A Compendium of Mutational Cancer Driver Genes

Francisco Martinez-Jimenez[1], Ferran Muiños[1], Inés Sentís[1], Jordi Deu-Pons[1], Iker Reyes-Salazar[1], Claudia Arnedo-Pac[1], Loris Mularoni[1], Oriol Pich[1], Jose Bonet[1], Hanna Kranas[1], Abel Gonzalez-Perez[1,2]*, Nuria Lopez-Bigas[1,2,3]*

Affiliations:

1. Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, Baldiri Reixac, 10, 08028 Barcelona, Spain.
2. Research Program on Biomedical Informatics, Universitat Pompeu Fabra, Barcelona, Catalonia, Spain.
3. Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain

*Corresponding authors
Abel Gonzalez-Perez: abel.gonzalez@irbbarcelona.org
Nuria Lopez-Bigas: nuria.lopez@irbbarcelona.org

Abstract

A fundamental goal in cancer research is to understand the mechanisms of cell transformation. This is key to developing more efficient cancer detection methods and therapeutic approaches. One milestone in this path is the identification of all the genes with mutations capable of driving tumors. Since the 1970s, the list of cancer genes has been growing steadily. Because cancer driver genes are under positive selection in tumorigenesis, their observed patterns of somatic mutations across tumors in a cohort deviate from those expected from neutral mutagenesis. These deviations, or signals may be detected by carefully designed bioinformatics methods, which have become state-of-the-art in the identification of driver genes. A systematic approach combining several of these signals could lead to the compendium of mutational cancer genes. We present the IntOGen pipeline, an implementation of this approach to obtain the compendium of mutational drivers, available through intogen.org. Its application to somatic mutations of more than 28,000 tumors of 66 cancer types reveals 568 cancer genes and points to their mechanisms of tumorigenesis. The application of this approach to the ever-growing datasets of somatic tumor mutations will support the continuous refinement of our knowledge of the genetic basis of cancer.

Introduction

Cancer is a family of diseases characterized by abnormal and uncontrolled cellular growth caused primarily by genetic mutations[1,2]. These mutations, called drivers after their ability to drive tumorigenesis, confer somatic cells in a somatic tissue a selective advantage with respect to neighboring cells[1]. They occur in a set of genes (called cancer driver genes), the mutant forms of which affect the homeostatic development of a set of key cellular functions. One of the main goals of cancer research, since the establishment of genetics, has been the discovery of these cancer driver genes across tumor types[3–6]. Their identification has led to the development of the paradigm of targeted anti-cancer therapies and, more generally to the search for genomic biomarkers of prognosis and response to treatments[7].

The first part of this article presents a historical perspective of the evolution of our knowledge on cancer genes starting before the sequencing of the first whole-exome and whole-genome tumors to the present, and provides an outlook of the future. It focuses on mutational driver genes, i.e., those capable of driving tumorigenesis via single nucleotide variants and short indels, which we call collectively point mutations. On the other hand, it does not cover other types of somatic alterations that affect cancer genes and also contribute to tumorigenesis, such as amplifications or deletions, genomic rearrangements and epigenetic silencing. For comprehensive reviews on some of the types of driver alterations not covered here, see for example[8–10]. Also excluded are methods that identify driver genes based on their vicinity to significantly mutated genes in biochemical pathways or networks, which have also been reviewed elsewhere[11].

In the second part, we propose that the maturity of the analysis methods of mutational driver identification and the wealth of tumor mutational datasets currently available in the public domain advance the goal of uncovering the compendium of driver genes across all tumor types and also point to their tumorigenic mechanisms. To demonstrate this proposition, we developed the IntOGen[12,13] pipeline, aimed at the systematic identification of the compendium of mutational driver genes across tumor types. The snapshot of the compendium of driver genes described in the article has been obtained through the application of state-of-the-art driver discovery methods to 28,076 tumors grouped in 221 cohorts of 66 different tumor types. This snapshot of the compendium of driver genes (and newer versions), as well as the automatic system to produce it are hosted at the IntOGen platform (intogen.org).

The genetic basis of cancer

The search for the causes of cancer is firmly intertwined with the development of genetics[14]. The first scientific notions of the causes of cancer derive from the progression of systematic record-keeping in the 18th and 19th centuries, which linked the high incidence of specific types of tumors to the exposures derived from the practice of some professions[15,16]. The first known report on the heritability of cancer by Broca dates from the late 1800s, even before the genetic basis of inheritance developed by Mendel was widely recognized[17]. In the early 1900s Peyton Rous was able to transmit tumors to healthy birds using cell-free extracts obtained from a diseased animal[18], thus suggesting that units smaller than cells were responsible for tumorigenesis. Contemporary with this finding, and previous to Morgan's work on chromosomes

as the seat of genes, Theodor Boveri proposed that cancer could arise as a result of incorrect chromosomes combinations[19]. This brought the basis of cancer firmly within the realm of genetics. Experiments with chemical carcinogens also demonstrated that changes to the sequence of DNA promoted cellular transformation[20–23].

The improvement of biochemistry and molecular genetics in the decades spanning between 1940 and 1980 fostered the development of laboratory methods like positional cloning, retrotranscription and Sanger sequencing. The application of these methods to cancer research led to the identification of the first cancer driver genes, named after the ability of their mutant forms to drive tumorigenesis. A small portion of the genomes of several birds that hybridized with part of the DNA of the avian sarcoma virus was the first cancer gene to be identified, and was thus named *SRC*[24]. The existence of such viral DNA fragments, a variant of "normal" genes present in the avian genomes which had acquired the transforming capability had already given birth to the term "oncogene" in 1969[25]. Oncogenes such as *HRAS* were then identified in human tumors[26,27], and the change of a single nucleotide in the gene sequence was demonstrated to be enough to provide the transforming capability[28,29]. With these discoveries, the genetic basis of tumorigenesis (including the aforementioned professional exposures) could finally be explained.

As the introduction of defective copies of the driver gene variant, despite the presence of normal alleles in the recipient cell was enough to produce transformation, it was concluded that oncogenes act in a dominant way[30]. However, the analysis of the incidence of retinoblastoma, a pediatric tumor, had shown that two hits, i.e., genetic events inactivating both alleles of the gene (later named *RB1*, after the disease) are necessary for the development of the malignancy[31]. This apparent contradiction was solved by the mid-1980s with the acknowledgment of the existence of a second type of cancer genes, termed tumor suppressors[30]. Unlike in the case of oncogenes, transformation is caused by the inactivation of tumor suppressors, which in general requires loss of activity of both alleles of the gene. The discovery of tumor suppressors also provided an explanation to familial cancer cases[17]: an inherited mutation inactivating one of the alleles of a tumor suppressor increases the likelihood to develop a tumor as only the second hit is required.

Following this clear blueprint of two classes of cancer genes, between the 1980s and early 2000s dozens of genomic loci encoding oncogenes (such as *MYC*, *RET*, *PDGFRA*, *MET*, *KIT*, *FLT3*, *EGFR*, *BRAF*)[32–38] and tumor suppressors (like *TP53*, *TGFRB2*, *RB1*, *PTEN*, *CHEK2*, *CDKN2A*, *BRCA1*, *BRCA2*, *APC*)[39–51] were identified. Germline mutations in some of the latter were also shown to confer susceptibility to cancer development[39,47,51–55]. Further pioneering studies also established the importance of other types of alterations affecting these genes, such as amplifications, deletions, translocations or promoter hypermethylation, for cell transformation[34,53,54,56].

In 2004, a seminal article compiled a list of 291 cancer driver genes from the scientific literature[57], including genes altered through point mutations, translocations or copy-number changes. In an effort to conceptualize this heterogeneity, driver genes were recognized to affect

primarily a handful of essential cellular functions, termed cancer hallmarks[58] (reviewed and updated in 2011[59]). According to this generalization, as a result of driver alterations, malignant cells become capable of: i) resisting apoptosis; ii) maintaining proliferative signaling (even in the absence of extracellular signals); iii) evading suppressors of cell growth; iv) activating invasion and metastasis; v) enabling replicative immortality; vi) inducing angiogenesis; vii) achieving deregulation of energy metabolism; and viii) avoiding immune destruction. In addition, the promotion of tissue inflammation and the genomic instability of tumors were regarded as intrinsic features enabling their evolution.

Somatic mutation patterns reveal drivers

In the early 2000s, improvements introduced in DNA sequencing technologies and the rapid advance in the annotation of the human genome enabled projects aimed at revealing increasing shares of the landscape of somatic mutations in tumors. In 2005, a study sequencing 518 kinase-encoding genes found 76 non-silent mutations on average across 25 breast primary tumors and cell lines[60]. The following year, another group sequenced 13,023 genes of 11 breast and 11 colorectal tumors and found 519 and 673 with mutations, respectively[61]. The development of Next Generation Sequencing (NGS) technologies in the mid 2000s[62] catalyzed the beginning of cancer genomics. In 2008, two further analyses of 22 glioblastomas and 24 pancreatic tumors sequencing the entire exome found 1,007 and 685 mutated genes, respectively[63,64]. A similar landscape arose from the first tumor whole-genomes sequenced[65–68]. Nevertheless, the consensus viewpoint on tumorigenesis was that only a few mutational events affecting driver genes were expected at the root of malignization[30,69]. The vast majority of these mutated genes would, therefore, have no involvement whatsoever in tumorigenesis; in other words, they are passengers, rather than drivers. This finding first exposed the need for statistical tests that included the heterogeneity of mutation rate and mutation types to identify the unexpected mutational patterns that reveal cancer genes[70–72].

These first studies paved the way for the launch of large tumor sequencing initiatives in several countries, such as The Cancer Genome Atlas (TCGA)[67], aimed at sequencing the exomes of hundreds of tumors of two-dozen frequent cancer types. As sequencing technologies continued to expand, more ambitious projects, many grouped under the umbrella of the International Cancer Genomes Consortium (ICGC)[4], set their goal on sequencing the whole genome of thousands of samples. With the recent conclusion of many of these initiatives, comprehensive pan-cancer analyses laid out some of the most important findings of a little over a decade of cancer genomics research[74–77], including lists of identified driver genes[5,78]. The vast majority of these pioneering projects focused on the study of primary malignancies. It is only more recently that similar projects probing metastatic tumors have begun to reveal the landscape of driver alterations of advanced malignancies[79,80].

One of the main goals of all these projects was the identification of the set of genes driving the malignancies. This revolutionary idea provided a rationale for the systematic and comprehensive identification of mutational driver genes. This rationale is rooted in the notion that tumorigenesis follows a Darwinian evolution characterized by variation and selection[81,82]. Variation is provided by randomly appearing somatic mutations that introduce genetic

differences between somatic cells in a tissue. Positive selection then acts upon cells carrying driver mutations that confer selective advantages over neighboring cells leading to clonal expansion of the mutants (Fig. 1a). (A variety of types of selective advantages, described above as the hallmarks of cancer, may be provided by mutations of different genes.)

As a result of this evolutionary process, when a cohort of tumors of the same cancer type is analyzed, the deviation of patterns of mutations in some genes from their expectation may constitute signals that the mutations in those genes are under positive selection in tumorigenesis. For example, driver genes are mutated at abnormally high frequencies across the tumors of a cohort, and methods to detect this significant mutational recurrence were subsequently developed to analyze the mutational datasets produced by these projects[69,83]. Other signals of positive selection in tumorigenesis (Fig. 1b), such as the abnormal clustering of mutations in certain regions of the proteins[84–89], a bias towards the accumulation of mutations with high functional impact[90], or a bias in the frequency of tri-nucleotide changes[91] have been exploited by driver identification methods[92,93]. Over time, many of these methods have been validated and tested on a number of cohorts of different cancer types and shown to be highly reliable. For thorough lists of methods see, for example, refs. [5,78,94,95].

The analysis of the first large mutational datasets revealed that different types of mutations appear with varying frequencies in tumors of different origin and that the rate of mutations across the human genome is highly heterogeneous (see box 1 for details). It quickly became apparent that drivers detection methods are profoundly affected by the heterogeneity of the background mutation rate[96]. Building background models that accurately account for all the factors that affect the mutation rate in the absence of selection has become a hallmark of most driver identification methods developed in the past six years[91,96–102]. While several driver genes mutated at very high frequency may be spotted just by looking at their mutational pattern across tumors[69], the accurate modeling of the background mutation rate is key to avoid false positive drivers and identify those with lower mutation recurrence, and thus uncover the genetic basis of tumorigenesis. The fact that methods exploit different signals of positive selection, and that some cancer driver genes clearly may exhibit one signal but not others makes the combination of their outputs the best approach for a comprehensive identification of cancer driver genes. Spurious discoveries by individual methods also have a higher chance of being filtered out by such combination [5,13,95,103].

Systematic discovery of driver genes

In parallel to the development of drivers identification methods, the adoption of NGS by cancer research, fostered by pioneering initiatives as the ones mentioned in the previous section, has generated a great amount of cancer genomics data available in the public domain. The tally of tumor samples sequenced at the whole-exome or whole-genome level which are currently available for systematic driver discovery is in the tens of thousands. These two premises provide in theory the opportunity to identify the compendium of mutational driver genes (compendium, for short), that is the list of genes driving each malignancy upon mutations.

An implementation of the system

To build a snapshot of this compendium, we have collected somatic mutations across 221 cohorts (comprising between 10 and 973 samples) of 66 different cancer types totaling 28,076 samples (Fig. 2a; Supplementary Methods; Supplementary Table 1). We define as cohort a set of tumor samples of the same cancer type analyzed within a project with a uniform sequencing and calling pipeline. Most samples are contributed by large sequencing efforts, such as the ICGC[4,104] (3,988 samples), TCGA[73] (10,010 samples), PCAWG[105] (2,554), the Hartwig Medical Foundation[79] (3,742) and TARGET[106] (246 samples). Importantly, the mutations across 60 other cohorts comprising 3,570 adult and 1,087 pediatric tumor samples sequenced by individual institutions were obtained via the cBioPortal and PedcBioPortal[107], respectively. This highlights the importance of developing and maintaining centralized efforts to collect sequencing data produced within small projects. Finally, the mutations of 2,257 tumors sequenced as part of eight independent cohorts were obtained from the original studies. In summary, most cohorts (157) comprise primary tumors, but 33 of them are composed of metastatic or relapse samples (4,340). A special effort has been made to include pediatric malignancies (1,972 grouped in 25 cohorts), traditionally underrepresented in driver discovery efforts.

The number of coding mutations in tumors varies depending on the cancer type, and an important degree of variability across the samples of a given malignancy is also observed (Fig. 2b, top). For example, some breast adenocarcinomas bear mutations in several hundred genes, while other samples of the same malignancy exhibit only a dozen mutated genes. Part of this heterogeneity may be explained by differences in sequencing technology or depth, or in mutation calling methods. Nevertheless, most of the heterogeneity in mutation burden has a biological basis, owing to differences in time or intensity of exposure to mutational processes, arising, for example, from faulty DNA repair[108–112]. While re-calling all mutations across the cohorts would eliminate part of the variability of technical origin, this is not yet possible for such large numbers of samples due to limitations in computational power. It is thus necessary, in the effort of systematic discovery of driver genes across cancer types, to analyze each cohort of tumors separately. Larger cohorts provide more statistical power to detect the signals of positive selection that characterize driver genes. In this systematic discovery, therefore, one expects that certain recurrently mutated driver genes appear across many cohorts of the same malignancy, while others will be detected in larger cohorts.

The construction of the compendium by exploiting these datasets of tumor mutations requires an efficient computational system that systematically runs state-of-the-art driver discovery methods. Our implementation of this system, which we refer to as the IntOGen[12,13] (Integrative OncoGenomics; Box 2) pipeline consists of three basic steps, illustrated in Figure 2c, and explained at length in the Supplementary Methods. A first pre-processing step guarantees that each method receives its input in the correct format and within operational parameters, e.g., deduplicating samples taken from the same tumor, or removing those with abnormal missense-to-synonymous mutations ratio or with hypermutator phenotype. Seven recently published complementary methods of driver identification --dNdScv[98], OncodriveFML[99], cBaSE[102], OncodriveCLUSTL[100], a re-implementation of HotMAPS accounting for tri-nucleotide contexts mutation types[88], smRegions[101] and Mutpanning[91]-- are executed next. Then, the lists of candidate drivers identified by each method are combined through a weighted vote in which the

weight awarded to each method is based on its perceived credibility (Supp. Fig. 1). The combination yields lists of driver genes per cohort that are more sensitive than those produced by individual methods without loss of specificity (Supp. Fig. 2). In a final post-processing step, spurious candidate driver genes that may appear due to known confounders are automatically filtered out (Supplementary Methods). The IntOGen pipeline is designed to scale smoothly as the datasets of tumor mutations continue to grow into the hundreds of thousands, advancing our view of the compendium.

Each driver discovery method focuses on one or more features of the mutational pattern of genes across tumors. To identify the signals of positive selection, it assesses the deviation between the observed and expected values of the feature under the assumption of neutral mutations (Fig. 2c). These mutational features, collected by the IntOGen pipeline for all driver genes, provide key insights into the mechanisms of tumorigenesis of each of these cancer genes (see below), and are an integral part of the compendium (Supplementary Methods). They comprise i) the clusters of mutations (both linear or 3D which may arise due to intra- or inter-protein interactions), ii) domains in the protein that are preferentially affected by mutations, and iii) the excess of mutations with different consequence types.

Linear clusters are local accumulations of mutations along the sequence of a gene across tumors, such as those formed by mutations at codons 12 and 13 of *KRAS*. On the other hand, 3D clusters involve amino acid residues which are separated in the primary structure of the protein but close in its tertiary structure (e.g., mutations contributed by amino acids at positions 26,39-42,57 and 59-62 of *RHOA*). Preferentially affected domains bear a significant accumulation of mutations, such as MH2 in *SMAD4*. The excess of mutations with different consequence types (100% and 50% of nonsense and missense mutations, respectively for *ARID1A*) informs about the mode of action (tumor suppressor or oncogene) of a driver gene. The relationship of the excess of nonsense and missense mutations in all drivers of the cohort is thus a good proxy to establish their mode of action in the onset of this malignancy. An excess of observed missense mutations in the absence of an excess of nonsense mutations indicates the activating mode of action of oncogenes. Tumor suppressor (or loss-of-function) genes, on the other hand, tend to exhibit an excess of nonsense mutations. While the mode of action of some genes is very clear-cut, some cases are harder to place within the binary oncogene-tumor suppressor model (close to the diagonal in the graph). Furthermore, the mode of action of some genes may differ between tumor types.

A snapshot of the compendium

How much does the systematic compendium, or more appropriately, the current snapshot obtained from these 221 cohorts of tumors (Box 2) add to the current knowledge of the genetic basis of tumorigenesis? A systematic mining of the literature to establish a thorough and reliable catalogue of validated cancer genes is beyond the scope of our analysis. To address this question, thus we employed the set of driver genes in the Cancer Gene Census[113] (CGC, version 87) as the “ground truth” of the genes involved in the development of the 66 malignancies represented in the compendium. While the CGC is incomplete and may contain some false positives, it is, to our knowledge, the most complete and accurate set of validated

cancer genes annotated from the literature, and it thus serves this purpose. One part of the answer (Fig. 3a,b), then is that almost three quarters of the 568 mutational driver genes in the compendium are already annotated in the CGC (which also provides a strong validation of the compendium). However, because the compendium identifies the signals of positive selection unbiasedly across the cohorts of all cancer types, more than 80% of all identified associations between a driver gene and a malignancy are not annotated in the CGC (Fig. 3a,b). For example, while 21 known CGC drivers of breast adenocarcinomas are in the compendium, 75 genes annotated in the CGC, but not previously recognized to drive this malignancy are shown to be under positive selection across one or more of the 12 breast cancer cohorts analyzed (Fig. 3a). In other words, for many well-known driver genes, the compendium reveals that their role across cancer types is much more widespread than previously documented (Fig. 3c). For example, the pattern of somatic mutations in *KMT2C* shows signals of positive selection across 31 tumor types. However, it is only annotated in the CGC as a driver of medulloblastomas. The unbiased discovery of cancer genes through the IntOGen pipeline is thus an essential complement to the annotation of experimentally validated drivers.

