Understanding the genomic makeup of tumors to guide personalized medicine

Carlota Rubio Perez

DIRECTORS DE LA TESI

Dra. Nuria López Bigas

Dr. Abel González Pérez

TESI DOCTORAL UPF 2017

DEPARTAMENT OF EXPERIMENTAL AND HEALTH SCIENCES

upf. Universitat Pompeu Fabra
Barcelona

Pels que hi són
i pels que han marxat

Agraïments

Sense el Bernat aquesta tesi no hauria sigut possible. Gràcies per aguantar els moments de nervis, de desmotivació, de plors i de bogeria i per fer que mai em doni per vençuda. També per compartir i exprimir al màxim els bons i moments de desconnexió. Has sigut el meu màxim suport aquests quatre anys, i falten paraules per agrair-ho. Però tot això i més tu ja ho saps.

Evidentment ma mare, tot i que ja no sigui al dia a dia, també ha hagut d'aguantar tots els meus estats emocionals, com porta fent tota la vida. Perquè tot i que els últims anys han sigut complicats, has seguit allà donant-me suport amb la meva feina i la meva vida... Gràcies es queda curt.

I tot i que ja no hi sigui, també li vull dedicar a l'Àvia. Perquè d'alguna forma tot lo orgullosa que va estar sempre de mi, es plasma aquí.

Al meu pare pels seus consells i per fer que no hagi deixat medicina. Que realment hauria sigut un error.

Al tiet Jesús, la tia nina, la Pati i l'Álvaro, pels estius a Cambrils i pels menjars i resposteries de Master Chef que sempre fem.

A la Marta, l'Oriol i el Guillem, per ser una família encantandora que m'ha acollit des del primer dia sense dubtar-ho.

I passada la família, vull agrair tots els genial moments que m'han permès desconnectar i agafar energies i ganes als meus amics. Començant pels *let's* o amics de la uni o amics de bio o amics una mica frikis (sí, en part sabeu que és així: Visca el let's symposium!). M'encantaria colar-vos un paràgraf genèric i quedar-me tan tranquil·la perquè m'heu posat tensa amb això dels agraïments, però va, faré algunes personificacions. Agraeixo a ...

> ... la Silvia pels seus moments de bogeria que animen el dia a qualsevol (MCPollo). Per ser pura alegria i positivisme que han ajudat a contrarestar el meu negativisme en nombroses ocasions i per aguantar les meves xapes pacientment.
>
> ... la Gasull perquè també te moments de bogeria i envia selfies que també animen el dia a qualsevol. Per ser alegre i reflexiva, escoltar les meves històries, aportar-hi positivisme i ajudar-me a analitzar-les.
>
> ... equip padilla o Pano, Juna (i Tango). Perquè va ser genial el temps que vam viure juntes, tots els marujeos abans d'anar a dormir, els runnings pel matí per fer tard, les neteges a les 12 de la nit, els sopars que vam fer, les series que vam mirar juntes (ai, no! que no ho vam aconseguir mai...) i tots els moments que vam compartir! Concretament, al Sgt. Pano li agraeixo el seu toc de realisme en nombroses ocasions, que fa tocar de peus a terra, i per ser al meu costat en les bones i les

dolentes des de que ens vam conèixer a la matrícula de la uni, a l'erasmus a Londres fins que ens anem a USA! I a la Juna, pels moments de locura que hem tingut juntes, per tot el que hem arribat a desfogar-nos l'una amb l'altre, el mutu bon humor pel matí, les reflexions que hem compartit els plors i els riures.

... la Marina, pel moments de frikisme compartit, entre HPs i LOTRs. Per les birres, els riures i els moments de serietat perquè quan són necessaris també hi són.

... la Cuca, pels moments de cuques, que són boníssims. Per la seva naturalitat i la seva personalitat cauenca, que crea com un ambient de bon rollo i positivisme molt guais.

... la Mariona, per tots els marujeos i les xapes compartides. Per saber escoltar i també explicar, sempre amb un somriure; i per fer-me descobrir ofertes i destinacions de viatges que jo sempre desconec.

... a l'Ali per la seva alegria contagiosa que et fa riure encara que no vulguis. Perquè acostuma a generar els moments èpics, tant necessaris per recordar durant generacions i generar més alegria encara.

... la Pebel, perquè tot i que hagi sigut a la distància, quan xerrem és com si fóssim al costat, aportant-me punts de pau i calma en les meves bogeries i suport en els meus pous.

... a la Vicky pels bons moments que tenim quan ve i pels frikismes varis que compartim en ocasions, que amb algú s'han de compartir!

... a l'Abel per fer que m'agradi LDS. Per les reflexions científiques i no científiques, pels Still alive i totes les birres!

... la Lara, per donar una visió no conformista i d'alguna manera més reivindicativa, que és 100% necessària pel meu conformisme.

... a la House per la seva naturalitat i espontaneïtat i per intentar repetidament fer-me creure que l'IRB no és tan lleig (sense èxit pel moment).

... al Marc per fer-me venir ganes de seguir amb Medicina i pel seu caràcter tan afable i tranquil, que sempre em fa molta enveja i m'agradaria que fos el meu.

... a la Farras pels seus híper-resumenes fabulosos que encara m'ha de deixar i per fer-me veure que la meva por als avions és una tonteria en comparació a la seva als pinyols.

... al Pepe per les seves excentricitats i bogeries, que aporten bons riures i per ensenyar-me que hi ha vida fora l'acadèmia.

No m'agradava massa la idea de puntualitzar perquè em deixo coses a comentar.... així que deixo una llista de memorables per la posteritat: "que te costaba tirar un poco para adelante 😉", cuques de persones varies, "que os jodan", Rototoms, Frankfurt una gran ciutat per fer turisme, la primera boda (J&I 2014), bidet, let's Tuesdays varis, múltiples regals personalitzats i múltiples birres.

A les meves amigues del cole+1, pels soparets de desconnexió que fem, pels diners que mai quadrem, pels mojitos i cap d'any, l'esdeveniment que mai fa pal.

> A Laura por su empatia y comprensión durante tantos años y sus locuras que siempre son divertidas de escuchar.
>
> A Isa por su sinceridad, que aunque a veces es radical, nunca está de más; y lo más importante, por no dejarme nunca sola pidiendo una cerveza.
>
> A l'ely per la seva espontaneïtat i naturalitat que sempre són refrescants.
>
> A l'eli per aportar un toc de racionalitat a la meva vida quan ens veiem.
>
> I al Vie, per ser el +1 més integrat i alegre. Per fer que no quedem soles la Isa i jo anant a les Santes (tot i que potser ja ens hem fet una mica massa grans...).

I als nous amics que he incorporat amb el Bernat, que són tots genials i divertits. Especialment al Masclans, al Pons, al Turi i al Joaquim que són amb els que he compartit més moments i sempre m'ho he passat molt be! I inclús hem parlat de coses series algunes vegades i tot... Entre voleys no competitius fins a bowlings tampoc competitius.

I finalment als BGs, als de cada dia. Aquests quatre anys han sigut tota una experiència per mi, tant perquè ha sigut la primera vegada que he treballat, com perquè per fi he conegut que és la ciència a través de vosaltres. Com a tota feina hi han hagut moments genials, moments bons i moments no tan bons, però he après moltíssim tant a nivell professional com personal, i això és el que considero més important. El grup m'ha ajudat a créixer moltíssim, i tinc la sensació que vaig entrar una mica com una espècie d'adolescent però surto, o això espero, com una persona adulta.

Agraeixo a l'Abel i la Núria tot el que m'han ensenyat durant aquests quatre anys. Pels projectes en que m'han fet participar i els investigadors que he arribat a conèixer gràcies a això, que em permet expandir els meus horitzons i obrir-me portes. També pel seu suport durant la tesi, per la seva perseverança i consell en la meva feina i les meves idees, que sempre han sigut escoltades. Per proposar-me tasques de forma equilibrada en funció del meu treball, però fent també que m'exigís a mi mateixa; per deixar desenvolupar les meves propostes més encertades i ajudar a descartar les més errònies, sempre amb consells constructius. I a la Nuria per creure que jo podia amb tot i acceptar que pogués compaginar fer Medicina i organitzar simpòsiums.

Al David per totes les hores compartides aquests quatre anys, incloent les de Hangout. Per tots els projectes que hem fet, per les vegades que ens hem posat d'acord i per les que no també. Crec que entre els dos hem tirat endavant coses molt xules, i que totes les hores invertides han tingut i tindran recompensa. També per totes les penes, que compartides han pesat menys, i per les alegries també (com la publicació del CGI el 20xx!).

Als PhDs al Joan, l'Oriol i la Inés. Perquè feu més divertits els dies de feina i els dinars. Sí, no puc parlar massa d'extraescolars perquè sembla que mai ens posarem d'acord per fer un sopar... però al retreat ho vam partir bastant. Al Joan per totes les beer sessions on ho hem petat i l'endemà hem anat a treballar. A la Inés, per totes els moments frikis de series, llibres, pelis (i tràilers) dels que no em cansaria de parlar. A l'Oriol per totes les xapes que t'he clavat i per haver-les aguantat, moltes gràcies! Tens futur com coach, però no tinc dubtes que al doctorat ho petaràs! (els 3 ho fareu).

Al Jordi per tots els milers de dubtes que t'he preguntat, per no haver-te'n cansat i tenir sempre ganes i disponibilitat d'ajudar. També per inculcar-me bons hàbits de feina, que crec em seran molt útils de cara al futur sens dubte! I més important, per haver-me fet descobrir que la Mola mola.

A Loris por hacer me ver que un lenguaje de programación puede ser algo guai, y descubrirme todo el mundillo del Python. Aunque no me haya unido al Py-frikismo, almenos ahora sé que existe!

To Sabari for all the beer sessions, mixed with some Volleywood and occasional partying together with learning good science from time to time.

A Iker, Fran y Ferran por vuestra disponibilidad en ayudar en cualquier cosa, en escuchar todas las dudas y ayudar a razonarlas.

A la Martina, la Mertixell i l'Erika pel toc d'alegria genial que donen al grup i al despatx.

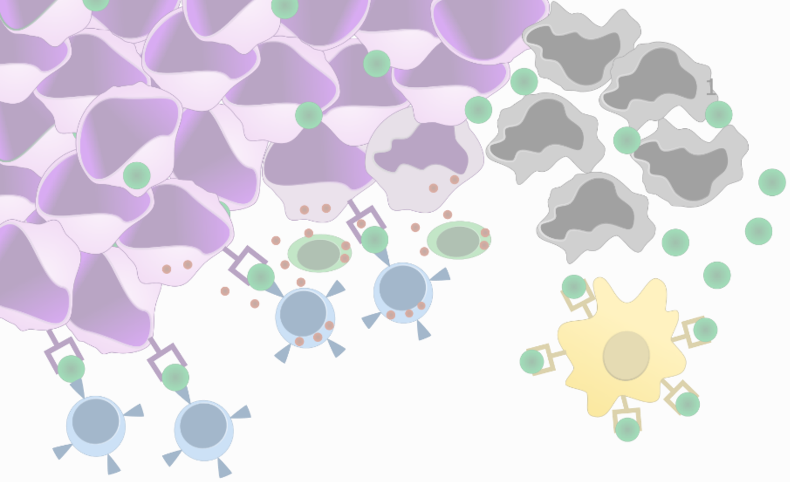
Abstract

Cancer is a disease of the genome. The study of tumor genomic alterations is used to guide several precision medicine strategies, some approved and a large number under clinical development. On the other hand, the study of tumor immunity is recently becoming the key for the success of other personalized strategies, named immunotherapies. Along this thesis I have made several contributions towards the advance of cancer precision medicine, based on the study of tumor “omics” data. First, I evinced the landscape of genomic-guided anti-cancer therapies. Second, I developed OncoPaD, a tool for the rational design of cost-effective cancer gene panels. Third, I contributed to the development of Cancer Genome Interpreter, a tool for the biological and therapeutic interpretation of variants found in newly sequenced tumors. Forth, I identified tumor intrinsic molecular mechanisms involved in tumor immune evasion.

Resum

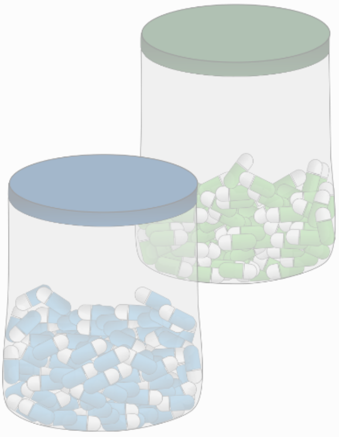
El càncer és una malaltia del genoma. L'estudi de les alteracions genòmiques dels tumors s’utilitza com a guia en varies estratègies de medicina de precisió, algunes d'elles aprovades i d'altres en assajos clínics. D'altra banda, l’estudi de la immunitat tumoral està esdevenint una peça clau per l’èxit d’altres estratègies terapèutiques, anomenades immunoteràpies. Al llarg d'aquesta tesi, mitjançant l'estudi de les dades “òmiques” tumorals, he contribuït de varies maneres cap a l'avenç de la medicina de precisió pel càncer. Primer, he identificat el panorama de les teràpies anticanceroses guiades per alteracions genòmiques. Segon, he desenvolupat OncoPaD, una eina pel disseny cost-efectiu i racional de panells de seqüenciació per càncer. A més, he contribuït al desenvolupament del Cancer Genome Interpreter, una eina que ajuda a la interpretació biològica i terapèutica de les variants presents a tumors novament seqüenciats. Per últim, he identificat diversos mecanismes mitjançant els quals els tumors són capaços d'evadir l’atac del sistema immunològic.

INDEX



PART I

INTRODUCTION



1.Overview of cancer disease

The first time the term cancer appeared in the literature was in 400 BC when Hippocrates nominated the tumors of his patients karkinos -crab in greek- because they resembled him a crab. However, it was not the first time in human history that cancer was described. The first documented reference to a disease which could be cancer dates from 2625 BC, described by the Egyptian physician Imhotep as a "disease without cure that showed protuberances in the chest". The first scientific evidence of cancer appeared many years later, an abdominal tumor dating from the 7th century AC discovered by the paleopathologist Arthur Aufderheide in 1990 in the Peruvian Chiribaya mummies[1]*. Therefore, cancer is not a new nor a modern disease. It has been part of human history for many years. Yet, the knowledge about its pathophysiology or about the best therapeutic strategies for most of the cancer diseases is incomplete.*

1.1 Definition and epidemiology

Cancer is defined as a group of diseases characterized by uncontrolled cellular proliferation, that in solid tissues forms a mass named tumor, which leads to the *exitus* of the patient if untreated. A cancer can begin in a specific organ but eventually it can propagate either by invading nearby structures or by migrating (i.e. metastasize), through bloodstream dissemination or through lymphatic metastases, to farther parts in the human body. The reasons why cancer leads to the *exitus* of the patient depends on the affected organ (e.g. renal carcinoma may lead to renal failure while liver carcinoma may lead to severe blood toxicity).

There are hundreds of different cancer types and subtypes described, going beyond the affected organ. The standard nomenclature for referring to the different cancer types is based on the International Classification of Diseases for Oncology (ICD-O-3)[2]. It classifies the cancer types according to the tissue of origin (i.e. histological type) in five major categories (carcinomas, sarcomas, leukemias, lymphomas and mesotheliomes); or according to the origin body location (i.e. primary site; e.g. breast, lung, stomach, etc)[2].

Cancer diseases are a major cause of morbidity, with 14 million new cases every year, 182/100,000 incidence rate; and mortality, with 8 million deaths, 102/100,000 mortality rate (data from 2012, according to the most recent study on cancer distribution worldwide)[3]. The incidence of the different cancer types is different, even at gender level. Among men the most prevalent cancer types are the ones affecting the lung, representing the 16.7% of all diagnosed men cancers; and among women the most prevalent cancer type is breast cancer, representing the 25.2% of the women diagnosed cancers. Moreover, there are differences between incidence rate and mortality rate across the different cancer types[3], mostly depending on each cancer type aggressiveness and the available therapeutic options.

1.2 Etiology and pathophysiology

Genomic alterations can be somatic -acquired in specific tissues during the lifetime of the individual- or germline -present in all body cells since birth. The main cause of tumor development (i.e tumorigenesis) is meant to be the accumulation of somatic alterations. Although some germline alterations are also known to play a role in cancer development. This idea of cancer as a consequence of somatic DNA alterations (e.g. mutations, copy number alterations...) has gained general acceptance during the last 25 years. Convincing evidence over many years has been provided by: systematic studies of X-rays, work on chemical mutagenesis and the large amount of data demonstrating smoke as the causative agent of lung cancer[4].

Somatic DNA alterations appear in cells due to errors caused by endogenous or exogenous processes which generate DNA damage. Most of these errors are repaired through several complex cellular mechanisms. However, if some of these errors are not properly-repaired, they give rise to somatic alterations which can, eventually, give rise to malignant cells over time. Therefore, the rate at which somatic mutations accumulate in the cells depends on the interplay between the errors generated by endogenous and exogenous processes and the rate at which they are repaired[5].

The endogenous processes generating DNA damage can be: random errors during DNA replication in the preparation for cell division; DNA repair machinery errors, because of the faulty recognition of DNA damaged regions; or spontaneous chemical changes in DNA bases (e.g. deamination of cytosine to uracil)[6]. Besides, some germline alterations (e.g. *BRCA* mutation) may have an influence on these endogenous processes too. Among the exogenous factors generating DNA damage, the better described ones for cancer are: exposure to carcinogens such as tobacco smoke or UV light radiation; and viral infections, such as hepatitis virus or papillomavirus[4].

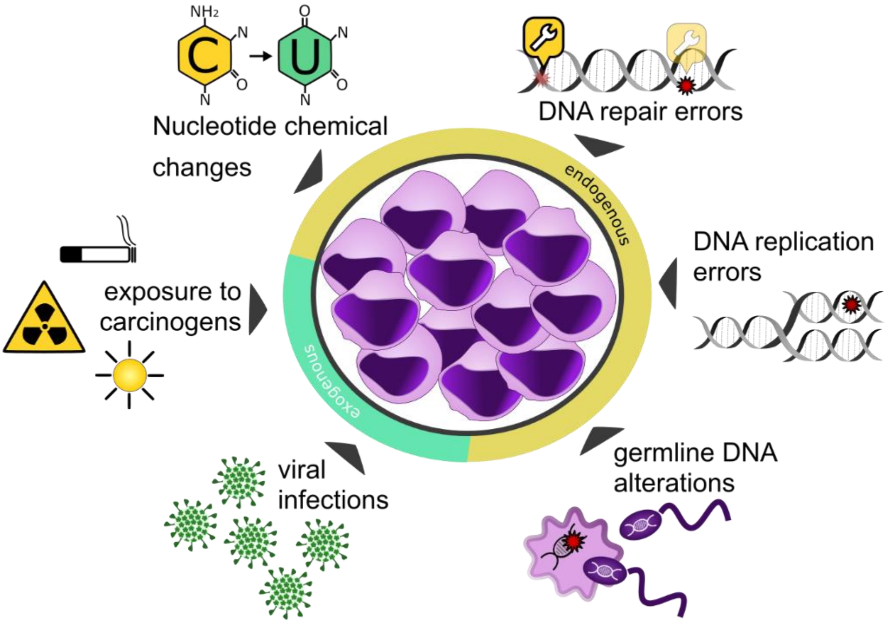


Figure 1 | Causative agents of cancer. Diagram of the main causative agents of DNA somatic alterations: exogenous (exposure to carcinogens, viral infections) and endogenous (DNA replication errors, DNA repair errors, germline DNA alterations, viral infections or spontaneous nucleotide chemical changes). Purple cells in the central circle represent a tumor.

2.Tumor molecular alterations

Cancer is driven by somatic alterations in the cellular DNA (i.e. genomic alterations). During each person's lifetime DNA genomic alterations accumulate in the genome, along with alterations in the epigenome -the set of DNA modifications not affecting the sequence which may influence gene activity- and the transcriptome of the cell- the set of transcribed RNA molecules. These genomic somatic alterations affect the function of key genes mostly related to cell growth and survival, de-regulating cellular biological process which lead to uncontrolled cellular proliferation. There are several types of genomic alterations which have varying sizes, ranging from mutations that affect a single nucleotide to chromosomal rearrangements or whole-genome duplications[7,8]. Even if I acknowledge that epigenomic alterations are relevant for tumorigenesis I will focus on genomic and transcriptomic alterations, as are the ones extensively analyzed in this thesis.

2.1 Mutations

Mutations are a type of genomic alterations which affect one or few nucleotides in the DNA sequence. When mutations occur in the coding part of the genome -the DNA regions called exons that are transcribed into messenger RNA- they can affect the protein aminoacid sequence. Depending on whether mutations change the aminoacid sequence, they can be classified into: synonymous (i.e. silent), that are single nucleotide variants (SNVs) (e.g. C to T) which do not change the aminoacid sequence because of codon degeneracy; and non-synonymous mutations (also known as PAMs, protein affecting mutations), that are DNA variants which alter the aminoacid sequence of a protein[9–11]. There are six main types of non-synonymous mutations[1]:

1) Missense mutations: SNVs which cause an aminoacid substitution (e.g. histidine to arginine). This type of mutations are the ones classically more frequently associated to tumorigenesis (e.g. *BRAF* V600E, *KRAS* G12D or *EGFR* L858R).

2) Nonsense mutations: SNVs where any aminoacid codon is replaced by a stop codon, leading to a premature end of translation.

3) Splice site mutations: SNVs at splice site sequences in the intron-exon junction. Splice sites are sequence elements essential for splicing - the process which removes the non-coding parts of the DNA sequence (introns) and selects the exons to be transcribed. These SNVs in splice sites may alter the ratio of alternative splicing patterns or affect the splicing of constitutive exons[12].

4) Translation start mutations: SNVs where at least one base of the starting codon is changed, they may affect the start of the transcription.

5) Translation stop mutations: SNVs where at least one base of the stop codon is changed, they may result in an elongated transcript, the transcribed RNA molecule.

6) Indels: insertions and deletions of few nucleotides (<100 base pairs). Indels maintain the open reading frames of the proteins when

[1] The types of PAMs 2 to 5 and the frameshift indels are usually referred to as truncating mutations[10]

they are multiples of three, causing the insertion or deletion of specific aminoacids. However, when not divisible by three, they induce a change in the reading frame (commonly referred to as frameshift mutations), generating a dramatically different transcript[10,13].

Of note, some types of non-coding alterations have also been associated with cancer development, even if they have been less explored until the date. The better described ones are SNVs in the promoter region of protein coding genes supporting uncontrolled cell growth, leading to their overexpression (e.g. mutations in the TERT promoter or mutations in LMO2 promoter)[14–16].

2.2 Chromosomal rearrangements

Chromosomal rearrangements or structural variants (SVs), are defined as alterations of the DNA sequence of approximately 1 kilobase (1,000 nucleotide base pairs) or larger size, in which DNA has been broken and rejoined elsewhere in the genome[17]. Chromosomal rearrangements can be balanced, preserving the amount of genetic information, or unbalanced, not preserving it[5,10,18].

The number of rearrangements that can be found in a chromosome can vary from a single rearrangement in a specific genomic region to thousands of clustered chromosomal rearrangements. This last phenomenon is termed chromotripsis and during the last years it has been described as a prevalent genomic aberration in several cancer types such as colorectal carcinoma[19].

2.2.1 Balanced rearrangements

Balanced rearrangements maintain the two copies of each DNA region, but re-order them across the genome. They can be caused by: (i) insertions, of one chromosomal region into the same or another chromosome; (ii) translocations, interchange of regions between chromosomes; and (iii) inversions, 180-degree chromosomal rotations[5,10].

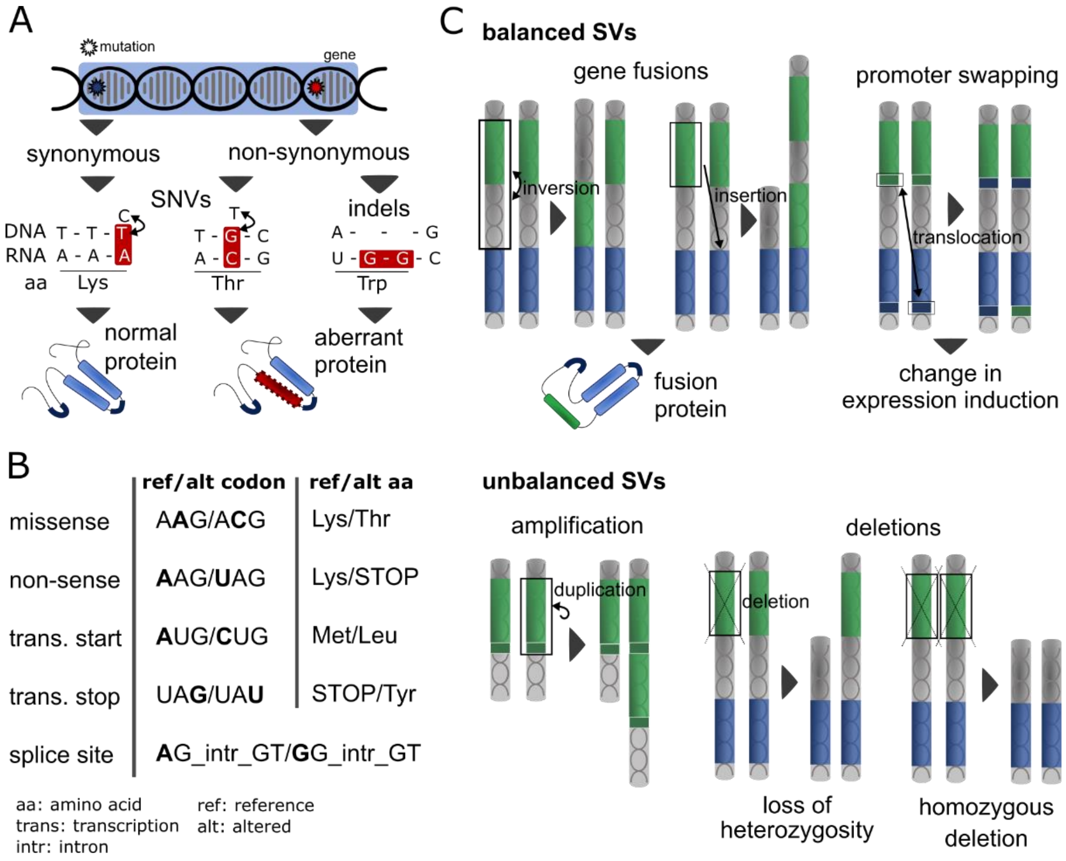
Among the types of balanced rearrangements, those with the most thoroughly described outcome in cancer are gene fusions. Gene fusions are translocations of two genes that join their coding sequences into a new fusion gene which encodes a fusion protein, with a function different to the one of each initial gene[10]. The first genic fusion in association with tumorigenesis was discovered in 1962, *BCR-ABL* (also known as Philadelphia chromosome) in chronic myeloid leukemia[20]. Over the last decades, several gene fusions have been associated to tumorigenesis, more frequently in hematologic malignancies (e.g. *PML-RARA* fusions in acute myelocytic leukemia[21] or *ALK* fusions in lung carcinomas[22–25]).

Additionally, there are other types of outcomes for balanced rearrangements that have also been described in cancer such as: swapping of 5’ ends, including the promoter, causing a change in the transcriptional induction either to an enhance or repression; incorporation of *trans* (i.e. distant) regulatory elements close to the transcription start sites, changing the transcriptional induction too; and gene truncation, generating aberrant transcripts[10,26].

2.2.2 Unbalanced rearrangements

Unbalanced rearrangements can be caused by duplications and multiple repeats, which imply an increase of the genomic content; or deletions, which cause a loss of genomic content. These rearrangements can vary in size, affecting from focal regions to whole chromosomal arms. If the duplications/repeats (i.e. amplifications) or deletions happen in coding regions of the genome they induce changes in the number of copies of the genes, referred to as Copy Number Alterations (CNAs). In turn, CNAs affecting genes may change their expression levels, either leading to an overexpression or to an underexpression[10,27].

Deletions can cause the loss of the two copies of the gene, homozygous loss, or just the loss of one of the copies, loss of heterozygosity. In some cases, the loss of heterozygosity is repaired, in terms of genomic content, and the remaining copy is duplicated, this phenomenon is called copy number neutral loss of heterozygosity. If it is not repaired, in some cases it can lead to an impaired phenotype due to haploinsufficiency[28].



	ref/alt codon	ref/alt aa
missense	A**A**G/A**C**G	Lys/Thr
non-sense	**A**AG/**U**AG	Lys/STOP
trans. start	**A**UG/**C**UG	Met/Leu
trans. stop	UA**G**/UA**U**	STOP/Tyr
splice site	**A**G_intr_GT/**G**G_intr_GT	

aa: amino acid
trans: transcription
intr: intron
ref: reference
alt: altered

Figure 2 | Main genomic alterations found in the tumor genome. Schematic representation of the main tumor genomic alterations. (A) Types of mutations according to their effect on the coding sequence: (i) synonymous, mutations not causing any aminoacid change, producing a normal protein; (ii) non-synonymous, mutations changing the aminoacid sequence of the protein. (B) Table presenting the types of non-synonymous Single Nucleotide Variants (SNVs). (C) Structural variants (SVs) classified into balanced and unbalanced. Balanced SVs can be: inversions, insertions or translocations, which can cause different outcomes (such as gene fusion or promoter swapping). Unbalanced SVs can generate a gene amplification, through duplications or multiple repetitions (not represented) or a gene deletion, through genomic deletions.

2.3 Gene expression changes

Changes in the transcriptome have been widely associated to cancer. It has been observed that many genes, even thousands, are differentially expressed, either overexpressed or underexpressed, between normal and tumor samples[29,30].

Expression differences at tissue type level have been observed across normal tissues, comprehensively explored within the framework of the Genotype-Tissue Expression (GTEx) project which has sequenced around 8000 normal tissues from autopsies[31]; as well as across different tumor tissues, within the framework of The Cancer Genome Atlas[32]. However, it has been shown that the differences across cancer types are not only explained because of the tissue of origin, but they are also related to the genomic, transcriptomic and epigenomic alterations (such as different DNA methylation patterns)[29,33] of the tumor samples[32].

On one hand, genomic alterations can modify the expression of genes through the alteration of regulatory structures, both *cis* (i.e. nearby) or *trans,* that result in either overexpression or underexpression of the genes (e.g. mutations or chromosomal rearrangements in the promoter region of the gene or in the transcription factors); or through alterations in the number of copies of the gene, either amplifications or deletions.

On the other hand, non-coding RNA molecules, such as long-non-coding RNAs (lncRNA) and microRNAs, can also modulate the expression of protein-coding genes. An example of this is the lncRNA *MALAT1* whose expression has been associated to a transcriptional and post-transcriptional regulation of cytoskeleton and extracellular matrix genes in various cancer types, promoting invasiveness and metastases[34].

3. Genomics of the tumorigenesis

Cancer development (i.e. tumorigenesis) is a Darwinian evolutionary process where tumor cell populations mimic a specie and tumor microenvironment represents nature environment. Analogous to neo-Darwinian evolution, also referred to as Darwinism in the context of genetics, cancer evolution is a branched evolutionary process based on: (i) the acquisition of heritable genetic variants in single cells by random alterations; (ii) natural selection acting on the phenotypic cellular diversity, either wiping out cells with acquired deleterious alterations (negative selection) or fostering cells with alterations that proliferate and survive more effectively than neighboring cells (positive selection); and (iii) gradual accumulation of the selected variants across each individual life-span[5,35]*.*

3.1 Driver and passenger alterations

As explained before, somatic alterations are the cause underlying virtually all cancers. However, not all somatic alterations lead to a malignization process. Indeed, because most tumors are genomically unstable[36] -they possess a tendency to accumulate genomic alterations along cell cycles- and may bear thousands of genomic alterations, it is likely that not all of them contribute to tumorigenesis[5,10].

Only those alterations conferring the cells biological capabilities which improve their survival in the microenvironmental context would be the ones positively selected. These biological capabilities, named hallmarks, go beyond from uncontrolled cell growth. There are eleven well-known cancer hallmarks that represent the complexity of all the processes that may be altered in cancer[37]: evasion of cell growth regulators (e.g. by loss-of-function alterations in *TP53* or *RB*[38]); sustained proliferative signaling (e.g. by activating alterations in *EGFR*[39]); deregulation of cellular energetics, mostly through of Warburg effect (i.e. aerobic glycolysis which leads to the production of lactic acid[40]); resistance to apoptosis (e.g. by alterations in *BCL-2* family members[41]); generation of genomic instability (e.g. by loss-of-function mutations in *BRCA1* or *BRCA2*[42]); induction of angiogenesis (e.g. alterations in *VEGF*[43]); invasion and metastasis (e.g. by alteration extracellular matrix components[44]); promotion of inflammation; replicative immortality (e.g. by altering telomerases[45]); and avoidance of immune destruction (e.g through exposure of immune checkpoint proteins[46]).

Therefore, alterations found in tumor cells can be divided in two types, depending on whether they contribute or not to tumorigenesis. Two different terms have been coined to differentiate them:

- Passenger alterations[2]: alterations not implicated in tumorigenesis, which do not exhibit signals of positive selection. They occur due to the interplay between DNA damage processes and DNA repair mechanisms. Once a tumor is established, their generation increases due to tumor genomic instability[5].

- Driver alterations[3]: the alterations implicated in tumorigenesis, which confer a growth advantage to the cell bearing them. These

[2] The term passenger also extends to the genes which only bear passenger alterations, the passenger genes[10].

[3] The term driver is also used for genes. Driver genes are those bearing at least one driver alteration, but can also bear passenger alterations[10].

alterations exhibit signals of positive selection, and they either activate genes, known as oncogenes (OGs) (e.g. *KRAS*, *BRAF*), that promote processes which lead to cell proliferation and survival; or cause the loss-of-function of genes that prevent the previous processes to happen, known as tumor suppressor genes (TSGs) (e.g. *APC*, *TP53*). In detail, the alterations of these two types of driver genes, OGs and TSGs, are different. While missense mutations in specific regions of the protein (known as mutational hotspots), gene amplifications and gene fusions confer gain-of-function properties to OGs; truncating mutations and deletions lead to the loss-of-function of TSGs[10].

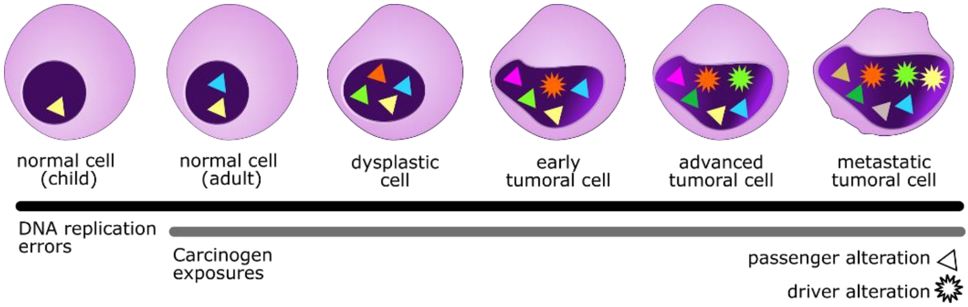


Figure 4 | Accumulation of somatic alterations across a tumor cell life-span. The accumulation of somatic alterations in the body cells starts right after birth, at a rate which depends on the DNA errors that emerge and the extent to which they are repaired. Initially, somatic alterations are benign, passengers, but under environmental selective pressure they could be positively selected and became drivers. Even after the malignization of the cells, they keep accumulating more mutations along its life-span, both passenger and drivers [This figure is an adaptation from Stratton et al. (2009)].

3.2 Computational identification of driver genes

During the past four decades, one of the main goals of the genetic study of cancer has been the identification of all the mutational driver genes[47]. The first driver genes were revealed by individual low throughput genetic and biochemical studies[48,49]. After two decades of experiments identifying cancer driver genes (CDs), Futreal and colleagues produced in 2004 the first manually curated consensus list of CDs, as reported in publications, named Cancer Gene Census (CGC), now containing approximately 600 genes[50].

More recently, international cancer genomics initiatives sequencing large cohorts of tumors served the purpose of identifying many more driver genes using statistical analyses (see below). These large cohort analyses were made possible by using next-generation sequencing technologies in large tumor cohorts, and provided the opportunity of expanding the catalog of CDs. Of note, among the sequencing consortia, the biggest ones, in terms of number of sequenced samples and number of different cancer types included, are:

(i) The Cancer Genome Atlas (TCGA), an American collaborative effort that began ten years ago and until now has generated multi-dimensional genomics data -through transcriptome profiling, exome sequencing, copy number alteration profiling and DNA methylation analysis and other techniques- for 33 cancer types and 11 thousand patient tumor samples[51].

(ii) International Cancer Genome Consortium (ICGC), a world-wide collaborative effort which started close in time to TCGA and until now has collected data for 16 thousand patient tumor samples from 21 cancer types. In contrast to TCGA, ICGC is mainly devoted to study of mutations, through exome sequencing. However, it has also data of transcriptome profiling along with a recent subset of tumor samples (around 2500) analyzed through whole genome sequencing[52].

It has been shown that most of the CDs are altered at low frequency, and that the set of genes driving tumorigenesis varies between cancer types. Therefore, data of large cohorts (such as TCGA and/or ICGC) is needed to evince a comprehensive catalog of CDs[5,10,53]. Currently, there is not a gold-standard computational approach for the detection of CDs (either mutational or with chromosomal rearrangements), as most of the developed methods have some drawbacks and/or biases[47]. However, most current approaches are based on the same principle, the detection of signals of positive selection through the evaluation of somatic alterations across tumor cohorts[47,54].

3.2.1 Identification of mutational driver genes

The first methods aimed to detect mutational cancer driver genes date from 2006. These methods were based on the detection of genes more mutated than a background mutation rate, that was corrected for gene size, among other variables, aimed to represent the mutational processes ongoing in the cell that may influence the mutational rate[55,56]. Later, similar approaches have been developed, mostly focused on improving the mutational background model, adding other variables known to affect the mutation rate such as gene expression or replication timing (e.g. MuSiC, MutSig)[57,58]. However, these methods are biased towards the detection of frequently mutated CDs, making difficult the detection of the lowly frequently mutated ones[47].

Other alternative approaches, not focused on the detection of more frequently mutated driver genes, have been developed. These mutational driver identification methods are aimed to detect genes with a particular composition of mutations, with respect to the total of mutations in the gene, named ratiometric methods[54]. Thus, there are several types of ratiometric methods depending on the mutational composition evaluated:

- Mutations with specific consequence types. These methods consider as CDs the genes with a certain ratio of mutations with a specific consequence type(s). Examples of these methods are: 20/20 rule from Vogelstein et al. (2013), that considers as CDs those above the 20% threshold of the oncogene score (proportion of recurrent missense o indel mutations out of the total of mutations) and tumor suppressor score (proportion of truncating mutations out the total of mutations); 20/20+ from Karchin et al. (2016), a RandomForest classifier based on the mutational attributes evaluated by Vogelstein et al. (2013); and TUSON from Davoli et al. (2014)[59], another machine learning approach which considers similar mutational ratios (e.g. proportion of truncating mutations out of synonymous mutations) and also classifies the CDs as OGs or TSGs.

- Clustered mutations. These methods identify the genes that tend to accumulate mutations in certain regions (i.e. clusters or mutational hotspots) with a higher frequency than expected from the background mutation model. There are two main types of clustering methods, those which perform clustering in the 2D sequence of a protein and those which perform it in the 3D protein structure. For example, OncodriveCLUST considers that a gene is a CD if the distribution of its PAMs in its 2D protein structure tends to be more clustered than the distribution of its synonymous mutations[60]; while CLUMPS identifies as CDs those genes with an overall enrichment of mutated residues spatially close to each other in the 3D protein structure[61].

- Functional impacting mutations. These methods identify as CDs the genes which accumulate more high impacting mutations than expected given a background mutational model. Examples of these methods are OncodriveFM[62] and OncodriveFML[63]. Both aggregate the functional impact scores of individual gene mutations to identify the gene functional impact bias but the background models used are different: OncodriveFM builds the background model by sampling of the observed mutations in the analyzed cohort, whereas OncodriveFML builds the background model by simulating a set of mutations according to the mutational processes occurring in the cohort under analysis, or cohorts of the same cancer type. Both methods use functional impact scores either based on the effect of the mutation on the protein function (i.e. SIFT[64], Polyphen-2[65] and Mutation Assessor[66]); or based on the effect of the mutation in non-coding regions such as microRNA targets and transcription factor binding sites; allowing the discovery of non-coding driver genes (i.e. CADD[67]).

- Mutations in special residues. These methods identify CDs that are biased towards the accumulation of mutations in functionally important residues. Examples of them are: ActiveDriver, which identifies genes that tend to accumulate mutations in phosphorylation sites[68]; and eDriver, which identifies genes accumulating mutations in protein functional regions[69].

Recently, some benchmarking studies have appeared, aimed to compare the performance of CDs detection methods[54,70]. However, these benchmarking efforts are biased towards the prioritization of certain methods (e.g. methods trained with CGC genes), producing contradictory results. In contrast to these studies aimed to identify the best performing method, integrative approaches using several methods have also been proposed. These approaches are based on the assumptions that (i) different driver genes bear different signals of positive selection that can be identified through different approaches; and (ii) each method presents various sources of biases that can be reduced when combining their results. Postulating that the combination of complementary methods is thought to provide a more comprehensive catalog of CDs[47].

