Preusser et al. 2018).

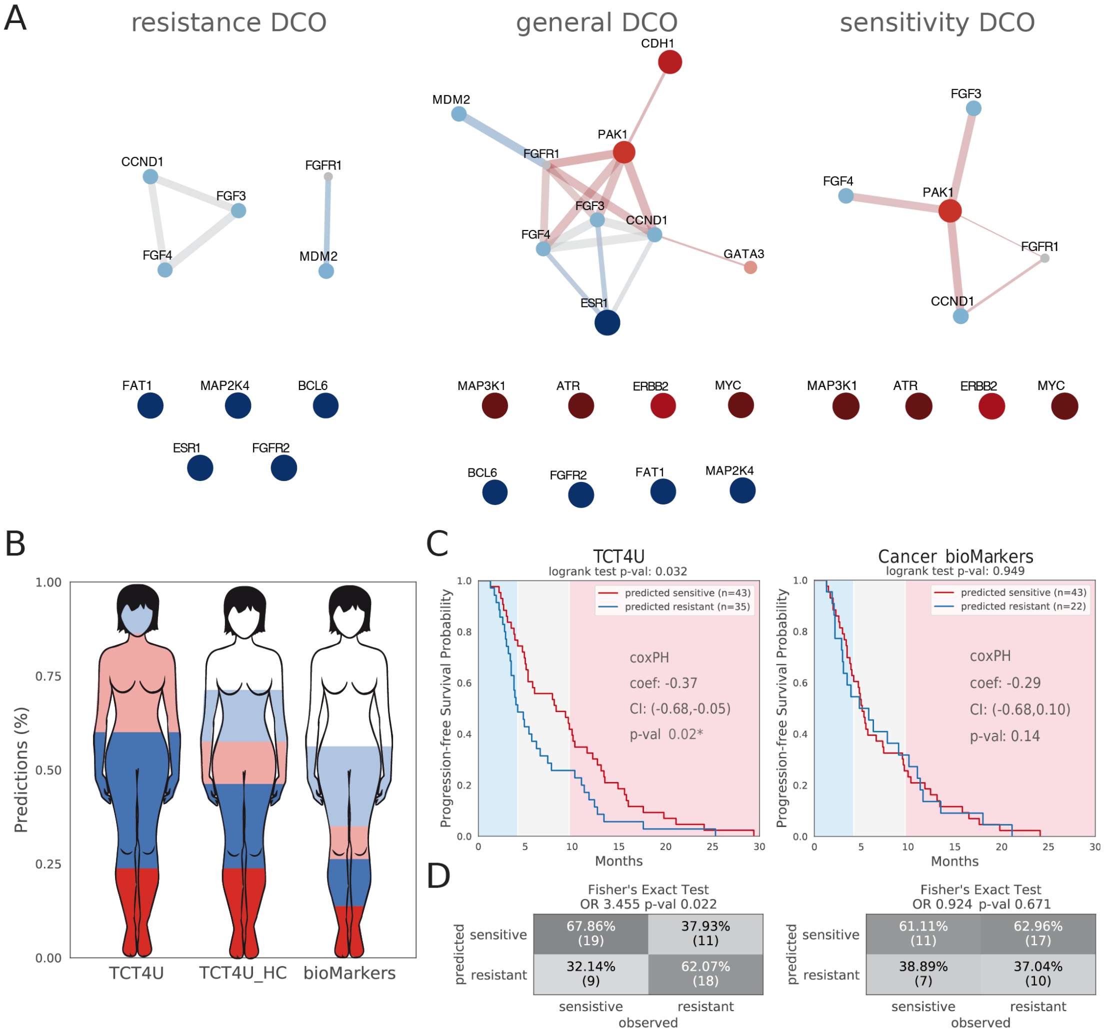
Regarding driver co-occurrences, the triplet formed by FGF3-FGF4-CCND1 oncogenes, located in the 11q13.3 genomic region, is co-altered more often in resistant than in sensitive patients. However, those three oncogenes tend to be co-altered with PAK1 (in 11q14.1), more often in sensitive than in resistant patients (e-values of 0.006, 0.007 and 0.005). It has been suggested that the amplitude of the regions affected by copy number changes strongly determines patient prognosis (Smith and Sheltzer 2018). Broader amplifications of this region (i.e spanning both 11q13 and 11q14) are likely to modify the dosage of multiple genes, which could have a cost in terms of cancer fitness and might contribute to the clinical benefit of the treatment.

On the other hand, FGFR1 (8p11.23) alteration also tends to be co-altered with the four aforementioned genes (e-values of 0.008, 0.006, $6 \cdot 10^{-4}$ and 0.002) although this does not seem to affect drug response. It rather seems to reflect the putative synergy between the gain of function of FGF3 and FGF4 ligands and their cognate receptor FGFR1. Another interesting co-occurrence is FGFR1-MDM2 co-alteration, which occurs more often than expected in resistant patients (e-value 0.020).

Overall, we saw that, although they shared a CDK4/6 inhibitor, the DCO networks were indeed

Figure 3.6 *(following page)*: Application of TCT4U to predict treatment outcome in a clinical cohort of HR+/HER2- metastatic breast cancer patients. (A) Driver Co-Occurrence networks (DCO) representing the oncogenic alterations and pairs of alterations that are overrepresented in patients that relapsed early (resistant) or in in patients that derived a durable clinical benefit (sensitive) from CDK4/6 inhibition in combination with hormonal therapy combined with a. (B) Stacked bar plots representing the precision and recall of TCT4U cross-validated predictions and that of approved and experimental biomarkers. The blue and red sections of the stacked bar plots represent the proportion of correct predictions, in terms of the classification into early or late relapse. Analogously, incorrect predictions are represented in faint colors. Missing predictions are represented in white, offering a comparative overview of the recall. (C) Kaplan-Meier analysis of progression free survival (PFS). TCT4U high-confidence predictions are better able to discriminate between patients that would experience early and late relapse than known biomarkers, with a median time to progression of 4.2 and 8.3 months respectively (log-rank test p-value 0.03 and Cox's PH coefficient of -0.37, p-val 0.02). (D) The contingency tables show the concordance between observed and predicted clinical benefit.

Figure 3.6: (continued)



very different to those derived in response to LEE011, with only 3 of the 35 driver genes (BCL6, FAT1, and MYC) and none of the co-occurrences in common. We then used the DCO networks to derive the corresponding TCT4U models, which should be able to predict whether a given patient will obtain a significant clinical benefit.

In a leave-one-out cross validation, TCT4U models yielded confident scores for 78 out of the 116 patients (see Materials and Methods). Of these, we predicted that 43 patients would be sensitive and 35 resistant to the treatment. Indeed, we validated 19 of the 30 sensitive and 18 of the 27 resistant predictions (Figure 6), while the remaining 21 patients obtained an uncertain clinical benefit (i.e. TTP between 5 and 10 months). Put together, we obtained a significant association between the predicted and the observed clinical benefit (OR 3.45, p-value 0.022), with an overall accuracy of 67% (38 of 57).

Additionally, a Kaplan-Meier analysis of the cross-validation showed that the 35 patients predicted to relapse early, with a median time to progression of 4.2 months, derived little clinical benefit compared to the 43 patients predicted to relapse later, whose median time to progression was significantly longer (8.3 months, log-rank test p-value 0.030). We obtained consistent results when fitting a Cox proportional hazards regression model (correlation coefficient -0.37, p-value 0.022), indicating that TCT4U scores are correlated with progression free survival. The performance of TCT4U models clearly surpassed that of known biomarkers for this drug combination. Although 56% (65 of 116) of patients had at least one annotated biomarker, which is a good coverage compared to other treatments, we could not find a significant association between observed and predicted outcomes, at least in terms of PFS (Figure 6).

Our results suggest that the proposed methodology could be used to derive DCO networks and train predictive models from the kind of data obtained from interim analyses in oncological clinical trials. Moreover, whenever the time to detect a clinical benefit is reasonable, such the 10 months in this study, TCT4U models could be derived with the first patients and used in population enrichment strategies to establish the bases for new recruitments in adaptive trials.

Concluding remarks

Cancer sequencing projects have unveiled hundreds of gene alterations driving tumorigenesis, enabling precision oncology. Indeed, current efforts now focus on the analyses of oncogenomic patterns to identify actionable alterations, drugs to modulate them and biomarkers to monitor response. Of particular interest are computational platforms such as OncoKB (Chakravarty et al. 2017) or the Cancer Genome Interpreter (Tamborero et al. 2018), which not only identify oncogenic alterations and potential targets, but also estimate their potential clinical applicability.

Most current strategies focus on the identification of a single vulnerability (i.e. driver gene) whose activity can be modulated by a drug. However, given the complexity and heterogeneity in tumors, and the high connectivity between cellular processes, every cancer might respond differently to a certain treatment, depending on its global oncogenomic profile. Indeed, the analysis of the mutational landscape of cancer has also uncovered the existence of mutual exclusivity and co-occurrence patterns among driver gene alterations (Kandoth et al. 2013; Sanchez-Vega et al. 2018). Many computational tools have been developed to identify those combinatorial patterns experimentally (i.e via CRISPR-Cas9 screens (Behan et al. 2019; Szlachta et al. 2018)) or computationally (Wu et al. 2015; Szczurek and Beerenwinkel 2014; Kim, Madan, and Przytycka 2017; Dao et al. 2017; Canisius, Martens, and Wessels 2016; Mina et al. 2017; Lee et al. 2018; Vandin, Upfal, and Raphael 2012).

Patterns of mutual exclusivity can arise from functional redundancy, context-specific dependencies (i.e tumor type or sub-type specific driving alterations), or synthetic lethality interactions. While functional redundancy has been used to reveal unknown functional interactions (Vandin, Upfal, and Raphael 2012), the synthetic lethality concept has been very successfully applied to the identification of novel therapeutic targets (Behan et al. 2019; Szlachta et al. 2018) or rational drug combinations (Szlachta et al. 2018), and to the prediction of drug response in cell lines (Szlachta et al. 2018) and patients (Lee et al. 2018).

Although less studied, driver co-occurrences are often interpreted as a sign of synergy and in some cases they have shown to be functionally relevant (Huun, Lonning, and Knappskog 2017; Ulz, Heitzer, and Speicher 2016; Sanchez-Vega et al. 2018; Lauber, Klink, and Seifert 2018; Tu et al. 2018; Dembla et al. 2018; Laroche-Clary et al. 2017). However, they have not yet been exploited for drug response prediction. With the methodology presented in this manuscript, we compared the mutational profiles of tumors that are sensitive or resistant to a certain drug to define Driver Co-Occurrence (DCO) networks, which capture both genomic structure and putative oncogenic synergy. We then used the DCO networks to train classifiers to identify the best possible treatment for each tumor based on its oncogenomic profile.

The development of tools for personalized treatment prioritization based on genomic profiles is an active field of research. Recently Al-Shahrour and colleagues presented PanDrugs, an in silico drug prescription tool that uses genomic information, pathway context and pharmacological evidence to prioritize the drug therapies that are most suitable for individual tumor profiles (Pineiro-Yanez et al. 2018). PanDrugs goes beyond the single-gene biomarker by taking into account the collective gene impact and pathway context of the oncogenic alterations identified in a given patient. However, it combines clinical evidence with *in vitro* drug screening data gathered from cancer cell line panels, which have limited clinical translatability (Bruna et al. 2016; Domcke et al. 2013; Gillet, Varma, and Gottesman 2013; Jaeger, Duran-Frigola, and Aloy 2015).

PANOPLY is another computational framework that uses machine learning and knowledge-driven network analysis approaches to predict patient-specific drug response from multiomics profiles (Kalari et al. 2018). This tool shows a great potential but the method strongly depends on whole genome and transcriptome patient data, which is not routinely acquired in clinical practice. Other methods like iCAGES have been developed mainly to identify patient-specific driver genes from somatic mutation profiles, which are later used to prioritize drug treatments (Dong et al. 2016). However, iCAGES only considers drugs that directly target the identified driver alterations based on current FDA prescription guidelines. All those methods rely on prior knowledge, which is incomplete and biased, and have not been conceived to identify novel co-occurrence patterns from the data and to exploit them for drug response prediction.

With the current implementation of TCT4U, we present a collection of drug-response predictive models for 53 treatments belonging to 20 drug classes, including targeted and more conventional chemotherapies. In a cross-validation setting, our drug-response models attained a global accuracy similar to that of approved biomarkers, but could be applied to twice as many samples, including drug classes for which no biomarker is currently available. Moreover, in an in vivo prospective validation, our models correctly predicted 12 out of 16 responses to 6 drugs tested on 15 tumors.

Obviously, our approach also suffers from some limitations. Due to the lack of systematic reporting of treatment history of the patients enrolled in genomic studies (J. Liu et al. 2018), it is difficult to match response to a drug with individual molecular profiles from clinical data. This practically impairs the systematic assessment of the prediction accuracy in patients for computational frameworks like TCT4U, PanDrugs(Pineiro-Yanez et al. 2018), PANOPLY (Kalari et al. 2018), iCAGES (Dong et al. 2016), or other in silico drug prescription tools such as the Cancer Genome Interpreter (Tamborero et al. 2018) or OncoKB (Chakravarty et al. 2017). Experimental validation of computational approaches is time-intensive and very expensive. Therefore, beyond the thorough experimental validation presented in this manuscript, only PanDrugs and PANOPLY predictions were experimentally validated, although on a single case study performed on a PDX model that was treated with 5 drugs (PanDrugs) or 2 drugs (PANOPLY).

Given the limited clinical representativity of drug screens performed on cell lines (Domcke et al. 2013; Gillet, Varma, and Gottesman 2013; Jaeger, Duran-Frigola, and Aloy 2015), we relied on patient-derived xenografts (PDXs) to implement our strategy and to identify biomarkers of drug response. Although PDXs have shown a good level of agreement with the course of disease evolution and treatment response observed in the tumors in the patient (Bruna et al. 2016; Byrne et al. 2017; Hidalgo et al. 2014; Krepler et al. 2017; Pompili et al. 2016; Izumchenko et al. 2017), they present some important drawbacks, such as the eventual loss of intratumoral heterogeneity (Villacorta-Martin, Craig, and Villanueva 2017; Eirew et al. 2015) or certain engraftment bias (Bruna et al. 2016; Willyard 2018).

Additionally, we have to consider that PDXs might not completely recapitulate the influence of the tissue of origin in tumors that have been implanted subcutaneously in immunodeficient mice and whose stroma has possibly regressed and/or been replaced by mouse stroma, altering thus their subclonal evolution and response to treatments (Hidalgo et al. 2014; M. Wang et al. 2018). However, our strategy can be readily adapted to derive drug-response models from continuous clinical outcome measures, such as progression free survival, which better represent the data acquired during routine clinical practice and in clinical trials. Indeed, we derived response models on a clinical cohort of breast cancer metastatic patients being treated with a combination of CDK4/6 and aromatase inhibitors, showing a good correlation with progression free survival.

Most importantly, TCT4U drug-response DCO networks are interpretable, and provide clear hints to identify the potential mechanisms of sensitivity or resistance present in each tumor. However, one key challenge in interpreting driver alteration co-occurrence patterns is that they can also emerge without necessarily being synergistic if a pair of genes is affected by a common mutagenic process. This commonly happens when several oncogenes are co-amplified as part of the same genomic region and our method already accounts for this. However, co-occurrence patterns can also emerge as a result of the exposure to other mutagenic processes that increase the mutational burden, the chromosomal instability, or that leave specific mutational signatures (Alexandrov et al. 2013; Canisius, Martens, and Wessels 2016; Dao et al. 2017).

Context or tumor type specific dependencies can also be a source of indirect associations with drug response. Although those confounding factors can obscure the biological interpretation of the DCO networks, they certainly provide valuable information for drug response prediction, especially in the case of ER status, which in most of the cases cannot be determined from somatic DNA alterations. Therefore, DCO networks are a valuable asset for hypothesis generation that need to be complemented with orthogonal sources of evidence, and functional validation will always be needed to demonstrate synergy. Indeed, we could find literature support for many of the candidate biomarkers identified, such as the loss of function of FAT1 and NF2 and their role in the development of resistance to CDK4/6 inhibitors (Li et al. 2018) (see Figure5, Figure6).

We also showed that our methodology is well suited to work with any custom gene panel, provided that the selected genes contribute to the differences in response to the drug being analyzed. As the cost of clinical molecular profiling continues to drop it is very likely that more types of data can be integratively analyzed to improve drug response prediction. However, in order to ensure the clinical translatability of our method in the short term, we decided to focus on well-supported oncogenic alterations that are readily detectable by cost effective methods in the clinical setting. We acknowledge that this is a very conservative decision and we accept that we might be missing biologically relevant information (i.e. non-coding alterations, methylation events or expression changes).