Not only does the systematic nature of the compendium add to our knowledge of the role of well-known cancer genes, but it also points at 152 potential new driver genes (Fig. 3a,c) --i.e., not previously annotated in the CGC. Note that since the CGC is most likely an incomplete proxy of the full catalogue of cancer genes, some of these potential new drivers may have been reported before in the literature. Indeed, we present and discuss below five of these unannotated genes which exhibit signals of positive selection in their mutational pattern across tumors, and have been suggested by independent studies to be involved in tumorigenesis (Fig. 3c, bottom).

The pattern of mutations in *RASA1* across lung and head and neck squamous cell carcinomas exhibits several signals of positive selection probed in the system. Its decreased expression or loss-of-function mutations have been recognized to increase RAS-mediated signaling in human bronchial epithelial[114] and melanoma[115] cell lines. It has also been linked to tumorigenic promoting functions in triple-negative breast cancer[116]. Because *RASA1*, like *NF1*, negatively regulates the RAS/MAPK pathway, both genes are thought to function as tumor suppressors, which is also suggested by their mutational patterns. *KDM3B*, a H3K9me2 demethylase that promotes the transcriptional activation of target genes exhibits significant excess of mutations and functional bias across two cohorts of pilocytic astrocytomas and medulloblastomas. *KDM3B* has been shown to be involved in cell cycle regulation in hepatocellular carcinomas[117], and to function as an activator of the Wnt signaling pathway in colorectal cancer stem cells[118]. Although these two studies suggest that *KDM3B* acts as an oncogene in tumorigenesis, a separate report proposes that some of its germline mutations cause susceptibility to Wilms tumors[119]. Its exact mode of action in tumorigenesis thus remains to be determined. Several Forkhead Box transcription factors are annotated in the CGC as drivers of several malignancies (e.g., *FOXA1* of breast and prostate carcinomas and *FOXR1* of neuroblastomas). Nevertheless, *FOXA2*, with several signals of positive selection across uterine carcinomas is not. *FOXA2* mutations frequently found in uterine carcinomas tend to affect the DNA binding domain or cause the truncation of the protein product[120], causing its failure to localize to the nucleus[121]. Some of

these mutant forms are known to cause a decrease of *CDH1*, and have been thus associated with epithelial to mesenchymal transition in the progression of certain tumors[122,123]. *KLF5*, a transcription factor involved in the regulation of human development identified as a cancer driver gene, altered through different mechanisms[124,125], exhibits signals of positive selection across cervical squamous, bladder and lung squamous cell carcinomas. We also identified *BRD7*, a bromodomain-containing protein with several paralogs already annotated in the CGC, which has been postulated to act as a coactivator of the SMAD transcription factors[126] as a driver of melanomas and liver carcinomas.

Some genes act as drivers across several cancer types, while others tend to be more specific. The compendium provides an opportunity to assess the specificity of driver genes across tumor types in a systematic manner (Fig. 3d). Most genes (360) act as drivers in one or two tumor types, and only a small group of 12 genes (cancer wide drivers) are able to drive more than 20 malignancies through mutations. Some very specific mutational drivers (upper left outliers) are very frequently mutated in only one or two cancer types. For example, 60% and 47% of all Burkitt lymphomas bear driver mutations in *MYC*[127] or *CCND3*[128], respectively. Half of the cases of uveal melanoma bear activating mutations in one of two hotspots of *GNAQ*, while almost all other cases bear mutations at one of two homologous hotspots of its paralog *GNA11*[129]. Interestingly, the transcription factor *GTF2I*, which drives virtually half of all thymomas[130] is not yet annotated in the CGC.

Mutational features of driver genes

We propose that the mutational features (exemplified in Fig. 2c) of a driver gene provide a unique opportunity to shed light on its tumorigenic function (Box 2). Below, we describe the mutational features of six driver genes as examples of the information they provide on the role they play in cell transformation.

The oncogene *PTPN11* (encoding a phosphatase) shows excessive missense mutations across multiple myelomas[131] (Fig. 4a) and other tumor types[132,133], which significantly cluster within its SH2 domain. Inhibitory contacts between this domain and the phosphatase domain are abrogated upon phosphorylation by a receptor tyrosine kinase in the wild-type or by mutations in the domain[134]. The activated *PTPN11* then dephosphorylates inhibitors of several signaling pathways, such as MAPK or AKT pathways[135]. *NFE2L2*, another classic oncogene is a transcription factor key in the control of the redox state of the cell and its response to stress[136–138]. Across lung squamous cell carcinomas[139], two narrow clusters of missense mutations appear at its N-terminal portion (Fig. 4b). These mutations affect sequences recognized by the cognate E3-ubiquitin ligase *KEAP1* (i.e., degrons), and cause the abnormal stabilization of the *NFE2L2* protein[101], as do mutations affecting its recognition domain in *KEAP1*, and cause the constitutive activation of *NFE2L2*-regulated genes[101].

The mutational features are radically different for tumor suppressors like *RB1* across bladder adenocarcinomas[140] (Fig. 4c), with greater excess of nonsense and splice affecting than missense mutations. Most nonsense mutations trigger nonsense mediated decay (NMD) of *RB1* mRNA[141], thus causing a depletion of the protein and abrogating its functions in the regulation of

cell cycle progression and the cell division cycle, the response to cellular stress, differentiation, cellular senescence, programmed cell death and maintenance of chromatin structure[142–144]. *PTEN*, encoding another tumor suppressor, also shows an excess of both nonsense and missense mutations across glioblastomas[73,145] (Fig. 4d). Nonsense mutations trigger NMD, preventing the production of a functional *PTEN* protein product, while missense mutations, hinder either its enzymatic activity or its recruitment to the membrane, or increase its susceptibility to ubiquitination for proteasome-mediated degradation[146,147]. These outcomes, in turn, interfere with its role in the regulation of a host of cellular functions, such as cell cycle progression, apoptosis, and protein synthesis[148–150].

Different tumorigenic mechanisms of the same driver across tumor types may also be revealed by their mutational features. For example, in glioblastomas[73], missense mutations of *EGFR* (an oncogene involved in the activation of multiple signaling pathways) tend to cluster in the extracellular domains of the protein (Fig. 4e). These act as gain-of-function alterations, likely through the stabilization of the open conformation of the receptor, which stimulates its autophosphorylation in the absence of ligand[151,152]. Across lung adenocarcinomas[153], on the other hand, missense mutations tend to cluster in the tyrosine kinase domain (Fig. 4f), altering its 'on-off' equilibrium and increasing its activity at the expense of a reduced affinity for ATP[154,155].

Overall, several protein domains across multiple genes appear as preferentially affected by mutations across more than 10 different tumor types (Fig. 5a,b). The P53 domain appears significantly enriched for somatic mutations across cohorts of a larger share of different cancer types (42) than any other protein domain, although this is driven only by *TP53*. In second place, the tyrosine kinase domain of 13 different genes is significantly enriched for mutations across cohorts of 24 tumor types. *BRAF* is the gene with the tyrosine kinase domain exhibiting a significant enrichment of somatic mutations across the largest number of tumor types (14). The RAS, cadherin and C2H2 zinc finger domains exhibit significant enrichment of mutations across 13 cancer types.

An overview of significant clusters reveals that those in tumor suppressors tend to be wider, while those in oncogenes are sharp and tend to accumulate a larger share of the mutations observed in the gene (Fig. 5c-g). Particularly narrow clusters are observed, for example in *KRAS* (5 nucleotides stretching codons 12 and 13 of the protein) accumulating 85% of the mutations in the gene across a cohort of 496 colorectal adenocarcinomas, or in IDH1 with all mutations in a cohort of 257 acute myeloid leukemias affecting one single nucleotide in codon 132. Wider clusters accumulate 28% of mutations of *TP53* (28 nucleotides between codons 266 and 275) across a cohort of 439 pilocytic astrocytomas or 83% of the mutations of *SPOP* (44 nucleotides between codons 119 and 133) across a cohort of 444 prostate adenocarcinomas (Fig. 5c-f). The width of clusters and the fraction of mutations of the protein that fall within them differ depending on the mode of action of cancer genes in tumorigenesis (Fig. 5g). The relatively narrow clusters of oncogenes reflect the existence of relatively few available gain-of-function mutations along their sequence. This is also the reason why these clusters tend to concentrate large shares of all the mutations observed in oncogenes across a cohort of tumors. Wider

clusters in tumor suppressor genes are observed because as a rule more loss-of-function mutations are available in their sequence (e.g., mutations affecting several amino acids of a functionally important domain).

Conclusions and perspectives

Much like ancient manuscripts, in which newer layers of writing have been superimposed onto older texts, or cities with long history of human dwelling, such as Rome, in which certain edifices exhibit rows of brick and mortar dating from different ages, the somatic mutations in tumor genomes constitute a record of their history. Therefore, borrowing the name given to these ancient scripts, somatic mutations in tumors may be considered a palimpsest[108], the study of which may render extremely useful information about itself and its environment. These palimpsests contain the footprints of all the mutational processes to which somatic cells in the tumor have been exposed during the life of the patient, as well as the signals of positive selection reminiscent of successive selective sweeps caused by driver mutations. Cleverly designed bioinformatics analyses applied to tumor genomes are able to reveal such footprints and traces. This article has shown that the systematic application of such bioinformatics analyses to the detection of positive selection from the palimpsest of tumor somatic mutations are able to reveal the compendium of cancer driver genes.

Before the inception of cancer genomics, a few dozen cancer driver genes were identified (Fig. 6). In the span of two or three decades, these genes were intensively studied and functionally characterized through an array of biochemical assays and the laborious dedication of several research groups. In contrast, in a little over one decade elapsed since the sequencing of the first tumor genomes, several hundred more cancer genes have been identified. This “era” of cancer genomics has been made possible by advances in DNA sequencing and the development of bioinformatics methods to cope with the challenges of genomics data analysis it poses. As shown here, the compendium of mutational driver genes derived from the analysis of the cancer exomes currently in the public domain (~28,000) comprises between 500 and 600 mutational drivers. The completion of the compendium constitutes a milestone in the road of our understanding of tumor biology. Probably genes mutated at frequencies above 10% have already been discovered[97], and systematic analyses reveal their involvement in tumorigenesis across cancer types.

We are also now in a position to project the evolution of the compendium into the future. The number of datasets of tumor somatic mutations deposited in the public domain is foreseen to increase quickly as initiatives to share data generated internationally, such as the Global Alliance for Genomics and Health (www.ga4gh.org), the 1M genomes initiative[156], and others come to fruition. As new snapshots of the compendium are uncovered exploiting these data, the trend described above is predicted to continue into the future, with the identification of i) new drivers mutated at frequencies below 10% across malignancies (owing to more statistical power[97]); ii) drivers of conditions not profiled before; iii) drivers in specific populations or ethnicities that have been biased against so far in tumor genome sequencing projects; and iv) drivers of new clinical entities, such as metastatic or relapse tumors, which have been comparatively less explored to date. For instance, a search through the current snapshot of the

compendium shows that *ESR1* and *AR*, while rarely mutated across primary breast and prostate tumors (respectively) are clear drivers of resistance to treatment.

In this article we have purposefully focused on driver mutations affecting protein-coding genes and left out driver mutations in non-coding elements. As mentioned in the Introduction, this excludes other types of somatic alterations affecting driver genes. Special mention must be made of short insertions and deletions (indels), whose probability of occurrence likely involve features beyond their immediate sequence context and the background rate of which is thus more difficult to calculate[111,112,157]. It also excludes the potential role in tumorigenesis of mutations affecting non-coding genomic elements, of which recent studies have identified few in comparison with coding genes[78,103]. Focusing on known cancer genes and their cis-regulatory regions, one of these surveys revealed that non-coding driver mutations are much less frequent than protein-coding ones (with the exception of *TERT*), even after correcting for differences of statistical power between whole-genome and whole-exome sequencing datasets[78]. Nevertheless, it has also become apparent from whole-genome sequenced tumors that our current knowledge of the distribution of mutations in non-coding regions is still incomplete to allow for a correct modeling of their background mutation rate. Furthermore, our knowledge of the biological function of most of the non-coding areas of the genome still lags far behind that of coding genes[105]. Solving these issues will be key to fully exploring the catalogue of driver non-coding genomic elements. A holistic compendium of all these types of driver alterations (coding and non-coding somatic point mutations, structural variants, epigenetic silencing events and germline susceptibility variants) is needed to uncover their panorama across tumors[103].

A detailed description of the precise involvement of each gene in tumor development is absent from the current snapshot of the compendium of driver genes. Understanding the precise mode of alteration of each driver gene (i.e., which of its mutations have a potential to drive tumorigenesis and why) and the specific biological function it perturbs in tumorigenesis are thus one of the major challenges of cancer genomics in the near future.

A first challenge is to precisely identify the mechanisms that alter the function of driver genes turning them capable of driving tumorigenesis. This is the same as identifying all the mutations of cancer driver genes that are capable of driving the malignancy and understanding how they do it[7,98,103]. As explained above, we propose that the mutational features computed within the compendium may aid in this endeavor. Furthermore, while the perturbation of several key biological processes (the hallmarks of cancer detailed above) are required for tumorigenesis, the specific process --e.g., evading apoptosis, maintaining proliferative signaling, escaping the immune system-- touched by mutations in many of the genes in the compendium is still unknown. The interpretation of the significance of driver mutations is also confounded by intratumoral heterogeneity and by the complexity of the ecology of the microenvironment of cancer cells[158,159]. Profiling other dimensions of tumors, such as transcriptomics, proteomics and methylomics, as well as systematic assays on the function of individual genes and their interactions[160–162], and single-cell profiling approaches[163–166] will contribute to bridging this gap.

A second challenge arises from the fact that while driver genes are identified in isolation by their signals of positive selection it is in fact a set of driver mutations that causes tumorigenesis[98,103]. For example, driver mutations affecting four specific pathways are known to occur in the vast majority of colorectal adenocarcinomas and are required from the progression of a healthy cell to an adenoma and finally an invasive carcinoma[69]. Furthermore, while the signals of positive selection in all driver genes in a tumor cohort are equivalent, driver mutations probably occur at different stages of the evolution of a tumor. Again, the clever application of bioinformatics to the analysis of the cancer genome palimpsest has allowed us to start resolving this temporal order[167]; nevertheless, more work is needed to understand it.

There is finally the challenge to fully understand how other features besides somatic mutations cooperate in tumorigenesis. While virtually all tumors contain genomic driver mutations[103], those are not sufficient to explain the history of cell transformation. Studies of somatic mutations from healthy donors have shown that many cancer drivers are already mutated in non transformed cells across somatic tissues[168–171]. The same has been shown in studies of premalignant stages of tumorigenesis[172,173] (e.g, in clonal hematopoiesis) or benign tumors[174,175]. This has led to the conclusion that a certain level of positive selection is present in healthy somatic tissues in a continuum, without reaching the level of cell transformation. In this continuum, positive selection occurs on mutations that confer a fitness advantage, which likely vary between somatic tissues and over time. A mutation thus can only be a driver when presented against a background of specific selective constraints. In some cases to reach the level of cell transformation non-genetic phenotypic changes may also be important. Such changes have been documented in processes such as resistance to drugs and metastasis[176–180].

In summary, closing the gap between the list of genes in the compendium and our complete knowledge of the process of tumorigenesis is one of the big challenges of cancer genomics for the near future. Applying clever bioinformatics analyses to the integrated analysis of cancer genomics and other dimensions of tumor cells will once again be key in this endeavor. In turn, gaining this insight into tumorigenesis will be fundamental to translate our knowledge of cancer genomics into precision cancer medicine.

Box 1. The background mutation rate of genes

The background mutation rate of a gene (i.e. the rate and distribution of mutations) in a somatic cell is determined by its sequence, the tissue and the mutational processes the person has been exposed to during their lifetime. A correct assessment of the background mutation rate of genes requires to accurately model the variability introduced by all these factors. This is key to identify which observed mutational patterns are actually unexpected and attributable to positive selection.

The mutational processes active in a tissue in an individual define a set of probabilities of each nucleotide in the gene to change taking into account its immediate sequence context[108,111,112,181,182]. These probabilities may be learned from the observed mutational profile of each tumor in a cohort, or derived from the activity of a set of relevant mutational processes across the samples of a cohort[183].

The probability that a specific nucleotide change occurs in the gene is also influenced by the specific features adopted by the chromatin of the cell both at the large and the small scales[96,184,185]. At the large scale, the time at which the gene is replicated relative to an origin[186], the level of compaction of the chromatin[187,188] at its locus, and the level at which the gene is expressed[96] influence its mutation rate. The effect of these large-scale factors may be carefully modeled for each gene in each relevant tissue[96,98]. Alternatively, a background model within each gene may be built by permuting the mutations observed in the gene following their local probability[99–101].

At the small scale, factors such as the occupancy by nucleosomes[189,190] and other proteins[191], the distribution of certain chromatin marks along the gene body[192,193], and the formation of local non B-DNA structures[194–196] may alter the mutation rate locally at sequence stretches within the gene.

Box 2: Accessing the compendium of mutational driver genes

The snapshot of the compendium of driver genes described in this article, as well as the automatic system to produce it are hosted at the IntOGen platform (intogen.org). Cancer researchers may explore the compendium, comprising the list of driver genes across tumor types and their mutational features, via the web interface of the platform. All the information contained in it is also downloadable. Furthermore, the automatic system (the IntOGen pipeline) can be obtained by researchers from the platform for local installation and application to datasets of somatic mutations across cohorts of tumors. Details on the current implementation of the IntOGen pipeline appear in Supplementary Methods. Building upon a practice that dates back to 2013, when the IntOGen platform for the analysis of cancer driver genes was first established[12,13], we will continue to collect tumor sequencing data as it becomes available in the public domain, and to produce more complete snapshots of the compendium. For future versions of the pipeline and the compendium, regular updates may be found at www.intogen.org.

Figure legends

Figure 1. Signals of positive selection identify driver genes
(a) Cells in somatic tissues accumulate mutations. Somatic mutations in certain genes provide the cell where they occur a selective advantage and are thus positively selected. Following a Darwinian process, over time, a clonal expansion occurs and thus the cells carrying mutations in these genes become predominant within the population.
(b) Deviations of the observed pattern of mutations of genes across samples of the same cancer types from the expectation reveal the genes under positive selection in tumorigenesis. Two samples are taken from a cancer patient: one from the tumor and another from a healthy tissue (e. g., peripheral blood in solid malignancies). Comparing the sequences of these two samples, the somatic point mutations in the tumor are identified. The number of somatic mutations identified in the exome or the genome of tumors varies one or two orders of magnitude (respectively), depending on the cancer type and the exposure of donors to specific mutational processes. As a result, between a few dozen and several thousand genes appear mutated in each tumor. The driver genes are those that exhibit one or more signals of positive selection across the tumors of a cohort.