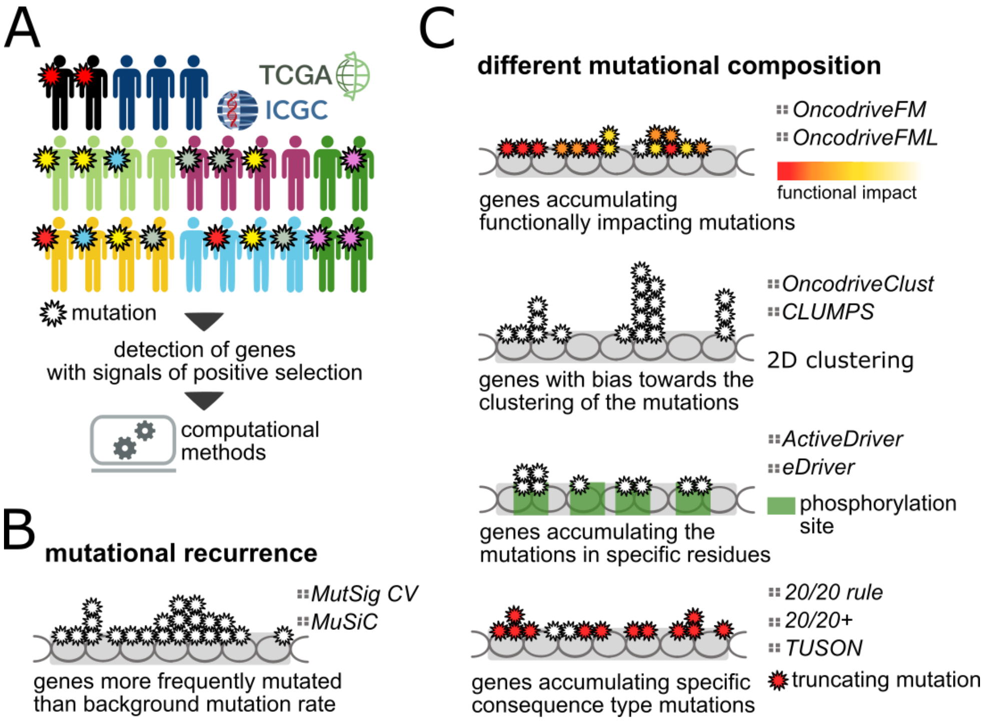

A
TCGA
ICGC
mutation
detection of genes
with signals of positive selection
computational
methods
B
mutational recurrence
MutSig CV
MuSiC
genes more frequently mutated
than background mutation rate
C
different mutational composition
OncodriveFM
OncodriveFML
functional impact
genes accumulating
functionally impacting mutations
OncodriveClust
CLUMPS
2D clustering
genes with bias towards the
clustering of the mutations
ActiveDriver
eDriver
phosphorylation
site
genes accumulating the
mutations in specific residues
20/20 rule
20/20+
TUSON
truncating mutation
genes accumulating specific
consequence type mutations

Figure 5 | Computational detection of cancer driver genes. (A) Schema of the common principle in which all computational methods detecting for cancer driver genes (CDs) are based, the identification of signals of positive selection across large tumor cohorts. There are two main types of methods for the identification of mutational CDs: those based on detecting genes more frequently mutated than a background mutation rate (B) and those which detect genes with specific mutational compositions (C); which can be: high functional impact mutations, clustered mutations, mutations localized in particular residues or mutations from a specific consequence type. (B) and (C) show a cartoon example on how mutations in a gene would be distributed so that each given method detects them as driver. Furthermore, examples of different methods of CD detection are also included.

3.2.2 Identification of driver genes bearing chromosomal rearrangements

Even if less abundant, there are also methods for the computational identification of cancer driver genes bearing chromosomal rearrangements.

On one hand, there are methods which detect cancer driver genes with CNAs, either amplifications or deletions. As for mutational approaches, the first developed methods detecting CNA drivers were based on frequency. These methods aimed to identify DNA regions with CNAs occurring at a significant frequency in a specific amplitude, when compared to a background rate (e.g. GISTIC)[71]. However, these methods identify DNA regions with CNAs that may contain large numbers of genes, not being clear which ones are the genes with CNAs providing the selective advantage to the tumor. To solve this hurdle, other methods aimed to simultaneously identify regions of focal copy number alterations together with gene expression changes have been developed (e.g. OncodriveCIS[72], FocalScan[73]).

On the other hand, several computational methods have been developed to detect the presence of fusion transcripts through the analysis of RNA-seq data[74–76]. However, even if fusion transcripts may be driver events, there are no computational methods based on the detection of signals of positive selection of gene fusions.

3.3 Tumor genomic heterogeneity

The study of TCGA and ICGC data has not only expanded the catalog of cancer driver genes but also deepen our understanding of how tumor genomes function. One of the first striking observations when analyzing the genome of thousands of tumors was the heterogeneity in the repertoire of altered CDs both between and within cancer types[53,77]. On one hand, it was observed that few cancer types were driven by a unique type of alteration. The observed general trend was intra-cancer type heterogeneity with most of the patients bearing alterations in a set of frequently altered CDs and additional alterations in a set of lowly frequently altered CDs[77] (e.g. TCGA analysis of ovarian serous carcinomas showed that, with the exception of *TP53* the genes identified are mutated in 10% or less of the patients)[78]. On the other, the frequently altered driver genes varied across cancer types (e.g. while most cutaneous melanoma samples are *BRAF* mutant, ovarian serous carcinomas frequently bear mutations in *TP53*)[78,79].

In addition to the heterogeneity at the level of driver genes, it was also observed the alteration level. The heterogeneity of alterations involves diversity in terms of: (i) alteration type, (ii) number and (iii) distribution across the genome[77]. (i) Alteration type heterogeneity (e.g. predomination of chromosomal rearrangements vs mutations) has been observed at inter-cancer type level; for example, while chromosomal rearrangements are frequent in leukemia, the tumorigenesis of cutaneous melanomas is mostly driven by mutations[77]. (ii) Alteration number heterogeneity are differences in the burden of alterations found inter- and intra-cancer type. For example, Lawrence et al. (2013)[80] observed, across 27 cancer types, a high variation of the mutational frequencies, ranging from a median of 0.1 mutations per megabase in the genome (i.e. one change across the entire exome) in pediatric cancers to a median of 100 mutations per megabase in cutaneous melanoma and lung carcinomas, related to exposure to carcinogens. Moreover, intra-cancer type mutational frequency heterogeneity was also observed (e.g. cutaneous melanomas and lung carcinomas showed a mutational frequency ranging from 0.1 to 100 mutations per megabase). At last, (iii) alteration distribution heterogeneity has been observed as changes in the distribution of mutations across the tumor genome. Lawrence et al also observed that certain mutation types (e.g. C to T) were not homogeneously distributed across the tumor genome between cancer types. For example, cutaneous melanomas showed a mutational spectrum dominated by C to T mutations, caused by unrepaired pyrimidine dimers induced by UV light; conversely C to A dominated the spectrum of lung carcinomas, caused by the exposure to tobacco smoke. Further study on the biological processes underlying different mutations led to the definition of cancer mutational signatures -different combinations of mutation types generated by different mutational processes[81].

4. Selective pressure during tumorigenesis

Tumors are not formed by a single population of cells -all of them genomically and phenotypically equal- but from multiple cell populations. Tumor clones are subpopulations of tumor cells with the same phenotype and genomic driver alterations that emerge because of the accumulation of different driver alterations along the tumorigenesis[82]*. These tumor clones are shaped by the tumor microenvironment, which causes a selective pressure and consequent competition where the clones with the best biological capabilities survive*[77]*.*

4.1 Tumor dynamic clonal evolution

The clonal architecture of tumors -the number of clones, their nature and their preponderance- is not constant across tumor evolution. The theory of clonal evolution states that a tumor starts with a founder clone, that arises as consequence of the accumulation of driver mutations. After the first cells became tumorigenic, additional alterations accumulate over time. The alterations conferring the tumor cells biological capabilities (i.e. cancer hallmarks), that can be shaped by the tumor microenvironment at different time points, will be selected and new clones of the tumor may then appear and expand (becoming major clones)[77]. On the contrary, those clones without biological capabilities that allow them to survive will shrink (becoming subclonal) and eventually they may disappear. For example, when a tumor starts an invasive process it acquires capabilities of invasion and metastasis but after it reaches the new microenvironment (i.e. metastasizes), cells with mutations providing capabilities that allow a good implantation (such as angiogenesis induction) will be selected, the clones bearing them will expand, and those clones with migration capabilities will shrink[37,77,82].

In addition to intrinsic tumor adaptation to the human body, there are external factors, such as anti-cancer therapies, which exert a selective pressure on the tumor[77,83]. Hence, once a patient starts a round of treatment, all tumor clones may die because of the treatment -generating a complete disease remission- but some cancer cells with mutations that allow them to survive the treatment may remain. In this last scenario, clones with driver alterations that confer them resistance capabilities to the treatment will expand, causing a disease relapse[5,77]**.** Many studies are emerging providing solutions to overcome specific drug resistances (described in further sections). Well-known examples of anti-cancer therapies resistance alterations include: *EGFR* T790M resistance mutation to first generation *EGFR* inhibitors (e.g. Erlotinib)[84], *ABL* T315I resistance mutation to *BCR-ABL* inhibitors (e.g. Imatinib, Nilotinib, Dasatinib)[85] or *KRAS* resistance mutations to *EGFR* antibody inhibitors (e.g. Cetuximab, Panitumumab)[86].

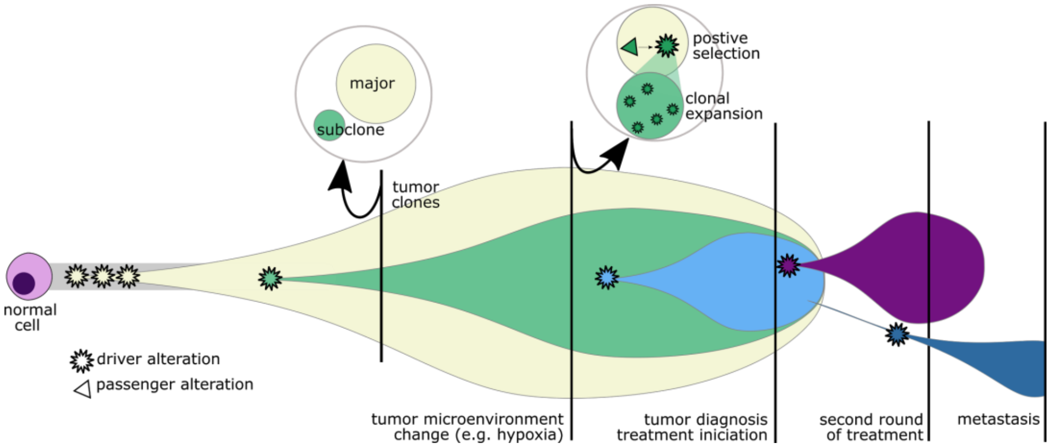


Figure 6 | Tumor dynamic clonal evolution. Tumor clones, each one represented as a colored bubble, emerge due to the accumulation of driver alterations. The dynamic clonal evolution is mostly shaped by clonal competition and environmental selection pressures. Thus, depending on the time-point of tumor life-span some clones will be larger (major) and some smaller (subclones). Note that changes in the environment (i.e treatment initiation) induce the positive selection (from passenger to driver) of alterations which confer a growth advantage, next undergoing a clonal expansion [This figure is an adaptation from Yates and Campbell (2012)].

4.2 Selective pressure from the immune system

Many studies have been devoted to get insights into the molecular basis underlying cancer hallmarks (some examples have already been cited previously). However, during the last few years, due to the success of anti-cancer immunotherapies (discussed in further sections), cancer research community is shifted towards the study of the interaction between the tumor and the immune system, involving “tumor promoting inflammation” and “avoiding immune destruction” Hanahan and Weinberg hallmarks.

The first observation that the immune system could recognize and attack tumor cells dates from the 1950s[87]. Beyond that, we know that relationship between the immune system and tumor cells is dual. On one hand, the immune system suppresses tumor growth by attacking tumor cells (e.g. through CD8+ cytotoxic T cells). On the other hand, the immune system exerts a selective pressure on the tumors that leads to the selection of tumor cells capable of surviving the immune system attack (e.g. tumor cells presenting *PDL-1*). This dual process is known as immunoediting[88].

The interaction between the immune system and tumor cells is a complex state of dynamic equilibrium[87]. Being a cyclic process where pro-stimulatory immune factors can enhance anti-tumor immune responses; but regulatory mechanisms, triggered by the tumor and its microenvironment, can in turn limit the immunological response[89,90]. According to Chen and Mellman (2013) this cycle can be divided in 7 major steps:

1) Release of antigens by tumor cells and capturing of those by dendritic cells (DCs). It is worth to point out that tumors release antigens different to the ones naturally exposed in normal tissues to which the immune system is self-tolerant. Non-normal tumor antigens come from three different sources: mutated peptides with aberrant conformations, named neoantigens; cancer-germline antigens, which are not expressed in normal tissues but tumor cells may express them due to DNA methylation changes; and viral proteins, expressed if the tumorigenic processes is influenced by a viral infection.

2) Presentation of the tumor antigens by the DCs and migration to the lymph node.

3) In the lymph node, priming and activation of effector cells, CD8 T cells and NK cells. Of note, effector CD8 T cells are primed with tumor antigens[89].

4) Migration of the activated effector cells to the tumor through the bloodstream, named trafficking.

5) Infiltration of the effector cells into the tumor bed (i.e. the normal tissue in which the tumor is located).

6) Recognition of tumor antigens through HLA molecules and binding by effector T cells.

7) Cytotoxic killing of cancer cells by effector cells which produces the release of tumor antigens (that again leads to step 1).

As mentioned, in cancer patients this cycle is impaired. Inhibition or impairment of the cycle can happen at any step: (1-2) tumor antigens may not be detected by dendritic cells; (2) priming of dendritic cells may treat the antigens as self, triggering T cell regulatory responses; (3) activation of T cells may not effectively traffic to the tumors; (5) effector populations might also be inhibited to infiltrate (e.g. due an immunosuppressive tumor microenvironment through the release of pro-angiogenic factors); (6) tumor cells may not be recognized (e.g. by occultation of the tumor antigens); or (7) the killing of the tumor cells by the immune effector cells could be impaired (e.g. through the inhibition of bind effector T cells by the tumor through checkpoint molecules such as *PDL-1* or *PDL-2*)[90,91].

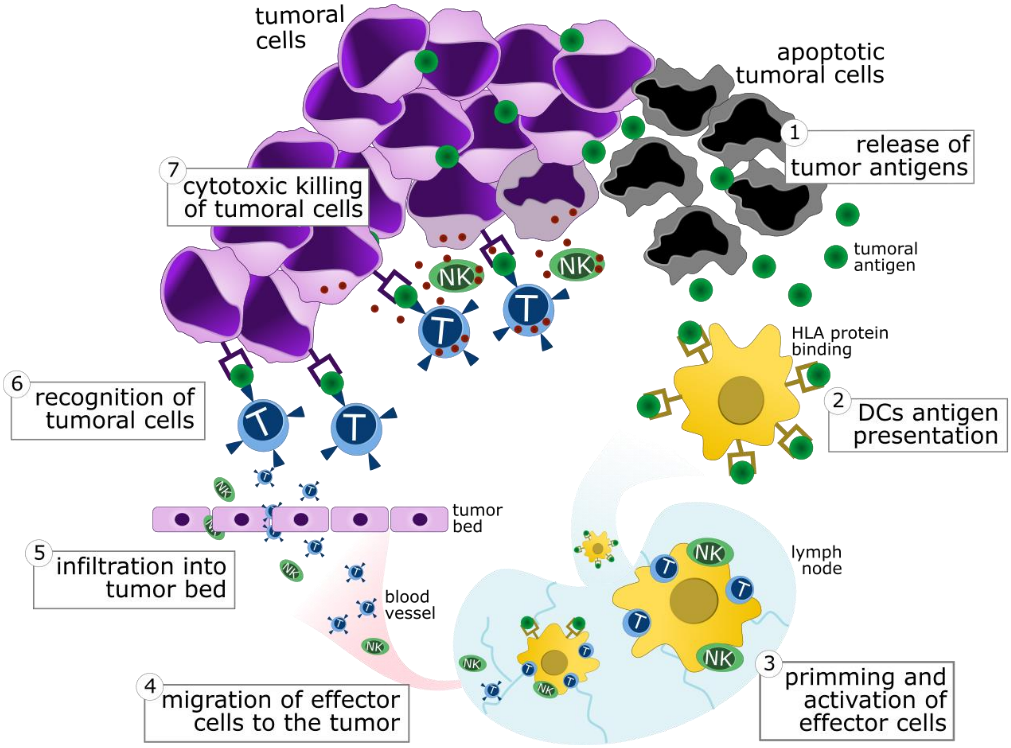

tumoral cells
apoptotic tumoral cells
1 release of tumor antigens
tumoral antigen
HLA protein binding
2 DCs antigen presentation
lymph node
3 priming and activation of effector cells
4 migration of effector cells to the tumor
blood vessel
5 infiltration into tumor bed
tumor bed
6 recognition of tumoral cells
7 cytotoxic killing of tumoral cells
NK
T

Figure 7 | Tumor-immune system interaction cycle. Schema of the cyclic interaction between the tumor and the immune system which can be divided in seven steps: (1) release of tumor antigens (represented as green circles) from the apoptosis of tumor cells, (2) presentation of tumor antigens to dendritic cells (DCs) which recognize them as non-self, (3) migration of the DCs into the lymph node and activation of effector cells, (4) trafficking of the active effector cells to the tumor through the bloodstream (represented T cells as T and NK cells as NK), (5) infiltration of the effector cells into the tumor bed, (6) recognition of tumor cells by effector cells, precisely T cells recognize tumor cells through HLA molecules; and (7) cytotoxic killing of tumor cells by NK cells and CD8 T cells through perforin and granzyme molecules (represented as red circles).

The exact mechanism underlying the tumor evasion of several cycle steps or the reason why the impairment of the cycle is heterogeneous across cancer patients is still a key problem to be solved. Tumor genomic heterogeneity has been postulated as a possible explanation for the heterogeneous response of cancer patients to immunotherapies[92]. Indeed, molecular tumor heterogeneity could have an impact on most of the steps of the tumor-immune system cycle. For example, it has been observed that tumors with truncating mutations in *B2M* lose the expression of HLA molecules in the cell surface, escaping from the recognition of T cells (step 6)[93], or that tumors with activation of *WNT/bCatenin* pathway show an impaired recruitment and activation of dendritic cells from a specific lineage (step 1-2)[94]. Thus, some studies are shedding light on the tumor evasion of the immune system but still a lot of effort is needed to understand the whole picture.

5. Cancer patient tumor profiling

The starting point of the detection of tumor genomic, transcriptomic and proteomic alterations is a tumor biopsy. After it is obtained, there are bunch of different experimental techniques that can be used to identify tumor alterations. Some of these techniques are currently being used in the clinical practice and others are mostly devoted to cancer research. Tumor profiling techniques have been typically classified into cytogenetic and molecular techniques, according to the molecular structures where they identify alterations, from chromosomes to DNA/RNA sequences, respectively.

5.1 Detection of chromosomal rearrangements

Cytogenetic techniques detect alterations at chromosome level, i.e., chromosomal rearrangements. Karyotyping is the most simple and cheapest cytogenetic technique, it classifies the 23 pairs of human chromosomes, allowing a study of the DNA amount in the whole genome (e.g. it allows identifying a deletion of an entire chromosome). However, its resolution is low, so short alterations cannot be visualized[95].

Fluorescence In Situ Hybridization (FISH) is another cytogenetic technique that allows to localize DNA specific sequences on chromosomes, cells or tissues; through the use of known fluorescent probes (i.e. specific DNA sequences). It is more suitable to identify chromosome translocations than karyotyping, as its results are easier to interpret[96]. Indeed, it is used in oncology clinical practice to identify chromosome translocations (e.g. identification of *BCR-ABL* translocation in chronic myeloid leukemia patients to prescribe treatment with imatinib[97]).

Comparative Genomic Hybridization, known as CGH array, appeared as novel cytogenetic technique, with better resolution. It used in the research and clinical context and allows the identification of unbalanced chromosomal rearrangements. CGH array consists of a series of wells with probes that map to different genome regions, covering the whole genome. When the DNA of study is added into the wells, it reacts with the probes and generates an assorted color light depending on the amount of DNA of study hybridized, distinguishing between amplifications and deletions[98]. A variant of CGH array named Single Nucleotide Polymorphism (SNP) array, has gained interest during the last decade. The technique is the same but probes contain a series of human polymorphisms. Hence, this technique also allows to perform polymorphism genotyping and can discriminate heterozygous alterations from homozygous ones[99].

5.2 Detection of DNA sequence alterations

The classical technique to detect DNA mutations is Sanger sequencing. This technique appeared in 1977 and is based on the selective incorporation of modified nucleotides by a DNA polymerase during an *in vitro* DNA replication. Next, these modified nucleotides are detected through gel electrophoresis and fluorescence, being the DNA sequence revealed[100]. Sanger sequencing is still the gold-standard of sequencing, currently being used in the clinical setting for the detection of point mutations (e.g. identification of *BRAF V600E* mutation in cutaneous melanoma patients to prescribe vemurafenib[101]).

In 2001 the first human genome was published using Sanger sequencing, as a 13-years multinational and metacentric project[102]. After that, a necessity of sequencing more genomes emerged and innovative technologies less costly and more efficient were required. This is how by 2005 next-generation sequencing (NGS) technologies, also known as high-throughput sequencing techniques, appeared. Since then, several platforms of NGS have been developed (e.g. Roche pyrosequencing, Illumina sequencing, Life Technologies SOLiD, etc), being all of them capable of sequencing simultaneously millions of DNA fragments in a massive parallel way. The principle underlying all NGS techniques is the same: first, the whole genome or exome of study (e.g. a tumor) is fragmented into millions of pieces; next, each of them is sequenced independently in parallel; generating a large volume of short-read sequencing data[103]. This read-based approach allows detecting not only mutations but also copy number alterations, by analyzing the amount of reads in a gene or DNA region, and gene fusions, through the identification of fusion genes.

A complex computational framework is needed to store, manage and analyze all the short-read data generated after sequencing. Using as an example the sequencing of the whole exome of a patient's tumor in a clinical context: (1) the resulting reads of the sequencing have to be aligned to the reference human genome; (2) tumor somatic variants have to be distinguished from the germline variants of the patient; (3) the quality of the somatic calls quality has to be assessed and bad quality variants filtered out; (4) functional impact of the somatic variants has to be annotated; and (5) among all somatic variants functionally relevant, a prioritization of the driver tumorigenic variants or the ones which can benefit from a therapy is needed to transfer the knowledge obtained back into the patient care[104]. However, from all this computational framework, there is no a gold-standard methodology or resource, particularly for step (6), becoming a bottleneck when trying to apply NGS strategies into the clinical context.

Because of that, whole genome and exome NGS is mostly used for research purposes. However, targeted sequencing through gene panels, another NGS strategy, is already becoming a standard tool for clinical oncology in some reference hospitals (e.g. MD Anderson, Vall d'Hebron). Gene panels possess a higher sensitivity and they are cheaper than performing a whole genome or exome[105]. Moreover, with respect to Sanger sequencing they allow to identify not a single mutation but a set of mutations, still limited, facilitating its interpretation. However, its design -the decision of which genes and gene regions should be included- is not trivial, it requires a laborious search in the literature, and even though there are some commercial solutions there is a not a gold standard to design them in each specific context.

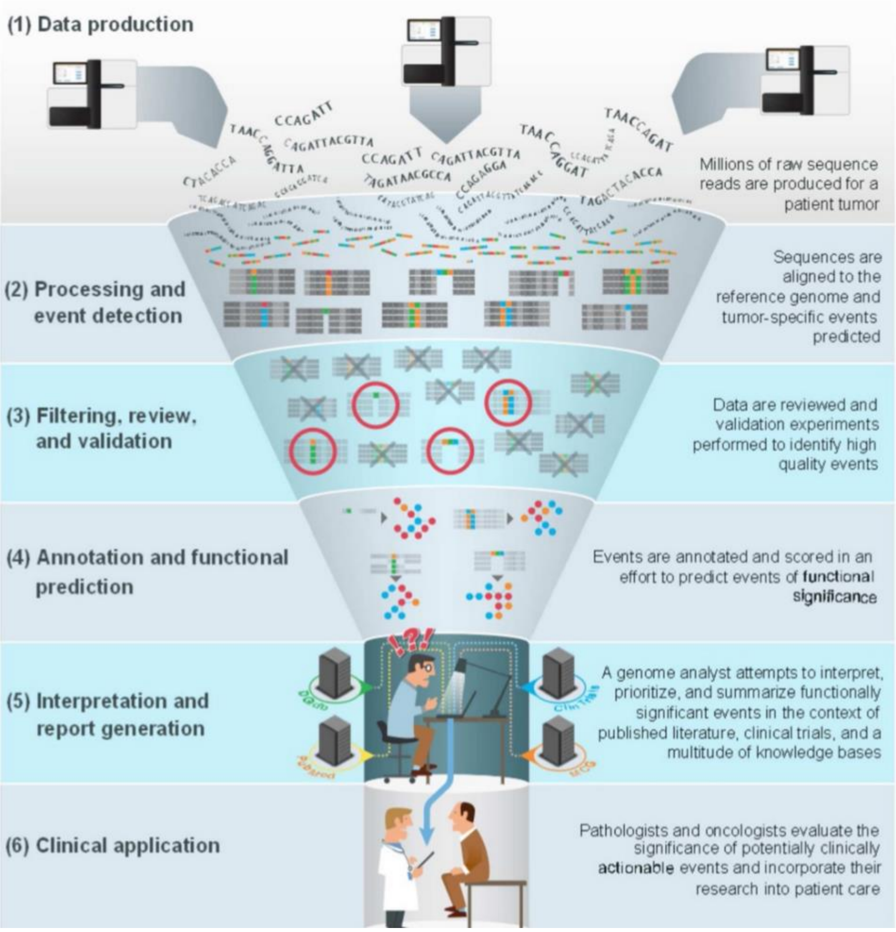

(1) Data production
Millions of raw sequence reads are produced for a patient tumor
(2) Processing and event detection
Sequences are aligned to the reference genome and tumor-specific events predicted
(3) Filtering, review, and validation
Data are reviewed and validation experiments performed to identify high quality events
(4) Annotation and functional prediction
Events are annotated and scored in an effort to predict events of functional significance
(5) Interpretation and report generation
A genome analyst attempts to interpret, prioritize, and summarize functionally significant events in the context of published literature, clinical trials, and a multitude of knowledge bases
(6) Clinical application
Pathologists and oncologists evaluate the significance of potentially clinically actionable events and incorporate their research into patient care

Figure 8 | From NGS of a patient's tumor to precision medicine. Common computational workflow from DNA NGS to its application on the clinics. Steps one to four are relatively automatic, while step 5 is currently a hardly manual step where a genome analysis should search in multiple scattered resources and integrate all the information in the context of the tumor being analyzed. Thus, step 5, as represented in the figure, is the bottleneck between NGS and the application of its results into the clinics [Adopted from Good et al 2014] .

5.3 Measure of gene expression

The gold standard technique to measure gene expression has been during several years microarrays, they are still being used but in recent years RNA sequencing has become also a widely used technique to measure gene expression.

On the one hand, DNA microarray techniques (also known as DNA chips) are based on the collection of series of probes into a surface that when mixed with cellular RNA they hybridize. The quantification of the hybridization events, through the incorporation of fluorescence or biotin labeled nucleotides, allows to measure gene expression as well as genotype a number of DNA regions[106].

On the other hand, RNA sequencing (RNA-seq) is based on the fragmentation of the cellular RNA, which is next converted into cDNA (complementary DNA) that after is prepared as a library (including adaptor proteins) and lastly sequenced in a high-throughput manner[107]. These techniques allow the quantification of all cell transcripts, including the product of fusion genes -chimeric transcripts. As mentioned before for DNA NGS, a complex computational framework is required after RNA-seq results are obtained. This framework, which starts also with read counts, involves: (1) a quality control for detecting sequencing errors or contamination artifacts; (2) an alignment of the reads to the reference genome; (3) a quantification of the read counts aligned to each transcript; and (4) a normalization by transcript length, which may have some variations. The most common normalizations are: (i) RPKMs (reads per kilobase per million mapped read), which in addition to transcript length also normalize by the cDNA library used[108]; (ii) its derived measurement FPKM (fragments per kilobase per million mapped read), used for paired-end sequencing, consider that 2 reads correspond to a single fragment[109]; and (iii) TPMs (transcripts per million mapped read), which normalize by a constant variable instead than cDNA library, 1 million transcripts[110].

5.4 Profiling of the tumor microenvironment

A tumor biopsy, the starting point for its genomic analysis, is not only formed by tumor cells. It is an admixture which also contains the tumor microenvironment (fibroblasts, immune cells, endothelial cells and normal epithelial cells)[111]. The fraction of cells from the admixture that are tumor determines the purity of the sample.

The sample purity is usually inferred by a pathologist, through a slide image analysis of the biopsy. When analyzing tumor sequencing data, a minimum of purity is usually required (e.g. the international sequencing consortium The Cancer Genome Atlas set the threshold at 60% of purity) to consider that the signal from the tumor can be distinguished from the one of the microenvironment[112]. However, it has been proved that differences in the level of purity across tumor samples have an impact on the interpretation of the genomic analysis, especially in RNA-seq data analysis. Thus, the correction of RNA-seq data by purity has been shown to reveal masked pathways or decrease the expression of pathways mostly overactivated in the microenvironment, not tumorigenic[112,113].

On the other hand, because tumor samples are admixtures, once sequenced we can analyze not only tumor cells but also the cells of the tumor microenvironment. Because of this, computational methods capable of quantifying the immune infiltrate (e.g. ESTIMATE[114]), and the individual cell populations infiltrating tumor samples have emerged during the last years[115–119]. Among the latter two main types of approaches have been developed: deconvolution and gene set enrichment methods. Deconvolution methods compute the proportion of each immune population within the overall set of immune cells infiltrating the tumor[118,119], while enrichment methods provide the relative estimate of the overall enrichment of the immune populations of interest[115–117]. There is no a gold standard methodology among both approaches. However, several caveats have recently been reported for deconvolution methods. For example, they have not been validated for RNA-seq or have been shown to be not robust when the expression from few genes of their training matrix cannot be assessed[120].

6. Personalized cancer medicine

Classical pharmacological cancer therapies, chemotherapies, do not consider the intrinsic features of the patient's tumor because their mechanism of action is not specific. These treatments kill all the cells (i.e. they are cytotoxic compounds) in the human body with a high replicative rate, causing an important toxicity. Personalized cancer medicine has emerged as a new therapeutic strategy that looks for the most suitable pharmacological treatment that, considering the biology of the tumors, is capable of blocking cell proliferation (i.e. they are cytostatic compounds). Therefore, when compared to classical therapies, the effectivity of personalized strategies is meant to be higher and the toxicity lower, improving patient care.

6.1 Genomics-driven personalized treatments

Cancer personalized treatments are not new, the earliest strategies date from the 1977, when the first hormone therapy for breast cancer, tamoxifen, was approved[121]. Hormone therapies are effective, and still are being used, for the treatment of cancer types whose growth is dependent on hormones (i.e. breast and prostate cancer). But even if effective, this kind of treatments could not be applied to most the cancer types. That is why other strategies, based on the inhibition of specific oncogenes promoting tumor growth was undertaken. These new strategies, referred to as targeted therapies, were based on what was described some years after as "oncogene addiction" principle[122].

"Oncogene addiction" is defined as the genomic dependency of the tumor on specific and few alterations to maintain the tumorigenesis active (i.e. tumor genomic *Achilles Heel*), it is based on the observations, across many years and experiments, that reversing one or few driver alterations in OGs inhibited the tumor growth[123,124].

Trastuzumab -a monoclonal antibody that selectively inhibits *HER2*- was the first successful targeted therapy introduced into the clinical setting in 1998 for breast cancer patients *HER2 positive*[125]. However, the mechanism of action of trastumzumab is not exactly as expected. The drug works through a dual mechanism that in one hand relies on the principle of “oncogene addiction” because inhibition of *HER2* arrests tumor growth, but on the other, it flags the cells for immune destruction when the drug binds to the tumor cell. It was not until the approval of imatinib in 2001, that the first prove of “oncogene addiction” principle to a particular genomic aberration -BCR-ABL gene fusion in chronic myeloid leukemia- was clinically used[126,97]. The success of imatinib provided compelling evidence that OGs could be potentially good drug targets and targeted treatments for other cancer types started emerging. Erlotinib was the next one to be approved, initially introduced in 2004 for the treatment of non-small cell lung which exhibited *EGFR* mutations[127–129].

Currently, there are more than 80 anti-cancer targeted therapies approved for specific cancer types and at least a third of them are used depending on the presence (or absence) of a genomic biomarker (e.g. gene fusion, mutation, deletion...) or proteomic aberration (e.g. HER2+)[130].

In most of the cases, targeted therapies are prescribed to patients in combination with chemotherapies. Besides, there are also combinations of targeted therapies which are prescribed to cancer patients. These combinatorial therapies are mostly designed to overcome drug resistance (see below) and have synergistic effects. An example is *BRAF* and *MEK* targeting in *BRAF* V600E mutant melanoma patients[131]. Of note, the main bottleneck of drug combinations is the emergence of additive toxicities[132].

6.1.1 Bottlenecks of anti-cancer therapies development

Even if 80 seems a large number of approved therapies, it is much smaller than the number of all therapies under investigation. Drug development is a slow process, having its biggest bottleneck when facing the clinical trials phase[133]. Specifically, most of the drugs fail due to unpredicted clinical toxicity, during clinical phase I testing[134,135].

Drug repurposing is a strategy aimed to solve the phase I bottleneck and has been widely used for cancer targeted therapies. Drug repurposing refers the re-use of approved drugs for prescriptions other than the approved one, either different molecular targets or different diseases. The approval of these strategies is usually faster, because they do not need to go over phase I again[133]. The first drug with a successful repurposing among cancer targeted therapies was imatinib, which in 2008 underwent fast approval for a new molecular target in another disease, *KIT* mutant gastrointestinal tumors[97].

The classical clinical trial design relies on the fact that molecular targets are associated to a specific disease, which has been demonstrated inaccurate now that the molecular profiling of large tumor cohorts is available. The molecular heterogeneity across tumors decreases the number of patients in which a genomic-directed therapy could be beneficial for a specific cancer type. This impairs the power to assess the outcome of these approaches. An innovative design of clinical trials emerged as a solution to this problem, named “basket trials”. Clinical basket trials design assumes that drug response is shaped by the tumor genomic alterations, not the cancer type, and thus a larger number of samples across different cancers can be pooled together. Moreover, there is variant of basket trials where several genomic alterations are analyzed together, allowing to explore several biomarker hypotheses and recruit more patients[136]. An example is CUSTOM (Molecular Profiling and Targeted Therapies in Advanced Thoracic Malignancies) where after testing 5 drugs across 11 types of molecular alterations the authors showed that certain genomic alterations shaped the response to erlotinib and selumetinib[137].

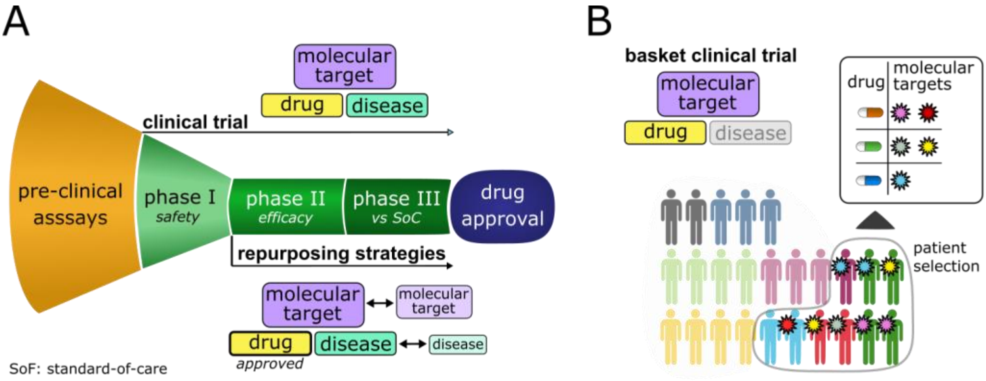


Figure 8 | Drug development process. (A) Diagram of the drug development process which starts with preclinical assays, where a lot of potential targets and compounds are investigated; followed by clinical trials (phases I to III) with a smaller number of investigational compounds that drops even more after phase I. In phase I drugs are tested for safety and toxicity, in phase II for its efficacy shaped by the molecular target and in phase III the efficacy is compared to the one of the standard-of-care. Drug repurposing are strategies meant to speed-up this process. They consist of re-using already approved drugs are for different molecular targets or diseases than the one approved, they do not need to go over safety tests again. (B) Schematic representation of the basis of basket clinical trials.

These trials select patients according their molecular alterations, rather than their disease type, and test different drugs according to the molecular alterations beard.

6.1.2 Resistance mechanisms to anti-cancer pharmacological treatment

Resistance mechanisms have been extensively described for cancer treatment, either for chemotherapies and targeted therapies. Resistance mechanisms can be intrinsic, when they are present already before the treatment (such as mutations in the target if it has to be wild type for an appropriate drug binding); or acquired, developed after treatment due to tumor adaptation[138].

While intrinsic resistances are easier to foresee, acquired resistances are more difficult to anticipate. That is why most of the research efforts have been put in these last type of resistance mechanisms. The main types of acquired resistance mechanism are the following:

1) Drug efflux and activation resistances. The first consist of an overexpression of cell membrane *ABC* proteins, which regulate the flux across the plasma membrane, preventing the cells from the internalization of the compounds. The second has been mostly observed in chemotherapy resistance (e.g. capecitabine) by epigenetic inactivation of enzymes which catalyze the conversion pro-drug (inactive form) to drug (active form) in the tumor cells. Both resistances can be solved through drug combination therapies, ABC family inhibitors and DNA methyltransferase inhibitors, respectively[138,139].

2) Drug target resistance mechanisms. This type of resistance is caused by alterations in the target of the drug either by a change in its conformation or an overexpression, leading to impaired drug effectivity[138]. On the one hand, changes in drug target protein conformation, caused by mutations, may prevent drug binding or allow the activation of the protein after drug binding. This latter type is produced by mutations in the so-called gatekeeper residues, which can stabilize the protein after drug binding, so that it keeps being functional[84,140]. The mechanisms of the gatekeeper mutations are different depending on the target and have been mostly described for *BCR-ABL, KIT, PDGFRA* and *EGFR*[141]. For example, *EGFR* T/M mutations in the gatekeeper residue 790 are reported to be responsible of the 50% of resistances to *EGFR* small molecule inhibitors; by increasing the binding affinity for ATP by *EGFR*. As another example of a different mechanism, resistance to *BCR-ABL* inhibitors by *ABL1* T315I mutation is suggested to cause resistance through the stabilization of the ATP-binding active conformation of the protein[142]. On the other hand, target overexpression reduces drug effectivity rather than generating a resistance. One example is the overexpression of the androgen receptor (*AR*) after the use of *AR* antagonists in prostate carcinoma[143].

3) Alternative pathway activation. This resistance arises as an adaptation to the oncogenic addiction of a tumor after inhibition of a pathway, leading to maintenance of the function. It is known as “oncogenic bypass” and it is becoming the major mechanism of resistance to targeted therapies[138]. *MET* amplifications acquired as a resistance mechanism to *EGFR* inhibitors serve as an example of these mechanisms. *EGFR* inhibitors hinder the activation of the PI3K-AKT pathway, by acquiring the cells amplifications in *MET* they

can make this pathway become active again[144]. A variant of this mechanism is the activation of pathways that evade the pro-apoptotic signal triggered by the drug, impairing the apoptosis process[138]. For example, changes in the expression levels of *BIM* pro-apoptotic molecule have been associated to varying degrees of response to *EGFR, HER2* and *PI3K* inhibitors[145].

6.1.3 Therapeutic strategies for tumor suppressor genes

All previously mentioned therapeutic strategies work under the principle that the tumor cell is addicted to the alteration of a gene that drives tumorigenesis through its activation, an OG. In 2000 a new dimension of "oncogene addiction" principle was described by Weinstein[146]. He called it "tumor suppressor hypersensitivity", and based its definition on the observation that reintroducing the wild-type version (i.e. non-genomically altered) of a TSG led to inhibition of cell growth and/or induction of apoptosis. Nevertheless, strategies for restoring tumor suppressor function are not as straightforward as the inhibition of oncogenes[147]. Several approaches can be tackled to therapeutically exploit genomic vulnerabilities presented by the loss-of-function of a TSG.

The first mechanism consists in indirectly targeting the TSG by inhibiting its negative regulators or oncogenes downstream its effect. On one hand, targeting negative regulators leads to an increase of the expression levels of the TSG. An example of these inhibitors are *MDM2* inhibitors which increase *TP53* protein levels and the activity of its targets (e.g. *CDKN1A*)[148] these inhibitors are currently in phase I/II trials for several solid tumors. Of note, this therapeutic strategy does not work if the TSG to be restored is mutated, as it will lead to expression of the non-functional form of the TSG[148]. On the other hand, inhibiting the targets negatively regulated by the TSG, would mimic the regulatory function of it. For example, the inhibitor of the PI3K-AKT-MTOR pathway everolimus has been shown to be effective upon *PTEN* deletions in prostate adenocarcinoma (a negative regulator of the pathway), being tested in phase I/II clinical trials[149]. Indeed, indirect targeting strategy is not only used for TSGs but also for OGS, as an example *MEK* inhibitors have been shown to elicit response in *BRAF* V600E mutant thyroid carcinomas[150].