Indeed, current clinical biomarkers for patient stratification are mostly based on the detection of histopathological, cytogenetic and immunohistochemical changes that are not always detectable at DNA sequence level. For example, breast cancer patient stratification strategies based on ER/PR and ERBB2 status have proven to been very informative, both in terms of prognosis and response to treatment (Onitilo et al. 2009). Accordingly, TCT4U predictions should be regarded as a complementary source of information for clinical decision-making.

We believe that the computational framework presented, which goes beyond the single gene approach by exploiting co-occurrence patterns, could represent a significant advance towards the development of effective methods for personalized cancer treatment prioritization, with potential applications in population enrichment strategies in the context of adaptive clinical trials. Overall, our strategy represents an opportunity to accelerate the identification and validation of complex biomarkers with the potential to increase the impact of genomic profiling in precision oncology.

Materials and Methods

Genomic data processing

A total of 1,075 PDX models were established as part of a large pharmacogenomics screening that used the 'one animal per model per treatment' (1x1x1) experimental design to assess the population responses to 62 treatments (H. Gao et al. 2015). We collected somatic mutations and copy number alterations for 375 of them, and used the Cancer Genome Interpreter resource (Tamborero et al. 2018) to classify protein-coding somatic mutations and copy number variants into predicted passenger or known/predicted oncogenic.

In order to obtain comparable gene-wise oncogenic alteration rate estimates from a larger dataset of cancer patients, we downloaded the Catalog of Driver Mutations (2016.5), a curated dataset of known and predicted oncogenic coding mutations identified after analyzing 6,792 exomes of a Pan-Cancer cohort of 28 tumor types(Rubio-Perez et al. 2015). We complemented somatic driver mutations with copy number variation data for 4,058 patients representing 16 tumor types, accessed through cBioPortal (Cerami et al. 2012).

We considered as oncogenic the deletion (GISTIC score ≤ -2) of tumor suppressor genes and the amplification (GISTIC score ≥ 2) of oncogenes. The role of driver genes was established by inspecting the Catalog of Cancer Genes (Tamborero et al. 2018). In order to increase the clinical translatability, we subsampled both datasets to consider those oncogenic alterations covered by IMPACT410 (Cheng et al. 2015) or by Foundation Medicine (Frampton et al. 2013) targeted gene panels to obtain DCO networks and TCT4U models that could be directly used with those kind of molecular profiles, which are becoming widely used in the clinical setting.

Drug response data

In the original dataset, a total of 62 treatment groups were tested in 277 PDXs across six indications. Drug response was determined by analyzing the change in tumor volume with respect to the baseline along time. They combined two metrics (Best Response and Best Average Response) into a modified RECIST classification (mRECIST) with four classes: PD (progressive disease), SD (stable disease), PR (partial response) and CR (complete response).

For our analyses, we considered PDXs whose tumors progressed upon treatment (PD) as resistant, and PDXs whose tumors stopped growing (SD) or regressed (PR, CR) as sensitive. After applying this binary classification, we had to exclude 9 treatments for which there were less than 5 PDXs in one of the two response groups, lacking thus enough interindividual heterogeneity to model drug response. A total of 276 PDXs were treated in at least one of the 53 treatment groups considered, each treatment being tested in 29 to 246 animals, with a median of 43 (IQR: 38-93). We could obtain the molecular profile for 187 of them, which had been treated with a median of 18 (IQR: 14-20) drugs.

The final dataset consisted on 3,127 experiments performed on 187 PDXs and 53 treatment responses, across 5 tumor types: BRCA (breast cancer, n=38), CM (cutaneous melanoma, n=32), COREAD (colorectal carcinoma, n=51), NSCLC (non-small cell lung carcinoma, n=27), PAAD (pancreatic adenocarcinoma, n=38), and 1 PDX without tumor type annotation).

Molecular representativity of PDXs

We used the OncoGenomic Landscapes tool (Mateo et al. 2018) to obtain a 2D representation of the molecular heterogeneity of the 187 PDXs being analyzed, and compared it to that of large reference cohorts of cancer patients. We downloaded the precomputed 2D projections of the following reference cohorts from the OncoGenomic Landscapes webserver (oglandscapes.irbbarcelona.org): Pan-Cancer (n=15,212), BRCA (breast cancer, n=2,021), CM (cutaneous melanoma, n = 492), COREAD (colorectal carcinoma, n = 1,442), LUAD (lung adenocarcinoma, n = 1,486), LUSC (lung squamous cell carcinoma, n = 352), and PAAD (pancreatic adenocarcinoma, n = 442).

We selected the 2D coordinates of the subset of TCGA and MSKCC patients of each reference cohort and represented their distribution in the PanCancer landscape as a level plot using the 2D kernel density estimate function of the 'seaborn' python library with 20 levels and a custom heatmap colorscheme as background.

Drug response prediction based on Cancer bioMarkers database

We manually mapped the set of 53 drugs and drug combinations tested in the cohort of PDXs to the corresponding drug families in the Cancer bioMarkers database (Tamborero et al. 2018) using drug target information available in ChEMBL and DrugBank (see Supplementary Table S4). We successfully assigned 50 out of the 53 treatments, spanning 29 drug family annotations. We considered those genomic alterations showing a 'complete match' with any of the reported predictive biomarkers and collapsed them at gene level.

We considered as 'approved' biomarkers those ones that are currently approved by the FDA or by the main clinical guidelines in the field, such as the National Comprehensive Cancer Network (NCCN), the College of American Pathologists (CAP), the Clinical Pharmacogenetics Implementation Consortium (CPIC), or the European LeukemiaNet guidelines. We considered the rest of biomarkers, with varying supporting evidence, as 'experimental' biomarkers. The Cancer bioMarkers database usually reports more than one biomarker per drug or drug family, and often a single patient (or PDX) harbors several biomarkers of response and/or resistance for the same drug or drug family. We grouped response and resistance biomarkers at gene level and calculated the balanced accuracy (BAcc; average between sensitivity and specificity) of the prediction made by each gene in each treatment arm.

We weighted the binary predictions made by each gene and combined them to obtain a final prediction per treatment and PDX ($wComb_{bmk}$).

$$wComb_{bmk} = \sum_{i \in S} BAcc_i \cdot s_i - \sum_{j \in R} BAcc_j \cdot s_j \qquad (3.1)$$

S:Set of genes with or without predictive biomarkers of sensitivity (s_i,binary).

R:Set of genes with or without predictive biomarkers of resistance (s_j,binary).

$BAcc$: balanced accuracy of the predictive biomarker in a given treatment arm.

Driver Co-Occurrence (DCO) Networks

Differentially altered drivers (DiffD)

For each treatment, we aimed at identifying single gene biomarkers by selecting those driver genes with a significant differential alteration rate (DiffD) between sensitive and resistant PDXs. To this end, we compared the posterior probability distribution of the alteration rate of a given gene in sensitive versus resistant PDXs. We used a gene-specific informative prior based on the alteration rate observed in the cohort of 4,058 TCGA patients described above. In order to set a prior information contribution on the posterior inference to 5%, we set the effective population size of the prior to a 5% of the population size of the sample. These are the parameters of the beta posterior probability distribution of the alteration rate of a given gene in a given response group:

$$p(g_i|R) \sim Beta\left((K_R + (\frac{\alpha}{\alpha+\beta} \cdot \varepsilon \cdot n_R), n_R - K_R + (1 - \frac{\alpha}{\alpha+\beta} \cdot \varepsilon \cdot n_R)\right) \qquad (3.2)$$

$p(g_i|R)$:oncogenic alteration probability of gene g_i in the response group R

k_R: number of PDXs in response group R with alterations in g_i

n_R: number of PDXs in response group R

α, β: number of patients in TCGA with and without oncogenic alterations in g_i,respectively.

ε: constant representing the relative contribution of the prior to the posterior inference.

We obtained an empirical distribution of DiffD by sampling 10,000 times the sensitive and resistant alteration rate posterior probability distributions and then obtained the probability that DiffD differs from 0 (DiffD e-value). We repeated this procedure considering the whole treatment arm and separately for each of the two response groups in order to identify three sets of genes per treatment arm: (i) sens_DiffD are those genes with more than 95% probability of showing higher alteration rate in the sensitive PDXs, (ii) res_DiffD are those genes with more than 95% probability of showing higher alteration rate in the resistant PDXs, and (iii) global_DiffD are those genes with more than 95% probability of showing differential alteration rate between the two response groups. Additionally, we required that the selected genes were altered more than once in the corresponding group, with a minimum inferred alteration rate of 5%.

Driver Pairs (Ps)

To identify pairs of driver gene alterations occurring more often than expected in each response group of a given treatment arm, for each pair of co-altered drivers observed more than once in a given set of PDXs, we compared the observed probability of co-occurrence to the expected one under the independence assumption. To obtain the posterior probability distribution of the observed driver co-occurrence ($p(P_{ij}|R)$), we used a pair-specific informative prior based on the co-occurrence rate of this pair in the cohort of 4,058 TCGA patients, as described above. When this information was not available, we used a generic prior reflecting the average co-occurrence rate of any pair of drivers in TCGA. Again, we set a prior information contribution on the posterior inference to 5% by setting the effective population size of the prior to a 5% of the population size of the sample. These are the parameters of the beta posterior probability distribution of the co-occurrence rate of a given pair of gene alterations:

$$p(P_{ij}|R) \sim Beta\left((K_R + (\frac{\alpha}{\alpha+\beta} \cdot \varepsilon \cdot n_R), n_R - K_R + (1 - \frac{\alpha}{\alpha+\beta} \cdot \varepsilon \cdot n_R)\right) \tag{3.3}$$

$p(P_{i,j}|R)$: oncogenic co-alteration probability of pair of genes g_i-g_j in the response group R

k_R: number of PDXs in response group R with co-alterations in g_i-g_j

n_R: number of PDXs in response group R

α, β: number of patients in TCGA with and without oncogenic co-alterations in g_i-g_j,respectively.

ε: constant representing the relative contribution of the prior to the posterior inference.

To obtain the expected probability distribution of co-occurrence if genes g_i and g_j were independent, we sampled 10,000 times the posterior probability distribution of the alteration rate of each gene in the corresponding response group and computed their product ($p(g_i|R) \cdot p(g_j|R)$, see Eq. 2). We then obtained an empirical distribution of the difference between the observed and the expected co-occurrence rate ($p(P_{ij}|R) - p(g_i|R) \cdot p(g_j|R)$, see Eq. 3) and determined the probability that this difference was larger than 0 (Ps e-value).

We repeated this procedure considering the whole treatment arm and separately for each of the two response groups in order to identify three sets of co-occurring drivers per treatment arm: (i) sens_Ps are those pairs of drivers with more than 95% probability of co-occurrence in the sensitive PDXs, (ii) res_Ps are those pairs of drivers with more than 95% probability of co-occurrence in the resistant PDXs, and (iii) global_Ps, which are those pairs of drivers with more than 95% probability of co-occurrence in the whole treatment arm. Additionally, we required that the selected pairs were altered more than once in the corresponding group, with a minimum inferred alteration rate of 5%. In the case of sens_Ps and res_Ps, we additionally required that the estimated co-occurrence rate was larger in the response group being considered than in the other one.

Driver Co-Occurrence (DCO) networks

The differentially-altered drivers (global_DiffD, sens_DiffD, res_DiffD) and pairs of co-altered drivers (global_Ps, sens_Ps and res_Ps) can be expressed in terms of co-occurrence networks, in which nodes representing differentially altered driver genes (DiffD) or driver genes involved in a pair of co-altered drivers (DiP) are connected according to significant co-occurrences (Ps). For each treatment arm, we obtained three of such networks: (i) a global network (global_DCO), (ii) a sensitivity network (sens_DCO), and (iii) a resistance network (res_DCO).

Genome adjacency clustering

We analyzed the topology of the DCO networks to characterize the genomic features of large densely connected modules (i.e Supplementary Figure S1C). We downloaded the genomic coordinates of human genes from UCSC genome browser to assess whether genomic linkage was influencing the probability of co-occurrence. For each DCO network, we computed the pairwise mutual information content between any pair of genes as follows.

$$MI(A,B) = \sum_{i=1}^{|A|} \sum_{j=1}^{|B|} P(i,j) log \frac{P(i,j)}{P(i),P(j)} \tag{3.4}$$

$MI(A,B)$: mutual Information content calculated for the pair of genes A and B.
$P(i),P(j)$: probability that the status of gene A falls in class i in a PDX picked at random. Likewise for $P(j)$.
$P(i,j)$: probability that the status of genes A and B fall in classes i and j in a PDX picked at random.

We then represented the pairwise mutual information of all driver genes in the DCO network sorted by genomic coordinates and computed the Spearman's rank correlation between MI and physical distance in the genome for pairs of genes belonging to the same chromosome (i.e Supplementary

Figure S2A). We applied an unsupervised clustering algorithm based on pairwise mutual information relative to genomic distance for pairs of drivers located in the same chromosome. More specifically, we retrieved a similarity graph from each DCO network after connecting every pair of driver genes located in the same chromosome by an edge weighted as follows.

$$AdjClust(A, B) = \begin{cases} log\frac{MI(A,B)}{|TSS_a - TSS_b|}, & chr_a = chr_b \\ NaN, & chr_a \neq chr_b \end{cases} \tag{3.5}$$

$AdjClust(A, B)$:adjacency clustering metric for the pair of genes A and B

TSS_g: chromosomal coordinates of the Transcriptional Start Site of gene g

chr_g: chromosome where gene g is located.

Finally, we ran the MCL algorithm (Enright, Van Dongen, and Ouzounis 2002) with an inflation value of 2.5 and used the clusters with three or more driver genes for dimensionality reduction of the feature vectors describing the DCO networks.

Functional Analysis of the collection of DCO networks

We performed a functional analysis of the collection of DCO networks by calculating the enrichment of the 10 canonical cancer pathways identified and curated from the analysis of 9,125 samples from 33 cancer types(Sanchez-Vega et al. 2018). The pathways analyzed are: cell cycle ('CellCycle'), Hippo signaling ('HIPPO'), Myc signaling ('MYC'), Notch signaling ('NOTCH'), oxidative stress response/Nrf2 ('NRF2'), PI-3-Kinase signaling ('PI3K'), receptor-tyrosine kinase (RTK)/RAS/MAP-Kinase signaling ('RTK-RAS'), TGFβ signaling ('TGF-Beta'), p53 ('TP53') and β-catenin/Wnt signaling ('WNT'). Those pathways capture key genes that are recurrently altered in cancer and are, therefore relatively small and specific, involving a total of 334 genes and 3 to 85 genes per pathway.

We performed a Fisher's Exact test to assess whether the functions they represent were enriched or not among the set of biomarkers, the set of differentially altered drivers (DiffD), or the set of drivers in the DCO networks (DiffD_DiP) inferred for each treatment. We also checked whether each treatment was targeting a given pathway or not by mapping the drug target(s) of each treatment to the canonical cancer pathways. Finally, we performed a Fisher's exact test to assess whether the pathways that are enriched in each DCO network are also the pathways that are associated to the known mechanism of action of each treatment, in terms of drug targets.

TCT4U drug response classifiers

We described the DCO networks with a matrix of Boolean vectors (1: altered, 0: unaltered) encoding the alteration status of differentially altered drivers, drivers in co-occurring pairs, and pairs of drivers (DiffD_DiP_Ps) in each PDX. When needed, we adjusted for genomic linkage by reducing the dimensionality of the feature vectors and aggregating all drivers into clusters, which we considered to be altered when one or more of its constituent drivers were altered. We put together all those vectors in the form of a matrix and used it to train a Bernoulli Naïve Bayes (NB) classifier based on the observed responses to the treatment, also encoded as a Boolean vector (1: SD, PR, or CR; 0: PD). Please, note that we repeated the same procedure for each treatment arm with each of the three DCO networks described before (global_DiffD_DiP_Ps, sens_DiffD_DiP_Ps and res_DiffD_DiP_Ps). We assessed the accuracy and robustness of each of the three NB classifiers by performing an external leave-one-out cross validation (LOOCV) that involved both the inference of DCO networks and the prediction of drug response. We used the balanced accuracy of the LOOCV as weight to combine the global_DiffD_DiP_Ps, sens_DiffD_DiP_Ps and res_DiffD_DiP_Ps predictions generated for each drug-PDX pair into a final score, as described in Equation 5.

$$\begin{aligned} wComb =& BAcc_sNBC \cdot I_{sNBC=1} \cdot P_s \\ & - BAcc_{rNBC} \cdot I_{rNBC=1} \cdot P_r \\ & + BAcc_{NBC} \cdot (I_{NBC=1} \cdot P_s - I_{NBC=0} \cdot P_r) \end{aligned} \tag{3.6}$$

sNBC: binary prediction made by sens_DiffD_DiP_Ps NBC. A value of 1 indicates sensitivity.

rNBC: binary prediction made by res_DiffD_DiP_Ps NBC. A value of 1 indicates resistance.