Figure 2. Application of the IntOGen pipeline to datasets of tumor mutations
(a) Datasets of tumor mutations collected from the public domain for the construction of the current snapshot of the compendium of driver genes. Both donut plots represent all datasets classified by source (left) or cancer type (right). In both plots, the innermost ring signals the cohorts from primary or metastatic/relapse tumors, while the second highlights cohorts of adult or pediatric tumors.
(b) Mutation burden (top) and mutation type (bottom) of tumors from cancer types represented by at least two cohorts. Below the plot, the number of cohorts and samples contributing to the distribution of each cancer type are shown.
(c) Schematic representation of the IntOGen pipeline exemplified through the flow of data resulting from its application to a cohort of tumors. The two outcomes of the pipeline, i.e, the catalog of driver genes in the cohort and the mutational features computed in each of them integrate the compendium of driver genes.

Figure 3. A snapshot of the compendium of mutational driver genes
(a) Overlap between the genes in the compendium in each cancer type and the Cancer Gene Census (CGC). The word clouds illustrate the genes driving tumorigenesis in three example cancer types, with the size of the driver name following its mutational frequency. The three-color scale to denote genes annotated in the CGC in the same tumor type or in a different tumor type, or genes not annotated in the CGC is used throughout the figure.
(b) Overlap between the genes in the compendium and the CGC (top bar) and between driver gene-tumor type associations in the compendium and the CGC (bottom bar).
(c) The landscape of tumorigenic associations between 25 well known mutational driver genes and the tumor types in the compendium is much denser than that annotated in the CGC. The bottom of the plot lays out the involvement in tumorigenesis of five previously unannotated

drivers across tumor types. The size of the dots represents the percentage of all cohorts of the tumor type in which the gene is identified as a driver.
(d) Distribution of prevalence of driver genes across cancer types in the compendium. Each driver gene is represented as a single dot in the graph. The abscissa represents the number of tumor types where a gene has been identified as driver and the ordinate, its maximum mutational frequency across them. The distribution of these two variables separately are represented through one-dimensional histograms above and at the right side of the graph. Two sets of drivers mutated at high frequency across one or very few tumor types (cancer-specific highly prevalent) and mutated across more than 20 cancer types (cancer wide drivers) are circled and denoted by their names. While most cancer wide drivers are *bona fide* long-known cancer genes, *LRP1B* has long been suspected to be a potential spurious finding. The discussion is not settled, since some studies have found its loss of function may be related to enhanced cell migration in several tissues[197–199]. The barplots at the right present the mutational frequency across tumor types of selected cancer-specific highly prevalent and cancer wide drivers. The maximum mutational frequency of each of them appears beside the corresponding row.

Figure 4. Interpreting the mutational patterns of driver genes
Six exemplary mutational patterns computed for five genes across five cohorts, including multiple myelomas (obtained from a study published in 2018[131]), and lung squamous cell carcinomas, bladder adenocarcinomas, glioblastomas and lung adenocarcinomas obtained from TCGA. Clusters and their boundaries are defined by methods that assess the significant clustering of mutations. In all plots N denotes the number of mutations of each consequence type observed in the gene across the cohort.

Figure 5. Recurrent cancer driver domains and mutational clusters
(a) Dots represent domains with significant enrichment for mutations in a number of different driver genes across a number of different tumor types.
(b) Genes with significant enrichment for mutations in domains colored in (a) across tumor types.
(c-f) Number of mutations and prevalence in the cohort of linear mutational clusters identified in several drivers across (c) colorectal adenocarcinomas (obtained from TCGA), (d) AML (obtained from the Beat AML project[200]), (e) prostate adenocarcinomas (obtained from a publication), and (f) pilocytic astrocytomas (obtained from ICGC). The fraction of mutations of each protein in each cohort that appear in clusters and their width appear in the heatmaps below each figure.
(g) Linear clusters detected in tumor suppressors (blue) and oncogenes (red) across all cohorts in the compendium.

Figure 6. Past, present and future of cancer genomics
Conceptual representation of the evolution of the compendium starting with the identification of the first cancer genes before the start of the cancer genomics era, through the sequencing of the first cancer tumors to the moment of writing this review, and outlook of consolidation and future trends of cancer genomics research.

References

1. Stratton, M. R., Campbell, P. J. & Futreal, P. A. The cancer genome. *Nature* **458**, 719–24 (2009).
2. Stratton, M. R. Exploring the genomes of cancer cells: progress and promise. *Science* **331**, 1553–8 (2011).
3. Mwenifumbo, J. C. & Marra, M. A. Cancer genome-sequencing study design. *Nat. Rev. Genet.* **14**, 321–332 (2013).
4. ICGC. International network of cancer genome projects. *Nature* **464**, 993–998 (2010).
5. Bailey, M. H. *et al.* Comprehensive Characterization of Cancer Driver Genes and Mutations. *Cell* **173**, 371-385.e18 (2018).
6. Weinstein, J. N. *et al.* The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.* **45**, 1113–20 (2013).
7. Tamborero, D. *et al.* Cancer Genome Interpreter annotates the biological and clinical relevance of tumor alterations. *Genome Med.* **10**, 25 (2018).
8. Mertens, F., Johansson, B., Fioretos, T. & Mitelman, F. The emerging complexity of gene fusions in cancer. *Nat. Rev. Cancer* **15**, 371–381 (2015).
9. Santarius, T., Shipley, J., Brewer, D., Stratton, M. R. & Cooper, C. S. A census of amplified and overexpressed human cancer genes. *Nat. Rev. Cancer* **10**, 59–64 (2010).
10. Baylin, S. B. & Ohm, J. E. Epigenetic gene silencing in cancer – a mechanism for early oncogenic pathway addiction? *Nat. Rev. Cancer* **6**, 107–116 (2006).
11. Kuenzi, B. M. & Ideker, T. A census of pathway maps in cancer systems biology. *Nat. Rev. Cancer* **20**, 233–246 (2020).
12. Gonzalez-Perez, A. *et al.* IntOGen-mutations identifies cancer drivers across tumor types. *Nat. Methods* **10**, 1081–2 (2013).
13. 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).

14. Faguet, GB. A brief history of cancer: Age□old milestones underlying our current knowledge database - Faguet - 2015 - International Journal of Cancer - Wiley Online Library. https://onlinelibrary.wiley.com/doi/epdf/10.1002/ijc.29134.

15. Greenberg M., Selikof IJ. LUNG CANCER IN THE SCHNEEBERG MINES: A REAPPRAISAL OF THE DATA REPORTED BY HARTING AND HESSE IN 1879 | Annals of Work Exposures and Health | Oxford Academic. https://academic.oup.com/annweh/article-abstract/37/1/5/160211?redirectedFrom=fulltext.

16. Waldron, H. A. A brief history of scrotal cancer. *Occup. Environ. Med.* **40**, 390–401 (1983).

17. Rahman, N. Realising the Promise of Cancer Predisposition Genes. *Nature* **505**, 302–308 (2014).

18. Martin, G. S. The road to Src. *Oncogene* **23**, 7910–7917 (2004).

19. Boveri, Theodor. *Zur Frage der Entstehung Maligner Tumoren*. (Gustav Fischer, 1914).

20. Bouck, N. & di Mayorca, G. Somatic mutation as the basis for malignant transformation of BHK cells by chemical carcinogens. *Nature* **264**, 722–727 (1976).

21. Shih, C., Shilo, B. Z., Goldfarb, M. P., Dannenberg, A. & Weinberg, R. A. Passage of phenotypes of chemically transformed cells via transfection of DNA and chromatin. *Proc. Natl. Acad. Sci.* **76**, 5714–5718 (1979).

22. Krontiris, T. G. & Cooper, G. M. Transforming activity of human tumor DNAs. *Proc. Natl. Acad. Sci.* **78**, 1181–1184 (1981).

23. Cooper, G. M., Okenquist, S. & Silverman, L. Transforming activity of DNA of chemically transformed and normal cells. *Nature* **284**, 418–421 (1980).

24. Stehelin, D., Varmus, H. E., Bishop, J. M. & Vogt, P. K. DNA related to the transforming gene(s) of avian sarcoma viruses is present in normal avian DNA. *Nature* **260**, 170–173 (1976).

25. Huebner, R. J. & Todaro, G. J. Oncogenes of RNA tumor viruses as determinants of

cancer. *Proc. Natl. Acad. Sci. U. S. A.* **64**, 1087–1094 (1969).

26. Parada, L. F., Tabin, C. J., Shih, C. & Weinberg, R. A. Human EJ bladder carcinoma oncogene is homologue of Harvey sarcoma virus ras gene. *Nature* **297**, 474–478 (1982).

27. Santos, E., Tronick, S. R., Aaronson, S. A., Pulciani, S. & Barbacid, M. T24 human bladder carcinoma oncogene is an activated form of the normal human homologue of BALB- and Harvey-MSV transforming genes. *Nature* **298**, 343 (1982).

28. 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).

29. Tabin, C. J. *et al.* Mechanism of activation of a human oncogene. *Nature* **300**, 143–149 (1982).

30. Klein, G. & Klein, E. Evolution of tumours and the impact of molecular oncology. *Nature* **315**, 190–195 (1985).

31. Knudson, A. G. Mutation and Cancer: Statistical Study of Retinoblastoma. *Proc. Natl. Acad. Sci.* **68**, 820 (1971).

32. Meyer, N. & Penn, L. Z. Reflecting on 25 years with MYC. *Nat. Rev. Cancer* **8**, 976–990 (2008).

33. Eng, C. & Mulligan, L. M. Mutations of theRET proto-oncogene in the multiple endocrine neoplasia type 2 syndromes, related sporadic tumours, and Hirschsprung disease. *Hum. Mutat.* **9**, 97–109 (1997).

34. Zhuang, Z. *et al.* Trisomy 7-harbouring non-random duplication of the mutant MET allele in hereditary papillary renal carcinomas. *Nat. Genet.* **20**, 66–69 (1998).

35. Hirota, S. *et al.* Gain-of-Function Mutations of c-kit in Human Gastrointestinal Stromal Tumors. *Science* **279**, 577 (1998).

36. Gilliland, D. G. & Griffin, J. D. The roles of FLT3 in hematopoiesis and leukemia. *Blood* **100**, 1532–1542 (2002).

37. Wong, A. J. *et al.* Structural alterations of the epidermal growth factor receptor gene in human gliomas. *Proc. Natl. Acad. Sci.* **89**, 2965–2969 (1992).

38. Davies, H. *et al.* Mutations of the BRAF gene in human cancer. *Nature* **417**, 949–954 (2002).

39. Laken, S. J. *et al.* Familial colorectal cancer in Ashkenazim due to a hypermutable tract in APC. *Nat. Genet.* **17**, 79–83 (1997).

40. Nigro, J. M. *et al.* Mutations in the p53 gene occur in diverse human tumour types. *Nature* **342**, 705–708 (1989).

41. Baker, S. J. *et al.* Chromosome 17 deletions and p53 gene mutations in colorectal carcinomas. *Science* **244**, 217–221 (1989).

42. Grady, W. M. *et al.* Mutational Inactivation of Transforming Growth Factor β Receptor Type II in Microsatellite Stable Colon Cancers. *Cancer Res.* **59**, 320 (1999).

43. Friend, S. H. *et al.* A human DNA segment with properties of the gene that predisposes to retinoblastoma and osteosarcoma. *Nature* **323**, 643–646 (1986).

44. Dunn, J. M., Phillips, R. A., Becker, A. J. & Gallie, B. L. Identification of germline and somatic mutations affecting the retinoblastoma gene. *Science* **241**, 1797–1800 (1988).

45. Li, J. PTEN, a Putative Protein Tyrosine Phosphatase Gene Mutated in Human Brain, Breast, and Prostate Cancer. *Science* **275**, 1943–1947 (1997).

46. Kamb, A. *et al.* A cell cycle regulator potentially involved in genesis of many tumor types. *Science* **264**, 436–440 (1994).

47. Meijers-Heijboer, H. *et al.* Low-penetrance susceptibility to breast cancer due to CHEK2*1100delC in noncarriers of BRCA1 or BRCA2 mutations. *Nat. Genet.* **31**, 55–59 (2002).

48. Tavtigian, S. V. *et al.* The complete BRCA2 gene and mutations in chromosome 13q-linked kindreds. *Nat. Genet.* **12**, 333–337 (1996).

49. Miki, Y. *et al.* A strong candidate for the breast and ovarian cancer susceptibility gene

BRCA1. *Science* **266**, 66–71 (1994).

50. Hall, J. M. *et al.* Linkage of early-onset familial breast cancer to chromosome 17q21. *Science* **250**, 1684–1689 (1990).

51. Wooster, R. *et al.* Localization of a breast cancer susceptibility gene, BRCA2, to chromosome 13q12-13. *Science* **265**, 2088–2090 (1994).

52. Malkin, D. *et al.* Germ line p53 mutations in a familial syndrome of breast cancer, sarcomas, and other neoplasms. *Science* **250**, 1233–1238 (1990).

53. Merlo, A. *et al.* 5′ CpG island methylation is associated with transcriptional silencing of the tumour suppressor p16/CDKN2/MTS1 in human cancers. *Nat. Med.* **1**, 686–692 (1995).

54. Miki, Y. *et al.* A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. *Science* **266**, 66–71 (1994).

55. Hall, J. *et al.* Linkage of early-onset familial breast cancer to chromosome 17q21. *Science* **250**, 1684–1689 (1990).

56. Powers, M. P. The ever-changing world of gene fusions in cancer: a secondary gene fusion and progression. *Oncogene* **38**, 7197–7199 (2019).

57. Futreal, A. *et al.* A census of human cancer genes. *Nat. Rev. Cancer* **4**, 177–183 (2004).

58. Hanahan, D. & Weinberg, R. A. The hallmarks of cancer. *Cell* **100**, 57–70 (2000).

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

60. Stephens, P. *et al.* A screen of the complete protein kinase gene family identifies diverse patterns of somatic mutations in human breast cancer. *Nat. Genet.* **37**, 590–592 (2005).

61. Sjöblom, T. *et al.* The consensus coding sequences of human breast and colorectal cancers. *Science* **314**, 268–274 (2006).

62. Goodwin, S., McPherson, J. D. & McCombie, W. R. Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.* **17**, 333–351 (2016).

63. Jones, S. *et al.* Core signaling pathways in human pancreatic cancers revealed by global

genomic analyses. *Science* **321**, 1801–6 (2008).

64. Parsons, D. W. *et al.* An Integrated Genomic Analysis of Human Glioblastoma Multiforme. *Science* **321**, 1807–1812 (2008).

65. Ley, T. J. *et al.* DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature* **456**, 66–72 (2008).

66. Pleasance, E. D. *et al.* A small-cell lung cancer genome with complex signatures of tobacco exposure. *Nature* **463**, 184–190 (2010).

67. Pleasance, E. D. *et al.* A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature* **463**, 191–196 (2010).

68. Shah, S. P. *et al.* Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution. *Nature* **461**, 809–813 (2009).

69. Vogelstein, B. *et al.* Cancer genome landscapes. *Science* **339**, 1546–58 (2013).

70. Forrest, W. F. & Cavet, G. Comment on 'The consensus coding sequences of human breast and colorectal cancers'. *Science* **317**, 1500; author reply 1500 (2007).

71. Getz, G. *et al.* Comment on 'The consensus coding sequences of human breast and colorectal cancers'. *Science* **317**, 1500 (2007).

72. Rubin, A. F. & Green, P. Comment on 'The consensus coding sequences of human breast and colorectal cancers'. *Science* **317**, 1500 (2007).

73. The Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455**, 1061–1068 (2008).

74. Hoadley, K. A. *et al.* Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer. *Cell* **173**, 291-304.e6 (2018).

75. Sanchez-Vega, F. *et al.* Oncogenic Signaling Pathways in The Cancer Genome Atlas. *Cell* **173**, 321-337.e10 (2018).

76. Ding, L. *et al.* Perspective on Oncogenic Processes at the End of the Beginning of

Cancer Genomics. *Cell* **173**, 305-320.e10 (2018).

77. Campbell, P. J. *et al.* Pan-cancer analysis of whole genomes. *Nature* **578**, 82–93 (2020).

78. Rheinbay, E. *et al.* Analyses of non-coding somatic drivers in 2,658 cancer whole genomes. *Nature* **578**, 102–111 (2020).

79. Priestley, P. *et al.* Pan-cancer whole-genome analyses of metastatic solid tumours. *Nature* **575**, 210–216 (2019).

80. Pleasance, E. *et al.* Pan-cancer analysis of advanced patient tumors reveals interactions between therapy and genomic landscapes. *Nat. Cancer* **1**, 452–468 (2020).

81. Greaves, M. & Maley, C. C. Clonal evolution in cancer. *Nature* **481**, 306–313 (2012).

82. McGranahan, N. & Swanton, C. Clonal Heterogeneity and Tumor Evolution: Past, Present, and the Future. *Cell* **168**, 613–628 (2017).

83. Dees, N. D. *et al.* MuSiC: Identifying mutational significance in cancer genomes. *Genome Res.* **22**, 1589–98 (2012).

84. Tamborero, D. *et al.* OncodriveCLUST: exploiting the positional clustering of somatic mutations to identify cancer genes. *Rev.* **29**, 2238–44 (2013).

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

86. Kamburov, A. *et al.* Comprehensive assessment of cancer missense mutation clustering in protein structures. *Proc. Natl. Acad. Sci.* **112**, E5486 (2015).

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

88. Tokheim, C. *et al.* Exome-scale discovery of hotspot mutation regions in human cancer using 3D protein structure. *Cancer Res.* **76**, 3719–3731 (2016).

89. Niu, B. *et al.* Protein-structure-guided discovery of functional mutations across 19 cancer types. *Nat. Genet.* **48**, 827–837 (2016).

90. Gonzalez-Perez, A. & Lopez-Bigas, N. Functional impact bias reveals cancer drivers.

Nucleic Acids Res. **40**, e169 (2012).

91. Dietlein, F. *et al.* Identification of cancer driver genes based on nucleotide context. *Nat. Genet.* 1–11 (2020) doi:10.1038/s41588-019-0572-y.

92. Davoli, T. *et al.* Cumulative haploinsufficiency and triplosensitivity drive aneuploidy patterns and shape the cancer genome. *Cell* **155**, 948–62 (2013).

93. 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 (2016).

94. Porta-Pardo, E. *et al.* Comparison of algorithms for the detection of cancer drivers at subgene resolution. *Nat. Methods* **14**, 782–788 (2017).

95. Tamborero, D. *et al.* Comprehensive identification of mutational cancer driver genes across 12 tumor types. *Sci. Rep.* **3**, 2650 (2013).

96. Lawrence, M. S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214–218 (2013).

97. Lawrence, M. S. *et al.* Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* **505**, 495–501 (2014).

98. Martincorena, I. *et al.* Universal Patterns of Selection in Cancer and Somatic Tissues. *Cell* **171**, 1029-1041.e21 (2017).

99. Mularoni, L. *et al.* OncodriveFML: a general framework to identify coding and non-coding regions with cancer driver mutations. (2016) doi:10.1186/s13059-016-0994-0.

100. Arnedo-Pac, C., Mularoni, L., Muiños, F., Gonzalez-Perez, A. & Lopez-Bigas, N. OncodriveCLUSTL: a sequence-based clustering method to identify cancer drivers. *Bioinformatics* doi:10.1093/bioinformatics/btz501.