The second mechanism, the reintroduction of the TSG, even if it seems the most straightforward strategy, it is difficult in terms of the therapeutic approaches to be used. In this direction, gene therapies for *TP53* have been investigated during many years. However, non-stable levels of efficacy have been reached, reason why *TP53* did not passed phase III trials in USA[147].

The third and last mechanism, is based on principle called "synthetic lethality". Synthetic lethality relies on "oncogenic bypass" phenomena mainly focused on the context of DNA damage repair pathways. If a tumor cell has an impairment of a DNA damage repair pathway (e.g. through a homozygous loss of the TSG *BRCA1*), it becomes addicted to another DNA damage repair pathway. If it becomes impaired too, this will lead to a cell death. Therefore, these lethal combinations can be therapeutically exploited either by inhibiting directly the pathway or enhancing DNA damage[147]. A successful example of synthetic lethality mechanism was the approval, in 2014, of the first *PARP* inhibitor, olaparib. Olaparib is an inhibitor of *PARP* enzymes, which are involved in cellular homeostasis, that when inhibited in *BRCA* deficient cells - that have impaired DNA damage repair- cause their death (in contrast with the other targeted therapies, it is a cytotoxic compound)[151].

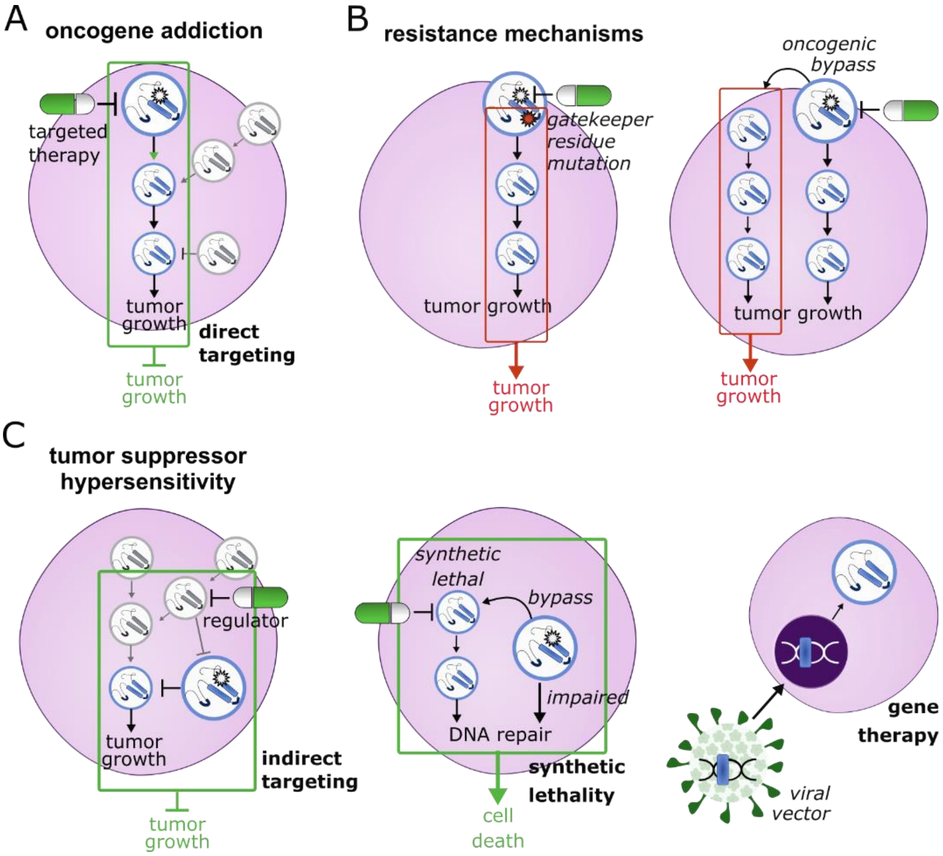


Figure 9 | Molecular mechanisms of targeted therapies. (A) Oncogene addiction principle, which explains the effectivity of targeted therapies that directly bind to OGs. (B) Mechanisms of resistance to targeted therapies. Only the two mechanisms widely associated to targeted therapies are represented: left, drug resistance by target alterations; right, drug resistance by activation of a pathway with the same effect than the one inhibited. (C) Targeting mechanisms of TSGs, based on the principle of tumor hypersensitivity: left, indirect targeting by inhibition of up-stream negative regulators; middle, synthetic lethality; and right, gene therapy to restore TSG expression. Each purple circle represents a cell and the picture inside and the blue circle represents a protein.

6.2 Personalized immunotherapies

Anti-cancer immunotherapies are a family of treatments aimed at stimulating the immune system of the patients to fight their tumors. The rationale behind immunotherapies is not new, it appeared in the 19th century, after the observation that the infectious processes, using bacterial vaccines, could provide anti-cancer therapeutic benefit[152]. Research on immunotherapies followed this direction through several decades, and in 1990 the first bacterial vaccine, with *bacillus calmette guerin*, was approved for *in situ* bladder carcinoma[153]. Some years before, in 1964, another strategy of immunotherapy was explored by infusing immune lymphocytes into a rat sarcoma, the positive results obtained were the beginning of the adoptive cell transfer (ACT) therapies[154].

ACT consists in administering tumor-specific T cells which have been expanded *ex vivo* and after are infused back to attack the tumor. The rationale for its effectiveness is based on T cell response robust specificity for tumor cells, as T cells can move to the tumor (even likely reaching distant metastases) and have a memory, guaranteeing the maintenance of the therapeutic effect after initial treatment[155]. The selection of the T cells that may attack the tumor for the *ex vivo* expansion can be done by selecting the T cells infiltrating the tumor (TILs), and then transferring a synthetic T cell receptor or a chimeric antigen receptor (CAR) into the T cells. CAR strategies are recently emerging as powerful therapeutics that have achieved remission rates up to 70-80% in hematologic malignancies[156,157]. Briefly, CAR-T cell strategies can recognize specifically programmed antigens independent of the HLA-complex, such as *CD19*, expressed in the cell surface of B cells, making T cells attack aberrant B cells from hematologic B cell malignancies[158].

Another type of strategy that emerged several years after ACT is the inhibition, through monoclonal antibodies (mAbs) of immune checkpoint molecules. As has been briefly described, checkpoint molecules are receptors or ligands which mediate the activation of T cells in the different steps of the cancer immunity cycle. The first step of the cycle involves the activation of T cells are through their interaction with DCs in the lymph node. This activation is a three-step process that requires *CTLA4* exposure on the T cell surface in the last step. If this happens at an earlier stage, it competes with co-stimulant molecules, leading to T cell inactivation. Moreover, T regulatory cells also use *CTLA4* to suppress the T cell function. *CTLA4* is therefore a checkpoint molecule whose inhibition contributes to T cell activation[ref]. Another well-known checkpoint is *PD1* that if interacting with *PDL1* or *PDL2* in the tumor cell leads to T cell inactivation. Again, the inhibition of either *PD1*, *PDL1* or *PDL2* may lead to T cell activation. During the last six years, five (nivolumab, avelumab, pembrolizumab, atezolizumab, and ipilimumab) checkpoint inhibitors have been approved, alone or in combination, for at least 6 different malignancies (merkel cell carcinoma, head and neck carcinoma, urothelial, non-small cell lung carcinoma, renal clear cell carcinoma and melanoma), mostly for its advanced or metastatic stages[130]. Moreover, additional checkpoints are being investigated as potential therapeutic targets (e.g. LAG3 or TIM3).

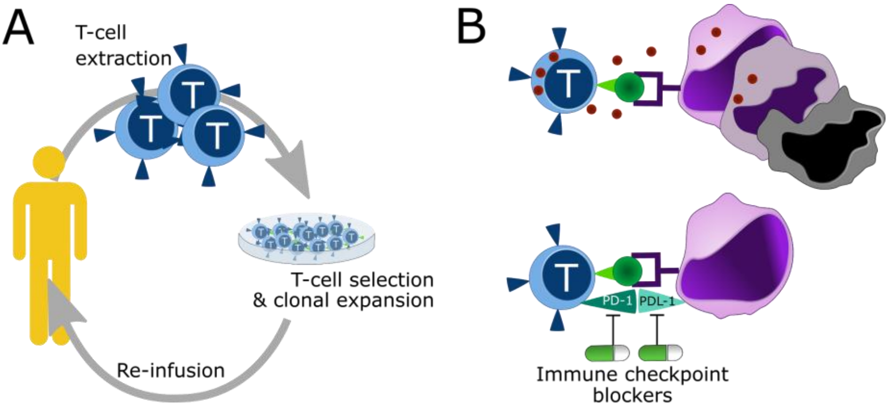
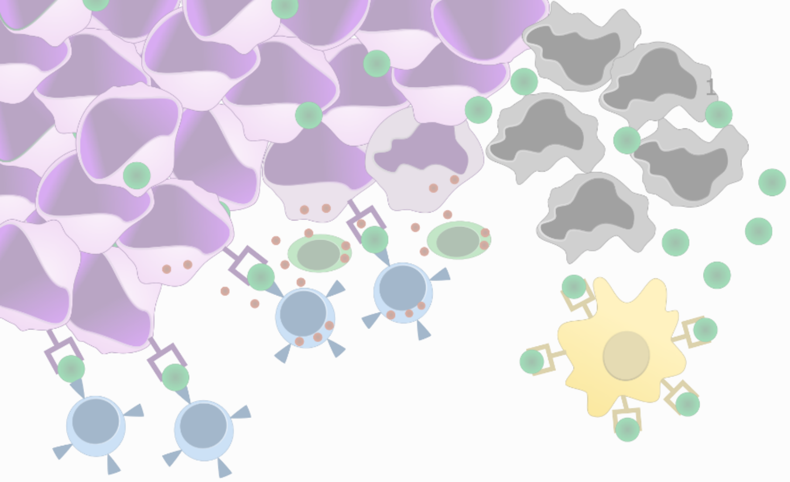


Figure 10 | Main types of immunotherapies used in the clinical setting. (A) Diagram of the workflow of an adoptive T cell-transfer treatment (B) Molecular rationale underlying immune checkpoint blocker therapies. Top diagram, attack of T cells to tumor cells causing tumor cell apoptosis through the release of cytotoxic molecules. Bottom diagram, inhibition of T cell attack by tumor cells through PD-1/PDL-1 interaction.

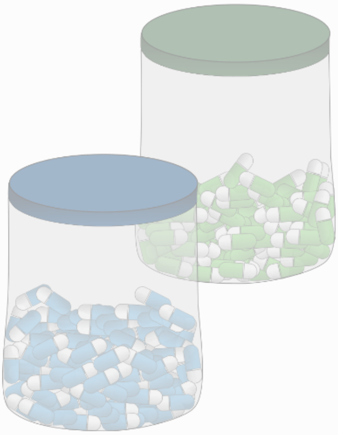
Concurrently with the development of more checkpoint inhibitors new combinatorial strategies of targeted therapies with checkpoint inhibitors are being tackled. On the one hand, checkpoint inhibitors have been evidence to be mostly effective only in high immunogenic tumors (such as melanoma and lung carcinomas) with high mutational burden that results into more tumor neoantigens and, in turn, into more TILs[158]. On the other hand, some targeted therapies have been shown to boost cancer immunity by influencing T cell trafficking or T cell tissue infiltration, such as MEK inhibitors or VEGF inhibitors[159]. Therefore, combinatorial therapies of checkpoint inhibitors and these targeted therapies may elicit a synergistic response, expanding the spectrum of patients who could benefit from checkpoint inhibitors. Some clinical trials in early phases are already exploring this possibility (NCT01940809, NCT01673854, NCT02224781, NCT02130466)

Finally, it is important to mention that resistances have been described for immunotherapies, just as the ones long known for targeted therapies. These mechanisms are less known than those for targeted therapies but some, such as up-regulation of checkpoint molecules different to the one inhibited[160] or acquisition of mutations which impair immunological response[93] have been described.



PART II

OBJECTIVES



The focus of my thesis is generating knowledge that contributes to the progress of cancer precision medicine through the analysis of cancer genomics data.

Genomics have been proven to be useful for guiding the treatment of cancer targeted therapies and there are large collections of tumor sequencing data currently available that provide a comprehensive view of tumor genomics across several malignancies. The **first objective** of my thesis is to *understand the current scope of genomic-guided personalized therapies*. This objective involves the following tasks:

- Identify the genes driving tumorigenesis in each cancer type via mutations, CNAs and chromosomal rearrangements.
- Build a comprehensive database of anti-cancer targeted therapies and the biomarkers of their effect on tumors.
- Develop a method to associate drug response and drug resistance biomarkers to tumor samples.
- Develop strategies of *in silico* drug repurposing.

Nowadays there is an urgent need of sequencing tumors in the clinical and research community. Cancer gene panels have emerged as a cost-effective solution to this necessity. However, with no guide to design these panels adapted to the specific needs of researchers, it is a manual and highly laborious task. The knowledge generated in the first objective on cancer type cancer driver genes, the therapeutic options and the mutational data compiled could be exploited to aid the design of cancer sequencing panels. Therefore, my **second objective** is the *development of an easy-to-use tool to support a rational design of cancer gene panels according to the user's needs,* which includes:

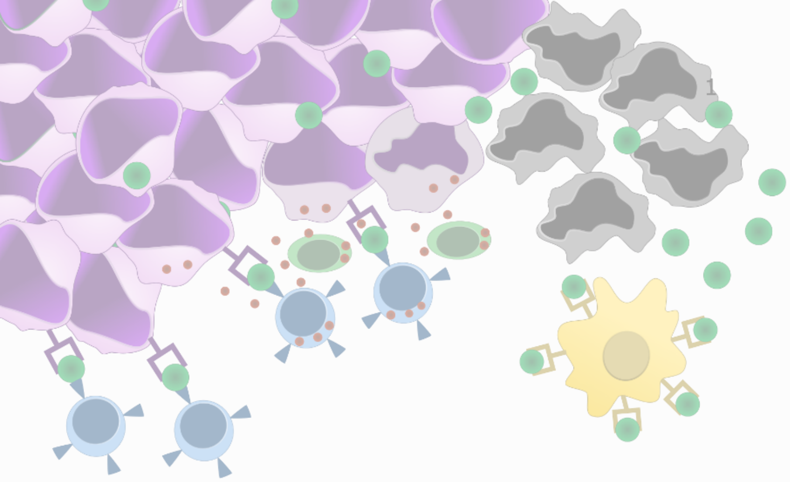
- Develop a method that prioritizes the genes and/or mutational hotspots with the highest mutational coverage in a cancer type or a group of them.
- Build a user-friendly web-tool to carry out the panel design.
- Design interactive reports integrating ancillary information that aids panel design.

Nevertheless, the bottleneck when sequencing tumor cells is to interpret the resulting data. We realized that the methodology developed in the first objective could be a starting point to solve it. Consequently, my **third objective** consists in *developing a tool capable of interpreting the relevance of somatic variants observed in a tumor, with a focus on the identification of those with therapeutic significance*. It includes these tasks:

- Improve the database of anti-cancer targeted therapies including the level of curation of the database and its extension with more biomarkers of drug response, resistance and toxicity.
- Build systematic nomenclatures for the classification of the genomic biomarkers, drugs and cancer types in the therapies database.
- Develop a method for matching drug biomarkers to tumor driver alterations considering interactions between genomic biomarkers and drugs.

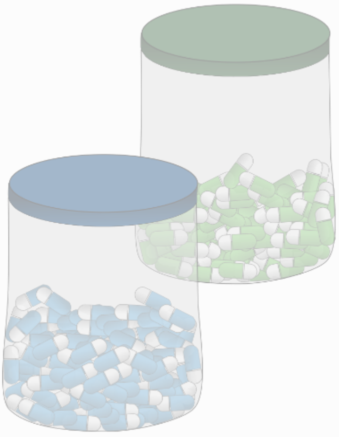
Finally, with the emergence of immunotherapies, its success and the lack of detailed knowledge in most steps of the interaction between the tumor and the immune system I directed my **fourth thesis objective** to *understanding molecular mechanisms related to tumorigenesis that modulate the anti-tumor action of the immune system*. The tasks to fulfill this last objective are:

- Develop and apply a method to identify immune subpopulations from the expression data of the tumor bulk sample.
- Define immunophenotypes given the profile of immune subpopulations in the tumor infiltrate.
- Identify correlates between the tumor architectures with the immunophenotypes.



PART III

RESULTS



Chapter 1

GENOMIC-GUIDED THERAPEUTIC LANDSCAPE OF CANCER

In the first chapter I present a comprehensive landscape of the therapeutic opportunities of a large cohort of cancer patients based on their genomic alterations. I have carried out this work together with Tamborero D, the other first co-author of the publication.

The work done in this chapter has been divided into three main steps: (i) identification of genes driving tumorigenesis across the 28 cancer types via mutations, copy number alterations and gene fusions; (ii) identification of drugs targeting the driver protein products; and (iii) *in silico* prescription of drugs to patients based on the driver events observed in each patient's tumor. The implementation of these three steps, in a cohort of 6792 samples from 28 different cancer types, has identified the most comprehensive therapeutic landscape of anti-cancer targeted therapies to date. In turn, this landscape has revealed interesting messages, such as that 40.2% of all cancer patients could benefit from drug repurposing opportunities.

From these three steps, I have developed step (ii) and step (iii); and Tamborero D has also contributed on the integration of some sources in step (ii). Specifically, I have built a comprehensive database of anti-cancer therapies targeting driver protein products. This database also includes genomic biomarkers of response to approved therapies, as stated in their clinical guidelines, and genomic biomarkers of drug resistance, either approved or in clinical trials. After building the database, I have developed a set of rules for: prescribing approved drugs according to their clinical guidelines pertaining genomic and cancer type annotations, repurposing approved drugs to different cancer types or genomic alterations, and considering the resistance biomarkers.

Next, I have developed a decision-algorithm, referred to as *in silico drug prescription*, that matches the drugs in the database to the alterations affecting driver genes, as identified by Tamborero D. The *in silico drug prescription* is able to take into account the genomic biomarkers of drug response and resistance, the cancer types to which the drugs are prescribed, the oncogenic role of the driver genes, and the mechanism of action of the drug.

Additionally, I participated in drafting the manuscript and preparing most of the figures and supplementary information.

Rubio-Perez* C, Tamborero* D, Schroeder MP, Antolín AA, Deu-Pons J, Perez-Llamas C, Mestres J, Gonzalez-Perez† A, Lopez-Bigas† N. (2015). In silico prescription of anticancer drugs to cohorts of 28 tumor types reveals targeting opportunities. ***Cancer cell*****,** *27*(3), 382-396.

* co-first authors † co-corresponding authors

Rubio-Perez C, Tamborero D, Schroeder MP, Antolín AA, Deu-Pons J, Perez-Llamas C, et al. In silico prescription of anticancer drugs to cohorts of 28 tumor types reveals targeting opportunities. Cancer Cell. 2015 Mar 9;27(3):382–96. DOI: 10.1016/j.ccell.2015.02.007

Chapter 2

RATIONAL DESIGN OF CANCER SEQUENCING PANELS

In the second chapter, I present a web-application aimed to rationally design next generation sequencing (NGS) mutational cancer panels. The web-application, named OncoPaD, designs NGS cancer panels for specific cancer types, or groups thereof, considering the role in cancer and therapeutic actionability of the genes included. By means of its prioritization algorithm, OncoPaD is able to identify which genes or mutational hotspots would increase more the coverage of the panel, converging to the most cost-effective solution. Moreover, the performance of OncoPaD panels, in terms of cost-effectiveness, is higher than that of commercially available panels, especially when focused on panels for specific cancer types or groups of them. OncoPaD is open-source and is available at http:/www.intogen.org/oncopad.

The work presented here was divided into two parts: the development of the web platform and the algorithm of selection and prioritization of genes and mutational hotspots. I have conceived and implemented both parts, with technical assistance by the second author of the publication, Deu-Pons J, in the web platform development. Additionally, I drafted the manuscript and prepared all the figures and supplementary information.

First, I implemented the algorithm, which I divided into five different parts: (1) sub-setting the pan-cancer cohort (7298 samples) by the cancer type(s) of interest, or panel cohort; (2) selection of the genes driving tumorigenesis in the panel cohort; (3) identification of the mutational hotspots in each gene in (2), by identifying the minimum number of base pairs regions across the sequence of the gene that contain most of its mutations; (4) computing the cumulative mutational frequency (CMF) distribution of the panel cohort (coverage), as the number of tumors bearing protein affecting mutations in the genes and mutational hotspots identified in (2) and (3), and prioritization of the genes or mutational hotspots which contribute more to tumorigenesis in Tiers 1 and 2; and (5) collection of the data to be displayed into the web platform.

Next, I designed and developed the web tool, mainly formed by an input and results sections. The input section allows the introduction of all the parameters needed to run the algorithm, or fine-tune its configuration. The results section is based on five reports: (i) CMF distribution in the panel cohort with additional information for each gene and mutational hotspot about the actionability; (ii) CMF considering more than one gene or mutational hotspots per tumor in the panel cohort; (iii) mutational distribution per each gene in Tiers 1 and 2; (iv) drug actionability details of the genes and mutational hotspots; and (v) general features of the genes, such as the mode of action in cancer or its clonality.

As a snapshot in time of its use, from 13th October 2016 until 15th of May 2017 OncoPaD has been accessed 794 times by 521 users; with a median of 12 sessions per week.

Rubio-Perez C, Deu-Pons J, Tamborero D, Lopez-Bigas N, & Gonzalez-Perez A. (2016). Rational design of cancer gene panels with OncoPaD. ***Genome Medicine***, *8*(1), 98.

Rubio-Perez C, Deu-Pons J, Tamborero D, Lopez-Bigas N, Gonzalez-Perez A. Rational design of cancer gene panels with OncoPaD. Genome Med. 2016 Dec 3;8(1):98. DOI: 10.1186/s13073-016-0349-1

Chapter 3

BIOLOGICAL AND THERAPEUTIC INTERPRETATION OF CANCER VARIANTS

In the third chapter of my thesis I present another web tool, named Cancer Genome Interpreter (CGI). CGI interprets the oncogenic relevance of tumor variants and identifies suitable therapies to target them according to several levels of evidence. On the one hand, it interprets the role in cancer of the input variants, mutations, copy number alterations and chromosomal rearrangements. On the other hand, it identifies and *in silico* prescribes the most suitable therapies according to the driver alterations present in the tumor. CGI uses include a broad range of applications that range from basic research to translational oncology. It has been implemented as a freely available online resource at http://cancergenomeinterpreter.org.

The output of the CGI is divided into two different analysis: the alteration analysis, that predicts the significance of the analyzed variants; and the prescription analysis, that identifies the best therapeutic options of the previously identified driver variants. My contribution to this project is limited to the prescription analysis. The prescription analysis is based on two steps: (i) building a comprehensive database of drugs including genomic biomarkers of response, resistance or toxicity and distinct levels of evidence and (ii) developing a method to prescribe the anti-cancer therapies in (i) to the identified driver alterations.

To generate a reliable database of anti-cancer therapies with drug biomarkers we took advantage of a pre-existing manually curated effort (Drug Knowledge Database) and extended it by including external cancer research curators. My work has consisted in integrating all data sources; generating a systematic nomenclature for the genomic biomarkers, the drugs and the cancer types; and keeping it up-to-date. Besides, I have generated another resource with ligands targeting altered driver genes according to different levels of potency of the interaction, which has been also integrated within CGI web tool. Additionally, I have conceived the algorithm for the prescription of driver alterations to drug biomarkers and ligands. The complex part of this algorithm is the proper handling of drugs with different genomic types of alterations (e.g. copy number alteration and mutation biomarkers); drugs with wild type genomic biomarkers; drugs with more than one genomic biomarker of response; and drugs with response and resistance biomarkers that can be simultaneously present in a tumor. The conception of the algorithm also includes the decision of which drug repurposing opportunities should be considered and consequently shown to the user.

As a snapshot in time, from 13th October 2016 until 15th of May 2017 CGI has been accessed 7200 times by 2600 users.

Tamborero D, **Rubio-Perez C,** Deu-Pons J, Schroeder MP, Vivancos A, Rovira A, Tusquets I, Albanell J, Rodon, J, Tabernero J, de Torres C, Dienstmann R, Gonzalez-Perez A, Lopez-Bigas N. (***Submitted***) .Cancer Genome Interpreter Annotates The Biological And Clinical Relevance Of Tumor Alterations

BioRxiv pre-print: https://doi.org/10.1101/140475

Tamborero D, Rubio-Perez C, Deu-Pons J, Schroeder MP, Vivancos A, Rovira A, et al. Cancer Genome Interpreter annotates the biological and clinical relevance of tumor alterations. Genome Med. 2018 Dec 28;10(1):25. DOI: 10.1186/s13073-018-0531-8

Chapter 4

TUMOR MOLECULAR MECHANISMS OF IMMUNE EVASION

In this last chapter, together with Tamborero D, we present a comprehensive identification of tumor molecular mechanisms which may allow evade the immune system.

The basement of this work is the identification of the infiltration patterns of sixteen immune populations across 28 solid tumors (9403 samples) through a sample-level enrichment method. Upon it, we first analyzed the infiltration patterns of the immune populations, revealing that the immune infiltration patterns did not correlate with the tissue of origin. This suggested that tumor intrinsic features may be responsible of the different infiltration patterns. To further explore this, we refined the immune infiltrates at cancer type level and grouped the tumors in immune-clusters, which represented the effectivity of the immune system attack. We observed that similar levels of cytotoxicity across immune-clusters showed different immune infiltrating patterns across cancer types. Next, we evinced if clinical features could explain the immune clusters and observed a tendency of low cytotoxicity in advanced stage tumors, suggesting this phenotype as a possible pre-requisite for tumors to progress. At last, we looked for pathways active in the tumor across the different immune-clusters, adjusting expression data for its immune component. High cytotoxic clusters were mostly enriched by pathways related with high immune infiltration and energy cell metabolism, intermediate cytotoxic clusters were enriched in angiogenesis and extracellular matrix pathways and low cytotoxic immune-clusters were mostly enriched by cell division pathways and others (such as TGFb) described to lead to low cytotoxicity. We finally integrated all the results into a reasoned biological model.

I explored different methodologies aimed to assess the infiltration of immune populations. I made a comparison with a deconvolution method (CIBERSORT) and a comparison between two sample-level enrichment methods (ssGSEA *vs* GSVA), to rationally decide which approach would meet our needs the better (everything described in the supplementary methods). After deciding for GSVA, I explored which gene sets to use for the identification of the immune populations. Next, I identified the immune infiltration pan-cancer and per-cancer type. Then, I explored the infiltration patterns across cancer types, comparing them with the infiltration of its normal tissues, and exploring them within the immune-clusters. I have carried out and explored the results of the pathway analysis and integrated them into a reasoned biological model. At last, I have written most of the sections in this chapter, including the supplementary, and done most of the figures and supplementary tables.

Here I present a first draft of a manuscript in preparation that will include all the results presented and additional ones where we are working on: analysis of the mutational load, copy number alterations and driver genomic correlates across immune-clusters.

Rubio-Perez C*, Tamborero D*, Muiños F, Lopez-Bigas N, Gonzalez-Perez A†. Identification of tumor immune avoidance processes across 28 solid tumors (*In preparation*)

* co-first authors † co-corresponding authors

Identification of tumor immune avoidance processes across 28 solid tumors

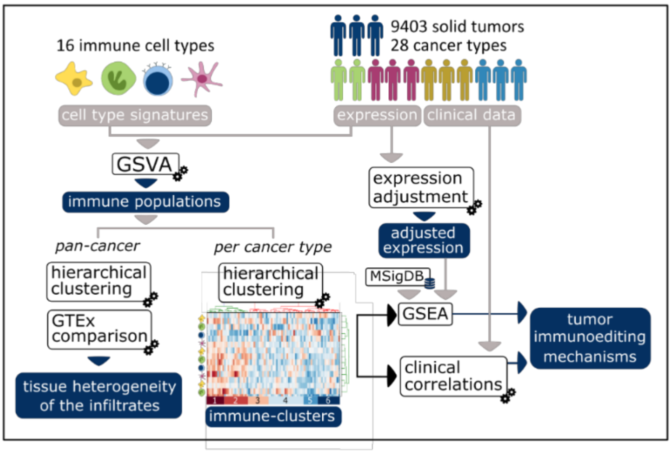
Carlota Rubio-Perez[1,2*], David Tamborero[1,2*], Ferran Muiños[1], Nuria Lopez-Bigas[1,3], Abel Gonzalez-Perez[1]

[1]Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, Baldiri Reixac, 10, 08028 Barcelona, Spain.
[2]Pompeu Fabra University, Barcelona, Spain.
[3]Institució Catalana de Recerca i Estudis Avançats (ICREA), Passeig Lluís Companys 23, 08010 Barcelona, Spain
[*]*co-first author*

ABSTRACT

There is a growing need of in depth understanding the tumor mechanisms that modulate the immune response. The availability of large tumor cohorts with transcriptome profiling, together with the development of methods which detect tumor immune infiltrating cell types from the transcriptome has opened the possibility of comprehensively studying tumor immune avoidance mechanisms. Here, we present the first comprehensive assessment of the tumor intrinsic pathways underlying the avoidance to the infiltration of sixteen different immune populations across 28 solid tumors. After discovering that immune infiltration profiles could not be explained by the tissue of origin, we hypothesized that tumor intrinsic mechanisms may be shaping the different immune profiles. Grouping cancer type specific profiles in three different scenarios of immune cytotoxicity revealed different tumor pathways active across the scenarios. The identified tumor specific pathways of immune avoidance attack may be good mechanisms for exploring combinatorial therapies with immunotherapies.

GRAPHICAL ABSTRACT



INTRODUCTION

The recent success of cancer immunotherapies and the observation that some patients do not respond to these treatments have increased the necessity of in depth understanding the relationship between tumor and immune cells. The mechanisms of interaction between tumor and immune cells are complex. They are usually referred to as a dynamic cycle in which immune system modulates tumor development and tumors can modulate the immune system response[1]. The idea that tumors can avoid the immune system response is not new, being the first hypotheses formulated by Sir Macfarlane Burnet in 1957[2]. However, this idea did not become accepted until the beginning of the 21st century[3], when data supporting it appeared. In 2011 tumor avoidance of immune destruction was formally considered as one of the cancer hallmarks[4].

Since then, several mechanisms of tumor avoidance of the immune system attack have been identified. An example is the presentation by tumor cells of immune checkpoint molecules in their cell surface (e.g. *PDL-1*), that prevent from T cell cytotoxicity when binded[5]. Indeed, inhibitory molecules have been designed for some of these checkpoints and have been shown to be successful in the clinical setting[6–10], emphasizing the impact that the identification of new immunomodulatory targets could have in patient care. Nevertheless, even if some successful mechanisms of immune avoidance have already been identified, cancer research community has not yet evinced a comprehensive view of the tumor mechanisms avoiding the immune system response.

During the last years, several large-scale studies aimed to comprehensively identify tumor immune infiltration profiles are emerging[11–16]. These studies have benefited from the systematic profiling of the transcriptome from tumor bulk samples, mostly from RNA-seq data. Tumor bulk samples contain not only tumor but also the cells of its microenvironment, which includes immune infiltrating cells. Several computational approaches, mainly sample-level enrichment[12,13,16] or deconvolution strategies[11,15], have been developed to reconstruct the immune infiltrate from the transcriptome profiling of a tumor bulk sample. Thus, the development of these methodologies together with the large collection of tumor RNA-seq data available, have opened the possibility of identifying the mechanisms of tumor immune evasion by analyzing tumor features across the different immune infiltrates.

Recent efforts, even if have produced interesting results, to our knowledge are not carried out in a comprehensive way. They are mainly focused: (i) on the study of the clinical impact of the immune infiltrates across different cell populations[17,18]; (ii) on the study of specific diseases (e.g. Senbabaoglu et al. 2017 focused their work in kidney clear cell carcinoma, Ali et al. 2016 in breast cancer and Angelova et al. 2015; in colorectal cancer)[16,19,20]; (iii) on the analysis of specific tumor alterations (e.g. Davoli et al. 2017; mainly focus on the study of tumor mutational burden and aneuploidy in relationship with immune evasion)[21] or (iv) both (e.g. Chaorentong et al. 2017; only give insights for two cancer types and their molecular subtypes)[13]. To our knowledge the most comprehensive assessment of the tumor mechanisms underlying immune response tumor evasion was the work done by Rooney et al (2015)[22]. However, they only considered genomic tumor alterations (not considering transcriptomic changes) and measured only the cytolytic activity, not considering the role of the other immune populations in relationship with the tumor.

Here, we present, to our knowledge, the first comprehensive assessment of tumor pathways modulating the immune system response. Briefly, using a sample-level enrichment analysis we reconstructed the infiltration profiles of sixteen immune populations across 9403 patients from 28 different solid tumors, revealing that the immune infiltration patterns could not be explained by the tissue of origin. Thus, we hypothesized that tumor intrinsic features could be shaping the immune profiles. With the aim of exploring this possibility, we refined the immune infiltration patterns at cancer type level and grouped the tumors in three immune cytotoxicity scenarios. Finally, after adjusting the expression by the immune infiltration we observed a heterogeneous activity of several tumor pathways across the three scenarios. Namely, we identified different pathways enriched in high, intermediate and low cytotoxic scenarios, and summarized all the knowledge generated into a plausible biological model of tumor immune evasion. We hypothesize that the set of pathways identified to be active in the tumors across different immune infiltration patterns are potentially good mechanisms for exploration in the context of new therapeutic interventions.

METHODS

Patient data collection

TCGA RNA-seq data for 28 solid tumors and 9403 patients (see Table S1 for a summary) was downloaded Firebrowse (20160128 version, rnaseqv2 RSEM genes normalized data) along with clinical data (20160128 version). In the clinical data we did a manual annotation of those samples with ambiguous stage. When there was more than one sample per patient, we kept only one, following the guidelines available in Firebrowse. TCGA patient virus infection data was obtained from Rooney et al 2015. To identify the infected tumors we followed the criteria stated in the same publication. Data on normal donors was retrieved from GTEx (v6). RPKM sample level matrix was downloaded from GTExportal (https://gtexportal.org). We additionally downloaded RNAseq data from melanoma patients treated with: anti-CTLA4[23] (provided by the authors, 42 patients) and anti-PD1[24] (retrieved from GEO: GSE78220, 28 patients).

Identification of immune populations

We have estimated the infiltration within tumors of 16 different immune populations. We have obtained the gene signatures of each cell type (Table S2A) from two different publications[12,13] (see Supplementary Methods). We have used the names of the cell types as used in the corresponding publications.

Following the rationale of[12,13,16,19] we have used a sample-level enrichment method, GSVA[25], to measure the infiltration of each immune population in tumor RNA-seq data for 28 solid tumors (9403 samples). We have used the GSVA implementation available in *R Bioconductor* package *gsva* (see Supplementary methods for details on the selection of the enrichment method and the comparison of this approach with a deconvolution method). Pan-cancer GSVA results, which implies a pan-cancer normalization, and per-cancer type level, which implies a different normalization in each cancer type, are available in Table S3.

Identification of immune-clusters

Hierarchical clustering was performed minimizing the squared Euclidean distance between the agglomerated samples by using the Ward method, through *hierarchy* module from *SciPy clustering python* library[26]. Samples were assigned to one of the "n" clusters according to the resulting linkage matrix: where "n" represents the total number of clusters of the partition. The number of clusters was selected observing the percentage of variance of the data explained as a function of a range of "n" values (see Supplementary Methods). The pan-cancer GSVA cluster analysis was unbiased, whereas the per-cancer type GSVA clustering was performed by setting a weight of 3 to the cytotoxic cell levels. Sample grouping in immune-clusters is available in Table S4, either pan-cancer (Table S4A) and per cancer type (Table S4B).

Statistical tests

Pearson's correlation was performed using *linregress* module from *python SciPy* library[26]. Multiple testing correction, when necessary, was applied using *multipletests* module from *python statsmodel* library. In all the cases where we corrected for multiple testing we used a Benjamini-Hochberg False Discovery Rate. Survival analysis was done building a Cox regression model, using *python lifelines* library, for each cancer type, adjusted by stage, age and gender when applicable (e.g. ovarian cancer was not adjusted by gender as all patients are females). Fisher exact test was performed using *fisher_exact* module from *python SciPy* library[26].

Expression adjustment

To adjust expression values of tumor bulk samples by its immune component, we followed the rationale described by Aran et al. (2016)[27] (Figure S1). Briefly, we adjusted the expression levels of each gene, per each sample, according to the contribution of *CD45* in the expression of each gene in normal tissues (see Supplementary Methods). Gene expression values which turned negative or zero after the adjustment process were set to -6 log2(RSEM) in downstream analysis and plots.

Pathway enrichment analysis

We performed a Gene Set Enrichment Analysis (GSEA)[28] to identify enriched pathways per immune-cluster, comparing against the other immune-clusters pooled together. We used GSEA R software available for download at http://software.broadinstitute.org/. We downloaded the pathway gene sets from MSigDB[28]. From the available gene set collections, we included all broad hallmark gene sets and specific pathways of interest from the curated gene sets, canonical pathways collection. Moreover, we also included some manually curated pathways, not found in the mentioned collections or from other publications. We minimized including pathways with high gene overlap (Figure S2, Table S2B). All GSEA results are available in Table S5.

RESULTS & DISCUSSION

Identification of sixteen tumor infiltrating immune populations

First, we sought to characterize the immune infiltrate of solid tumors by means of the prevalence of 16 cell populations: B cells, eosinophils, macrophages, mast cells, NK CD56bright cells (NKbright), NK CD56dim cells (NKdim), neutrophils, T helper cells, Tcm cells, Tem cells, Tfh cells, iDC, aDC, activated CD8 T cell (CD8), gamma delta T cell (Tgd) and regulatory T cell (Treg). We obtained the gene signatures from these gene sets from two different publications: Bindea et al. (2013) and Charoentong et al. (2016) (see Methods and Supplementary methods for details). Of note, the gene sets of the different immune populations showed no overlap among them or very few genes overlapping (Table S2A).

We employed a sample-level enrichment method, following the rationale of previous studies[12,13,16,19], to quantify the infiltration of the immune cell populations in each tumor in a cohort of 9403 solid tumors from 28 different cancer types. Nevertheless, we did a comparison with a deconvolution method which suggested that sample-level enrichment methods would be more suitable for the current work, as it is based on RNA-seq data analysis (see details in Supplementary Methods). Indeed, this statement has recently been reported by Senbabaoglu et al (2017) where they discuss the advantages of enrichment methods over deconvolution approaches for analyzing RNA-seq data.

Among the available sample-level enrichment methods, we used Gene Set Variation Analysis (GSVA)[25] (scored from -1 to +1). Comparing GSVA with single-sample Gene Set Enrichment Analysis (ssGSEA)[29], another widely used sample-level enrichment method, we yielded a high correlation (Pearson's correlation coefficient, 0.87). However, we decided to use GSVA over ssGSEA due to its intrinsic normalization step, which helps to reduce the noise.

Immune cells tend to co-infiltrate across tumors and are clinically relevant

We first investigated the different prevalence of cell populations of the immune infiltrate across cancer types in the cohort through a pan-cancer GSVA.

We first studied how correlated different cell populations are in the immune infiltrate. Most cell populations, independent of their specific immune functions were positively correlated (Pearson's correlations from 0.63 to 0.1, with few cases of weak negative correlations; Figure 1A). This suggests that an important force driving the enrichment is the overall magnitude of infiltration across tumors. Macrophages, iDCs and neutrophils exhibited the strongest correlations (0.63 Pearson's correlation), followed by NKdim and CD8s (0.59 Pearson's correlation). The other well-known effector immune population, Tgd, was highly correlated with the latter two (0.42 with NKdim and 0.52 with CD8s),

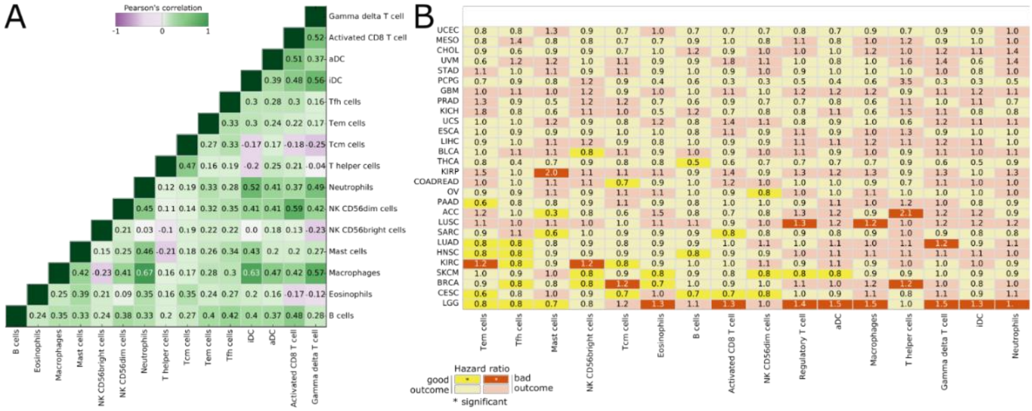


Figure 1 | Co-infiltration and clinical impact of sixteen immune populations. (A) Heatmap with the median Pearson's correlation values between the enrichment of the sixteen immune cell types. Negative correlations are colored in a violet color-scale and positive correlations in a green color-scale. (B) Heatmap with the Hazard ratio of the influence of each immune cell type on patient survival. It is sorted by the cancer types and cell types with more significant associations. Cell type infiltration associated to a bad outcome is colored orange and infiltration associated to good outcome is colored yellow, colored pale in both cases if not significant.

suggesting that effector immune populations (NKdim, NKbright, CD8 and Tgd) tend to co-infiltrate. The same pattern of co-infiltration with high correlations (0.57 median Pearson's correlation) was observed in suppressor immune populations (macrophages and Treg).