NBC: binary prediction made by DiffD_DiP_Ps NBC. A value of 1 or 0 indicate sensitivity or resistance, respectively.

P_s,P_r: probability estimate for sensitivity or resistance,respectively.

$I_{NBC=j}$: indicator function that takes a value of 1 when NBC predicts class j or a value of 0 otherwise.

$BAcc_NBC$:Balanced accuracy of the NBC in the LOOCV.

Experimental validation in PDXs

We collected all the available molecular profiles of the VHIO collection of breast cancer PDXs. Most PDXs were profiled using a hybridization-based capture panel of 410 genes (IMPACT410) (Cheng et al. 2015). As we did for the training set, we used the Cancer Genome Interpreter resource (Bailey et al. 2018) in order to filter out as many passenger alterations as possible. In the same way we did for the LOOCV, we described the molecular profile of each PDX according to the DiffD_DiP_Ps feature vectors associated to each DCO network and used them to predict the response to the 53 treatments in the TCT4U collection. For each PDX, we ranked all treatments based on the predicted response and focused on the 10 highest-scoring predictions of sensitivity and resistance.

In order to increase the novelty of our findings, we excluded those predictions that were in agreement with predictions made by known predictive biomarkers. Then, we selected the 5 highest-scoring predictions of sensitivity and resistance per treatment. At this point, we had 51 novel, high-confidence predictions involving 32 PDXs and the following treatments: MEK inhibitor (n=15), Pi3K inhibitor (n=14), taxane (n=7), Pi3K inhibitor + CDK4/6 inhibitor (n=5), CDK4/6 inhibitor (n=5), and ER antagonist (n=5). We could recover raw experimental data for 10 drug-PDX pairs, including PDXs treated with Pi3K and/or CDK4/6, alone or in combination. We also found out that STG201 had already been reported to be resistant to Tamoxifen (ER antagonist) (70).

In order to cover the remaining treatment classes, we picked 6 additional drug-PDX pairs for experimental validation involving a MEK inhibitor (n=2), a taxane (n=2), and an ER antagonist (n=1). For each drug-PDX pair, 2 to 10 tumors were subcutaneously implanted in immunocompromised mice and grown until they reached a volume of 120-150 mm3. Tumors were treated with either vehicle or the corresponding drug or combination at a clinically relevant dose. Tumor growth was measured at least twice per week for approximately 20 to 40 days, when typically tumor volume in the control group had doubled twice or more.

Caliper measurements were converted into tumor volume estimates using the formula $(l \cdot w \cdot w) \cdot (\frac{}{6})$, where l and w are the major and minor tumor axes, respectively. The response was determined following the mRECIST guidelines that were used in the PDX screening that we used as training set (H. Gao et al. 2015). Basically, we calculated the percentage change in tumor volume from baseline $(\Delta Vol_t = \frac{V_t - V_i}{V_i} \cdot 100)$ and determined the BestResponse as the minimum value of Δ□Vol□_t after 10 or more days of treatment. In order to capture tumor growth dynamics, we also calculated the BestAverageResponse as the minimum value of $\frac{1}{n} \cdot \sum_{i=1}^{n} \Delta Vol_i$ after 10 or more days of treatment. PDXs were classified into response groups according to the mRECIST criteria applied in the following order:

CR: BestResp $\leq -95\% andBestAvgResp \leq -40\%$

$PR : BestResp \leq -50\% andBestAvgResp \leq -20\%$

$SD : BestResp \leq 35\% andBestAvgResp \leq 30\%$

$PD : BestResp \geq 35\% and BestAvgResp \geq 30\%$

Adaptation of TCT4U to use continuous clinical outcome measurements

We obtained both genomic and clinical data for a total of 116 patients with HR+/HER2- metastatic breast cancer that were treated with a CDK4/6 inhibitor in combination with an Aromatase Inhibitor in metastatic setting (Li et al. 2018). All patients underwent prospective clinical genomic profiling consisting on the identification of single nucleotide variants, small indels and copy number alterations detected from matched tumor-normal sequence data using the MSK-IMPACT targeted gene panels. We used the Cancer Genome Interpreter (Tamborero et al. 2018) to filter out passenger mutations and CNVs. and keep only known or predicted driver mutations or copy number alterations.

Detailed treatment history data was collected for each patient and included all lines of systemic therapy from the time of diagnosis of invasive carcinoma to the study data lock in September 2017. The exact regimen, as well as the dates of start and stop of therapy were also recorded. For the current analysis, we considered the treatment duration time as a measure of clinical benefit derived by patients whose biopsies were collected prior to or within the first 60 days of therapy initiation.

We used the TCT4U model of response to LEE011 to predicted response to CDK4/6 inhibition, as described before. Due to the differences in clinical outcome measurements between the training and the clinical cohort, we decided to adapt the TCT4U methodology to use continuous clinical outcome measurements as training set, instead of binary classification of drug response based on tumor growth. Our strategy consisted on comparing extreme populations both to derive the DCO networks and to train the classifier. We partitioned the population into three equally sized sets and applied the methodology described above. In this exercise, we set the cut-offs at 4.2 months and 9.7 months.

We selected as sens_DiffD or res_DiffD those genes with more than 95% probability of showing higher alteration rate in the one third of patients showing the most durable or shortest clinical benefit, respectively, compared to the third of patients at the other extreme of the distribution. Additionally, we selected as global_DiffD all those genes with more than 95% probability of showing differential alteration rate between the two extreme populations. The same strategy was applied in the identification of pairs of driver gene alterations occurring more often than expected considering all patients (global_Ps) or separately for the one third of patients that relapsed the latest (sens_Ps) or the earliest (res_Ps). The remaining steps were applied exactly as described for the binary TCT4U methodology. In this setting with only one treatment per patient, high confidence predictions were selected by optimizing the threshold of the global score to get a maximum false discovery rate of 30% in the LOOCV.

Supplementary Figures

S1.png S1.bb

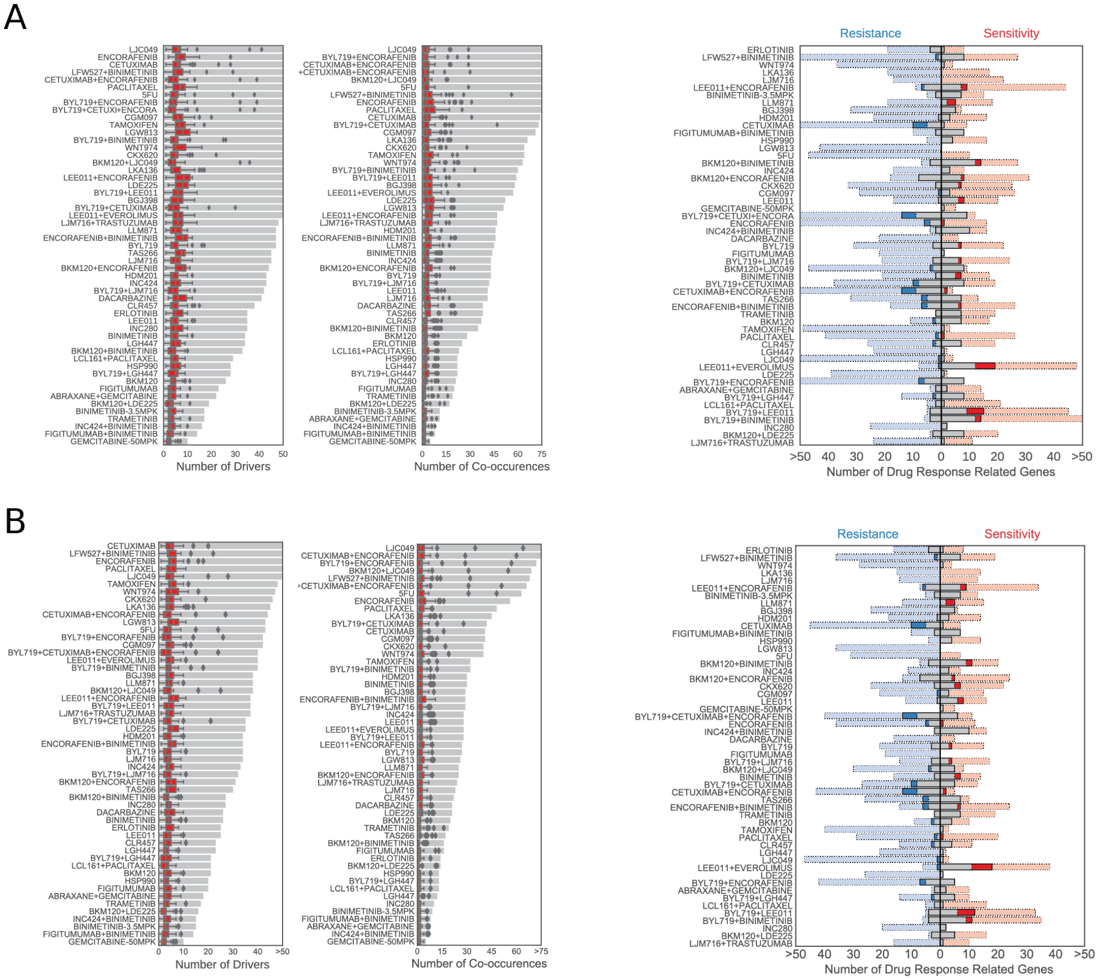


Figure S1: Characterization of the DCO networks derived from targeted gene panels. (A) DCO networks derived from IMPACT410 gene panel have 10 to 79 driver genes (median of 47, IQR: 34-57) and 5 to 230 pairs of drivers (median of 45, IQR: 25-64). Each PDX has a median of median of 5 altered drivers (IQR: 3-7) and 2 driver alteration co-occurrences (IQR: 1-4). (B) DCO networks derived from Foundation Medicine gene panel have 10 to 56 driver genes (median of 34, IQR: 23-42) and 5 to 85 pairs of drivers (median of 27, IQR: 16-40). Each PDX has a median of 4 altered drivers (IQR: 2-5) and 1 driver alteration co-occurrence (IQR: 0-3).

S2-02.png S2-02.bb

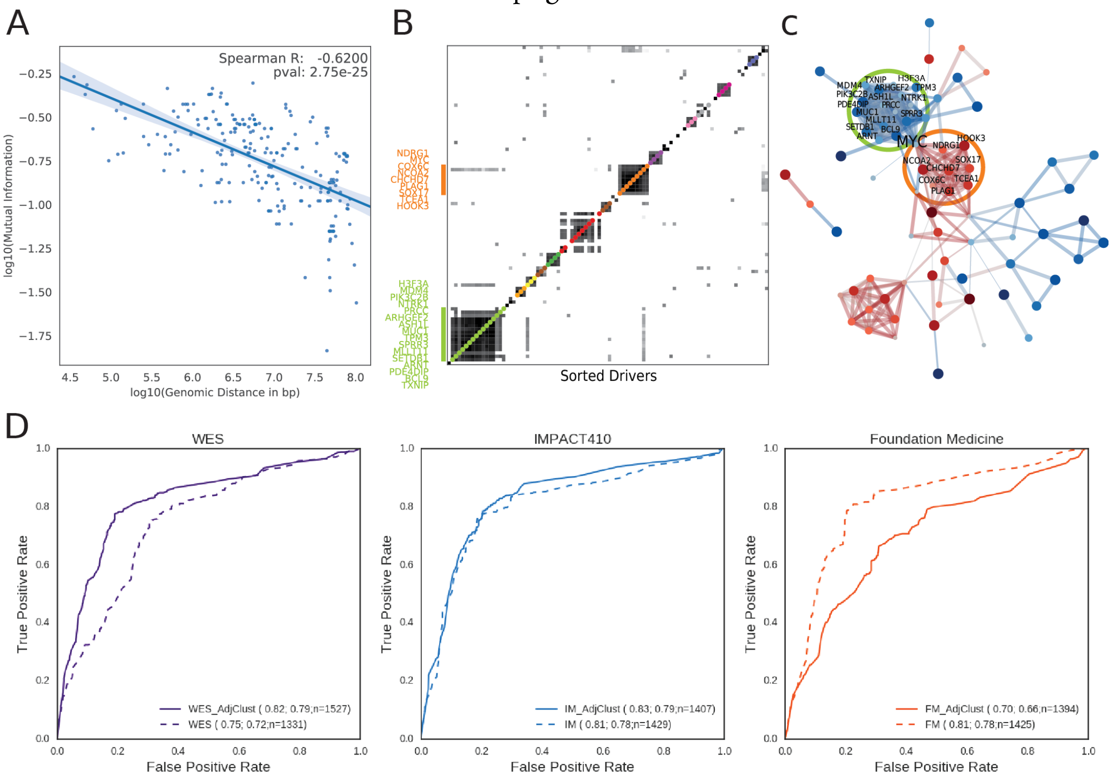


Figure S2: Genomic linkage in whole exome sequencing profiles. As an illustrative example, we show the genomic linkage of driver genes considered in binimetinib (MEK inhibitor) DCO network. (A) Correlation between pairwise mutual information content (MI) with respect to genomic distance of intra-chromosomal pairs of drivers. (B) Heatmap representation of mutual information content of pairs of drivers sorted by genomic coordinates. The different colors indicate the labels of a MCL clustering of adjacent drivers with high MI. Highlighted genes belong to two large clusters that tend to be co-altered with MYC. (C) Partial view of binimetinib DCO network, showing the pivotal role of MYC alteration, which is associated with sensitivity when co-altered with its adjacent genes in chr8 p11-12 and chr8 q11-24 (orange cluster), or with resistance when co-altered with a cluster of distant genes that are mostly located in chr1 q21-23 region. (D) Adjusting for genomic linkage improved the overall performance of the whole exome sequencing (WES) derived models, had very little impact on IMPACT410 derived models, and worsened the performance of Foundation Medicine derived models.

S3-01.png S3-01.bb

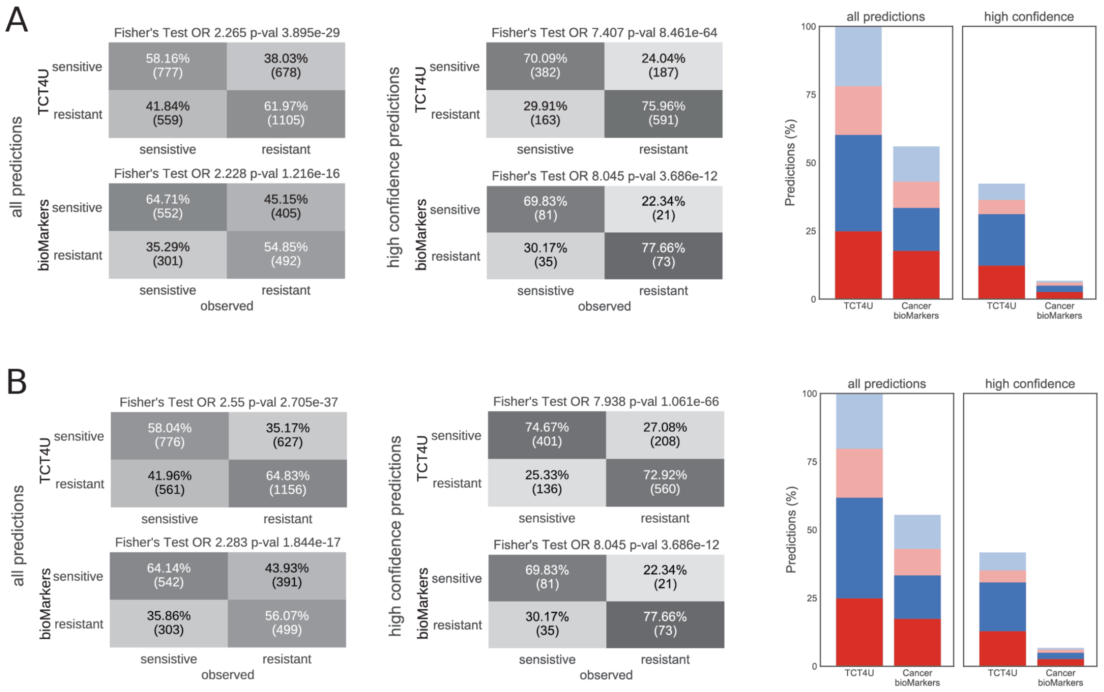


Figure S3: Evaluation of TCT4U models derived from the subset of alterations that would be detectable by two widely used targeted gene panels. (A) IMPACT410 (Cheng et al. 2015) and (B) Foundation Medicine (Frampton et al. 2013). The contingency tables show the association between the observed and the predicted drug responses. The precision and recall of each set of predictions is illustrated by the red and blue sections of the stacked bar plots, which represent the proportion of correct sensitivity and resistance predictions. Analogously, incorrect predictions are represented in faint colors. Missing predictions are represented in white to offer a comparative overview of the recall.