101. Martínez-Jiménez, F., Muiños, F., López-Arribillaga, E., Lopez-Bigas, N. & Gonzalez-Perez, A. Systematic analysis of alterations in the ubiquitin proteolysis system reveals its contribution to driver mutations in cancer. *Nat. Cancer* (2019) doi:10.1038/s43018-019-0001-2.

102. Weghorn, D. & Sunyaev, S. Bayesian inference of negative and positive selection in human cancers. *Nat. Genet.* **49**, 1785–1788 (2017).

103. Sabarinathan, R. *et al.* The whole-genome panorama of cancer drivers. *bioRxiv* (2017) doi:10.1101/190330.

104. Zhang, J. *et al.* International Cancer Genome Consortium Data Portal--a one-stop shop for cancer genomics data. *Database* **2011**, bar026–bar026 (2011).

105. Pan-cancer analysis of whole genomes. *Nature* **578**, 82–93 (2020).

106. Ma, X. *et al.* Pan-cancer genome and transcriptome analyses of 1,699 paediatric leukaemias and solid tumours. *Nature* **555**, 371–376 (2018).

107. Cerami, E. *et al.* The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data. *Cancer Discov.* **2**, 401–404 (2012).

108. Nik-Zainal, S. *et al.* The Life History of 21 Breast Cancers. *Cell* **149**, 994–1007 (2012).

109. Ma, J., Setton, J., Lee, N. Y., Riaz, N. & Powell, S. N. The therapeutic significance of mutational signatures from DNA repair deficiency in cancer. *Nat. Commun.* **9**, 3292 (2018).

110. Phillips, D. H. Mutational spectra and mutational signatures: Insights into cancer aetiology and mechanisms of DNA damage and repair. *DNA Repair* **71**, 6–11 (2018).

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

112. Alexandrov, L. B. *et al.* The repertoire of mutational signatures in human cancer. *Nature* **578**, 94–101 (2020).

113. Sondka, Z. *et al.* The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. *Nat. Rev. Cancer* **18**, 696 (2018).

114. Hayashi, T. *et al.* RASA1 and NF1 are Preferentially Co-Mutated and Define A Distinct Genetic Subset of Smoking-Associated Non–Small Cell Lung Carcinomas Sensitive to MEK Inhibition. *Clin. Cancer Res.* (2017) doi:10.1158/1078-0432.CCR-17-2343.

115. Sung, H. *et al.* Inactivation of RASA1 promotes melanoma tumorigenesis via R-Ras

activation. *Oncotarget* **7**, 23885–23896 (2016).

116. Suárez-Cabrera, C. *et al.* A Transposon-based Analysis Reveals *RASA1* Is Involved in Triple-Negative Breast Cancer. *Cancer Res.* **77**, 1357–1368 (2017).

117. An, M.-J. *et al.* Histone demethylase KDM3B regulates the transcriptional network of cell-cycle genes in hepatocarcinoma HepG2 cells. *Biochem. Biophys. Res. Commun.* **508**, 576–582 (2019).

118. Li, J. *et al.* KDM3 epigenetically controls tumorigenic potentials of human colorectal cancer stem cells through Wnt/β-catenin signalling. *Nat. Commun.* **8**, 1–15 (2017).

119. Mahamdallie, S. *et al.* Identification of new Wilms tumour predisposition genes: an exome sequencing study. *Lancet Child Adolesc. Health* **3**, 322–331 (2019).

120. Smith, B. *et al.* The mutational spectrum of FOXA2 in endometrioid endometrial cancer points to a tumor suppressor role. *Gynecol. Oncol.* **143**, 398–405 (2016).

121. Neff, R. *et al.* Functional characterization of recurrent FOXA2 mutations seen in endometrial cancers. *Int. J. Cancer* **143**, 2955–2961 (2018).

122. Song, Y., Washington, M. K. & Crawford, H. C. Loss of FOXA1/2 is essential for the epithelial-to-mesenchymal transition in pancreatic cancer. *Cancer Res.* **70**, 2115–2125 (2010).

123. Zhang, Z. *et al.* FOXA2 attenuates the epithelial to mesenchymal transition by regulating the transcription of E-cadherin and ZEB2 in human breast cancer. *Cancer Lett.* **361**, 240–250 (2015).

124. Zhang, X. *et al.* Somatic Superenhancer Duplications and Hotspot Mutations Lead to Oncogenic Activation of the KLF5 Transcription Factor. *Cancer Discov.* **8**, 108–125 (2018).

125. Jia, L. *et al.* KLF5 promotes breast cancer proliferation, migration and invasion in part by upregulating the transcription of TNFAIP2. *Oncogene* **35**, 2040–2051 (2016).

126. Liu, T. *et al.* Tumor suppressor bromodomain-containing protein 7 cooperates with Smads to promote transforming growth factor-β responses. *Oncogene* **36**, 362–372 (2017).

127. Cowling, V. H., Turner, S. A. & Cole, M. D. Burkitt's lymphoma-associated c-Myc mutations converge on a dramatically altered target gene response and implicate Nol5a/Nop56 in oncogenesis. *Oncogene* **33**, 3519–3527 (2014).

128. Jayaraman, S. S., Rayhan, D. J., Hazany, S. & Kolodney, M. S. Mutational Landscape of Basal Cell Carcinomas by Whole-Exome Sequencing. *J. Invest. Dermatol.* **134**, 213–220 (2014).

129. Van Raamsdonk, C. D. *et al.* Mutations in GNA11 in Uveal Melanoma. *N. Engl. J. Med.* **363**, 2191–2199 (2010).

130. Radovich, M. *et al.* The integrated genomic landscape of thymic epithelial tumors. *Cancer Cell* **33**, 244-258.e10 (2018).

131. Hoang, P. H. *et al.* Whole-genome sequencing of multiple myeloma reveals oncogenic pathways are targeted somatically through multiple mechanisms. *Leukemia* **32**, 2459–2470 (2018).

132. Prahallad, A. *et al.* PTPN11 Is a Central Node in Intrinsic and Acquired Resistance to Targeted Cancer Drugs. *Cell Rep.* **12**, 1978–1985 (2015).

133. Hill, K. S. *et al.* PTPN11 Plays Oncogenic Roles and Is a Therapeutic Target for BRAF Wild-Type Melanomas. *Mol. Cancer Res.* **17**, 583–593 (2019).

134. Keilhack, H., David, F. S., McGregor, M., Cantley, L. C. & Neel, B. G. Diverse Biochemical Properties of Shp2 Mutants IMPLICATIONS FOR DISEASE PHENOTYPES. *J. Biol. Chem.* **280**, 30984–30993 (2005).

135. Östman, A., Hellberg, C. & Böhmer, F. D. Protein-tyrosine phosphatases and cancer. *Nat. Rev. Cancer* **6**, 307–320 (2006).

136. Qian, Z. *et al.* Nuclear factor, erythroid 2-like 2-associated molecular signature predicts lung cancer survival. *Sci. Rep.* **5**, 1–10 (2015).

137. Kerins, M. J. & Ooi, A. A catalogue of somatic NRF2 gain-of-function mutations in cancer. *Sci. Rep.* **8**, 1–13 (2018).

138. Stewart, P. A. *et al.* Proteogenomic landscape of squamous cell lung cancer. *Nat. Commun.* **10**, 1–17 (2019).

139. Hammerman, P. S. *et al.* Comprehensive genomic characterization of squamous cell lung cancers. *Nature* **489**, 519–525 (2012).

140. The Cancer Genome Atlas Research Network. Comprehensive molecular characterization of urothelial bladder carcinoma. *Nature* **507**, 315–22 (2014).

141. Lindeboom, R. G. H., Supek, F. & Lehner, B. The rules and impact of nonsense-mediated mRNA decay in human cancers. *Nat. Genet.* **48**, 1112–1118 (2016).

142. Di Fiore, R., D'Anneo, A., Tesoriere, G. & Vento, R. RB1 in cancer: different mechanisms of RB1 inactivation and alterations of pRb pathway in tumorigenesis. *J. Cell. Physiol.* **228**, 1676–1687 (2013).

143. Dick, F. A., Goodrich, D. W., Sage, J. & Dyson, N. J. Non-canonical functions of the RB protein in cancer. *Nat. Rev. Cancer* **18**, 442–451 (2018).

144. Goodrich, D. W. The retinoblastoma tumor-suppressor gene, the exception that proves the rule. *Oncogene* **25**, 5233–5243 (2006).

145. Brennan, C. W. *et al.* The Somatic Genomic Landscape of Glioblastoma. *Cell* **155**, 462–477 (2013).

146. Yang, J.-M. *et al.* Characterization of PTEN mutations in brain cancer reveals that pten mono-ubiquitination promotes protein stability and nuclear localization. *Oncogene* **36**, 3673–3685 (2017).

147. Nguyen, H.-N. *et al.* A new class of cancer-associated PTEN mutations defined by membrane translocation defects. *Oncogene* **34**, 3737–3743 (2015).

148. Hollander, M. C., Blumenthal, G. M. & Dennis, P. A. PTEN loss in the continuum of common cancers, rare syndromes and mouse models. *Nat. Rev. Cancer* **11**, 289–301 (2011).

149. Yin, Y. & Shen, W. H. PTEN: a new guardian of the genome. *Oncogene* **27**, 5443–5453

(2008).

150. Keniry, M. & Parsons, R. The role of PTEN signaling perturbations in cancer and in targeted therapy. *Oncogene* **27**, 5477–5485 (2008).

151. Furnari, F. B., Cloughesy, T. F., Cavenee, W. K. & Mischel, P. S. Heterogeneity of epidermal growth factor receptor signalling networks in glioblastoma. *Nat. Rev. Cancer* **15**, 302–310 (2015).

152. Xu, H. *et al.* Epidermal growth factor receptor in glioblastoma. *Oncol. Lett.* **14**, 512–516 (2017).

153. Collisson, E. A. *et al.* Comprehensive molecular profiling of lung adenocarcinoma. *Nature* **511**, 543–550 (2014).

154. Gazdar, A. F. Activating and resistance mutations of EGFR in non-small-cell lung cancer: role in clinical response to EGFR tyrosine kinase inhibitors. *Oncogene* **28**, S24–S31 (2009).

155. Sharma, S. V., Bell, D. W., Settleman, J. & Haber, D. A. Epidermal growth factor receptor mutations in lung cancer. *Nat. Rev. Cancer* **7**, 169–181 (2007).

156. Saunders, G. *et al.* Leveraging European infrastructures to access 1 million human genomes by 2022. *Nat. Rev. Genet.* **20**, 693–701 (2019).

157. Degasperi, A. *et al.* A practical framework and online tool for mutational signature analyses show intertissue variation and driver dependencies. *Nat. Cancer* **1**, 249–263 (2020).

158. Reiter, J. G. *et al.* An analysis of genetic heterogeneity in untreated cancers. *Nat. Rev. Cancer* **19**, 639–650 (2019).

159. Maley, C. C. *et al.* Classifying the evolutionary and ecological features of neoplasms. *Nat. Rev. Cancer* **17**, 605–619 (2017).

160. Dempster, J. M. *et al.* Agreement between two large pan-cancer CRISPR-Cas9 gene dependency data sets. *Nat. Commun.* **10**, 1–14 (2019).

161. Tsherniak, A. *et al.* Defining a Cancer Dependency Map. *Cell* **170**, 564-576.e16 (2017).

162. Behan, F. M. *et al.* Prioritization of cancer therapeutic targets using CRISPR-Cas9 screens. *Nature* **568**, 511–516 (2019).

163. Lawson, D. A., Kessenbrock, K., Davis, R. T., Pervolarakis, N. & Werb, Z. Tumour heterogeneity and metastasis at single-cell resolution. *Nat. Cell Biol.* **20**, 1349–1360 (2018).

164. Baslan, T. & Hicks, J. Unravelling biology and shifting paradigms in cancer with single-cell sequencing. *Nat. Rev. Cancer* **17**, 557–569 (2017).

165. Levitin, H. M., Yuan, J. & Sims, P. A. Single-Cell Transcriptomic Analysis of Tumor Heterogeneity. *Trends Cancer* **4**, 264–268 (2018).

166. Wagner, J. *et al.* A Single-Cell Atlas of the Tumor and Immune Ecosystem of Human Breast Cancer. *Cell* **177**, 1330-1345.e18 (2019).

167. Gerstung, M. *et al.* The evolutionary history of 2,658 cancers. *Nature* **578**, 122–128 (2020).

168. Martincorena, I. *et al.* Tumor evolution. High burden and pervasive positive selection of somatic mutations in normal human skin. *Science* **348**, 880–886 (2015).

169. Martincorena, I. *et al.* Somatic mutant clones colonize the human esophagus with age. *Science* **362**, 911–917 (2018).

170. Lee-Six, H. *et al.* The landscape of somatic mutation in normal colorectal epithelial cells. *Nature* **574**, 532–537 (2019).

171. Jaiswal, S. & Ebert, B. L. Clonal hematopoiesis in human aging and disease. *Science* **366**, (2019).

172. Weaver, J. M. J. *et al.* Ordering of mutations in preinvasive disease stages of esophageal carcinogenesis. *Nat. Genet.* **46**, 837–843 (2014).

173. Gregson, E. M., Bornschein, J. & Fitzgerald, R. C. Genetic progression of Barrett's oesophagus to oesophageal adenocarcinoma. *Br. J. Cancer* **115**, 403–410 (2016).

174. Kanojia, D. *et al.* Identification of somatic alterations in lipoma using whole exome sequencing. *Sci. Rep.* **9**, (2019).

175. Ye, L. *et al.* The genetic landscape of benign thyroid nodules revealed by whole exome and transcriptome sequencing. *Nat. Commun.* **8**, 1–8 (2017).

176. Pisco, A. O. & Huang, S. Non-genetic cancer cell plasticity and therapy-induced stemness in tumour relapse: 'What does not kill me strengthens me'. *Br. J. Cancer* **112**, 1725–1732 (2015).

177. Holohan, C., Van Schaeybroeck, S., Longley, D. B. & Johnston, P. G. Cancer drug resistance: an evolving paradigm. *Nat. Rev. Cancer* **13**, 714–726 (2013).

178. Payne, J. L. & Wagner, A. The causes of evolvability and their evolution. *Nat. Rev. Genet.* **20**, 24 (2019).

179. Nguyen, D. X., Bos, P. D. & Massagué, J. Metastasis: from dissemination to organ-specific colonization. *Nat. Rev. Cancer* **9**, 274–284 (2009).

180. Ganesh, K. *et al.* L1CAM defines the regenerative origin of metastasis-initiating cells in colorectal cancer. *Nat. Cancer* **1**, 28–45 (2020).

181. Alexandrov, L. B. *et al.* Mutational signatures associated with tobacco smoking in human cancer. *Science* **354**, 618–622 (2016).

182. Alexandrov, L. B. *et al.* Clock-like mutational processes in human somatic cells. *Nat. Genet.* **47**, 1402–1407 (2015).

183. Rosenthal, R., McGranahan, N., Herrero, J., Taylor, B. S. & Swanton, C. DeconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. *Genome Biol.* **17**, 31 (2016).

184. Gonzalez-Perez, A., Sabarinathan, R. & Lopez-Bigas, N. Local Determinants of the Mutational Landscape of the Human Genome. *Cell* **177**, 101–114 (2019).

185. Supek, F. & Lehner, B. Scales and mechanisms of somatic mutation rate variation across the human genome. *DNA Repair* **81**, 102647 (2019).

186. Stamatoyannopoulos, J. A. *et al.* Human mutation rate associated with DNA replication timing. *Nat. Genet.* **41**, 393–395 (2009).

187. Polak, P. *et al.* Reduced local mutation density in regulatory DNA of cancer genomes is linked to DNA repair. *Nat. Biotechnol.* **32**, 71–5 (2014).

188. Polak, P. *et al.* Cell-of-origin chromatin organization shapes the mutational landscape of cancer. *Nature* **518**, 360–364 (2015).

189. Pich, O. *et al.* Somatic and Germline Mutation Periodicity Follow the Orientation of the DNA Minor Groove around Nucleosomes. *Cell* **175**, 1074-1087.e18 (2018).

190. Brown, A. J., Mao, P., Smerdon, M. J., Wyrick, J. J. & Roberts, S. A. Nucleosome positions establish an extended mutation signature in melanoma. *PLOS Genet.* **14**, e1007823 (2018).

191. Sabarinathan, R. *et al.* Nucleotide excision repair is impaired by binding of transcription factors to DNA. *Nature* **532**, 264–267 (2016).

192. Frigola, J. *et al.* Reduced mutation rate in exons due to differential mismatch repair. *Nat. Genet.* **49**, 1684–1692 (2017).

193. Supek, F. & Lehner, B. Clustered Mutation Signatures Reveal that Error-Prone DNA Repair Targets Mutations to Active Genes. *Cell* **170**, 534-547.e23 (2017).

194. Buisson, R. *et al.* Passenger hotspot mutations in cancer driven by APOBEC3A and mesoscale genomic features. *Science* **364**, (2019).

195. Hess, J. M. *et al.* Passenger Hotspot Mutations in Cancer. *Cancer Cell* **36**, 288-301.e14 (2019).

196. Georgakopoulos-Soares, I., Morganella, S., Jain, N., Hemberg, M. & Nik-Zainal, S. Noncanonical secondary structures arising from non-B DNA motifs are determinants of mutagenesis. *Genome Res.* **28**, 1264–1271 (2018).

197. Tanaga, K. *et al.* LRP1B attenuates the migration of smooth muscle cells by reducing membrane localization of urokinase and PDGF receptors. *Arterioscler. Thromb. Vasc. Biol.* **24**, 1422–1428 (2004).

198. Li, Y. *et al.* Low density lipoprotein (LDL) receptor-related protein 1B impairs urokinase

receptor regeneration on the cell surface and inhibits cell migration. *J. Biol. Chem.* **277**, 42366–42371 (2002).

199. Wang, Z. *et al.* Down-regulation of LRP1B in colon cancer promoted the growth and migration of cancer cells. *Exp. Cell Res.* **357**, 1–8 (2017).

200. Tyner, J. W. *et al.* Functional genomic landscape of acute myeloid leukaemia. *Nature* **562**, 526–531 (2018).

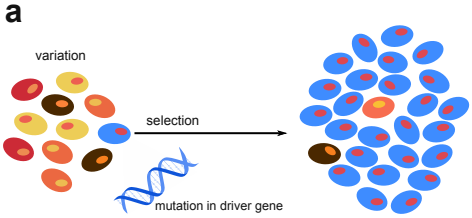
Acknowledgements
First and foremost, we must acknowledge the contribution of cancer patients, their families and a myriad of medical doctors and cancer genomics researchers who laboriously gather, process and sequence tens of thousands of tumor samples. Without them, the compendium of mutational driver genes would not be possible. We are also greatly indebted to generations of researchers who laid the foundations of cancer genomics, generated and shared data and developed methods for driver identification. N.L-B. acknowledges funding from the European Research Council (consolidator grant 682398) and Spanish Ministry of Economy and Competitiveness (SAF2015-66084-R, MINECO/FEDER, UE). 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). The results shown here are in whole or part based upon data generated by the TCGA Research Network, the Pan-Cancer Analysis of Whole Genomes (PCAWG), the cBioPortal, the Hartwig Medical Foundation, the International Cancer Genomes Consortium (ICGC), the St. Jude pediatric hospital, the Pediatric cBioPortal, TARGET projects, the BEAT AML study, and several other studies scattered throughout the scientific literature.