Additionally, we characterized the clinical significance of the immune population infiltration, as done in13,18. An adjusted Cox regression (see Methods) per cancer type showed that the enrichment of all immune populations had significant (Q-value < 0.1) influence in the clinical outcome in at least one cancer type (Figure 1B). Suggesting that the identified immune infiltration patterns are biologically meaningful. In half of the cell types we observed heterogeneity in the survival influence across cancer types (e.g. CD8s, Tem and NKbright). Cell types described as suppressors (Treg and Macrophages) tended to be associated to bad outcome and effector cell types to be associated to good outcome (NK cells and CD8s). Nevertheless, we observed a degree of heterogeneity across cancer types. Counterintuitively, gamma delta T cells were associated to bad outcome in two cancer types. This heterogeneity suggests that using a cytotoxicity measure may not be enough to capture the effect of the immune system infiltration, being more accurate using the infiltration patterns across immune populations.

Immune infiltration is not explained by the tissue of origin

We next carried out a more detailed analysis of the immune infiltration patterns across the different cancer types (Figure S3). We re-covered the previously observed pattern of co-infiltration across cancer types for some immune-populations (e.g. B cells, Tcm, Tem). But for most of them we observed heterogeneity of infiltration, alone or in combination with other immune populations, across cancer types. Consequently, we hypothesized whether the different patterns of immune infiltrate across cancer types were primarily driven by differences observed in the tissue of origin of the tumors. We carried out two different analyses to answer this question.

First, we compared the infiltration of tumors (TCGA data) with those from healthy donor samples (GTEx data) corresponding to their respective tissue of

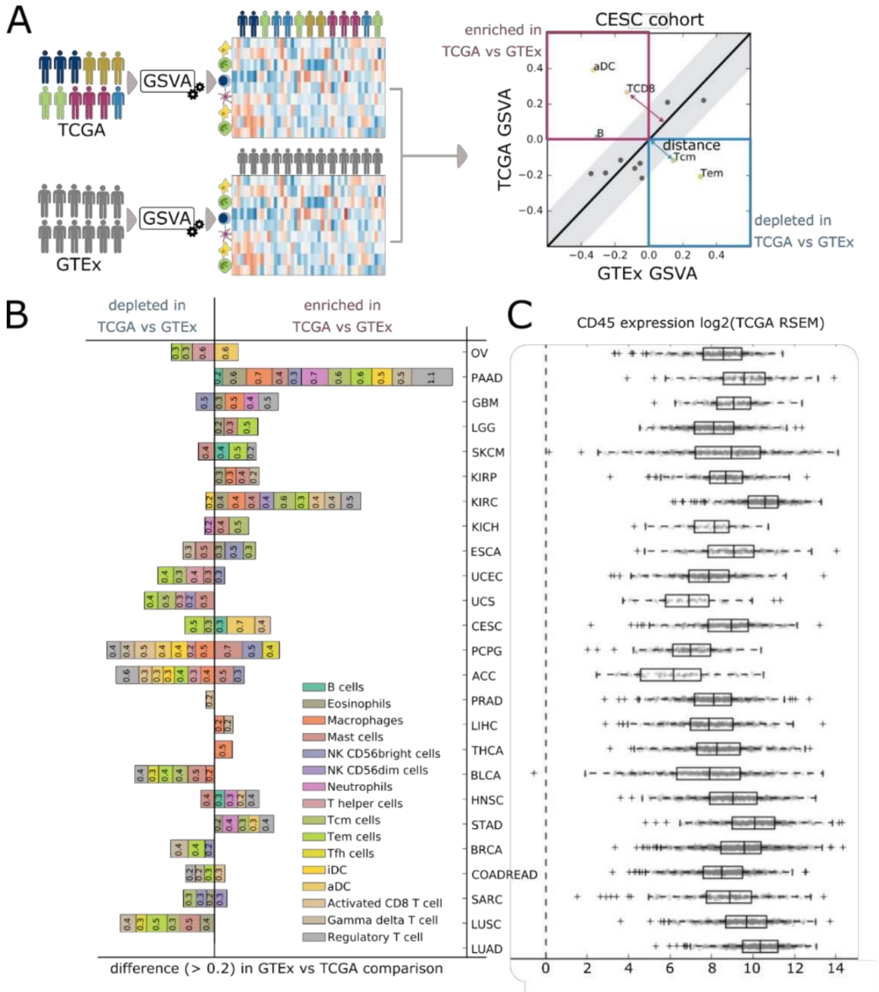


Figure 2 | Immune infiltration patterns in normal donors versus tumors. (A) Schema of the GTEx comparison analysis. We correlated GTEX and TCGA data GSVA scores (right) for each cancer type – tissue (Figure S4B), here only Cervix squamous carcinoma (CESC) correlation is shown. For each correlation, we considered that a cell type was enriched/depleted in TCGA vs GTEx if the difference of enrichment was > 0.2 GSVA score. (B) Bar plot summarizing the difference of enrichment for each immune cell types in GTEX vs TCGA GSVA correlations. Cancer types are ordered according its overall infiltration in normal tissues (Figure S4B), being lung the highest infiltrated tissue and ovary the lowest infiltrated one. (C) Boxplots representing the distribution of *CD45* expression across TCGA solid tumors log2 RSEM, each dot represents a sample.

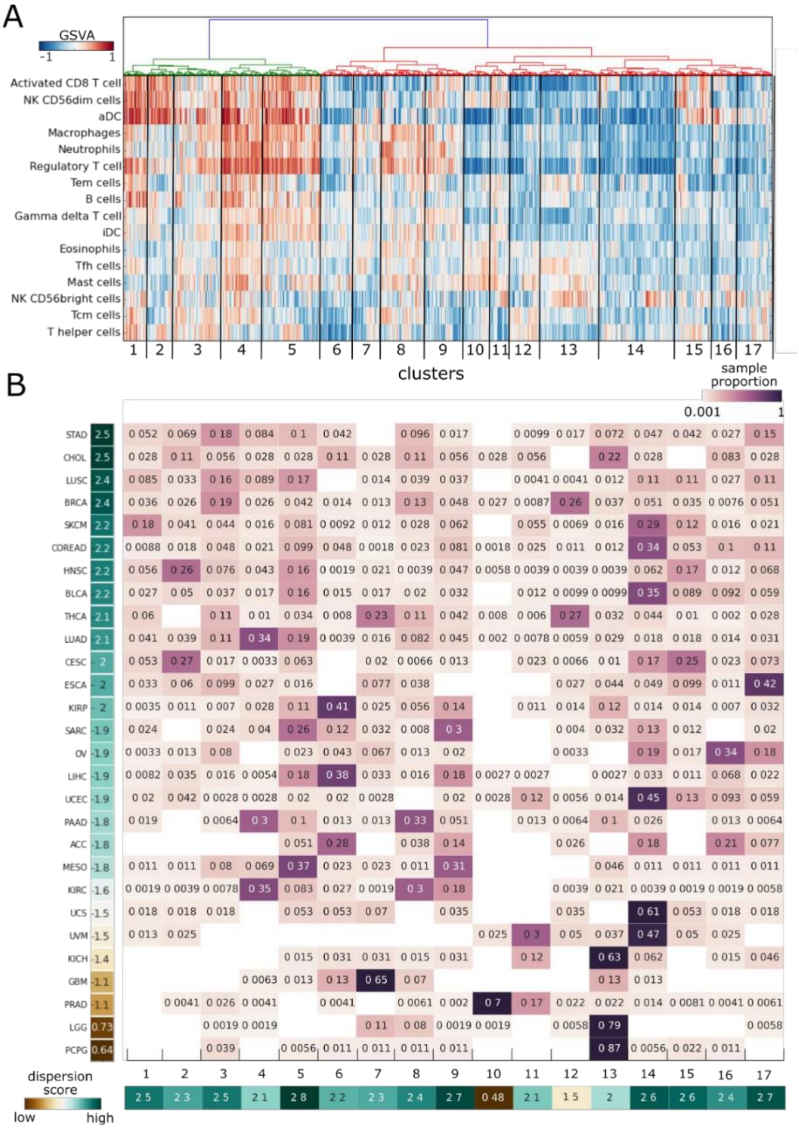


Figure 3 | Pan-cancer distribution of immune infiltration patterns. (A) Heatmap with the GSVA score (from -1 to 1 and colored from blue to red) per sample across the sixteen immune populations. GSVA values have been clustered in the x-axis, identifying 17 clusters. (B) Heatmap representing the proportion of samples from each cancer type that are found in each cluster (from 0 to 1, from pale pink to dark purple). On the left, there is a representation of the entropy score of each cancer type across the clusters, where brown corresponds to low entropy (most of the cancer type samples are in a single or few clusters) and dark green corresponds to high entropy (more heterogeneous distribution of cancer type samples across clusters). Bottom, same dispersion score is show, at cluster level.

origin (see Supplementary Methods). We carried out the comparison both at the level of overall immune infiltrate (measured as the level of *CD45*; Figure 2A and B and Figure S4B), and for the immune infiltration pattern provided by the 16 populations. At the overall level, we observed few cancer types (e.g lung carcionomas) infiltrated at levels comparable to their tissue of origin. In most cases, we found little coherence. For example, pancreas adenocarcinoma (PAAD) and kidney clear cell carcinoma (KIRC) were among the highest infiltrated tumors but their normal tissues (pancreas and kidney, respectively) were among the lower infiltrated ones. Similar low coherence was observed when we compared the pattern of immune infiltration of tumors and their tissue of origin (Figure 2A and B and S4A). For example, adrenal cortical carcinoma (ACC) and pheocromocytoma and paraganglioma (PCPGC) showed high enrichment for seven different immune cell types in normal tissues which were not enriched in their tumor tissues, which were enriching other cell types.

Second, we clustered the immune infiltrate of all tumors in the pan-cancer cohort (Figure 3A, Table S4A). We obtained 17 separate groups reflecting distinct immune infiltrate profiles in cancer. The clusters captured patterns of immune infiltration raging from overall low infiltration (e.g. cluster 12) to others with a high infiltration (e.g. 1). In the middle, there was a range of clusters with mixed infiltration. Most of the cancer types showed a heterogeneous immune infiltrate, being the samples of these cancers grouped in different clusters (Figure 3B).

Stomach adenocarcinoma (STAD), cholangiocarcinoma (CHOL), lung squamous carcinoma (LUSC), breast cancer (BRCA), cutaneous melanoma (SKCM) and colorectal adenocarcinoma (COADREAD) samples were among those more distributed across distinct immune infiltrates. As opposite, prostate adenocarcinoma (PRAD), PCPG, lower grade glioma (LGG) and glioblastoma (GBM) tumors exhibited a more homogeneous infiltrate and most (>65%) of the samples of these cancers appeared grouped in a single immune cluster (Figure 3B). Besides, large subsets of certain cancer types shared the same immune profiles, converging in the same immune-clusters. For example, most (65%) of the GBM tumors were in cluster 7, together with a significant percentage (23%) of thyroid carcinoma (THCA) tumors, sharing an immune infiltrate that exhibited low levels of Nkdim and CD8 and higher abundance of Tregs and macrophages.

Taken together, these results suggest that differences observed in the immune infiltrate across cancer types are not driven by their tissue of origin, but rather by the molecular features of the tumors. Moreover, they provide a general overview of the specific immune profiles of specific cancer types, and therefore provide a description of that infiltrate which may be used to study the interaction of the molecular characteristics of the tumors and the immune system.

Three different scenarios of immune infiltration are found across solid tumors

Given our hypothesis that tumor intrinsic molecular characteristics are shaping the immune infiltration profile, we next fine-tuned the immune infiltration profiles by identifying them at cancer type level (see Methods). Then, we grouped the tumors in each cancer type according to their pattern of immune infiltrate, with an overweight given to the population of cytotoxic cells (see Methods). The rationale behind this weighted clustering aims to separate tumors by their pattern of effective immune infiltration. We hypothesize that different patterns of immune infiltration effectivity may correlate with different mechanisms of immune evasion.

The clustering approach identified six groups of tumors with distinct pattern of immune infiltration in each cancer type (Table S4B). These six groups, as expected, reflected three scenarios of immune infiltration across solid tumors: low cytotoxic infiltrate (groups 1/2), mid cytotoxic infiltrate (groups 3/4) and high cytotoxic infiltrate (groups 5/6). Figure 4 shows an example of two of the 28 per-cancer type clustering (Figure S5 shows the 28) for two cancer types at the extremes of overall infiltration range, by means of *CD45*: KIRC and uveal melanoma (UVM).

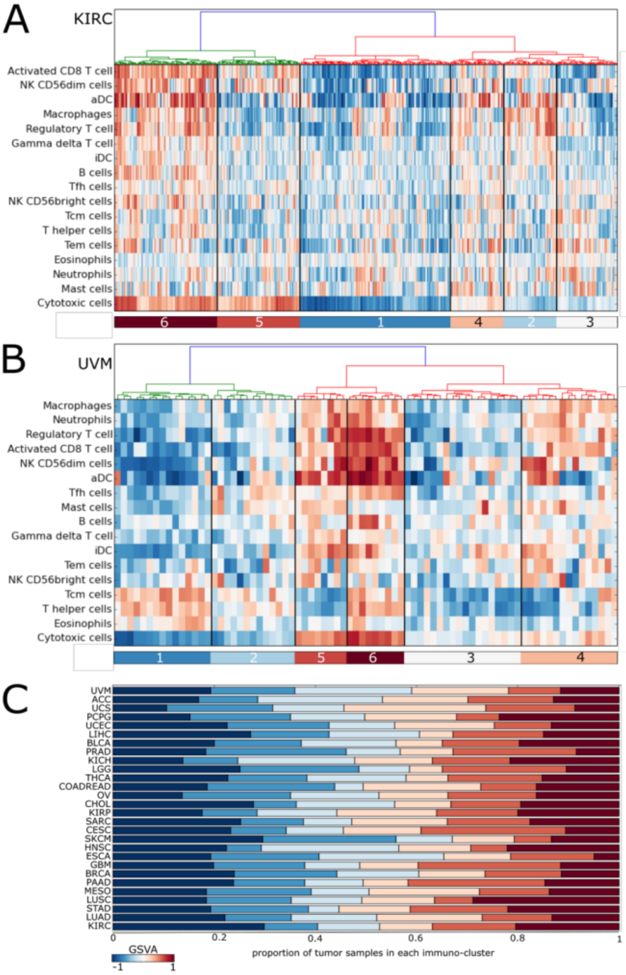


Figure 4 | Cancer type immune-cluster distribution. (A) and (B) show heatmaps with the distribution of GSVA values across the 16 immune populations and cytotoxic cells, giving an overweight to cytotoxic cells to perform the hierarchical clustering. Immune-clusters are numbered according the median enrichment of the latter across samples, from 1 to 6. (A) corresponds to GSVA in KIRC and (B) in UVM. (C) Bar plot with the proportion of samples in each immune-cluster across cancer types. Colors correspond to the 6 immune-clusters from the lowest cytotoxic (immune-cluster 1, dark blue) to the highest cytotoxic (immune-cluster 6, dark red).

Note that, despite the differences of overall infiltration (Figure 2C), in both cancer types the three scenarios of immune infiltration are apparent. These three scenarios are detected across all cancer types, although their relative prevalence varies in each of them (Figure 4C), likely due to the heterogeneity of immune infiltration patterns across cancer types.

Tumors in immune-cluster 1 showed a depletion of most immune cell types across cancer types, which resemble an immune desert phenotype (as described by[30,31]), except for an enrichment of aDC in THCA (Figure S6). Tumors in immune-cluster 2, as the previous ones, showed low infiltration of most immune cell types. Interestingly, 7/28 cancer types showed a high infiltration (GSVA > 0.2) of Treg and/or macrophages -both suppressor immune populations-, while no cancer type showed high infiltrations of effector immune cells. Thus, collectively, tumors in immune clusters 1 and 2 possess either a very low infiltrate or mostly suppressive infiltrate, respectively.

Immune-clusters 3 and 4 showed varying degrees of enrichment and depletion of different cell types across cancer types, producing heterogeneous infiltrates. Of note, while tumors of immune-cluster 3 exhibited depletion of cell populations across the board, those of immune-cluster 4 showed a median higher infiltration of suppressor *vs* effector populations (median GSVA across cancer types

Tumors of immune-cluster 5 showed enrichment of effector populations and depletion of suppressor ones in seven cancer types (suppressor populations showing a median GSVA < 0 and at least 3 out of the 4 effector populations with GSVA > 0). Moreover, the median GSVA of effector populations was higher than that of suppressor population across all cancer types (0.16 and 0.06, respectively). Immune-cluster 6 showed a high infiltration of most immune cell populations, coherent with an inflamed phenotype (as described by[30,31]). Hence, clusters 6 and 5 collectively grouped tumors with an infiltrate of high immune infiltration or majority effector infiltrate, respectively.

Stage and viral infection as clinical correlates across immune-clusters

Our next objective was to identify the characteristics of the tumors in each cancer type immune-cluster that could explain the features of their immune infiltrates.

First, we hypothesized that the presence of viral infections in patients could yield more highly immunogenic tumors due to the expression of viral antigens32. To evaluate this question, we collected the tumor expression of four known oncogenic viruses (see Methods): Epstein-Barr virus (EBV), human papillomavirus (HPV), hepatitis B and C (HBV and HCV) and observed the distribution of infected tumors across immune-clusters (Figure 5A).

We observed that EBV-infected STAD and HPV-infected head and neck squamous carcinoma (HNSC) tumors were significantly (P-value < 0.05) enriched (according to Fisher exact test) in high cytotoxic immune-clusters, whereas HPV-infected cervical squamous carcinoma (CESC) tumors did not exhibited any relationship. HCV-infected hepatocellular carcinoma (LIHC) was more frequent across tumors in high cytotoxic immune clusters, although it did not reach statistical significance, but strikingly, HBV infection appeared significantly enriched across LIHC tumors with a lower immune infiltrate. These results show a heterogeneous landscape of immunogenicity of viral-infected tumor samples, influenced by both the tumor and virus type.

Second, we hypothesized that the clinical stage of the tumor could also be a major determinant of its immune infiltrate. As a general trend, we observed that tumors of advanced stages (III/IV) were depleted for tumors in highly cytotoxic clusters (Figure 5B). This, suggests that tumor progression correlates with the immune infiltrate, either by directly influencing the status of the cell populations present, or because a less cytotoxic infiltrate is needed for tumor development. We observed a significant (P-value < 0.05) depletion of high cytotoxic immune-clusters (5-6) in early (stages I/II) versus advanced stages in seven cancer types, according to Fisher's exact test. The opposite relationship was observed in KIRC, where tumors with the highest cytotoxic infiltrate significantly enriched for tumors of stage III/IV, which may explain why the presence of immune infiltrate has been found to be a factor of bad prognosis in this cancer type when not adjusted by other clinical variables[19].

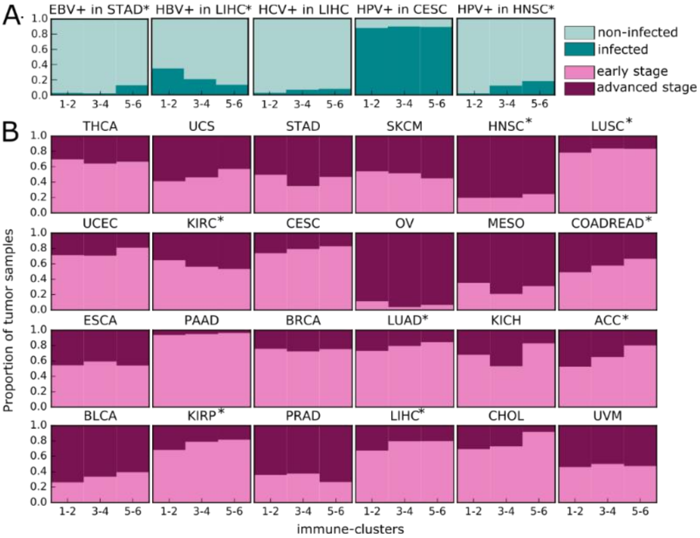


Figure 5 | Effect of clinical variables in the immune infiltrates. (A) Bar plot showing the proportion of infected patients (dark turquoise) versus non-infected (turquoise) with four different viruses in five different cancer types across immune-clusters (grouped by its cytotoxic phenotype in 1-2, 3-4 and 5-6). (B) Proportion of samples in early (stages I or II) and advanced clinical stages (stages III or IV) across immune-clusters (grouped as in A). Significant associations (Fisher exact test P-value < 0.05) in both (A) and (B) are marked with an asterisk.

Transcriptomic tumor programs modulate their immune infiltrate.

Next, we wanted to explore which transcriptomic programs of the tumors are active during immune evasion. To use tumor bulk expression to delineate these transcriptomic programs across tumors, we first needed to adjust the expression measured for each gene in each tumor sample to eliminate the contribution of the immune content (see Methods, Figure S1A). We then used the adjusted tumor expression of all genes to perform a Gene Set Enrichment Analysis (GSEA) of 52 selected pathways (Figure S2, Table S2B) to dissect the molecular mechanisms of tumors that may influence their immune infiltrate.

First, we checked that overall gene expression was homogeneous across immune-clusters, so no cluster biases were present (Figure S7). We next asked how much the adjustment of tumors expression changed the enrichment of each pathway across tumors (Figure 6A, detailed examples in Figure S1C). We observed that immune related pathways suffered the most with the adjustment process, e.g. more than 70% of the genes in the chemokines reactome pathway had lower expression levels after the adjustment process (n=57 genes). On the other hand, pathways related to oncogenic processes remained mostly unchanged after the adjustment; e.g. the expression of any gene in G2M checkpoint hallmark (n=200 genes) was lowered or increased after expression adjustment. Interestingly, the expression of some pathways such as cancer germline antigens would be masked if the expression was not adjusted, as the expression levels of the genes in this gene set were increased after the adjustment.

After adjusting the expression data, we ran the GSEA[28]. Thus, we only performed GSEA analysis on cancer types whose expression we could correct: i.e., all but UVM, mesothelioma (MESO) and cholangiocarcinoma (CHOL) (see Supplementary Methods). The GSEA computed the enrichment of up-regulated genes in tumors of each cluster in each cancer type with respect to the other clusters in the same cancer type (Figure 6B and C, Table S4). The heterogeneity of tumor pathways enriched for up-regulated genes in each cluster suggested that

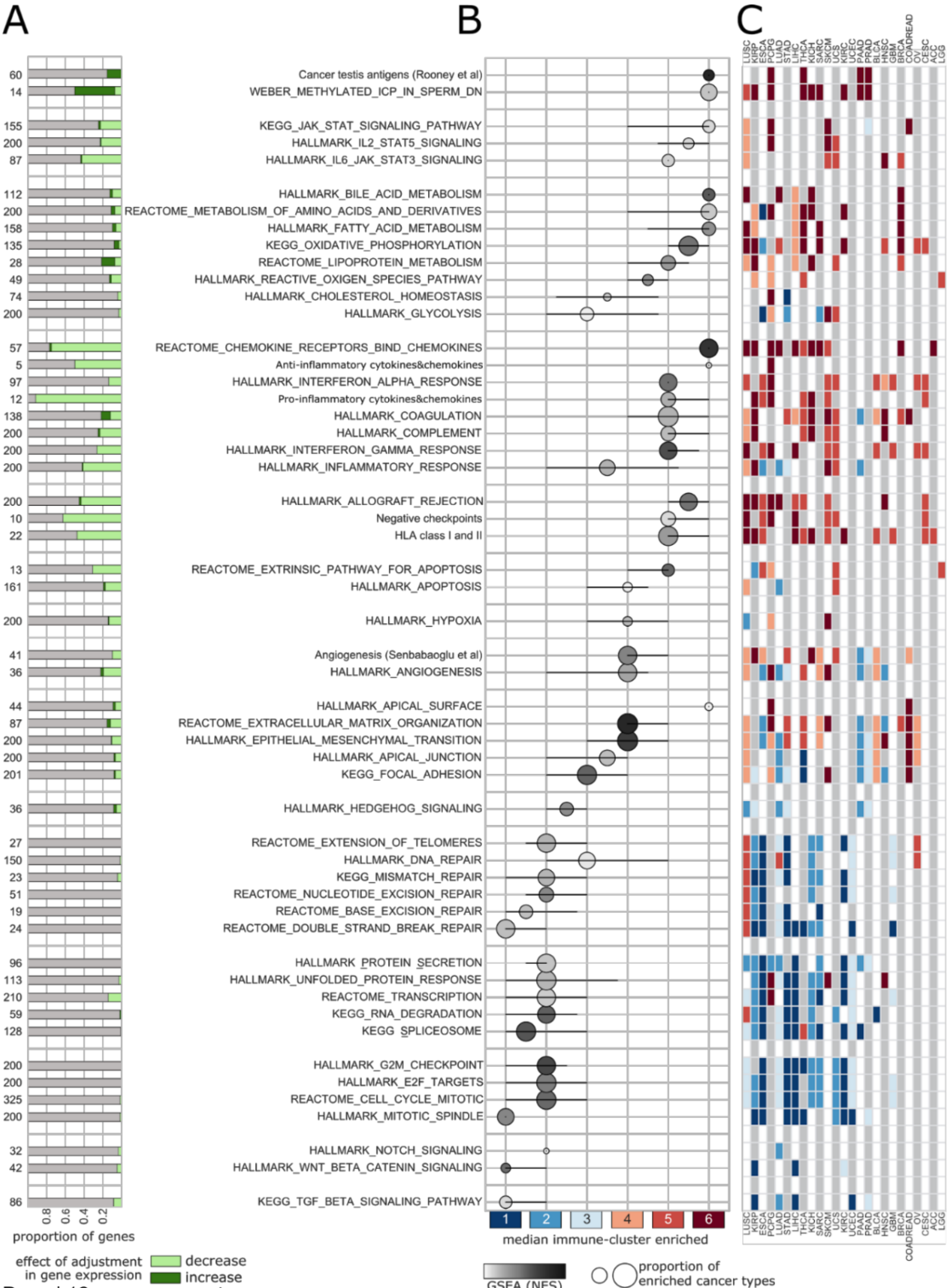

A
B
C
60
14
155
200
87
112
200
158
135
28
49
74
200
57
5
97
12
138
200
200
200
200
10
22
13
161
200
41
36
44
87
200
200
201
36
27
150
23
51
19
24
96
113
210
59
128
200
200
325
200
32
42
86
0.8
0.6
0.4
0.2
proportion of genes
effect of adjustment in gene expression
decrease
increase
equal
Cancer testis antigens (Rooney et al)
WEBER_METHYLATED_ICP_IN_SPERM_DN
KEGG_JAK_STAT_SIGNALING_PATHWAY
HALLMARK_IL2_STAT5_SIGNALING
HALLMARK_IL6_JAK_STAT3_SIGNALING
HALLMARK_BILE_ACID_METABOLISM
REACTOME_METABOLISM_OF_AMINO_ACIDS_AND_DERIVATIVES
HALLMARK_FATTY_ACID_METABOLISM
KEGG_OXIDATIVE_PHOSPHORYLATION
REACTOME_LIPOPROTEIN_METABOLISM
HALLMARK_REACTIVE_OXIGEN_SPECIES_PATHWAY
HALLMARK_CHOLESTEROL_HOMEOSTASIS
HALLMARK_GLYCOLYSIS
REACTOME_CHEMOKINE_RECEPTORS_BIND_CHEMOKINES
Anti-inflammatory cytokines&chemokines
HALLMARK_INTERFERON_ALPHA_RESPONSE
Pro-inflammatory cytokines&chemokines
HALLMARK_COAGULATION
HALLMARK_COMPLEMENT
HALLMARK_INTERFERON_GAMMA_RESPONSE
HALLMARK_INFLAMMATORY_RESPONSE
HALLMARK_ALLOGRAFT_REJECTION
Negative checkpoints
HLA class I and II
REACTOME_EXTRINSIC_PATHWAY_FOR_APOPTOSIS
HALLMARK_APOPTOSIS
HALLMARK_HYPOXIA
Angiogenesis (Senbabaoglu et al)
HALLMARK_ANGIOGENESIS
HALLMARK_APICAL_SURFACE
REACTOME_EXTRACELLULAR_MATRIX_ORGANIZATION
HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION
HALLMARK_APICAL_JUNCTION
KEGG_FOCAL_ADHESION
HALLMARK_HEDGEHOG_SIGNALING
REACTOME_EXTENSION_OF_TELOMERES
HALLMARK_DNA_REPAIR
KEGG_MISMATCH_REPAIR
REACTOME_NUCLEOTIDE_EXCISION_REPAIR
REACTOME_BASE_EXCISION_REPAIR
REACTOME_DOUBLE_STRAND_BREAK_REPAIR
HALLMARK_PROTEIN_SECRETION
HALLMARK_UNFOLDED_PROTEIN_RESPONSE
REACTOME_TRANSCRIPTION
KEGG_RNA_DEGRADATION
KEGG_SPLICEOSOME
HALLMARK_G2M_CHECKPOINT
HALLMARK_E2F_TARGETS
REACTOME_CELL_CYCLE_MITOTIC
HALLMARK_MITOTIC_SPINDLE
HALLMARK_NOTCH_SIGNALING
HALLMARK_WNT_BETA_CATENIN_SIGNALING
KEGG_TGF_BETA_SIGNALING_PATHWAY
1
2
3
4
5
6
median immune-cluster enriched
GSEA (NES)
proportion of enriched cancer types
LUSC
KIRP
ESCA
PCPG
LUAD
STAD
LIHC
THCA
KICH
SARC
SKCM
UCS
KIRC
UCEC
PAAD
PRAD
BLCA
HNSC
GBM
BRCA
COADREAD
OV
CESC
ACC
LGG

Figure 6 | Tumor pathway enrichment across immune-clusters. (A) Bar plot representing the proportion of genes in each pathway gene set that after expression adjustment its expression has been lowered (pale green), increased (dark green) or remains stable (grey). (B) Dot plot with the distribution of enriched pathways in adjusted expression data (Q-value < 0.25) across the immune-clusters. Each dot size represents the proportion of cancer types enriching the pathway and the x-axis the median immune-cluster enriched across cancer types. Dots also show the dispersion of enriched immune-clusters as a black line and they are colored according to the median NES across cancer types. Bottom bar plot represents the total of pathways enriched in each cluster. (C) Heatmap representing the cancer types enriching each pathway colored by the cluster enriched in each cancer type (see in B bottom bar plot the color legend).

different tumors may influence their immune infiltrate through the activation of a panoply of molecular mechanisms.

Pathways with a median enrichment in high cytotoxic tumors were consistently enriched across cancer types in immune-clusters 5 or 6 (80.3% of the significant enrichments). Six diverse types of tumor pathways were enriched in a high cytotoxic immune context. In turn, these pathways may be classified in pro-immunogenic and immune-resistant pathways.

Pro-immunogenic pathways include: cancer germline antigen (CGA) pathways, which have been described to lead to an immunogenic phenotype and postulated as candidate targets for immunotherapies because of that[33,34]; some cytokines and chemokines pathways (e.g. Reactome chemokine receptors bind chemokines or Pro-inflammatory cytokines & chemokines), HLA class I and II and apoptosis pathways (e.g. hallmark apoptosis and Reactome extrinsic apoptosis pathway). On the other hand, likely immune-resistant pathways involve: *STAT* signaling (e.g. hallmark *IL6 JAK/STAT3* signaling), associated to an inhibition in the production of pro-inflammatory cytokines and chemokines[35,36]; energy metabolism (e.g. hallmark fatty acid metabolism and KEGG oxidative phosphorylation among others), which may deplete the stroma of nutrients, and consequently prevent from effector immune population differentiation[37,38]; some inhibitory cytokines and chemokines and negative checkpoints. These findings suggest that tumors highly infiltrated survive in a dynamic equilibrium between pro-immunogenic and anti-immunogenic signals.

Pathways with a median enrichment in intermediate cytotoxic clusters (hypoxia, angiogenesis and extracellular matrix pathways) were lowly enriched in these immune-clusters (3 or 4) across cancer types (46% of all significant enrichment were in cluster 3 or 4). We hypothesized that cancer types enriched in the same pathway, even if in different immune-clusters, showed a similar immune profile, but at different cytotoxicity levels across cancer types.

Hypoxia hallmark was significantly enriched in LUSC, SKCM and PCPG for immune clusters with different cytotoxicities (2, 6 and 4 respectively). When observing in detail the immune populations mostly enriched in these immune-clusters (Figure S6), we observed that all of them showed high infiltration of Treg and neutrophils, across different cytotoxic scenarios by the tumors. This observation was coherent with the literature, as hypoxia can induce the recruitment of Tregs, via the expression of specific chemokines[39,40]. Thus, hypoxic conditions of the microenvironment, generating an up-regulation of the hypoxia response pathway in the tumor, may be associated to an increased recruitment of Tregs across different cytotoxic scenarios by the tumors. Angiogenesis pathways, according to literature are intimately linked to hypoxia[41–43], they were enriched in seven cancer types in immune-clusters 3 and 4, all of them exhibiting high infiltration by macrophages (median GSVA of 0.297), which are known to be recruited during angiogenic processes[44–46]. Besides, the cancer types enriched in other immune-clusters showed also infiltration by macrophages (median GSVA of 0.3). Extracellular matrix (ECM) pathways were enriched in 15/25 cancer types. Changes in ECM proteins have been shown to have an influence in the leukocyte trafficking into the tumor (either acting as a biophysiological barrier, promoting certain cell types to migrate or even modulating the migration mechanisms of the immune cells) and in the leukocyte polarization. These ECM changes have been associated to both pro-inflammatory and anti-inflammatory phenotypes, having a dual role[47–49].

Low cytotoxic clusters were consistently enriched across cancer types (68.5% of all enrichments were in immune clusters 1 or 2). There were six pathways enriched in low-cytotoxic clusters: *TGFβ* signaling, *WNT-βCatenin* pathway, cell cycle pathways, DNA damage repair and telomerase pathways, protein synthesis pathways and Hedgehog (SHH) signaling. *TGFβ* signaling has been described to prevent immune infiltration via chemokine and cytokine suppression[50], or stromal proliferation[50,51], and also triggering immune regulatory response by favoring the development of Treg according to literature[51]. LUAD and PRAD showed Treg infiltration (median GSVA 0.23) in clusters enriched for *TGFβ* pathway (2 and 3, respectively), suggesting that the activation of *TGFβ* pathway in both cancer types may promote Treg infiltration, while other *TGFβ* enriched cancer types, with depletion of Tregs, (KIRP, LIHC and UCEC) may activate *TGFβ* pathway to avoid immune infiltration. *WNT-βCatenin* pathway, enriched only in four cancer types, has been suggested to lead to a T cell immune-excluded phenotype[52,53]. Concordantly we found median depletion (GSVA = -0.05) of all T cells (comprising 7 cell types) across the cancer immune-clusters enriched in this pathway. Cell cycle pathways (e.g. E2F targets or G2M checkpoints) are enriched across 14 cancer types, high cellular proliferation is associated to a reduction of immunogenicity (and so cytotoxicity) due to the generation of many new non-recognized tumor antigens by the immune system[54]. DNA damage repair and telomerase pathways (e.g. DNA repair or mismatch repair), were likely enriched due to the identified increased rate of cellular proliferation as they are tightly bound to DNA replication, 11/15 cancer types with enrichment of immune-clusters for DNA damage pathways showed consistent enrichment for cell cycle. Protein synthesis pathways, similar as DNA damage could be linked to an increased rate of cellular proliferation, as a cell division implies a duplication of the protein content; 12 out of the 18 cancer types with enrichment of immune-clusters for these pathways showed also enrichment for the same immune clusters of at least one cell cycle pathway. Hedgehog (SHH) signaling was enriched in 5 cancer types. An activation of this pathway has been associated to a decreased immunogenicity by regulation of *HLA* molecules (downregulation, with a consequent decrease in CD8 and CD4 T cell infiltration) and *STAT* proteins[55–58]. Concordantly, enriched cancer types showed a depletion in CD8s.

Finally, we evaluated whether the analysis of the 52 selected pathways may inform of tumors response to immune-checkpoint blockade. We performed a GSEA of the 52 pathways comparing the up-regulation of genes between responders and non-responder tumors of the two available melanoma cohorts with transcriptome profiling and data of response to immune checkpoint blockade therapy response[23,24]. We found three significant enrichments considering a P-value < 0.05 in the anti-*PD1* treated cohort. We observed that up-regulated genes in non-responders were enriched in angiogenesis pathway from Senbabaoglu et al. (2017) (P-value = 0.02), extracellular matrix organization (Reactome) (P-value = 0.04) and *Wnt-βCatenin* hallmark (MSigDB hallmark) (P-value = 0.02). The enrichment of the first two pathways was already reported in the original publication of the cohort, as part of the IPRES signature[24], but the enrichment for *Wnt-βCatenin* had not been previously reported by performing this type of analyses, although it has been reported in the literature[52,53]. Even if limited, the pathway enrichment analysis of patients responding and not to immunotherapies sheds light into the underlying mechanisms of therapy no response. As more datasets with RNAseq data and the response to immune-checkpoint blockers appear, a more comprehensive analysis could be done than the snapshot provided here.

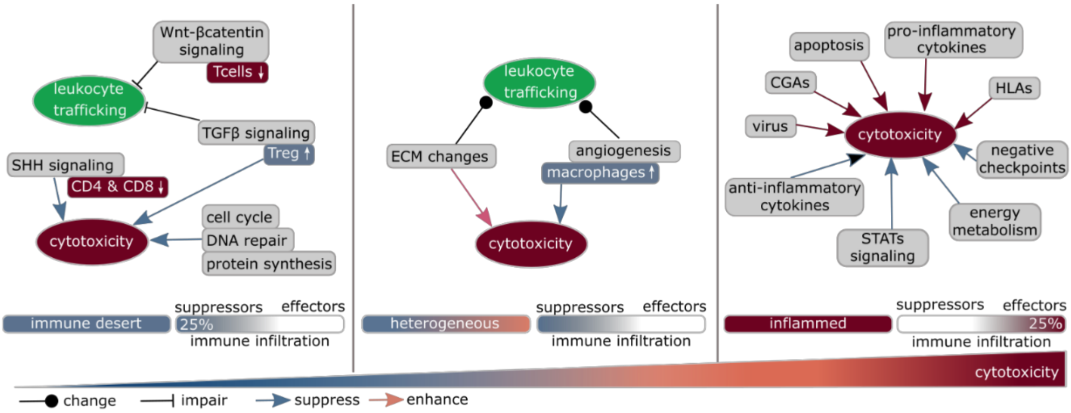


Figure 7 | Representation of the mechanisms underlying the tumor cytotoxic scenarios. A detailed explanation is found in the results section. SHH: sonic hedgehog; CGA: cancer germline antigens; Treg: T regulatory cell; CD8: Active CD8 T cells; ECM: extracellular matrix.

Integrative biological model of tumor intrinsic immune evasion mechanisms

Finally, we propose some potential mechanistic links between tumor molecular characteristics and their immune infiltrate pattern based on the results of our analyses. We use the three cytotoxicity scenarios already described to illustrate it: low cytotoxic infiltrate (clusters 1 and 2), intermediate cytotoxic infiltrate (3 and 4) and high cytotoxic infiltrate (5 and 6) (Figure 7).

The low cytotoxicity infiltrate scenario in 25% of the tumors in immune-cluster 2 showed an increased infiltration of suppressor populations in comparison to effector ones. While, tumors in immune-cluster 1 presented a phenotype of immune desert[31] (i.e .low cytotoxic). Tumors in these clusters tend to be of advanced clinical stages, suggesting that this immune phenotype could be a pre-requisite for tumor progression. The high proliferative state identified in 14 cancer types (marked by the enrichment of cell cycle, DNA damage and protein synthesis pathways) could be also responsible of lowering the immunogenicity by generating new non-recognized antigens because of the massive cellular proliferation[54]. Besides, the phenotype of immune exclusion could be explained in some tumors by the activation of pathways which impair leukocyte recruitment (*SHH, TGFβ* and *Wnt-βCatenin* signaling).

The intermediate cytotoxic scenario showed an heterogenous profile of immune infiltrates, with predominance of suppressor populations in 25% of the cancer types in immune-cluster 4. These tumors showed consistently enrichment for only two pathways: angiogenesis, which may impair leukocyte trafficking and contribute to a decrease in the cytotoxicity via recruitment of macrophages[41–43] and ECM changes. In turn, ECM changes could either promote or suppress cytotoxicity and leukocyte recruitment through different mechanisms[47–49].