4

Patient-derived signatures for oncogenic pathway signaling

AUTHORS: Lidia Mateo, Miquel Duran-Frigola and Patrick Aloy

PUBLICATION: In preparation.

Introduction

Cancer is a genomic disease that is driven by evolutionary forces. During life somatic cells accumulate genomic alterations due to the exposure to endogenous and exogenous mutational processes (Alexandrov et al. 2013). Natural selection has constrained somatic evolution to try to maintain genomic stability and homeostasis. This has led to the emergence of tumor suppressor mechanisms but has also left some oncogenic vulnerabilities in our genomes (Merlo et al. 2006). For this reason, oncogenic alterations tend to converge into driver genes, altering cell signaling and conferring cancer cells the ability to divide and grow uncontrollably (Hanahan and Weinberg 2000; Hanahan and Weinberg 2011; Vogelstein, Papadopoulos, et al. 2013).

Evolutionary forces work on many levels in biology and selection can operate both at gene-level and pathway-level. Indeed, many driver genes that are only rarely mutated converge into key signaling pathways linked to proliferation and survival (Leiserson et al. 2015; Senft et al. 2017). Moreover, it has been shown that mutations in the same cancer pathway cause the same, or similar, cancer subtypes and associated clinical outcomes (Hofree et al. 2013). This rational motivated the aggregation of recurrent driver alterations in the TCGA PanCancer Atlas into higher-level structures, yielding as a result the systematic characterization and manual curation of a set of 10 well-known mitogenic signaling pathways: (1) cell cycle progression, (2) Hippo signaling, (3) Myc signaling, (4) Notch signaling, (5) NRF2 mediated response to oxidative stress, (6) PI3K signaling, (7) RTK/RAS_MAPK pathway, (8) TGF-beta signaling, (9) TP53 signaling, and (10) WNT signaling (Sanchez-Vega et al. 2018).

Knowing which are the main pathways that are driving tumor progression in each patient could have very important implications for risk assessment and therapeutic management in personalized oncology. Many computational methods and resources have been developed to distinguish patient-specific driver alterations from passenger DNA alterations (Chakra- varty et al. 2017; Dong et al. 2016; Tamborero et al. 2018). However, identifying the subset of likely driver alterations at DNA level might not be enough because there are many factors (i.e tumor microenvironment, pathway cross-talk, level of protein expression, etc.) that can severely impact on the phenotypic expression of a given genotype. To overcome this limitation, other approaches integrate somatic alterations with gene expression in order to identify patient-specific driver genes based on their estimated functional impact (Bertrand et al. 2015; Hou and Ma 2014).

Given the complex genetic architecture of cancer, methods operating at pathway-level should be more robust than those operating at gene-level. Indeed, the genetic perturbation of signaling pathways cause detectable transcriptional changes that capture both the direct effect of the perturbagen and the effect of downstream signaling (i.e posttranslational modifications). PROGENy (Pathway RespOnsive GENes,(Schubert et al. 2018)) exploited this concept to accurately infer pathway activity from gene expression in a variety of contexts, with a special focus on cancer. PROGENy is different from pathway expression mapping methods because the set of genes that respond to a pathway perturbation are not necessarily the pathway members themselves, but are often involved in biological processes downstream.

In the work we present, we brought the PROGENy methodology an step further and applied it to derive transcriptional signatures directly from 7,789 cancer patients, considering driver alterations as naturally occurring in vivo perturbations of the 10 oncogenic signaling pathways defined in the ”Pan-Cancer Atlas project” (Sanchez-Vega et al. 2018). With those signatures, we inferred patient-specific oncogenic pathway activities as a way of measuring the cancer endophenotype. We then explored how oncogenicity profiles correlate with the tissue of origin, aggressiveness and therapeutic vulnerabilities of the tumors, being able to bridge driver alterations occurring at DNA level with their clinical expression.

Results

A pan-cancer analysis of oncogenic pathway activity

To infer the patient-specific activity of 10 main oncogenic signaling pathways, we first identified a collection of 185 transcriptional signatures (Table S1) associated to the oncogenic alteration of each pathway within histological and molecularly homogeneous subpopulations of cancer patients. To this end, we aggregated the oncogenic alterations detected in 7,789 cancer patients at the level of pathways and applied the methodology described in PROGENy (Schubert et al. 2018) to select the 100 genes that were most strongly associated to the perturbation of each pathway in the context of a given tumor type and molecular subtype (Materials and Methods).

Next, we used the signatures to infer the oncogenic pathway activity in each patient, referred to as oncogenicity, by looking at the expression level of pathway-responsive genes. As expected, the inferred oncogenicity is able to discriminate between patients with and without oncogenic pathway alterations in a leave-one-out cross validation (LOOCV) experiment, attaining satisfactory AUROC (Area Under the Receiver Operating Characteristics curve) in most of the cases (median of 0.69, IQR: 0.61 - 0.79, see Table S2). However, not all pathways performed equally well. The median AUROC per pathway ranged from 0.62 in RTK_RAS to 0.91 in NRF2 (Table S2). Finally, we obtained a patient-specific pan-cancer oncogenicity score by summarizing the oncogenicity scores of each pathway across tumor types. In the score aggregation procedure, we considered the oncogenomic distance of each patient to the corresponding reference cohort, and the predictive value of each signature in the LOOCV (Materials and Methods).

Pathway oncogenicities reflects tumor type specific dependencies

We used TumorMap (Newton et al. 2017) to obtain a 2D representation of the pathway activation profile of the patient cohort. In this map, pairs of patients that are similar in terms pathway activity are positioned in close proximity to each other. When we colored each sample according to the tumor type we observed that the predominant activation of certain pathways was specific to certain tumor types, recapitulating well-known tumor type specific dependencies (Figure 1).

For example, the 10% of patients with highest MYC oncogenicity is strongly enriched in breast (22%; OR 3.61, p-value $1.39\cdot10^{-33}$, Figure 1B,C) and ovarian cancer patients (18%; OR 11.06, p-value $7.47\cdot10^{-69}$), followed by colorectal (9.5%; OR 2.39 p-value $2.11\cdot10^{-9}$) and stomach adenocarcinoma (9%; OR 3.00, p-value $1.40\cdot10^{-12}$). Our observations coincide with the tumor types that were recently reported to have frequent MYC alteration (Sanchez-Vega et al. 2018), specifically colorectal (17-52%), ovarian (40%), breast (29-39%), and esophagogastric (7-22%) cancer patients.

We observed similar trends for patients with high WNT oncogenicity, which are enriched in gastrointestinal cancer types. Both stomach adenocarcinoma and colon adenocarcinoma represent a 16% of those patients (OR 7.91, p-value $8.91\cdot10^{-52}$ and OR 5.16 p-value $1.06\cdot10^{-36}$, respectively, Figure1B,C). Again, this coincides with the ubiquitous activation of this pathway in colorectal and esophagogastric cancers, with up to 95% and 70% alteration rates reported in the MSI-POLE subtypes (Sanchez-Vega et al. 2018). The same trend is observed in patients with high RTK-RAS oncogenicity, enriched in pancreas adenocarcinomas (17.46%; OR 70.39 p-value $1.21\cdot10^{-116}$), which is one of the tumor types with the highest reported alteration rates (78%) (Sanchez-Vega et al. 2018).

Finally, patients with high NRF2 oncogenicity mainly represent lung squamous cell carcinoma (25%; OR 16.03, p-value $1.51 \cdot 10^{-112}$, Figure1B,C), liver hepatocellular carcinoma (19%; OR 9.08, p-value $4.73 \cdot 10^{-65}$) and head and neck squamous cell carcinoma (16%; OR 3.93, p-value $6.69 \cdot 10^{-29}$). In parallel with our observations, the highest alteration rates reported correspond to lung (25%), esophagogastric (23%), and head and neck (13%) squamous cell carcinomas, and lung adenocarcinoma (15%).

Pan-cancer pathway oncogenicity also recapitulates reported inter-pathway co-occurrences. For instance, p53 signaling and cell-cycle control, which were found to be frequently co-altered across multiple tumor types (Mina et al. 2017; Sanchez-Vega et al. 2018) also show a significant overlap in terms of oncogenicity (OR 3.84, p-value $3.23 \cdot 10^{-41}$) and occupy a shared region at the bottom right corner of the TumorMap (Figure 1D).

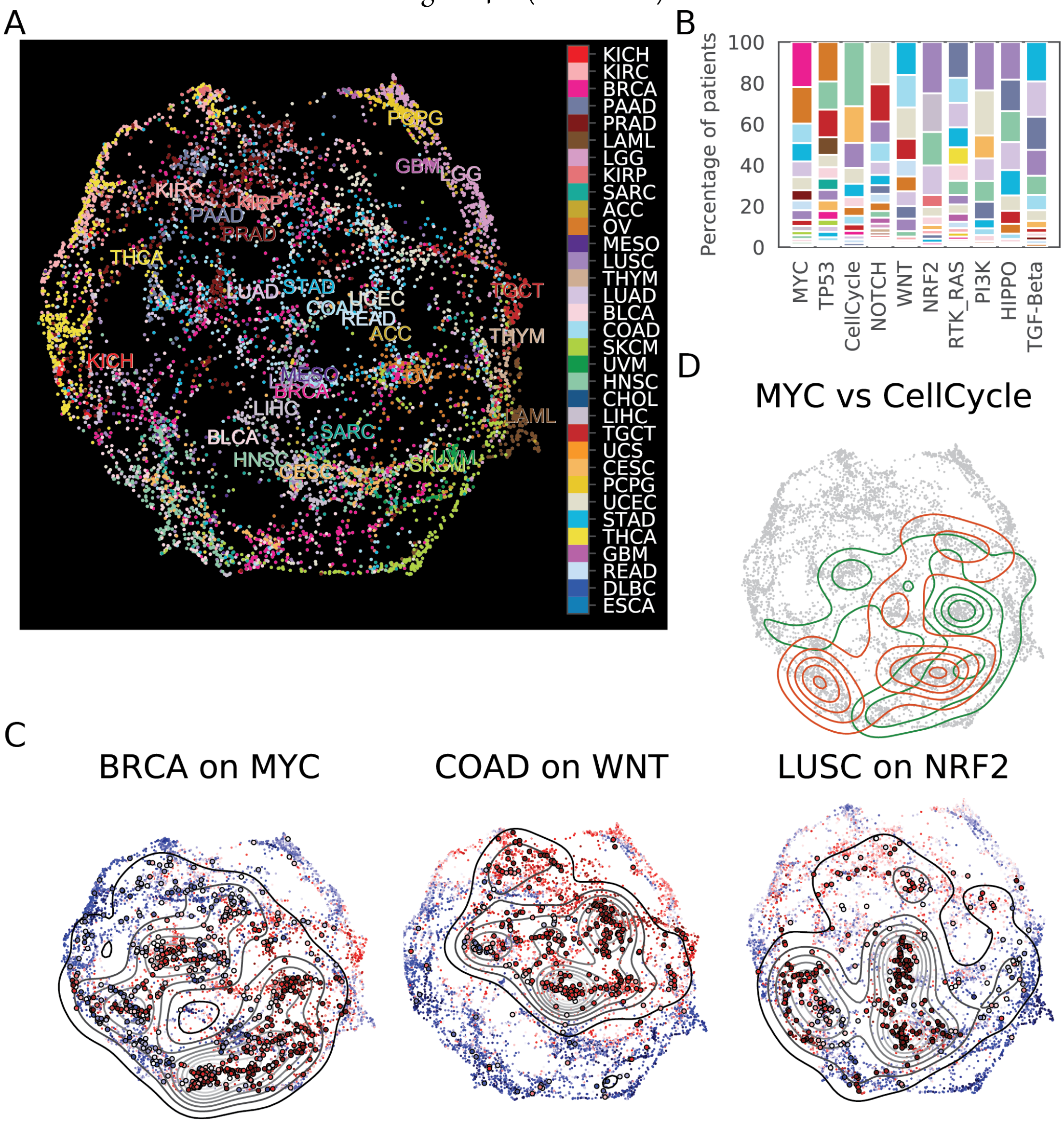
Prognostic significance of oncogenic pathway signaling

Determining which are the pathways that are driving cancer progression in each patient can carry important prognostic information. Typically, prognostic biomarkers are based on histopathological and immunohistochemical markers such as lymph-node infiltration (Engel et al. 2019) or ER/PR/HER2 status in breast cancer (Onitilo et al. 2009). Less is known about how somatic DNA alterations correlate with patient survival but, while driver gene mutations seem to carry very limited prognostic value, copy number alterations demonstrated stronger association with patient survival (Smith and Sheltzer 2018).

Pathway oncogenicity scores capture gene expression changes associated to pathway-level oncogenic DNA alterations and, therefore, they should be more robust than single gene DNA alterations, and easier to interpret than transcriptome-wide associations. To explore this hypothesis, we built a Cox univariate proportional hazards model linking the oncogenicity score of each pathway with the clinical outcome of patients separately for each tumor type. We then obtained a pan-cancer Z score across tumor types and compared it against the Z scores obtained similarly from pathway-level DNA alterations. Pathway oncogenicity proved to be more informative than pathway-level DNA alterations (Figure 2A,C). Consistent with previous findings, dysregulation of TP53 and CellCycle pathways were strongly correlated with poor prognosis, both when analyzing pathway-level DNA alterations (meta-Z score 9.94 and 8.58, p-value $2.64 \cdot 10^{-23}$ and $9.68 \cdot 10^{-18}$) and oncogenicity (meta-Z score 9.94 and 8.58, p-value $2.64 \cdot 10^{-23}$ and $9.68 \cdot 10^{-18}$,Figure2B).

Figure 4.1 *(following page)*: The oncogenicity TumorMap. (A) Patients with similar oncogenicity profiles are located close to each other in this two-dimensional projection of the pairwise similarity matrix. Each dot represents a tumor sample and it is colored according to its tissue of origin. Labels are located at the centroid of the patients of each tumor type. (B) Representation of each tumor type among the 10 of patients with the highest oncogenicity score per pathway. (C) The same oncogenicity TumorMap is used to represent the oncogenictiy score for MYC, WNT, and NRF2 pathways. Each patient is colored according to its inferred oncogenicity using a color scale that goes from blue (low) to red (high). The density plots highlight the distribution of breast (BRCA), colon (COAD) and lung squamous cell carcinoma patients (LUSC), which are individually represented with larger points.

Figure 4.1: (continued)



The oncogenicity of certain pathways only had prognostic value in specific tumor types (Figure 2C). For example, while TGF-Beta oncogenicity was not significantly correlated with survival in the pan-cancer analysis (meta-Z score 1.79, p-value 0.0745), it seems to be critical in renal clear cell carcinoma (KIRC, meta-Z score 4.54, p-value $5.59 \cdot 10^{-6}$). Our finding coincides with the established role of TGF-Beta signaling in the progression and aggressiveness of KIRC (Sitaram et al. 2016). The strongest tumor type specific correlation associates TP53 oncogenicity with increased risk of death of adrenocortical carcinoma patients (ACC, Z score 5.50, p-value $3.85 \cdot 10^{-8}$), which was also one of the strongest associations reported by PROGENy authors (Schubert et al. 2018).

Taken together, our results suggest that the oncogenicity of certain pathways is strongly associated with patient survival. Our resource, as well as other pathway expression mapping methods, might provide complementary information with potential applications in patient risk stratification.

Pathway oncogenicity correlates with drug efficacy in preclinical breast cancer models

Identifying pathway-level cancer dependencies in patients might help to identify therapeutic vulnerabilities, even in the absence of known biomarkers. To explore this possibility, we performed a correlative analysis of pathway oncogenicty and drug efficacy in the Breast Cancer PDTX Encyclopaedia (BCaPE). This pre-clinical drug-screening platform consists of 1,566 drug-tumor combinations, which measured the impact of 91 approved or experimental cancer treatments on the viability of 10-20 PDX-derived short-term cell line cultures (PDTCs) per drug.