Author contributions
IntOGen pipeline conceptualization: F.M-J., F.M, A. G-P. and N.L-B. Combination approach development: F.M. and F.M-J. Driver identification methods re-implementation: C.A-P., L.M. and F.M-J. Downstream analyses: F.M-J., F.M., O.P., H.K., J.B., C.A-P. Analysis and discussion of the snapshot of the compendium: F.M-J., F.M., O.P., A.G-P., N.L-B. Figure preparation: F.M-J., F.M., L.M., J.B.. Data collection and annotation: I.S., F.M-J., L.M. IntOGen pipeline development and maintenance: J.D-P., F.M-J., L.M. and I.R-S. IntOGen website development and maintenance: I.R-S., F.M-J. Project supervision: A.G-P., N.L-B. Manuscript preparation: A.G-P., N.L-B.

Competing interests
None to declare.

Figure 1

a



b

Somatic mutations from tumor sample

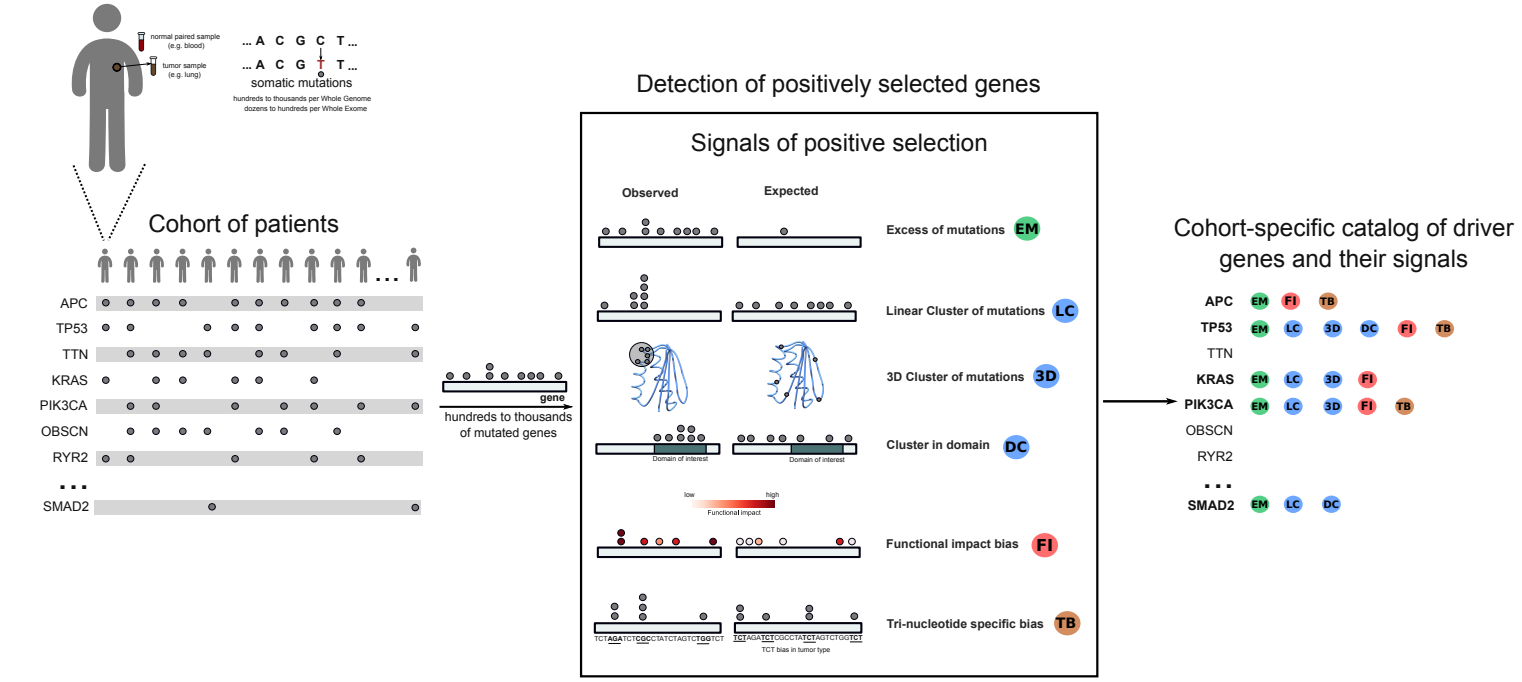
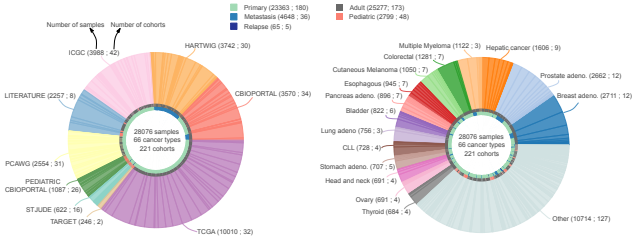
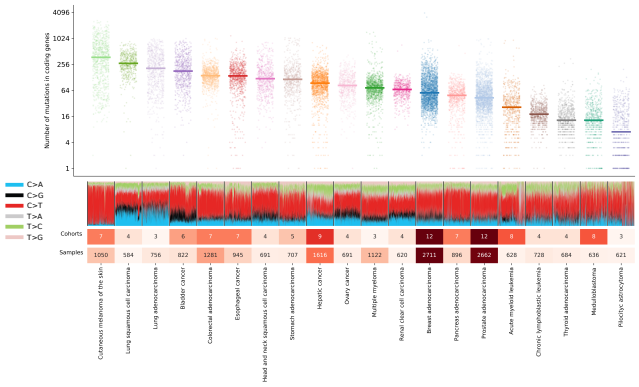


Figure 2

a


Primary (23363 ; 180)
Metastasis (4648 ; 36)
Relapse (65 ; 5)
Adult (25277; 173)
Pediatric (2799 ; 48)
Number of samples
Number of cohorts
ICGC (3988 ; 42)
HARTWIG (3742 ; 30)
CBIOPORTAL (3570 ; 34)
LITERATURE (2257 ; 8)
PCAWG (2554 ; 31)
PEDIATRIC CBIOPORTAL (1087 ; 26)
STJUDE (622 ; 16)
TARGET (246 ; 2)
TCGA (10010 ; 32)
28076 samples
66 cancer types
221 cohorts
Multiple Myeloma (1122 ; 3)
Hepatic cancer (1606 ; 9)
Colorectal (1281 ; 7)
Cutaneous Melanoma (1050 ; 7)
Esophagous (945 ; 7)
Pancreas adeno. (896 ; 7)
Prostate adeno. (2662 ; 12)
Bladder (822 ; 6)
Breast adeno. (2711 ; 12)
Lung adeno (756 ; 3)
CLL (728 ; 4)
Stomach adeno. (707 ; 5)
Head and neck (691 ; 4)
Ovary (691 ; 4)
Thyroid (684 ; 4)
Other (10714 ; 127)

b


Number of mutations in coding genes
4096
1024
256
64
16
4
1
C>A
C>G
C>T
T>A
T>C
T>G
Cohorts
Samples
Cutaneous melanoma of the skin
Lung squamous cell carcinoma
Lung adenocarcinoma
Bladder cancer
Colorectal adenocarcinoma
Esophageal cancer
Head and neck squamous cell carcinoma
Stomach adenocarcinoma
Hepatic cancer
Ovary cancer
Multiple myeloma
Renal clear cell carcinoma
Breast adenocarcinoma
Pancreas adenocarcinoma
Prostate adenocarcinoma
Acute myeloid leukemia
Chronic lymphoblastic leukemia
Thyroid adenocarcinoma
Medulloblastoma
Pilocytic astrocytoma

c

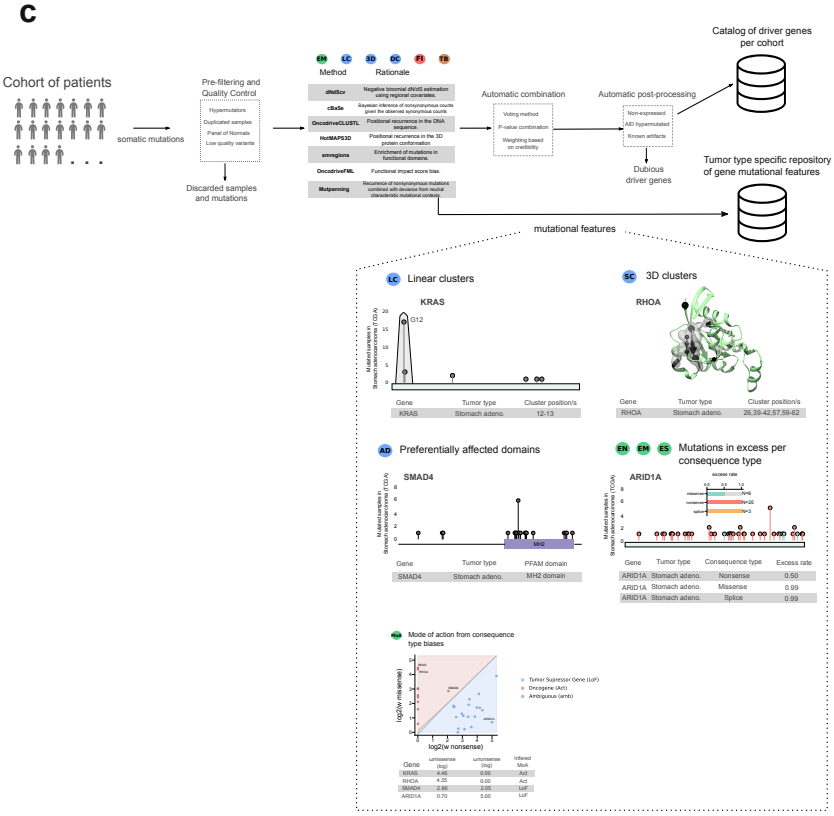

Cohort of patients
somatic mutations
Pre-filtering and Quality Control
Hypermutators
Duplicated samples
Panel of Normals
Low quality variants
Discarded samples and mutations
EM LC 3D DC FI TB
Method
Rationale
dNdScv
Negative binomial dN/dS estimation using regional covariates.
cBaSe
Bayesian inference of nonsynonymous counts given the observed synonymous counts
OncodriveCLUSTL
Positional recurrence in the DNA sequence.
HotMAPS3D
Positional recurrence in the 3D protein conformation
smregions
Enrichment of mutations in functional domains.
OncodriveFML
Functional impact score bias.
Mutpanning
Recurrence of nonsynonymous mutations combined with deviance from neutral characteristic mutational context.
Automatic combination
Voting method
P-value combination
Weighting based on credibility
Automatic post-processing
Non-expressed
IAD hypermutated
Known artifacts
Dubious driver genes
Catalog of driver genes per cohort
Tumor type specific repository of gene mutational features
mutational features
LC Linear clusters
KRAS
G12
Gene
Tumor type
Cluster position/s
KRAS
Stomach adeno.
12-13
SC 3D clusters
RHOA
RHOA
Stomach adeno.
28,39-42,57,59-62
AD Preferentially affected domains
SMAD4
MH2
SMAD4
Stomach adeno.
MH2 domain
EN EM ES Mutations in excess per consequence type
ARID1A
excess rate
ARID1A
Stomach adeno.
Nonsense
0.50
ARID1A
Stomach adeno.
Missense
0.99
ARID1A
Stomach adeno.
Splice
0.99
Mode of action from consequence type biases
log2(w missense)
log2(w nonsense)
Tumor Supressor Gene (LoF)
Oncogene (Act)
Ambiguous (amb)
Gene
Inferred MoA
KRAS
Act
RHOA
Act
SMAD4
LoF
ARID1A
LoF

Figure 3

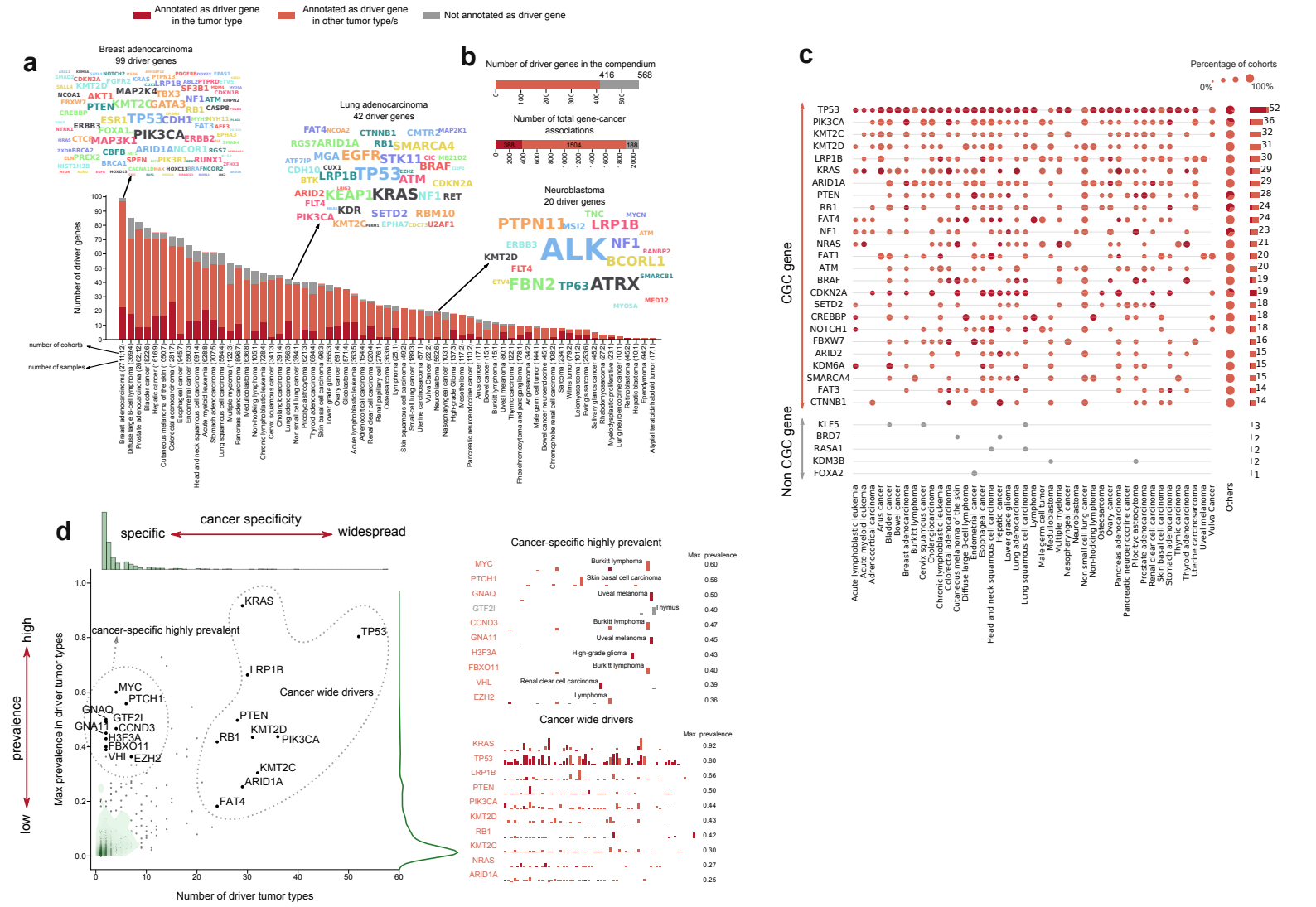

Annotated as driver gene in the tumor type
Annotated as driver gene in other tumor type/s
Not annotated as driver gene
a
Breast adenocarcinoma
99 driver genes
Lung adenocarcinoma
42 driver genes
Neuroblastoma
20 driver genes
Number of driver genes
number of cohorts
number of samples
b
Number of driver genes in the compendium
416
568
Number of total gene-cancer associations
388
1504
188
c
Percentage of cohorts
0%
100%
CGC gene
Non CGC gene
TP53
PIK3CA
KMT2C
KMT2D
LRP1B
KRAS
ARID1A
PTEN
RB1
FAT4
NF1
NRAS
FAT1
ATM
BRAF
CDKN2A
CREBBP
NOTCH1
FBXW7
ARID2
KDM6A
SMARCA4
FAT3
CTNNB1
KLF5
BRD7
RASA1
KDM3B
FOXA2
Others
d
cancer specificity
specific
widespread
Max prevalence in driver tumor types
Number of driver tumor types
high
prevalence
low
cancer-specific highly prevalent
Cancer wide drivers
MYC
PTCH1
GNAQ
GTF2I
CCND3
GNA11
H3F3A
FBXO11
VHL
EZH2
Max. prevalence

Figure 4

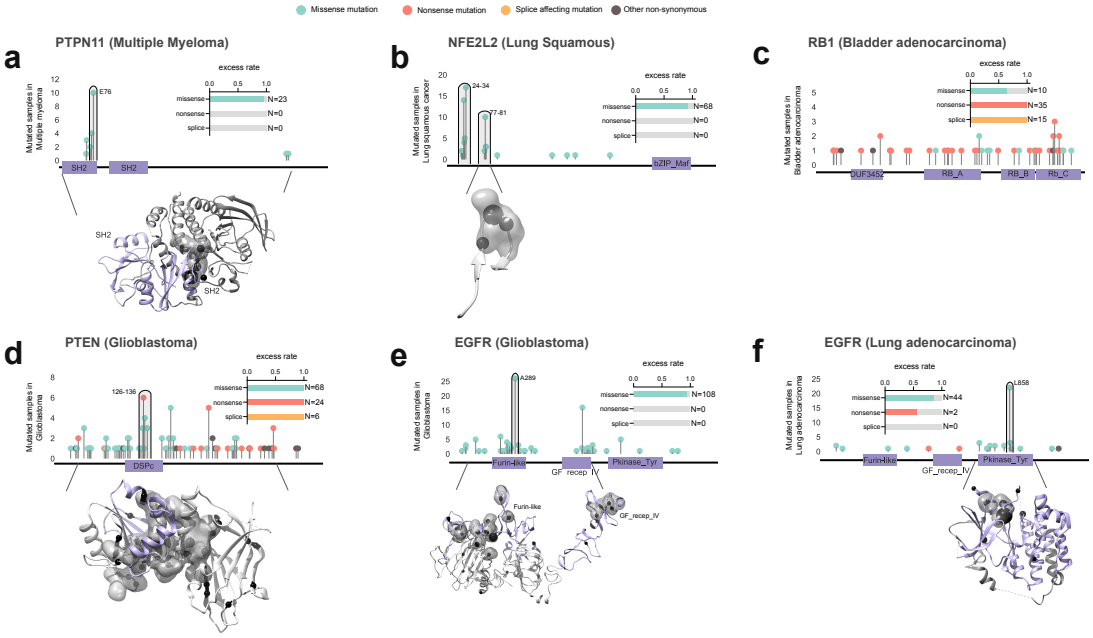

Missense mutation
Nonsense mutation
Splice affecting mutation
Other non-synonymous
a
PTPN11 (Multiple Myeloma)
b
NFE2L2 (Lung Squamous)
c
RB1 (Bladder adenocarcinoma)
d
PTEN (Glioblastoma)
e
EGFR (Glioblastoma)
f
EGFR (Lung adenocarcinoma)
excess rate
SH2
bZIP_Maf
DUF3452
RB_A
RB_B
Rb_C
DSPc
Furin-like
GF_recep_IV
Pkinase_Tyr