The high cytotoxic infiltrate scenario could be explained in immune-cluster 5 by a predominance of effector populations over suppressor ones in 25% of the cancer types, and in immune-cluster 6 due to their inflamed phenotype (i.e. high cytotoxic)[31]. In LIHC and HNSC high cytotoxicity could be caused by viral infection too. In tumor from these clusters we found activation of processes that promoted the cytotoxic infiltrate (e.g. viral processes, high expression of HLA molecules and CGAs) and processes that presumably allow tumor cells to survive in it (e.g. negative checkpoints, anti-inflammatory cytokines). Additionally, we identified an enrichment of up-regulated energy metabolism pathways in tumors with high cytotoxic infiltrate that may establish a competition for nutrients with immune effector cells that would impair the differentiation of the latter[37,38].

Herein we have identified many biological processes whose activation we suggest may influence the pattern of their immune infiltrate. We thus contribute to shed light onto the mechanisms of immune evasion by tumors. Besides, the identified pathways active in the tumor evasion of the immune system could be good candidates for further investigation as potential targets for a combinatorial immunotherapy, as has already been suggested for *Wnt-βCatenin* pathway[52].

A clear limitation of this study is causality of the tumor pathways and its effect in the immune infiltration could not be inferred. Besides, we have not considered genetic and epigenetic tumor alterations. These may be positively selected due to the environmental pressure of the immunologic microenvironment, providing the tumor capabilities of immune resistance by activation of the identified pathways or other mechanisms, such as the described mutation of *HLA* and *B2M* molecules, preventing from the recognition of T cells or the mutations in *CASP8* which avoid apoptosis induced by immune cells[22] . Besides, we are aware that other genomic alterations, tumor mutational load and aneuploidy, have been described as markers of increased cytotoxicity[21,22], likely represented in the tumors of our high cytotoxic scenario.

CONCLUSIONS

In this study we explored the pan-cancer immune infiltration profiles of sixteen different immune populations across 9403 samples and 28 cancer types. The identified immune infiltration profiles were of clinical relevance and could not be explained by the tissue of origin. Given that, we explored tumor intrinsic features that may be shaping the immune infiltration of the different cancer types. To do so, we refined the immune infiltration profiles at cancer type level and described three different scenarios of cytotoxicity, divided in six immune-clusters. We showed that different profiles of infiltration may lead to the same levels of cytotoxicity across cancer types, to our knowledge not previously comprehensively described.

For the first time, we did a comprehensive analysis on the tumor mechanisms and clinical features underlying immune system evasion by tumors across 28 solid tumors. By adjusting the expression levels, we prevented highly infiltrated tumors from appearing only enriched in immune system related pathways and we identify other interesting ongoing processes, such as cellular metabolism, that otherwise would be masked. Albeit incomplete here we have presented a biological model which explains three different cytotoxic scenarios (high, intermediate and low) by the activation of tumor signaling pathways. These pathways, may allow the tumor resist and module a high cytotoxicity microenvironment, such as enhanced energy metabolism pathways in the tumor, or may lead to low cytotoxic environment, like a high cellular proliferative rate. The identified tumor mechanisms of immune avoidance are potential candidates to be explored as potential targets for combinatorial therapy with immunotherapies.

ACKNOWLEDGEMENTS

The results of tumor samples published here are in whole based upon data generated by the TCGA Research Network: http://cancergenome.nih.gov/. Normal donor samples were obtained from the Genotype-Tissue Expression (GTEx) Project, supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS.

IRB Barcelona is a recipient of a Severo Ochoa Centre of Excellence Award from the Spanish Ministry of Economy and Competitiveness (MINECO) (Government of Spain) and is supported by CERCA (Generalitat de Catalunya). DT is supported by project SAF2015-74072- JIN funded by the Agencia Estatal de Investigacion (AEI) and Fondo Europeo de Desarrollo Regional (FEDER). CR-P is funded by FPI MINECO grant (BES-2013-063354). AG-P is supported by a Ramon y Cajal fellowship (RYC-2013-14554).

REFERENCES

1. Chen, D. S. & Mellman, I. Oncology meets immunology: the cancer-immunity cycle. *Immunity* **39,** 1–10 (2013).
2. Burnet, M. Cancer--A Biological Approach: III. Viruses Associated with Neoplastic Conditions. IV. Practical Applications. *BMJ* **1,** 841–847 (1957).
3. Dunn, G. P., Bruce, A. T., Ikeda, H., Old, L. J. & Schreiber, R. D. Cancer immunoediting: from immunosurveillance to tumor escape. *Nat. Immunol.* **3,** 991–998 (2002).
4. Hanahan, D. & Weinberg, R. A. Hallmarks of Cancer: The Next Generation. *Cell* **144,** 646–674 (2011).
5. Mahoney, K. M., Freeman, G. J. & McDermott, D. F. The Next Immune-Checkpoint Inhibitors: PD-1/PD-L1 Blockade in Melanoma. *Clin. Ther.* **37,** 764–782 (2015).
6. Prieto, P. A. *et al.* CTLA-4 blockade with ipilimumab: long-term follow-up of 177 patients with metastatic melanoma. *Clin. Cancer Res.* **18,** 2039–2047 (2012).
7. McDermott, D. *et al.* Efficacy and safety of ipilimumab in metastatic melanoma patients surviving more than 2 years following treatment in a phase III trial (MDX010-20). *Ann. Oncol.* **24,** 2694–2698 (2013).
8. Ascierto, P. A. *et al.* Clinical experience with ipilimumab 3 mg/kg: real-world efficacy and safety data from an expanded access programme cohort. *J. Transl. Med.* **12,** 116 (2014).
9. McDermott, D. F. *et al.* Atezolizumab, an Anti-Programmed Death-Ligand 1 Antibody, in Metastatic Renal Cell Carcinoma: Long-Term Safety, Clinical Activity, and Immune Correlates From a Phase Ia Study. *J. Clin. Oncol.* **34,** 833–842 (2016).
10. Weber, J. S. *et al.* Safety, Efficacy, and Biomarkers of Nivolumab With Vaccine in Ipilimumab-Refractory or -Naive Melanoma. *J. Clin. Oncol.* **31,** 4311–4318 (2013).
11. Becht, E. *et al.* Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression. *Genome Biol.* **17,** 218 (2016).
12. Bindea, G. *et al.* Spatiotemporal dynamics of intratumoral immune cells reveal the immune landscape in human cancer. *Immunity* **39,** 782–795 (2013).
13. Charoentong, P. *et al.* Pan-cancer Immunogenomic Analyses Reveal Genotype-Immunophenotype Relationships and Predictors of Response to Checkpoint Blockade. *Cell Rep.* **18,** 248–262 (2017).
14. Yoshihara, K. *et al.* Inferring tumour purity and stromal and immune cell admixture from expression data. *Nat. Commun.* **4,** 2612 (2013).
15. Newman, A. M. *et al.* Robust enumeration of cell subsets from tissue expression profiles. *Nat. Methods* **12,** 453–457 (2015).
16. Angelova, M. *et al.* Characterization of the immunophenotypes and antigenomes of colorectal cancers reveals distinct tumor escape mechanisms and novel targets for immunotherapy. *Genome Biol.* **16,** 64 (2015).
17. Gentles, A. J. *et al.* The prognostic landscape of genes and infiltrating immune cells across human cancers. *Nat. Med.* **21,** 938–945 (2015).
18. Li, B. *et al.* Comprehensive analyses of tumor immunity: implications for cancer immunotherapy. *Genome Biol.* **17,** 174 (2016).
19. Şenbabaoğlu, Y. *et al.* Tumor immune microenvironment characterization in clear cell renal cell carcinoma identifies prognostic and immunotherapeutically relevant messenger RNA signatures. *Genome Biol.* **17,** 231 (2016).
20. Ali, H. R., Chlon, L., Pharoah, P. D. P., Markowetz, F. & Caldas, C. Patterns of Immune Infiltration in Breast Cancer and Their Clinical Implications: A Gene-Expression-Based Retrospective Study. *PLoS Med.* **13,** e1002194 (2016).
21. Davoli, T., Uno, H., Wooten, E. C. & Elledge, S. J. Tumor aneuploidy correlates with markers of immune evasion and with reduced response to immunotherapy. *Science* **355,** (2017).
22. Rooney, M. S., Shukla, S. A., Wu, C. J., Getz, G. & Hacohen, N. Molecular and genetic properties of tumors associated with local immune cytolytic activity. *Cell* **160,** 48–61 (2015).
23. Van Allen, E. M. *et al.* Genomic correlates of response to CTLA-4 blockade in metastatic melanoma. *Science* **350,** 207–211 (2015).
24. Hugo, W. *et al.* Genomic and Transcriptomic Features of Response to Anti-PD-1 Therapy in Metastatic Melanoma. *Cell* **165,** 35–44 (2016).
25. Hänzelmann, S., Castelo, R. & Guinney, J. GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinformatics* **14,** 7 (2013).
26. Jones, E., Oliphant, T. & Peterson, P. SciPy: Open Source Scientific Tools for Python. (2001). Available at: http://www.scipy.org/. (Accessed: May 2017)
27. Aran, D. *et al.* Widespread parainflammation in

human cancer. *Genome Biol.* **17,** 145 (2016).

28. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* **102,** 15545–15550 (2005).
29. Barbie, D. A. *et al.* Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. *Nature* **462,** 108–112 (2009).
30. Mellman, I., Coukos, G. & Dranoff, G. Cancer immunotherapy comes of age. *Nature* **480,** 480–489 (2011).
31. Gajewski, T. F., Schreiber, H. & Fu, Y.-X. Innate and adaptive immune cells in the tumor microenvironment. *Nat. Immunol.* **14,** 1014–1022 (2013).
32. Galluzzi, L., Buqué, A., Kepp, O., Zitvogel, L. & Kroemer, G. Immunogenic cell death in cancer and infectious disease. *Nat. Rev. Immunol.* **17,** 97–111 (2017).
33. Gjerstorff, M. F., Andersen, M. H. & Ditzel, H. J. Oncogenic cancer/testis antigens: prime candidates for immunotherapy. *Oncotarget* **6,** 15772–15787 (2015).
34. Akers, S. N., Odunsi, K. & Karpf, A. R. Regulation of cancer germline antigen gene expression: implications for cancer immunotherapy. *Future Oncol.* **6,** 717–732 (2010).
35. Rabinovich, G. A., Gabrilovich, D. & Sotomayor, E. M. Immunosuppressive strategies that are mediated by tumor cells. *Annu. Rev. Immunol.* **25,** 267–296 (2007).
36. Wang, T. *et al.* Regulation of the innate and adaptive immune responses by Stat-3 signaling in tumor cells. *Nat. Med.* **10,** 48–54 (2004).
37. Ho, P.-C. & Liu, P.-S. Metabolic communication in tumors: a new layer of immunoregulation for immune evasion. *J Immunother Cancer* **4,** 4 (2016).
38. Wong, W. Metabolic competition between tumors and T cells. *Sci. Signal.* **8,** ec281–ec281 (2015).
39. McNamee, E. N., Korns Johnson, D., Homann, D. & Clambey, E. T. Hypoxia and hypoxia-inducible factors as regulators of T cell development, differentiation, and function. *Immunol. Res.* **55,** 58–70 (2013).
40. Facciabene, A. *et al.* Tumour hypoxia promotes tolerance and angiogenesis via CCL28 and Treg cells. *Nature* **475,** 226–230 (2011).
41. Krock, B. L., Skuli, N. & Simon, M. C. Hypoxia-Induced Angiogenesis: Good and Evil. *Genes Cancer* **2,** 1117–1133 (2011).
42. Pugh, C. W. & Ratcliffe, P. J. Regulation of angiogenesis by hypoxia: role of the HIF system. *Nat. Med.* **9,** 677–684 (2003).
43. Moeller, B. J. *et al.* The relationship between hypoxia and angiogenesis. *Semin. Radiat. Oncol.* **14,** 215–221 (2004).
44. Willenborg, S. *et al.* CCR2 recruits an inflammatory macrophage subpopulation critical for angiogenesis in tissue repair. *Blood* **120,** 613–625 (2012).
45. Rahat, M. A., Hemmerlein, B. & Iragavarapu-Charyulu, V. The regulation of angiogenesis by tissue cell-macrophage interactions. *Front. Physiol.* **5,** 262 (2014).
46. Ribatti, D., Nico, B., Crivellato, E. & Vacca, A. Macrophages and tumor angiogenesis. *Leukemia* **21,** 2085–2089 (2007).
47. Nourshargh, S., Hordijk, P. L. & Sixt, M. Breaching multiple barriers: leukocyte motility through venular walls and the interstitium. *Nat. Rev. Mol. Cell Biol.* **11,** 366–378 (2010).
48. Morwood, S. R. & Nicholson, L. B. Modulation of the immune response by extracellular matrix proteins. *Arch. Immunol. Ther. Exp.* **54,** 367–374 (2006).
49. Sorokin, L. The impact of the extracellular matrix on inflammation. *Nat. Rev. Immunol.* **10,** 712–723 (2010).
50. Pickup, M., Novitskiy, S. & Moses, H. L. The roles of TGFβ in the tumour microenvironment. *Nat. Rev. Cancer* **13,** 788–799 (2013).
51. Feig, C. *et al.* Targeting CXCL12 from FAP-expressing carcinoma-associated fibroblasts synergizes with anti-PD-L1 immunotherapy in pancreatic cancer. *Proc. Natl. Acad. Sci. U. S. A.* **110,** 20212–20217 (2013).
52. Spranger, S. & Gajewski, T. F. A new paradigm for tumor immune escape: β-catenin-driven immune exclusion. *J Immunother Cancer* **3,** 43 (2015).
53. Sweis, R. F. *et al.* Molecular Drivers of the Non-T-cell-Inflamed Tumor Microenvironment in Urothelial Bladder Cancer. *Cancer Immunol Res* **4,** 563–568 (2016).
54. Vinay, D. S. *et al.* Immune evasion in cancer: Mechanistic basis and therapeutic strategies. *Semin. Cancer Biol.* **35 Suppl,** S185–98 (2015).
55. Gonnissen, A., Isebaert, S. & Haustermans, K. Targeting the Hedgehog signaling pathway in cancer: beyond Smoothened. *Oncotarget* **6,** 13899–13913 (2015).
56. Otsuka, A. *et al.* Hedgehog pathway inhibitors promote adaptive immune responses in basal cell carcinoma. *Clin. Cancer Res.* **21,** 1289–1297 (2015).
57. Yoshimoto, A. N. *et al.* Hedgehog pathway

signaling regulates human colon carcinoma HT-29 epithelial cell line apoptosis and cytokine secretion. *PLoS One* **7,** e45332 (2012).

58. Hanna, A. & Shevde, L. A. Hedgehog signaling: modulation of cancer properies and tumor mircroenvironment. *Mol. Cancer* **15,** 24 (2016).

SUPPLEMENTARY MATERIAL

Identification of tumor immune avoidance processes across 28 solid tumors.

SUPPLEMENTARY METHODS

1.Selection of immune population identification method

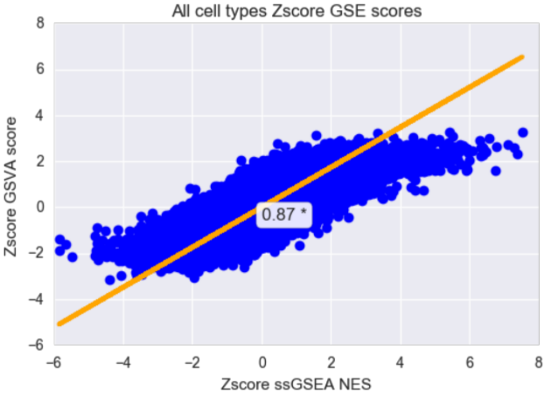
Immune populations can be estimated from RNA-seq bulk tumors through computational methods. There are two different types of methods currently being used to identify immune cell populations: (i) gene set enrichment methods [1],[2],[3],[4] used to estimate the absolute amount of each cell population in each tumor sample, so taking into account not only the relative amount of the population but also the overall immune infiltrate; and (ii) deconvolution methods [5],[6], used to estimate the relative proportion of each cell type in each tumor sample.

Deconvolution methods have several pros and cons. As a “pro”, they identify the exact proportion of each cell type within a sample. As drawbacks (i) they have not been validated for RNAseq data, (ii) they rely on reference expression matrices that cannot be customized, as the populations and genes in the matrix are fixed; and (iii) they should be used with another method/estimate to assess the overall infiltration within the samples. On the contrary, gene set enrichment methods (i) can be customized, at the level of which gene signatures and which genes in each signature should be included, (ii) they have been validated for RNAseq data[1], and (iii) they identify the overall infiltration for each cell type. Therefore, given the aim of this work and the sequencing data available, we decided to use gene set enrichment methods to estimate immune population infiltration within the tumors.

2.Selection of sample level enrichment method

The most widely used methods to perform a sample level enrichment are Gene Set Varation Analysis (GSVA)[7] and single sample Gene Set Enrichment Analysis (ssGSEA[8]). Both are unsupervised Gene Set Enrichment (GSE) methods that compute an enrichment score for each gene set in each individual sample. The main difference is that GSVA first normalizes gene expression profiles over the analyzed samples, helping to reduce the noise, while ssGSEA does not normalize. Moreover, the comparison with ssGSEA in Hanzelmann et al. (2013) shows that GSVA performs better than ssGSEA when modeling gene set enrichment over a sample population

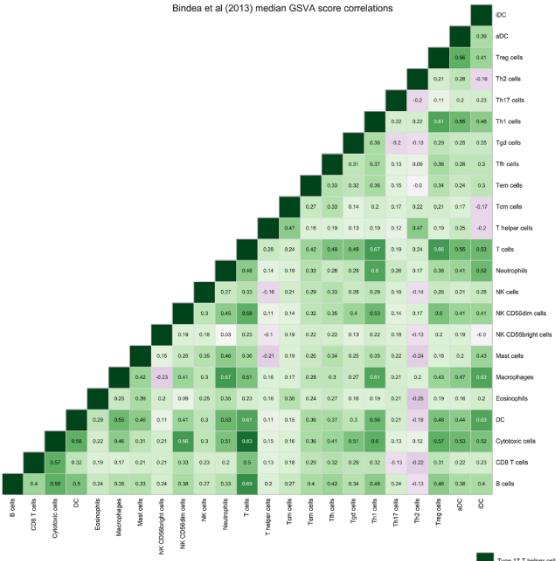
Using the gene sets of Bindea et al. (2013) we compared the gene set enrichment scores (GSEs) computed through both methods, ssGSEA and GSVA, across cancer types. We run them using the R Bioconductor package *gsva* 3.5 with default parameters. We found a positive significant (P-value < 0.05) Pearson's correlation of 0.87 between both methodologies. The scatter plot below represents the correlation between GSEs of ssGSEA and GSVA, with the regression line. The immune population *Tgd* was discarded for the sake of representation for being a depleted outlier according ssGSEA. The GSEs were normalized across each cell type in each cancer type through a Z-score to make the scales comparable.

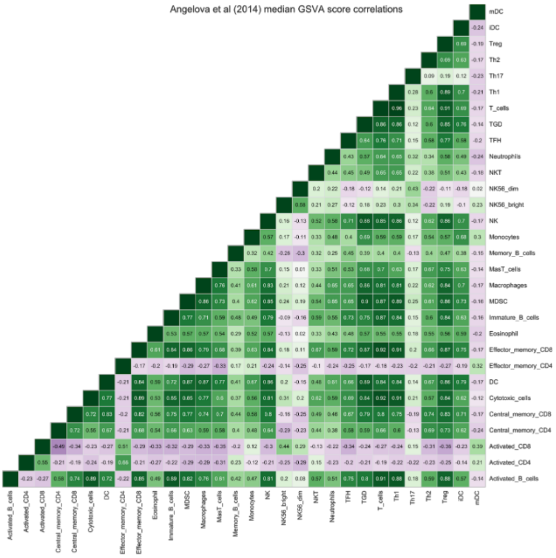


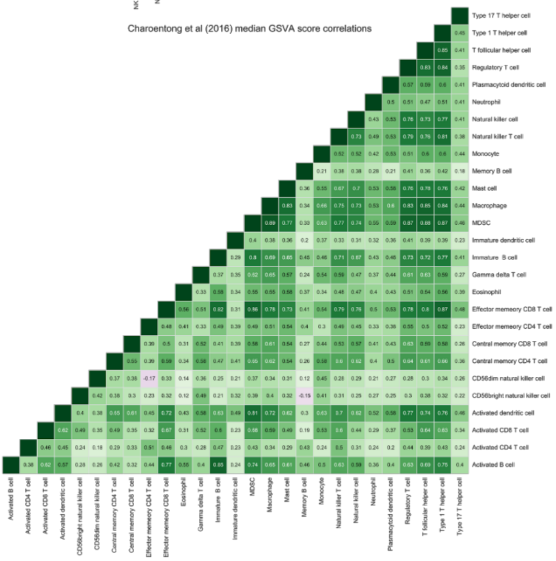
3.Selection of immune signatures for GSE

We downloaded immune signatures from the supplemental material of three different studies: (1) Bindea et al. (2013), (2) Angelova et al. (2014) and (3) Charoentong et al. (2016). The methodology followed to identify the gene signatures by the three works was similar. The only dataset with experimental validation was (1). Senbabaoglu et al. validated the enrichment of 5 immune populations (NK, T cell CD8, T cell CD4, T regulatory and Macrophages) from (1) according to ssGSEA enrichment score with FACS and/or immunofluorescence. To decide which gene signature dataset would be better to use, we observed the performance of the enrichment of the different signatures.

First, we tested if the enrichment between the different populations in each gene signature dataset was highly correlated among all cell types, to evince if each dataset could discriminate better between cell populations. We observed that GSEs from (3) were highly correlated, with a median Pearson correlation across all cell types 0.5; (2) dataset showed very highly correlated GSEs across cell types and anti-correlated ones too for certain populations across most of the other cell types (e.g. Activated CD8); GSEs from (1) were mostly positively correlated among cell populations in a lower degree than (3). After this analysis, we decided to exclude dataset (2) from further selection due to the anti-correlations among certain cell populations, a priori without a biological explanation. Moreover, we decided to prioritize the selection of dataset (1) as seemed better to distinguish between cell types, for showing lower correlations.



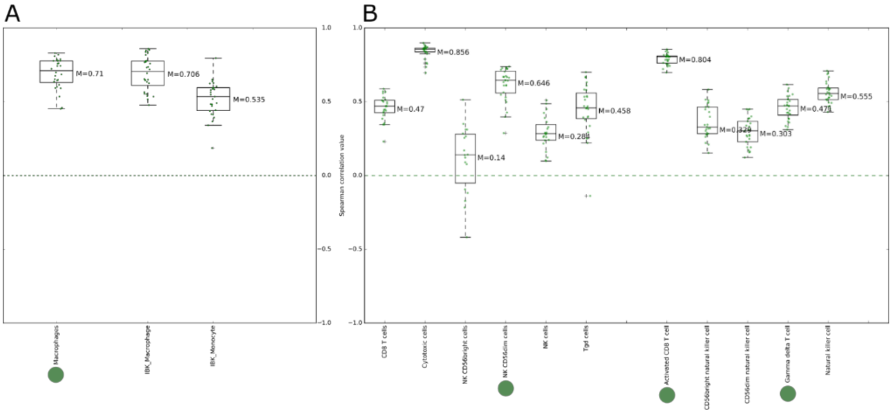




Next, we correlated the GSEs from (1) and (3) with well-known markers of certain immune populations:

- Macrophages with the myeloid marker *CD68*
- Effector populations (T cells CD8, NK cells and T cells gamma delta) with cytolytic activity as described by Rooney et al. (2015)[8,9] (computing the geometric mean between the expression of *PRF1* and *GZMA*).

The boxplots below show the correlation values with macrophage gene signatures and *CD68* (A) and effector immune populations and cytolytic activity (B), measured as the geometric mean between GZMA and PRF1 expression. Only significant correlations are shown (Q value < 0.1). Beside each boxplot the median of all spearman correlation values is shown.



We observed how among Macrophages gene signatures the one performing better corresponded to gene set (1) and among effector populations NK cells dim subtype was also performing better in (1), while CD T cells and T cells gamma delta performed better in (3).

Due to all argued reasons we decided to focus on immune populations from (1) but changing the gene signatures of CD8 T cells and T cells gamma delta for those from (3). Besides, we also considered the (3) gene signature for T regulatory cells instead of the one in (1) because in (1) it was formed by a single gene. GSE methods are known to overestimate enrichment if they are run with very small gene sets, not being suitable for using a single gene. Following this rationale, we also discarded the gene signature of pDC from (1). At last, we discarded some cell types from (1) for being redundant (e.g. NK cells include NK CD56 bright and dim).

We ended up with a dataset of sixteen gold immune signatures (hereafter named GImmS). From them, three were from data set (3) and 13 from dataset (1): B cells, Eosinophils, Macrophages, Mast cells, NK CD56bright cells, NK CD56dim cells, Neutrophils,T helper cells, Tcm cells, Tem cells, Tfh cells, iDC, aDC, Activated CD8 T cell, Gamma delta T cell and Regulatory T cell.

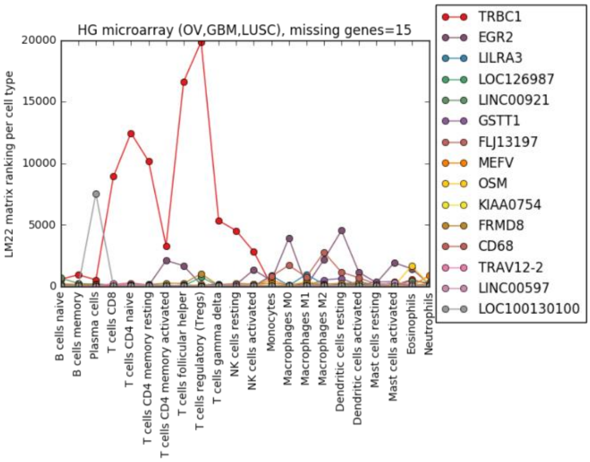
Cytotoxic cells gene signature was discarded from the GImmS dataset for being redundant with other immune populations (Gamma delta T cell, NK CD56bright cells, NK CD56dim cells and Activated CD8 T cell). However, it was used after for the identification of immune-clusters at cancer type level (see Methods).

4.Comparison between a deconvolution method and a GSE method

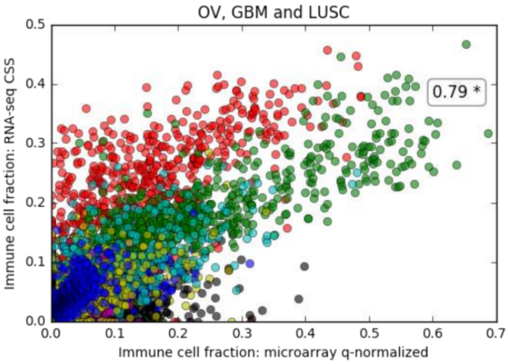
To be exhaustive, even if we have already decided to use a GSE method, we compared the performance of GSVA with the one of a deconvolution method, CIBERSORT[5]. We did not consider comparing with MCP-counter[6] because they only consider 8 immune lineages (and fibroblast). The selected GImmS were more specific and did not only include many immune lineages, being better comparable with CIBERSORT. CIBERSORT is a deconvolution method that estimates the fraction of infiltration for 22 cell types (LM22 matrix). It is trained with microarray data. Because the distribution between microarrays and RNA-seq data is different, input directly RNA-seq data into CIBERSORT impairs the fitting with LM22 matrix, worsen the performance of the method. Moreover, not having data for all the genes in LM22 matrix (548 genes) also impairs the performance of the method.

There is only microarray data for three cancer types (OV, GBM and LUSC) out of the 28 solid tumors currently available with RNA-seq data in TCGA. Following the procedure described in Charoentong et al 2016, we build a model with the TCGA microarray data to transform the RNA-seq data of all the cancer types into microarray-like.

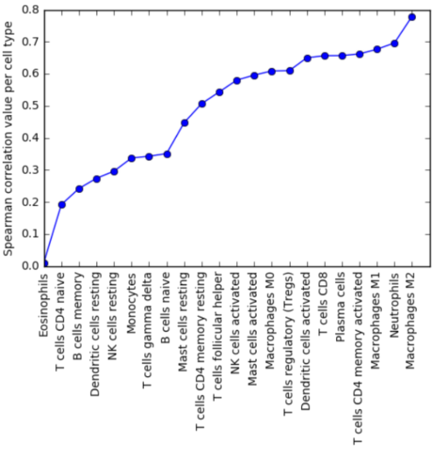
First, we downloaded HG platform (level 2) data on probe-sample-intensities from the GEO repository for OV, GBM and LUSC. If one probe was mapping to more than one gene, the probe with the highest mean intensity was kept. Probe-symbol mapping was done with Ensembl v79, we also re-mapped manually some gene symbols to maximize LM22 matrix coverage. We retrieved intensities for 13872 genes, 533/548 were found in the LM22 matrix. Additionally, we performed a quantile-normalization of microarray data, following the standard procedure of microarray data analysis. The plot below shows the contribution of the genes from LM22 matrix not found in TCGA microarray data to the identification of each cell type. It can be observed how some of the missing genes (specially *TRBC1*) may impair the detection of several cell types.



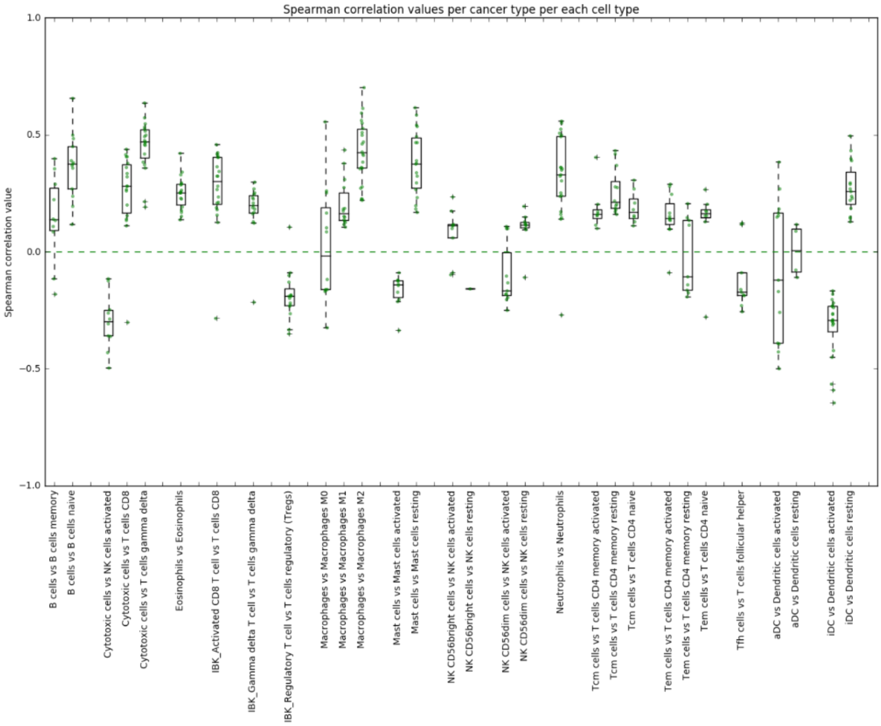
Second, we built an univariate cubic smoothing spline (CSS) model with four degrees of freedom with TCGA microarray data using the *interp1d* library from *scipy interpolate* python module. Then, we used this model to transform the RNA-seq data into microarray-like data. The performance of the model was assessed for OV, GBM and LUSC with a leave-one-out cross-validation. Considering all the cell types and cancer types together, we obtained a good overall performance of the method. We found a 0.79 significant (P-value < 0.05) Spearman correlation between the immune cell fractions of RNA-seq transformed data and microarray data. The plot below shows the correlation, each dot represents the sample fraction of a tumor sample, and each color represents a different cell population.



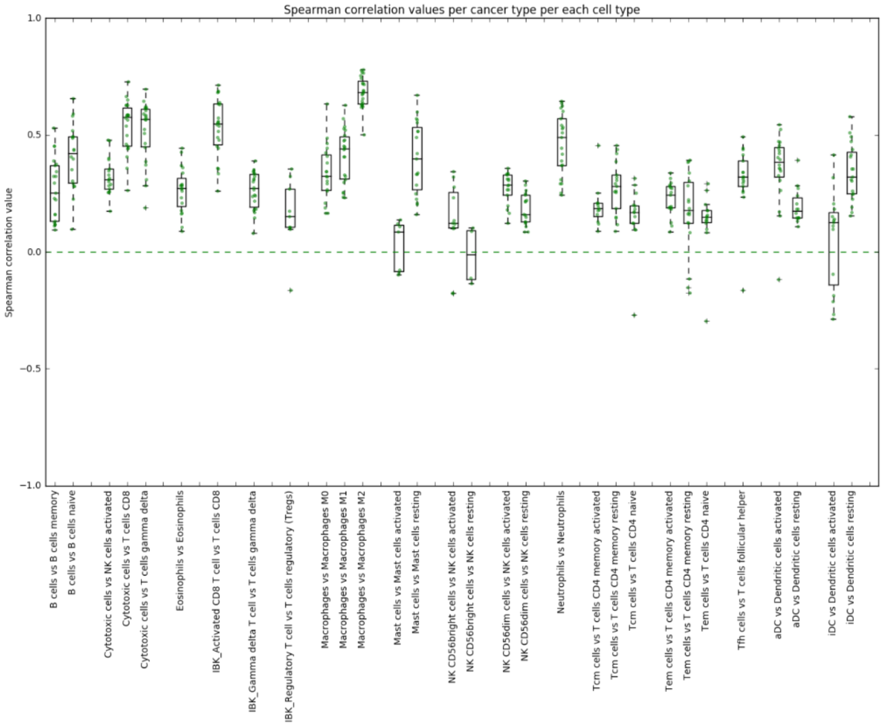
Going into detail, we observed how using the CSS to transform RNA-seq data did not performed equally for the estimation of all cell populations. The plot below shows the correlation value of immune cell fraction values per each cell type using microarray data and RNA-seq transformed data. Hence, even if the overall modeling of RNA-seq data looks good, we cannot conclude that the identification of all cell types through CIBERSORT with RNA-seq is adequate.



Third, we compared the enrichment values of the GImmS by using GSVA with the cell fractions obtained after running CIBERSORT with RNA-seq transformed data. To do so, we first mapped the cell types of the GImmS to the ones of the LM22 matrix. We did not found good correlation for most the cell types. Indeed, we found anti-correlations for some cell types (e.g. NK cells activated from CIBERSORT with Cytotoxic cells from GSE). The box plots below represent the significant (Q-value < 0.1) median Pearson correlation values in each cancer type across cell types. Correlations are labeled as GImmS GSVA cell type vs CIBERSORT cell type.



At last, we repeated the same comparison considering that CIBERSORT and GSE methods do not measure the same. While GSEA methods infer the overall infiltration for each cell type, CIBERSORT infers the relative abundance of each cell type infiltrating the tumor. Thus, we looked again the same correlation but multiplying CIBERSORT cell proportions times overall infiltration of the tumor. We used ESTIMATE Immune score[10] to measure the overall infiltration in each tumor. As can be observed in the boxplots below, the correlations improved, dramatically for some cell type comparisons.



From these analyses, we concluded that after modeling RNA-seq data into microarray-like, CIBERSORT can be run. However, the relative proportions of some cell types, such as NK cells, should be considered cautiously, as according the CSS modeling they do not highly correlate with their identification in microarray data. Besides, the correlation with a GSE approach shows positive concordance between both methodologies, even if not quite high across all cell types.

5. Immune-cluster hierarchical clustering

Hierarchical clustering was performed minimizing the squared Euclidean distance between the agglomerated samples by using the Ward method. Samples were assigned to one of the *n* clusters according to the resulting linkage matrix, where *n* represents the total number of clusters of the partition. To determine the number of clusters in which the cohort is divided, we measured the percentage of variance (VAR) of the data explained as a function of a range of *n* values:

$$VAR = 1 - (SSE/SST) \qquad SST = \sum_{j=1}^{m} d(x_j, \bar{x})^2 \qquad SSE = \sum_{i=1}^{n} SSE_i \qquad SSE_i = \sum_{x \in C_i} d(x, \bar{x}_i)^2$$

where *m* is the overall number of samples of the cohort, and *i represents* one of the *n* clusters; *d* states the euclidean distance and the cohort and cluster centroids, respectively.

In the case of the pan-cancer pooled analysis, the VAR increased at a similar rate (<1%) for each additional cluster after *n*=9, and a clear cutoff (e.g. following the elbow approach) could not be established. Therefore, and as an orthogonal observation, we evaluated how the cancer cohorts distributed across the clusters when the *n* was larger than 9. First, we estimated the degree of dispersion in the distribution of cancer types in each of the *n* clusters by the entropy score:

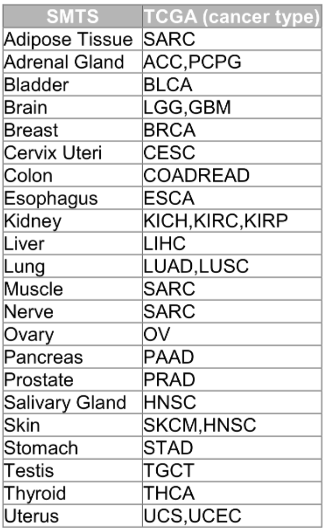
$$H_i = -\sum_{t=1}^{k} p_t \log_2(p_t)$$

where *p* is the proportion of tumors of a given cancer cohort grouped in the cluster *i* , and *k* is the total number of cancer cohorts.

As a result, we found that the first tercile of the *Hi* values did not decrease significantly after *n*=15; in other words, the samples grouped in those clusters more enriched by one or more cancer type(s) tend to remain stable even if the overall cohort is further split in more clusters. Second, we identified those clusters grouping large proportions (>25% of the samples) of different cancer types. We found that these pan-cancer groups do not separate in more cancer-specific clusters after *n*=17. Taken all these results together, and with the aim of favoring the creation of clusters capturing specific cancer-type immune signatures rather than to group the tumors in wider pan-cancer clusters, we opted for dividing the overall immune infiltrates in 17 groups.

6. GTEx immune infiltration

GTEX data v6 was downloaded from GTEXportal (https://gtexportal.org/home/) (sample level RPKM matrix). When there was more than one entry per gene symbol, we kept the one with the highest median expression (RPKM) across samples. GTEX tissues were mapped to TCGA cancer types according tissue SMTS GTEX annotation correspondence (e.g. Lung was mapped to LUAD and LUSC) below:



SMTS	TCGA (cancer type)
Adipose Tissue	SARC
Adrenal Gland	ACC,PCPG
Bladder	BLCA
Brain	LGG,GBM
Breast	BRCA
Cervix Uteri	CESC
Colon	COADREAD
Esophagus	ESCA
Kidney	KICH,KIRC,KIRP
Liver	LIHC
Lung	LUAD,LUSC
Muscle	SARC
Nerve	SARC
Ovary	OV
Pancreas	PAAD
Prostate	PRAD
Salivary Gland	HNSC
Skin	SKCM,HNSC
Stomach	STAD
Testis	TGCT
Thyroid	THCA
Uterus	UCS,UCEC

We ran GSVA with GImmS across all normal tissues in GTEX and compared the immune infiltration pattern with the one of pan-cancer GImmS GSVA. We also looked at the overall immune infiltration in normal tissues by looking at *CD45* (*PTPRC*) expression in GTEx samples.

We correlated the GSVA scores of the GImms with the GSVA of the corresponding tumors with a Pearson’s correlation. We considered that an immune cell type was more enriched in tumors or normals when it deviated +/-0.2 of the correlation diagonal (Figure S3).

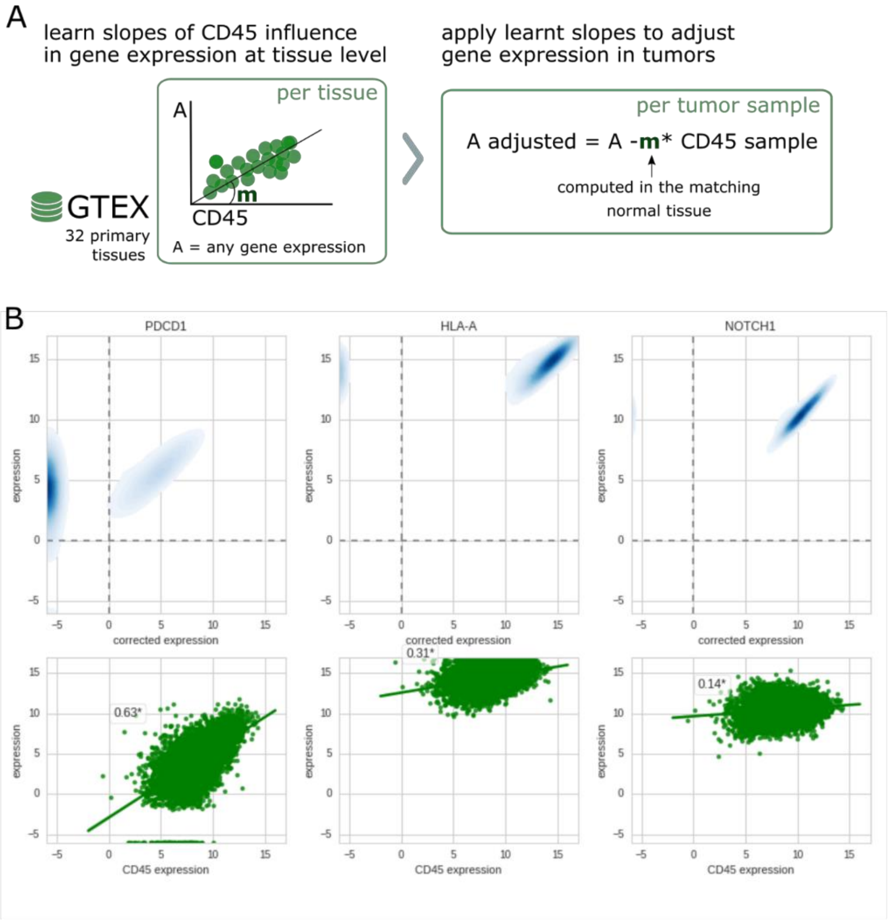
7. TCGA expression adjustment for its immune component

To adjust TCGA expression for its immune component we have followed the rationale suggested by Aran et al. (2016)[11] (Figure S1A). We have adjusted the expression levels of each gene according the expression value of *CD45.* From GTEx normal tissues, we first learnt how much each gene expression was influenced by *CD45* in each tissue. We extracted the slope of the correlation between each gene and *CD45* in each tissue, by fitting it in a degree one polynomial. Next, we mapped the GTEx tissues to TCGA tumors (see above) and for each gene expression value in each sample we subtracted the expression value of CD45 in the sample times the slope learnt from GTEx. Note that we were not able to adjust expression for UVM, CHOL and MESO for not having normal tissue data in GTEx.