For more than one third of the drugs (36 out of the 91), the efficacy of the treatment significantly correlated with the oncogenicity of at least one of the 10 pathways. We identified a total of 49 significant associations (Spearman's rank p-value <0.05, median |rho| 0.54 IQR: 0.51 - 0.65). Most of the drugs were associated to a single pathway (n=26), whereas some of them were associated to two pathways (n=8) or even three or four pathways, as was the case for 'AICAR' and 'PF-4708671' (see Figure 3, Table S4). On the other hand, the oncogenicity score of each of the 10 pathways was associated with the efficacy of at least one drug, with WNT pathway being associated to the largest number of drugs (n=15).

Overall, our analysis reveals that oncogenic signaling through these 10 important mitogenic pathways is associated with sensitivity in a high proportion of the cancer drugs screened in BCaPE.

Pan-AKT inhibition reverts breast cancer HIPPO signature in cell lines

So far, we have shown that patient-derived transcriptional signatures not only are detectable in breast cancer PDXs but they also correlate with drug response. The observed statistical associations do not necessarily imply mechanistic connections. However, if we could show that the exposure to a given drug induces a transcriptional response opposed to the oncogenicity signature, the most plausible hypothesis would be that the oncogenic alterations and the drug target are mechanistically connected. To explore this hypothesis, we used the Connectivity Map (Subramanian et al. 2017), which is a large catalog of transcriptional responses of human cells to chemical and genetic perturbations.

The strongest drug-signature correlation observed in pre-clinical breast cancer models links oncogenic HIPPO signaling with response to 'AKT inhibitor VIII', with a Spearman's rho 0.84 and p-value of $2.22 \cdot 10^{-3}$ (see Figure 3, Table S4). We queried CMap to obtain the list of perturbagens that induce a transcriptional response opposed to our HIPPO oncogenicity signatures, which would be reflected into negative connectivity scores. In order to do that, we had to narrow down our focus to the tumor type specific signatures contributing to the pan-cancer aggregated score. We found that only the HIPPO oncogenicity score obtained from luminal breast cancer patients' signature was significantly correlated with sensitivity to 'AKT inhibitor VIII' (BRCA_c19_HIPPO: Spearman's rho 0.77; p-value $9.22 \cdot 10^{-3}$).

In CMap, 'AKT inhibitor VIII' was one of the perturbagens that induced a transcriptional signature opposed to the BRCA_c19_HIPPO signature, with a connectivity score of -94.64. In general, CMap connectivity scores of +90 or higher, and of -90 or lower are considered strong scores. In addition to 'AKT inhibitor VIII', we found two additional pan-Akt inhibitors, MK-2206 and triciribine, that had been tested in the same cell line. Sensitivity to MK-2206 was also significantly associated BRCA_c19_HIPPO oncogenicity in BCaPE (Spearman's rho 0.60, p-value 5.45e-3) and it also showed a significant signature reversion in CMap (-98.99). On the other hand, triciribine was not tested in BCaPE and its enantiomers showed inconsistent transcriptional outcomes in CMap, with connectivity scores ranging from -57.53 to 94.11. In CMap, perturbagens that share the same mechanism of action or biological function have been grouped in perturbational classes to simplify the biological interpretation of the results. The compound classes that induced the strongest reversions of BRCA_c19_HIPPO signatures were 'DNA Replication LOF', 'FLT3 inhibitor', and 'JAK inhibitor', with connectivity scores of -99.9 or below. The complete list of chemical and genetic perturbagens that induced a transcriptional response opposed to that of BRCA_c19_HIPPO oncogenic signaling are reported in Table S5.

Although the involvement of HIPPO pathway in breast cancer is not yet fully understood, it has been recently reported that signaling through PI3K/PDK1/AKT pathway increases the nuclear translocation and activation of the HIPPO effector proteins YAP and TAZ (Zhao et al. 2018). In turn, the intranuclear accumulation of YAP and TAZ transcription factors has been associated to the induction of components of the cyclin D-CDK4/6 complex, specially CDK6 (Li et al. 2018), which was shown to promote cell cycle progression and resistance to CDK4/6 inhibitors in metastatic breast cancer patients. In this context, AKT blockage would uncouple PI3K from HIPPO mediated cell cycle progression, providing a plausible mechanistic support for this drug-signature association (Figure 4).

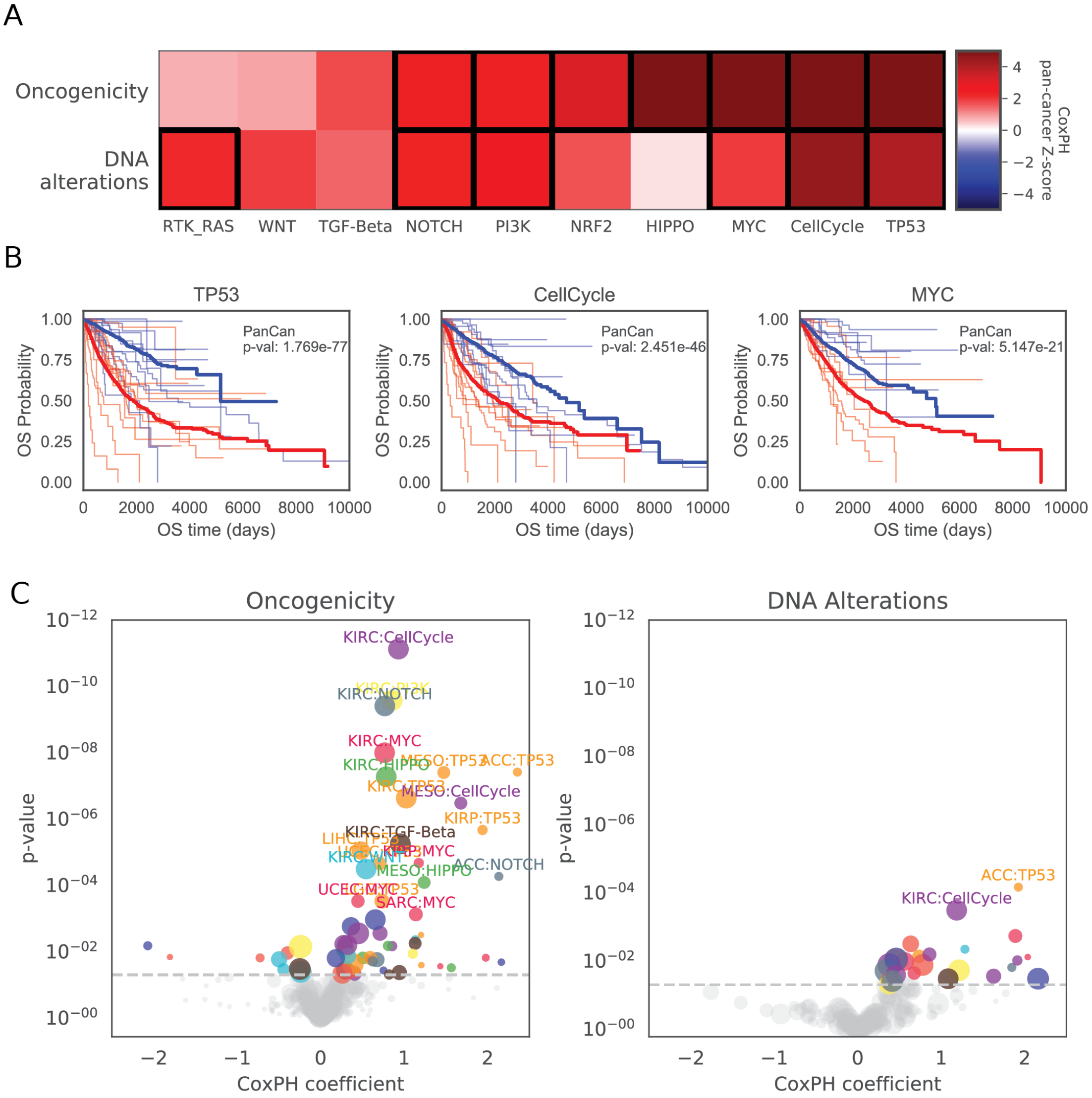
In agreement with the reported role of HIPPO signaling in CDK4/6 inhibitor resistance, we also found that HIPPO oncogenicity correlated negatively with response to Palbociclib (PD-0332991), although this association was only significant for the tumor type specific signatures derived from uterine and bladder cancer patients (UCEC_c08_HIPPO: rho -0.59, p-value 0.007 and BLCA_c27_HIPPO: rho -0.47, p-value 0.036). Our results provide mechanistic support for the combined inhibition of PI3K/AKT and CDK4/6, a drug combination that has already shown promising results in PIK3CA mutant breast cancer PDXs (Vora et al. 2014).

WNT signature reversion does not explain the growth-inhibitory effect of Embelin

The second strongest drug-signature correlation identified in breast cancer PDTCs links oncogenic WNT signaling with response to Embelin, with a Spearman's rho of 0.78 and p-value of $5.70 \cdot 10^{-4}$ (Figure 3 and Table S4). The two tumor-type specific WNT signatures that significantly correlated with response to Embelin were derived from adrenocortical carcinoma (ACC_c06_WNT: Spearman's rho 0.79; p-value $4.22 \cdot 10^{-4}$), and uterine corpus endometrial carcinoma patients (UCEC_c08_WNT: Spearman's rho 0.57; p-value 0.03). The WNT signature derived from breast cancer patients did not correlate with response to Embelin (BRCA_c19_WNT: Spearman's rho 0.37; p-value 0.169). In CMap, Embelin did not induce a transcriptional response significantly opposing to ACC_c06_WNT or UCEC_c08_WNT signatures (-59.99 and 90.54 connectivity scores) in MCF7 cell line. The connectivity score in the remaining reference cell lines ranged from -87.91 in PC3 to 35.76 in VCAP for ACC_c06_WNT, and from -88.93 in HT29 to 97.94 in A549 for UCEC_c08_WNT signature. Taken together, our results do not provide compelling functional evidence for the association between oncogenic WNT signaling and Embelin efficacy, at least in the context of breast cancer.

Figure 4.2 *(following page)*: . Implications for patient survival. (A) Comparison of the prognostic significance of inferred pathway oncogenicities with respect to pathway-level DNA alterations. The color scale represents the pan-cancer Z score corresponding to the Cox proportional hazards regression coefficients aggregated across tumor types. Statistically significant correlations are squared in black (p-value < 0.05). (B) Kaplan-Meier curves for the three pathways whose oncogenicity is most significantly associated with poor clinical outcome. Pan-cancer curves are represented with thick lines to distinguish them from the tumor type specific ones (thin lines).Patients were split according to their oncogenicity score into the top (red) and bottom (blue) quartiles and compared by a rank-sums test. (C) Vulcano plots representing the associations between pathway oncogenicity or pathway-level DNA alterations and overall survival. The x-axis represents the regression coefficients of Cox univariate proportional hazards models built separately for each pathway and tumor type. The color and size of the dots represent the pathway and the number of observations, respectively. The most significant findings (p-value <0.001) are labeled.

Figure 4.2: (continued)



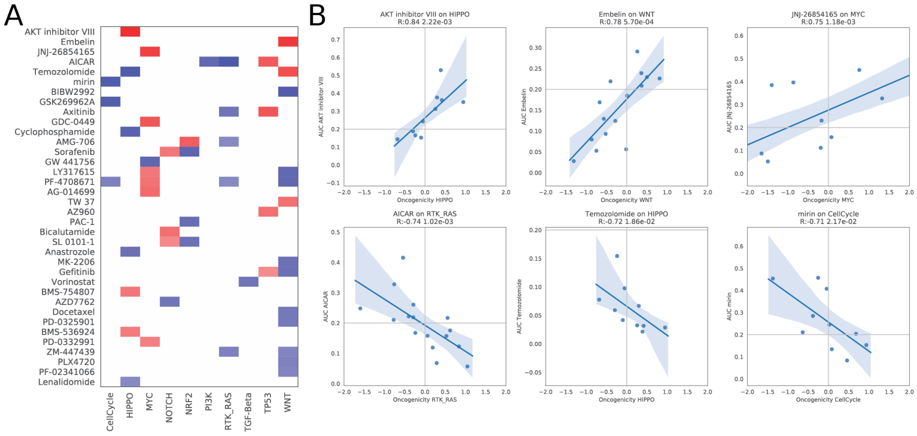


Figure 4.3: Correlative analysis of pathway oncogenicity and drug efficacy in preclinical models of breast cancer. (A) Heatmap representation of the 49 significant drug-pathway correlations identified across 10 pathways and 36 drugs. The color scale represents the strength of the association (Spearman's rho), which goes from -1 (blue) to 1 (red). (B) Lineal regression representation of the six strongest drug-pathway correlations.

Targeting MYC oncogenicity via its synthetic lethal interaction with DNA damage response

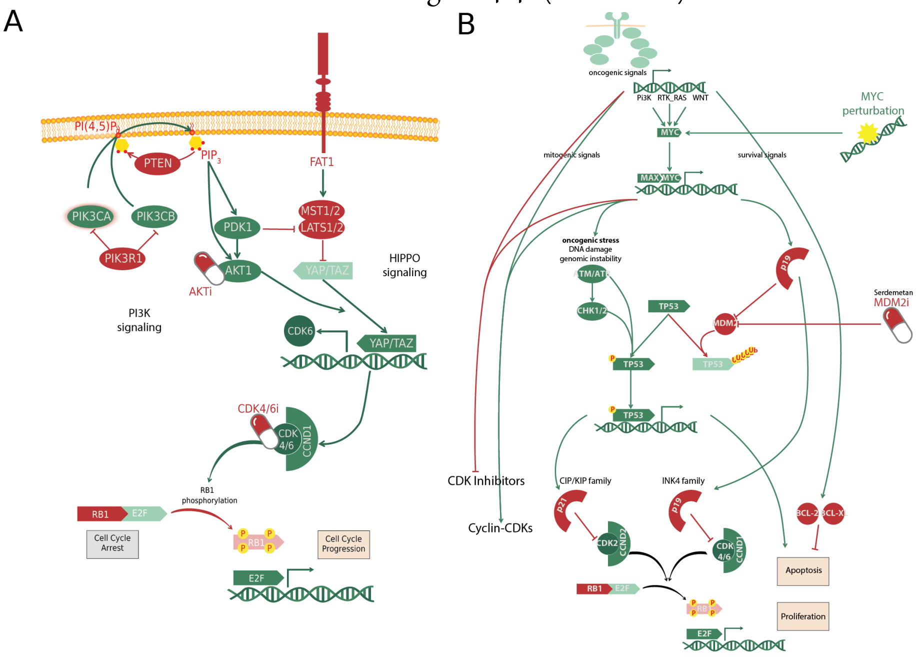
The third strongest drug-signature correlation was observed between oncogenic MYC signaling and response to 'JNJ-26854165' (Serdemetan), with a Spearman's rho of 0.75 and a p-value of $1.18 \cdot 10^{-3}$ (see Figure 3 and Table S4). MYC oncogenicity significantly correlated with response to 'Serdemetan' across signatures derived from several tumor types: colorectal adenocarcinoma (COADREAD_c04_MYC: Spearman's rho 0.81; p-value $2.19 \cdot 10^{-4}$ and COADREAD_c18_MYC: Spearman's rho 0.67; p-value $6.13 \cdot 10^{-3}$), breast cancer (BRCA_c19_MYC: Spearman's rho 0.77; p-value $6.90 \cdot 10^{-4}$ and BRCA_c17_MYC: Spearman's rho 0.73; p-value $1.91 \cdot 10^{-3}$), prostate cancer (PRAD_c16_MYC: Spearman's rho 0.77; p-value $6.90 \cdot 10^{-4}$), liver cancer (LIHC_c26_MYC: Spearman's rho 0.71; p-value $3.20 \cdot 10^{-3}$), ovarian cancer (OV_c06_MYC: Spearman's rho 0.58; p-value $2.38 \cdot 10^{-2}$) and lung adenocarcinoma patients (LUAD_c14_MYC: Spearman's rho 0.51; p-value 0.05).

The MYC signature that showed the strongest reversion in MCF7 cell lines treated with Serdemetan was BRCA_c19_MYC (-98.78), followed by COADREAD_c04_MYC (-91.10). We also observed more modest signature reversion scores for PRAD_c16_MYC (-82.58), and COADREAD_c18_MYC (-77.07), and no evidence for signature reversion in MCF7 for the remaining signatures (BRCA_c17_MYC: 0.00, LUAD_c14_MYC: 0.00, LIHC_c26_MYC: 97.86, and OV_c06_MYC: 99.96). In addition to 'Serdemetan', we found two additional MDM2 inhibitors ('HLI-373' and 'MDM2-inhibitor') and a MDM2 knock-out experiment that had been tested in this cell line. While the MDM2 knock-out showed a similar BRCA_c19_MYC signature reversion score (-93.02), the other two MDM2 inhibitors, which are chemically different from Serdemetan, did not revert BRCA_c19_MYC signature (11.85 and 72.77).