Figure 5

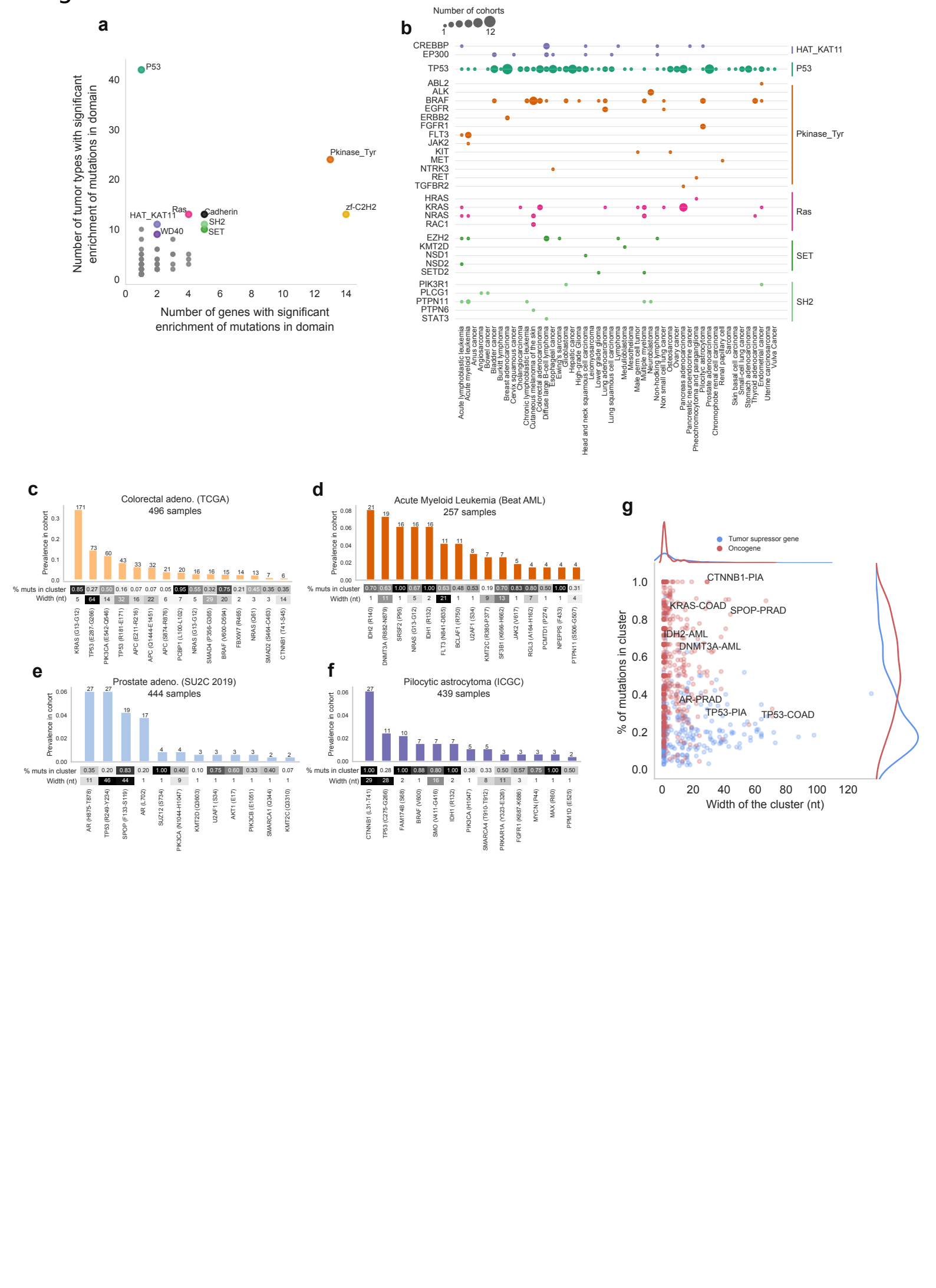
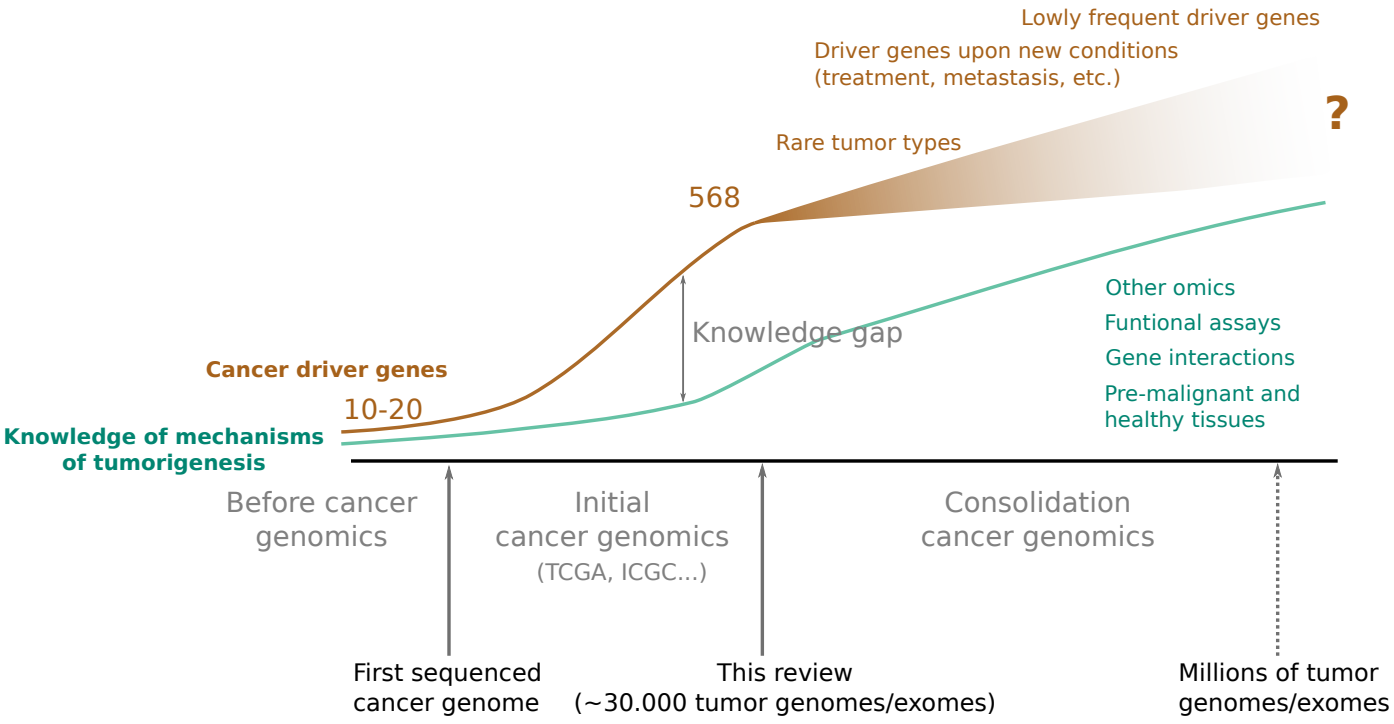

a
Number of tumor types with significant enrichment of mutations in domain
Number of genes with significant enrichment of mutations in domain
P53
Pkinase_Tyr
zf-C2H2
HAT_KAT11
Ras
Cadherin
SH2
SET
WD40
b
Number of cohorts
1
12
CREBBP
EP300
TP53
ABL2
ALK
BRAF
EGFR
ERBB2
FGFR1
FLT3
JAK2
KIT
MET
NTRK3
RET
TGFBR2
HRAS
KRAS
NRAS
RAC1
EZH2
KMT2D
NSD1
NSD2
SETD2
PIK3R1
PLCG1
PTPN11
PTPN6
STAT3
HAT_KAT11
P53
Pkinase_Tyr
Ras
SET
SH2
c
Colorectal adeno. (TCGA)
496 samples
Prevalence in cohort
% muts in cluster
Width (nt)
KRAS (G13-G12)
TP53 (E287-G286)
PIK3CA (E542-Q546)
TP53 (R181-E171)
APC (E211-R216)
APC (Q1444-E1451)
APC (S874-R876)
PCBP1 (L100-L102)
NRAS (G13-G12)
SMAD4 (P356-G365)
BRAF (V600-D594)
FBXW7 (R465)
NRAS (Q61)
SMAD2 (S464-C465)
CTNNB1 (T41-S45)
d
Acute Myeloid Leukemia (Beat AML)
257 samples
Prevalence in cohort
% muts in cluster
Width (nt)
IDH2 (R140)
DNMT3A (R882-N879)
SRSF2 (P95)
NRAS (G13-G12)
IDH1 (R132)
FLT3 (N841-D835)
BCLAF1 (R750)
U2AF1 (S34)
KMT2C (R380-P377)
SF3B1 (K666-H662)
JAK2 (V617)
RGL3 (A184-H162)
PCMTD1 (P274)
NPEPPS (F433)
PTPN11 (S506-G507)
e
Prostate adeno. (SU2C 2019)
444 samples
Prevalence in cohort
% muts in cluster
Width (nt)
AR (H875-T878)
TP53 (R249-Y234)
SPOP (F133-S119)
AR (L702)
SUZ12 (S734)
PIK3CA (N1044-H1047)
KMT2D (Q3803)
U2AF1 (S34)
AKT1 (E17)
PIK3CB (E1051)
SMARCA1 (Q344)
KMT2C (Q3310)
f
Pilocytic astrocytoma (ICGC)
439 samples
Prevalence in cohort
% muts in cluster
Width (nt)
CTNNB1 (L31-T41)
TP53 (C275-G266)
FAM174B (S68)
BRAF (V600)
SMO (V411-G416)
IDH1 (R132)
PIK3CA (H1047)
SMARCA4 (T910-T912)
PRKAR1A (Y323-E326)
FGFR1 (K687-K686)
MYCN (P44)
MAX (R60)
PPM1D (E525)
g
Tumor supressor gene
Oncogene
% of mutations in cluster
Width of the cluster (nt)
CTNNB1-PIA
KRAS-COAD
SPOP-PRAD
IDH2-AML
DNMT3A-AML
AR-PRAD
TP53-PIA
TP53-COAD

Figure 6


Lowly frequent driver genes
Driver genes upon new conditions
(treatment, metastasis, etc.)
?
Rare tumor types
568
Knowledge gap
Other omics
Funtional assays
Gene interactions
Pre-malignant and
healthy tissues
Cancer driver genes
10-20
Knowledge of mechanisms
of tumorigenesis
Before cancer
genomics
Initial
cancer genomics
(TCGA, ICGC...)
Consolidation
cancer genomics
First sequenced
cancer genome
This review
(~30.000 tumor genomes/exomes)
Millions of tumor
genomes/exomes

Supplementary Information

This document contains Supplementary Information to Martinez-Jimenez *et al*., *Nat. Rev. Cancer*, 2020, and is composed of three main sections. The first, a document of **Supplementary Methods** to the main manuscript contains technical details of the development of the IntOGen pipeline and its application to collected and annotated datasets of tumor somatic mutations from the public domain. Secondly, two **Supplementary Figures** illustrate specific aspects of the IntOGen pipeline, i.e., the combination of the output of driver identification methods and the comparison of the performance of this combination with that of individual driver identification methods. Finally, a **Supplementary Table** lists relevant information on the cohorts employed to produce the snapshot of the compendium of mutational cancer genes that is described in the main manuscript.

Table of Contents

Supplementary Methods

Data collection and annotation

TCGA

TCGA somatic mutations (mc3.v0.2.8 version) were downloaded from (https://gdc.cancer.gov/about-data/publications/pancanatlas). We then grouped mutations according to their patient's cancer type into 32 different cohorts. Additionally, we kept somatic mutations passing the somatic filtering from TCGA (i.e., column FILTER == "PASS").

PCAWG

PCAWG somatic mutations were downloaded from the International Cancer Genome Consortium (ICGC) data portal (https://dcc.icgc.org/releases/PCAWG/consensus_snv_indel/). Note that only mutations of ICGC samples can be freely downloaded from this site. The TCGA portion of the callsets is controlled data. To obtain them, we followed the instructions to dowload them that can be found in the same webpage.

cBioPortal

Somatic mutations from Whole Exome Sequencing (WXS) and Whole Genome Sequencing (WGS) cohorts uploaded into cBioPortal that were not part of any other projects included in the analysis (i.e., TCGA, PCAWG, St. Jude or HARTWIG) were downloaded on 2020/01/15 (http://www.cbioportal.org/datasets). We then created cohorts following these criteria:

1. Cohorts with a limited number of samples (i.e., lower than 30 samples) associated to cancer types with extensive representation (such as Breast cancer, Prostate cancer or Colorectal adenocarcinoma) across the compendium of cohorts were removed.
2. Samples were uniquely mapped to a cohort. If the same sample was originally included in two cohorts, we removed the sample from one of them.
3. Mutations from samples that were not obtained from human tumor biopsies were discarded (cell lines, xenografts, normal tissue, etc.).
4. When patient information was available, only one sample of each patient was selected. The criteria to prioritize samples from the same patient was: WXS over WGS; untreated over treated, primary over metastasis or relapse and, finally, by alphabetical order. When there is no patient information we assumed that all patients have only one sample in the cohort.
5. When sequencing platform information was available, samples from the same study but with different sequencing platforms were further subclassified into WXS and WGS datasets (only if the resulting cohorts fulfilled the requirements herein described; otherwise, the samples were discarded).
6. When variant calling information was available, samples from the same cohort and sequencing type were further classified according to their calling algorithm (VarScan, MuTect, etc.). If the resulting cohorts for each subclass fulfilled the requirements herein described, the samples were included; otherwise, the samples were discarded. When variant calling information was not available we assumed that all the samples went through the same calling pipeline.
7. When treatment information was available, samples from the same cohort, sequencing type, calling algorithm were further classified according to their treatment status (i.e, treated versus untreated). If the resulting cohorts from the subclassification fulfilled the requirements herein described, the samples were included; otherwise, the samples were discarded. When information was not available we assumed that samples had not been treated.

8. When biopsy information was available, samples from the same cohort, sequencing type, calling algorithm, treatment status were further classified according to their biopsy type (i.e, primary, relapse or metastasis). If the resulting datasets from the subclassification fulfilled the requirements herein described, the samples were included; otherwise, the samples were discarded. When information was not available we assumed that the biopsy type of the sample was primary.

Hartwig Medical Foundation

Somatic mutations of metastatic WGS from Hartwig Medical Foundation https://www.hartwigmedicalfoundation.nl/en/database/ were downloaded on 2020/01/17 through their platform. Datasets were split according to their primary site. Samples from unknown primary sites (i.e., None, Nan, Unknown, Cup, Na), double primary or aggregating into cohorts of fewer than 7 samples were not considered. A total of 30 different cohorts were thus created.

ICGC

Somatic mutations from Whole Exome Sequencing (WXS) and Whole Genome Sequencing (WGS) studies uploaded in ICGC Data Portal (https://dcc.icgc.org/repositories) not overlapping with other projects included in the analysis (i.e., TCGA, PCAWG, CBIOP or St. Jude) were downloaded on 2018/01/09. We then created cohorts following the criteria used for the cBioPortal datasets (cBioPortal).

St. Jude

Somatic mutations from Whole Exome Sequencing (WXS) and Whole Genome Sequencing (WGS) of Pediatric Cancer Genome Project uploaded in the St. Jude Cloud

(https://www.stjude.cloud/data.html) were downloaded on 2018/07/16. Cohorts were created according to their primary site and their biopsy type (i.e., primary, metastasis and relapse). Resulting datasets with fewer than 5 samples were discarded.

PedcBioPortal

Somatic mutations from Whole Exome Sequencing (WXS) and Whole Genome Sequencing (WGS) studies uploaded in PedcBioPortal that were not part of any other projects included in the analysis (i.e., St. Jude or CBIOP) were downloaded on 2020/01/15 (http://www.pedcbioportal.org/datasets). We then created cohorts following the criteria described in the cBioPortal dataset (cBioPortal).

TARGET

Somatic SNVs from WXS and WGS of two TARGET studies, Neuroblastoma (NB) and Wilms Tumor (WT), from the TARGET consortium were downloaded on 2019/03/07 from the Genomic Data Commons Porta (https://gdc.cancer.gov/).

Beat AML

We downloaded unfiltered somatic mutations from samples included in the Beat AML study from the Genomic Data Commons Porta (https://gdc.cancer.gov/). We next applied the following criteria to create our Beat AML cohort:

1. We focused on somatic single nucleotide variants from VarScan2 using skin as normal control. All samples that did not belong to this class were discarded.
2. Samples from relapses were filtered out.
3. Samples from bone-marrow transplants were discarded.
4. If there were several samples per patient fulfilling the points 1-3, we selected the first in chronological order.

257 independent samples of Beat AML tumors composed our Beat AML cohort.

Literature

We also manually collected publicly available cohorts from the literature. Each cohort was filtered following the same steps mentioned above for the cBioPortal dataset (see above).

Preprocessing

Given the heterogeneity of the datasets analyzed in the current release of intOGen (resulting from differences in the genome aligners, variant calling algorithms, sequencing coverage, sequencing strategy, etc.), we implemented a pre-processing strategy aiming at reducing possible biases. Specifically, we conducted the following filtering steps:

1. The pipeline is configured to run using GRCh38 as reference genome. Therefore, for each input dataset the pipeline requires that the reference genome is defined. Datasets using GRCh37 as reference genome were lifted over using PyLiftover (https://pypi.org/project/pyliftover/; version 0.3) to GRCh38. Mutations failing to liftover from GRCh37 to GRCh38 were discarded.
2. We removed mutations with equal alternate and reference alleles, duplicated mutations within the sample sample, mutations with 'N' as reference or alternative allele, mutations with a reference allele not matching the nucleotide in the reference genome and mutations outside autosomes or sexual chromosomes.
3. Additionally, we removed mutations with low pileup mappability, i.e. mutations in regions that could potentially map elsewhere in the genome. For each position of the genome we computed the pileup mappability, defined as the average uniqueness of all the possible reads of 100bp overlapping a position and allowing up to 2 mismatches. This value is equal to 1 if all the reads overlapping a

mutation are uniquely mappable while it is close to 0 if most mapping reads can map elsewhere in the genome. Positions with pileup mappability lower than 0.9 were removed from further analyses.

4. We filtered out multiple samples from the same donor. The analysis of positive selection in tumors requires that each sample in a cohort is independent from the other samples. That implies that if the input dataset includes multiple samples from the same patient –resulting from different biopsy sites, time points or sequencing strategies– the pipeline automatically selects the first according to its alphabetical order. Therefore, all mutations in the discarded samples are not considered anymore.
5. We also filtered out hypermutated samples. WXS samples carrying more than 1000 mutations or WGS samples with more than 10000 mutations were filtered out if they also exhibited a mutation count greater than 1.5 times the interquartile range above the third quartile of the mutation burden of the cohort were considered hypermutated and therefore removed from further analyses.
6. Datasets without synonymous variants were discarded. Most cancer driver identification methods require synonymous variants to fit a background mutation model. Therefore, datasets with less than 5 synonymous and datasets with a missense/synonymous ratio greater than 10 were excluded .
7. When the Variant Effect Predictor (VEP) mapped one mutation into multiple transcripts associated with different HUGO symbols, we selected the canonical transcript of the first HUGO symbol in alphabetical order.
8. We also discarded mutations mapping into genes without canonical transcript in VEP.92[1].