To validate the results of the adjustment method we looked at specific genes described to be expressed: (1) mostly in immune but not tumor cells (e.g. PD1), (2) in both immune and tumor cells (e.g. HLA molecules), and (3) mostly in tumor but not immune cells (e.g. NOTCH1). The adjustment method worked as expected. The expression levels of genes from (1) to (3) were lowered, ranging from a dramatic decrease to almost the same expression level (Figure S1B).

Moreover, we also explored how the expression adjustment influenced the enrichment of gene sets. Again, as expected, we observed that the expression of most genes in immune-related gene sets was lowered while in oncogenic gene sets it remained stable (Figure S1C). We considered that the expression of a gene was lowered/increased if it deviated -/+2 from the diagonal of the correlation plot between adjusted and unadjusted expression.

SUPPLEMENTARY FIGURES



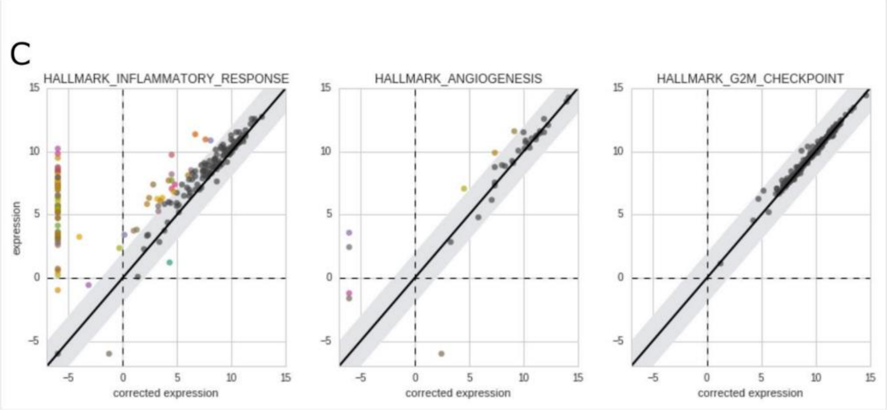


Figure S1. Tumor expression adjustment for its immune component. (A) Schema of the adjustment method which consists mainly of two steps, (i) learning slopes on *CD45* influence in gene expression in normal tissues and (ii) applying the learnt slopes to TCGA data (see supplementary methods). (B) Top. Correlation plots of the pan-cancer expression and adjusted pan-cancer expression, in log2 RSEM, of three different genes: PDCD1 (known as PD-1) which is mostly expressed in immune cells, HLA-A expressed in immune and tumor cells and NOTCH1 mostly expressed in tumor cells. Bottom. Correlation plots of the pan-cancer expression of each gene with *CD45*. (C) Gene expression changes after expression adjustment in three different pathways, from left to right inflammatory response (tightly related to the immune component of the tumors), angiogenesis (mildly related) and G2M checkpoint (mostly tumor cell specific). Each plot represents the median pan-cancer expression of each gene (represented as a dot) in the gene set before adjustment (y-axis) and after adjustment (x-axis). The genes outside the diagonal grey area, are the ones where expression is changed after adjustment (+/- 2 log2 RSEM deviation from the diagonal), leading to a gene expression decrease (on the left of the diagonal) or increase (on the right of the diagonal).

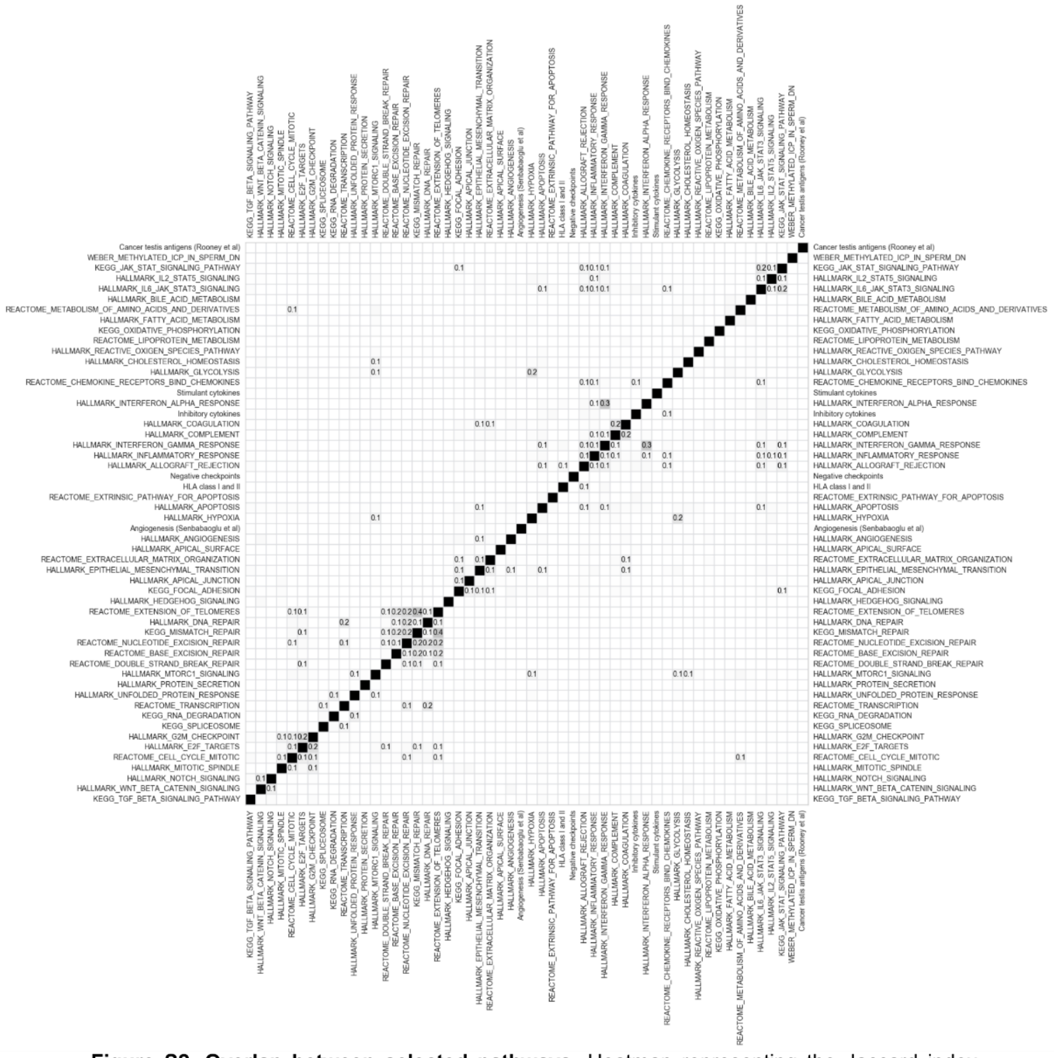


Figure S2. Overlap between selected pathways. Heatmap representing the Jaccard index between the selected pathways. Grey color gradient corresponds to the Jaccard index too. Note that the highest Jaccard index is 0.4 between REACTOME_EXTENSION_OF_TELOMERES and KEGG_MISS_MATCH REPAIR while in most of the comparisons there is no overlap or very low.

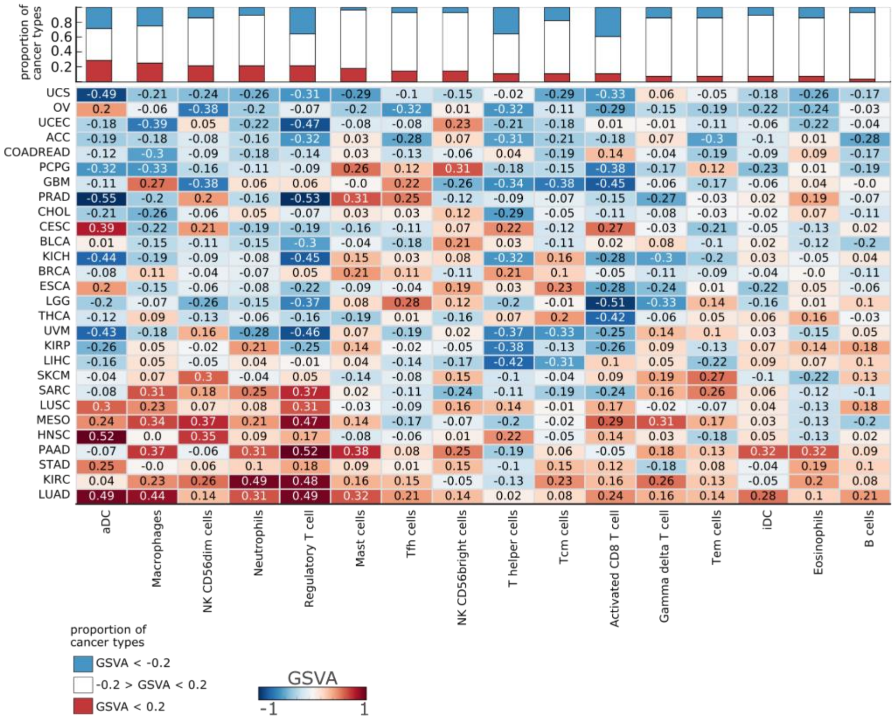
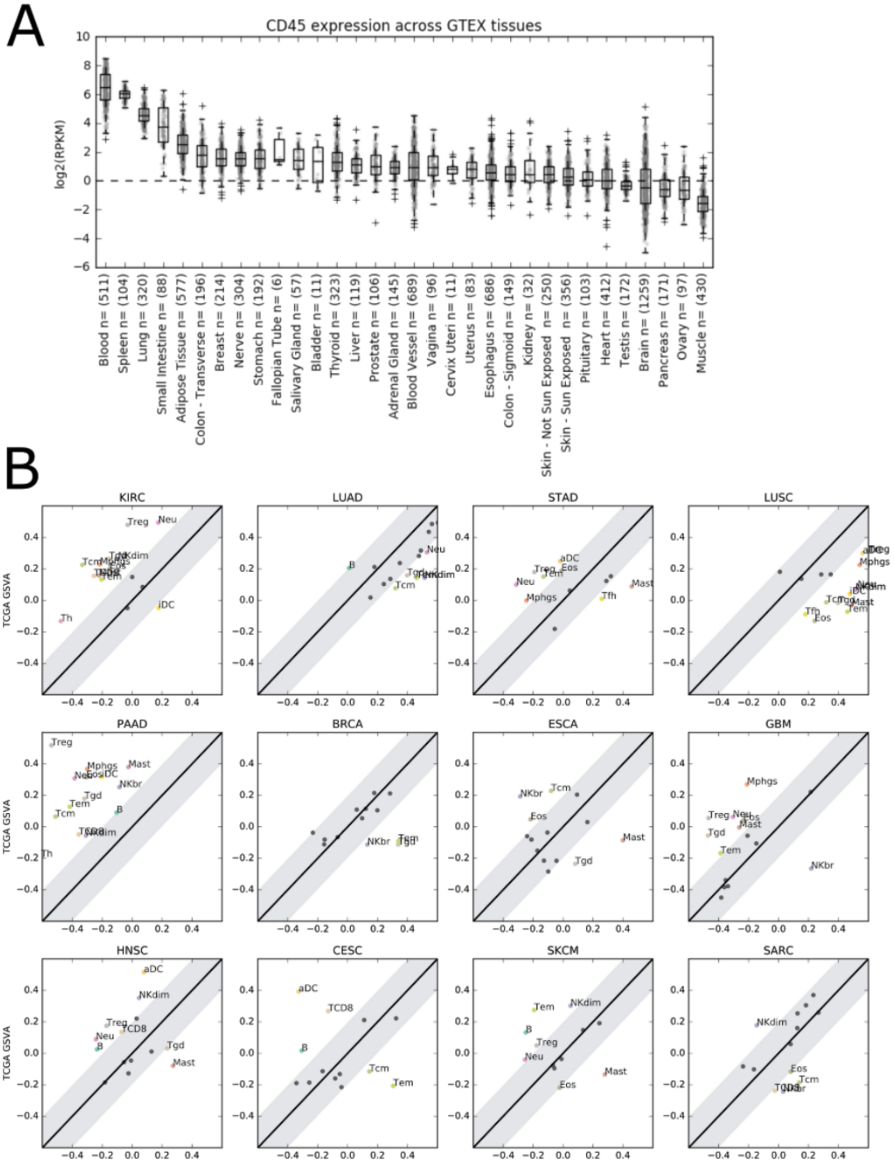


Figure S3. Enrichment distribution of cell type across cancer types.

Bottom panel, heatmap with the median pan-cancer GSVA score per cell type across cancer types. GSVA scores are colored from blue (depletion, -1) to red (enrichment, +1). X-axis represents the sixteen immune cell types and y-axis the 28 cancer types analyzed. X-axis is sorted according the number of cancer types with high enrichment in the cell type (see below). Y-axis is sorted according the number of infiltrating cell types per cancer type (GSVA > 0). Top panel, barplot of the proportion of cancer types that show a high enrichment (GSVA > 0.2), a strong depletion (GSVA < -0.2) or an intermediate infiltration pattern (-0.2 < GSVA < 0.2) across cancer cell types.


A
CD45 expression across GTEX tissues
log2(RPKM)
10
8
6
4
2
0
−2
−4
−6
Blood n= (511)
Spleen n= (104)
Lung n= (320)
Small Intestine n= (88)
Adipose Tissue n= (577)
Colon - Transverse n= (196)
Breast n= (214)
Nerve n= (304)
Stomach n= (192)
Fallopian Tube n= (6)
Salivary Gland n= (57)
Bladder n= (11)
Thyroid n= (323)
Liver n= (119)
Prostate n= (106)
Adrenal Gland n= (145)
Blood Vessel n= (689)
Vagina n= (96)
Cervix Uteri n= (11)
Uterus n= (83)
Esophagus n= (686)
Colon - Sigmoid n= (149)
Kidney n= (32)
Skin - Not Sun Exposed n= (250)
Skin - Sun Exposed n= (356)
Pituitary n= (103)
Heart n= (412)
Testis n= (172)
Brain n= (1259)
Pancreas n= (171)
Ovary n= (97)
Muscle n= (430)
B
KIRC
LUAD
STAD
LUSC
PAAD
BRCA
ESCA
GBM
HNSC
CESC
SKCM
SARC
TCGA GSVA
0.4
0.2
0.0
−0.2
−0.4
−0.4 −0.2 0.0 0.2 0.4
Treg
Neu
Tcm
Tgd
NKdim
Mphgs
Eos
TCD8
Tem
iDC
Th
B
aDC
Mast
Tfh
NKbr
Tfb

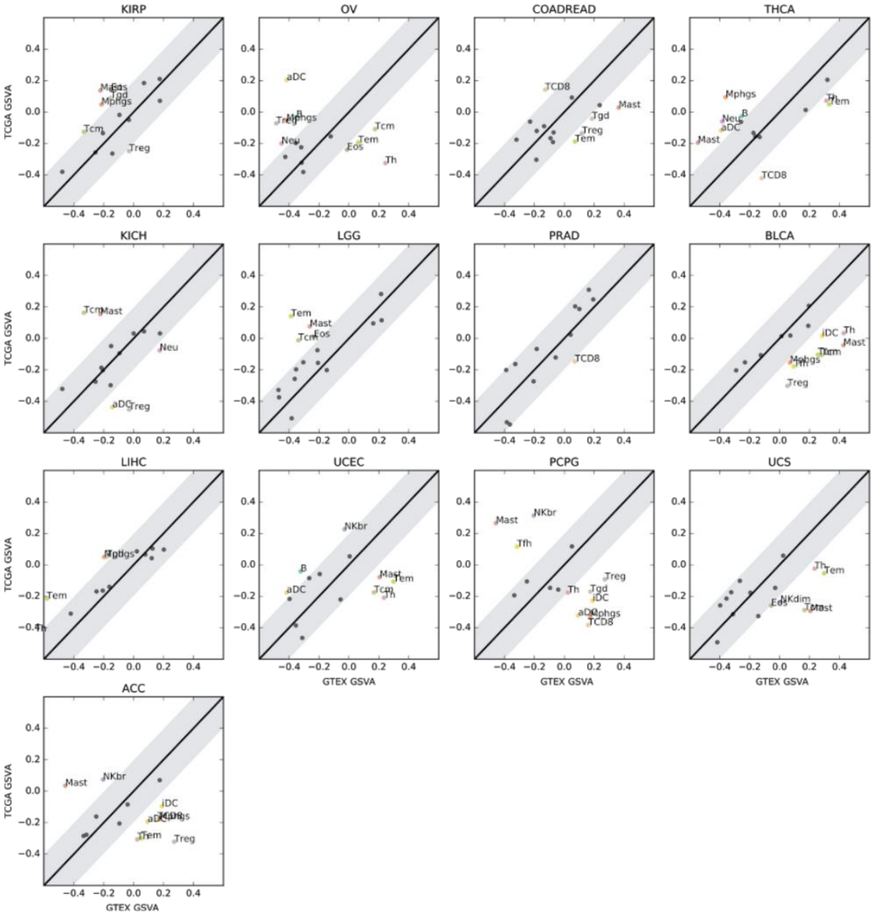
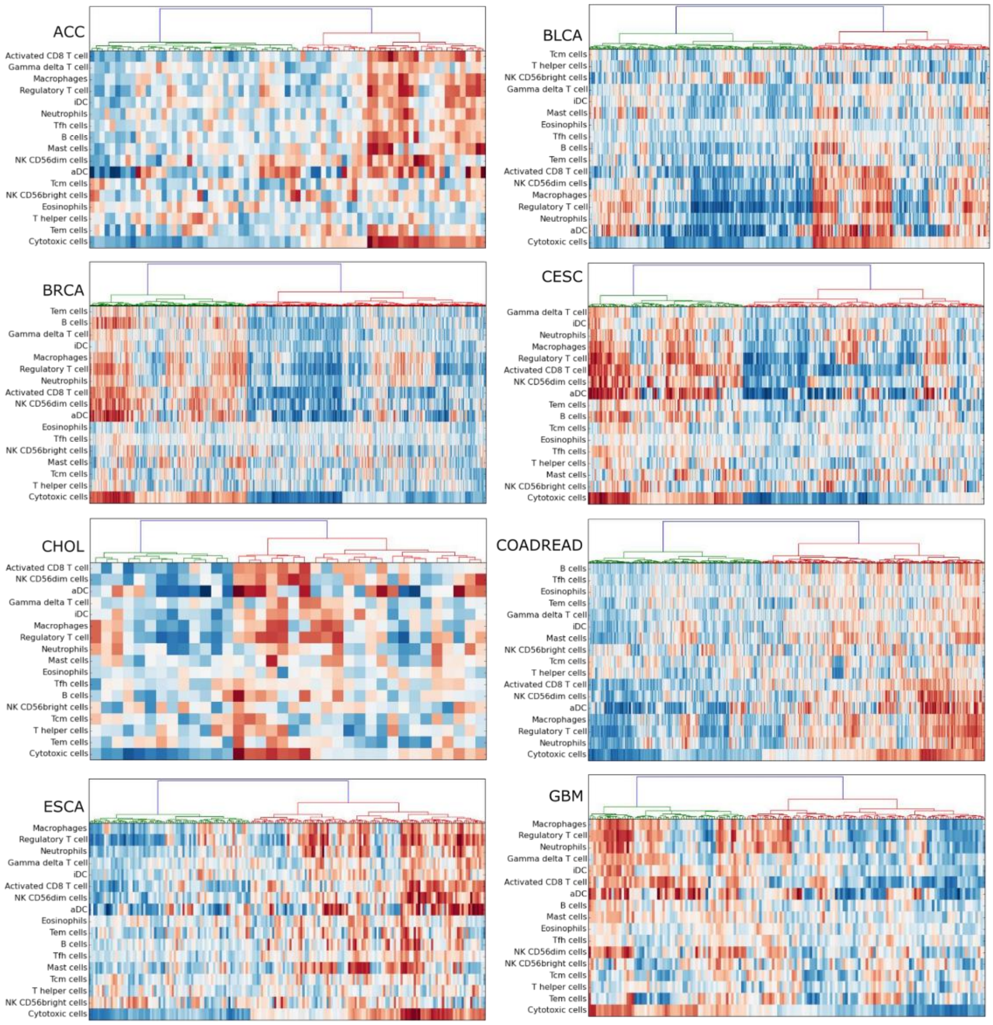
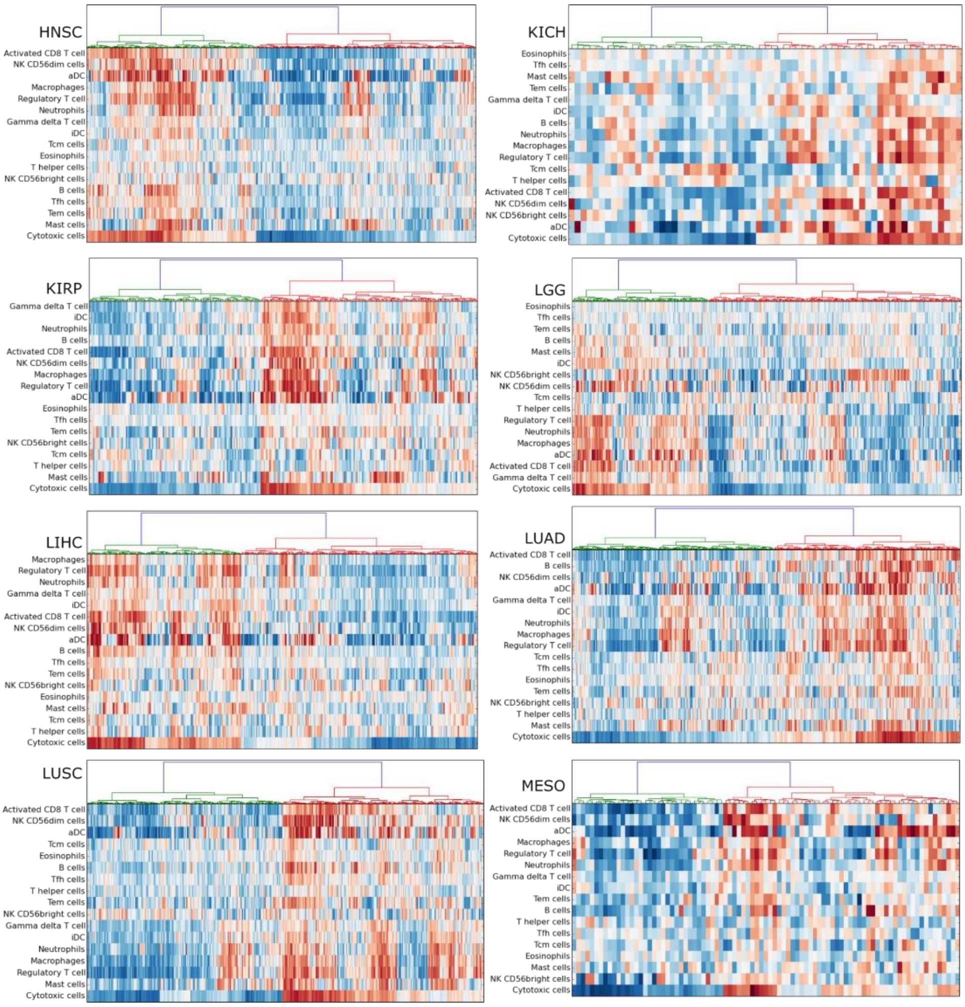
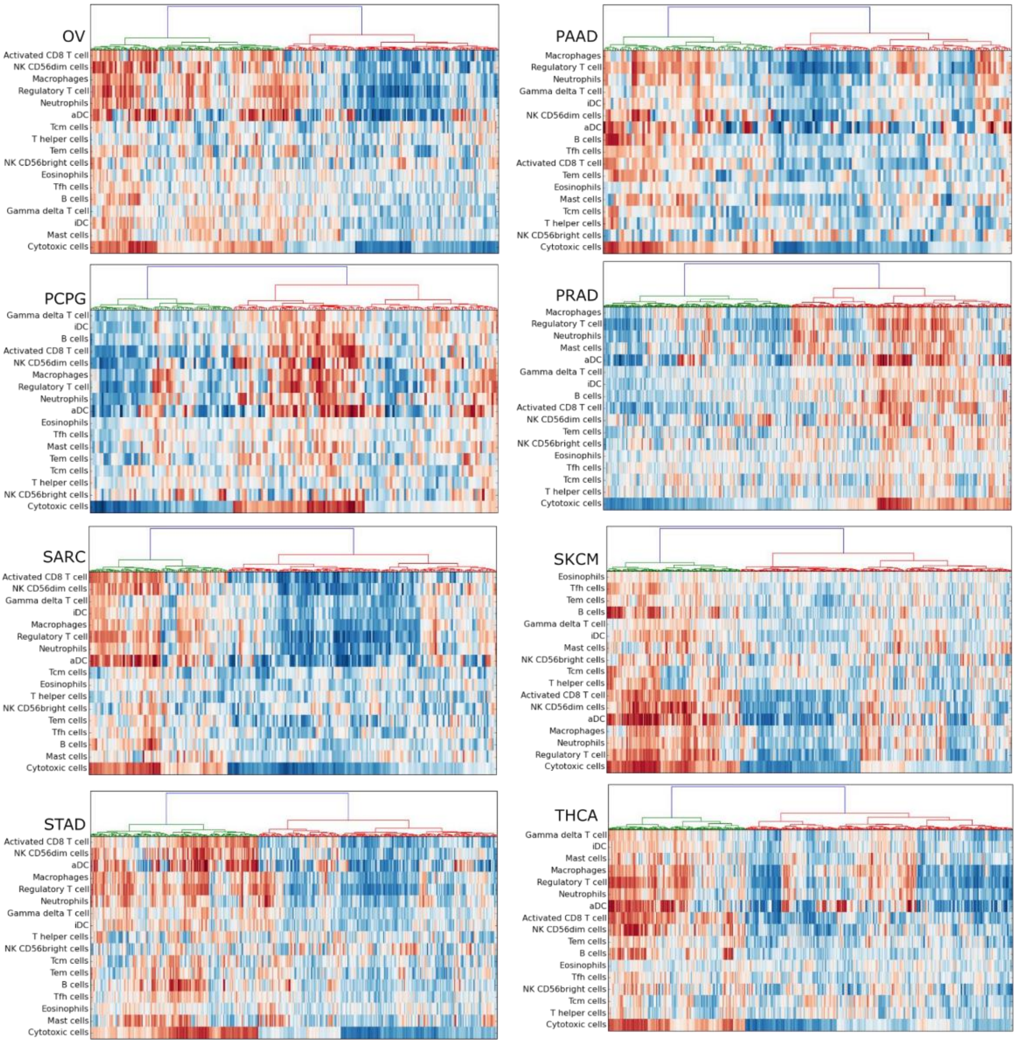


Figure S4. Comparison of immune population infiltration in normal tissues versus tumors. (A) Boxplots representing the distribution of CD45 expression across GTEX tissues in log2 RPKM, each dot represents a sample. (B) GSVA score of the sixteen immune cell types in tumors (y-axis) versus normal tissues (x-axis) across 25 cancer types (the ones with a matching normal tissue in GTEx). Each dot represents the median GSVA score of each immune cell type in each tumor-normal tissue pair. Dots outside the grey area, are the ones more differently enriched/depleted in tumors vs normal tissues (+/- 0.2 median GSVA). Scatter plots are sorted from the highest infiltration tumor (KIRC) to the lowest infiltrated one (ACC).


ACC
Activated CD8 T cell
Gamma delta T cell
Macrophages
Regulatory T cell
iDC
Neutrophils
Tfh cells
B cells
Mast cells
NK CD56dim cells
aDC
Tcm cells
NK CD56bright cells
Eosinophils
T helper cells
Tem cells
Cytotoxic cells
BLCA
Tcm cells
T helper cells
NK CD56bright cells
Gamma delta T cell
iDC
Mast cells
Eosinophils
Tfh cells
B cells
Tem cells
Activated CD8 T cell
NK CD56dim cells
Macrophages
Regulatory T cell
Neutrophils
aDC
Cytotoxic cells
BRCA
Tem cells
B cells
Gamma delta T cell
iDC
Macrophages
Regulatory T cell
Neutrophils
Activated CD8 T cell
NK CD56dim cells
aDC
Eosinophils
Tfh cells
NK CD56bright cells
Mast cells
Tcm cells
T helper cells
Cytotoxic cells
CESC
Gamma delta T cell
iDC
Neutrophils
Macrophages
Regulatory T cell
Activated CD8 T cell
NK CD56dim cells
aDC
Tem cells
B cells
Tcm cells
Eosinophils
Tfh cells
T helper cells
Mast cells
NK CD56bright cells
Cytotoxic cells
CHOL
Activated CD8 T cell
NK CD56dim cells
aDC
Gamma delta T cell
iDC
Macrophages
Regulatory T cell
Neutrophils
Mast cells
Eosinophils
Tfh cells
B cells
NK CD56bright cells
Tcm cells
T helper cells
Tem cells
Cytotoxic cells
COADREAD
B cells
Tfh cells
Eosinophils
Tem cells
Gamma delta T cell
iDC
Mast cells
NK CD56bright cells
Tcm cells
T helper cells
Activated CD8 T cell
NK CD56dim cells
aDC
Macrophages
Regulatory T cell
Neutrophils
Cytotoxic cells
ESCA
Macrophages
Regulatory T cell
Neutrophils
Gamma delta T cell
iDC
Activated CD8 T cell
NK CD56dim cells
aDC
Eosinophils
Tem cells
B cells
Tfh cells
Mast cells
Tcm cells
T helper cells
NK CD56bright cells
Cytotoxic cells
GBM
Macrophages
Regulatory T cell
Neutrophils
Gamma delta T cell
iDC
Activated CD8 T cell
aDC
B cells
Mast cells
Eosinophils
Tfh cells
NK CD56dim cells
NK CD56bright cells
Tcm cells
T helper cells
Tem cells
Cytotoxic cells


HNSC
Activated CD8 T cell
NK CD56dim cells
aDC
Macrophages
Regulatory T cell
Neutrophils
Gamma delta T cell
iDC
Tcm cells
Eosinophils
T helper cells
NK CD56bright cells
B cells
Tfh cells
Tem cells
Mast cells
Cytotoxic cells
KICH
Eosinophils
Tfh cells
Mast cells
Tem cells
Gamma delta T cell
iDC
B cells
Neutrophils
Macrophages
Regulatory T cell
Tcm cells
T helper cells
Activated CD8 T cell
NK CD56dim cells
NK CD56bright cells
aDC
Cytotoxic cells
KIRP
Gamma delta T cell
iDC
Neutrophils
B cells
Activated CD8 T cell
NK CD56dim cells
Macrophages
Regulatory T cell
aDC
Eosinophils
Tfh cells
Tem cells
NK CD56bright cells
Tcm cells
T helper cells
Mast cells
Cytotoxic cells
LGG
Eosinophils
Tfh cells
Tem cells
B cells
Mast cells
iDC
NK CD56bright cells
NK CD56dim cells
Tcm cells
T helper cells
Regulatory T cell
Neutrophils
Macrophages
aDC
Activated CD8 T cell
Gamma delta T cell
Cytotoxic cells
LIHC
Macrophages
Regulatory T cell
Neutrophils
Gamma delta T cell
iDC
Activated CD8 T cell
NK CD56dim cells
aDC
B cells
Tfh cells
Tem cells
NK CD56bright cells
Eosinophils
Mast cells
Tcm cells
T helper cells
Cytotoxic cells
LUAD
Activated CD8 T cell
B cells
NK CD56dim cells
aDC
Gamma delta T cell
iDC
Neutrophils
Macrophages
Regulatory T cell
Tcm cells
Tfh cells
Eosinophils
Tem cells
NK CD56bright cells
T helper cells
Mast cells
Cytotoxic cells
LUSC
Activated CD8 T cell
NK CD56dim cells
aDC
Tcm cells
Eosinophils
B cells
Tfh cells
T helper cells
Tem cells
NK CD56bright cells
Gamma delta T cell
iDC
Neutrophils
Macrophages
Regulatory T cell
Mast cells
Cytotoxic cells
MESO
Activated CD8 T cell
NK CD56dim cells
aDC
Macrophages
Regulatory T cell
Neutrophils
Gamma delta T cell
iDC
Tem cells
B cells
T helper cells
Tfh cells
Tcm cells
Eosinophils
Mast cells
NK CD56bright cells
Cytotoxic cells


OV
Activated CD8 T cell
NK CD56dim cells
Macrophages
Regulatory T cell
Neutrophils
aDC
Tcm cells
T helper cells
Tem cells
NK CD56bright cells
Eosinophils
Tfh cells
B cells
Gamma delta T cell
iDC
Mast cells
Cytotoxic cells
PAAD
Macrophages
Regulatory T cell
Neutrophils
Gamma delta T cell
iDC
NK CD56dim cells
aDC
B cells
Tfh cells
Activated CD8 T cell
Tem cells
Eosinophils
Mast cells
Tcm cells
T helper cells
NK CD56bright cells
Cytotoxic cells
PCPG
Gamma delta T cell
iDC
B cells
Activated CD8 T cell
NK CD56dim cells
Macrophages
Regulatory T cell
Neutrophils
aDC
Eosinophils
Tfh cells
Mast cells
Tem cells
Tcm cells
T helper cells
NK CD56bright cells
Cytotoxic cells
PRAD
Macrophages
Regulatory T cell
Neutrophils
Mast cells
aDC
Gamma delta T cell
iDC
B cells
Activated CD8 T cell
NK CD56dim cells
Tem cells
NK CD56bright cells
Eosinophils
Tfh cells
Tcm cells
T helper cells
Cytotoxic cells
SARC
Activated CD8 T cell
NK CD56dim cells
Gamma delta T cell
iDC
Macrophages
Regulatory T cell
Neutrophils
aDC
Tcm cells
Eosinophils
T helper cells
NK CD56bright cells
Tem cells
Tfh cells
B cells
Mast cells
Cytotoxic cells
SKCM
Eosinophils
Tfh cells
Tem cells
B cells
Gamma delta T cell
iDC
Mast cells
NK CD56bright cells
Tcm cells
T helper cells
Activated CD8 T cell
NK CD56dim cells
aDC
Macrophages
Neutrophils
Regulatory T cell
Cytotoxic cells
STAD
Activated CD8 T cell
NK CD56dim cells
aDC
Macrophages
Regulatory T cell
Neutrophils
Gamma delta T cell
iDC
T helper cells
NK CD56bright cells
Tcm cells
Tem cells
B cells
Tfh cells
Eosinophils
Mast cells
Cytotoxic cells
THCA
Gamma delta T cell
iDC
Mast cells
Macrophages
Regulatory T cell
Neutrophils
aDC
Activated CD8 T cell
NK CD56dim cells
Tem cells
B cells
Eosinophils
Tfh cells
NK CD56bright cells
Tcm cells
T helper cells
Cytotoxic cells

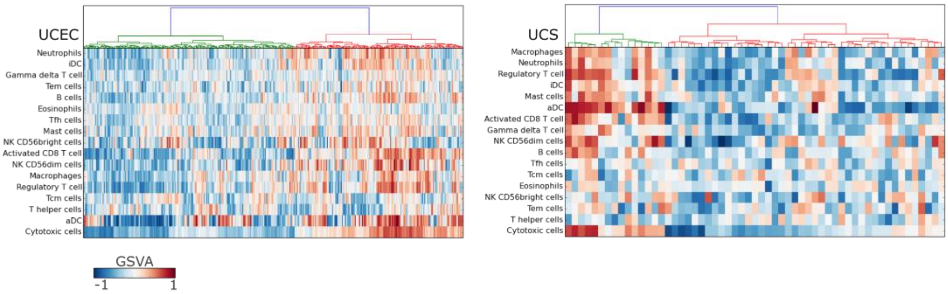


Figure S5. Per-cancer type immune-clustering.

Heatmaps with the distribution of GSVA values (from -1 to 1, from blue to red), across the 16 immune cell populations and cytotoxic cells, giving an overweight to cytotoxic cells to perform the hierarchical clustering. Hierarchical clustering is found in the top-panel of each heatmap. A heatmap for each of the remaining 26 cancer types is displayed (not including KIRC and UVM as they are found in Figure 4), each cancer type acronym is found in the top-left corner of each heatmap.

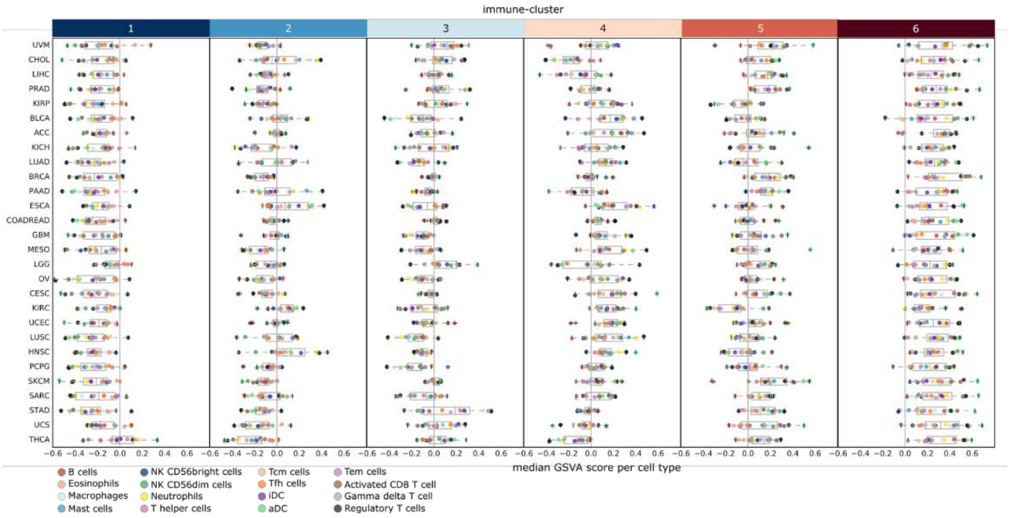


Figure S6. Heterogeneity of infiltration profiles across cancer type immune-clusters.

Boxplots with the median enrichment of each cell type (see color legend) in each cancer type immune-cluster (from 1 to 6, from left to right). Each dot represents the median GSVA enrichment of the immune population and each population is colored different (see legend).

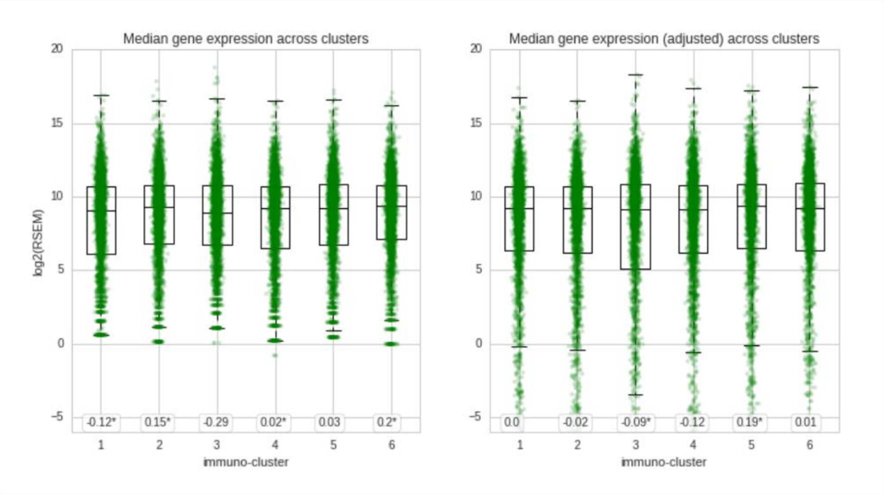


Figure S7. Expression homogeneity across immune-clusters.

Boxplots representing the median expression for each gene analyzed in the pathway analysis (3901 genes) across each immune-cluster. Below each boxplot the log Fold Change between each immune-cluster and the rest of immune-cluster pooled together is shown. An asterisk is added if the Mann Whitney U test of the mentioned comparison is significant (P-value < 0.05). Note how, even if some differences are significant any Fold Change is higher than 0.3.

SUPPLEMENTARY TABLES

Cancer type acronym	Cancer type full name	Number of patients (RNA-seq)
ACC	Adrenocortical carcinoma	78
BLCA	Bladder Urothelial Carcinoma	404
BRCA	Breast invasive carcinoma	1082
CESC	Cervical squamous cell carcinoma and endocervical adenocarcinoma	301
CHOL	Cholangiocarcinoma	36
COADREAD	Colorectal adenocarcinoma	599
ESCA	Esophageal carcinoma	182
GBM	Glioblastoma multiforme	163
HNSC	Head and Neck squamous cell carcinoma	515
KICH	Kidney Chromophobe	65
KIRC	Kidney renal clear cell carcinoma	515
KIRP	Kidney renal papillary cell carcinoma	285
LGG	Brain Lower Grade Glioma	514
LIHC	Liver hepatocellular carcinoma	368
LUAD	Lung adenocarcinoma	511
LUSC	Lung squamous cell carcinoma	485
MESO	Mesotelioma	87
OV	Ovarian serous cystadenocarcinoma	300
PAAD	Pancreatic adenocarcinoma	156
PCPG	Pheochromocytoma and Paraganglioma	178
PRAD	Prostate adenocarcinoma	493
SARC	Sarcoma	254
SKCM	Skin Cutaneous Melanoma	434
STAD	Stomach adenocarcinoma	405
THCA	Thyroid carcinoma	499
UCEC	Uterine Corpus Endometrial Carcinoma	357
UCS	Uterine Carcinosarcoma	57
UVM	Uveal Melanoma	80

Table S1. Tumor cohort details
Table with the details of the analyzed cohort. It contains the acronym, full name and number of samples with RNAseq data of the 28 solid tumors.

Table S2. Immune cell type and pathway gene sets.

(Table S2A). Table including the genes used to identify each immune populations and the source of each gene signature. It includes the gene signatures of the 16 immune populations and the gene signature of Cytotoxic cells.

(Table S2B). Table including the genes used to identify each tumor pathway and the source of each gene set.

Table S3. Pan-cancer and per-cancer type GSVA scores across immune populations.

GSVA enrichment scores per tumor samples of pan-cancer GSVA (Table S3A) and per-cancer type GSVA analyses (Tables S3B to S3ZC). Tumor samples are represented in columns and the immune cell types in rows. Each cell shows the GSVA enrichment score.

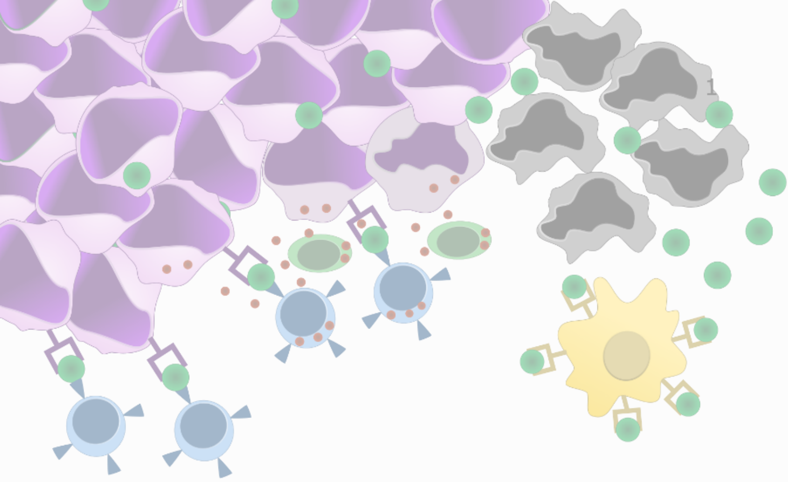
Cancer type	Table
Pan-cancer	TableS3A
THCA	TableS3B
UCS	TableS3C
STAD	TableS3D
SARC	TableS3E
SKCM	TableS3F
PCPG	TableS3G
HNSC	TableS3H
LUSC	TableS3I
UCEC	TableS3J
KIRC	TableS3K
CESC	TableS3L
OV	TableS3M
LGG	TableS3N
MESO	TableS3O
GBM	TableS3P
COADREAD	TableS3Q
ESCA	TableS3R
PAAD	TableS3S
BRCA	TableS3T
LUAD	TableS3U
KICH	TableS3V
ACC	TableS3W
BLCA	TableS3X
KIRP	TableS3Y
PRAD	TableS3Z
LIHC	TableS3ZA
CHOL	TableS3ZB
UVM	TableS3ZC

Table S4. Pan-cancer and per-cancer type level immune-clusters.
(Table S4A). Classification of each patient across the seventeen pan-cancer immune clusters.
(Table S4B). Classification of each patient across the six immune-clusters in each cancer type.

Table S5. Results of the GSEA enrichment
Table containing the output of GSEA analysis. Each row corresponds to an enrichment of a pathway in a cancer type immune-cluster compared to the other immune-clusters of the cancer type. Only significant enrichments (Q-value < 0.25) are shown. GSEA NES corresponds to the Normalized Enrichment Score of the pathway in the immune-cluster, as provided by GSEA software. GSEA FDR Q-value corresponds the P-value of the enrichment adjusted for gene set size and multiple testing, as provided by GSEA software.

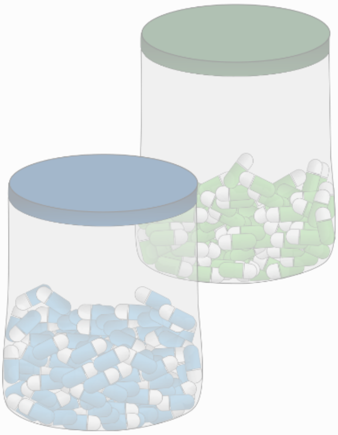
SUPPLEMENTARY REFERENCES

1. Şenbabaoğlu, Y. *et al.* Tumor immune microenvironment characterization in clear cell renal cell carcinoma identifies prognostic and immunotherapeutically relevant messenger RNA signatures. *Genome Biol.* **17,** 231 (2016).
2. Charoentong, P. *et al.* Pan-cancer Immunogenomic Analyses Reveal Genotype-Immunophenotype Relationships and Predictors of Response to Checkpoint Blockade. *Cell Rep.* **18,** 248–262 (2017).
3. Angelova, M. *et al.* Characterization of the immunophenotypes and antigenomes of colorectal cancers reveals distinct tumor escape mechanisms and novel targets for immunotherapy. *Genome Biol.* **16,** 64 (2015).
4. Bindea, G. *et al.* Spatiotemporal dynamics of intratumor immune cells reveal the immune landscape in human cancer. *Immunity* **39,** 782–795 (2013).
5. Newman, A. M. *et al.* Robust enumeration of cell subsets from tissue expression profiles. *Nat. Methods* **12,** 453–457 (2015).
6. Becht, E. *et al.* Erratum to: Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression. *Genome Biol.* **17,** (2016).
7. Hänzelmann, S., Castelo, R. & Guinney, J. GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinformatics* **14,** 7 (2013).
8. Barbie, D. A. *et al.* Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. *Nature* **462,** 108–112 (2009).
9. Rooney, M. S., Shukla, S. A., Wu, C. J., Getz, G. & Hacohen, N. Molecular and genetic properties of tumors associated with local immune cytolytic activity. *Cell* **160,** 48–61 (2015).
10. Yoshihara, K. *et al.* Inferring tumour purity and stromal and immune cell admixture from expression data. *Nat. Commun.* **4,** 2612 (2013).
11. Aran, D. *et al.* Widespread parainflammation in human cancer. *Genome Biol.* **17,** 145 (2016).



PART IV

DISCUSSION



Cancer precision medicine is aimed to choose the most suitable anti-cancer therapy for each patient based on the study of the biology of its tumor. In this direction, the evolution of the cancer molecular knowledge has provided the community remarkable advances, such as the development of the first targeted therapy, trastuzumab, which has significantly improved the prognosis of HER2+ breast cancer patients[125,161]. Besides, cancer genomics studies have proven useful for guiding targeted therapeutic strategies. Examples of this are the clinical use of vemurafenib in *BRAF* V600E mutant melanoma patients[101], cetuximab in *EGFR* L858R mutant non-small cell lung carcinoma patients[162] or imatinib in BCR-ABL chronic myeloid leukemia patients[97].

Therefore, the availability of the sequences of the exomes or genomes of tumors from thousands of patients has opened the possibility of not only comprehensively identifying the genes driving tumorigenesis, but also of estimating the scope of current and future cancer targeted therapies. The first work presented here (Chapter 1) aimed to contribute to both objectives: uncovering the landscape of cancer driver genes and the landscape of genomic-guided therapeutic opportunities available for cancer patients, based on the analysis of a large pan-cancer cohort of tumor samples (n=6792). Of note, by the start of the project, similar efforts on the direction of the pan-cancer comprehensive identification of cancer driver genes had been carried out, mostly within the framework of The Cancer Genome Atlas (TCGA) consortia and predominantly based on the study of mutational cancer driver genes[47,57,163]. Even if cancer type level comprehensive integrations of driver genes bearing several types of genomic alterations had been published within TCGA framework[51], to our knowledge we evinced one of the first pan-cancer driver integrative landscapes, by analyzing driver genes bearing mutations, copy number alterations (CNA) and chromosomal translocations (referred to as fusion drivers in Chapter 1) in 28 different cancer types. Besides, as far as we knew, only one previous study (Van Allen et al 2014) had attempted to identify the landscape of genomic-guided therapeutic opportunities in cancer, even though several strategies for compiling anti-cancer drug targets and response biomarkers were emerging[164–166]. Of note, that previous study did not consider the genes driving tumorigenesis in each cohort analyzed but pan-cancer cancer driver genes, nor considered most of the rules of our database of targeted therapies and the cohort analyzed was less comprehensive, with a smaller number of samples and cancer types analyzed.

In Rubio-Perez and Tamborero et al (2015) (Chapter 1) we developed an *in silico* drug prescription approach which linked targeted drugs to the driver genes altered in each patient-tumor sample considered. First, we identified the driver genes through their signals of positive selection in each cancer type. On detail, we used several methods following complementary criteria with the rationale that the combination of their results minimizes the number of false positives derived from each model[47]. Additionally, we classified the cancer driver genes according to their role in tumorigenesis in activating (i.e. oncogene) or loss of function (i.e. tumor suppressor)[167] (see Appendix 1). This identification of the mode of action of the driver genes was essential for exploring their therapeutic opportunities, since different targeting molecular mechanisms are associated to either type of driver genes[130,137,147]. Second, we compiled drugs able to interact with the driver gene products in distinct phases of their development, including therapies approved for their clinical use, therapies tested in clinical trials and ligands). We also included rules for: prescribing approved drugs according to their guidelines of use; considering resistance biomarkers co-occurring or not with other drug prescriptions in the same tumor; repurposing approved drugs for other prescriptions than the one approved; and for prescribing ligands, taking into account the ligand mechanism of action and the driver gene role. Third, we used the generated information on cancer drivers (named Drivers Database) and anti-cancer targeted therapies (named Drivers Actionability Database) to *in silico* prescribe treatments based on the genomic alterations of each sample, revealing a snapshot in time of the therapeutic landscape of cancer patients.

The landscape of driver genes revealed in our work, although thorough according to state-of-the-art methods, was still likely incomplete. On one hand, because of the low recurrence of some driver genes[53]. On the other hand, at the level of alteration type, neither epigenomic alterations nor non-coding elements were considered. Of note, recent advances pertaining the identification of driver non-coding elements, have recently expanded our results (see Appendix 5). Regarding the driver gene alterations with drug prescriptions, we are aware that not all alterations found in a driver gene are drivers[5,10]; thus, a prioritization of these alterations would refine the results and produce more accurate landscapes. We also consider that a better exploration of tumor clonality would refine the prescription of drugs to alterations found in tumor major clones which would exert a higher therapeutic benefit. Furthermore, in this analysis we have considered that more than one drug can be prescribed to a single patient (i.e. drug combinations). Combinatorial approaches have to be cautiously considered as drug combination toxicities, not addressed in the former work, have been described for targeted therapies combinations[132]. At last, we acknowledge that the incompleteness of the Drivers Actionability Database would have revealed an incomplete landscape, and that a more exhaustive manual curation of it would provide more accurate results.

Despite the presented limitations, that must be taken into consideration, our work produced a proof of principle strategy of comprehensively exploiting cancer genomic data to identify personalized medicine strategies; becoming one of the first comprehensive and integrative analyses of cancer driver genes and targeted therapies that has shed light into the molecular understanding of tumorigenesis and the scope and future perspectives of cancer genome-guided personalized medicine. As a snapshot in time we observed that very few cancer patients (5.9%) could benefit from approved therapies based on their tumor genomic alterations, but that this small fraction could be expanded up to a 40% when considering repurposing options and up to 73% when considering treatments undergoing clinical trials, not before estimated. Besides, we provided the cancer research community all the results generated, including the database of cancer driver genes and of anti-cancer therapies, as well as a prioritization list of 80 therapeutically unexploited driver genes with druggability features.

The evolution of the work described until now became the seed for other two of my projects (Chapter 2 and Chapter 3). On the one hand, observing that virtually all cancer patients (90%) bore at least an alteration of a driver gene, led us to explore the use of the Drivers Database on the design of informative sequencing cancer gene panels (Chapter 2). On the other, upon mounting evidence that not all mutations in driver genes are necessarily tumorigenic, we worked on overcoming the limitation of driver gene level analyses and moved to a better strategy to identify individual driver alterations in each single tumor. Besides, we thought that if we were able to overcome this limitation, an implementation into a tool of the driver alteration identification and the *in silico* drug prescription, refined at alteration level, would have potential broad applications from its use in pre-clinical to translational research (Chapter 3). Indeed, the *in silico* drug prescription strategy was useful not only for the work produced in our research group, but provided an extra value to other projects (see Appendix 2-4).

As already mentioned, alterations in the tumor genome may have an influence not only in drug response, but they can also inform about patient prognosis (e.g. different structural variants in chronic myeloid leukemia)[168,169] contribute to disease early diagnosis through liquid biopsies and be used as a way to monitor relapse also through liquid biopsies[170–172]. That is why profiling the tumor genome is becoming a standard tool in current clinical oncology. However, deciding the sequencing technique is not trivial. To identify a single predictive alteration, such as *BRAF* V600E mutation in a melanoma patient to prescribe vemurafenib[101], using Sanger sequencing could be enough. However, to enter a refractory patient into a clinical trial, which may have accumulated several relevant genomic alterations, or to investigate the tumor genome of a tumor cohort; sequencing few specific mutations with Sanger will not be enough. At that point, Next Generation Sequencing (NGS) techniques should be considered, but even among them a decision on whether sequencing the whole tumor genome, exome or only a set of genes and gene regions, by using a gene panel; must be taken. Of note, sequencing through gene panels possess a higher sensitivity and specificity in the variant detection step, when compared to whole exome sequencing[105]; which makes it effectively the most cost-effective option for both the translational research and clinical setting.

Several cancer gene panels are commercially available; relying most of these panels on manually gathered lists of genes or genomic regions decided at pan-cancer level. To design a panel for a specific question (e.g. identify the genomic markers of tumor relapse in a specific cancer type), a laborious search in the literature, extended to bioinformatic resources to estimate its cost-effectiveness (i.e. estimate the proportion of patients bearing the alteration in the gene or gene region in the disease of study, the panel coverage); needs to be carried out. Exploiting the resources generated in my previous work, discussed above, in Rubio-Perez et al (2016) we developed OncoPaD, the first tool aimed to the rational design of NGS sequencing mutational cancer gene panels (www.intogen.org/oncopad).

Through a user-friendly interface, OncoPaD suggests researchers sets of genes and/or gene regions to be included in a gene panel tailored for one or several cancer types, based on its cost-effectiveness. The genes suggested by OncoPaD either are: well-known cancer driver genes, bear mutations that are biomarkers of drug response, or have been identified as drivers via the detection of signals of positive selection across large tumor cohorts. Additionally, the user may decide to use its own list of genes. Next, OncoPaD estimates the cost-effectiveness of including each of the genes in the panel, on the basis of the selected cancer type(s), either considering all the exons in the gene, or only mutational hotspots (referred to as gene regions). From the cost-effectiveness estimation, the user obtains a prioritization of the genes and gene regions in three tiers. From 1 to 3 these include the genes and/or gene regions which increase the most the coverage (tier 1) up to those which do not increase the coverage at all (tier 3). Finally, OncoPaD results on the gene prioritization are shown to the researcher together with reports on the relevance of individual mutations for tumorigenesis or for anti-cancer treatment, supporting the interpretation of the generated results.

We acknowledge that before OncoPaD three approaches with similar aim than ours were already available[173–175] (see Chapter 3 Table 1 for the exhaustive comparison). However, they either make no previous selection of the gens based on driver gene identification[175], or consider only genes with high impacting mutations or frequently mutated[173,174], respectively. However, it is known that not all genes bearing high impacting mutations or frequently mutated are relevant for cancer development[10], leading to the inclusion of likely false positive candidates. OncoPaD, as well as the two latter approaches, have a common limitation: no considering other alteration type drivers than mutational (e.g. drivers bearing structural variants) and among mutational no considering non-coding alterations. The limitation to mutational coding drivers is inherited by OncoPaD from the Cancer Drivers database. However, as more comprehensive lists of driver genes bearing structural variants or non-coding mutations emerge, we will incorporate them to OncoPaD.

Nevertheless, even if OncoPaD is limited to the design of coding mutational cancer gene panels, we expect that it can become a useful and used tool in the cancer research and clinical community, because of the necessity of tumor genome sequencing, all the features included that minimize the inclusion of false positive candidates (either non driver genes and/or gene regions with very low coverage), the results reports generated and the outperformance, in terms of cost-effectivity, when compared to commercially available panels. Indeed, since we started the user tracking (from October 2016 until May 2017) we have registered around 800 accessions from 521 different users, showing that OncoPaD is used by the community, although more diffusion effort is needed.

Considering the opposite scenario, where the whole exome of the tumor is profiled instead of a gene panel, the interpretation of the obtained results may be more challenging, as a plethora of variants of unknown significance may be identified. The prioritization and interpretation of tumor somatic variants (i.e. the interpretation of the tumor genome), mostly in the context of exome sequencing, is still a non-resolved problem, being a bottleneck in the clinical and translational setting[104]. Once a tumor is sequenced most of the reported variants, even if located in cancer driver genes, are of uncertain significance and querying several scattered bioinformatic resources is needed to identify the variants driving the tumorigenesis. Moreover, once we identify the relevant tumor variants if we want to obtain information about its actionability we also need to go through different and scattered resources. Therefore, there is a necessity of developing new computational tools aimed to solve both hurdles, including the identification of driver variants among all variants found in a tumor and the identification of actionable variants[104,166,176].

To meet these necessities, we developed the Cancer Genome Interpreter (CGI) a web platform aimed to aid the interpretation of tumor genomes by contributing to solve the two hurdles (www.cancergenomeinterpreter.org) (Chapter 3). The specific aim of the CGI is, first, the identification of tumor variants more likely to drive the tumorigenesis, including those already validated as oncogenic and computational estimations of the effect of the remaining variants of unknown significance. And second, it also aims to identify the variants which shape the response to anti-cancer therapies (either response, resistance or toxicity), according to several levels of clinical evidence (either approved prescriptions, advanced clinical trials, early clinical trials, case reports or pre-clinical assays).

On detail, the CGI workflow starts with the set of alterations of a patient's tumor -either mutations, CNAs and/or chromosomal rearrangements- and the cancer type. The first step is the identification of the genes that putatively drive the tumorigenesis in the analyzed tumor. We based their identification on manually curated lists of cancer genes (e.g. Cancer Gene Census[50]) and catalogs of driver genes obtained from bioinformatic analyses of large tumors cohorts[177]. In the case of mutations, the CGI performs an additional step to evaluate each individual variant, since not all the mutations in cancer genes are equally relevant[10]. As in the first step, we used as basis *a priori* knowledge, by compiling mutations with a clinically or experimentally validated oncogenic effect, including cancer-predisposing germline variants. However, most of the variants observed in tumors are of unknown significance, and the estimation of their effect still relies on computational approaches. We did this using a novel tool, OncodriveMUT, which distinguishes from other methods with similar purpose, because it combines the mutation-centric measurements of the gene (or gene region) with the knowledge generated from the analyses of thousands of tumors. This provides statistically robust information that refines the evaluation of individual mutations.

Next, the CGI is aimed to identify which of these tumor alterations may shape the response to anti-cancer therapies. Scattered and unstructured information on the identification of genomic biomarkers which shape the response to anti-cancer therapies is continuously generated in clinical trials and/or pre-clinical assays, being the compilation and maintenance of this information a laborious task. We developed an expert curated resource, named Cancer Biomarkers database, as an extension of the Driver Actionability Database from Rubio-Perez and Tamborero et al (2015). Here, we increased the number and level of curation (through collaboration with oncology experts) of the genomic biomarkers. We added new types of biomarkers (i.e. toxicity, no response) and increased its degree of complexity (e.g. we added more alteration types such as biallelic inactivation, considered mutation consequence types or wild type variants). Moreover, we also added more multi-biomarker drug associations, even including biomarkers from different genomic types and stratified all the biomarkers according to the level of evidence of the biomarker-drug association, not only the drug status of approval (as done before). However, we acknowledge that such a manually curated database is costly to maintain. The mid-term maintenance of the Cancer Biomarkers database is supported by the collaborative H2020 MedBioinformatics project and we expect that its long-term maintenance will be supported by the Global Alliance for Health and Genomics (GA4H), which has the aim of unifying efforts such as the Cancer Biomarkers database and similar resources: CIViC[178], JAX-CKB[179], MyCancerGenome[130], OncoKB[180], PMKB[181] and PCT (https://pct.mdanderson.org). Of note, besides the Cancer Drivers Database we also developed a database containing the interactions of cancer driver genes with ligands (named Cancer Bioactivities database), with distinct levels of binding affinity. We suggest this resource as an interesting annotation for driver genes without biomarkers of drug response.

Beyond Cancer Biomarkers database maintenance, CGI inherits the limitations of Chapter 1, as it is based on the knowledge generated there. Additionally, we acknowledge that the assessment of the mutational signatures would be also of interest either for interpreting the biology of the tumor as well as for its therapeutic interpretation. Hence, we are planning to add this feature in next CGI updates.

Even if the pipeline for the integration of all the steps in CGI is complex. One of the main advantages of the CGI is the intuitiveness of its interface, including the visualization of all the variants identified in the tumor, the assessment of whether they are tumorigenic, and all annotations employed to classify them; which is not a trivial issue[176]. The actionable variants in tumors are stratified following levels of confidence and/or evidence, and presented to the user through interactive reports which help analyzing the results obtained. Several flexible input formats for the alterations are accepted, and an Application Programming Interface (API) has been developed to allow programmatic access. Additionally, we provide all resources supporting the CGI, including the Cancer Biomarkers Database and the catalog of driver genes and validated driver alterations, which may be of interest beyond its use in CGI. Taking all that into account, we think of the CGI as a versatile platform which automatizes highly laborious steps in the interpretation of cancer genomes. Due to CGI characteristics and our commitment of keeping it up to date with the evolving knowledge we expect that CGI will become a widely used tool either in the clinical, translational and basic research settings. Indeed, since October 2016 until May 2017, CGI website has had 2600 users and 7200 accessions, giving support to our expectations.

The CGI, however is not currently focused on assessing the extent of response to immunotherapies, since comparably much less is known about them, than about targeted therapies biomarkers. Due to the remarkable success of cancer immunotherapies, both T cell adoptive cell transfer in haematologic malignancies and immune checkpoint blockers in solid tumors[158], the translational cancer research community is recently shifting its focus towards the study of the tumor immune system interaction, to identify new immunotherapy strategies. The discovery of immune checkpoint molecules, among others, has shown that tumors have active mechanisms to resist the immune attack. However, even if some examples have been identified, there is an incomplete knowledge on the tumor mechanisms that modulate the action of the immune system.

Trying to contribute to fill up this gap, we evinced the last section of my thesis (Chapter 4), where we identified tumor pathways that become activated in correspondence with different patterns of cell populations in the immune infiltrate in the tumor. These pathways are candidates to tumor mechanisms through which the tumor may evade or counteract the activity of the immune infiltrate. First, we measured the degree of infiltration of sixteen different immune populations using sample-level gene set enrichment analysis of gene signatures representing each of immune population. To that end, we analyzed the bulk tumor RNA-seq data, following the rationale of previous works[115–117]. Gene set enrichment analysis is not the only approach used to identify immune populations, deconvolution methods have also been developed with the same aim[118,119] and both methods have been equally used in several works[116,120,182–184]. However, deconvolution methods are mostly machine learning approaches that have been trained on DNA microarray data, being unclear to which extent they can be applied to RNA-seq data[120]. That is the reason why we decided to use a gene set enrichment approach. Besides, an analysis carried out by us using one of these deconvolution methods, CIBERSORT, on the same cohort of tumors under study, demonstrated that the infiltration patterns of all immune cell populations could not properly reproduced. Giving strength to our decision of using a sample-level enrichment method.

After deciding for a sample-level enrichment method, we estimated the fraction of each cell population in the infiltrate of each tumor across the entire pan-cancer cohort through the enrichment of their representative gene signatures. We observed inter- and intra- tumor heterogeneity of immunological infiltrates. This finding, together with a comparison with data from normal donors, suggested that the differences in the immune infiltration patterns observed across cancer types cannot be explained solely by their tissue of origin. In turn, this observation further advocate that tumor intrinsic features may be responsible of the different infiltration patterns. To explore this hypothesis, we refined the identification of immune infiltration patterns at cancer type level and grouped the tumors of each cancer type in immune-clusters weighted by the effectivity of the immune system attack, as we hypothesized that the tumor mechanisms would be different depending on immune system effectivity, based on previous evidences[185,186]. After applying this approach to the 28 solid tumors, three different scenarios of cytotoxicity emerged across all cancer types, with different immune infiltration patterns. Suggesting that different immune cell compositions may lead to an equivalent level of cytotoxicity.

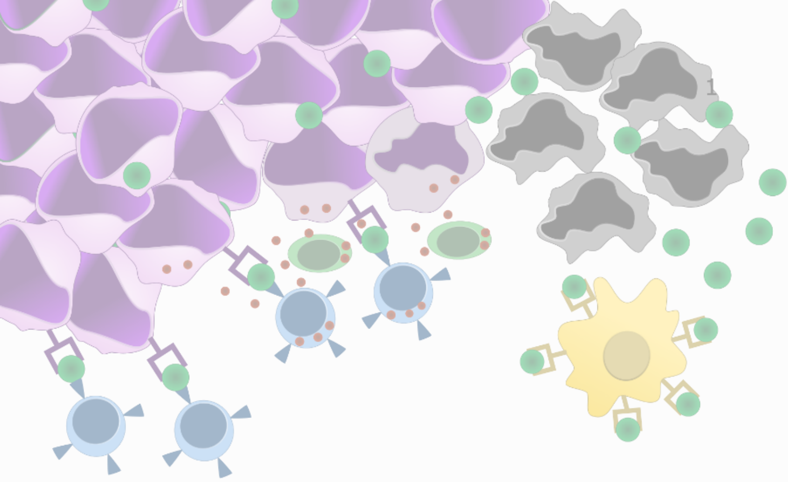
Next, we proceed to identify the tumor intrinsic features across these three scenarios. We first observed that, as already described[185], viral infections correlated with a higher cytotoxicity of the infiltrate in several cases, probably due to the increase of cell stress signaling and the expression of viral antigens[187]. On the other hand, we found that the group of tumors with lowest cytotoxic infiltrate was enriched for tumors of later stages, which may account for part of their increment in aggressiveness and worse prognosis. Next, we investigated the tumor pathways active in the different scenarios of cytotoxicity. Here we introduced a relevant methodological change with respect to other studies that made similar analyses[120,188]. We adjusted the expression levels of the tumor bulk samples for its immune component, following the rationale of Aran et al (2016)[113]. This adjustment revealed pathways that otherwise would be masked due to the contribution of the microenvironment to the bulk RNA. We checked that the failure to do this adjustment affects the results of enrichment analyses, as has already been proposed [112].

We found a heterogeneous enrichment of pathways across the three different scenarios that we integrated into a reasoned biological model. In the low cytotoxicity scenario, we identified an overexpression of high proliferative pathways identified in 14 of the cancer types (by the enrichment of cell cycle, DNA damage and protein synthesis pathways) which could be responsible of lowering the immunogenicity due to the generation of a large number of new non-recognized tumor antigens, as product of the newly emerging cells[189]. Besides, the phenotype of immune exclusion could be explained in some tumors by the activation of pathways which impair leukocyte recruitment (SHH, TGFb and Wnt-bCatenin signaling)[94,190,191]. The intermediate cytotoxic scenario appeared consistently enriched for only two pathways: angiogenesis, which may impair leukocyte trafficking and contribute to a decrease in the cytotoxicity by the recruitment of macrophages[192,193]; and ECM changes, that could either promote or suppress cytotoxicity and leukocyte recruitment[194,195]. In the high cytotoxic scenario, we found activation of processes leading to the cytotoxic phenotype (e.g. viral processes, high expression of HLA molecules and CGAs) and processes that presumably allowed tumor cells to survive on it (e.g. negative checkpoints, anti-inflammatory cytokines). Of note, we identified an enrichment of energy metabolism pathways in high cytotoxic tumors that may establish a competition by nutrients between tumors and immune effector cells that would impair its differentiation[196–198].

A limitation of this work is that we have not considered genetic and epigenetic tumor alterations. These may be positively selected due to the selective pressure of the immunologic microenvironment, providing the tumor capabilities of immune resistance by activation of the identified pathways or other mechanisms, such as the described mutation of *HLA* and *B2M* molecules, preventing from the recognition of T cells or the mutations in *CASP8* which avoid apoptosis induced by immune cells[185]. However, albeit incomplete, we have identified several biological processes which constitute potential good mechanisms for further research, and could be explored in the context of the combination of a targeted therapy and immunotherapies, as has already been suggested for Wnt-bCatenin pathway[199]. Besides, to our knowledge, this analysis is one of the most comprehensive landscapes on the tumor mechanisms related to immune evasion. Previous efforts are limited either because they focus in a single disease[115,120,182], or simplify the tumor molecular mechanisms to be analyzed[116,200] or the measure of immune infiltration[185].

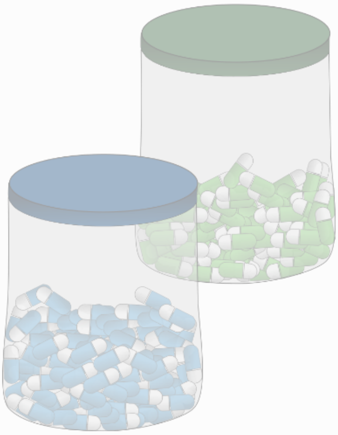
To sum up, I have acknowledged several limitations present across the thesis chapters. A common point is the lack of integration of epigenomic data. The consideration of epigenomic data could have increased the scope and comprehensiveness of the analyses carried out in Chapters 1 and 3, and it could explain some of the transcriptomic changes identified in Chapter 4. Non-coding mutations have not been considered either, mostly because to date there is no a comprehensive identification of non-coding driver alterations, even though efforts on this direction are ongoing (see Appendix 5). The incompleteness of the anti-cancer drug databases is also a limitation affecting more than one chapter (1, 2 and 3), and could decrease the comprehensiveness of the results obtained.

However, even if I acknowledge all these limitations, the work done in this thesis has been carried out methodically, comprehensively and integrative; generating knowledge and resources that contribute to the advance of cancer precision medicine. Chapter 1 has provided the cancer research community one of the first, if not the first most comprehensive, therapeutic landscape, shedding light into the scope of anti-cancer therapies in the most prevalent cancer types together with a list of good candidates targets for the design of new anti-cancer therapies. Chapter 2 has given the community a tool for the rational and cost-effective design of cancer gene panels, which may contribute to move a step forward the sequencing of new tumor cohorts either in the research and clinical field. Chapter 3 has given the community another tool that aids the interpretation of newly sequenced tumors, allowing interpret variants of unknown significance which could improve the patient handling, for example prioritizing patients for entering clinical trials, along with giving insights in the molecular mechanisms underlying newly sequenced tumor cohorts. At last, the tumor mechanisms identified in Chapter 4 shed light into a hot topic of current cancer research, the mechanisms of immune evasion by tumors, which can be potential targets to be explored for combinatorial therapies with immunotherapies.



PART V

CONCLUSIONS



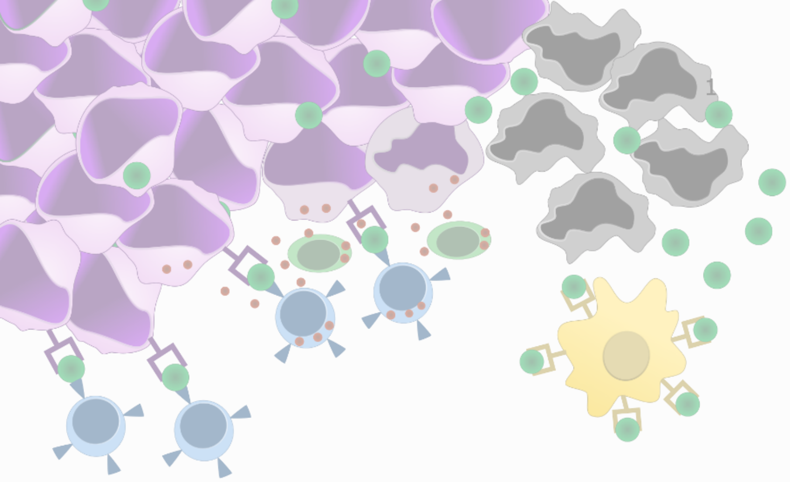
The work produced in this thesis, with limitations acknowledged in the discussion, and to the extent of the state-of-the art of cancer genomics research, has made several contributions towards the advance of cancer precision medicine.

Together with Tamborero D. we have identified the therapeutic landscape of anti-cancer therapies, based on cancer patient genomic alterations in driver genes (mutational and with chromosomal rearrangements), as a snapshot in time. It has provided information on the extent of targeted therapies and their potential future progress (through the analysis of treatments in clinical trials) across 28 prevalent cancer types. Besides, we have identified a list of potential good targets for anti-cancer drug design as well as several drug repurposing opportunities.

I have developed a tool for the rational design of cancer NGS mutational panels, that works at cancer type level or in groups thereof. OncoPaD maximizes the coverage of tumors in a cohort that a panel can achieve and minimizes the amount of DNA to be sequenced to obtain that result. Additionally, it provides the user ancillary annotations (such as which genes have biomarkers of drug response) that helps to decide which candidates include in the panel. OncoPaD is open source and freely available at www.intogen.org/oncopad.

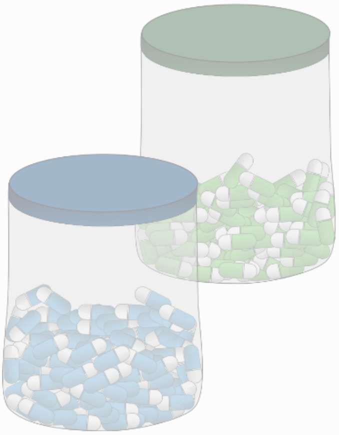
Complementary to OncoPaD, I contributed to the development of the Cancer Genome Interpreter, a tool for guiding the interpretation of newly sequenced tumors, to identify which of the alterations observed in a tumor are oncogenic and which may inform a therapeutic benefit. The Cancer Genome Interpreter has been developed along with a database of drug response biomarkers and cancer target ligands, where I contributed the most. The Cancer Genome interpreter is freely available at www.cancergenomeinterpreter.org

Finally, I focused on the study of tumor mechanisms of immune evasion. In this part, through the analysis of tumor RNA-seq bulk data, together with Tamborero D., we identified the patterns of infiltration of sixteen immune cell populations. Next, we grouped the infiltrating immune populations in clusters reflecting their cytotoxicity. We then performed ad in-depth study of the clusters and identified clinical correlates and tumor active pathways involved in the evasion of the immune system.



PART VI

APPENDIX



In this section, I cite and attach other publications where I contributed, thanks to the knowledge acquired and produced during my thesis.

Appendix 1

In Schroeder et al. (2014) we developed OncodriveROLE, a machine learning approach that classifies genes according to their role in tumorigenesis, either Activating or Loss of function. The classifier uses the distribution of genomic alterations in the genes (mutations and/or copy number alterations) to classify them. We achieved a 0.93 accuracy when applying the classifier to Cancer Gene Census gene list and a Matthew Correlation Coefficient of 0.84. The classifier is available at http://bg.upf.edu/oncodrive-role .

Here, I contributed in the exploration of the machine learning approaches which could be used for building the classifier and in the selection and generation of the genomic attributes to classify the driver genes.

Schroeder MP, **Rubio-Perez C**, Tamborero D, Gonzalez-Perez A, Lopez-Bigas N. OncodriveROLE classifies cancer driver genes in loss of function and activating mode of action. ***Bioinformatics.*** 2014 Sep 1;30(17):i549-55.

Briefly, in Biton et al. (2014) they analyse the transcriptome of bladder cancer and identify bladder-specific biological components. They also characterized bladder subtypes (luminal, basal-like and muscle-invasive). The study of the urothelial differentiation in luminal bladder carcinomas revealed a pro-tumorigenic role of *PPARG* in these tumors.

In this publication, I contributed, together with my supervisor N. Lopez-Bigas in the discussion of *PPARG* therapeutic implications.

Biton A, Bernard-Pierrot I, Lou Y, Krucker C, Chapeaublanc E, **Rubio-Pérez C,** López-Bigas N, Kamoun A, Neuzillet Y, Gestraud P, Grieco L, Rebouissou S, de Reyniès A, Benhamou S, Lebret T, Southgate J, Barillot E, Allory Y, Zinovyev A, Radvanyi F. Independent Component Analysis Uncovers the Landscape of the Bladder Tumor Transcriptome and Reveals Insights into Luminal and Basal Subtypes. ***Cell Rep.*** 2014 Nov 20;9(4):1235-45.

In brief, Puente et al. (2016) did a comprehensive identification of the genomic driver alterations, coding and non-coding, in a cohort of 452 chronic lymphocytic leukemia (CLL) cases and 54 with monoclonal B-lymphocytosis, a stage previous to CLL. They identified novel recurrent genomic alterations in the disease of study, such as *NOTCH1* 3' alterations.

In this project, I explored the therapeutic implications of the identified driver alterations in CLL patients (Figure S6, Table S9). Tamborero D. helped me in the curation of the drug response biomarkers.

Puente XS, Beà S, Valdés-Mas R, Villamor N, Gutiérrez-Abril J, Martín-Subero JI, Munar M, **Rubio-Pérez C**, Jares P, Aymerich M, Baumann T, Beekman R, Belver L, Carrio A, Castellano G, Clot G, Colado E, Colomer D, Costa D, Delgado J, Enjuanes A, Estivill X, Ferrando AA, Gelpí JL, González B, González S, González M, Gut M, Hernández-Rivas JM, López-Guerra M, Martín-García D, Navarro A, Nicolás P, Orozco M, Payer ÁR, Pinyol M, Pisano DG, Puente DA, Queirós AC, Quesada V, Romeo-Casabona CM, Royo C, Royo R, Rozman M, Russiñol N, Salaverría I, Stamatopoulos K, Stunnenberg HG, Tamborero D, Terol MJ, Valencia A, López-Bigas N, Torrents D, Gut I, López-Guillermo A, López-Otín C, Campo E. Non-coding recurrent mutations in chronic lymphocytic leukaemia. ***Nature***. 2015 Oct 22;526(7574):519-24

Shortly, Karube et al., (Submitted) genomically characterized a large cohort of diffuse large B-cell lymphoma (DLBCL). They found that germinal center B-cell and activated B-cell DLBCL had a differential profile of mutations, altered pathogenic pathways and CNA; recognizing potential targets for new intervention strategies

In this project I have applied the *in silico prescription strategy*, exploring the therapeutic options, including drug repurposing opportunities, of the DLBCL cohort analyzed (Figure 5).

Karube K, Enjuanes A, Dlouhy I, Jares P, MartinGarcia D, Nadeu F, Ordóñez GR, Rovira J, Clot G, Royo C, Navarro A, Gonzalez-Farre B, Vaghefi A, Castellano G, **Rubio-Perez C**, Tamborero D, Briones J, Salar A, Sancho JM, Mercadal S, Gonzalez-Barca E, Escoda L, Miyoshi H, Ohshima K, Miyawaki K, Kato K, Akashi K, Mozos A, Colomo L, Alcoceba M, Valera A, Carrió A, Costa D, Lopez-Bigas N, Schmitz R, Staudt LM, Salaverria I, LópezGuillermo A, Campo E. Integrating genomic alterations in diffuse large B-cell lymphoma identifies new relevant pathways and potential therapeutic targets**.** (*Submitted*)

Briefly, Sabarinathan and Pich et al. (*submitted*) did a comprehensive analysis of the driver alterations (mutations and chromosomal rearrangements) that contributed to tumorigenesis in a cohort of 2583 whole-genome sequenced tumors from 37 different cancer types. They find a genomic driver alteration in more than 90% of the patients, proving that cancer is driven by genetic events. Besides, they observed that the average of driver events per patient (around 4.6) was stable across tumors even if they showed huge differences of mutational burden.

In this publication, I explored the therapeutic landscape of the tumor cohort, including either driver coding and non-coding alterations (Figure 6).

Radhakrishnan Sabarinathan*, Oriol Pich*, Iñigo Martincorena, **Carlota Rubio-Perez**, Malene Juul Rasmussen, Jeremiah Wala, Steven Schumacher, Ofer Shapira, Nikos Sidiropoulos, Sebastian Waszak, David Tamborero, Loris Mularoni, Esther Rheinbay, Henrik Hornshøj, Jordi Deu-Pons, Ferran Muiños, Johanna Bertl, Qianyun Guo, PCAWG-2,5,9,14, Joachim Weischenfeldt, Jan Korbel, Gad Getz, Peter Campbell, Jakob Skou Pedersen, Rameen Beroukhim, Abel Gonzalez-Perez, Núria López-Bigas. The whole-genome panorama of cancer drivers. (*Submitted*)

* Co-first author

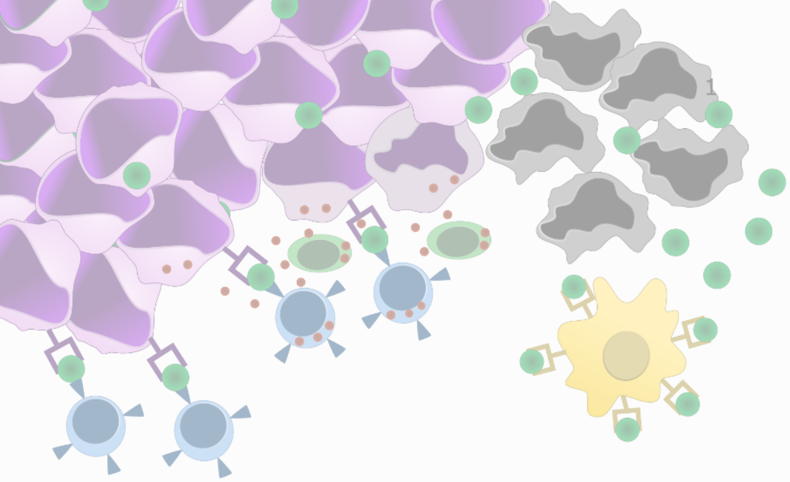
Schroeder MP, Rubio-Perez C, Tamborero D, Gonzalez-Perez A, Lopez-Bigas N. OncodriveROLE classifies cancer driver genes in loss of function and activating mode of action. Bioinformatics. 2014 Sep 1;30(17):i549-55. DOI: 10.1093/bioinformatics/btu467

Biton A, Bernard-Pierrot I, Lou Y, Krucker C, Chapeaublanc E, Rubio-Pérez C, et al. Independent component analysis uncovers the landscape of the bladder tumor transcriptome and reveals insights into luminal and basal subtypes. Cell Rep. 2014 Nov 20;9(4):1235–45. DOI: 10.1016/j.celrep.2014.10.035

Puente XS, Beà S, Valdés-Mas R, Villamor N, Gutiérrez-Abril J, Martín-Subero JI, et al. Non-coding recurrent mutations in chronic lymphocytic leukaemia. Nature. 2015 Oct 22;526(7574):519–24. DOI: 10.1038/nature14666

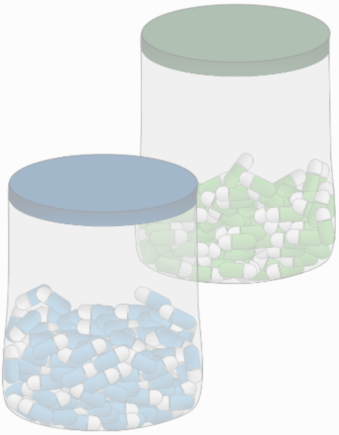
Karube K, Enjuanes A, Dlouhy I, Jares P, Martin-Garcia D, Nadeu F, et al. Integrating genomic alterations in diffuse large B-cell lymphoma identifies new relevant pathways and potential therapeutic targets. Leukemia. 2018 Mar 14;32(3):675–84. DOI: 10.1038/leu.2017.251

Sabarinathan R, Pich O, Martincorena I, Rubio-Perez C, Juul M, Wala J, et al. The whole-genome panorama of cancer drivers. bioRxiv. 2017 Sep 20;190330. DOI: 10.1101/190330



PART VII

BIBLIOGRAPHY



1. Mukherjee, S. *The Emperor of All Maladies: A Biography of Cancer*. (Paperback, 2011).
2. WHO. *International Classification of Diseases for Oncology*. (WHO, 2013).
3. Ferlay, J. *et al.* GLOBOCAN 2012 v1.0, Cancer Incidence and Mortality Worldwide: IARC CancerBase No. 11. *Lyon, France: International Agency for Research on Cancer; 2013*
4. Morin, P., Vogelstein, B., Trent, J. M. & Collins, F. S. Chapter 83. Cancer Genetics. in *Harrison's Principles of Internal Medicine* (Mc Graw Hill, 2013).
5. Stratton, M. R., Campbell, P. J. & Andrew F, P. The cancer genome. *Nature* **458,** 719–724 (2009).
6. Alberts, B. *et al.* DNA Repair. (2002).
7. Raphael, B. J. *et al.* Identifying driver mutations in sequenced cancer genomes: computational approaches to enable precision medicine. *Genome Med.* **6,** 5 (2014).
8. Mermel, C. H. *et al.* GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol.* **12,** R41 (2011).
9. Lodish, H. F. *Molecular cell biology*. (W.H. Freeman, 2000).
10. Vogelstein, B. *et al.* Cancer genome landscapes. *Science* **339,** 1546–58 (2013).
11. Eilbeck, K. & Lewis, S. E. Sequence ontology annotation guide. *Comp. Funct. Genomics* **5,** 642–7 (2004).
12. Ward, A. J. & Cooper, T. A. The pathobiology of splicing. *J. Pathol.* **220,** 152–63 (2010).
13. Mullaney, J. M., Mills, R. E., Pittard, W. S. & Devine, S. E. Small insertions and deletions (INDELs) in human genomes. *Hum. Mol. Genet.* **19,** R131-6 (2010).
14. Huang, F. W. *et al.* Highly Recurrent TERT Promoter Mutations in Human Melanoma. *Science (80-.).* **339,** 957–959 (2013).
15. Horn, S. *et al.* TERT promoter mutations in familial and sporadic melanoma. *Science* **339,** 959–61 (2013).

16. Rahman, S. *et al.* Activation of the LMO2 Oncogene in T-ALL through a Somatically Acquired Neomorphic Promoter. *Blood* **128,** (2016).

17. Freeman, J. L. *et al.* Copy number variation: New insights in genome diversity. *Genome Res.* **16,** 949–961 (2006).

18. Hasty, P. & Montagna, C. Chromosomal Rearrangements in Cancer: Detection and potential causal mechanisms. *Mol. Cell. Oncol.* **1,** (2014).

19. Forment, J. V., Kaidi, A. & Jackson, S. P. Chromothripsis and cancer: causes and consequences of chromosome shattering. *Nat. Rev. Cancer* **12,** 663–670 (2012).

20. NOWELL, P. C. The minute chromosome (Phl) in chronic granulocytic leukemia. *Blut* **8,** 65–6 (1962).

21. Kakizuka, A. *et al.* Chromosomal translocation t(15;17) in human acute promyelocytic leukemia fuses RAR alpha with a novel putative transcription factor, PML. *Cell* **66,** 663–74 (1991).

22. Kwak, E. L. *et al.* Anaplastic Lymphoma Kinase Inhibition in Non?Small-Cell Lung Cancer. *N. Engl. J. Med.* **363,** 1693–1703 (2010).

23. Koivunen, J. P. *et al.* EML4-ALK Fusion Gene and Efficacy of an ALK Kinase Inhibitor in Lung Cancer. *Clin. Cancer Res.* **14,** 4275–4283 (2008).

24. Soda, M. *et al.* Identification of the transforming EML4-ALK fusion gene in non-small-cell lung cancer. *Nature* **448,** 561–6 (2007).

25. Takeuchi, K. *et al.* RET, ROS1 and ALK fusions in lung cancer. *Nat. Med.* **18,** 378–81 (2012).

26. Alaei-Mahabadi, B., Bhadury, J., Karlsson, J. W., Nilsson, J. A. & Larsson, E. Global analysis of somatic structural genomic alterations and their impact on gene expression in diverse human cancers. *Proc. Natl. Acad. Sci. U. S. A.* **113,** 13768–13773 (2016).

27. Cooper, G. *The Cell: a molecular approach*. (2000).

28. Ryland, G. L. *et al.* Loss of heterozygosity: what is it good for? *BMC Med. Genomics* **8,** 45 (2015).

29. Gnad, F., Doll, S., Manning, G., Arnott, D. & Zhang, Z. Bioinformatics analysis of thousands of TCGA tumors to determine the involvement of epigenetic regulators in human cancer. *BMC Genomics* **16,** S5 (2015).

30. Peng, L. *et al.* Large-scale RNA-Seq Transcriptome Analysis of 4043 Cancers and 548 Normal Tissue Controls across 12 TCGA Cancer Types. *Sci. Rep.* **5,** 13413 (2015).

31. The GTEx Consortium *et al.* The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science (80-.).* **348,** 648–60 (2015).

32. Hoadley, K. A. *et al.* Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. *Cell* **158,** 929–44 (2014).

33. Bird, A. DNA methylation patterns and epigenetic memory. *Genes Dev.* **16,** 6–21 (2002).

34. Kung, J. T. Y., Colognori, D. & Lee, J. T. Long Noncoding RNAs: Past, Present, and Future. *Genetics* **193,** 651–669 (2013).

35. Koonin, E. V. Darwinian evolution in the light of genomics. *Nucleic Acids Res.* **37,** 1011 (2009).

36. Wei Dai, Y. Y. Genomic Instability and Cancer. *J. Carcinog. Mutagen.* **5,** (2014).

37. Hanahan, D. & Weinberg, R. A. Hallmarks of Cancer: The Next Generation. *Cell* **144,** 646–674 (2011).

38. Sherr, C. J. & McCormick, F. The RB and p53 pathways in cancer. *Cancer Cell* **2,** 103–112 (2002).

39. Rude Voldborg, B., Damstrup, L., Spang-Thomsen, M. & Skovgaard Poulsen, H. Epidermal growth factor receptor (EGFR) and EGFR mutations, function and possible role in clinical trials. *Ann. Oncol.* **8,** 1197–1206 (1997).

40. Ferreira, L. M. R. Cancer metabolism: The Warburg effect today. *Exp. Mol. Pathol.* **89,** 372–380 (2010).

41. Kang, M. H. & Reynolds, C. P. Bcl-2 Inhibitors: Targeting Mitochondrial Apoptotic Pathways in Cancer Therapy. *Clin. Cancer*

Res. **15,** (2009).

42. Gudmundsdottir, K. & Ashworth, A. The roles of BRCA1 and BRCA2 and associated proteins in the maintenance of genomic stability. *Oncogene* **25,** 5864–5874 (2006).
43. Ferrara, N., Gerber, H.-P. & LeCouter, J. The biology of VEGF and its receptors. *Nat. Med.* **9,** 669–676 (2003).
44. Sundquist, E. *et al.* Neoplastic extracellular matrix environment promotes cancer invasion in vitro. *Exp. Cell Res.* **344,** 229–240 (2016).
45. Kim, N. W. *et al.* Specific Association of Human Telomerase Activity with Immortal Cells and Cancer. *Science (80-.).* **266,** 2011–2015
46. Freeman, G. J. *et al.* Engagement of the PD-1 immunoinhibitory receptor by a novel B7 family member leads to negative regulation of lymphocyte activation. *J. Exp. Med.* **192,** 1027–34 (2000).
47. Tamborero, D. *et al.* Comprehensive identification of mutational cancer driver genes across 12 tumor types. *Sci. Rep.* **3,** 2650 (2013).
48. Reddy, E. P., Reynolds, R. K., Santos, E. & Barbacid, M. A point mutation is responsible for the acquisition of transforming properties by the T24 human bladder carcinoma oncogene. *Nature* **300,** 149–152 (1982).
49. Tabin, C. J. *et al.* Mechanism of activation of a human oncogene. *Nature* **300,** 143–149 (1982).
50. Futreal, P. A. *et al.* A census of human cancer genes. *Nat. Rev. Cancer* **4,** 177–83 (2004).
51. The Cancer Genome Atlas Research Network. TCGA. Available at: https://cancergenome.nih.gov/.
52. (ICGC), I. C. G. C. ICGC. Available at: http://icgc.org/.
53. Garraway, L. A. & Lander, E. S. Lessons from the Cancer Genome. *Cell* **153,** 17–37 (2013).
54. Tokheim, C. J., Papadopoulos, N., Kinzler, K. W., Vogelstein, B. & Karchin, R. Evaluating the evaluation of cancer driver genes. *Proc.*

Natl. Acad. Sci. **113,** 14330–14335 (2016).

55. Parmigiani, G. *et al.* Design and analysis issues in genome-wide somatic mutation studies of cancer. *Genomics* **93,** 17 (2009).
56. Sjoblom, T. *et al.* The Consensus Coding Sequences of Human Breast and Colorectal Cancers. *Science (80-.).* **314,** 268–274 (2006).
57. Lawrence, M. S. *et al.* Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* **505,** 495–501 (2014).
58. Dees, N. D. *et al.* MuSiC: Identifying mutational significance in cancer genomes. *Genome Res.* **22,** 1589–1598 (2012).
59. Davoli, T. *et al.* Cumulative Haploinsufficiency and Triplosensitivity Drive Aneuploidy Patterns and Shape the Cancer Genome. *Cell* **155,** 948–962 (2013).
60. Tamborero, D., Gonzalez-Perez, A. & Lopez-Bigas, N. OncodriveCLUST: exploiting the positional clustering of somatic mutations to identify cancer genes. **29,** 2238–2244 (2013).
61. Kamburov, A. *et al.* Comprehensive assessment of cancer missense mutation clustering in protein structures. *Proc. Natl. Acad. Sci. U. S. A.* **112,** E5486 (2015).
62. Gonzalez-Perez, A. & Lopez-Bigas, N. Functional impact bias reveals cancer drivers. *Nucleic Acids Res.* **40,** e169–e169 (2012).
63. Mularoni, L., Sabarinathan, R., Deu-Pons, J., Gonzalez-Perez, A. & L?pez-Bigas, N. OncodriveFML: a general framework to identify coding and non-coding regions with cancer driver mutations. *Genome Biol.* **17,** 128 (2016).
64. Kumar, P., Henikoff, S. & Ng, P. C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat. Protoc.* **4,** 1073–1081 (2009).
65. Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. *Nat. Methods* **7,** 248 (2010).
66. Reva, B., Antipin, Y. & Sander, C. Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Res.* **39,** e118–e118 (2011).

67. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46,** 310–315 (2014).

68. Reimand, J. & Bader, G. D. Systematic analysis of somatic mutations in phosphorylation signaling predicts novel cancer drivers. *Mol. Syst. Biol.* **9,** 637 (2013).

69. Porta-Pardo, E. & Godzik, A. e-Driver: a novel method to identify protein regions driving cancer. *Bioinformatics* **30,** 3109–3114 (2014).

70. Hofree, M. *et al.* Challenges in identifying cancer genes by analysis of exome sequencing data. *Nat. Commun.* **7,** (2016).

71. Beroukhim, R. *et al.* Assessing the significance of chromosomal aberrations in cancer: Methodology and application to glioma. *Proc. Natl. Acad. Sci.* **104,** 20007–20012 (2007).

72. Tamborero, D., Lopez-Bigas, N., Gonzalez-Perez, A., Down, T. & Hubbard, T. Oncodrive-CIS: A Method to Reveal Likely Driver Genes Based on the Impact of Their Copy Number Changes on Expression. *PLoS One* **8,** e55489 (2013).

73. Karlsson, J. *et al.* FocalScan: Scanning for altered genes in cancer based on coordinated DNA and RNA change. *Nucleic Acids Res.* **512,** gkw674 (2016).

74. Piazza, R. *et al.* FusionAnalyser: a new graphical, event-driven tool for fusion rearrangements discovery. *Nucleic Acids Res.* **40,** e123–e123 (2012).

75. Benelli, M. *et al.* Discovering chimeric transcripts in paired-end RNA-seq data by using EricScript. *Bioinformatics* **28,** 3232–3239 (2012).

76. Kim, D. & Salzberg, S. L. TopHat-Fusion: an algorithm for discovery of novel fusion transcripts. *Genome Biol.* **12,** R72 (2011).

77. Yates, L. R. and C. Evolution of the cancer genome. *Trends Genet.* **28,** 155–163 (2012).

78. Cancer Genome Atlas Research Network, {fname}. Integrated genomic analyses of ovarian carcinoma. *Nature* **474,** 609–15

(2011).

79. Network, T. C. G. A. Genomic Classification of Cutaneous Melanoma. *Cell* **161,** 1681 (2015).

80. Lawrence, M. S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499,** 214–8 (2013).

81. Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500,** 415–21 (2013).

82. Nowell, P. The clonal evolution of tumor cell populations. *Science (80-.).* **194,** (1976).

83. Venkatesan, S. & Swanton, C. Tumor Evolutionary Principles: How Intratumor Heterogeneity Influences Cancer Treatment and Outcome. *Am. Soc. Clin. Oncol. Educ. B.* **36,** e141–e149 (2016).

84. Pao, W. *et al.* Acquired Resistance of Lung Adenocarcinomas to Gefitinib or Erlotinib Is Associated with a Second Mutation in the EGFR Kinase Domain. *PLoS Med.* **2,** e73 (2005).

85. Soverini, S. *et al.* BCR-ABL kinase domain mutation analysis in chronic myeloid leukemia patients treated with tyrosine kinase inhibitors: recommendations from an expert panel on behalf of European LeukemiaNet. *Blood* **118,** 1208–1215 (2011).

86. Misale, S. *et al.* Emergence of KRAS mutations and acquired resistance to anti-EGFR therapy in colorectal cancer. *Nature* **486,** 532–6 (2012).

87. Quail, D. F. & Joyce, J. A. Microenvironmental regulation of tumor progression and metastasis. *Nat. Med.* **19,** 1423–37 (2013).

88. Schreiber, R. D., Old, L. J. & Smyth, M. J. Cancer Immunoediting: Integrating Immunity's Roles in Cancer Suppression and Promotion. *Science (80-.).* **331,** 1565–1570 (2011).

89. Mellman, I., Coukos, G. & Dranoff, G. Cancer immunotherapy comes of age. *Nature* **480,** 480–489 (2011).

90. Chen, D. S. & Mellman, I. Oncology MeetsImmunology:TheCancer-ImmunityCycle. 1–10 (2013). doi:10.1016/j.immuni.2013.07.012

91. Motz, G. & Coukos, G. Deciphering and Reversing Tumor Immune Suppression. *Immunity* **39,** 61–73 (2013).

92. Gajewski, T. F., Schreiber, H. & Fu, Y.-X. Innate and adaptive immune cells in the tumor microenvironment. *Nat. Immunol.* **14,** 1014–22 (2013).

93. Zaretsky, J. M. *et al.* Mutations Associated with Acquired Resistance to PD-1 Blockade in Melanoma. *N. Engl. J. Med.* **375,** 819–829 (2016).

94. Spranger, S. & Gajewski, T. F. A new paradigm for tumor immune escape: β-catenin-driven immune exclusion. *J. Immunother. cancer* **3,** 43 (2015).

95. Bates, S. E. Classical Cytogenetics: Karyotyping Techniques. in 177–190 (2011). doi:10.1007/978-1-61779-201-4_13

96. T?nnies, H. Modern molecular cytogenetic techniques in genetic diagnostics. *Trends Mol. Med.* **8,** 246–250 (2002).

97. FDA. *Gleevec Prescribing Information*. (2008).

98. Kallioniemi, A. *et al.* Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. *Science* **258,** 818–21 (1992).

99. Keren, B. The advantages of SNP arrays over CGH arrays. *Mol. Cytogenet.* **7,** I31 (2014).

100. Hagemann, I. S. Chapter 1 – Overview of Technical Aspects and Chemistries of Next-Generation Sequencing. in *Clinical Genomics* 3–19 (2015). doi:10.1016/B978-0-12-404748-8.00001-0

101. FDA. *Zelboraf Prescribing Information*. (2011).

102. Hoy, M. A. & Hoy, M. A. Chapter 7 – DNA Sequencing and the Evolution of the '-Omics'. in *Insect Molecular Genetics* 251–305 (2013). doi:10.1016/B978-0-12-415874-0.00007-X

103. Su, Z. *et al.* Next-generation sequencing and its applications in molecular diagnostics. *Expert Rev. Mol. Diagn.* **11,** 333–43 (2011).

104. Good, B. M., Ainscough, B. J., McMichael, J. F., Su, A. I. & Griffith, O. L. Organizing knowledge to enable personalization of medicine in cancer. *Genome Biol.* **15,** 1–9 (2014).

105. Jones, S. *et al.* Personalized genomic analyses for cancer mutation discovery and interpretation. *Sci. Transl. Med.* **7,** 283ra53 (2015).

106. Bumgarner, R. Overview of DNA microarrays: types, applications, and their future. *Curr. Protoc. Mol. Biol.* **Chapter 22,** Unit 22.1. (2013).

107. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **10,** 57–63 (2009).

108. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* **5,** 621–628 (2008).

109. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28,** 511–515 (2010).

110. Wagner, G. P., Kin, K. & Lynch, V. J. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory Biosci.* **131,** 281–285 (2012).

111. Joyce, J. A. & Pollard, J. W. Microenvironmental regulation of metastasis. *Nat. Rev. Cancer* **9,** 239–252 (2009).

112. Aran, D., Sirota, M., Butte, A. J., Modrusan, Z. & Clark, H. F. Systematic pan-cancer analysis of tumour purity. *Nat. Commun.* **6,** 8971 (2015).

113. Aran, D. *et al.* Widespread parainflammation in human cancer. *Genome Biol.* **17,** 145 (2016).

114. Yoshihara, K. *et al.* Inferring tumour purity and stromal and immune cell admixture from expression data. *Nat. Commun.* **4,** 2612 (2013).

115. Angelova, M. *et al.* Characterization of the immunophenotypes and antigenomes of colorectal cancers reveals distinct tumor escape mechanisms and novel targets for immunotherapy. *Genome Biol.* **16,** 64 (2015).

116. Charoentong, P. *et al.* Pan-cancer Immunogenomic Analyses Reveal Genotype-Immunophenotype Relationships and Predictors

of Response to Checkpoint Blockade. *Cell Rep.* **18,** 248–262 (2017).

117. Bindea, G. *et al.* Spatiotemporal Dynamics of Intratumoral Immune Cells Reveal the Immune Landscape in Human Cancer. *Immunity* **39,** 782–795 (2013).

118. Becht, E. *et al.* Estimating the?population abundance of tissue-infiltrating immune and stromal cell populations using gene expression. *Genome Biol.* **17,** 218 (2016).

119. Newman, A. M. *et al.* Robust enumeration of cell subsets from tissue expression profiles. *Nat. Methods* **12,** 453–457 (2015).

120. Şenbabaoğlu, Y. *et al.* Tumor immune microenvironment characterization in clear cell renal cell carcinoma identifies prognostic and immunotherapeutically relevant messenger RNA signatures. *Genome Biol.* **17,** 231 (2016).

121. Fisher, B. *et al.* Tamoxifen for prevention of breast cancer: report of the National Surgical Adjuvant Breast and Bowel Project P-1 Study. *J. Natl. Cancer Inst.* **90,** 1371–88 (1998).

122. Garraway, L. A. Genomics-driven oncology: framework for an emerging paradigm. *J. Clin. Oncol.* **31,** 1806–1814 (2013).

123. Weinstein, I. B. CANCER: Enhanced: Addiction to Oncogenes--the Achilles Heal of Cancer. *Science (80-.).* **297,** 63–64 (2002).

124. Weinstein, I. B. & Joe, A. K. Mechanisms of Disease: oncogene addiction?a rationale for molecular targeting in cancer therapy. *Nat. Clin. Pract. Oncol.* **3,** 448–457 (2006).

125. FDA. *Herceptin Perscribing Information*. (2010).

126. Weisberg, E. *et al.* AMN107 (nilotinib): a novel and selective inhibitor of BCR-ABL. *Br. J. Cancer* **94,** 1765–9 (2006).

127. Lynch, T. J. *et al.* Activating Mutations in the Epidermal Growth Factor Receptor Underlying Responsiveness of Non?Small-Cell Lung Cancer to Gefitinib. *N. Engl. J. Med.* **350,** 2129–2139 (2004).

128. Paez, J. G. *et al.* EGFR mutations in lung cancer: correlation with clinical response to gefitinib therapy. *Science* **304,** 1497–500 (2004).

129. Pao, W. *et al.* EGF receptor gene mutations are common in lung cancers from &quot;never smokers&quot; and are associated with sensitivity of tumors to gefitinib and erlotinib. *Proc. Natl. Acad. Sci.* **101,** 13306–13311 (2004).

130. Abramson, R. Overview of Targeted Therapies for Cancer - My Cancer Genome. Available at: https://www.mycancergenome.org/content/molecular-medicine/overview-of-targeted-therapies-for-cancer/. (Accessed: 26th June 2017)

131. Johnson, D. B. *et al.* Combined BRAF (Dabrafenib) and MEK inhibition (Trametinib) in patients with BRAFV600-mutant melanoma experiencing progression with single-agent BRAF inhibitor. *J. Clin. Oncol.* **32,** 3697–704 (2014).

132. Soria, J.-C., Massard, C. & Izzedine, H. From theoretical synergy to clinical supra-additive toxicity. *J. Clin. Oncol.* **27,** 1359–61 (2009).

133. Strittmatter, S. M. Overcoming Drug Development Bottlenecks With Repurposing: Old drugs learn new tricks. *Nat. Med.* **20,** 590–591 (2014).

134. DiMasi, J. A., Feldman, L., Seckler, A. & Wilson, A. Trends in Risks Associated With New Drug Development: Success Rates for Investigational Drugs. *Clin. Pharmacol. Ther.* **87,** 272–277 (2010).

135. Arrowsmith, J. & Miller, P. Trial Watch: Phase II and Phase III attrition rates 2011–2012. *Nat. Rev. Drug Discov.* **12,** 569–569 (2013).

136. Redig, A. J. & Jänne, P. A. Basket trials and the evolution of clinical trial design in an era of genomic medicine. *J. Clin. Oncol.* **33,** 975–7 (2015).

137. Lopez-Chavez, A. *et al.* Molecular profiling and targeted therapy for advanced thoracic malignancies: a biomarker-derived, multiarm, multihistology phase II basket trial. *J. Clin. Oncol.* **33,** 1000–7 (2015).

138. Holohan, C., Van Schaeybroeck, S., Longley, D. B. & Johnston, P.

G. Cancer drug resistance: an evolving paradigm. *Nat. Rev. Cancer* **13,** 714–26 (2013).

139. Szakacs, G. *et al.* Predicting drug sensitivity and resistance: profiling ABC transporter genes in cancer cells. *Cancer Cell* **6,** 129–137 (2004).

140. Nurwidya, F., Murakami, A., Takahashi, F. & Takahashi, K. Molecular mechanisms contributing to resistance to tyrosine kinase-targeted therapy for non-small cell lung cancer. *Cancer Biol. Med.* **9,** 18–22 (2012).

141. Zhang, J., Yang, P. L. & Gray, N. S. Targeting cancer with small molecule kinase inhibitors. *Nat. Rev. Cancer* **9,** 28–39 (2009).

142. Azam, M., Seeliger, M. A., Gray, N. S., Kuriyan, J. & Daley, G. Q. Activation of tyrosine kinases by mutation of the gatekeeper threonine. *Nat. Struct. Mol. Biol.* **15,** 1109–1118 (2008).

143. Palmberg, C. *et al.* Androgen receptor gene amplification in a recurrent prostate cancer after monotherapy with the nonsteroidal potent antiandrogen Casodex (bicalutamide) with a subsequent favorable response to maximal androgen blockade. *Eur. Urol.* **31,** 216–9 (1997).

144. Bardelli, A. *et al.* Amplification of the *MET* Receptor Drives Resistance to Anti-EGFR Therapies in Colorectal Cancer. *Cancer Discov.* **3,** 658–673 (2013).

145. Faber, A. C. *et al.* BIM Expression in Treatment-Naive Cancers Predicts Responsiveness to Kinase Inhibitors. *Cancer Discov.* **1,** 352–365 (2011).

146. Weinstein, I. B. Disorders in cell circuitry during multistage carcinogenesis: the role of homeostasis. *Carcinogenesis* **21,** 857–64 (2000).

147. Morris, L. G. T. & Chan, T. A. Therapeutic targeting of tumor suppressor genes. *Cancer* **121,** 1357–68 (2015).

148. Burgess, A. *et al.* Clinical Overview of MDM2/X-Targeted Therapies. *Front. Oncol.* **6,** 7 (2016).

149. Templeton, A. J. *et al.* Phase 2 Trial of Single-agent Everolimus in

Chemotherapy-naive Patients with Castration-resistant Prostate Cancer (SAKK 08/08). *Eur. Urol.* **64,** 150–158 (2013).

150. Hayes, D. N. *et al.* Phase II Efficacy and Pharmacogenomic Study of Selumetinib (AZD6244; ARRY-142886) in Iodine-131 Refractory Papillary Thyroid Carcinoma with or without Follicular Elements. *Clin. Cancer Res.* **18,** 2056–2065 (2012).

151. FDA. *Lynparza Prescribing Information*. (2014).

152. Wiemann, B. & Starnes, C. O. Coley's toxins, tumor necrosis factor and cancer research: A historical perspective. *Pharmacol. Ther.* **64,** 529–564 (1994).

153. FDA. *BCG Vaccine Perscribing Information*. (1990).

154. DELORME, E. J. & ALEXANDER, P. TREATMENT OF PRIMARY FIBROSARCOMA IN THE RAT WITH IMMUNE LYMPHOCYTES. *Lancet (London, England)* **2,** 117–20 (1964).

155. Perica, K., Varela, J. C., Oelke, M. & Schneck, J. Adoptive T cell immunotherapy for cancer. *Rambam Maimonides Med. J.* **6,** e0004 (2015).

156. Wei, G., Ding, L., Wang, J., Hu, Y. & Huang, H. Advances of CD19-directed chimeric antigen receptor-modified T cells in refractory/relapsed acute lymphoblastic leukemia. *Exp. Hematol. Oncol.* **6,** 10 (2017).

157. Kochenderfer, J. N. *et al.* B-cell depletion and remissions of malignancy along with cytokine-associated toxicity in a clinical trial of anti-CD19 chimeric-antigen-receptor-transduced T cells. *Blood* **119,** 2709–2720 (2012).

158. Khalil, D. N., Smith, E. L., Brentjens, R. J. & Wolchok, J. D. The future of cancer treatment: immunomodulation, CARs and combination immunotherapy. *Nat. Rev. Clin. Oncol.* **13,** 273–290 (2016).

159. Hughes, P. E., Caenepeel, S. & Wu, L. C. Targeted Therapy and Checkpoint Immunotherapy Combinations for the Treatment of Cancer. *Trends Immunol.* (2016). doi:10.1016/j.it.2016.04.010

160. Koyama, S. *et al.* Adaptive resistance to therapeutic PD-1

blockade is associated with upregulation of alternative immune checkpoints. *Nat. Commun.* **7,** 10501 (2016).

161. Moja, L. *et al.* Trastuzumab containing regimens for early breast cancer. *Cochrane database Syst. Rev.* CD006243 (2012). doi:10.1002/14651858.CD006243.pub2

162. FDA. *Tarceva Prescribing Information*. (2004).

163. Kandoth, C. *et al.* Mutational landscape and significance across 12 major cancer types. *Nature* **502,** 333–9 (2013).

164. Halling-Brown, M. D., Bulusu, K. C., Patel, M., Tym, J. E. & Al-Lazikani, B. canSAR: an integrated cancer public translational research and drug discovery resource. *Nucleic Acids Res.* **40,** D947–D956 (2012).

165. Griffith, M. *et al.* DGIdb: mining the druggable genome. *Nat. Methods* **10,** 1209–1210 (2013).

166. Dienstmann, R. *et al.* Standardized decision support in next generation sequencing reports of somatic cancer variants. *Mol. Oncol.* **8,** 859–873 (2014).

167. Schroeder, M. P., Rubio-Perez, C., Tamborero, D., Gonzalez-Perez, A. & Lopez-Bigas, N. OncodriveROLE classifies cancer driver genes in loss of function and activating mode of action. *Bioinformatics* **30,** i549-55 (2014).

168. Furman, R. R. Prognostic markers and stratification of chronic lymphocytic leukemia. *Hematol. Am. Soc. Hematol. Educ. Progr.* **2010,** 77–81 (2010).

169. Mertens, D. & Stilgenbauer, S. Prognostic and predictive factors in patients with chronic lymphocytic leukemia: relevant in the era of novel treatment approaches? *J. Clin. Oncol.* **32,** 869–72 (2014).

170. Garcia-Murillas, I. *et al.* Mutation tracking in circulating tumor DNA predicts relapse in early breast cancer. *Sci. Transl. Med.* **7,** 302ra133-302ra133 (2015).

171. Forshew, T. *et al.* Noninvasive identification and monitoring of cancer mutations by targeted deep sequencing of plasma DNA. *Sci. Transl. Med.* **4,** 136ra68 (2012).

172. Dawson, S.-J. *et al.* Analysis of Circulating Tumor DNA to Monitor Metastatic Breast Cancer. *N. Engl. J. Med.* **368,** 1199–1209 (2013).

173. Martinez, P., McGranahan, N., Birkbak, N. J., Gerlinger, M. & Swanton, C. Computational optimisation of targeted DNA sequencing for cancer detection. *Sci. Rep.* **3,** 3309 (2013).

174. Alemán, A., Garcia-Garcia, F., Medina, I. & Dopazo, J. A web tool for the design and management of panels of genes for targeted enrichment and massive sequencing for clinical applications. *Nucleic Acids Res.* **42,** W83-7 (2014).

175. Illumina Inc. DesignStudio. Available at: www.illumina.com/%0Adesignstudio.

176. Van Allen, E. M., Wagle, N. & Levy, M. A. Clinical analysis and interpretation of cancer genome data. *J. Clin. Oncol.* **31,** 1825–33 (2013).

177. Rubio-Perez, C. *et al.* In Silico Prescription of Anticancer Drugs to Cohorts of 28 Tumor Types Reveals Targeting Opportunities. *Cancer Cell* **27,** 382–396 (2015).

178. Griffith, M. *et al.* CIViC is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer. *Nat. Genet.* **49,** 170–174 (2017).

179. Patterson, S. E. *et al.* The clinical trial landscape in oncology and connectivity of somatic mutational profiles to targeted therapies. *Hum. Genomics* **10,** 4 (2016).

180. Chakravarty, D. *et al.* OncoKB: A Precision Oncology Knowledge Base. *JCO Precis. Oncol.* 1–16 (2017). doi:10.1200/PO.17.00011

181. Huang, L. *et al.* The cancer precision medicine knowledge base for structured clinical-grade mutations and interpretations. *J. Am. Med. Informatics Assoc.* **24,** ocw148 (2016).

182. Ali, H. R. *et al.* Patterns of Immune Infiltration in Breast Cancer and Their Clinical Implications: A Gene-Expression-Based Retrospective Study. *PLOS Med.* **13,** e1002194 (2016).

183. Li, B. *et al.* Comprehensive analyses of tumor immunity:

implications for cancer immunotherapy. *Genome Biol.* **17,** 174 (2016).

184. Gentles, A. J. *et al.* The prognostic landscape of genes and infiltrating immune cells across human cancers. *Nat. Med.* **21,** 938–45 (2015).

185. Rooney, M. S. *et al.* Molecular and genetic properties of tumors associated with local immune cytolytic activity. *Cell* **160,** 48–61 (2015).

186. Spranger, S., Bao, R. & Gajewski, T. F. Melanoma-intrinsic b-catenin signalling prevents anti-tumour immunity. *Nature* **523,** 231–235 (2015).

187. Gjerstorff, M. F., Andersen, M. H. & Ditzel, H. J. Oncogenic cancer/testis antigens: prime candidates for immunotherapy. *Oncotarget* **6,** 15772–15787 (2015).

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

189. Vinay, D. S. *et al.* Immune evasion in cancer: Mechanistic basis and therapeutic strategies. *Semin. Cancer Biol.* **35,** S185–S198 (2015).

190. Hanna, A. & Shevde, L. A. Hedgehog signaling: modulation of cancer properies and tumor mircroenvironment. *Mol. Cancer* **15,** 24 (2016).

191. Pickup, M., Novitskiy, S. & Moses, H. L. The roles of TGF? in the tumour microenvironment. *Nat. Rev. Cancer* **13,** 788–799 (2013).

192. Rahat, M. A., Hemmerlein, B. & Iragavarapu-Charyulu, V. The regulation of angiogenesis by tissue cell-macrophage interactions. *Front. Physiol.* **5,** 262 (2014).

193. Riabov, V. *et al.* Role of tumor associated macrophages in tumor angiogenesis and lymphangiogenesis. *Front. Physiol.* **5,** 75 (2014).

194. Morwood, S. R. & Nicholson, L. B. Modulation of the immune response by extracellular matrix proteins. *Arch. Immunol. Ther. Exp. (Warsz).* **54,** 367–374 (2006).

195. Sorokin, L. The impact of the extracellular matrix on inflammation. *Nat. Rev. Immunol.* **10,** 712–723 (2010).

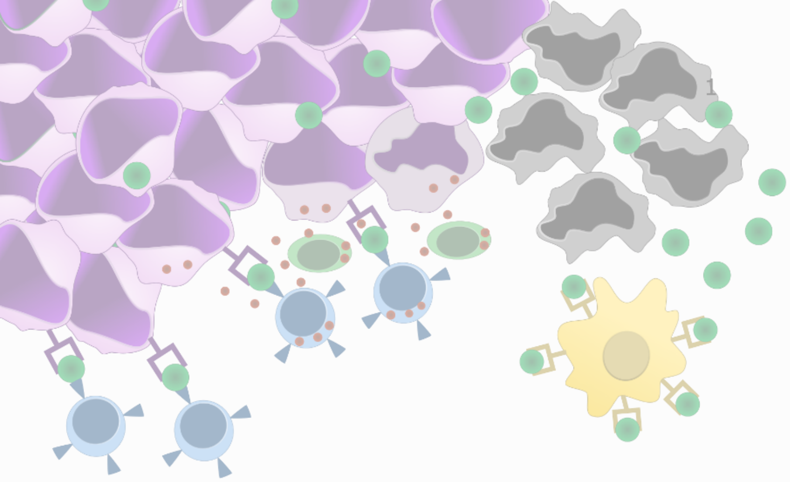
196. Wong, W. Metabolic competition between tumors and T cells. *Sci. Signal.* **8,** (2015).

197. Chang, C. H. *et al.* Metabolic Competition in the Tumor Microenvironment Is a Driver of Cancer Progression. *Cell* **162,** 1229–1241 (2015).

198. Ho, P.-C. C. & Liu, P.-S. S. Metabolic communication in tumors: a new layer of immunoregulation for immune evasion. *J. Immunother. Cancer* **4,** 1 (2016).

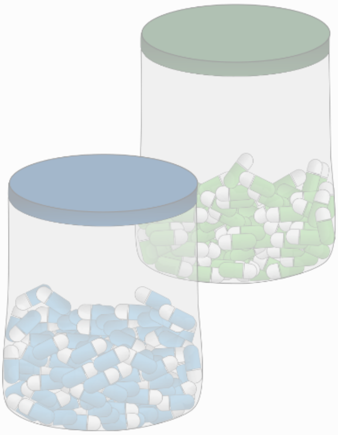
199. Pai, S. G. *et al.* Wnt/beta-catenin pathway: modulating anticancer immune response. *J. Hematol. Oncol.* **10,** 101 (2017).

200. Davoli, T., Uno, H., Wooten, E. C. & Elledge, S. J. Tumor aneuploidy correlates with markers of immune evasion and with reduced response to immunotherapy. *Science (80-.).* **355,** eaaf8399 (2017).



PART VIII

ACRONYMS



CAR	Chimeric Antigen Receptor
CD	Cancer Driver
cDNA	Complementary DNA
CGC	Cancer Gene Census
CGH	Comparative Genomic Hybridization
CNA	Copy Numer Alteration
DC	Dendritic cell
DNA	Desoxiribonucleic acid
DNA-seq	DNA sequencing
FISH	Fluorescence In Situ Hybridization
FPKM	Fragments per kilobase per million mapped read
GTEx	Genotype-Tissue Expression
ICD-O	Interantional Classification of Diseases for Oncology
ICGC	International Cancer Genome Consortium
lncRNA	Long-non coding RNA
mAB	Monoclonal Antibody
NGS	Next Generation Sequencing
OG	Oncogene
PAM	Protein Affecting Mutation
RNA	Ribonucleic acid
RNA-seq	RNA sequencing
RPKM	Reads per kilobase per million mapped read
SNP	Single Nucleotide Polymorphism
SNV	Single Nucleotide Variant
SV	Structural Variant
TCGA	The Cancer Genome Atlas
TIL	Tumor infiltrating lymphocyte
TPM	Transcripts per million mapped read
TSG	Tumor suppressor gene
UV	Ultraviolet