The compound classes that induced the strongest reversion of BRCA_c19_MYC signature are 'Structural maintenance of chromosomes proteins LOF', 'Norepinephrine reuptake inhibitor', and

Figure 4.4 *(following page)*: Hypothetical mechanistic connections underlying drug-signature correlations. (A) AKT inhibition mediates the suppression of HIPPO signaling. PI3K signaling positively regulate the HIPPO effector proteins YAP and TAZ (Zhao et al. 2018). The YAP and TAZ transcription factors induce the expression CDK6 (Li et al. 2018), which was shown to contribute to cell cycle progression and CDK4/6 inhibitor resistance. AKT blockage would uncouple PI3K from HIPPO mediated cell cycle progression. (B) MDM2 inhibition potentiates TP53-induced apoptosis in MYC-driven tumors. Mitogenic signals converge to increase energy production and macromolecule biosynthesis, induce cyclin-dependent kinases (CDKs), and the repress CDK inhibitors to prepare cells for S-phase entry (H. Liu et al. 2015). They can also trigger survival signals that activate antiapoptotic proteins (BCL/BAX). The MYC transcription factor integrates and responds to these mitogenic signals by orchestrating a broad transcriptional response that promotes both proliferation and apoptosis. Several mechanisms have been proposed to couple deregulated MYC expression with TP53-induced apoptosis. On one hand, MYC signaling causes genomic instability and DNA damage, leading to the activation of the ATM/ATR axes and their checkpoint effectors (CHK1/2) (Pusapati et al. 2006). In response to genotoxic stress, ATM/ATR axes activate p53 signaling both directly, by p53 N-terminal phosphorylation, and indirectly, through MDM2 phophorylation (Pusapati et al. 2006). On the other hand, MYC induces the expression of the CDK inhibitor p19ARF (CDKN2A), which in turn binds to and inhibits MDM2 and stabilizes TP53 (Pusapati et al. 2006; Phesse et al. 2014). In this context, the pharmacological inhibition of MDM2 with Serdemetan would potentiate the endogenous apoptotic role of MYC.

Figure 4.4: (continued)



’Thymidylate synthase inhibitor’, with connectivity scores of -99.83 or below. The compound showing a strongest signature reversion score is Sunitinib (-100), which is a pan-RTK inhibitor. The complete list of chemical and genetic perturbagens that induce a transcriptional signature opposed to BRCA_c19_MYC oncogenic signaling can be found in Supplementary Table S6.

It is noteworthy that Serdemetan showed even stronger COADREAD_c04_MYC and BRCA_c19_MYC signature reversion scores in A375 colorectal cancer cell line (-99.94, and -99.78), suggesting that MDM2 inhibition might represent a shared opportunity for MYC actionability in BRCA and COADREAD tumors. MYC signaling plays a key central role in cancer because it integrates a plethora of mitogenic and developmental signals and orchestrates the broad transcriptional response that is required to support cell growth and proliferation (H. Liu et al. 2015; Schaub et al. 2018). However, MYC overexpression also triggers tumor suppressor mechanisms that promote apoptosis (Pusapati et al. 2006; Phesse et al. 2014; Jamerson, Johnson, and Dickson 2000). Several mechanisms have been proposed to couple deregulated MYC expression with TP53-induced apoptosis and one of them depends on the indirect inhibition of MDM2, which stabilizes TP53 (Pusapati et al. 2006; Phesse et al. 2014). The pharmacological inhibition of MDM2 with Serdemetan would potentiate this endogenous apoptotic role of MYC, which would explain why tumors with higher MYC oncogenicity experience stronger tumor growth inhibition (see Figure 3, Table S6).

Materials and Methods

Pathway perturbation data

We evaluated 9,487 samples across 32 cancer types from the TCGA PanCancer Atlas for which we could obtain both whole-exome sequencing data and gene expression levels (RNA-Seq). In order to account for molecular heterogeneity we further stratified those samples into distinct subtypes by taking into account both the histological and molecular identity assigned by the integrative clustering algorithm (iCluster) (Mo et al. 2013; Hoadley, Yau, Hinoue, et al. 2018). For each tumor type, we only considered the samples that grouped together in any of the dominant clusters (i.e clusters that covered at least 50 samples and representing 10% or more of the total). This resulted in a set of 7,789 samples from 37 distinct histological and molecular subtypes.

We used the Cancer Genome Interpreter (Tamborero et al. 2018) to filter out passenger mutations and only worked with driver somatic mutations and copy number alterations. We mapped driver alterations to the 10 oncogenic signaling pathways defined by the PanCancer Atlas project (Sanchez-Vega et al. 2018). A pathway was considered to be perturbed in a sample if one or more genes in that pathway contained an oncogenic alteration. Pathways that were perturbed in more than 10 and less than two thirds of patients per tumor type were considered for further analyses, representing a total of 185 context specific pathway perturbations. With this information, we built pathway perturbation matrix (P) where each pathway has a 1 if it is perturbed in a patient and 0 otherwise.

Multiple linear regression modeling

We adapted the methodology described in PROGENy (Schubert et al. 2018) to extract transcriptional signatures for oncogenic pathway perturbations from cancer patients. To this end, we obtained the transcriptional profile of the 7,789 samples considered from the Broad GDAC (Center 2016) (RNASeq Level 3, 2016_01_28 release). In order to increase the robustness of our signatures, we focused on genes that were expressed across all tumor subtypes. To do this, we fitted a Gaussian Mixture model (k=5) for each tumor type with the average basal gene expression (labeled as 'mRNAseq_RSEM_normalized_log2' in GDAC and herein referred as to bG). We filtered out the genes assigned to the first component, containing non- or very lowly expressed genes in a given tumor type, and worked with the list of 16,010 genes that were not excluded in any tumor type.

To extract oncogenicity transcriptional signatures we used the gene expression Z-score (labeled as 'mRNAseq_RSEM_Z_score' in GDAC and herein referred as to zG), which represents the relative expression of a given gene in a sample with respect to the reference population. This reference population is either the rest of tumors of the same tissue that are diploid for that gene, or, when available, normal adjacent tissue. We used the generalized least squares model from StatsModels (Seabold and Perktold 2010) to built a linear regression model for each gene and tumor subtype separately, with pathway perturbations as explanatory variable and gene expression Z-score as response variable.

For each pathway in a given tumor subtype, we sorted all genes according to the strength of their association with pathway perturbation. We picked the 100 genes with the smallest p-value and largest absolute Beta coefficient (>=0.25) as the most significant ones. As a result, we obtained a matrix of coefficients (B) where each gene (in rows) has the corresponding Beta correlation coefficient if it responds to the perturbation of a pathway or a 0 otherwise.

Patient-specific pathway oncogenicity score

We obtained patient-specific oncogenicity scores by calculating the scalar product of the basal gene expression profile of that patient (Gi) and the vector of coefficients (Bp) of the corresponding pathway. This can be efficiently done as a matrix multiplication between the gene expression matrix (bG) and the Beta coefficients matrix (B). Next, we centered and scaled the pathway oncogenicity score of each tumor type to have a mean of zero and a standard deviation of one. This enables the comparison of scores across pathways and samples and would be conceptually analogous to performing an average of the gene expression Z-scores of a given patient for a given gene signature, weighted by the basal level of expression and the strength of the association of each gene with pathway perturbation.

For the leave-one-out cross validation (LOOCV) experiment, we set aside one patient, recalculated the lineal regression models with the reminder patients, and used the resulting Beta coefficients matrix to infer the oncogenicity profile for the test patient. We quantified the AUROC (Area Under the Receiver Operating Characteristics curve) to assess whether the inferred oncogenicities were able to discriminate between patients with and without pathway perturbations.

Pan-cancer oncogenicity score aggregation

For each patient, we inferred a total of 185 context specific pathway oncogenicities, comprising at least one of the 10 initial pathways across 37 molecularly homogeneous tumor subtypes. We obtained a pan-cancer oncogenicity score for each of the 10 pathways by aggregating the context specific oncogenicity scores across tumor subtypes. The simplest strategy would have been to obtain the average context specific oncogenicity per pathway. However, this strategy would dilute tumor type dependencies that might be biologically relevant.

To account for this, we weighted the context specific oncogenicity scores by the similarity of each patient to the reference population of each tumor subtype, in terms of basal gene expression. In more detail, we first obtained all pairwise correlation distances between individual patients and the centroids of the 37 histological and molecularly defined subtypes. We then obtained a similarity score per patient and tumor subtype by calculating the fraction of centroids with a similarity less than or equal to that of the current centroid. The resulting score ranges from 0 to 1 and reflects how different or similar a patient is with respect to the cohort of patients that was used to derive each transcriptional signature.

Additionally, we considered the predictive value of each signature in the LOOCV. We rescaled the AUROC to the interval $[-1, 1]$ with the transformation $2 \cdot (AUROC - 0.5)$ and truncated it to the range $[0, 1]$ to disregard the inferences made by models with an AUROC lower than 0.5. To obtain the pan-cancer oncogenicity score, we multiplied the context specific oncogenicity scores by their corresponding transcriptional similarity scores and scaled AUROC and finally averaged them out across tumor types.

Oncogenicity TumorMap

We used the UCSC TumorMap portal (Newton et al. 2017) to generate a custom two-dimensional representation of pairwise patient similarity based on pan-cancer oncogenicity profiles. We uploaded a 'Feature data' matrix with the pathway activation profiles and downloaded the output file with the XY-coordinates of the patients before applying the hexagonal binning process (xyPreSquiggle.tsv). We also downloaded the tumor type classification and color scheme to enable the visual comparison with previously published TumorMaps (Hoadley, Yau, Hinoue, et al. 2018).

For each pathway, we used Fisher's exact test to determine which tumor types were enriched among the 10% of samples with highest oncogenicity score. To illustrate tumor type dependencies, we colored samples according to the inferred oncogenicity score of a given pathway and, on top of that, we represented the distribution of patients of a given tumor type. More specifically, we used the kdeplot function with 10 levels to visualize the 2D kernel density estimate of the samples locations. We used the same approach to represent the overlap between the 10% of patients with the highest MYC and CellCycle oncogenicity scores.

Survival analysis

We used lifelines Davidson-(Davidson-Pilon 2019) python library to build a Cox univariate proportional hazards model separately for each pathway and tumor type. We obtained the overall survival time (OS) from the TCGA Pan-Cancer Clinical Data Resource (J. Liu et al. 2018). As previously reported (Smith and Sheltzer 2018), we used Stouffer's method to obtain a single Z score per pathway after aggregating the Z scores obtained from each cancer type. The application of Stouffer's method consisted on dividing the sum of the tumor type specific Z scores by the square root of the number of cancer types for each pathway (Stouffer et al. 1949).

Cox proportional hazards regression is suitable to work both with continuous and discrete input data, allowing us to analyze both the association of pathway-level DNA alterations (discrete) and oncogenicity (continuous) with patient survival. Kapplan-Meier plots and the associated log-rank p-values were also generated with lifelines with illustrative purposes only.

Correlative analysis of pathway oncogenicity and drug sensitivity in BCaPE

The BCaPE consists of 84 molecularly annotated breast cancer patient derived xenografts (PDXs). Analogously to what we did with patients, we used the 185 context specific pathway oncogenicity signatures to infer PDX-specific oncogenicity scores. We calculated the scalar product of the normalized basal gene expression profile of a given PDX (bGi) and the vector of coefficients of each pathway (Bp). We centered and scaled the oncogenicity score of each pathway to have a mean of zero and a standard deviation of one. We then aggregated context specific pathway oncogenicities into a pan-cancer oncogenicity score per pathway as described above. To obtain the transcriptional similarity scores we obtained a matrix of pairwise correlation distances between the 84 PDXs and the 37 centroids of the reference patient populations, based on the expression of the 14,259 genes that there were in common. In BCaPE, a selection of 20 PDXs were used to establish PDX-derived short-term cell line cultures (PDTCs) and were used as a pre-clinical drug-screening platform. A total of 91 approved or experimental cancer treatments were tested on 10–20 PDTC models. For each drug, we measured the Spearman's rank correlation between the area under the dose response curve (AUC) and both the pan-cancer and the tumor-type specific oncogenicity score of the 10 pathways.

Signature reversion in Connectivity Map

We used the CLUE analysis environment (https://clue.io) to query the Connectivity Map reference dataset (CMap Touchstone) for chemical or genetic perturbagens that induce a transcriptional response opposed to patient-derived oncogenicity signatures. Out of the 100 genes per signature, we only considered the ones belonging to the set of Best Inferred Genes (BING), which are >10,000 genes whose expression can be safely inferred from the 978 landmark genes measured in CMap. We obtained a list of perturbagens rank-ordered by the CMap connectivity score (tau).

We were specially interested in perturbagens with a strong negative connectivity, indicating that the perturbagen's signature and a given patient-derived oncogenicity signature are opposing (ie. genes that are upregulated in patients with a given pathway perturbation are downregulated by treatment with the perturbagen and vice versa). The connectivity score ranges from -100 to 100, with scores $\geq$90 and $\leq$-90 being generally considered as strong enough for hypotheses generation.

General Discussion

The main objective pursued in this thesis was the development of a computational strategy to tailor a cancer patient's treatment to the specific molecular alterations of their tumor (Targeted Cancer Therapy For You or TCT4U, chapter 3). To aid in this process, we also developed two accompanying tools that enable the contextualization and visualization of the molecular portraits of individual patient tumors (Cancer PanorOmics, chapter 1) and cohorts of patients (Oncogenomic Landscapes, chapter 2). We believe that those tools are already a valuable asset for clinical researchers and oncologists willing to interpret and communicate molecular findings to their colleagues or, even more importantly, to their patients.

We have concluded the development phase and future efforts should be directed to increase their online visibility. This could be achieved by investing in web marketing and positioning services that could help us meet the precise needs of our intended users, or we could even offer them training sessions and tailored solutions. Unfortunately, those solutions are out of reach in the context of this thesis. An alternative approach could be the integration into more comprehensive clinical-genomics curation pipelines (i.e. in the context of The Variant Interpretation for Cancer Consortium, https://cancervariants.org/).

Another important aspect of the whole thesis was the tradeoff between complexity and biological interpretability. It was clear that we needed to go beyond the very few FDA approved single-gene biomarkers. However, we were not willing to do so at the expenses of considering variants without a clear implication in cancer. In order to avoid developing "black box" predictors, we decided to focus on combinations of driver alterations. This way, we reached a compromise between complexity and interpretability. We also decided to aggregate alterations at gene level, exploiting the fact that oncogenic alterations in a given driver gene are very likely to exert a common endophenotype (i.e hyperactivation of an oncogene or inactivation of a tumor suppressor). However, it is well known that the outcome of mutations in different regions of the same protein might be radically different. The loss of sub-gene resolution is the price we have to pay for the increased statistical power needed to detect genotype-phenotype associations. This is crucial when dealing with limited sample sizes of genomically diverse subgroups of patients or PDXs.

In order to ensure the direct clinical translatability of our tools, we decided to focus on alterations that are readily detectable using affordable state of the art theranostics, such as targeted gene panels or, at most, whole exome sequencing. We know that we might be missing valuable information coming from expression changes, epigenetic marks, post-translational modifications, or non-coding mutations, among many others. However, we prioritized the potential clinical translatability and decided to do the most we could do with the data that is currently feasible to acquire in the clinical setting.

All those *ab initio* considerations strongly determined the choices I have made during the development of this thesis, and reflect the challenges that translational research will have to face if genomic "big data" is to improve patient outcomes.

Concluding Remarks

In this thesis we presented four computational developments for the integration, analysis, and visualization of oncogenomic profiles. We have also extracted genotype-phenotype associations related to drug response and oncogenic pathway signaling, with the aim to improve patient stratification in personalized medicine. As a result, we have reached the following conclusions:

- Using PanorOmics, we can get an integrative visualization of the genomic alterations detected in a cancer patient. This allows for the detection of recurrent alterations in driver genes and larger structural variants of different amplitudes.
- The molecular context of provides important information to identify those mutations that are more likely to interfere with the protein fold, catalytic activity, or with its capacity to interact with important partners. Moreover, it can reveal therapeutic opportunities operating through its functional neighborhood.
- With OncoGenomic Landscapes, we can obtain global representations of tumor genome similarities within and across cohorts of patients or cancer models. Those comparisons can be used to assess the molecular representativity of one with respect to the other, or to characterize the

genomic architecture of subpopulations of patients with specific clinico-pathological features (i.e metastatic or drug resistant tumors).

- We used driver genes as landmarks for the interpretation of the OncoGenomic Landscapes, reaching a reasonable compromise between complexity and biological interpretability.
- We showed that patient's tumor similarity with respect to other successfully engrafted tumors is an indicator of poor prognosis. We believe that oncogenomic similarity is capturing the genomic basis of tumor aggressivenes and the resulting engraftment bias.
- With Targeted Cancer Therapy For You (TCT4U), we identified recurrent driver co-occurrences that reflected both genomic structure and putative synergistic interactions, such as the resistance associated co-alteration of TP53-CCND2 in ribociclib, or the PIK3C2B-PIK3CA in alpelisib.
- We exploited Driver Co-Ocurrence networks to predict treatment outcome in PDXs and also in patients. In a cross-validation setting, our drug-response models attained a global accuracy similar to that of approved biomarkers, but could be applied to twice as many samples, including drug classes for which no biomarker is currently available.
- Finally, I presented a collection of patient-derived transcriptional signatures for monitoring the activity of 10 oncogenic signaling pathways (oncogenicity). We show that patient-specific oncogenicity profile captures well known tumor type dependencies and carries important prognostic information.
- We propose a strategy to identify drugs targeting pathway-level dependencies in preclinical models. This approach might help to identify therapeutic vulnerabilities, even in the absence of known biomarkers.
- Contextualizing individual mutations and patients in PanorOmics and OncoGenomic Landscapes can be a first step to identify treatment opportunities for patients that ran out of stan-

dard therapeutic options. We can complement this information With TCT4U predictions, based on driver co-occurrences, and with the potential to revert oncogenic signatures. With all this, our toolkit contributes to expand the applicability domain of precision oncology.

Future perspectives

We developed the Targeted Cancer Therapy for You (TCT4U) methodology (chapter 3) with the expectation of applying it to large-scale repositories of cancer patient's molecular profiles with precisely annotated clinical outcomes and associated treatment history. Unfortunately, the annotation of the clinical outcome of patients treated with targeted agents is very scarce in currently available cohorts of profiled patients.

We are aware that annotation and validation of clinical and epidemiologic data remain expensive and time consuming (Clinical Cancer Genome Task Team of the Global Alliance for et al. 2017), but we still hope that the promise made by the National Cancer Institute (NCI) becomes a reality. The day that the genomic and clinical data from the NCI Molecular Analysis for Therapy Choice (MATCH) becomes available we will have an unprecedented opportunity to obtain predictive models with direct clinical applicability.

On the other hand, I would like to expand the applicability domain of our methodological developments in order to investigate other human diseases that are less studied. The oncogenic pathway signatures obtained in the latter work (chapter 4) represent a perfect opportunity to transfer the knowledge extracted from richly annotated cancer patients to other diseases caused by the dysregulation of the same fundamental mitogenic pathways.

Molecular profiles are becoming cheaper and easier to acquire than ever before, and novel technological improvements are continuously appearing. It will only be when molecular profiling is coupled to high-quality clinical data collection, sharing, and harmonization that genomic 'Big Data' will revolutionize health care. I hope that my future work and the work of my colleagues will contribute to realize our common goal, which should be transitioning from cancer diagnosis to molecular discovery to patient recovery (Clinical Cancer Genome Task Team of the Global Alliance for et al. 2017).

A

Drug repositioning beyond the low-hanging fruits

Duran-Frigola M, Mateo L, Aloy P. Drug repositioning beyond the low-hanging fruits. Current Opinion in Systems Biology. 2017;3:95–102. DOI: 10.1016/j.coisb.2017.04.010

Bibliography

Alexandrov, L. B. et al. (2013). “Signatures of mutational processes in human cancer”. In: *Nature* 500.7463, pp. 415–21.

Andre, F. et al. (2019). “Alpelisib for PIK3CA-Mutated, Hormone Receptor-Positive Advanced Breast Cancer”. In: *N Engl J Med* 380.20, pp. 1929–1940.

Arnedo-Pac, C. et al. (2019). “OncodriveCLUSTL: a sequence-based clustering method to identify cancer drivers”. In: *Bioinformatics*.

Arriola, E. et al. (2008). “Genomic analysis of the HER2/TOP2A amplicon in breast cancer and breast cancer cell lines”. In: *Lab Invest* 88.5, pp. 491–503.

Bailey, M. H. et al. (2018). “Comprehensive Characterization of Cancer Driver Genes and Mutations”. In: *Cell* 173.2, 371–385 e18.

Barretina, J. et al. (2012). “The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity”. In: *Nature* 483.7391, pp. 603–7.

Baselga, J. et al. (2014). “Phase III trial of nonpegylated liposomal doxorubicin in combination with trastuzumab and paclitaxel in HER2-positive metastatic breast cancer”. In: *Ann Oncol* 25.3, pp. 592–8.

Behan, F. M. et al. (2019). “Prioritization of cancer therapeutic targets using CRISPR-Cas9 screens”. In: *Nature* 568.7753, pp. 511–516.

Berger, A. H. et al. (2016). “High-throughput Phenotyping of Lung Cancer Somatic Mutations”. In: *Cancer Cell* 30.2, pp. 214–228.

Bertrand, D. et al. (2015). “Patient-specific driver gene prediction and risk assessment through integrated network analysis of cancer omics profiles”. In: *Nucleic Acids Res* 43.7, e44.

Biankin, A. V. (2017). “The road to precision oncology”. In: *Nat Genet* 49.3, pp. 320–321.

Bruna, A. et al. (2016). “A Biobank of Breast Cancer Explants with Preserved Intra-tumor Heterogeneity to Screen Anticancer Compounds”. In: *Cell* 167.1, 260–274 e22.

Burgucu, D. et al. (2012). “Tbx3 represses PTEN and is over-expressed in head and neck squamous cell carcinoma”. In: *BMC Cancer* 12, p. 481.

Byer, S. J. et al. (2011). “Tamoxifen inhibits malignant peripheral nerve sheath tumor growth in an estrogen receptor-independent manner”. In: *Neuro Oncol* 13.1, pp. 28–41.

Byrne, A. T. et al. (2017). “Interrogating open issues in cancer precision medicine with patient-derived xenografts”. In: *Nat Rev Cancer* 17.4, pp. 254–268.

Canisius, S., J. W. Martens, and L. F. Wessels (2016). “A novel independence test for somatic alterations in cancer shows that biology drives mutual exclusivity but chance explains most co-occurrence”. In: *Genome Biol* 17.1, p. 261.

Center, Broad Institute TCGA Genome Data Analysis (2016). “Firehose stddata$_{2016_{0}1_{2}8run}$”. In: *Broad Institute of MIT and Harvard*.

Cerami, E. et al. (2012). “The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data”. In: *Cancer Discov* 2.5, pp. 401–4.

Chakravarty, D. et al. (2017). “OncoKB: A Precision Oncology Knowledge Base”. In: *JCO Precis Oncol* 2017.

Chang, M. T. et al. (2018). “Accelerating Discovery of Functional Mutant Alleles in Cancer”. In: *Cancer Discov* 8.2, pp. 174–183.

Cheng, D. T. et al. (2015). “Memorial Sloan Kettering-Integrated Mutation Profiling of Actionable Cancer Targets (MSK-IMPACT): A Hybridization Capture-Based Next-Generation Sequencing Clinical Assay for Solid Tumor Molecular Oncology”. In: *J Mol Diagn* 17.3, pp. 251–64.

Clinical Cancer Genome Task Team of the Global Alliance for, Genomics et al. (2017). “Sharing Clinical and Genomic Data on Cancer - The Need for Global Solutions”. In: *N Engl J Med* 376.21, pp. 2006–2009.

Conte, N. et al. (2019). “PDX Finder: A portal for patient-derived tumor xenograft model discovery”. In: *Nucleic Acids Res* 47.D1, pp. D1073–D1079.

Cui, Y. et al. (2016). “BioCircos.js: an interactive Circos JavaScript library for biological data visualization on web applications”. In: *Bioinformatics* 32.11, pp. 1740–2.

Dao, P. et al. (2017). “BeWith: A Between-Within method to discover relationships between cancer modules via integrated analysis of mutual exclusivity, co-occurrence and functional interactions”. In: *PLoS Comput Biol* 13.10, e1005695.

Das, S. and A. W. Lo (2017). “Re-inventing drug development: A case study of the I-SPY 2 breast cancer clinical trials program”. In: *Contemp Clin Trials* 62, pp. 168–174.

Davidson-Pilon, C. (2019). “lifelines: survival analysis in Python”. In: *ournal of Open Source Software* 4.40, p. 1317.

Dembla, V. et al. (2018). “Prevalence of MDM2 amplification and coalterations in 523 advanced cancer patients in the MD Anderson phase 1 clinic”. In: *Oncotarget* 9.69, pp. 33232–33243.

Domcke, S. et al. (2013). “Evaluating cell lines as tumour models by comparison of genomic profiles”. In: *Nat Commun* 4, p. 2126.

Dong, C. et al. (2016). “iCAGES: integrated CAncer GEnome Score for comprehensively prioritizing driver genes in personal cancer genomes”. In: *Genome Med* 8.1, p. 135.

Einarsdottir, B. O. et al. (2014). “Melanoma patient-derived xenografts accurately model the disease and develop fast enough to guide treatment decisions”. In: *Oncotarget* 5.20, pp. 9609–18.

Eirew, P. et al. (2015). “Dynamics of genomic clones in breast cancer patient xenografts at single-cell resolution”. In: *Nature* 518.7539, pp. 422–6.

Engel, J. et al. (2019). “Lymph node infiltration, parallel metastasis and treatment success in breast cancer”. In: *Breast* 48, pp. 1–6.

Enright, A. J., S. Van Dongen, and C. A. Ouzounis (2002). “An efficient algorithm for large-scale detection of protein families”. In: *Nucleic Acids Res* 30.7, pp. 1575–84.

Fearon, E. R. and B. Vogelstein (1990). “A genetic model for colorectal tumorigenesis”. In: *Cell* 61.5, pp. 759–67.

Forbes, S. A. et al. (2017). “COSMIC: somatic cancer genetics at high-resolution”. In: *Nucleic Acids Res* 45.D1, pp. D777–D783.

Frampton, G. M. et al. (2013). “Development and validation of a clinical cancer genomic profiling test based on massively parallel DNA sequencing”. In: *Nat Biotechnol* 31.11, pp. 1023–31.

Franz, M. et al. (2016). “Cytoscape.js: a graph theory library for visualisation and analysis”. In: *Bioinformatics* 32.2, pp. 309–11.

Gao, H. et al. (2015). “High-throughput screening using patient-derived tumor xenografts to predict clinical trial drug response”. In: *Nat Med* 21.11, pp. 1318–25.

Gao, J. et al. (2013). “Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal”. In: *Sci Signal* 6.269, pl1.

Garnett, M. J. et al. (2012). “Systematic identification of genomic markers of drug sensitivity in cancer cells”. In: *Nature* 483.7391, pp. 570–5.

Garraway, L. A. and E. S. Lander (2013). “Lessons from the cancer genome”. In: *Cell* 153.1, pp. 17–37.

GENIE Consortium, Aacr Project (2017). “AACR Project GENIE: Powering Precision Medicine through an International Consortium”. In: *Cancer Discov* 7.8, pp. 818–831.

Gennari, A. et al. (2008). “HER2 status and efficacy of adjuvant anthracyclines in early breast cancer: a pooled analysis of randomized trials”. In: *J Natl Cancer Inst* 100.1, pp. 14–20.

Gerstung, M. et al. (2017). “Precision oncology for acute myeloid leukemia using a knowledge bank approach”. In: *Nat Genet* 49.3, pp. 332–340.

Gillet, J. P., S. Varma, and M. M. Gottesman (2013). “The clinical relevance of cancer cell lines”. In: *J Natl Cancer Inst* 105.7, pp. 452–8.

Gonzalez-Perez, A. et al. (2013). “IntOGen-mutations identifies cancer drivers across tumor types”. In: *Nat Methods* 10.11, pp. 1081–2.

Guinney, J. and J. Saez-Rodriguez (2018). “Alternative models for sharing confidential biomedical data”. In: *Nat Biotechnol* 36.5, pp. 391–392.

Haarberg, H. E. and K. S. Smalley (2014). “Resistance to Raf inhibition in cancer”. In: *Drug Discov Today Technol* 11, pp. 27–32.

Hanahan, D. and R. A. Weinberg (2000). “The hallmarks of cancer”. In: *Cell* 100.1, pp. 57–70.

– (2011). “Hallmarks of cancer: the next generation”. In: *Cell* 144.5, pp. 646–74.

Hidalgo, M. et al. (2014). “Patient-derived xenograft models: an emerging platform for translational cancer research”. In: *Cancer Discov* 4.9, pp. 998–1013.

Hoadley, K. A., C. Yau, T. Hinoue, et al. (2018). “Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer”. In: *Cell* 173.2, 291–304 e6.

Hoadley, K. A., C. Yau, D. M. Wolf, et al. (2014). “Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin”. In: *Cell* 158.4, pp. 929–944.

Hofree, M. et al. (2013). “Network-based stratification of tumor mutations”. In: *Nat Methods* 10.11, pp. 1108–15.

Hortobagyi, G. N. et al. (2016). “Correlative Analysis of Genetic Alterations and Everolimus Benefit in Hormone Receptor-Positive, Human Epidermal Growth Factor Receptor 2-Negative Advanced Breast Cancer: Results From BOLERO-2”. In: *J Clin Oncol* 34.5, pp. 419–26.

Hou, J. P. and J. Ma (2014). “DawnRank: discovering personalized driver genes in cancer”. In: *Genome Med* 6.7, p. 56.

Hsiehchen, D. and A. Hsieh (2018). “Nearing saturation of cancer driver gene discovery”. In: *J Hum Genet* 63.9, pp. 941–943.

Huun, J., P. E. Lonning, and S. Knappskog (2017). “Effects of concomitant inactivation of p53 and pRb on response to doxorubicin treatment in breast cancer cell lines”. In: *Cell Death Discov* 3, p. 17026.

Hyman, D. M. et al. (2015). “Vemurafenib in Multiple Nonmelanoma Cancers with BRAF V600 Mutations”. In: *N Engl J Med* 373.8, pp. 726–36.

International Cancer Genome, Consortium et al. (2010). “International network of cancer genome projects”. In: *Nature* 464.7291, pp. 993–8.

Iorio, F. et al. (2016). “A Landscape of Pharmacogenomic Interactions in Cancer”. In: *Cell* 166.3, pp. 740–754.

Izumchenko, E. et al. (2017). “Patient-derived xenografts effectively capture responses to oncology therapy in a heterogeneous cohort of patients with solid tumors”. In: *Ann Oncol* 28.10, pp. 2595–2605.

Jaeger, S., M. Duran-Frigola, and P. Aloy (2015). “Drug sensitivity in cancer cell lines is not tissue-specific”. In: *Mol Cancer* 14, p. 40.

Jamerson, M. H., M. D. Johnson, and R. B. Dickson (2000). "Dual regulation of proliferation and apoptosis: c-myc in bitransgenic murine mammary tumor models". In: *Oncogene* 19.8, pp. 1065–71.

Jardim, D. L. et al. (2015). "Impact of a Biomarker-Based Strategy on Oncology Drug Development: A Meta-analysis of Clinical Trials Leading to FDA Approval". In: *J Natl Cancer Inst* 107.11.

Juric, D., P. Castel, et al. (2015). "Convergent loss of PTEN leads to clinical resistance to a PI(3)Kalpha inhibitor". In: *Nature* 518.7538, pp. 240–4.

Juric, D., F. Janku, et al. (2019). "Alpelisib Plus Fulvestrant in PIK3CA-Altered and PIK3CA-Wild-Type Estrogen Receptor-Positive Advanced Breast Cancer: A Phase 1b Clinical Trial". In: *JAMA Oncol* 5.2, e184475.

Juric, D., J. Rodon, et al. (2018). "Phosphatidylinositol 3-Kinase alpha-Selective Inhibition With Alpelisib (BYL719) in PIK3CA-Altered Solid Tumors: Results From the First-in-Human Study". In: *J Clin Oncol* 36.13, pp. 1291–1299.

Kalari, K. R. et al. (2018). "PANOPLY: Omics-Guided Drug Prioritization Method Tailored to an Individual Patient". In: *JCO Clin Cancer Inform* 2, pp. 1–11.

Kandoth, C. et al. (2013). "Mutational landscape and significance across 12 major cancer types". In: *Nature* 502.7471, pp. 333–339.

Kim, Y. A., S. Madan, and T. M. Przytycka (2017). "WeSME: uncovering mutual exclusivity of cancer drivers and beyond". In: *Bioinformatics* 33.6, pp. 814–821.

Knudsen, E. S. and A. K. Witkiewicz (2017). "The Strange Case of CDK4/6 Inhibitors: Mechanisms, Resistance, and Combination Strategies". In: *Trends Cancer* 3.1, pp. 39–55.

Krepler, C. et al. (2017). "A Comprehensive Patient-Derived Xenograft Collection Representing the Heterogeneity of Melanoma". In: *Cell Rep* 21.7, pp. 1953–1967.

Krogan, N. J. et al. (2015). "The cancer cell map initiative: defining the hallmark networks of cancer". In: *Mol Cell* 58.4, pp. 690–8.

Laroche-Clary, A. et al. (2017). "Combined targeting of MDM2 and CDK4 is synergistic in dedifferentiated liposarcomas". In: *J Hematol Oncol* 10.1, p. 123.

Lauber, C., B. Klink, and M. Seifert (2018). "Comparative analysis of histologically classified oligodendrogliomas reveals characteristic molecular differences between subgroups". In: *BMC Cancer* 18.1, p. 399.

Lawrence, M. S. et al. (2014). "Discovery and saturation analysis of cancer genes across 21 tumour types". In: *Nature* 505.7484, pp. 495–501.

Le Tourneau, C. et al. (2015). "Molecularly targeted therapy based on tumour molecular profiling versus conventional therapy for advanced cancer (SHIVA): a multicentre, open-label, proof-of-concept, randomised, controlled phase 2 trial". In: *Lancet Oncol* 16.13, pp. 1324–34.

Ledford, H. (2016). "US cancer institute to overhaul tumour cell lines". In: *Nature* 530.7591, p. 391.

Lee, J. S. et al. (2018). "Harnessing synthetic lethality to predict the response to cancer treatment". In: *Nat Commun* 9.1, p. 2546.

Leiserson, M. D. et al. (2015). "Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes". In: *Nat Genet* 47.2, pp. 106–14.

Li, Z. et al. (2018). "Loss of the FAT1 Tumor Suppressor Promotes Resistance to CDK4/6 Inhibitors via the Hippo Pathway". In: *Cancer Cell* 34.6, 893–905 e8.

Liu, H. et al. (2015). "Redeployment of Myc and E2f1-3 drives Rb-deficient cell cycles". In: *Nat Cell Biol* 17.8, pp. 1036–48.

Liu, J. et al. (2018). "An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics". In: *Cell* 173.2, 400–416 e11.

Mateo, L. et al. (2018). "Exploring the OncoGenomic Landscape of cancer". In: *Genome Med* 10.1, p. 61.

Mayer, I. A. et al. (2017). "A Phase Ib Study of Alpelisib (BYL719), a PI3Kalpha-Specific Inhibitor, with Letrozole in ER+/HER2- Metastatic Breast Cancer". In: *Clin Cancer Res* 23.1, pp. 26–34.

McGlynn, L. M. et al. (2009). "Ras/Raf-1/MAPK pathway mediates response to tamoxifen but not chemotherapy in breast cancer patients". In: *Clin Cancer Res* 15.4, pp. 1487–95.

Merlo, L. M. et al. (2006). "Cancer as an evolutionary and ecological process". In: *Nat Rev Cancer* 6.12, pp. 924–35.

Mina, M. et al. (2017). "Conditional Selection of Genomic Alterations Dictates Cancer Evolution and Oncogenic Dependencies". In: *Cancer Cell* 32.2, 155–168 e6.

Mo, Q. et al. (2013). "Pattern discovery and cancer gene identification in integrated cancer genomic data". In: *Proc Natl Acad Sci U S A* 110.11, pp. 4245–50.

Mosca, R., A. Ceol, and P. Aloy (2013). "Interactome3D: adding structural details to protein networks". In: *Nat Methods* 10.1, pp. 47–53.

Mosca, R., J. Tenorio-Laranga, et al. (2015). "dSysMap: exploring the edgetic role of disease mutations". In: *Nat Methods* 12.3, pp. 167–8.

Mularoni, L. et al. (2016). "OncodriveFML: a general framework to identify coding and noncoding regions with cancer driver mutations". In: *Genome Biol* 17.1, p. 128.

Nakanishi, Y. et al. (2016). "Activating Mutations in PIK3CB Confer Resistance to PI3K Inhibition and Define a Novel Oncogenic Role for p110beta". In: *Cancer Res* 76.5, pp. 1193–203.

Newton, Y. et al. (2017). "TumorMap: Exploring the Molecular Similarities of Cancer Samples in an Interactive Portal". In: *Cancer Res* 77.21, e111–e114.

Nicolau, M., A. J. Levine, and G. Carlsson (2011). "Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival". In: *Proc Natl Acad Sci U S A* 108.17, pp. 7265–70.

Onitilo, A. A. et al. (2009). "Breast cancer subtypes based on ER/PR and Her2 expression: comparison of clinicopathologic features and survival". In: *Clin Med Res* 7.1-2, pp. 4–13.

Pallmann, P. et al. (2018). "Adaptive designs in clinical trials: why use them, and how to run and report them". In: *BMC Med* 16.1, p. 29.

Parris, T. Z. et al. (2014). "Frequent MYC coamplification and DNA hypomethylation of multiple genes on 8q in 8p11-p12-amplified breast carcinomas". In: *Oncogenesis* 3, e95.

Patnaik, A. et al. (2016). "Efficacy and Safety of Abemaciclib, an Inhibitor of CDK4 and CDK6, for Patients with Breast Cancer, Non-Small Cell Lung Cancer, and Other Solid Tumors". In: *Cancer Discov* 6.7, pp. 740–53.

Pedregosa, F. et al. (2011). "Scikit-learn: Machine Learning in Python". In: *JMLR* 12.Oct, pp. 2825–2830.

Pergolini, I. et al. (2017). "Tumor engraftment in patient-derived xenografts of pancreatic ductal adenocarcinoma is associated with adverse clinicopathological features and poor survival". In: *PLoS One* 12.8, e0182855.

Phesse, T. J. et al. (2014). "Endogenous c-Myc is essential for p53-induced apoptosis in response to DNA damage in vivo". In: *Cell Death Differ* 21.6, pp. 956–66.

Pineiro-Yanez, E. et al. (2018). "PanDrugs: a novel method to prioritize anticancer drug treatments according to individual genomic data". In: *Genome Med* 10.1, p. 41.

Pompili, L. et al. (2016). "Patient-derived xenografts: a relevant preclinical model for drug development". In: *J Exp Clin Cancer Res* 35.1, p. 189.

Porta-Pardo, E., T. Hrabe, and A. Godzik (2015). "Cancer3D: understanding cancer mutations through protein structures". In: *Nucleic Acids Res* 43.Database issue, pp. D968–73.

Prasad, V. (2016). "Perspective: The precision-oncology illusion". In: *Nature* 537.7619, S63.

Preusser, M. et al. (2018). "CDK4/6 inhibitors in the treatment of patients with breast cancer: summary of a multidisciplinary round-table discussion". In: *ESMO Open* 3.5, e000368.

Prokopenko, D. et al. (2016). "Utilizing the Jaccard index to reveal population stratification in sequencing data: a simulation study and an application to the 1000 Genomes Project". In: *Bioinformatics* 32.9, pp. 1366–72.

Pusapati, R. V. et al. (2006). "ATM promotes apoptosis and suppresses tumorigenesis in response to Myc". In: *Proc Natl Acad Sci U S A* 103.5, pp. 1446–51.

Razavi, P. et al. (2018). "The Genomic Landscape of Endocrine-Resistant Advanced Breast Cancers". In: *Cancer Cell* 34.3, 427–438 e6.

Rubio-Perez, C. et al. (2015). "In silico prescription of anticancer drugs to cohorts of 28 tumor types reveals targeting opportunities". In: *Cancer Cell* 27.3, pp. 382–96.

Salvadores, M., D. Mas-Ponte, and F. Supek (2019). "Passenger mutations accurately classify human tumors". In: *PLoS Comput Biol* 15.4, e1006953.

Sanchez-Vega, F. et al. (2018). "Oncogenic Signaling Pathways in The Cancer Genome Atlas". In: *Cell* 173.2, 321–337 e10.

Schaub, F. X. et al. (2018). "Pan-cancer Alterations of the MYC Oncogene and Its Proximal Network across the Cancer Genome Atlas". In: *Cell Syst* 6.3, 282–300 e2.

Schubert, M. et al. (2018). "Perturbation-response genes reveal signaling footprints in cancer gene expression". In: *Nat Commun* 9.1, p. 20.

Schutte, M. et al. (2017). "Molecular dissection of colorectal cancer in pre-clinical models identifies biomarkers predicting sensitivity to EGFR inhibitors". In: *Nat Commun* 8, p. 14262.

Schwaederle, M. et al. (2015). "Impact of Precision Medicine in Diverse Cancers: A Meta-Analysis of Phase II Clinical Trials". In: *J Clin Oncol* 33.32, pp. 3817–25.

Seabold, S. and J. Perktold (2010). "Statsmodels:Econometric and statistical modeling with python". In: *9th Python in Science Concerence*.

Senft, D. et al. (2017). "Precision Oncology: The Road Ahead". In: *Trends Mol Med* 23.10, pp. 874–898.

Shannon, P. et al. (2003). "Cytoscape: a software environment for integrated models of biomolecular interaction networks". In: *Genome Res* 13.11, pp. 2498–504.

Shapiro, G. I. (2017). "Genomic Biomarkers Predicting Response to Selective CDK4/6 Inhibition: Progress in an Elusive Search". In: *Cancer Cell* 32.6, pp. 721–723.

Shihab, H. A. et al. (2015). "An integrative approach to predicting the functional effects of non-coding and coding sequence variation". In: *Bioinformatics* 31.10, pp. 1536–43.

Shoemaker, R. H. (2006). "The NCI60 human tumour cell line anticancer drug screen". In: *Nat Rev Cancer* 6.10, pp. 813–23.

Shoemaker, R. H. et al. (1983). "Use of the KB cell line for in vitro cytotoxicity assays". In: *Cancer Treat Rep* 67.1, p. 97.

Simon, R. (2017). "Critical Review of Umbrella, Basket, and Platform Designs for Oncology Clinical Trials". In: *Clin Pharmacol Ther* 102.6, pp. 934–941.

Sitaram, R. T. et al. (2016). "Transforming growth factor-beta promotes aggressiveness and invasion of clear cell renal cell carcinoma". In: *Oncotarget* 7.24, pp. 35917–35931.

Smith, J. C. and J. M. Sheltzer (2018). "Systematic identification of mutations and copy number alterations associated with cancer patient prognosis". In: *Elife* 7.

Stockley, T. L. et al. (2016). "Molecular profiling of advanced solid tumors and patient outcomes with genotype-matched clinical trials: the Princess Margaret IMPACT/COMPACT trial". In: *Genome Med* 8.1, p. 109.

Stouffer, S.A. et al. (1949). "The American Soldier". In: *Princeton University Press, Princeton.* Vol1.Adjustment during Army Life.

Subramanian, A. et al. (2017). "A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles". In: *Cell* 171.6, 1437–1452 e17.

Sweeney, K. J. et al. (1997). "Cyclin D2 activates Cdk2 in preference to Cdk4 in human breast epithelial cells". In: *Oncogene* 14.11, pp. 1329–40.

Szczurek, E. and N. Beerenwinkel (2014). "Modeling mutual exclusivity of cancer mutations". In: *PLoS Comput Biol* 10.3, e1003503.

Szlachta, K. et al. (2018). "CRISPR knockout screening identifies combinatorial drug targets in pancreatic cancer and models cellular drug response". In: *Nat Commun* 9.1, p. 4275.

Tamborero, D. et al. (2018). "Cancer Genome Interpreter annotates the biological and clinical relevance of tumor alterations". In: *Genome Med* 10.1, p. 25.

Tanguy, M. L. et al. (2018). "Cdk4/6 inhibitors and overall survival: power of first-line trials in metastatic breast cancer". In: *NPJ Breast Cancer* 4, p. 14.

Thangavel, C. et al. (2011). "Therapeutically activating RB: reestablishing cell cycle control in endocrine therapy-resistant breast cancer". In: *Endocr Relat Cancer* 18.3, pp. 333–45.

Therasse, P. et al. (2000). "New guidelines to evaluate the response to treatment in solid tumors. European Organization for Research and Treatment of Cancer, National Cancer Institute of the United States, National Cancer Institute of Canada". In: *J Natl Cancer Inst* 92.3, pp. 205–16.

Thorlund, K. et al. (2018). "Key design considerations for adaptive clinical trials: a primer for clinicians". In: *BMJ* 360, k698.

Tokheim, C. J. et al. (2016). "Evaluating the evaluation of cancer driver genes". In: *Proc Natl Acad Sci U S A* 113.50, pp. 14330–14335.

Tu, Q. et al. (2018). "CDKN2B deletion is essential for pancreatic cancer development instead of unmeaningful co-deletion due to juxtaposition to CDKN2A". In: *Oncogene* 37.1, pp. 128–138.

Ulz, P., E. Heitzer, and M. R. Speicher (2016). "Co-occurrence of MYC amplification and TP53 mutations in human cancer". In: *Nat Genet* 48.2, pp. 104–6.

Vandin, F., E. Upfal, and B. J. Raphael (2012). "De novo discovery of mutated driver pathways in cancer". In: *Genome Res* 22.2, pp. 375–85.

Villacorta-Martin, C., A. J. Craig, and A. Villanueva (2017). "Divergent evolutionary trajectories in transplanted tumor models". In: *Nat Genet* 49.11, pp. 1565–1566.

Vogelstein, B. and K. W. Kinzler (2015). "The Path to Cancer –Three Strikes and You're Out". In: *N Engl J Med* 373.20, pp. 1895–8.

Vogelstein, B., N. Papadopoulos, et al. (2013). "Cancer genome landscapes". In: *Science* 339.6127, pp. 1546–58.

Vora, S. R. et al. (2014). "CDK 4/6 inhibitors sensitize PIK3CA mutant breast cancer to PI3K inhibitors". In: *Cancer Cell* 26.1, pp. 136–49.

Wang, M. et al. (2018). “Humanized mice in studying efficacy and mechanisms of PD-1-targeted cancer immunotherapy”. In: *FASEB J* 32.3, pp. 1537–1549.

Wang, S. et al. (2018). “Typing tumors using pathways selected by somatic evolution”. In: *Nat Commun* 9.1, p. 4159.

Weinstein, J. N. et al. (2013). “The Cancer Genome Atlas Pan-Cancer analysis project”. In: *Nat Genet* 45.10, pp. 1113–20.

Wertz, I. E. et al. (2011). “Sensitivity to antitubulin chemotherapeutics is regulated by MCL1 and FBW7”. In: *Nature* 471.7336, pp. 110–4.

Wheler, J. J. et al. (2016). “Presence of both alterations in FGFR/FGF and PI3K/AKT/mTOR confer improved outcomes for patients with metastatic breast cancer treated with PI3K/AKT /mTOR inhibitors”. In: *Oncoscience* 3.5-6, pp. 164–72.

Whittle, J. R. et al. (2015). “Patient-derived xenograft models of breast cancer and their predictive power”. In: *Breast Cancer Res* 17, p. 17.

Willmer, T. et al. (2015). “The T-Box factor TBX3 is important in S-phase and is regulated by c-Myc and cyclin A-CDK2”. In: *Cell Cycle* 14.19, pp. 3173–83.

Willyard, C. (2018). “The mice with human tumours: Growing pains for a popular cancer model”. In: *Nature* 560.7717, pp. 156–157.

Wu, H. et al. (2015). “Identifying overlapping mutated driver pathways by constructing gene networks in cancer”. In: *BMC Bioinformatics* 16 Suppl 5, S3.

Yang, W. et al. (2013). “Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells”. In: *Nucleic Acids Res* 41.Database issue, pp. D955–61.

Zehir, A. et al. (2017). “Mutational landscape of metastatic cancer revealed from prospective clinical sequencing of 10,000 patients”. In: *Nat Med* 23.6, pp. 703–713.

Zhao, Y. et al. (2018). “PI3K Positively Regulates YAP and TAZ in Mammary Tumorigenesis Through Multiple Signaling Pathways”. In: *Mol Cancer Res* 16.6, pp. 1046–1058.

Zhong, Q. et al. (2009). “Edgetic perturbation models of human inherited disorders”. In: *Mol Syst Biol* 5, p. 321.