Methods for cancer driver gene identification

The current version of the intOGen pipeline uses seven cancer driver identification methods to identify cancer driver genes from somatic point mutations: dNdScv[2], cBaSE[3] and MutPanning[4] which test for mutation count bias in genes while correcting for regional genomic covariates[5], mutational processes and coding consequence type; OncodriveCLUSTL[6], which tests for significant clustering of mutations in the protein sequence; smRegions[7], which tests for enrichment of mutations in protein functional domains; HotMAPS[8], which tests for significant clustering of mutations in the 3D protein structure; and OncodriveFML[9], which tests for functional impact bias of the observed mutations. Next, we briefly describe the rationale and the configuration used to run each driver identification method.

dNdScv

dNdScv assesses gene-specific positive selection by inferring the ratio of non-synonymous to synonymous substitutions (dN/dS) in the coding region of each gene. The method resorts to a Poisson-based hierarchical count model that can correct for: i) the mutational processes operative in the cohort determined by the mutational profile of single-nucleotide substitutions with its flanking nucleotides; ii) the regional variability of the background mutation rate explained by histone modifications – it incorporates information about 10 histone marks from 69 cell lines within the ENCODE project[5]; iii) the abundance of sites per coding consequence type across in the coding region of each gene.

We downloaded (release date 2018/10/12) and built a new reference database based on the list canonical transcripts defined by VEP.92 (GRCh38). We then used this

reference database to run dNdScv on all datasets of somatic mutations using the default setting of the method.

OncodriveFML

OncodriveFML is a tool that aims to detect genes under positive selection by analysing the functional impact bias of observed somatic mutations. Briefly, OncodriveFML consists of three steps: in the first step, it computes the average Functional Impact (FI) score (in our pipeline we used CADD[10] v1.4) of coding somatic mutations observed in a gene across a cohort of tumor samples. In the next step, sets of mutations of the same size as the number of mutations observed in the gene of interest are randomly sampled following trinucleotide context conditional probabilities consistent with the relative frequencies of the mutational profile of the cohort. This sampling is repeated N times (with $N = 10^6$ in our configuration) to generate expected average scores across all mutated genes. Finally, it compares the observed average FI score with the distribution expected from the simulations in the form of an empirical p-value. The p-values are then adjusted with a multiple testing correction using the Benjamini–Hochberg (FDR).

OncodriveCLUSTL

OncodriveCLUSTL is a sequence-based clustering algorithm to detect significant linear clustering bias of the observed somatic mutations. Briefly, OncodriveCLUSTL first maps somatic single nucleotide variants (SNVs) observed in a cohort to the genomic element under study. After smoothing the mutation count per position along its genomic sequence using a Tukey kernel-based density function, clusters are identified and scored taking into account the number and distribution of mutations observed. A score for each genomic element is obtained by adding up the scores of its clusters. To estimate the significance of the observed clustering signals, mutations are locally randomized using tri- or penta-nucleotide context conditional probabilities consistent with the relative frequencies of the mutational profile of the cohort.

Within the IntOGen pipeline, OncodriveCLUSTL version 1.1.2 is run for the set of defined canonical transcripts bearing 2 or more SNVs mapping the mutations file. Tuckey-based smoothing is conducted with 11bp windows. The different consecutive coding sequences contained on each transcript are concatenated to allow the algorithm to detect clusters of 2 or more SNVs expanding two exons in a transcript. Simulations are carried out using pre-computed mutational profiles. All cohorts are run using tri-nucleotide context SNVs profiles except for cutaneous melanomas, where penta-nucleotide profiles are calculated. Default randomization windows of 31bp are not allowed to expand beyond the coding sequence boundaries (e.g., windows overlapping part of an exon and an intron are shifted to fit inside the exon). A total number of $N = 10^3$ simulations per transcript are performed. Clustering signals are assessed using analytical p-values.

cBaSE

cBaSE asserts gene-specific positive and negative selection by measuring mutation count bias with Poisson-based hierarchical models. The method allows six different models based on distinct prior alternatives for the distribution of the regional mutation rate. As in the case of dNdScv, the method allows for correction by i) the mutational processes operative in the tumor, with either tri- or penta- nucleotide context; ii) the site count per consequence type per gene; iii) regional variability of the neutral mutation rate.

We run a modified version of the cBaSE script to fit the specific needs of our pipeline. The main modification is adding a rule to automatically select a regional mutation rate prior distribution. Based on the total mutation burden in the dataset, the method runs either an inverse-gamma (mutation count < 12,000), an exponential-inverse-gamma mixture (12,000 < mutation count < 65,000) or a gamma-inverse-gamma mixture (mutation count > 65,000) as mutation rate prior distributions – after communication with

Donate Weghorn, cBaSE's first author). We also skip the negative selection analysis part, as it is not needed for downstream analyses.

Mutpanning

Mutpanning resorts to a mixture signal of positive selection based on two components: i) the mutational recurrence realized as a Poisson-based count model reminiscent to the models implemented at dNdScv or cBaSE; ii) a measure of deviance from the characteristic tri-nucleotide contexts observed in neutral mutagenesis; specifically, an account of the likelihood that a prescribed count of non-synonymous mutations occur in their observed given a context-dependent mutational likelihood attributable to the neutral mutagenesis.

HotMaps3D

HotMAPS asserts gene-specific positive selection by measuring the spatial clustering of mutations in the 3D structure of the protein. The original HotMAPS method assumes that all amino acid substitutions in a protein structure are equally likely. We employed HotMAPS-1.1.3 and modified it to incorporate a background model that more accurately represents the mutational processes operative in a cohort of tumors.

Specifically, we implemented a modified version of the method where the trinucleotide context probability of mutation is compatible with the mutational processes operative in the cohort. Briefly, for each analyzed protein structure harbouring missense mutations, the same number of simulated mutations are randomly generated within the protein structure considering the precomputed mutation frequencies per tri-nucleotide in the cohort. This randomization is performed N times ($N = 10^5$ in our configuration) thereby leading to a background with which to compare the observed mutational data. The rest of the HotMAPS algorithm was not modified.

We downloaded the pre-computed mapping of GRCh37 coordinates into structure residues from the Protein Data Bank (PDB) (http://karchinlab.org/data/HotMAPS/mupit_modbase.sql.gz). We also downloaded (on 2019/09/20) all protein structures from the PDB alongside all human protein 3D models from Modeller (ftp://salilab.org/databases/modbase/projects/genomes/H_sapiens/2013/H_sapiens_2013.tar.xz). and (ftp://salilab.org/databases/modbase/projects/genomes/H_sapiens/2013/ModBase_H_sapiens_2013_refseq.tar.xz). We then annotated the structures following the steps described in HotMAPS tutorial (https://github.com/KarchinLab/HotMAPS/wiki/Tutorial-(Exome-scale)).

Since HotMAPS configuration files are pre-built in GRCh37 coordinates and our pipeline is designed to run using GRCh38, for each input cohort, we first converted input somatic mutations to GRCh37, executed the HotMAPS algorithm and transformed the output to coordinates to GRCh38. All conversions were done using the PyLiftover tool (https://pypi.org/project/pyliftover/).

smRegions

SmRegions is a method developed to detect linear enrichment of somatic mutations in user-defined regions of interest. Briefly, smRegions first counts the number of non-synonymous mutations overlapping a Pfam domain in a particular protein. Next, these non-synonymous variants are randomized N times (N = 10^3 in our configuration) along the nucleotide sequence of the gene, following the trinucleotide context probability derived from precomputed mutation frequencies per tri-nucleotide in the cohort. The observed and average number of simulated mutations in the Pfam domain and outside of it are compared using a G-test of goodness-of-fit, from which the smRegions p-value is derived. Within the IntOGen pipeline, smRegions discards domains with a number of observed mutations lower than the average from the randomizations. The p-values are adjusted with a multiple testing correction using the Benjamini–Hochberg procedure.

Therefore, the analysis is confined to Pfam domains with a number of observed mutations higher than or equal to the mean simulated number of mutations in the re-sampling.

To create the database of genomic coordinates of Pfam domains we followed the next steps: i) we gathered the first and last amino acid positions of all Pfam domains for canonical transcripts (VEP.92) from BioMart; ii) for each Pfam domain we mapped the first and last amino acid positions into genomic coordinates using TransVar –using GRCh38 as reference genome–; iii) we discarded Pfam domains failing to map either the first or last amino acid positions into genomic coordinates.

smRegions was conceptually inspired by e-driver[11], although significant enhancements to the approach have been introduced. Particularly, i) our background model accounts for the observed tri-nucleotide frequencies rather than assuming that all mutations are equally likely; ii) the statistical test is more conservative; iii) Pfam[12] domains are part of the required input and can be easily updated by downloading the last Pfam release; iv) the method can be configured to any other setting that aims to detect genes possibility selected by enrichment of mutations in pre-defined gene regions.

Combining the outputs of driver identification methods

Rationale

The IntOGen pipeline aims to provide a compendium of driver genes which appropriately reflects the consensus from these seven driver identification methods.

To combine the results of individual statistical tests, p-value combination methods continue to be a standard approach in the field: e.g., Fisher's[13], Brown's[14], and Stouffer's Z-score methods have been used for this purpose. These methods are useful for combining probabilities in meta-analyses, in order to provide a ranking based on combined significance under statistical grounds. However, the application of these methods may bear some caveats:

1. The ranking resulting from p-value combination may lead to inconsistencies when compared to the individual rankings, i.e., they may yield a consensus ranking that does not preserve recurrent precedence relationships found in the individual rankings.
2. Some methods, like Fisher's or Brown's method, may bear anti-conservative performance, thus leading to many likely false discoveries.
3. Balanced (non-weighted) p-value combination methods may lead to faulty results just because of the influence of one or more driver identification method performing poorly for a given dataset.

Weighted methods to combine p-values, like the weighted Stouffer's Z-score, provide some extra room for proper balancing, in the sense of incorporating the relative credibility of each driver identification method. We reasoned that in the context of the combination of the output of driver identification methods, the allocation of weights should account for differences in credibility between methods and across cohorts.

Our combination approach works independently for each cohort. To create a consensus list of driver genes for each cohort, we first determine how credible each driver identification method is when applied to this specific cohort (see Supplementary Figure 1 for a representation of the combinatorial workflow). We do so by tuning a voting weight for each driver identification method that yields a good enrichment of bona-fide cancer genes -- reported in the COSMIC Cancer Gene Census database[15] (CGC) -- in the highly ranked positions of the resulting consensus ranking upon letting each driver

identification method vote. Once a credibility score has been established, we use a weighted method for combining the p-values that each driver identification method gives for each candidate gene: this combination takes the driver identification methods credibility into account. Based on the combined p-values, we conduct FDR correction to conclude a ranking of candidate driver genes alongside q-values.

Weight Estimation by Voting

The relative credibility awarded to each method is based on the ability of the method to give precedence to well-known genes already collected in the CGC catalog of validated driver genes. As each method yields a ranking of driver genes, these lists can be combined using a voting system –Schulze's voting method. The method allows us to consider each method as a voter with some voting rights (weighting) which casts ballots containing a list of candidates sorted by precedence. Schulze's method takes information about precedence from each individual method and produces a new consensus ranking[16].

Instead of conducting balanced voting, we tune the voting rights of the methods so that the enrichment of CGC genes at the top positions of the consensus list is maximized. We limit the share each method can attain in the credibility simplex –up to a uniform threshold. The resulting voting rights are deemed the relative credibility of each method.

Ranking Score

Upon selection of a catalog of bona-fide known driver elements (the Cancer Gene Census, or CGC) we can provide a score for each ranking R of genes as follows:

$$E(R) = \sum_{i=1}^{N} \frac{p_i}{\log(i+1)}$$

where p(i) is the proportion of elements with rank smaller (closer to top) than i which belong to CGC and N is a suitable threshold to consider only the N top ranked

elements. Using E(R) we can define a function that maps each weighting vector w (in the simplex of methods weights) to a value E(R(w)) where R(w) denotes the consensus ranking obtained by applying Schulze's voting with voting rights given by the weighting vector w.

Optimization with constraints

Finally we are bound to find a good candidate for

$$\hat{w} = argmax\ E(R(w))$$

For each method to have chances to contribute to the consensus score, we impose the mild constraint of limiting the share of each method to 0.3.

Optimization is then carried out in two steps: we first find a good candidate $\hat{w}_0$ by exhaustive search in a rectangular grid satisfying the constraints defined above (with grid step=0.05); in the second step we take $\hat{w}_0$ as the seed for a stochastic hill-climbing procedure (we resort to Python's scipy.optimize "basinhopping", method=SLSQP and stepsize=0.05).

Estimation of combined p-values using weighted Stouffer's Z-score

Using the relative weight estimate that yields a maximum value of the objective function *f* we combined the p-values resorting to the weighted Stouffer's Z-score method. Thereafter we performed Benjamini-Hochberg FDR correction with the resulting combined p-values, yielding one q-value for each genomic element. If the element

belongs to the CGC, we computed its q-value using only the collection of p-values of CGC genes. Otherwise, we computed the q-value using the p-values computed for all genes.

Tiers of driver genes from sorted list of combined rankings and p-values

To finalize the analysis we considered only genes with at least two mutated samples in the cohort under analysis. These genes were classified into four groups according to the level of evidence in that cohort that the gene harbours signal of positive selection.

For the sake of simplicity, we give some conventions before proceeding to describe the groups. For each gene G we have defined a rank r(G) and a significance q-value q(G) according to the voting and p-value combinations described above. Given the final ranked list of genes we can define two rank cutoffs that depend on a prescribed significance level t:

$$R = \min_{G} \{r(G) \mid q(G) > t\} - 1$$

$$r = \max_{G} \{r(G) \mid q(G) < t\}$$

It is readily seen that r < R+1. By default the significance level t is set to 0.05.

1. The first group of genes, TIER1, contains genes showing high confidence and agreement in their positive selection signals. TIER1 comprises all the genes G such that r(G) < r.

2. The second group, TIER2, was devised to contain known cancer driver genes, showing mild signals of positive selection, that were not included in TIER1. More in detail, we defined TIER2 genes as those CGC genes, not included in TIER1, whose CGC q-value was lower than a prescribed significance level (default CGC q-value=0.25). The CGC q-value is computed by performing FDR of the combined p-values albeit restricted to CGC genes.
3. The third group, TIER3, encompasses genes G that are not included in TIER1 or TIER2 which fulfill that r(G) < R.
4. All genes not included in the aforementioned classes are considered non-driver genes.

Combination benchmark

We have assessed the performance of the combination compared to i) each of the seven individual methods and ii) other strategies to combine the output from cancer driver identification methods.

Finally, we evaluated the contribution of each of the individual methods to the consensus list of driver genes.

Datasets for evaluation

We decided to perform an evaluation based on the 32 Whole-Exome cohorts of the TCGA PanCanAtlas project (downloaded from *https://gdc.cancer.gov/about-data/publications/pancanatlas*). These cohorts sequence coverage, sequence alignment and somatic mutation calling were performed using the same methodology guaranteeing that biases due to technological and methodological artifacts are minimal.

The Cancer Genes Census –version v87– was downloaded from the COSMIC data portal (*https://cancer.sanger.ac.uk/census*) and used as a positive set of known cancer driver genes.

We created a catalog of genes with evidence of not involvement in cancerogenesis. This set includes very long genes (HMCN1, TTN, OBSCN, GPR98, RYR2 and RYR3), and a list of olfactory receptors from Human Olfactory Receptor Data Exploratorium (HORDE) (https://genome.weizmann.ac.il/horde/; download date 14/02/2018). In addition, for all TCGA cohorts, we added non-expressed genes, defined as genes where at least 80% of the samples showed a RSEM expressed in $\log_2$ scale smaller or equal to 0. Expression data for TCGA was downloaded from *https://gdc.cancer.gov/about-data/publications/pancanatlas*.

Metrics for evaluation

We defined a metric, referred to as CGC-Score, that is intended to measure the quality of a ranking of genes as the enrichment of CGC elements in the top positions of the ranking; specifically given a ranking R mapping each element to a rank, we define the CGC-Score of R as:

$$S(R) = \sum_{i=1}^{N} \frac{p(i)}{\log(i+1)} \Big/ \sum_{i=1}^{N} \frac{1}{\log(i+1)}$$

where p(i) is the proportion of elements with rank $\leq i$ that belong to CGC and N is a convenient threshold to consider just the top elements in the ranking (by default N=40). We estimated the CGC-Score across TCGA cohorts for all the rankings given by individual methods and by the consensus ranking.

Similarly, we defined a metric, referred to as Negative-Score, that aims to measure the proportion of non-cancer genes among the top positions in the ranking. Specifically, given a ranking R, we define the Negative-Score of R as:

$$N(R) = \sum_{i=1}^{N} \frac{n(i)}{\log(i+1)} \; / \; \sum_{i=1}^{N} \frac{1}{\log(i+1)}$$

where n(i) is the proportion of elements with rank $\leq i$ that belong to the negative set and N is a suitable threshold to consider just the top elements in the ranking (by default N = 40). We estimated the Negative-Score across TCGA cohorts for all the rankings given by individual methods and by the consensus ranking.

Comparison with individual methods

We compared the CGC-Score and Negative-Score of the combined lists of drivers with the individual outputs of the seven driver discovery methods integrated in the pipeline.

We observed a consistent increase in CGC-Score of the combinatorial strategy compared to any individual method across 23/32 (71%) of the TCGA cohorts (Supplementary Figure 2a and 2b). Similarly, we observed a consistent decrease in Negative-Score across TCGA cohorts, where the combinatorial strategy ranked the least enriched in non-cancer genes in 14 (43%) cohorts and in none of them was the most enriched in non-cancer genes (Supplementary Figure 2c).

In summary, the evaluation shows that the combinatorial strategy increases the True Positive Rate (measured using the CGC-Score) preserving lower False Positive Rate (measured using the Negative-Score) than the seven individual methods included in the pipeline.

Comparison with other combinatorial selection methods

We then computed the CGC-Score and Negative-Score based on the consensus ranking from the aforementioned combinatorial methods and compared them to our Schulze's weighted combination ranking across all TCGA cohorts. Our combinatorial approach met a larger enrichment in known cancer genes for 30/32 (93%) TCGA cohorts (Supplementary Figure 2d).

Leave-one-out combination

We aimed to estimate the contribution of each method's ranking to the final ranking after Schulze's weighted combination. To tackle this question, we measured the effect of removing one method from the combination by, first, calculating the CGC-Score of the combination excluding the aforementioned method and, next, obtaining its ratio with the original combination (i.e., including all methods). This was iteratively calculated for all methods across TCGA cohorts. Methods that positively contributed to the combined ranking quality show a ratio below one, while methods that negatively contributed to the combined ranking show a ratio greater than one.

We observed that across TCGA cohorts most of the methods contributed positively (i.e., ratio above one) to the weighted combination performance (Supplementary Figure 2e). Moreover, there is substantial variability across TCGA cohorts in the contribution of each method to the combination performance. In summary, all methods contributed positively to the combinatorial performance across TCGA supporting our methodological choice for the individual driver discovery methods (Supplementary Figure 2e).

Drivers postprocessing

The intOGen pipeline outputs a ranked list of driver genes for each input cohort. We aimed to create a comprehensive compendium of driver genes per tumor type from all the cohorts included in this version.

Then, we performed a filtering on automatically generated driver gene lists per cohort. This filtering is intended to reduce artifacts from the cohort-specific driver lists, due to e.g. errors in calling algorithms, local hypermutation effects, undocumented filtering of mutations.

We first created a collection of candidate driver genes by selecting either: i) significant non-CGC genes (q-value < 0.05) with at least two significant bidders (methods rendering the genes as significant); ii) significant CGC genes (either q-value < 0.05 or CGC q-value < 0.25) from individual cohorts. All genes that did not fulfill these requirements were discarded.

Additionally, candidate driver genes were further filtered using the following criteria:

1. We discarded non-expressed genes using TCGA expression data. For tumor types directly mapping to cohorts from TCGA –including TCGA cohorts– we removed non-expressed genes in that tumor type. We used the following criterion for non-expressed genes: genes where at least 80% of the samples showed a negative log2 RSEM. For those tumor types which could not be mapped to TCGA cohorts this filtering step was not done.
2. We also discarded genes highly tolerant to Single Nucleotide Polymorphisms (SNP) across human populations. Such genes are more susceptible to calling errors and should be taken cautiously. More specifically, we downloaded transcript specific constraints from gnomAD (release 2.1; 2018/02/14) and used the observed-to-expected ratio score (oe) of missense (mys), synonymous (syn) and loss-of-function (lof) variants to detect genes highly tolerant to SNPs. Genes enriched in SNPs (oe_mys > 1.5 or oe_lof > 1.5 or oe_syn > 1.5) with a number of mutations per sample greater than 1 were discarded. Additionally, we discarded mutations overlapping with germline variants (germline count > 5) from a panel of normals (PON) from Hartwig Medical Foundation

(https://nextcloud.hartwigmedicalfoundation.nl/s/LTiKTd8XxBqwaiC?path=%2FHMFTools-Resources%2FSage).

3. We also discarded genes that are likely false positives according to their known function from the literature. We convened that the following genes are likely false positives: i) known long genes such as TTN, OBSCN, RYR2, etc.; ii) olfactory receptors from HORDE (http://bioportal.weizmann.ac.il/HORDE/; download date 2018/02/14); iii) genes not belonging to Tier1 CGC genes lacking literature references according to CancerMine[17] (http://bionlp.bcgsc.ca/cancermine/).
4. We also removed non CGC genes with more than 3 mutations in one sample. This abnormally high number of mutations in a sample may be the result of either a local hypermutation process or cross contamination from germline variants.
5. Finally we discarded genes whose mutations are likely the result of local hypermutation activity. More specifically, some coding regions might be the target of mutations associated with COSMIC Signature 9 (https://cancer.sanger.ac.uk/cosmic/signatures) which is associated with non-canonical AID activity in lymphoid tumours. In those cancer types were Signature 9 is known to play a significant mutagenic role (i.e., AML, Non-Hodgkin Lymphomas, B-cell Lymphomas, CLL and Myelodysplastic syndromes) we discarded genes where more than 50% of mutations in a cohort of patients were associated with Signature 9.

Candidate driver genes that were not discarded composed the compendium of driver genes.

Classification according to annotation level from CGC

We then annotated the catalog of highly confident driver genes according to their annotation level in CGC version 87. Specifically, we created a three-level annotation: i) the first level included driver genes with a reported involvement in the source tumor type according to the CGC; ii) the second group included CGC genes lacking reported

association with the tumor type; iii) the third group included genes that were not present in CGC.

To match the tumor type of our analyzed cohorts and the nomenclature/acronyms of cancer types reported in the CGC we manually created a dictionary comprising all the names of tumor types from CGC and cancer types defined in our study, according to the following rules:

1. All the equivalent terms for a cancer type reported in the CGC using the Somatic Tumor Type field (e.g. “breast”, “breast carcinoma”, “breast cancer”), were mapped into the same tumor type.
2. CGC terms with an unequivocal mapping into our cancer types were automatically linked (e.g., “breast” with “BRCA”).
3. CGC terms representing fine tuning classification of a more prevalent cancer type that did not represent an independent cohort in our study, were mapped to their closest parent tumor type in our study (e.g., “malignant melanoma of soft parts” into “cutaneous melanoma” or “alveolar soft part sarcoma” into “sarcoma”).
4. Adenomas were mapped to carcinomas of the same cell type (e.g.,”hepatic adenoma” into “hepatic adenocarcinoma”, “salivary gland adenoma” into “salivary gland adenocarcinoma”).
5. CGC parent terms mapping to several tumor types from our study were mapped to each of the potential child tumor types. For instance, the term “non small cell lung cancer” was mapped to “LUAD” (lung adenocarcinoma) and “LUSC” (lung squamous cell carcinoma).
6. Finally, CGC terms associated with benign lesions, with unspecified tumor types (e.g., “other”, “other tumor types”, “other CNS”) or with tumor types with missing parents in our study were left unmatched.

Mode of action of driver genes

We computed the mode of action for highly confident driver genes. To do so, we first performed a pan-cancer run of dNdScv across all TCGA cohorts. We then applied the aforementioned algorithm (see Mode of action section below for more details on how the algorithm determines the role of driver genes according to their distribution of mutations in a cohort of samples) to classify driver genes into the three possible roles: Act (activating or oncogene), LoF (loss-of-function or tumor suppressor) or Amb (ambiguous or non-defined). We then combined these predictions with prior knowledge from the Cancer Genome Interpreter[18] according to the following rules: i) when the inferred mode of action matched the prior knowledge, we used the consensus mode of action; ii) when the gene was not included in the prior knowledge list, we selected the inferred mode of action; iii) when the inferred mode of action did not match the prior knowledge, we selected that of the prior knowledge list.

Repository of mutational features

Linear clusters

Linear clusters for each gene and cohort were identified by OncodriveCLUSTL. We defined as significant those clusters in a driver gene with a p-value lower than 0.05. The start and end of the clusters were retrieved from the first and last mutated amino acid overlapping the cluster, respectively.

3D clusters

Information about the positions involved in the 3D clusters defined by HotMAPS were retrieved from the gene specific output of each cohort. We defined as significant those amino acids in a driver gene with a q-value lower than 0.05.

Pfam Domains

Pfam domains for each driver gene and cohort were identified by smRegions. We defined as significant those domains in driver genes with a q-value lower than 0.1 and with positive log ratio of observed-to-simulated mutations (observed mutations / simulated mutations > 1). The first and last amino acids are defined from the start and end of the Pfam domain, respectively.

Excess of mutations

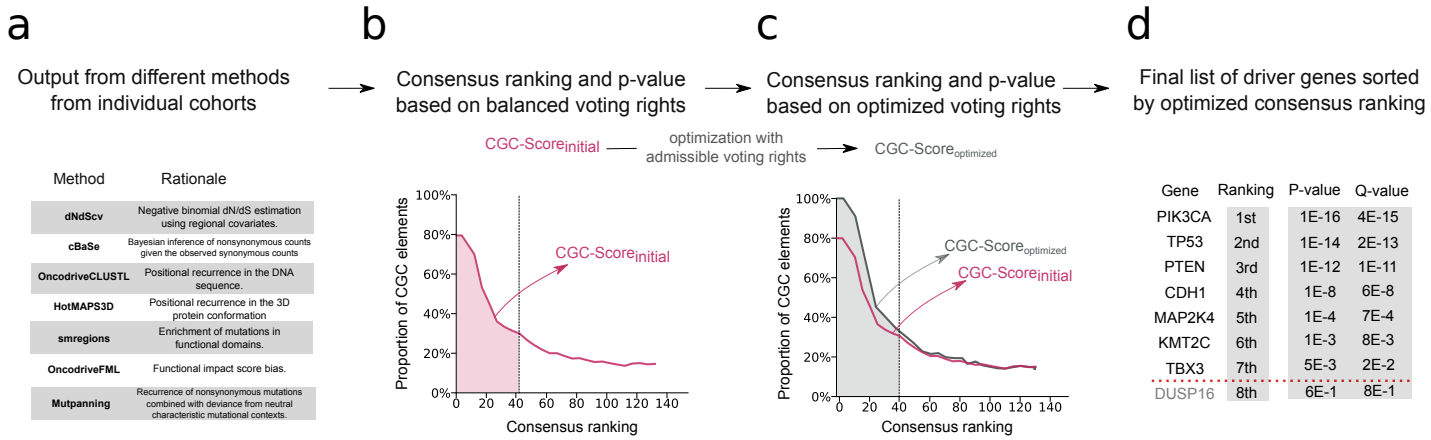
The so-called excess of mutations for a given coding consequence-type quantifies the proportion of observed mutations at this consequence-type that are not explained by the neutral mutation rate. The excess is computed from the consequence-type specific dN/dS estimates ω_c as $(\omega_c - 1) / \omega_c$. We computed the excess for missense, nonsense and splicing-affecting mutations according to the canonical transcript.

Mode of action

Upon the consequence-type specific dN/dS estimates for nonsense and missense mutations computed at each gene, denoted ω_{mis} and ω_{non}, we deemed a gene activating or Act (resp. Loss-of-function or LoF) if $\omega_{mis} - \omega_{non} > \varepsilon$ (resp. $\omega_{non} - \omega_{mis} > \varepsilon$) with $\varepsilon = 0.1$. Genes with $|\omega_{mis} - \omega_{non}| < \varepsilon$ as well as genes with $\omega_{mis} < 1$ were deemed to have an “ambiguous” mode of action.

Supplementary Figures

Figure Supplementary 1



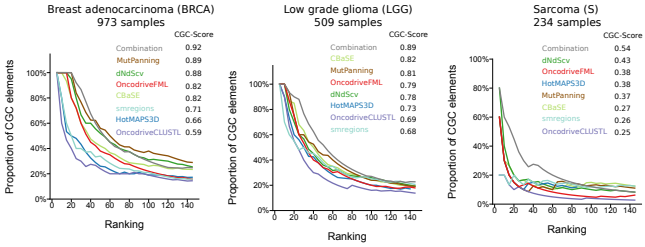
Supplementary Figure 1. Schematic representation of the approach to combine the output of driver discovery methods.

a) Given the output of the seven driver discovery methods integrated in intOGen, b) the pipeline dynamically estimates the credibility of the output of each method based on its enrichment for Cancer Gene Census genes. Then in c) it performs the combination of the outputs weighting each method output according to the credibility previously allocated. Finally in d), the resulting list of drivers is sorted by the optimized consensus ranking and their associated combined p-value.

Figure Supplementary 2

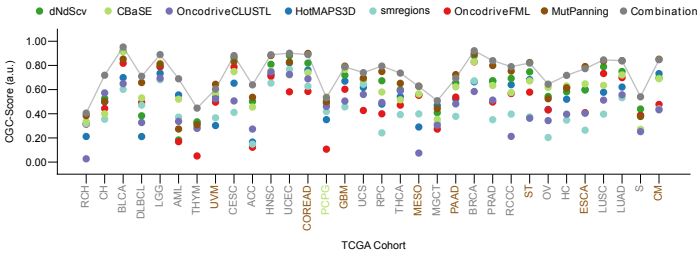
a

The combination shows higher specificity than individual methods



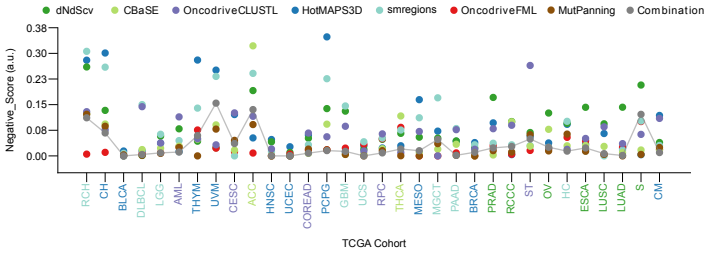
b

The combination generally improves
the specificity of individual methods across TCGA cohorts



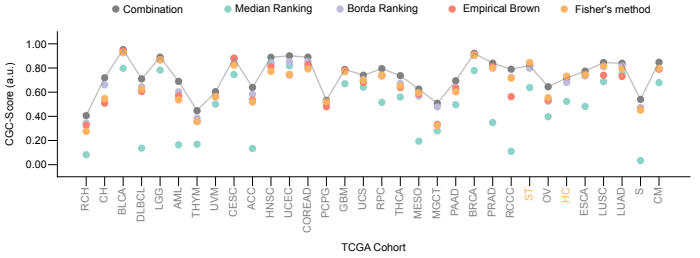
c

The combination keeps low a false positive ratio (measured by non-cancer genes)



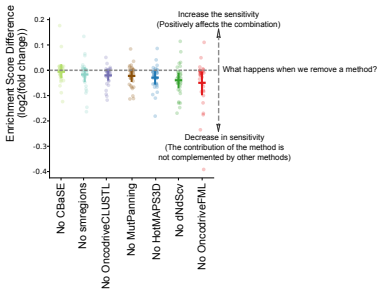
d

The combination improves the sensitivity compared
to other classical combinatorial strategies



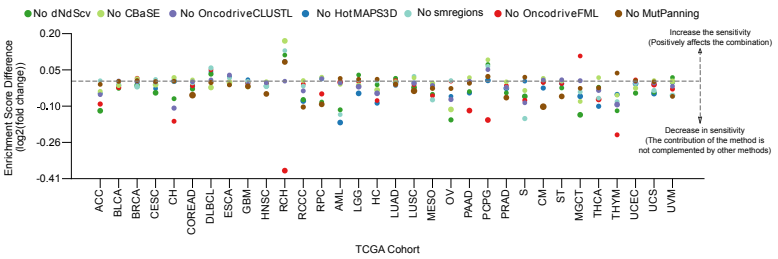
e

Removing invididual methods negatively affects the combination



f

Heterogeneous contribution of individual methods across TCGA cohorts



Supplementary Figure 2. Benchmark of the IntOGen combination using TCGA cohorts.

a) The proportion of CGC drivers among the top ranking genes in the combined list is greater than that of the lists of individual driver identification methods in three exemplary TCGA cohorts (BRCA, LGG and Sarcoma). The proportion of CGC drivers in each list of genes is measured across growing top ranked genes (x-axis). To summarize the proportion of CGC drivers obtained throughout all values of rank tested, a numeric value (CGC score) is derived (see Supplementary Methods).

b) CGC score of the output of all driver discovery methods and the combined list across 32 TCGA cohorts. Systematically, the combined list exhibits a CGC score which is at least equal to that of the best performing individual method. In most cases, the combined list exhibits a higher CGC score than that of any individual method.

c) For any drivers list we can also compute a potential false positives score or Negative Score, tracking the proportion of a set of non driver genes (known “fishy” genes of driver identification, and not expressed genes in each tissue) within the top-ranking elements of the list. The Negative Score of the combined list across all TCGA cohorts is comparable to that of methods with the lowest Negative Score. This means that the increase in sensitivity of drivers identification in the combined list that is documented in a) and b) does not come at the cost of a reduction of specificity.

d) Comparison of the CGC Score of the combined list with that obtained using classic combination strategies across all TCGA cohorts. The combination approach developed in the pipeline exhibits higher sensitivity than any other strategy across all cohorts.

e) To assess the contribution of each individual method to the combined list of drivers, we carried out a systematic leave-one-out analysis across all TCGA cohorts (dots in each distribution). We then evaluated the sensitivity of the new combination using the CGC Score. In most cohorts, the elimination of a method from the combination causes a decrease of sensitivity.

f) The effect of eliminating one method on the sensitivity of the combination changes across cohorts.

Supplementary Table

Supplementary Table 1. Summarized list of cohorts employed to produce the snapshot of the compendium of cancer genes described in the main manuscript.

The list of cohorts collected from the public domain and employed in the construction of subsequent snapshots of the compendium will be updated and published regularly in the IntOGen website (www.intogen.org).

Bibliography

1. McLaren, W. et al. The Ensembl Variant Effect Predictor. Genome Biol. 17, 1–14 (2016).
2. Martincorena, I. *et al.* Universal Patterns of Selection in Cancer and Somatic Tissues. *Cell* **171**, 1029-1041.e21 (2017).
3. Weghorn, D. & Sunyaev, S. Bayesian inference of negative and positive selection in human cancers. *Nat. Genet.* **49**, 1785–1788 (2017).
4. Dietlein, F. *et al.* Identification of cancer driver genes based on nucleotide context. *Nat. Genet.* **52**, 208–218 (2020).
5. Kundaje, A. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
6. Arnedo-Pac, C., Mularoni, L., Muiños, F., Gonzalez-Perez, A. & Lopez-Bigas, N. OncodriveCLUSTL: a sequence-based clustering method to identify cancer drivers. *Bioinformatics* **35**, 4788–4790 (2019).
7. Martínez-Jiménez, F., Muiños, F., López-Arribillaga, E., Lopez-Bigas, N. & Gonzalez-Perez, A. Systematic analysis of alterations in the ubiquitin proteolysis system reveals its contribution to driver mutations in cancer. *Nat. Cancer* **1**, 122–135 (2020).
8. Tokheim, C. *et al.* Exome-Scale Discovery of Hotspot Mutation Regions in Human Cancer Using 3D Protein Structure. *Cancer Res.* **76**, 3719–3731 (2016).
9. Mularoni, L., Sabarinathan, R., Deu-Pons, J., Gonzalez-Perez, A. & López-Bigas, N. OncodriveFML: a general framework to identify coding and non-coding regions with

cancer driver mutations. *Genome Biol.* **17**, 128 (2016).

10. Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J. & Kircher, M. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res.* **47**, D886–D894 (2019).
11. Porta-Pardo, E. & Godzik, A. e-Driver: a novel method to identify protein regions driving cancer. *Bioinforma. Oxf. Engl.* **30**, 3109–3114 (2014).
12. El-Gebali, S. *et al.* The Pfam protein families database in 2019. *Nucleic Acids Res.* **47**, D427–D432 (2019).
13. Mosteller, F. & Fisher, R. A. Questions and Answers. *Am. Stat.* **2**, 30–31 (1948).
14. Brown, M. B. 400: A Method for Combining Non-Independent, One-Sided Tests of Significance. *Biometrics* **31**, 987–992 (1975).
15. Sondka, Z. *et al.* The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. *Nat. Rev. Cancer* **18**, 696–705 (2018).
16. Schulze, M. The Schulze Method of Voting. *ArXiv180402973 Cs* (2019).
17. Lever, J., Zhao, E. Y., Grewal, J., Jones, M. R. & Jones, S. J. M. CancerMine: a literature-mined resource for drivers, oncogenes and tumor suppressors in cancer. *Nat. Methods* **16**, 505–507 (2019).
18. Tamborero, D. *et al.* Cancer Genome Interpreter annotates the biological and clinical relevance of tumor alterations. *Genome Med.* **10**, 25 (2018).

##IntOGen RELEASE 2020/02/01

Index	Column name	Explanation
1	COHORT	Name of the cohort.
2	CANCER_TYPE	Acronym of the cancer type associated with the cohort.
3	CANCER_TYPE_NAME	Long name of the cancer type associated with the cohort.
4	SOURCE	Source of the data (TCGA, PCAWG, Hartwig Medical Foundation, cBioPortal, PedCbioPortal, ICGC, etc.).
5	PLATFORM	Whole-exome sequencing (WXS) or Whole-genome sequencing (WGS).
6	REFERENCE	Pubmed ID of the publication.
7	TYPE	Type of cohort [“Primary”, “Metastatic”, “Relapse”]
8	TREATED	Whether the cohort of patients has undergone cancer treatment {“Treated”, “Untreated”}.
9	AGE	Age of the cohort, status {“Adult”, “Pediatric”}
10	SAMPLES	Number of samples (prior to any filtering by intOGen).
11	MUTATIONS	Number of total of mutations in the cohort (before filtering by intOGen).
12	WEB_SHORT_COHORT_NAME	Short name in the intOGen website.
13	WEB_LONG_COHORT_NAME	Long name in the intOGen website.

COHORT: