RESEARCH ARTICLE

Use of signals of positive and negative selection to distinguish cancer genes and passenger genes

László Bányai[1], Maria Trexler[1], Krisztina Kerekes[1], Orsolya Csuka[2], László Patthy[1]*

[1]Institute of Enzymology, Research Centre for Natural Sciences, Budapest, Hungary; [2]Department of Pathogenetics, National Institute of Oncology, Budapest, Hungary

Abstract A major goal of cancer genomics is to identify all genes that play critical roles in carcinogenesis. Most approaches focused on genes positively selected for mutations that drive carcinogenesis and neglected the role of negative selection. Some studies have actually concluded that negative selection has no role in cancer evolution. We have re-examined the role of negative selection in tumor evolution through the analysis of the patterns of somatic mutations affecting the coding sequences of human genes. Our analyses have confirmed that tumor suppressor genes are positively selected for inactivating mutations, oncogenes, however, were found to display signals of both negative selection for inactivating mutations and positive selection for activating mutations. Significantly, we have identified numerous human genes that show signs of strong negative selection during tumor evolution, suggesting that their functional integrity is essential for the growth and survival of tumor cells.

***For correspondence:** patthy.laszlo@ttk.mta.hu

Competing interests: The authors declare that no competing interests exist.

Funding: See page 33

Received: 03 June 2020
Accepted: 10 January 2021
Published: 11 January 2021

Reviewing editor: Eduardo Eyras, Australian National University, Australia

Introduction

Genetic, epigenetic, transcriptomic, and proteomic changes driving carcinogenesis

In the last two decades, the rapid advance in genomics, epigenomics, transcriptomics, and proteomics permitted an insight into the molecular basis of carcinogenesis. These studies have confirmed that tumors evolve from normal tissues by acquiring a series of genetic, epigenetic, transcriptomic, and proteomic changes with concomitant alterations in the control of the proliferation, survival, and spread of affected cells.

The genes that play key roles in carcinogenesis are usually assigned to two major categories: proto-oncogenes that have the potential to promote carcinogenesis when activated or overexpressed and tumor suppressor genes (TSGs) that promote carcinogenesis when inactivated or repressed.

Several alternative mechanisms can modify the structure or expression of a gene in a way that promotes carcinogenesis. These include subtle genetic changes (single nucleotide substitutions, short indels), major genetic events (deletion, amplification, translocation and fusion of genes to other genetic elements), as well as epigenetic changes affecting the expression of cancer genes. These mechanisms are not mutually exclusive: there are many examples illustrating the point that multiple types of the above mechanisms may convert the wild-type form of a cancer gene to a driver gene.

Exomic studies of common solid tumors revealed that usually several cancer genes harbor subtle somatic mutations (point mutations, short deletions, and insertions) in their translated regions but malignancy-driving subtle mutations can also occur in all genetic elements outside the coding region, namely in enhancer, silencer, insulator, and promoter regions as well as in 5′- and 3′-

eLife digest The DNA in the cells of the human body is usually copied correctly when a cell divides. However, errors (mutations) are sometimes introduced during the copying process. Although the majority of mutations have no major impact on cells, many mutations are harmful: they decrease the ability of cells to survive. There are, however, mutations that can lead to cells dividing more frequently or gaining the ability to spread, which can lead to cancer. These mutations are known as 'driver mutations' because they drive the growth of tumors. Since such 'driver mutations' provide a growth advantage to tumor cells, they are subject to positive selection, this is, their frequency in the tumor increases over time. Because of their selective advantage, driver mutations accumulate at significantly higher rates than the neutral 'passenger mutations' that do not play a role in tumor growth.

Genes that carry driver mutations are called driver genes, while genes that carry only passenger mutations are known as passenger genes. Certain genes, however, do not fit into either category. For example, some genes that are essential for tumor growth must get rid of harmful mutations to maintain activity. Mutations of such 'tumor essential genes' are thus subject to 'negative' or 'purifying selection'.

A major goal of cancer research is to identify genes that play critical roles in tumor growth. Earlier studies have identified numerous driver genes positively selected for driver mutations, exploiting the fact that driver genes show significantly higher mutation rates than passenger genes. Identification of tumor essential genes, however, is inherently more difficult since the paucity of mutations of negatively selected genes hinders the analysis of the mutation data. The failure to provide convincing evidence for negative selection in tumors has led to suggestions that it has no role in cancer evolution.

Bányai et al. used a novel approach to address the question of whether negative selection occurs in cancer. Based on characteristic differences in the patterns of mutations in cancer they distinguished clusters of passenger genes, driver genes and tumor essential genes. The group of tumor essential genes includes genes that serve to satisfy the increased demand of rapidly dividing tumor cells for nutrients' and genes that are essential for cell migration and metastasis (the spread of cancer cells to other areas of the body).

The tumor essential genes that Bányai et al. identified may prove to be valuable targets for cancer therapy, illustrating the importance of genome sequencing in cancer research. Identification of additional tumor essential genes is, however, hindered by the fact that they are likely to have low levels of mutations, which can exclude them from meaningful analyses. Progress with genomic sequencing of tumors is expected to overcome this limitation and help identify additional genes that are essential for cancer growth.

untranslated regions. Intron or splice site mutations that alter the splicing pattern of cancer genes can also drive carcinogenesis (***Diederichs et al., 2016***). A recent study has presented a comprehensive analysis of driver point mutations in non-coding regions across 2658 cancer genomes (***Rheinbay et al., 2020***). A noteworthy example of how subtle mutations in regulatory regions may activate proto-oncogenes is the telomerase reverse transcriptase gene *TERT* that encodes the catalytic subunit of telomerase. Recurrent somatic mutations in melanoma and other cancers in the *TERT* promoter cause tumor-specific increase of *TERT* expression, resulting in the immortalization of the tumor cell (***Heidenreich et al., 2014***).

In addition to subtle mutations, tumors also accumulate major chromosomal changes (***Li et al., 2020***). Most solid tumors display widespread changes in chromosome number, as well as chromosomal deletions and translocations (***Lengauer et al., 1998***). Homozygous deletions of a few genes frequently drive carcinogenesis and the target gene involved in such deletions is always a TSG (***Cheng et al., 2017***). Somatic copy-number alterations, amplifications of cancer genes are also widespread in various types of cancers. Amplifications usually contain an oncogene (OG) whose protein product is abnormally active simply because the tumor cell contains 10–100 copies of the gene per cell, compared with the two copies present in normal cells (***Beroukhim et al., 2010***; ***Verhaak et al., 2019***). Chromosomal translocations may also convert wild-type forms of TSGs into forms that drive

carcinogenesis if the translocation inactivates the genes by truncation or by separating them from their promoter. Similarly, translocations may activate proto-oncogenes by changing their regulatory properties (**Haller et al., 2019**).

Epigenetic mechanisms such as DNA methylation and histone modifications may also alter the activity of cancer genes. It is now widely accepted that genetic and epigenetic changes go hand in hand in carcinogenesis: numerous genes involved in shaping the epigenome are mutated in common human cancers, and epigenetic changes affect many genes carrying driver mutations (**Yang and Yu, 2013**; **Chen et al., 2017b**; **Di Domenico et al., 2017**; **Roussel and Stripay, 2018**; **Chatterjee et al., 2018**). For example, promoter hypermethylation events may promote carcinogenesis if they lead to silencing of TSGs; the tumor-driving role of promoter methylation is obvious in the case of TSGs that are frequently inactivated by mutations in cancer (**Pfeifer, 2018**). Conversely, there is now ample evidence that promoter hypomethylation can promote carcinogenesis if it leads to increased expression of proto-oncogenes (**Van Tongelen et al., 2017**).

Non-coding RNAs (ncRNAs) also play key roles in carcinogenesis (**Slack and Chinnaiyan, 2019**). An explosion of studies has shown that – based on complementary base pairing – ncRNAs may function as OGs (by inhibiting the activity of TSGs), or as tumor suppressors (by inhibiting the activity of OGs or tumor essential genes [TEGs]).

Alterations in the splicing of primary transcripts of protein-coding genes also contribute to carcinogenesis. Recent studies on cancer genomes have revealed that recurrent somatic mutations of genes encoding RNA splicing factors (e.g. *SF3B1*, *U2AF1*, *SRSF2*, *ZRSR2*) lead to altered splice site preferences, resulting in cancer-specific mis-splicing of genes. In the case of proto-oncogenes, changes in the splicing pattern may generate active oncoproteins, whereas abnormal splicing of TSGs is likely to generate inactive forms of the tumor suppressor protein (**Dvinge et al., 2016**).

There is now convincing evidence that dysregulation of processes responsible for proteostasis also contributes to the development and progression of numerous cancer types (**Mofers et al., 2017**; **Chen et al., 2017c**; **Voutsadakis, 2017**). Recent studies on tumor tissues have revealed that genetic alterations and abnormal expression of various components of the protein homeostasis pathways (e.g. *FBXW7*, *VHL*) contribute to progression of human cancers by excessive degradation of tumor-suppressor molecules or through impaired disposal of oncogenic proteins (**Ge et al., 2018**; **Bernassola et al., 2019**).

Hallmarks of cancer and the function of genes involved in carcinogenesis

Hanahan and Weinberg have defined a set of hallmarks of cancer that allow the categorization of cancer genes with respect to their role in carcinogenesis (**Hanahan and Weinberg, 2011**). These hallmarks describe the biological capabilities usually acquired during the evolution of tumor cells: these include sustained proliferative signaling, evasion of growth suppressors, evasion of cell death, acquisition of replicative immortality, acquisition of capability to induce angiogenesis and activation of invasion and metastasis. Underlying all these hallmarks are defects in genome maintenance that help the acquisition of the above capabilities. Additional emerging hallmarks of potential generality have been suggested to include tumor promoting inflammation, evasion of immune destruction and reprogramming of energy metabolism in order to most effectively support neoplastic proliferation (**Hanahan and Weinberg, 2011**).

Figure 1 summarizes our current view of the cellular processes that play key roles in tumor evolution to emphasize their contribution to the various major hallmarks of cancer. Changes in the maintenance of the genome, epigenome, transcriptome, and proteome occupy a central position because they increase the chance that various constituents of other cellular pathways will experience alterations that favor the acquisition of capabilities that permit the proliferation, survival, and metastasis of tumor cells.

Chronology of tumor evolution: initiation and progression

In the first phase of carcinogenesis, a cell may acquire a mutation that permits it to proliferate abnormally, and in the next phase, other mutations allow the expansion of cell number and this process of mutations (and associated epigenetic, transcriptomic and proteomic alterations) continues, thus generating a primary tumor that can eventually metastasize to distant organs. Recent studies on the

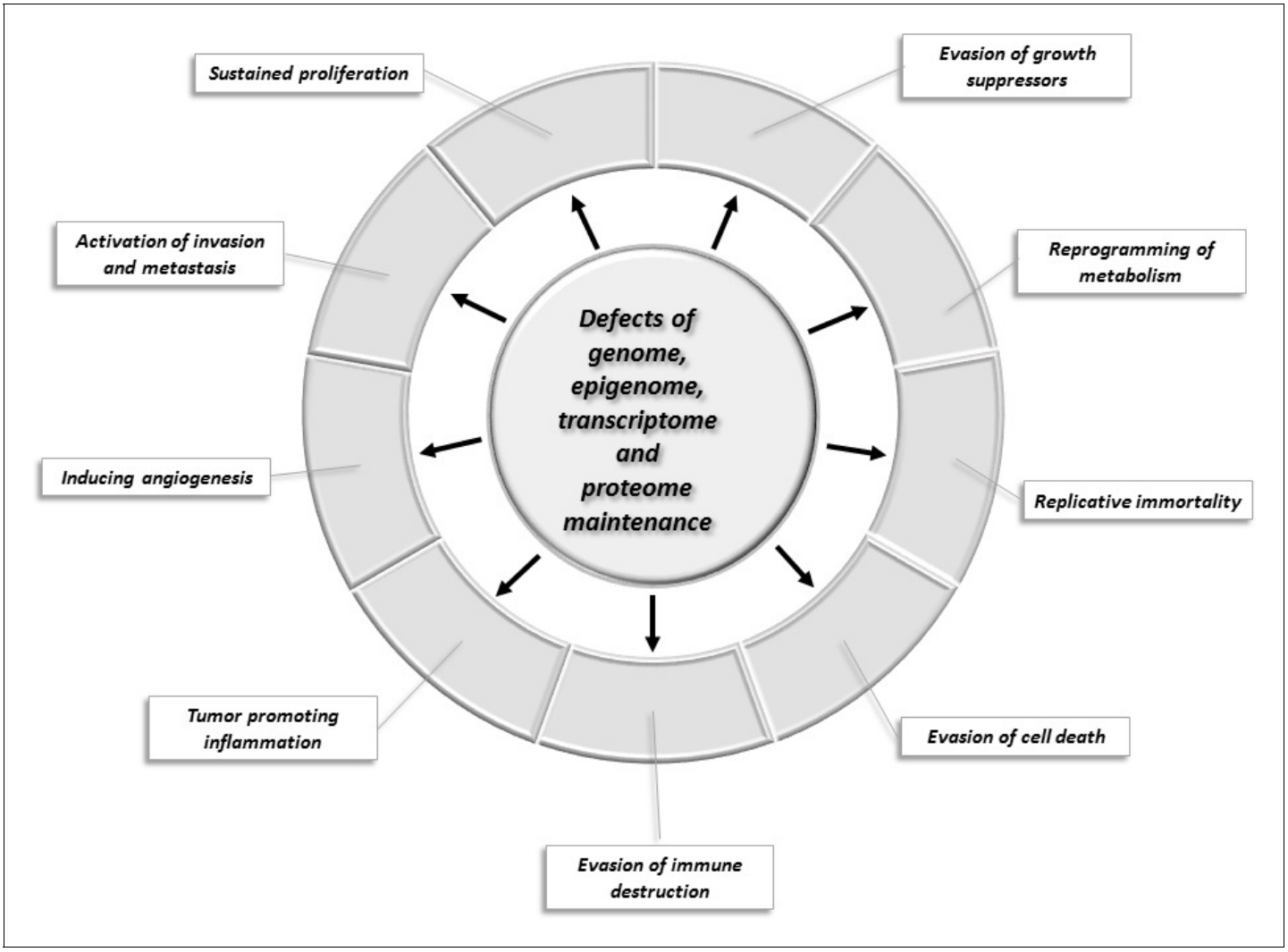


Figure 1. Changes of key cellular processes contributing to carcinogenesis. The central circle refers to processes involved in the maintenance of the integrity of the genome, epigenome, transcriptome, and proteome: defects in these processes increase the chance that genes and proteins of other cellular pathways (represented by segments of the outer circle) will suffer alterations that favor the acquisition of capabilities that permit the proliferation, survival, and metastasis of tumor cells.

chronology and genomic landscape of the events that drive carcinogenesis suggest that complex structural changes of the genome occur early, whereas point mutations occur in later disease phases (***Maura et al., 2019***; ***Voronina et al., 2020***).

According to current estimates, the number of cancer driving mutations needed for the full development of cancer ranges from two-eight depending on cancer type (***Vogelstein and Kinzler, 2015***; ***Anandakrishnan et al., 2019***). A recent integrative analysis of 2658 whole-cancer genomes and their matching normal tissues across 38 tumor types revealed that, on average, cancer genomes contain four to five driver mutations (***Campbell et al., 2020***).

Although the temporal order of the mutations affecting genes of key pathways differs among cancer types, it appears that a common feature is that mutations of genes that regulate apoptosis occur in the early phases of tumor progression, whereas mutations of genes involved in invasion pathways occur only in the last stages of carcinogenesis (***Gerstung et al., 2011***). It has been suggested that the reason why the loss of apoptotic control is a critical step for initiating cancer is that the larger the surviving cell population, the higher the number of cells at risk of acquiring additional mutations.

Analyses of the mutation landscapes and evolutionary trajectories of various tumor tissues have identified *BRAF, KRAS, TP53, RB*, or *APC* as the key genes whose mutation is most likely to initiate carcinogenesis, permitting the cell to divide abnormally (**Vogelstein and Kinzler, 2015**). In the case of ovarian cancers, *TP53* mutation is believed to be the earliest tumorigenic driver event, with presence in nearly all cases of ovarian cancer (**Bashashati et al., 2013**). The prevalence of *TP53* mutations and *BRCA* deficiency in these tumors leads to incompetent DNA repair promoting subsequent steps of carcinogenesis. Studies on the evolution of melanoma from precursor lesions have revealed that the vast majority of melanomas harbor *TERT* promoter mutations, indicating that these immortalizing mutations are selected at an unexpectedly early stage of neoplastic progression (**Shain et al., 2015**).

The life history and evolution of mutational processes and driver mutation sequences of 38 types of cancer has been analyzed recently by whole-genome sequencing analysis of 2658 cancers. This study has shown that early oncogenesis is characterized by mutations in a constrained set of driver genes and that the driver mutations that most commonly occur in a given cancer also tend to occur the earliest (**Gerstung et al., 2020**).

Cancer genes and passenger genes

The prominent role of *KRAS* and *TP53* genes in initiating carcinogenesis has been evident from the observation that their mutation rate in tumors far exceeds those of other genes, suggesting that their mutations are subject to positive selection during tumor evolution.

Several types of approaches exploit this principle for the identification of genes that drive carcinogenesis: the rate of mutation of 'driver genes' must be significantly higher in the tumor tissue than those of 'passenger genes' (PGs) that have no role in the development of cancer but simply happen to mutate in the same tumor (**Parmigiani et al., 2009**; **Meyerson et al., 2010**).

Unfortunately, methods based on mutation frequency alone cannot reliably indicate which genes are cancer drivers because the background mutation rates differ significantly as a consequence of intrinsic characteristics of DNA sequence and chromatin structure (**Michaelson et al., 2012**). Intrinsic mutation hotspots are mutation hotspots that depend on the nucleotide sequence context, the mechanism of mutagenesis and the action of the repair and replication machineries (**Rogozin and Pavlov, 2003**). Genes enriched in intrinsic mutation hotspots may accumulate mutations at a significantly higher rate than other genes, creating the illusion of positive selection; based on recurrent mutations they may be mistakenly identified as cancer driver genes (**Carter, 2019**; **Buisson et al., 2019**).

In principle, we can avoid this danger if we compare the mutation pattern of the gene in the tumor tissue with that in the normal tissue the tumor has originated from. However, since the rate of mutation in such hotspots depends not only on the nucleotide sequence but also on the mechanism of mutagenesis and the integrity of DNA repair pathways (**Buisson et al., 2019**; **Poulos et al., 2018**) mutation hotspots that arise during carcinogenesis could still create the illusion of positive selection.

Chromatin organization also has a major influence on regional mutation rates in human cancer cells (**Schuster-Böckler and Lehner, 2012**; **Gonzalez-Perez et al., 2019**). Since large-scale chromatin features, such as replication time and accessibility influence the rate of mutations, this may hinder the distinction of cancer driver genes whose high mutation rate reflects positive selection and PGs whose high mutation rate is the result of the distinctive features of the chromatin region in which they reside. Moreover, since the cell-of-origin chromatin organization shapes the mutational landscape, rates of somatic mutagenesis of genes in cancer are highly cell-type-specific (**Polak et al., 2015**). Actually, since regional mutation density of 'passenger' mutations across the human chromosomes correlates with the cell type the tumor had originated from, this feature may be used to classify human tumors (**Salvadores et al., 2019**).

Through the comparison of the exome sequences of 3083 tumor-normal pairs **Lawrence et al., 2013** have discovered an extraordinary variation in mutation frequency and spectrum within cancer types across the genome, which is strongly correlated with DNA replication timing and transcriptional activity. The authors have shown that by incorporating mutational heterogeneity into their analyses, they could eliminate many of the apparent artefactual findings, improving the identification of genes truly associated with cancer. In a more recent study **Lawrence et al., 2014** compared the frequency of somatic point mutations in exome sequences from 4742 human cancers and their matched normal-tissue samples across 21 cancer types and identified 33 genes that were not

previously known to be significantly mutated in cancer. They have concluded that 224 genes are significantly mutated in one or more tumor types.

However, since background mutational frequency estimates are not sensitive enough, the list of driver genes (defined as genes with increased somatic mutation rate) is likely to be incomplete, but may also contain false positives. To overcome these limitations of mutation rate-based approaches, several methods use additional features that may distinguish driver genes and PGs. A major group of such approaches incorporates observations about the impact of mutations on the structure and function of well-characterized proteins encoded by proto-oncogenes and TSGs. Several computational methods aim to identify driver missense mutations most likely to generate functional changes that causally contribute to tumorigenesis (**Kaminker et al., 2007**; **Carter et al., 2009**; **Nussinov et al., 2019**).

In a different type of approach **Youn and Simon, 2011** identified cancer driver genes as those for which the non-silent mutation rate is significantly greater than a background mutation rate estimated from silent mutations, indicating that the non-silent mutations are subject to positive selection. The authors have identified 28 genes as driver genes, the majority of the significant matches (e.g. *EGFR, CDKN2A, KRAS, STK11, TP53, NF1, RB1 PTEN*, and *NRAS*), were well-characterized OGs or TSGs known from earlier studies.

In a more recent study, **Zhou et al., 2017** have identified 365 genes for which the ratio of the nonsynonymous to synonymous substitution rate was significantly increased, suggesting that they are subject to the positive selection of driver mutations. However, an obvious limitation of such approaches is that they implicitly assume that synonymous substitutions are selectively neutral and therefore the ratio of the nonsynonymous to synonymous substitution rate properly monitors selection. This is not necessarily true: some synonymous mutations may have a significant impact on splicing, RNA stability, RNA folding and translation of the transcript of the affected gene and may thus actually act as driver mutations (**Supek et al., 2014**; **Hurst and Batada, 2017**; **Sharma et al., 2019**). Furthermore, some mutation hotspots may significantly increase the rate of synonymous mutations therefore a low ratio of nonsynonymous to synonymous substitution rate does not necessarily indicate the absence of positive selection or the action of purifying selection.

Vogelstein et al., 2013 have used a heuristic approach to identify cancer driver genes. Since the patterns of mutations in the first and best-characterized OGs and TSGs were found to be highly characteristic and nonrandom, the authors assumed that the same characteristics are generally valid and may be used to identify previously uncharacterized cancer genes. For example, since many known OGs were found to be recurrently mutated at the same amino acid positions, to classify a gene as an OG, it was required that >20% of the recorded mutations in the gene are at recurrent positions and are missense. Similarly, since in the case of known tumor suppressors the driver mutations most frequently truncate the tumor suppressor proteins, to be classified as a TSG, it was required that >20% of the recorded mutations in the gene are truncating (nonsense or frameshift) mutations. Along these lines, **Vogelstein et al., 2013** have analyzed the patterns of the subtle mutations in the Catalogue of Somatic Mutations in Cancer (COSMIC) database to identify driver genes. As a proof of the reliability of this '20/20 rule', the authors emphasized that all well-documented cancer genes passed these criteria (**Vogelstein et al., 2013**). Although this indicates that the approach detects known cancer genes, it does not guarantee that it detects all driver genes. Acknowledging that additional cancer driver genes might exist, the authors have introduced the term 'Mut-driver gene' for genes that contain a sufficient number or type of driver gene mutations to distinguish them from other genes, whereas for cancer genes that are expressed aberrantly in tumors but not frequently mutated they proposed the term 'Epi-driver gene'.

Based on these analyses, the authors have concluded that out of the 20,000 human protein-coding genes, only 125 genes qualify as Mut-driver genes, of these, 71 are TSGs and 54 are OGs (**Vogelstein et al., 2013**). The authors have expressed their conviction that nearly all genes mutated at significant frequencies had already been identified and that the number of Mut-driver genes is nearing saturation. This conclusion may not be justified since the criteria used to identify OGs and tumor suppressors appear to be too stringent and somewhat arbitrary.

In search of additional driver genes, **Tamborero et al., 2013** employed five complementary methods to find genes showing signals of positive selection and identified a list of 291 'high-confidence cancer driver genes' acting on 3205 tumors from 12 different cancer types. **Bailey et al., 2018** used multiple advanced algorithms to identify cancer driver genes and driver mutations. Based

on their PanCancer and PanSoftware analysis spanning 9423 tumor exomes, comprising all 33 of The Cancer Genome Atlas projects and using 26 computational tools they have identified 299 driver genes showing signs of positive selection. Their sequence and structure-based analyses detected >3400,400 putative missense driver mutations and 60–85% of the predicted mutations were validated experimentally as likely drivers.

Zhao et al., 2019a have developed driverMAPS (Model-based Analysis of Positive Selection), a model-based approach for driver gene identification that captures elevated mutation rates in functionally important sites and spatial clustering of mutations. The authors have identified 255 known driver genes as well as 170 putatively novel driver genes.

Currently, COSMIC (the Catalogue Of Somatic Mutations In Cancer, https://cancer.sanger.ac.uk/cosmic) is the most detailed and comprehensive resource for exploring the effect of subtle somatic mutations of driver genes in human cancer (***Tate et al., 2019***) but COSMIC also covers all the genetic mechanisms by which somatic mutations promote cancer, including non-coding mutations, gene fusions, and copy-number variants. In parallel with COSMIC's variant coverage, the Cancer Gene Census (CGC, https://cancer.sanger.ac.uk/census) describes a curated catalogue of genes driving every form of human cancer (***Sondka et al., 2018***). CGC has recently introduced functional descriptions of how each gene drives disease, summarized into the cancer hallmarks. CGC describes in detail the effect of a total of 719 cancer-driving genes, encompassing Tier 1 genes (574 genes) and a list of Tier 2 genes (145 genes) from more recent cancer studies that show less detailed indications of a role in cancer.

In a different type of approach, ***Torrente et al., 2016*** used comprehensive maps of human gene expression in normal and tumor tissues to identify cancer related genes. These analyses identified a list of genes with systematic expression change in cancer. The authors have noted that the list is significantly enriched with known cancer genes from large, public, peer-reviewed databases, whereas the remaining ones were proposed as new cancer gene candidates. A recent study has provided a comprehensive catalogue of cancer-associated transcriptomic alterations with the top-ranking genes carrying both RNA and DNA alterations. The authors have noted that this catalogue is enriched for cancer census genes (***Calabrese et al., 2020***).

Using transposon mutagenesis in mice, several laboratories have conducted forward genetic screens and identified thousands of candidate genetic drivers of cancer that are highly relevant to human cancer. The Candidate Cancer Gene Database (CCGD, http://ccgd-starrlab.oit.umn.edu/) is a manually curated database containing a unified description of all identified candidate driver genes (***Abbott et al., 2015***).

In summary, although a variety of approaches have been developed to identify 'cancer genes', there is significant disagreement as to the number of genes involved in carcinogenesis. Some of the studies argue that the number is in the 200–700 range, other approaches suggest that their number may be much higher. Since the ultimate goal of cancer genome projects is to discover therapeutic targets, it is important to identify all true cancer genes and distinguish them from PGs and candidates that do not play a significant role in the process of carcinogenesis.

We must point out, however, that the majority of genomics-based methods were biased as they defined the aim of cancer genomics as the identification of mutated driver genes (equating them with 'cancer genes') that are causally implicated in oncogenesis (***Futreal et al., 2004***). In all these studies, the underlying rationale for interpreting a mutated gene as causal in cancer development is that the mutations are likely to have been positively selected because they confer a growth advantage on the cell population from which the cancer has developed. An inevitable consequence of this focus on positive selection was that most studies neglected the possibility that negative selection may also play a significant role in tumor evolution.

Carcinogenesis as an evolutionary process

In principle, with respect to its effect on carcinogenesis, a somatic mutation may promote or may hinder carcinogenesis or may have no effect on carcinogenesis. In cancer genomics, the mutations that promote carcinogenesis (and are subject to positive selection during tumor evolution) are called 'driver mutations' to distinguish them from 'passenger mutations' that do not play a role in carcinogenesis (and are not subject to positive or negative selection during tumor evolution). Mutations that impair the growth, survival, and invasion of tumor cells have received much less attention, although they could also play a significant role in shaping the mutation pattern of genes during

carcinogenesis. Hereafter, we will refer to this category of mutations as 'cancer blocking mutations' because they are deleterious from the perspective of tumor growth.

As discussed above, in cancer genomics, genes are usually assigned to just two categories with respect to their role in carcinogenesis: (1) 'PGs' (or bystander genes) that play no significant role in carcinogenesis and their mutations are passenger mutations; (2) 'driver genes' that drive carcinogenesis when they acquire driver mutations.

The problem with this binary driver gene-PG categorization is that some genes with functions essential for the growth and survival of tumor cells (hereafter referred to as 'tumor essential genes') may not easily fit into either category. The coding sequences of driver genes (TSGs, proto-oncogenes), PGs, and TEGs are predicted to experience markedly different patterns of selection during tumor evolution.

The mutation patterns of selectively neutral, bona fide PGs are likely to reflect the lack of positive and negative selection, whereas in the case of TEGs purifying selection is predicted to dominate. In the case of TSGs, the mutation pattern is expected to reflect positive selection for inactivating driver mutations. Proto-oncogenes, however, are expected to show signs of both positive selection for activating mutations and negative selection for inactivating, 'cancer blocking' mutations as their activity is essential for their oncogenic role. In the coding regions of proto-oncogenes positive selection for driver mutations is expected to favor nonsynonymous substitutions over synonymous substitutions only at sites that are critical for the novel, oncogenic function. For these sites (and these sites only), the ratio of nonsynonymous to synonymous rates is expected to be significantly greater than one reflecting positive selection. If there are many such sites in a protein, or selection is extremely strong the overall nonsynonymous to synonymous ratio for the entire protein may also be significantly higher than one, otherwise the effect of positive selection on the synonymous to nonsynonymous ratio may be overridden by purifying selection at other sites (**Patthy, 1999**).

In harmony with some of these expectations, using just the ratio of the nonsynonymous to synonymous substitution rate as a measure of positive or negative selection, **Zhou et al., 2017** have shown that in cancer genomes, the majority of genes had nonsynonymous to synonymous substitution rate values close to one, suggesting that they belong to the PG category. The authors have identified a total of 365 potential cancer driver genes that had nonsynonymous to synonymous substitution rate values significantly greater than one (reflecting the dominance of positive selection). Conversely, 923 genes had nonsynonymous to synonymous substitution rate values significantly less than one (reflecting the dominance of negative selection), leading the authors to suggest that these negatively selected genes may be important for the growth and survival of cancer cells.

Pyatnitskiy et al., 2015 have also used the dN/dS ratio (the ratio of nonsynonymous and synonymous substitution rates) as an indicator of selective pressure and have identified 91 protein-coding genes ('essential cancer proteins') with amino acid sequences under negative selection.

Realizing that genes whose wild-type coding sequences are needed for tumor growth are also of key interest for cancer research, **Weghorn and Sunyaev, 2017** have also focused on the role of negative selection in human cancers. The authors have used an approach based on the principle that both positive and negative selection can be inferred by comparing the observed mutation rates to the expectation under the sole action of the mutation process. As the authors have pointed out, identification, and analysis of true negatively selected,' undermutated' genes is particularly difficult since the sparsity of mutation data results in lower statistical power, making conclusions less reliable. Although the signal of negative selection was exceedingly weak, the authors have noted that the group of negatively selected candidate genes is enriched in cell-essential genes identified in a CRISPR screen (**Wang et al., 2015a**), consistent with the notion that one of the potential causes of negative selection is the maintenance of genes that are responsible for basal cellular functions. Based on pergene estimates of negative selection inferred from the pan-cancer analysis the authors have identified 147 genes with significant negative selection. The authors have noted that among the 13 genes showing the strongest signs of negative selection there are several genes (*ATAT1, BCL2, CLIP1, GALNT6, CKAP5,* and *REV1*) that are known to promote carcinogenesis.

In a similar work, **Martincorena et al., 2017** have used the normalized ratio of non-synonymous to synonymous mutations, to quantify selection in coding sequences of cancer genomes. Using a nonsynonymous-to-synonymous substitution rate value >1 as a marker of cancer genes under positive selection, they have identified 179 cancer genes, with about 50% of the coding driver mutations being found to occur in novel cancer genes. The authors, however, have concluded that purifying

selection is practically absent in tumors since nearly all (>99%) coding mutations are tolerated and escape negative selection. The authors have suggested that this remarkable absence of negative selection on coding point mutations in cancer indicates that the vast majority of genes are dispensable for any given somatic lineage, presumably reflecting the buffering effect of diploidy and the inherent resilience and redundancy built into most cellular pathways.

The key message of ***Martincorena et al., 2017*** that negative selection has no role in cancer evolution had a major impact on cancer genomics research as reflected by several commentaries in major journals of the field that have propagated this conclusion (***Bakhoum and Landau, 2017***; ***Koch, 2017***; ***Vitale and Galluzzi, 2018***).

Some more recent studies, however, contradict this conclusion. Although ***Zapata et al., 2018*** have also used the ratio of nonsynonymous-to-synonymous substitutions to identify genes that are under selection, they have detected significant negative selection in the case of 25 genes. ***López et al., 2020***, focusing on dN/dS values for truncating mutations, have shown that purifying selection of essential genes is significant in early phases of tumor evolution (before whole genome duplications), whereas whole-genome doubling allows the accumulation of deleterious alterations. ***Tilk et al., 2020*** have shown that appreciable negative selection (dN/dS ~ 0.4) is present in tumors with a low mutational burden, while the majority of tumors exhibit dN/dS ratios approaching 1, suggesting that tumors with higher mutational burden do not remove deleterious mutations.

Van den Eynden and Larsson, 2017, however, cautioned that it is crucial to take into account mutational signatures when applying the dN/dS metric to cancer somatic mutation data. For example, the authors have shown that the low dN/dS values observed in malignant melanoma may be due to the predominance of C to T mutations in this tumor and do not necessarily indicate gene essentiality. The authors have also shown that purifying selection is very limited and similar in all tumor types if the dN/dS metric uses mutational signature-derived substitution probabilities.

In view of the contradicting conclusions about the significance of negative selection in tumor evolution, in the present work we have reexamined this question using an approach that attempts to overcome some of the problems highlighted by earlier studies.

First, most studies used a single dN/dS metric to measure nonsynonymous to synonymous substitution rates as indicators of selective pressure and paid less attention to the fact that the strength of purifying selection is an order of magnitude greater for nonsense mutations than for missense mutations (***Gorlov et al., 2006***). Furthermore, the use of a single dN/dS value for a transcript may preclude the simultaneous detection of positive and negative selection of activating and inactivating mutations, both of which might operate for a given gene. To overcome these limitations, in the present study we have used a clustering-based approach that can detect different signals of selection manifested in rates of nonsense, missense *versus* silent substitutions in the coding regions of genes.

Second, an inherent problem with the detection of purifying selection in tumor tissues is that putative TEGs are likely to be undermutated relative to PGs and driver genes, resulting in low statistical power of their analyses based on dN/dS metrics. We have reduced this problem by combining subtle somatic mutations from different tumors types and limiting our work to transcripts that have at least 100 somatic mutations in tumors. (Note that the requirement of a minimum number of mutations does not place a theoretical limit on this approach; progress with genome-wide screens and collection of more data is overcoming this limitation.)

In harmony with earlier observations, our analyses have confirmed that the vast majority of human genes are PGs that do not show detectable signals of selection, whereas known TSGs are positively selected for inactivating (primarily nonsense and frame-shift) mutations. Known OGs, however, were found to display signals of both negative selection for inactivating (nonsense, frame-shift) mutations and positive selection for activating (missense) mutations. Improved detection of signals of selection has permitted the identification of a number of novel driver genes that are likely to play important roles in carcinogenesis as TSGs or as OGs.

Significantly, we have identified a cluster of human genes that show clear signs of negative selection during tumor evolution, suggesting that their functional integrity is essential for the growth and survival of tumor cells. The group of negatively selected genes includes genes known to play critical roles in the Warburg effect of cancer cells, others are known to mediate invasion and metastasis of tumor cells, indicating that negatively selected TEGs may prove a rich source for novel targets for tumor therapy.

Results

Distinguishing PGs and cancer genes

The rationale of the analyses described in the present work is that — due to their different roles in carcinogenesis — proto-oncogenes, TSGs, TEGs, and PGs are expected to experience different patterns of selection during tumor evolution and this is reflected in the relative rates of missense, nonsense, and silent mutations of their protein-coding regions. To monitor these differences, we have calculated for each transcript the fraction of somatic substitutions that could be assigned to the silent (fS), misssense (fM), and nonsense (fN) category and analyzed their relative rates. (For details, the reader should consult the Materials and methods section).

Our analyses have shown that in 3D scatter plots of the fS, fM, and fN values of transcripts the majority of genes are present in a central cluster characterized by fS, fM, and fN values close to those expected assuming no mutation bias and absence of selection, consistent with the view that they correspond to PGs (***Figure 2***). Known OGs, however, were found in a separate cluster characterized by higher fM values, reflecting positive selection for missense mutations, whereas the cluster of known TSGs has higher fN values, reflecting positive selection for truncating nonsense mutations (***Figure 2B and C***).

Known cancer genes also separate from the majority of human genes in 3D scatter plots of rSM, rNM, rNS parameters, defined as the ratio of fS/fM, fN/fM, fN/fS, respectively (***Figure 3***). In these scatter plots, OGs separate from the central cluster in having lower rSM and rNM values, whereas TSGs have higher rNS and rNM values than those of the central cluster (***Figure 3***).

The separation of known cancer genes from the majority of human genes is even more manifest in 3D scatter plots of parameters rSMN, rMSN, and rNSM defined as the ratio of fS/(fM+fN), fM/(fS+fN), and fN/(fS+fM), respectively (***Figure 4***). In these plots, the transcripts form a three-pronged cluster, with known OGs and TSGs being present on separate spikes of this cluster, the rMSN and rNSM spikes, respectively (***Figure 4***).

There is, however, a fourth cluster of genes that deviates from the clusters of PGs, OGs, and TSGs (***Figures 2***, ***3*** and ***4***). The high fS, rSM, and rSMN values of the transcripts in this group

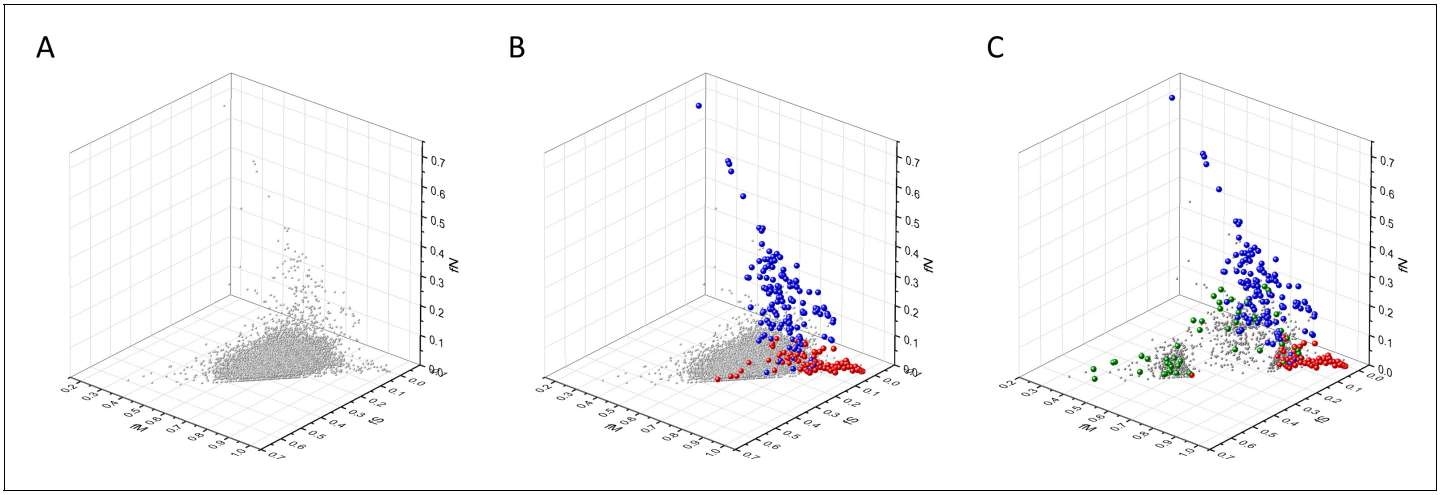


Figure 2. Analyses of fS, fM, and fN parameters of human protein-coding genes of tumor tissues. The figure shows the results of the analysis of 13,803 transcripts containing at least 100 subtle, confirmed somatic mutations from tumor tissues, including only mutations identified as not single-nucleotide polymorphisms (SNPs). Axes *x, y*, and *z* represent the fractions of somatic single-nucleotide substitutions that are assigned to the synonymous (fS), nonsynonymous (fM), and nonsense (fN) categories, respectively. In Panel A, each gray ball represents a human transcript; note that the majority of human genes are present in a dense cluster. Panel B highlights the positions of transcripts of the genes identified by ***Vogelstein et al., 2013*** as oncogenes (OGs, large red balls) or tumor suppressor genes (TSGs, large blue balls). It is noteworthy that these driver genes separate significantly from the central cluster and from each other: OGs have a significantly larger fraction of nonsynonymous, whereas TSGs have significantly larger fraction of nonsense substitutions. Panel C shows data only for candidate cancer genes present in the CG_SO^{2SD}_SSI^{2SD} list (see Materials and methods). The positions of novel cancer gene transcripts validated in the present work are highlighted as large green balls.

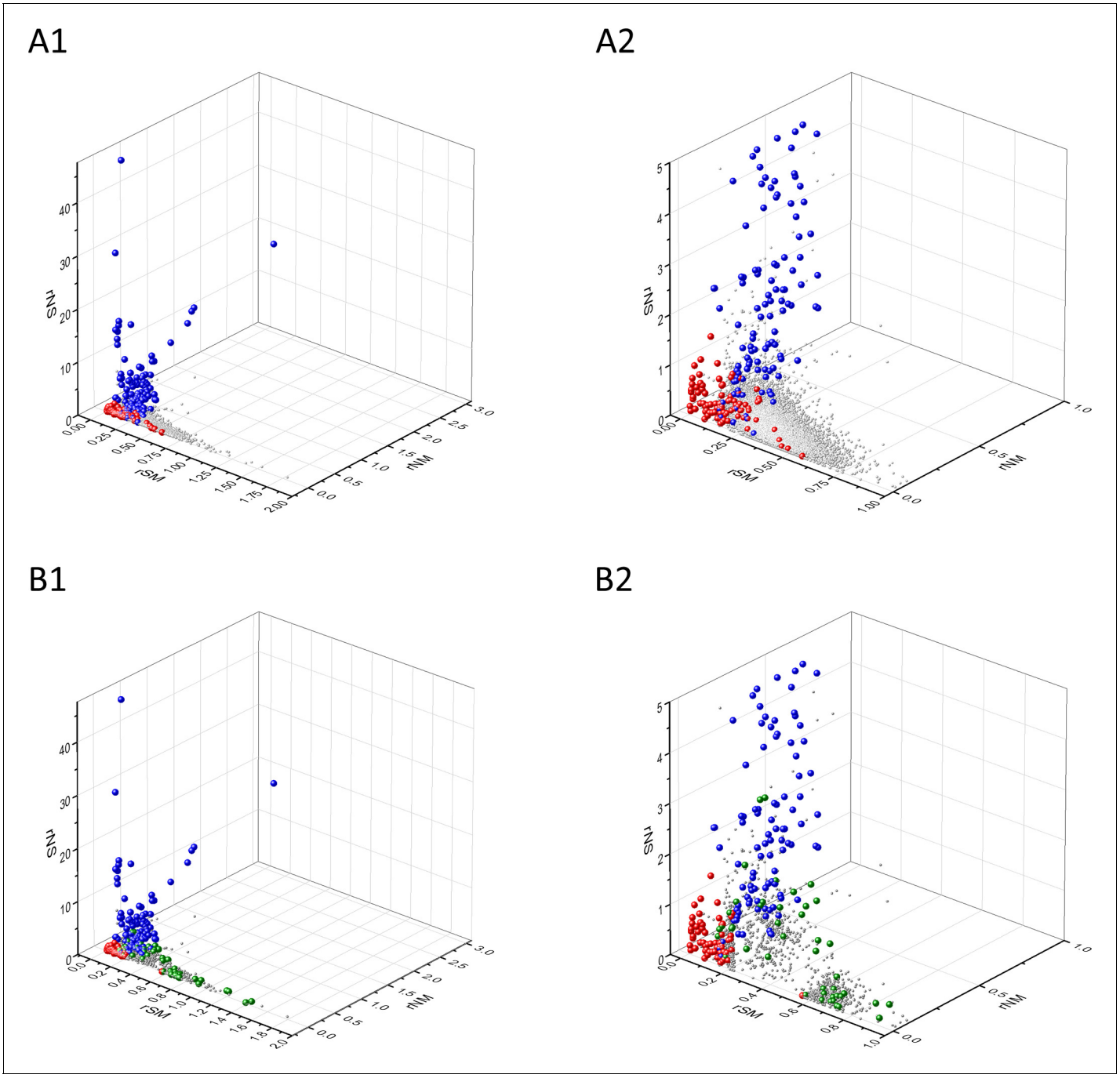


Figure 3. Analyses of rSM, rNM, rNS parameters of human protein-coding genes of tumor tissues. The figure shows the results of the analysis of 13,803 transcripts containing at least 100 subtle, confirmed somatic mutations from tumor tissues, including only mutations identified as not single-nucleotide polymorphisms (SNPs). Axes *x*, *y*, and *z* represent the rSM, rNM, rNS values defined as the ratio of fS/fM, fN/fM, fN/fS, respectively. Each ball represents a human transcript; the positions of transcripts of the genes identified by ***Vogelstein et al., 2013*** as oncogenes (OGs, large red balls) or tumor suppressor genes (TSGs, large blue balls) are highlighted. Panels A1, A2 show the distribution of the 13,803 transcripts at different magnification. Note that the majority of human genes are present in a dense cluster but known OGs and TSGs separate significantly from the central cluster and from each other. The rNS and rNM values of TSGs are higher, whereas the rSM and rNM values of OGs are lower than those of passenger genes. Panels B1, B2 show data only for candidate cancer genes present in the CG_SO^{2SD}_SSI^{2SD} list (see Materials and methods). The positions of novel cancer gene transcripts validated in the present work are highlighted as large green balls.

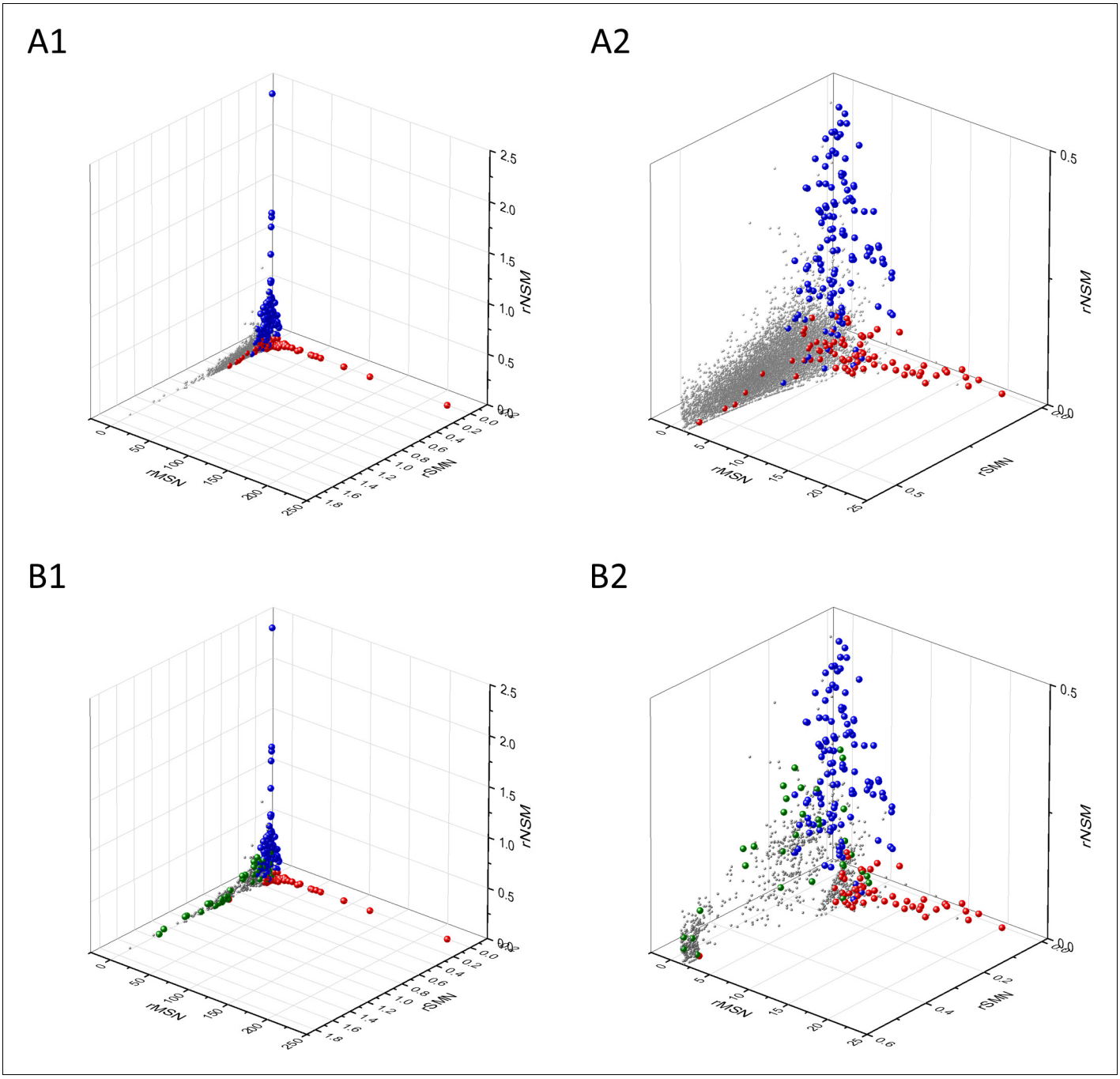


Figure 4. Analyses of rSMN, rMSN, and rNSM parameters of human protein-coding genes of tumor tissues. The figure shows the results of the analysis of transcripts containing at least 100 subtle, confirmed somatic mutations from tumor tissues, including only mutations identified as not single-nucleotide polymorphisms (SNPs). Axes *x*, *y*, and *z* represent the rSMN, rMSN, and rNSM defined as the ratio of fS/(fM+fN), fM/(fS+fN), and fN/(fS+fM). Each ball represents a human transcript; the positions of transcripts of the genes identified by ***Vogelstein et al., 2013*** as oncogenes (OGs, large red balls) or tumor suppressor genes (TSGs, large blue balls) are highlighted. Panels A1, A2 show the distribution of the 13,803 transcripts at different magnification. Note that the majority of human genes are present in a dense cluster but known OGs and TSGs separate significantly from the central cluster and from each other. The rNSM values of TSGs are higher, their rMSN and rSMN are lower than those of passenger genes (PGs). OGs also separate from PGs in that their rMSN values are higher and their rSMN and rNSM values are lower than those of PGs. Panels B1, B2 show data only for candidate cancer genes present in the CG_SO^{2SD}_SSI^{2SD} list (see Materials and methods). The positions of novel cancer gene transcripts validated in the present work are highlighted as large green balls.

suggest that they are subject to purifying selection during tumor evolution, raising the possibility that this group may contain genes essential for the survival of tumors.

The analyses discussed above did not take into account the impact of differences in mutation probability on the fN, fM, and fS values of transcripts. To check the influence of this factor, we have calculated the expected fN*, fM*, and fS* values for all human transcripts using the probabilities of the six substitution classes (C>A, C>G, C>T, T>A, T>C, and T>G) observed across tumors (for deatails the reader should consult the Materials and methods section).

The various types of observed/expected ratios (rN*, rM*, rS*; rSM*, rNM*, rNS*; rSMN*, rMSN* and rNSM*) were calculated for each transcript and the data were analyzed in 3D scatter plots as described above for the observed values. As shown in ***Figures 5***, ***6*** and ***7***, the distribution of transcripts in these 3D scatter plots are similar to those observed in the corresponding ***Figures 2***, ***3*** and ***4***, indicating that the separation of the clusters of PGs, OGs, TSGs, and TEGs is relatively insensitive to transcript-specific differences in mutation probabilities.

Analyses of candidate cancer gene sets

We assumed that the genes whose patterns of subtle mutations deviate significantly (by more than 2SD) from those of prototypical PGs are enriched in cancer genes that play important role in carcinogenesis. The patterns of subtle mutations of candidate cancer genes assign them to one of the three main clusters that show signs of positive and/or negative selection (see ***Figures 2–7***). (A) Genes positively selected for inactivating (nonsense and frame-shift) mutations – putative TSGs; (B) genes positively selected for missense mutations and negatively selected for inactivating mutations – putative proto-oncogenes; (C) negatively selected genes – putative TEGs.

The assumption that the cancer genes assigned to these three clusters play significant roles in carcinogenesis has strong support in the case of the first two categories: the approach used in the present study correctly assigned the known, 'gold standard' TSGs and OGs (***Supplementary file 1***). In the case of the third category, however, no similar gold standard exists for TEGs.

To check the validity and predictive value of the assumption that the genes assigned to the three clusters play critical roles in carcinogenesis, we have selected a number of genes at random from

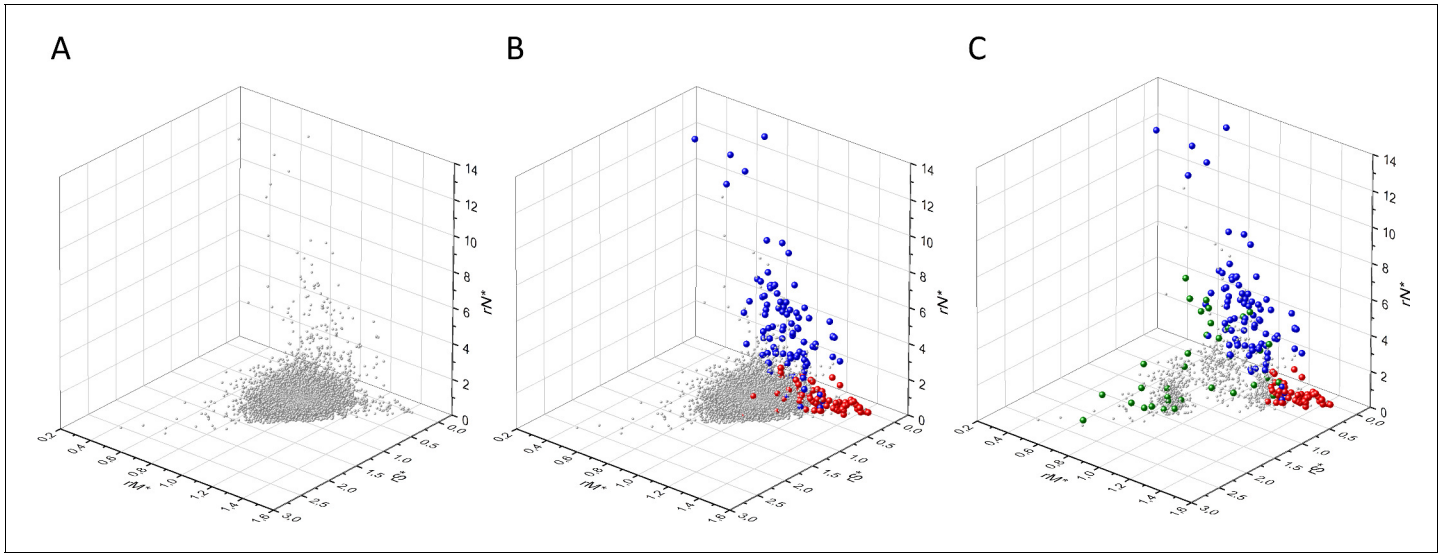


Figure 5. Analyses of rS*, rM*, and rN* parameters of human protein-coding genes of tumor tissues. The figure shows the results of the analysis of transcripts containing at least 100 subtle, confirmed somatic mutations from tumor tissues. Axes *x*, *y*, and *z* represent rS*, rM*, and rN* values, respectively. In Panel A, each gray ball represents a human transcript; note that the majority of human genes are present in a dense cluster. Panel B highlights the positions of transcripts of the genes identified by ***Vogelstein et al., 2013*** as oncogenes (OGs, large red balls) or tumor suppressor genes (TSGs, large blue balls). It is noteworthy that these driver genes separate significantly from the central cluster and from each other: OGs have a significantly larger fraction of nonsynonymous, whereas TSGs have significantly larger fraction of nonsense substitutions than expected. Panel C shows data only for candidate cancer genes present in the CG_SO^{2SD}_SSI^{2SD} list (see Materials and methods). The positions of novel cancer gene transcripts validated in the present work are highlighted as large green balls.

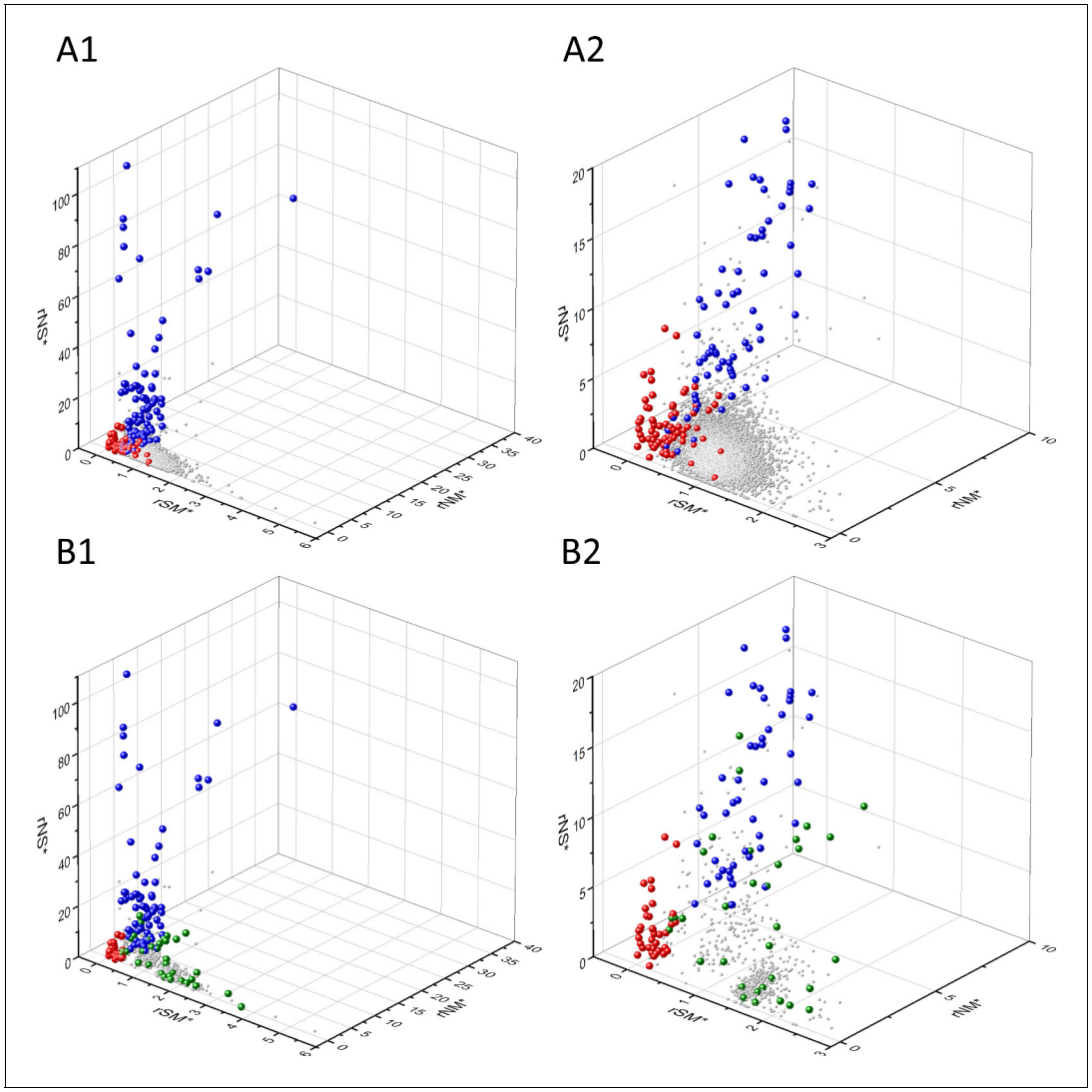


Figure 6. Analyses of rSM*, rNM*, rNS* parameters of human protein-coding genes of tumor tissues. The figure shows the results of the analysis of transcripts containing at least 100 subtle, confirmed somatic mutations from tumor tissues. Axes *x*, *y*, and *z* represent rSM*, rNM*, rNS* values, respectively. Each ball represents a human transcript; the positions of transcripts of the genes identified by ***Vogelstein et al., 2013*** as oncogenes (OGs, large red balls) or tumor suppressor genes (TSGs, large blue balls) are highlighted. Panels A1 and A2 show the distribution of the transcripts at different magnification. Note that the majority of human genes are present in a dense cluster but known OGs and TSGs separate significantly from the central cluster and from each other. The rNS* and rNM* values of TSGs are higher, whereas the rSM* and rNM* values of OGs are lower than those of passenger genes. Panels B1, B2 show data only for candidate cancer genes present in the CG_SO^{2SD}_SSI^{2SD} list (see Materials and methods). The positions of novel cancer gene transcripts validated in the present work are highlighted as large green balls.

each cluster for further in-depth analyses. We have used three criteria to select genes for detailed analyses from the combined list of candidate cancer genes that deviate from the central clusters of PGs by more than 2SD (see Materials and methods). (1) The candidate gene is among the genes showing the strongest signals of selection characteristic of the given group. (2) The candidate gene is novel in the sense that it is not listed among the 145' gold standard' OGs and TSGs of ***Vogelstein et al., 2013*** or among the 719 cancer genes of CGC (***Sondka et al., 2018***). (3) There is substantial experimental information in the scientific literature on the given gene to permit the assessment of its role in carcinogenesis.

The genes discussed below include genes positively selected for truncating mutations (putative TSGs), genes positively selected for missense mutations and negatively selected for inactivating mutations (putative proto-oncogenes) and negatively selected genes (putative TEGs). In the main text, we summarize only the major conclusions of our analyses; for annotations of the individual

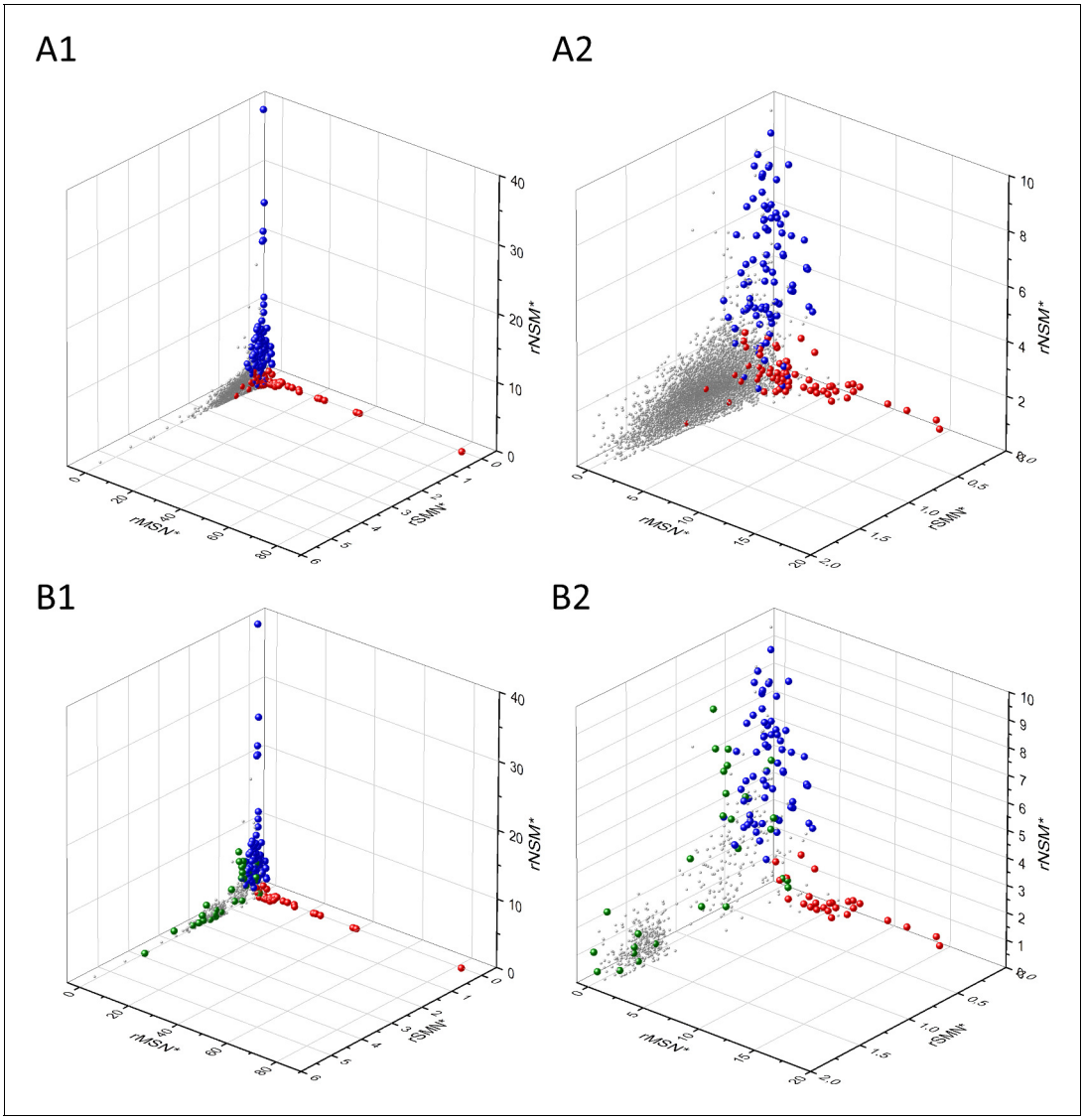


Figure 7. Analyses of rSMN*, rMSN*, and rNSM* parameters of human protein-coding genes of tumor tissues. The figure shows the results of the analysis of transcripts containing at least 100 subtle, confirmed somatic mutations from tumor tissues. Axes *x*, *y*, and *z* represent the rSMN*, rMSN*, and rNSM* values, respectively. Each ball represents a human transcript; the positions of transcripts of the genes identified by ***Vogelstein et al., 2013*** as oncogenes (OGs, large red balls) or tumor suppressor genes (TSGs, large blue balls) are highlighted. Panels A1, A2 show the distribution of the transcripts at different magnification. Note that the majority of human genes are present in a dense cluster but known OGs and TSGs separate significantly from the central cluster and from each other. The rNSM* values of TSGs are higher, their rMSN* and rSMN* are lower than those of passenger genes (PGs). OGs also separate from PGs in that their rMSN* values are higher and their rSMN* and rNSM* values are lower than those of PGs. Panels B1, B2 show data only for candidate cancer genes present in the CG_SO^{2SD}_SSI^{2SD} list (see Materials and methods). The positions of novel cancer gene transcripts validated in the present work are highlighted as large green balls.

genes, the reader should consult Appendix 1. We discuss examples of negatively selected genes in the main text in more detail since earlier studies that focussed on positive selection of driver mutations inevitably missed these genes. We also discuss some instructive examples of' false' hits, that is cases where the mutation parameters deviate significantly from those of PGs, but this deviation is not due to selection.

Novel cancer genes positively selected for nonsense mutations

We have selected genes positively selected for truncating mutations from the combined list of candidate transcripts, that is, transcripts whose parameters deviate from those of PGs by more than 2SD

(for details see Materials and methods). We have used the additional restriction that genes with indel_rNSM <0.125 were excluded (**Supplementary file 1**), thereby removing OGs and TEGs. Out of the 624 genes that satisfy these criteria, we have subjected *B3GALT1*, *BMPR2*, *BRD7*, *ING1*, *MGA*, *PRRT2*, *RASA1*, *RNF128*, *SLC16A1*, *SPRED1*, *TGIF1*, *TNRC6B*, *TTK*, *ZNF276*, *ZC3H13*, *ZFP36L2*, and *ZNF750* to further analysis.

Annotation of the majority of these genes (*BMPR2*, *BRD7*, *ING1*, *MGA*, *PRRT2*, *RASA1*, *RNF128*, *SLC16A1*, *SPRED1*, *TGIF1*, *TNRC6B*, *ZC3H13*, *ZFP36L2*, and *ZNF750*) has provided convincing evidence for their role in carcinogenesis as tumor suppressors. Interestingly, experimental evidence suggests that *TTK*, encoding dual specificity protein kinase TTK, is a proto-oncogene that may be converted to an OG by truncating mutations affecting the very C-terminal end of the protein, downstream of its kinase domain (for further details see Appendix 1). Our annotations suggest that *B3GALT1*, *ZNF276* are false positives whose apparent mutation pattern deviates significantly from those of PGs, but this deviation is not due to selection.

Based on functional annotation of the TSGs identified and validated in the present work (see Appendix 1), we have assigned them to various cellular processes of cancer hallmarks in which they are involved (**Table 1**).

Comparison of the list of 624 genes present in this dataset (CG_SSI^{2SD} rNSM >0.125) with lists identified by others (**Supplementary file 1**) revealed that ~60–100 of our candidate TSG-like genes are also found in several gene lists identified by others through analyses of somatic mutations of tumor tissues. Many of the genes selected for annotation are present in at least one of the candidate gene lists identified by others; the genes of *MGA*, *RASA1*, *TGIF1*, *ZFP36L2*, and *ZNF750* are present in multiple cancer gene lists (**Supplementary file 1**). It is noteworthy, however, that *RNF128*, *SLC16A1*, *SPRED1*, *TNRC6B*, and *TTK* are novel in that they are found only among the candidate cancer genes identified by forward genetic screens in mice (**Abbott et al., 2015**) or among the genes whose expression changes in cancer (**Torrente et al., 2016**).

We have also analyzed the genes present in dataset CG_SO*^{2SD}_rNSM >3, that is, candidate cancer genes for which the observed rNSM values are more than threefold higher than expected taking into account mutational signature-derived substitution probabilities of tumors (**Supplementary file 2**). We have found that 164 (100%) of the 164 genes present in this dataset are also present in the dataset CG_SSI^{2SD} rNSM >0.125. It is noteworthy that the majority of candidate TSGs selected for annotation (*B3GALT1*, *BMPR2*, *BRD7*, *ING1*, *MGA*, *PRRT2*, *RASA1*, *SLC16A1*, *SPRED1*, *TGIF1*,

Table 1. Assignment of novel positively or negatively selected cancer genes to key cellular processes of carcinogenesis.

Hallmarks of cancer	Gene symbol
Defects of genome, epigenome, transcriptome, or proteome maintenance	*CDK8, FOXG1, IDH3B, MARCH7, MGA,* **NOVA1**, **PNCK**, *RNF128, TGIF1, TNRC6B,* **TWIST1**, *ZC3H13, ZFP36L1, ZFP36L2, ZNF750*
Sustained proliferation	*AURKA, BRD7, ING1,* **FOXG1**, **MAPK13**, **PNCK**, *PRRT2, RASA1, RIT1, SPRED1,* **TRIB2**, *TTK, YAP1, YES1, ZFP36L1, ZFP36L2, ZNF750*
Evasion of growth suppressors	
Reprogramming of metabolism	*BRD7,* **G6PD**, *SLC16A1,* **SLC16A3**, **SLC2A1**, **SLC2A8**, *YAP1, YES1*
Replicative immortality	**NOVA1**
Evasion of cell death	*BRD7, ING1,* **MAPK13**, **PNCK**, *PRRT2,* **TP73**, *TRIB2, TTK, YAP1, YES1, ZNF750*
Evasion of immune destruction	
Tumor promoting inflammation	*BMP2R,* **CCR2**, **CCR5**, **CX3CR1**, **MAPK13**
Inducing angiogenesis	**CCR2**
Activation of invasion and metastasis	**CCR2**, **CCR5**, **CX3CR1**, *RASA1,* **TBXA2R**

For annotation of novel genes identified in the present study see Appendix 1. The names of negatively selected genes are marked by bold underline.

ZNF276, *ZFP36L2*, and *ZNF750*) are present among the genes shared by the two datasets that show the strongest signals of positive selection for nonsense substitutions.

Novel cancer genes positively selected for missense and negatively selected for nonsense mutations

We have selected genes positively selected for missense and negatively selected for inactivating mutations from the list of candidate transcripts using the restriction that genes with rMSN <3.00 (440) were excluded, thereby removing the majority of TSGs and TEGs (***Supplementary file 1***). Out of the 440 genes that satisfy these criteria, we have subjected *AURKA, CDK8, IDH3B, MARCH7, RIT1, YAP1*, and *YES1* to further analysis.

Annotation of these genes has confirmed that they play important roles in carcinogenesis as OGs. Three of these genes encode kinases (Aurora kinase A, also known as breast tumor-amplified kinase; cyclin-dependent kinase 8; tyrosine-protein kinase Yes, also known as proto-oncogene c-Yes) but unlike many other oncogenic kinases, these OGs do not show significant clustering of missense mutations. In fact, only in the case of *IDH3B* and *RIT1* did we observe clustering of missense mutations, indicating that recurrent mutation is not an obligatory property of proto-oncogenes.

Based on functional annotation of the novel OGs identified and validated in the present work (see Appendix 1), we have assigned them to various cellular processes of cancer hallmarks in which they are involved (***Table 1***).

Comparison of this list of 440 genes (CG_SO^{2SD} rMSN >3.00) with the lists of cancer genes identified by others (***Supplementary file 1***) revealed that ~60–100 of our candidate OG-like genes are present in cancer gene lists identified by others through analyses of somatic mutations of tumor tissues.

Out of the genes that we have selected for annotation only the *RIT1* gene has been identified by others as an OG, based on the analysis of somatic mutations (***Supplementary file 1***). *AURKA* and *IDH3B* are not present in any of the lists of cancer genes, whereas *CDK8, MARCH7*, *YAP1*, and *YES1* are found among the more than 9000 candidate cancer genes identified by forward genetic screens in mice (***Abbott et al., 2015***). Interestingly, *TTK*, identified as a gene positively selected for truncating mutations (see list CG_SSI^{2SD} rNSM >0.125), but annotated as an OG, is also present in the list of genes positively selected for missense mutations (CG_SO^{2SD} rMSN >3.00).

We have also analyzed the genes present in dataset CG_SO*^{2SD}–rMSN >1.50, that is, genes for which the observed rMSN values are more than 1.5-fold higher than expected taking into account mutational signature-derived substitution probabilities of tumors (***Supplementary file 2***). We have found that 119 (98.3%) of the 121 genes present in this dataset are also present in the dataset CG_SO^{2SD} rMSN >3.00. It should be noted that the majority of candidate OGs selected for annotation (*AURKA, RIT1, YAP1*, and *YES1*) are found among the genes shared by the two datasets, showing strong signals of positive selection for missense substitutions.

Negatively selected genes

We have selected putative TEGs from the list of candidate cancer genes using the restriction that we have excluded genes with rSMN <0.5 to eliminate OGs and TSGs (***Supplementary file 3***). Out of the 505 genes, we have subjected *CX3CR1, FOXG1, FOXP2, G6PD, MAPK13, MLLT3, NOVA1, PNCK, RUNX2, SLC16A3, SLC2A1, SLC2A8, TBP, TBXA2R, TP73*, and *TRIB2* to further analysis.

Our analyses have confirmed that in the majority of cases (*CX3CR1, FOXG1, G6PD, MAPK13, NOVA1, PNCK, SLC16A3, SLC2A1, SLC2A8, TBXA2R, TP73, TRIB2*) the high synonymous-to-nonsynonymous and nonsense mutation rates could be interpreted as evidence for purifying selection during tumor evolution. There were, however, several examples (e.g. *DSPP, FOXP2, MLLT3, RUNX2, TBP*) where high synonymous-to-nonsynonymous and nonsense mutation rates were found to reflect increased rates of synonymous substitution (due to the presence of mutation hotspots), rather than decreased rates of nonsynonymous and nonsense substitutions that could be due to purifying selection (for details see Appendix 1).

Annotations of the genes *CX3CR1, FOXG1, G6PD, MAPK13, NOVA1, PNCK, SLC16A3, SLC2A1, SLC2A8, TBXA2R, TP73*, and *TRIB2* have confirmed that all of them play important roles in carcinogenesis (see Appendix 1) permitting their assignment to various cellular processes of cancer hallmarks (***Table 1***.). As discussed below (and in Appendix 1), they fulfill pro-oncogenic functions by

promoting cell proliferation (*FOXG1, MAPK13, PNCK, TRIB2*), evasion of cell death (*MAPK13, PNCK, TP73*), replicative immortality (*NOVA1*), reprogramming of energy metabolism of cancer cells (*G6PD, SLC16A3, SLC2A1, SLC2A8*), inducing tumor promoting inflammation (*CX3CR1, MAPK13*) and invasion and metastasis (*CX3CR1, TBXA2R*). In view of the pro-oncogenic role of these proteins, it is noteworthy, that *G6PD, MAPK13, PNCK, SLC16A3*, and *SLC2A1* are among the candidate cancer genes identified by forward genetic screens in mice (**Abbott et al., 2015**).

Comparison of our list of 505 negatively selected genes (CG_SO^{2SD}_rSMN > 0.5) with those identified by others have revealed very little similarity (**Supplementary file 3**). Out of the 147 genes of **Weghorn and Sunyaev, 2017**, only one is present in the list of top-ranking negatively selected genes identified in the present study. Similarly, only four of the 25 genes of **Zapata et al., 2018** and only five of the 91 genes of **Pyatnitskiy et al., 2015** are found in our list of negatively selected genes (**Supplementary file 3**).

We observed a greater similarity when we compared our list of negatively selected genes with that of **Zhou et al., 2017**; 32 of the 112 genes identified by **Zhou et al., 2017** are also present among the 505 negatively selected genes identified in the present work (**Supplementary file 3**). It is noteworthy that top-ranking genes present in both lists include the *ACKR3*, *TBP*, and *MLLT3* genes. As discussed in Appendix 1, the apparent signals of negative selection (high synonymous-to-nonsynonymous rates) of genes like *DSPP, FOXP2, MLLT3, RUNX2*, and *TBP* may reflect the presence of mutation hotspots generating silent mutations and not purifying selection. **Zhou et al., 2017** have also noted that "some cancer genes also show negative selection in cancer genomes, such as the OG *MLLT3*" and that "interestingly, *MLLT3* has recurrent synonymous mutations at amino acid positions 166 to 168". Apparently, the authors did not realize that this observation of recurrent silent substitutions (in a poly-Ser region of the protein) questions the validity of the claim that the unusually low nonsynonymous to synonymous rate is due to negative selection (for more detail see Appendix 1).

In summary, the pro-oncogenic, negatively selected genes annotated and validated in the present work are missing from the earlier lists of negatively selected genes (**Zhou et al., 2017**; **Pyatnitskiy et al., 2015**; **Weghorn and Sunyaev, 2017**; **Zapata et al., 2018**). A possible explanation for the lack of similarity of top-ranking negatively selected genes identified in the present study with those identified by others is that we have limited our work to transcripts that have at least 100 somatic mutations. It is noteworthy that a large fraction of genes identified by others did not pass this requirement (see Materials and methods).

We have also analyzed the genes present in dataset CG_SO*^{2SD} rSMN >1.50, that is, candidate cancer genes for which the observed rSMN values are more than 1.5-fold higher than expected taking into account mutational signature-derived substitution probabilities of tumors (**Supplementary file 4**). We have found that 200 (86.5%) of the 231 genes present in this dataset are also present in dataset CG_SO^{2SD} rSMN >0.5. It should be noted that the majority of candidate TEGs selected for annotation (*CX3CR1, FOXG1, FOXP2, MAPK13, MLLT3, NOVA1, RUNX2, SLC16A3, SLC2A8, TBP, TBXA2R*, and *TRIB2*) are found among the 200 genes shared by the two datasets and that show the strongest signals of negative selection for missense and nonsense substitutions.

Negative selection, cell essentiality, and tumor essentiality of genes

As we have emphasized in the Introduction, the conclusions drawn from earlier studies searching for signs of negative selection are highly controversial. A highly publicized study has propagated the conclusion that negative selection has no role in tumor evolution (**Martincorena et al., 2017**; **Bakhoum and Landau, 2017**; **Koch, 2017**; **Vitale and Galluzzi, 2018**). **Martincorena et al., 2017** have argued that the practical absence of purifying selection during tumor evolution is due to the buffering effect of diploidy and functional redundancy of most cellular pathways.

A recent study has examined the influence of functional redundancy on the essentiality of genes (**De Kegel and Ryan, 2019**). The authors have used CRISPR score profiles of 558 genetically heterogeneous tumor cell lines and converted continuous values of gene CRISPR scores to binary essential and nonessential calls. These analyses have shown that 1014 genes belong to a category of 'broadly essential genes', that is, these genes were found to be essential in at least 90% of the 558 cell lines. **De Kegel and Ryan, 2019** have shown that, compared to singleton genes, paralogs are less frequently essential and that this is more evident when considering genes with multiple paralogs or

with highly sequence-similar paralogs. In harmony with these conclusions, ***López et al., 2020*** have found that purifying selection of essential genes is significant in early phases of tumor evolution but in later phases whole-genome doubling allows the accumulation of deleterious alterations.

Since the group of negatively selected genes identified by ***Weghorn and Sunyaev, 2017*** were shown to be enriched in cell-essential genes (***Wang et al., 2015a***), the authors have proposed that the major cause of negative selection during tumor evolution is the maintenance of genes that are responsible for basal cellular functions. Nevertheless, ***Weghorn and Sunyaev, 2017*** have pointed out that negative selection is also expected to act on neoantigens, expanding the possible scope of purifying selection beyond cell essentiality.

Although analyses of negatively selected genes have led ***Zapata et al., 2018*** to conclude, "Processes that are most strongly conserved are those that play fundamental cellular roles such as protein synthesis, glucose metabolism, and molecular transport" they also emphasized the possible importance of less basic functions. Since the immune system is capable of discriminating cancer cells by recognizing mutated epitope sequences the authors have hypothesized that native epitope sequences would be protected from nonsynonymous mutations during tumor evolution. In harmony with this hypothesis, the authors have observed signals of selection in the immunopeptidome and proteins of the epitope presentation machinery, arguing for their importance in the evasion of immune surveillance by tumors.

Gene Ontology analysis of the negatively selected 'essential cancer proteins' identified by ***Pyatnitskiy et al., 2015*** have revealed enrichment of essential proteins related to membrane and cell periphery, leading the authors to speculate that this could be a sign of immune system-driven negative selection of cancer neo-antigens.

In summary, there is some disagreement about the significance of purifying selection in tumor evolution and whether tumor essential functions can be equated with basic cellular functions.

In order to assess the contribution of cell-essentiality to purifying selection during tumor evolution, we have plotted various measures of negative selection of human genes as a function of their cell-essentiality scores determined by ***De Kegel and Ryan, 2019***. These analyses have shown that there is a very weak, positive correlation (Pearson's r = 0.05345, $p<0.05$) between rSMN (a measure of purifying selection) and the cell-essentiality scores of transcripts (***Figure 8***, ***Supplementary file 5***). Since, by definition, there is a negative correlation between the essentiality of genes and their cell-essentiality scores (***De Kegel and Ryan, 2019***), our data indicate that cell essentiality does not contribute significantly to purifying selection during tumor evolution.

It is also noteworthy that the cell essentiality scores of negatively selected genes (CG_SO^{2SD} rSMN >0.5) are not significantly different from those of PGs (***Figure 8***, ***Supplementary file 5***). Comparison of CRISPR scores (−0.07665 ± 0.17269) of the cluster of negatively selected genes of CG_SO^{2SD} rSMN >0.5 listed in ***Supplementary file 3*** with CRISPR scores (−0.09506 ± 0.24168) of the cluster of PGs ($PG\_SO^{r3\_1SD}$) revealed that they are not significantly different ($p>0.05$). This indicates that basic cell-essentiality per se does not explain the purifying selection observed for this cluster of genes.

Comparison of the lists of negatively selected genes identified in the present work with the 1014 'broadly essential genes' defined by ***De Kegel and Ryan, 2019*** has revealed that there is practically no overlap between the two groups. Only six of the 1014 broadly essential genes are included in our list of negatively selected genes (***Supplementary file 3***). This observation also suggests that cell-essentiality defined by CRISPR scores determined experimentally on cell lines is not relevant for negative selection during tumor evolution in vivo.

Our analyses of cases of strong purifying selection suggest that it has more to do with a function *specifically* required by the tumor cell for its growth, survival, and metastasis than with general basic cellular functions (***Table 1***). It is noteworthy in this respect, that the genes showing the strongest signals of negative selection include several plasma membrane receptor proteins (e.g. *ACKR3, CCR2, CCR5, CX3CR1, TBXA2R*) that cancer cells utilize to promote migration, invasion, and metastasis (Appendix 1). Significantly, these proteins exert their biological functions (in cell migration, inflammation, angiogenesis etc.) primarily at the organism level, therefore their cell-essentiality scores may have little to do with their overall essentiality for tumor growth and metastasis. Inspection of the data of ***De Kegel and Ryan, 2019*** shows that *ACKR3, CX3CR1, TBXA2R* were not assigned to the essential category in any of the 558 tumor cell lines tested.

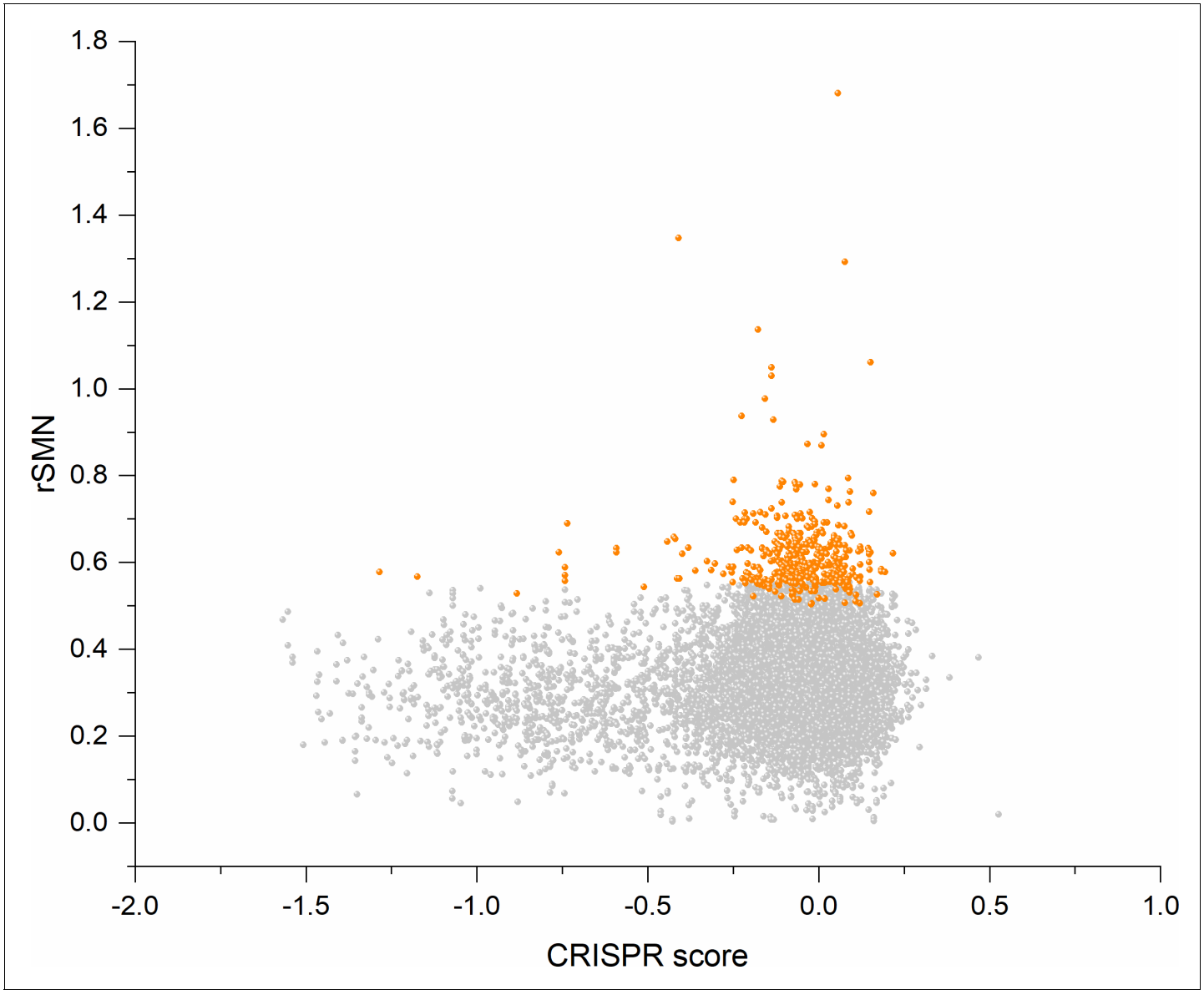


Figure 8. Cell-essentiality scores of human genes and negative selection during tumor evolution. The figure shows the results of the analysis of transcripts containing at least 100 subtle, confirmed somatic, non-polymorphic mutations from tumor tissues. The abscissa indicates the cell-essentiality score of the genes, the ordinate shows the rSMN parameters of the transcripts. Each ball represents a human transcript. Transcripts showing strongest signals of negative selection (CG_SO^{2SD} rSMN >0.5) are represented by dark orange balls.

Negatively selected, TEGs identified in the present study do include proteins involved in cell-level processes: they promote cell proliferation (*FOXG1, MAPK13, PNCK*, and *TRIB2*), evasion of cell death (*MAPK13, PNCK*, and *TP73*), replicative immortality (e.g. *NOVA1*), or they are crucial for the reprogramming of energy metabolism in cancer cells (e.g. *GAPD, SLC16A3, SLC2A1*, and *SLC2A8*). Nevertheless, their negative selection is unlikely to be a mere reflection of their basic cellular functions. Rather, it reflects the exceptional role of the corresponding cancer hallmarks (evasion of cell death, replicative immortality, reprogramming of metabolism) in carcinogenesis (***Figure 1***). In harmony with this conclusion *NOVA1, SLC16A3, SLC2A8*, and *TP73* were assigned to the essential category by ***De Kegel and Ryan, 2019*** in less than 10% of the 558 tumor cell lines tested. *SLC2A1* (glucose transporter 1) is an exception in as much as it was found to be cell-essential in 41% of the cell lines. Significantly, several nutrient transporter genes (*SLC16A3, SLC2A1*, and *SLC2A8)* were found among the genes showing the strongest signs of purifying selection. It must be mentioned

here that **Zapata et al., 2018** have also noted that the glucose transporters *SLC2A1* and *SLC2A8* and the lactate transporter *SLC16A3* show signs of purifying selection, although they did not list these genes among the 25 genes with significant negative selection.

The most likely explanation for the tumor essentiality of transporter protein genes *SLC16A3, SLC2A1*, and *SLC2A8* is that tumor cells have an increased demand for nutrients and this demand is met by enhanced cellular entry of nutrients through upregulation of specific transporters (**Ganapathy et al., 2009**). The uncontrolled cell proliferation of tumor cells involves major adjustments of energy metabolism in order to support cell growth and division in the hypoxic microenvironments in which they reside. Otto Warburg was the first to observe an anomalous characteristic of cancer-cell energy metabolism: even in the presence of oxygen, cancer cells limit their energy metabolism largely to glycolysis, leading to a state that has been termed 'aerobic glycolysis' (**Warburg, 1956a**; **Warburg, 1956b**). Cancer cells are known to compensate for the lower efficiency of ATP production through glycolysis than oxidative phosphorylation by upregulating glucose transporters, such as facilitated glucose transporter member 1, GLUT1 (encoded by the *SLC2A1* gene), thus increasing glucose import into the cytoplasm (**Jones and Thompson, 2009**; **DeBerardinis et al., 2008**; **Hsu and Sabatini, 2008**).

The markedly increased uptake of glucose has been documented in many human tumor types, by visualizing glucose uptake through positron emission tomography. The reliance of tumor cells on glycolysis is also supported by the hypoxia response system: under hypoxic conditions, not only glucose transporters but also multiple enzymes of the glycolytic pathway are upregulated (**Jones and Thompson, 2009**; **DeBerardinis et al., 2008**; **Semenza, 2010a**; **Semenza, 2010b**; **Kroemer and Pouyssegur, 2008**).

In our view, the central role of GLUT1 in cancer metabolism is reflected by the fact that the *SLC2A1* gene encoding this glucose transporter is among the genes that show the strongest signals of purifying selection. The key importance of GLUT1 in cancer may be illustrated by the fact that high levels of GLUT1 expression correlates with a poor overall survival and is associated with increased malignant potential, invasiveness, and poor prognosis (**Wang et al., 2017a**; **Deng et al., 2018**; **de Castro et al., 2019**). The strict requirement for GLUT1 in the early stages of mammary tumorigenesis highlights the potential for glucose restriction as a breast cancer preventive strategy (**Wellberg et al., 2016**). The tumor essentiality of GLUT1 may also be illustrated by the fact that knockdown of GLUT1 inhibits cell glycolysis and proliferation and inhibits the growth of tumors (**Xiao et al., 2018**). In view of its essentiality for tumor growth, GLUT1 is a promising target for cancer therapy (**Shibuya et al., 2015**; **Noguchi et al., 2016**; **Chen et al., 2017d**).

Recent studies suggest that the *YAP1-TEAD1-GLUT1* axis plays a major role in reprogramming of cancer energy metabolism by modulating glycolysis (**Lin and Xu, 2017**). These authors have shown that *YAP1* and *TEAD1* are involved in transcriptional control of the glucose transporter *GLUT1*, whereas knockdown of *YAP1* inhibited glucose consumption, and lactate production of breast cancer cells, overexpression of GLUT1 restored glucose consumption and lactate production.

Besides GLUT1 another glucose transporter, GLUT8 (encoded by the *SLC2A8* gene) also shows strong signals of negative selection, arguing for its importance in tumor survival. In harmony with this interpretation, there is evidence that GLUT8 is overexpressed in and is required for proliferation and viability of tumors (**Goldman et al., 2006**; **McBrayer et al., 2012**).

Due to abnormal conversion of pyruvic acid to lactic acid even under normoxia, the altered metabolism of glucose consuming tumors must rapidly efflux lactic acid to the microenvironment to maintain a robust glycolytic flux and to prevent poisoning themselves (**Mathupala et al., 2007**). Survival and maintenance of the glycolytic phenotype of tumor cells is ensured by monocarboxylate transporter 4 (MCT4, encoded by the *SLC16A3* gene) that efficiently transports L-lactate out of the cell (**Ganapathy et al., 2009**). Significantly, MCT4, encoded by the *SLC16A3* gene also shows strong signals of negative selection, in harmony with its importance in tumor survival. As high metabolic and proliferative rates in cancer cells lead to production of large amounts of lactate, extruding transporters are essential for the survival of cancer cells as illustrated by the fact that knockdown of MCT4 increased tumor-free survival and decreased in vitro proliferation rate of tumor cells (**Andersen et al., 2018**). Using a functional screen **Baenke et al., 2015** have also demonstrated that monocarboxylate transporter four is an important regulator of breast cancer cell survival: MCT4 depletion reduced the ability of breast cancer cells to grow, suggesting that it might be a valuable therapeutic target. In harmony with the essentiality of MCT4 for tumor growth, several studies

indicate that expression of the hypoxia-inducible monocarboxylate transporter MCT4 is increased in tumors and its expression correlates with clinical outcome, thus it may serve as a valuable prognostic factor (**Witkiewicz et al., 2012**; **Doyen et al., 2014**; **Baek et al., 2014**). Consistent with the key importance of MCT4 for the survival of tumor cells, its selective inhibition to block lactic acid efflux appears to be a promising therapeutic strategy against highly glycolytic malignant tumors (**Choi et al., 2016**; **Todenhöfer et al., 2018**; **Choi et al., 2018**; **Zhao et al., 2019b**).

Interestingly, the thromboxane A2 receptor gene (*TBXA2R*) as well as several chemokine receptor protein genes (*CCR2, CCR5, CX3CR1)* were also found among the genes showing strong signs of purifying selection (see Appendix 1). (Note that **Pyatnitskiy et al., 2015** have also identified *CCR5* as a negatively selected gene). The most likely explanation for their essentiality for tumor growth is that tumor cells rely on these receptors in various steps of invasion and metastasis (see Appendix 1). It is noteworthy in this respect that another member of the family of chemokine receptors, the atypical chemokine receptor 3, *ACKR3* is also among the genes showing very high values of rSMN, suggesting negative selection of missense and nonsense mutations (**Supplementary file 3**). (Note that **Zhou et al., 2017** have also identified *ACKR3* as a negatively selected gene). Significantly, *ACKR3* is a well-known OG, present in Tier 1 of the Cancer Gene Census. Several studies support the key role of *ACKR3* in tumor invasion and metastasis (**Li et al., 2014**; **Stacer et al., 2016**; **Zhao et al., 2017**; **Puddinu et al., 2017**; **Melo et al., 2018**; **Qian et al., 2018**). Since knock-down or pharmacological inhibition of *ACKR3* has been shown to reduce tumor invasion and metastasis, ACKR3 is a promising therapeutic target for the control of tumor dissemination (for further details see Appendix 1).

Negative selection of germline mutations in the human population versus negative selection of somatic mutations in cancer

The data discussed in the previous section indicate that the importance ('essentiality') of a given gene is a question of perspective. Cell-essential genes may be non-essential for tumor growth, whereas TEGs with tumor-specific functions do not necessarily have cell-essential functions. Similarly, we may assume that the importance of a gene might be quite different from the perspective of tumor cells and from the perspective of the entire organism. One could speculate that somatic mutations of genes with functions that have no relevance for tumor growth (PGs) experience neutral evolution during tumor growth, whereas germline mutations of the same genes may be subject to purifying selection at the level of organismal evolution, as is true for the majority of genes (**Gorlov et al., 2006**). One may also assume that genes with tumor essential, tumor-specific functions may be subject to purifying selection during both tumor evolution and organism evolution, but the strength of purifying selection of these genes is increased in tumors relative to those of genes that do not have tumor-specific functions.

To test these assumptions, we have determined the signals of selection of germline mutations (**Supplementary file 6**) and compared them with those determined for the same genes in the case of somatic mutations of cancer. Comparison of the patterns of germline and somatic mutations of human transcripts (**Supplementary file 7**) has revealed that the proportion of silent substitutions is significantly higher for germline mutations than for somatic mutations of tumors (fS^g: 0.33900 versus fS^s: 0.24604, $p<0.05$). Conversely, the proportions of nonsense and missense mutations are significantly lower for germline mutations than for somatic mutations of tumors (fN^g: 0.02329 versus fN^s: 0.04669, $p<0.05$; fM^g: 0.63771 versus fM^s: 0.70727, $p<0.05$). These observations are in harmony with the dominance of purifying selection in the human population (**Gorlov et al., 2006**).

As shown in **Figure 9**, the pattern of the distribution of transcripts in 3D scatter plots of fM, fN, and fS parameters for germline mutations are strikingly different from those observed in the case of fM, fN, and fS parameters of somatic mutations in cancer (compare **Figure 9A and B**). In addition to a general shift of germline mutations to lower fN and fM and higher fS values, in the case of germline mutations the fN, fM, and fS parameters of transcripts of TSGs, OGs, and TEGs do not separate from those of the central cluster of genes. Similarly, the distribution of transcripts in 3D scatter plots of rS**, rM**, and rN** parameters for germline mutations are different from those observed in the case of rS*, rM*, and rN* parameters of somatic mutations in cancer (compare **Figure 9C and D**): cancer genes do not separate from the central cluster of genes.

Comparison of the fS, rSM, and rSMN parameters of germline and somatic mutations of transcripts (**Figure 10**, **Supplementary file 7**) has shown that there is only weak correlation between the strength of purifying selection of genes during tumor evolution and organismal evolution. The

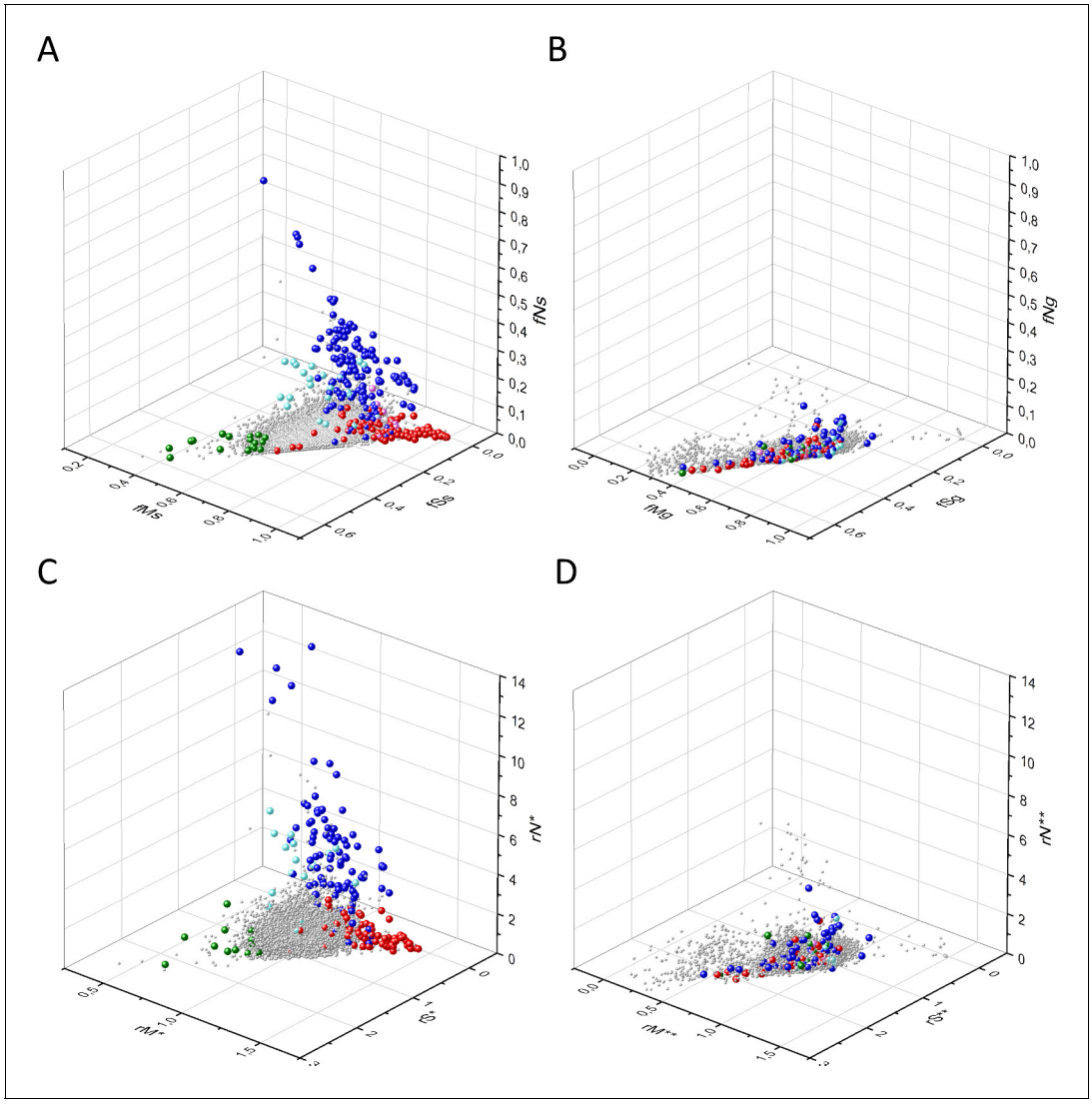


Figure 9. Comparison of the patterns of germline mutations of genes with those of somatic mutations observed during tumor evolution. Panel A: fS, fM, and fN scores of somatic mutations in cancer, Panel B: fS, fM, and fN scores of germline mutations. Panel C: rS*, rM*, and rN* scores of somatic mutations in cancer, Panel D: rS**, rM**, and rN** scores of germline mutations. Each ball represents a human transcript. The positions of transcripts of the genes identified by ***Vogelstein et al., 2013*** as oncogenes (OGs, large red balls) or tumor suppressor genes (TSGs, large blue balls) are highlighted. Novel proto-oncogenes, tumor suppressors and tumor essential genes identified in the present work are highlighted in magenta, cyan, and green, respectively.

Pearson's r values for the correlations of the fS, rSM, and rSMN parameters of germline and somatic mutations are 0.1127, 0.05757, and 0.02635, $p<0.05$, respectively.

These comparisons have also revealed that – relative to other genes – the candidate TEGs identified in the present study (CG_SO2SD_rSMN >0.5) display significantly stronger signals of purifying selection during tumor evolution than during organismal evolution (***Figure 10***, ***Supplementary file 7***). The fS, rSM, and rSMN parameters of somatic mutations of candidate TEGs are significantly higher than those of other genes (fS^s: 0.38322 versus 0.24045, $p<0.05$; rSM^s: 0.66013 versus 0.34375, $p<0.05$; $rSMN^s$: 0.62774 versus 0.32356, $p<0.05$). The fS, rSM, and rSMN parameters of the germline mutations of candidate TEGs, however, differ much less from the corresponding parameters of other genes (fS^g: 0.36487 versus 0.33831, $p<0.05$; rSM^g: 0.64054 versus 0.56394, $p<0.05$; $rSMN^g$: 0.61264 versus 0.56178, $p<0.05$). These observations indicate that the negative selection of candidate TEGs during tumor evolution is not a simple reflection of their essentiality at the organism level; it is more likely that they serve tumor-specific functions.

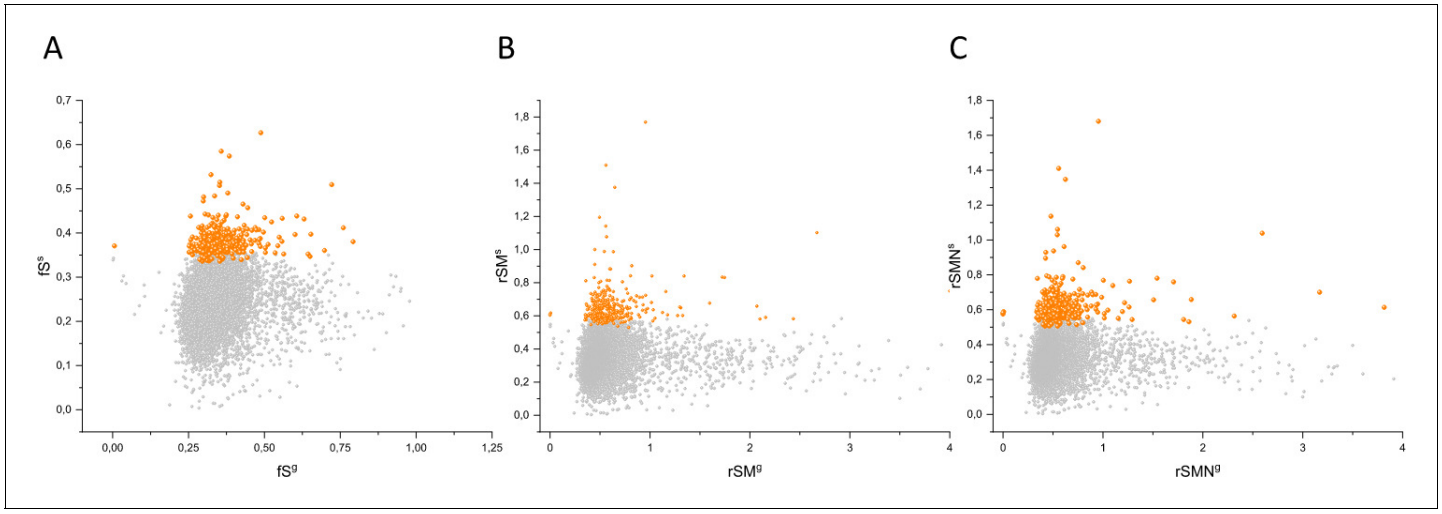


Figure 10. Comparison of fS, rSM, and rSMN scores of genes determined for somatic mutations in tumors with those determined for germline mutations. The abscissas indicate the fS^g (panel A), rSM^g (panel B), and $rSMN^g$ (panel C) scores of germline mutations of human genes and the ordinates shows the corresponding fS^s, rSM^s, and $rSMN^s$ scores of somatic mutations of tumors for the same genes. Each ball represents a human gene. Transcripts showing the strongest signals of negative selection during tumor evolution (CG_SO^{2SD} rSMN >0.5) are represented by dark orange balls.

In order to assess the contribution of cell-essentiality to purifying selection during organismal evolution we have plotted $rSMN^g$, a measure of negative selection of germline mutations of human genes, as a function of their cell-essentiality scores determined by ***De Kegel and Ryan, 2019***. These analyses have shown that there is a very weak negative correlation (Pearson's $r = -0.03662$, $p<0.05$) between the strength of purifying selection of transcripts ($rSMN^g$) and their cell-essentiality scores (***Figure 11***, ***Supplementary file 7***). This observation also indicates that essentiality of cell-level functions measured by cell-essentiality scores contribute to, but do not explain the strength of purifying selection observed during organismal evolution.

Discussion

One of the major goals of cancer research is to identify all 'cancer genes', that is genes that play a role in carcinogenesis. In the last two decades, several types of approaches have been developed to achieve this goal, but the implicit assumption of most of these studies was that a distinguishing feature of cancer genes is that they are positively selected for mutations that drive carcinogenesis. As a result of combined efforts, the PCAWG driver list identifies a total of 722 protein-coding genes as cancer driver genes and 22 non-coding driver mutations (***Rheinbay et al., 2020***; ***Campbell et al., 2020***).

In a recent editorial, commenting on a suite of papers on the genetic causes of cancer, Nature has expressed the view that the core of the mission of cancer-genome sequencing projects—to provide a catalogue of driver mutations that could give rise to cancer—has been achieved (***Editorial, 2020***). It is noteworthy, however, that, although on average, cancer genomes were shown to contain four to five driver mutations, in around 5% of cases no drivers were identified in tumors (***Campbell et al., 2020***). As pointed out by the authors, this observation suggests that cancer driver discovery is not yet complete, possibly due to failure of the available bioinformatic algorithms. The authors have also suggested that tumors lacking driver mutations may be driven by mutations affecting cancer-associated genes that are not yet described for that tumor type, however, using driver discovery algorithms on tumors with no known drivers, no individual genes reached significance for point mutations (***Campbell et al., 2020***).

In our view, these observations actually suggest that a rather large fraction of cancer genes remains to be identified. Assuming that tumors, on average, must have driver mutations affecting at least four or five cancer genes and that known and unknown cancer genes play similar roles in carcinogenesis, the observation that a 0.05 fraction of tumors has no known drivers (i.e. they are driven

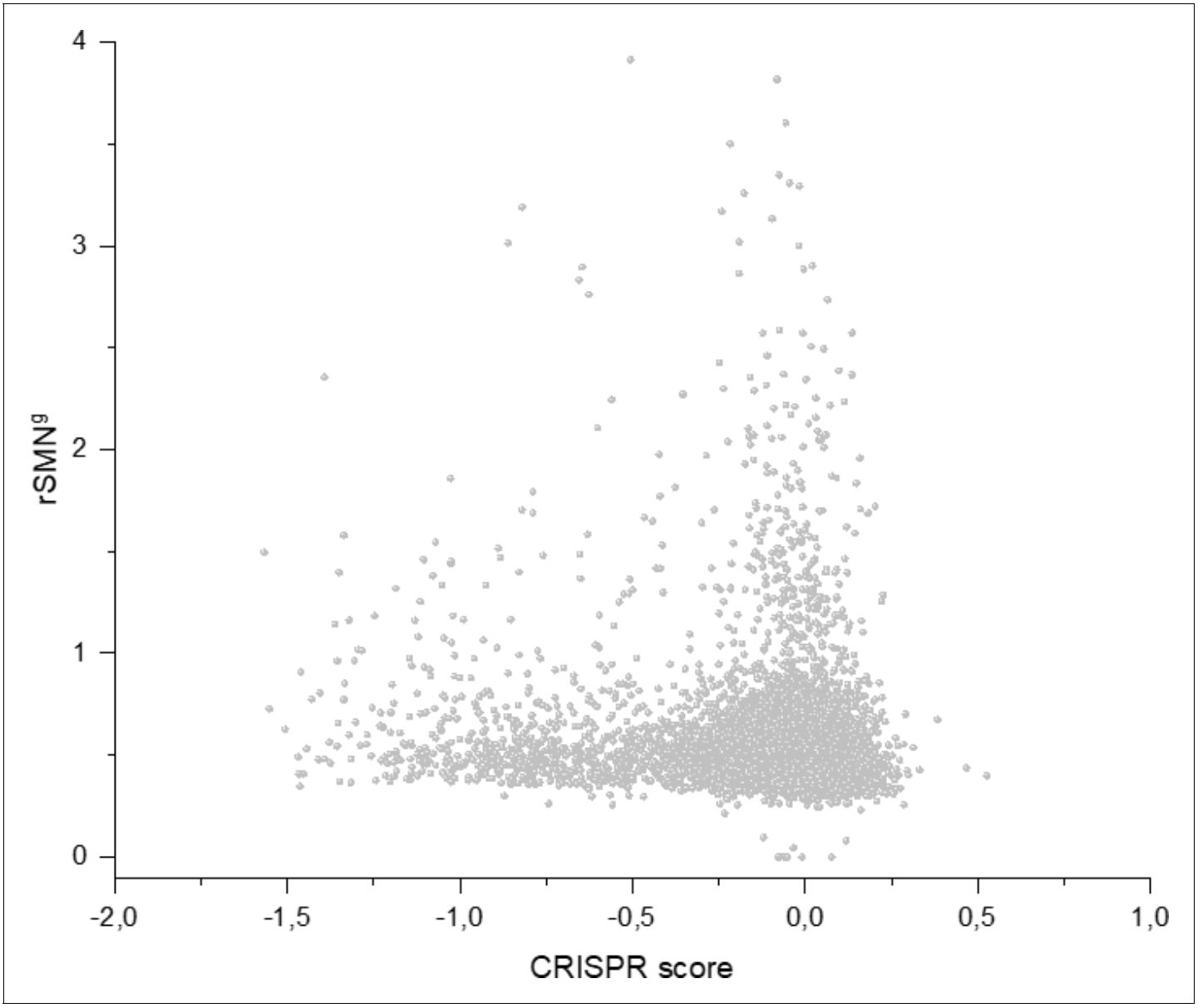


Figure 11. Cell-essentiality scores of human genes and negative selection on single-nucleotide polymorphisms (SNPs). The figure shows the results of the analysis of transcripts containing at least 100 polymorphic mutations. The abscissa indicates the cell-essentiality score of the genes, the ordinate shows the $rSMN^g$ parameters of the transcripts. Each ball represents a human transcript. Note that there is a weak negative correlation (Pearson's $r = -0.03662$, $p<0.05$) between the strength of purifying selection of transcripts ($rSMN^g$) and their cell-essentiality scores.

by four to five unknown cancer drivers) indicates that about half of the drivers is still unknown. If we assume that ~50% of cancer genes is still unknown 3–6% (0.5^5–0.5^4, i.e. 0.03125–0.0625 fraction) of tumors is expected to lack any of the known driver genes, and to be driven by four or five unknown driver mutations. Since the list of known drivers used in the study of the ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium (***Campbell et al., 2020***) comprises 722 driver genes, these observations suggest that hundreds of cancer driver genes remain to be identified.

In the present work, we have used analyses that combined multiple types of signals of selection, permitting improved detection of positive and negative selection. Our analyses have identified a large number of novel positively selected cancer gene candidates, many of which could be shown to play significant roles in carcinogenesis as tumor suppressors and OGs. Significantly, our analyses have identified a major group of human genes that show signs of negative selection during tumor

evolution, suggesting that the integrity of their function is essential for the growth and survival of tumor cells. Our analyses of representative members of negatively selected genes have confirmed that they play crucial pro-oncogenic roles in various cancer hallmarks (**Table 1**). It is important to emphasize that a survey of the group of OGs and pro-oncogenic TEGs reveals that they form a continuum in as much as there are numerous known OGs where negative selection also dominates (e.g. *ACKR3, BCL2*).

Although several groups have investigated the role of negative selection in tumor evolution earlier (**Zhou et al., 2017**; **Pyatnitskiy et al., 2015**; **Weghorn and Sunyaev, 2017**; **Martincorena et al., 2017**; **Zapata et al., 2018**; **López et al., 2020**; **Tilk et al., 2020**; **Van den Eynden and Larsson, 2017**), the study that received the greatest attention has reached the conclusion that negative selection has no role in tumor evolution (**Martincorena et al., 2017**; **Bakhoum and Landau, 2017**; **Koch, 2017**; **Vitale and Galluzzi, 2018**). The data presented here contradict the latter conclusion.

We believe that the approach reported here will promote the identification of numerous novel OGs, TSGs, and pro-oncogenic TEGs that may serve as therapeutic targets.

Materials and methods

Somatic mutation data

Cancer somatic mutation data were extracted from COSMIC v88, the Catalogue Of Somatic Mutations In Cancer (https://cancer.sanger.ac.uk/cosmic/download), which includes single nucleotide substitutions and small insertions/deletions affecting the coding sequence of human genes. The downloaded file (CosmicMutantExport.tsv, release v88) contained data for 29,415 transcripts (**Supplementary file 8**). For all subsequent analyses we have retained only transcripts containing mutations that were annotated under' Mutation description' as substitution or subtle insertion/deletion. This dataset contained data for 29,405 transcripts containing 6,449,721 mutations (substitution and short indels, SSI) and 29,399 transcripts containing 6,141,650 substitutions only (SO). **Supplementary file 9** contains the metadata for these SO and SSI datasets.

Since we were interested in the selection forces that operate during tumor evolution, only confirmed somatic mutations were included in our analyses. In COSMIC such mutations are annotated under' Mutation somatic status' as Confirmed Somatic, that is confirmed to be somatic in the experiment by sequencing both the tumor and a matched normal tissue from the same patient. **Supplementary file 10** indicates the contribution of major tumor types ('Tumor Primary site') to the somatic mutations of the dataset. As to' Sample Type, Tumor origin': we have excluded mutation data from cell-lines, organoid-cultures, xenografts since they do not properly represent human tumor evolution at the organism level. We have found that by excluding cell lines we have eliminated many artifacts of spurious recurrent mutations caused by repeated deposition of samples taken from the same cell-line at different time-points. To eliminate the influence of polymorphisms on the conclusions we retained only somatic mutations flagged 'n' for SNPs. (**Supplementary file 8**).

To increase the statistical power of our analyses, we have limited our work to transcripts that have at least 100 somatic mutations; **Supplementary file 5** contains the metadata for transcripts containing at least 100 confirmed somatic, non polymorphic mutations identified in tumor tissues. Hereafter, unless otherwise indicated, our analyses refer to datasets containing transcripts with at least 100 somatic mutations. This limitation eliminated ~38% of the transcripts that contain very few mutations but reduced the number of total mutations only by 9% (**Supplementary file 8**).

It should be noted that requiring a higher minimum number of somatic mutations increases the statistical power of the analyses but may disfavor the identification of negatively selected genes that tend to be undermutated. To assess the influence of the cut-off value of the minimum number of mutations on the robustness of the conclusions about negatively or positively selected genes, we have compared the results of analyses in which the minimum number of somatic mutation per gene was set as 0, 50, 100, or 500 (**Supplementary files 11–13**, **Figure 12**).

The choice of the minimum number of somatic mutations was found to have a strong influence on the pattern of observed fN, fS, and fM scores (**Supplementary files 11–13**, **Figure 12**). In the case of dataset N0 (no requirement for a minimum number of mutations), a large number of transcripts with less than 50 substitutions had scores of zero for one or two of the fN, fS, and fM parameters

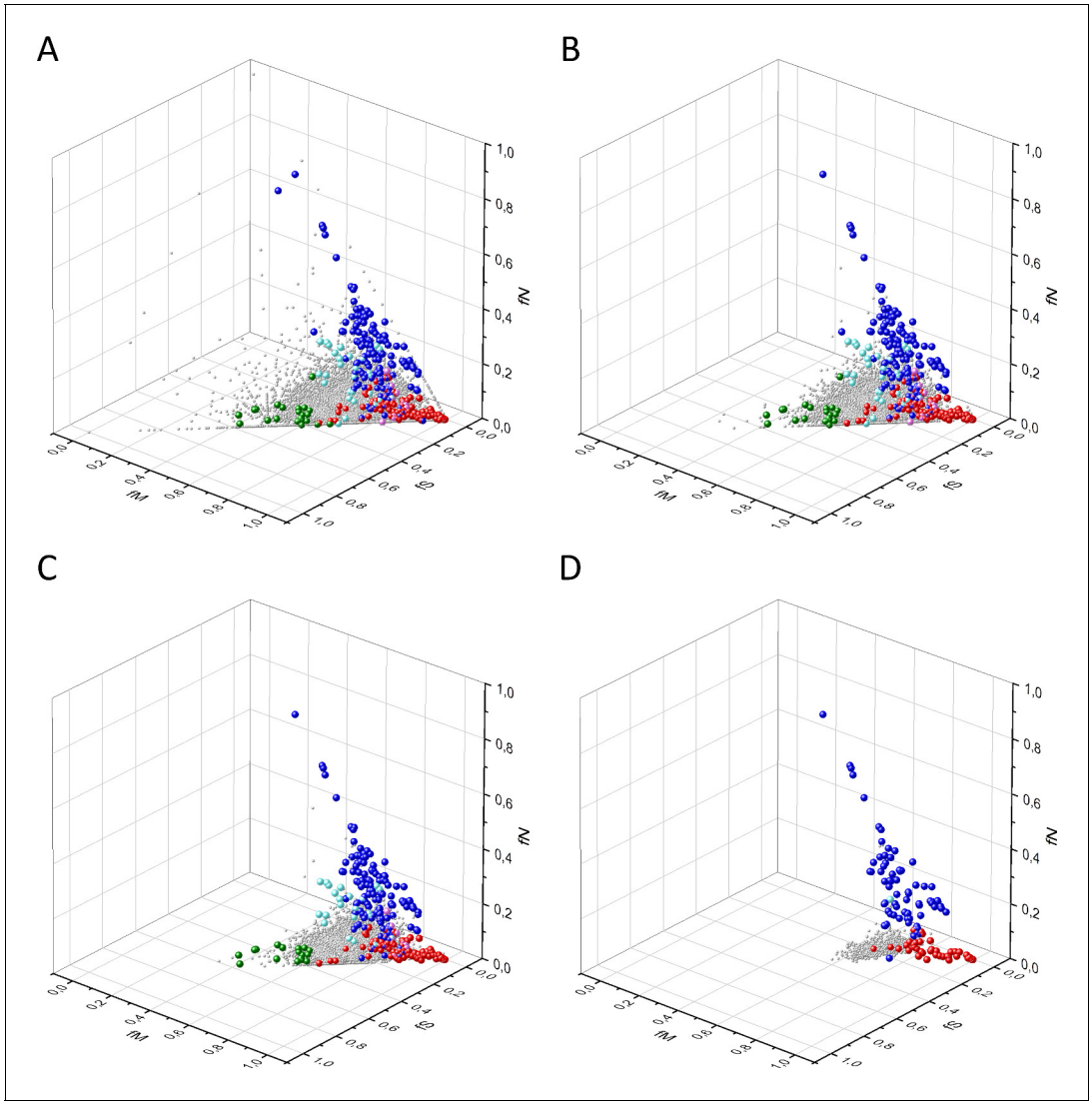


Figure 12. Analyses of fS, fM, and fN parameters of datasets N0, N50, N100, and N500 containing transcripts of human protein-coding genes with at least 0, 50, 100, or 500 somatic substitutions in tumors. The figure shows the results of the analysis of 29,333, 21,307, 13,803, and 997 transcripts present in datasets N0 (panel A), N50 (panel B), N100 (panel C), and N500 (panel D), respectively. Axes x, y, and z represent the fractions of somatic single nucleotide substitutions that are assigned to the synonymous (fS), nonsynonymous (fM), and nonsense (fN) categories. Each gray ball represents a human transcript. The positions of transcripts of the genes identified by ***Vogelstein et al., 2013*** as oncogenes (OGs, large red balls) or tumor suppressor genes (TSGs, large blue balls) are highlighted; novel proto-oncogenes, TSGs, and negatively selected tumor essential genes validated in the present work are represented by large magenta, cyan, and green balls, respectively. It is noteworthy that the requirement of at least 50 somatic mutations per transcript eliminates transcripts where the signal-to-noise ratio is too low to permit detection of signals of selection through the analysis of fS, fM, and fN parameters (compare panel A and panel B). It should also be noted that the requirement of at least 500 somatic mutations per transcript eliminates transcripts of negatively selected genes (compare panel C and panel D), consistent with the view that they tend to be undermutated.

due to the absence of somatic mutations in those categories (***Figure 12A***, ***Supplementary file 11***). Increasing the minimum number of somatic mutations per transcript to 50, 100, and 500 resulted in loss of these transcripts and elimination of a diffusely scattered group of transcripts that do not cluster with PGs, known OGs, TSGs, and TEGs (compare ***Figure 12A and B,C and D***, ***Supplementary file 11***). These observations indicate that we cannot draw valid conclusions about the significance of selection from fN, fS, and fM scores in cases where the number of mutations of a given transcript is too low to permit meaningful analyses. Our analyses have also revealed that a large proportion (22%) of the transcripts unique to the $N0^{2SD}$ dataset (containing fewer than 50

substitutions), correspond to short transcript fragments encoding less than 100 amino acids (***Supplementary file 12***). This finding suggests that the requirement for a minimum number of somatic mutations would not only increase statistical power, but also increases the biological relevance of the conclusions with the elimination of fragments that do not properly represent the full-length coding sequences.

Our analyses, however, have shown that the requirement of more than 500 somatic mutations per transcript (dataset N500) is too stringent. Although the majority of the 86 transcripts of the $N500^{2SD}$ dataset correspond to known OGs (25) or TSGs (48), most OGs and TSGs are not represented in the $N500^{2SD}$ dataset (***Supplementary file 12***). Furthermore, none of the negatively selected TEGs validated in the present study is present among the 86 transcripts of the $N500^{2SD}$ dataset or among the 997 transcripts of the N500 dataset (***Figure 12D***, ***Supplementary file 12***). This observation is consistent with the view that since negatively selected genes tend to have fewer mutations, they are less likely to pass the requirement for a high number of somatic mutations.

Our analyses suggest that the choice of 50 or 100 as the minimum number of somatic mutations per transcript represent acceptable trade-off between statistical power and loss of negatively or positively selected genes. As shown in ***Supplementary file 12***, both the $N50^{2SD}$ dataset (1846 transcripts) and the $N100^{2SD}$ dataset (1060 transcripts) contained the majority of known OGs, known TSGs, and TEGs validated in the present work, but since the choice of 100 offers higher statistical power, we have used this dataset in our analyses.

Our choice of 100 as the minimum number of somatic mutation per transcript may also have some relevance for the lack of more extensive similarity of our list of negatively selected genes with those identified by others. As shown in ***Supplementary file 13***, only 48%, 64%, 77%, and 89% of the negatively selected genes identified by ***Weghorn and Sunyaev, 2017***, ***Zapata et al., 2018***, ***Zhou et al., 2017***, and ***Pyatnitskiy et al., 2015***, respectively are present in dataset N100 containing 13,803 transcripts with at least 100 somatic mutations (***Supplementary file 11***). It thus appears that one of the reasons for the differences observed is that, with respect to minimum number of mutations, we have used rather stringent criteria to increase the robustness of our estimates. We wish to point out that, in order to obtain more reliable estimates of purifying selection, ***Pyatnitskiy et al., 2015*** have also excluded genes from their analysis that carried a low number of mutations but they excluded only those with less than 11 mutations.

The COSMIC database of somatic mutations used in the present study contains data obtained by three main types of sequencing: whole-genome sequencing (WGS), whole-exome sequencing (WES) and targeted sequencing. As shown in ***Supplementary file 8***, targeted screens provided substitution mutation data for only 13,120 transcripts of human genes, whereas genome wide screens covered 29,407 human transcripts as opposed to 29,415 transcripts covered by targeted plus genome wide screens. The contribution of targeted screens to somatic point mutations is even more restricted: only 508,124 (8.3%) of the 6,141,650 somatic point mutations of the entire COSMIC database were identified by targeted sequencing (***Supplementary file 8***). To check the impact of targeted sequencing on the dataset, in some analyzes we have used somatic mutation data only from genome-wide screens, excluding those obtained by targeted sequencing. We have found that omission of the data from targeted screens had no significant effect on the conclusions drawn from our analyses. Several factors may explain this observation. First, targeted screens usually focus on known cancer genes and they usually just reinforce the 'known cancer gene' status of the targeted genes. Second, since only a small fraction of the somatic mutations originates from targeted screens their impact is limited even in the case of the targeted genes. Finally, inclusion or omission of data from targeted screens has no impact on the number and pattern of mutations of non-targeted genes identified in genome wide screens.

Germline mutation data

Information on SNPs affecting the coding regions of human genes was downloaded from the dbSNP database (https://www.ncbi.nlm.nih.gov/snp/). For each SNP, we extracted nucleotide and amino acid variants from the original dbSNP file. In cases where two or three mutant variant was reported for a specific rsID, each variant was treated as an independent polymorphism. The retrieved SNPs were assigned to three functional categories: (i) Nonsense or Stop_gained mutations (N), which change an amino acid-encoding codon into a stop codon, (ii) Missense mutations (M), which change an amino acid into a mutant amino acid, and (iii) Synonymous or silent mutations (S), which do not

change the amino acid. We have focused only on SNPs of genes that were also found to contain at least 100 confirmed somatic, non polymorphic mutations in the COSMIC database (***Supplementary file 5***). ***Supplementary file 6*** shows the numbers and fractions of SNPs affecting the coding sequences of the various human genes, according to the functional categories of the point mutations.

Substitution metrics

The 61 sense codons can undergo 549 single base substitutions and, depending on the wild type and mutant codon, each substitution can be assigned to the silent, missense or nonsense mutation category. Out of the 549 single-base substitutions, 392 result in missense mutation, 134 lead to silent mutation, and 23 generate nonsense mutation, thus – assuming equal codon frequency, equal probability of the different types of substitutions and neutrality – the expected fractions of nonsense, missense and silent substituions are fN = 0.04189, fM = 0.71403, and fS = 0.24408, respectively.

Codons, however, differ significantly in the probability that their mutation would lead to nonsense (N), missense (M), or silent (S) mutation (***Supplementary files 14***, ***15***, ***16***) and since the 61 sense codons (amino acids) do not occur with the same frequency in the coding region of human genes this may have a significant influence on the expected fN, fM, and fS values. We have calculated the probability that a substitution would lead to nonsense, missense or silent mutation taking into account the codon frequency of the proteome of *Homo sapiens* (https://www.kazusa.or.jp/codon/cgi-bin/showcodon.cgi?species=9606). This calculation yielded values of fN = 0.0419, fM = 0.7299, fS = 0.2282 for the proteome, slightly different from the values of fN = 0.0419, fM = 0.7140, fS = 0.2441, assuming equal frequency of codons.

The amino acid composition and codon usage of some individual proteins (especially short fragments) may deviate significantly from average, therefore we have calculated the expected proportion of silent, missense, and nonsense mutations for all transcripts, assuming equal probability of different substitutions classes (***Supplementary file 17***). For these calculations, we have downloaded the coding sequences of 53,190 transcripts of human protein coding genes (All_COSMIC_Genes.fasta.gz) from the COSMIC database (https://cancer.sanger.ac.uk/cosmic) and their codon usage and amino acid composition were determined using the SMS server (https://www.bioinformatics.org/sms2/codon_usage.html, ***Stothard, 2000***).

Different classes of substitutions, however, do not occur with equal probability, moreover the various normal and tumor tissues show characteristic differences in the spectrum of substitutions classes (***Alexandrov et al., 2013***; ***Alexandrov et al., 2020***). Substitutions are assigned to six classes (C>A, C>G, C>T, T>A, T>C, and T>G) referred to by the pyrimidine of the mutated Watson–Crick base pair. It is of crucial importance to take differences in the probability of the six mutation classes into account since—due to the unique structure of the genetic code—the six types of substitutions differ markedly in the probability that they would lead to nonsense (N), missense (M), or silent (S) mutation of the coding region of protein-coding genes. As shown in ***Supplementary files 18–25***, there are significant differences in the impact of different substitution classes on the expected proportion of missense, silent, and nonsense mutations of codons (assuming equal codon frequency). For example, the dominance of C>G increases the proportion of missense substitutions, whereas higher rates of C>T and T>C substitutions increase the proportion of silent substitutions. Since mutation bias favoring C>T substitutions is expected to decrease the ratio of missense to silent mutations, decreased dN/dS values may not be taken as evidence for negative selection in the case of tumors, such as malignant melanoma, where the vast majority of all somatic mutations is C>T substitution (***Van den Eynden and Larsson, 2017***).

To take into account differences in mutation bias, we have calculated the contribution of the C>A, C>G, C>T, T>A, T>C, and T>G mutations to the pattern of single base substitutions in tumors. We have downloaded the files containing 'Mutational Signatures v3.1' and 'Attributions of the SBS Signatures to Mutations in Tumors' from the COSMIC website (https://cancer.sanger.ac.uk/cosmic/signatures/SBS/index.tt). The contributions of C>A, C>G, C>T, T>A, T>C, and T>G mutations to the pattern of Single Base Substitutions in the tumors listed in the PCAWG_sigProfiler_SBS_signatures_in_samples file are summarized in ***Supplementary file 26***. The C>T substitution accounts for the largest fraction of substitutions in most tumors (0.3726), followed by T>C (0.1842), C>A (0.1583), C>G (0.1162), T>G (0.0891), and T>A (0.0796). There are, however, differences in the relative contribution of the six mutation classes to different tumors. For example, the contribution of

C>A mutation is higher than average for colon cancer and lung cancer, the role of C>G mutation is above average for bladder cancer and some breast cancers. The contribution of C>T mutation is very high in the case of skin-melanoma, whereas the T>A mutation contributes significantly to some kidney cancers. The T>C mutation plays a significant role in biliary and liver cancer, whereas the T>G mutation is more significant in colon cancer and esophageal cancer than in other tumors (see ***Supplementary file 26***).

In order to correct for the influence of mutation bias on fN, fM, and fS values of transcripts in tumor tissues, we have calculated the expected fN*, fM*, and fS* values for all human transcripts using the average values of the six substitution types observed across tumors (***Supplementary file 27***). It is noteworthy that the average values of expected fN*, fM*, and fS* (fN*=0.04483, fM*=0.69114, and fS*=0.26402) are similar to those (fN = 0.04189, fM = 0.71403, and fS = 0.24408) assuming equal codon frequency and equal probability of the different types of substitutions.

In the case of germline cells, we have also calculated the expected fN**, fM**, and fS** values for all human transcripts using the mutation probabilities characteristic of these cells (***Supplementary file 28***). It has been shown earlier that the human germline mutation spectrum can be recapitulated by a combination of the cancer signatures SBS1 and SBS5 (***Alexandrov et al., 2015***; ***Rahbari et al., 2016***; ***Heredia-Genestar et al., 2020***). In the present work, we have combined the effect of mutation signatures SBS1 and SBS5 on the germline mutation spectrum of proteins according to the formula (0.1 × SBS1 + 0.9 × SBS5) recommended by ***Heredia-Genestar et al., 2020***. It is noteworthy that the average values of expected fN**, fM**, and fS** (fN**=0.03791, fM**=0.68653, and fS**=0.27556) are similar to those expected for tumor tissues (fN*=0.04483, fM*=0.69114, and fS*=0.26402).

Per-gene detection of selection signals in tumor tissues

We have used two approaches to determine the observed fM, fS, and fN values of transcripts: one in which we have restricted our analyses to single nucleotide substitutions (hereafter referred to as SO for 'substitution only') and a version in which we have also taken into account subtle indels (hereafter referred to as SSI for 'substitutions and subtle indels').

In the first case, we have calculated for each transcript the fraction of somatic substitutions that could be assigned to the synonymous (fS), nonsynonymous (fM), and nonsense mutation (fN) category (***Supplementary file 5*** and ***9***). In the version that also included data for subtle indels, we have calculated the fraction of mutations corresponding to synonymous substitutions (indel_fS), but have merged nonsynonymous substitutions and short inframe indels in the category of mutations that lead to changes in the amino acid sequence (indel_fM). Nonsense substitutions and short frame-shift indels were included in the third category of mutations (indel_fN) as both types of mutation lead eventually to stop codons that truncate the protein (***Supplementary file 5*** and ***9***).

Analyses of datasets (***Supplementary file 5***) containing substitutions only have shown that in 3D scatter plots transcripts form a cluster (***Figure 2A***) characterized by values of 0.2436 ± 0.0619, 0.7090 ± 0.0556, and 0.0475 ± 0.0322 for fractions of silent, missense, and nonsense substitutions, respectively. The mean fS, fM, and fN values of the transcripts in this cluster are close to those expected if we assume that the structure of the genetic code has the most important role in determining the probability of somatic substitutions during tumor evolution of human genes (***Supplementary file 29***). Based on the structure of the genetic code, assuming equal usage of the codons and equal probability of different point mutations, in the absence of selection one would expect that a fraction of 0.24408 would be silent, 0.71403 of the single-base substitutions would be missense and 0.04189 would be nonsense mutations.

It is noteworthy, however, that the fS, fM, and fN values of the best known cancer genes (***Vogelstein et al., 2013***) deviate from those characteristic of the majority of human genes (***Figure 2B***). The genes in the central cluster, deviating from mean fM, fS, and fN values by ≤1 SD, are characterized by fraction values of 0.24548 ± 0.03079, 0.71084 ± 0.0274, and 0.04368 ± 0.01572 for synonymous, nonsynonymous and nonsense substitutions, respectively. Note that these values are very close to those expected from the structure of the genetic code in the absence of selection, assuming equal frequency of codons and equal probability of the different classes of mutations (***Supplementary file 29***). This central cluster of genes (***Supplementary file 5***) is hereafter referred to as PG_SO$^{f_{-1SD}}$ (for Passenger Gene_Substitution Only deviating from mean fM, fS, and fN values by ≤1 SD) because it is likely to be enriched in genes that play no major role in carcinogenesis.

In harmony with earlier observations, the values for OGs show a significant ($p<0.05$) shift of fM to higher values (0.8563 ± 0.08224) relative to those of PGs (0.71084 ± 0.0274), reflecting positive selection for missense mutations (***Supplementary file 29***). On the other hand, the fN values of TSGs are significantly ($p<0.05$) higher (0.1964 ± 0.11063) than those of PGs (0.04368 ± 0.01572), reflecting positive selection for truncating nonsense mutations (***Supplementary file 29***).

The genes (1060 transcripts) with values that deviate from mean values of fS, fM, and fN by more than 2SD, however, are likely to be subject to selection. In harmony with this expectation, this group contains transcripts of the majority of known driver genes (62 OG and 119 TSG driver gene transcripts). This gene set, defined by 2SD cut-off value, is hereafter referred to as CG_SO^{f_2SD} (for Cancer Gene_Substitution Only deviating from mean fM, fS, and fN values by more than 2SD) because it is likely to be enriched in cancer genes (***Supplementary file 29***). Out of the 1060 transcripts present in CG_SO^{f_2SD}, 737 transcripts are derived from genes that are not included in the OG, TSG, and CGC cancer gene lists (***Supplementary files 5*** and ***29***). Since the majority of these 737 transcripts have parameters that cluster them with known OGs or TSGs, we assume that they qualify as candidate OGs or TSGs. However, a group of genes deviates from both the central PG cluster and the clusters of OGs and TSGs (***Figure 2C***). The high fS and low fM and fN values of the genes in this cluster suggest that they experience purifying selection during tumor evolution, raising the possibility that they may correspond to TEGs important for the growth and survival of tumors.

Known cancer genes (OGs and TSGs) also separate from the majority of human genes in 3D scatter plots of parameters rSM, rNM, rNS defined as the ratio of fS/fM, fN/fM, fN/fS, respectively (***Figure 3***). The central cluster of genes that deviate from mean rSM, rNM and rNS values by $\leq$1 SD is hereafter referred to as PG_SO^{r2_1SD} (Passenger Gene_Substitution Only deviating from mean rSM, rNM, and rNS values by $\leq$1 SD) since it is likely to be enriched in PGs. Conversely, the group of transcripts that deviate from mean rSM, rNM, and rNS values by more than 2SD is referred to as CG_SO^{r2_2SD} (Cancer Gene_Substitution Only deviating from mean rSM, rNM, rNS values by more than 2SD) because it is likely to be enriched in cancer genes (***Supplementary file 29***). The CG_SOr2_2SD gene set (780 transcripts) contains the majority of driver gene transcripts (40 transcripts of OGs, 103 transcripts of TSGs genes), 79 transcripts of CGC genes and 558 transcripts derived from 468 genes that are not found in the OG, TSG, and CGC cancer gene lists (***Supplementary file 29***).

In these scatter plots OGs separate from the central cluster in having significantly ($p<0.05$) lower rSM (0.13971 ± 0.10621) and rNM (0.03936 ± 0.0313) values than those of the central cluster of PGs (rSM: 0.34523 ± 0.06137; rNM: 0.0607 ± 0.02595, ***Supplementary file 29***), reflecting positive selection for missense mutations and negative selection of nonsense mutations. Interestingly, in these plots some OGs (e.g. BCL2) have unusually high values of rSM and low values of rNM (e.g. ***Figure 3A1, A2*** and ***Supplementary file 5***) suggesting that in the case of these OGs purifying selection may dominate over positive selection for amino acid changing mutations.

TSGs also separate from the central cluster: they have significantly ($p<0.05$) higher rNS (3.92588 ± 5.66261) and rNM (0.31524 ± 0.31575) values than those of PGs (rNS: 0.18403 ± 0.09138; rNM: 0.0607 ± 0.02595; ***Figure 3A1, A2***. ***Supplementary file 29***), reflecting the dominance of positive selection for inactivating mutations.

As mentioned above, the candidate cancer gene set defined by a cut-off value of 2SD also contains 558 transcripts derived from 468 genes that are not found in the OG, TSG, or CGC lists.

Since the majority of these 558 transcripts have parameters that cluster them with known OGs or TSGs, they can be regarded as candidate OGs or TSGs. There is, however, a group of genes that deviate from the clusters of PGs, OGs, and TSGs in that they have unusually high rSM values and low rNM and rNS values. Since these values may be indicative of purifying selection, we assumed that they might correspond to TEGs important for the growth and survival of tumors.

The separation of known cancer genes from the majority of human genes is even more obvious in 3D scatter plots of parameters rSMN, rMSN, and rNSM defined as the ratio of fS/(fM+fN), fM/(fS+fN), and fN/(fS+fM), respectively (***Figure 4 A1, A2***). In these plots, the gene transcripts are present in a three-pronged cluster, with OGs and TSG being present on separate spikes of this cluster (***Figure 4***).

We refer to the central cluster of genes, deviating from mean rSMN, rMSN, and rNSM values by $\leq$1 SD as PG_SO^{r3_1SD} (Passenger Gene_Substitution Only deviating from mean rSMN, rMSN, and rNSM values by $\leq$1 SD) as they are likely to be enriched in PGs. Similarly, we refer to the gene

set defined by 2SD cut-off value (***Supplementary files 5*** and ***29***) as CG_SO^{r3_2SD} (Cancer Gene_Substitution Only deviating from mean rSMN, rMSN, and rNSM values by more than 2SD) as it is likely to be enriched in candidate cancer genes. This gene set has 751 transcripts, containing the majority of transcripts of known driver genes (35 OGs, 103 TSGs), 80 transcripts of CGC genes and 533 transcripts (derived from 448 genes) not found in the OG, TSG, and CGC cancer gene lists (***Supplementary files 5*** and ***29***).

The mean parameters of TSGs differ significantly ($p<0.05$) from those of PGs in as much as rNSM values of TSGs are higher (0.27937 ± 0.2783) but rSMN (0.10865 ± 0.06128) values are lower than those of PGs (rNSM: 0.04812 ± 0.02561; rSMN: 0.3259 ± 0.09265, ***Supplementary file 29***), reflecting the dominance of positive selection for inactivating nonsense mutations.

In the case of OGs the rMSN values are significantly ($p<0.05$) higher (15.35971 ± 30.07472) and the rSMN values are significantly lower (0.13363 ± 0.10266) than those of PGs (rMSN: 2.58911 ± 0.68355; rSMN: 0.3259±0.09265 ***Supplementary file 29***), reflecting positive selection for missense mutations. The rNSM values of OGs (0.03394 ± 0.02621) are also significantly ($p<0.05$) lower than those of PGs (0.04812 ± 0.02561), reflecting purifying selection avoiding nonsense mutations. Interestingly, some OGs have unusually high scores of rSMN (***Figure 4 A1, A2***, ***Supplementary file 5***) suggesting that in these cases (e.g. *BCL2*) purifying selection dominates over positive selection for amino acid changing mutations.

As mentioned above, the candidate cancer gene set defined by a cut-off value of 2SD contains 533 transcripts (derived from 448 genes) not found in the OG, TSG, or CGC lists. Since the majority of these genes have parameters that assign them to the clusters containing OGs or TSGs, they can be regarded as candidate OGs or TSGs. There is, however, a group of genes that deviates from the clusters of PGs, OGs, and TSGs (***Figure 4***). Their high rSMN and low rMSN and rNSM values suggest that they experience purifying selection during tumor evolution, raising the possibility that this group may be enriched in genes essential for the survival of tumors as pro-oncogenes or TEGs.

The three types of analyses described for Substitutions Only (illustrated in ***Figures 2–4***) were also carried out for datasets in which both substitutions and subtle indels (Substitutions and Subtle Indels, SSI) were used (for details of these analyzes see Appendix 2).

Comparison of the data obtained by SO and SSI analyses (***Supplementary file 5***) revealed that inclusion of indels has only minor influence on the separation of the clusters of PGs and CGs. The lists of PGs identified with 1SD cut-off values for SO analyes (PG_SO^{f_1SD}, PG_SO^{r2_1SD}, PG_SO^{r3_1SD}) and SSI analyses (PG_SSI^{f_1SD}, PG_SSI^{r2_1SD}, PG_SSI^{r3_1SD}) show more than 90% identity in the case of the relevant SO/SSI pairs (***Supplementary file 30***). Similarly, the lists of CGs identified with 2SD cut-off values for SO analyses (CG_SO^{f_2SD}, CG_SO^{r2_2SD}, CG_SO^{r3_2SD}) and SSI analyses (CG_SSI^{f_2SD}, CG_SSI^{r2_2SD}, CG_SSI^{r3_2SD}) show 78%, 87%, and 92% identity, respectively, for the relevant SO/SSI pairs (***Supplementary file 30***).

The parameters of the 1158 transcripts present in at least one of the various CG_SO^{2SD} lists and the 1333 transcripts present in at least one of the various CG_SSI^{2SD} lists (***Supplementary file 31***) were used to assign them to three distinct clusters. (1) Cluster of genes positively selected for missense mutations and negatively selected for nonsense mutations; (2) Cluster of genes positively selected for nonsense mutations; (3) Clusters of negatively selected genes (see ***Figure 2C***, ***Figure 3 B1, B2*** and ***Figure 4 B1, B2***). To check the validity and predictive value of the assumption that the genes assigned to these clusters play significant roles in carcinogenesis, we have selected a number of genes for further analyses from the 1457 transcripts present in the combined list (CG_SO^{2SD}_SSI^{2SD}) of candidate cancer genes (***Supplementary file 31***). The results of these analyses are summarized in the Results section.

As outlined in the section on Substitution metrics, a limitation of the analyses discussed above is that they did not take into account the impact of differences in mutation probability on the fN, fM, and fS values of transcripts. In order to eliminate this source of error, we have calculated the expected fN*, fM*, and fS* values for all human transcripts using the probability of the six substitution types observed across tumors (***Supplementary file 27***). The various types of observed/expected ratios (rN*, rM*, rS*; rSM*, rNM*, rNS*; rSMN*, rMSN*, and rNSM*) of somatic mutations were calculated for all transcripts (***Supplementary file 32***) and the data were analyzed in 3D scatter plots as described above for the observed values.

As shown in ***Figures 5–7***, the distribution of transcripts in these 3D scatter plots are similar to those observed in the corresponding ***Figures 2–4***, in that known OGs, TSGs, and TEGs are separated from the central cluster of PGs as well as from each other (***Supplementary file 32***).

Per-gene detection of selection signals in the database of human single-nucleotide polymorphisms

As a reference, we have carried out similar analyses of the fN, fM, and fS parameters of germline mutations, through the analysis of the human database of human single-nucleotide polymorphisms (SNPs; ***Supplementary file 6***). ***Supplementary file 33*** contains the various types of observed/expected ratios (rN**, rM**, rS**; rSM**, rNM**, rNS**; rSMN**, rMSN**, and rNSM**) of germline mutations calculated for all transcripts. Data were analyzed in 3D scatter plots as described for somatic mutations. Details of these analyses are presented in the Results section.

Cancer gene list

As the gold standard of 'known' cancer genes we have used the lists of OG and TSGs identified by ***Vogelstein et al., 2013***. As another list of known cancer genes we have also used the genes of the Cancer Gene Census (***Sondka et al., 2018***).

Statistical analyses

The statistical package of Origin 2018 was used for all data processing and statistical analysis. We report details of statistical tests in the Supplementary files of the respective sections. Statistical significance was set as a p value of < 0.05.

Acknowledgements

LB, KK, MT, and LP are supported by the GINOP-2.3.2-15-2016-00001 grant of the Hungarian National Research, Development and Innovation Office (NKFIH), OC is supported by the NVKP_16-1-2016-0005 grant of the Hungarian National Research, Development and Innovation Office (NKFIH).

Additional information

Funding

Funder	Grant reference number	Author
Hungarian National Research, Development and Innovation Office	GINOP-2.3.2-15-2016-00001	László Bányai Maria Trexler Krisztina Kerekes László Patthy
Hungarian National Research, Development and Innovation Office	NVKP_16-1-2016-0005	Orsolya Csuka

The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

Author contributions

László Bányai, Formal analysis, Validation, Investigation, Methodology, Writing - original draft, Writing - review and editing; Maria Trexler, Krisztina Kerekes, Formal analysis, Validation, Writing - original draft, Writing - review and editing; Orsolya Csuka, Funding acquisition, Validation, Writing - original draft, Writing - review and editing; László Patthy, Conceptualization, Supervision, Funding acquisition, Validation, Methodology, Writing - original draft, Project administration, Writing - review and editing

Author ORCIDs

László Patthy https://orcid.org/0000-0003-1329-0484

Decision letter and Author response
Decision letter https://doi.org/10.7554/eLife.59629.sa1
Author response https://doi.org/10.7554/eLife.59629.sa2

Additional files

Supplementary files

- Supplementary file 1. Comparison of the lists of genes in datasets CG_SSI^{2SD}_rNSM > 0.125 and CG_SO^{2SD}_rMSN > 3.00 with the lists of cancer genes identified by others (VOG, ***Vogelstein et al., 2013***; TAM, ***Tamborero et al., 2013***; LAW, ***Lawrence et al., 2014***; ABB, ***Abbott et al., 2015***; TOR, ***Torrente et al., 2016***; ZHO, ***Zhou et al., 2017***; MAR, ***Martincorena et al., 2017***; BAI, ***Bailey et al., 2018***; SON, ***Sondka et al., 2018***; ZHA, ***Zhao et al., 2019a***). Transcripts of OGs (oncogenes) and TSGs (tumor suppressor genes) of the cancer gene list of ***Vogelstein et al., 2013*** are highlighted by brick red and blue backgrounds, respectively. Transcripts of CGC genes (SON, ***Sondka et al., 2018***) that do not correspond to OGs or TSGs of the cancer gene list of ***Vogelstein et al., 2013*** are highlighted by yellow background. Novel positively or negatively selected cancer genes validated in the present work are highlighted in dark green background.

- Supplementary file 2. Comparison of the lists of genes in datasets CG_SSI^{2SD}_rNSM > 0.125 and CG_SO^{2SD}_rMSN > 3.00 with the lists of genes in datasets CG_SO*^{2SD}_rNSM > 3 and CG_SO*^{2SD}_rMSN > 1.50, respectively.

- Supplementary file 3. Comparison of the list of negatively selected genes, CG^{2SD}_rSMN > 0.5 with the lists of negatively selected genes (WEG, ZHOU, ZAPATA, PYATNITSKIY), defined by ***Zhou et al., 2017***, ***Weghorn and Sunyaev, 2017***, ***Zapata et al., 2018***, ***Pyatnitskiy et al., 2015***, respectively as well as the list of genes (De Kegel) identified by ***De Kegel and Ryan, 2019*** as broadly essential genes. Negatively selected genes discussed in detail in the present work are highlighted in dark green background.

- Supplementary file 4. Comparison of the list of genes in dataset CG^{2SD}_rSMN > 0.5 with the list of genes in dataset CG_SO*^{2SD}_rSMN > 1.50.

- Supplementary file 5. SO (Substitution Only) and SSI (Substitutions and Subtle Indel) analyses of somatic mutations of transcripts of human protein coding genes that have at least 100 confirmed somatic, non-polymorphic mutations identified in tumor tissues. The table also contains lists of passenger genes (PG_SO^{f_1SD}, PG_SO^{r2_1SD}, PG_SO^{r3_1SD}, PG_SSI^{f_1SD}, PG_SSI^{r2_1SD}, PG_SSI^{r3_1SD}) whose parameters deviate from the mean values by $\leq$1 SD as well as lists of candidate cancer genes (CG_SO^{f_1SD}, CG_SO^{r2_1SD}, CG_SO^{r3_1SD}, CG_SSI^{f_1SD}, CG_SSI^{r2_1SD}, CG_SSI^{r3_1SD}) whose parameters deviate from the mean values by >1 SD. Table also contains lists of candidate cancer genes (CG_SO^{f_2SD}, CG_SO^{r2_2SD}, CG_SO^{r3_2SD}, CG_SSI^{f_2SD}, CG_SSI^{r2_2SD}, CG_SSI^{r3_2SD}) whose parameters deviate from the mean values by >2 SD as well as lists of passenger genes (PG_SO^{f_2SD}, PG_SO^{r2_2SD}, PG_SO^{r3_2SD}, PG_SSI^{f_2SD}, PG_SSI^{r2_2SD}, PG_SSI^{r3_2SD}) whose parameters deviate from the mean values by <2 SD. Transcripts of OGs (oncogenes) and TSGs (tumor suppressor genes) of the cancer gene list of ***Vogelstein et al., 2013*** are highlighted by brick red and blue backgrounds, respectively. Transcripts of CGC (Cancer Gene Census) genes (***Sondka et al., 2018***) that do not correspond to OGs or TSGs of the cancer gene list of ***Vogelstein et al., 2013*** are highlighted by yellow background.

- Supplementary file 6. Numbers and fractions of missense, nonsense, and silent single-nucleotide polymorphisms (SNPs) affecting the coding sequences of the human genes. Transcripts of OGs (oncogenes) and TSGs (tumor suppressor genes) of the cancer gene list of ***Vogelstein et al., 2013*** are highlighted by brick red and blue backgrounds, respectively. Transcripts of CGC genes (SON, ***Sondka et al., 2018***) that do not correspond to OGs or TSGs of the cancer gene list of ***Vogelstein et al., 2013*** are highlighted by yellow background. Novel positively or negatively selected cancer genes validated in the present work are highlighted in dark green background.

- Supplementary file 7. Comparison of fS, rSM, and rSMN scores of genes determined for somatic mutations in tumors with those determined for germline mutations.

• Supplementary file 8. Statistics of transcripts and subtle somatic mutations of human protein coding genes of the different datasets analyzed.

• Supplementary file 9. SO (Substitution Only) and SSI (Substitutions and Subtle Indel) analyses of somatic mutations of transcripts of human protein coding genes. Transcripts of OGs (oncogenes) and TSGs (tumor suppressor genes) of the cancer gene list of ***Vogelstein et al., 2013*** are highlighted by brick red and blue backgrounds, respectively. Transcripts of CGC (Cancer Gene Census) genes (***Sondka et al., 2018***) that do not correspond to OGs or TSGs of the cancer gene list of ***Vogelstein et al., 2013*** are highlighted by yellow background.

• Supplementary file 10. Contribution of major types of tumors ('Tumor Primary site') to subtle somatic substitutions of the human protein coding genes analyzed.

• Supplementary file 11. Analyses of fS, fM, and fN parameters of transcripts of human protein coding genes that have at least 0 (N0), 50 (N50), 100 (N100), or 500 (N500) somatic substitutions in tumors, respectively. Transcripts of OGs (oncogenes) and TSGs (tumor suppressor genes) of the cancer gene list of ***Vogelstein et al., 2013*** are highlighted by brick red and light blue backgrounds, respectively. Transcripts of CGC (Cancer Gene Census) genes (***Sondka et al., 2018***) that do not correspond to OGs or TSGs of the cancer gene list of ***Vogelstein et al., 2013*** are highlighted by yellow background. Novel proto-oncogenes, TSGs and negatively selected tumor essential genes validated in the present work are shown in brown, dark blue, and green colors, respectively. For 3D representations of the data, see ***Figure 12***.

• Supplementary file 12. Analyses of fS, fM, and fN parameters of transcripts of human protein coding genes that have at least 0 ($N0^{2SD}$), 50 ($N50^{2SD}$), 100 ($N100^{2SD}$), or 500 ($N500^{2SD}$) somatic substitutions in tumors and deviate from average values of fS, fM, and fN by more than 2SD (Sheet 'CG_SOf_2SD'). Transcripts of OGs (oncogenes) and TSGs (tumor suppressor genes) of the cancer gene list of ***Vogelstein et al., 2013*** are highlighted by brick red and light blue backgrounds, respectively. Transcripts of CGC (Cancer Gene Census) genes (SON, ***Sondka et al., 2018***) that do not correspond to OGs or TSGs of the cancer gene list of ***Vogelstein et al., 2013*** are highlighted by yellow background. Novel proto-oncogenes, TSGs and negatively selected tumor essential genes (TEGs) validated in the present work are shown in brown, dark blue, and green colors, respectively. Sheet 'statistics' contains a summary of the fS, fM, and fN parameters of datasets N0, N50, N100, N500, $N0^{2SD}$, $N50^{2SD}$, $N100^{2SD}$, $N500^{2SD}$ and indicates the number of known and novel OGs, TSGs and TEGs that are present in the different datasets.

• Supplementary file 13. Negatively selected genes in datasets N0, N50, N100, and N500. Sheet 'SO' lists the genes/transcripts in datasets N0, N50, N100, and N500 that contain transcripts of human protein coding genes with at least 0, 50, 100, or 500 somatic substitutions in tumors, respectively. The lists of negatively selected genes identified by others were taken from the publications of ***Weghorn and Sunyaev, 2017***, ***Zapata et al., 2018***, ***Zhou et al., 2017*** and ***Pyatnitskiy et al., 2015***. Sheet 'statistics' indicates the number of negatively selected genes identified by others that are present in the N0, N50, N100, and N500 datasets. Note that only 48%, 64%, 77%, and 89% of the negatively selected genes identified by ***Weghorn and Sunyaev, 2017***, ***Zapata et al., 2018***, ***Zhou et al., 2017*** and ***Pyatnitskiy et al., 2015***, respectively, are present in the dataset N100 that we have analyzed in the present work.

• Supplementary file 14. Expected fractions of nonsense, missense, and silent substitutions of various codons in the absence of selection assuming that there is no difference in the probability of the substitution classes C>A, C>G, C>T, T>A, T>C, and T>G.

• Supplementary file 15. Expected fractions of nonsense, missense, and silent substitutions of various codons in the absence of selection assuming that there is no difference in the probability of the substitution classes C>A, C>G, C>T, T>A, T>C, and T>G.

• Supplementary file 16. Expected fractions of nonsense, missense, and silent substitutions of various codons in the absence of selection assuming that there is no difference in the probability of the substitution classes C>A, C>G, C>T, T>A, T>C, and T>G.

• Supplementary file 17. Expected fraction of silent, missense, and nonsense mutations of coding sequences of human protein-coding genes, assuming equal probability of different substitutions classes.

- Supplementary file 18. Expected fractions of nonsense, missense, and silent substitutions of various codons in the absence of selection assuming that only C>A and G>T mutations occur.
- Supplementary file 19. Expected fractions of nonsense, missense, and silent substitutions of various codons in the absence of selection assuming that only C>G and G>C mutations occur.
- Supplementary file 20. Expected fractions of nonsense, missense, and silent substitutions of various codons in the absence of selection assuming that only C>T and G>A mutations occur.
- Supplementary file 21. Expected fractions of nonsense, missense, and silent substitutions of various codons in the absence of selection assuming that only T>A and A>T mutations occur.
- Supplementary file 22. Expected fractions of nonsense, missense, and silent substitutions of various codons in the absence of selection assuming that only T>C and A>G mutations occur.
- Supplementary file 23. Expected fractions of nonsense, missense, and silent substitutions of various codons in the absence of selection assuming that only T>G and A>C mutations occur.
- Supplementary file 24. Expected fractions of nonsense, missense and silent substitutions of various codons in the absence of selection assuming that only C>A or C>G or C>T or T>A or T>C or T>G mutations occur.
- Supplementary file 25. Expected fractions of nonsense, missense, and silent substitutions in the absence of selection assuming equal codon frequency and that only C>A or C>G or C>T or T>A or T>C or T>G mutations occur.
- Supplementary file 26. Contributions of C>A, C>G, C>T, T>A, T>C, and T>G mutations to the pattern of Single Base Substitutions in tumors.
- Supplementary file 27. Expected fractions of nonsense (fN*), missense (fM*), and silent (fS*) mutations of human protein-coding genes taking into account the probability of different substitutions classes in tumors.
- Supplementary file 28. Expected fractions of nonsense (fN**), missense (fM**), and silent (fS**) mutations of human protein-coding genes taking into account the probability of different substitutions classes in germline cells.
- Supplementary file 29. Statistics of the results of SO (Substitution Only) and SSI (Substitutions and Subtle Indel) analyses of the data presented in ***Supplementary file 5***. The column marked 'Expected' indicates the parameters expected if we assume that the structure of the genetic code determines the probability of somatic substitutions.
- Supplementary file 30. Comparison of the results of SO (Substitution Only) and SSI (Substitutions and Subtle Indel) analyses.
- Supplementary file 31. Lists of genes ($CG\_SO^{f\_2SD}$, $CG\_SO^{r2\_2SD}$, $CG\_SO^{r3\_2SD}$, $CG\_SSI^{f\_2SD}$, $CG\_SSI^{r2\_2SD}$, $CG\_SSI^{r3\_2SD}$) whose parameters deviate from the mean values by >2 SD. Transcripts of OGs (oncogenes) and TSGs (tumor suppressor genes) of the cancer gene list of ***Vogelstein et al., 2013*** are highlighted by brick red and blue backgrounds, respectively. Transcripts of CGC (Cancer Gene Census) genes (***Sondka et al., 2018***) that do not correspond to OGs or TSGs of the cancer gene list of ***Vogelstein et al., 2013*** are highlighted by yellow background.
- Supplementary file 32. Observed/expected parameters (rN*, rM*, rS*; rSM*, rNM*, rNS*; rSMN*, rMSN*, and rNSM*) of somatic mutations affecting the coding sequences of the human genes in cancer. Transcripts of OGs (oncogenes) and TSGs (tumor suppressor genes) of the cancer gene list of ***Vogelstein et al., 2013*** are highlighted by brick red and blue backgrounds, respectively.
- Supplementary file 33. Observed/expected parameters (rN**, rM**, rS**; rSM**, rNM**, rNS**; rSMN**, rMSN**, and rNSM**) of single-nucleotide polymorphisms (SNPs) affecting the coding sequences of the human genes. Transcripts of OGs (oncogenes) and TSGs (tumor suppressor genes) of the cancer gene list of ***Vogelstein et al., 2013*** are highlighted by brick red and blue backgrounds, respectively.
- Transparent reporting form

Data availability

All data generated or analysed during this study are included in the manuscript and supporting files.

The following datasets were generated:

References

Abbott KL, Nyre ET, Abrahante J, Ho YY, Isaksson Vogel R, Starr TK. 2015. The candidate Cancer gene database: a database of Cancer driver genes from forward genetic screens in mice. *Nucleic Acids Research* **43**: D844–D848. DOI: https://doi.org/10.1093/nar/gku770, PMID: 25190456

Ablain J, Xu M, Rothschild H, Jordan RC, Mito JK, Daniels BH, Bell CF, Joseph NM, Wu H, Bastian BC, Zon LI, Yeh I. 2018. Human tumor genomics and zebrafish modeling identify *SPRED1* loss as a driver of mucosal melanoma. *Science* **362**:1055–1060. DOI: https://doi.org/10.1126/science.aau6509, PMID: 30385465

Adesina AM, Nguyen Y, Mehta V, Takei H, Stangeby P, Crabtree S, Chintagumpala M, Gumerlock MK. 2007. FOXG1 dysregulation is a frequent event in medulloblastoma. *Journal of Neuro-Oncology* **85**:111–122. DOI: https://doi.org/10.1007/s11060-007-9394-3, PMID: 17522785

Adesina AM, Veo BL, Courteau G, Mehta V, Wu X, Pang K, Liu Z, Li XN, Peters L. 2015. FOXG1 expression shows correlation with neuronal differentiation in cerebellar development, aggressive phenotype in Medulloblastomas, and survival in a xenograft model of medulloblastoma. *Human Pathology* **46**:1859–1871. DOI: https://doi.org/10.1016/j.humpath.2015.08.003, PMID: 26433703

Aldinucci D, Casagrande N. 2018. Inhibition of the CCL5/CCR5 Axis against the progression of gastric Cancer. *International Journal of Molecular Sciences* **19**:1477. DOI: https://doi.org/10.3390/ijms19051477, PMID: 29772686

Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SA, Behjati S, Biankin AV, Bignell GR, Bolli N, Borg A, Børresen-Dale AL, Boyault S, Burkhardt B, Butler AP, Caldas C, Davies HR, Desmedt C, Eils R, Eyfjörd JE, Foekens JA, Greaves M, et al. 2013. Signatures of mutational processes in human Cancer. *Nature* **500**:415–421. DOI: https://doi.org/10.1038/nature12477, PMID: 23945592

Alexandrov LB, Jones PH, Wedge DC, Sale JE, Campbell PJ, Nik-Zainal S, Stratton MR. 2015. Clock-like mutational processes in human somatic cells. *Nature Genetics* **47**:1402–1407. DOI: https://doi.org/10.1038/ng.3441, PMID: 26551669

Alexandrov LB, Kim J, Haradhvala NJ, Huang MN, Tian Ng AW, Wu Y, Boot A, Covington KR, Gordenin DA, Bergstrom EN, Islam SMA, Lopez-Bigas N, Klimczak LJ, McPherson JR, Morganella S, Sabarinathan R, Wheeler DA, Mustonen V, Getz G, Rozen SG, et al. 2020. The repertoire of mutational signatures in human Cancer. *Nature* **578**:94–101. DOI: https://doi.org/10.1038/s41586-020-1943-3, PMID: 32025018

Alves IT, Cano D, Böttcher R, van der Korput H, Dinjens W, Jenster G, Trapman J. 2017. A mononucleotide repeat in PRRT2 is an important, frequent target of mismatch repair deficiency in Cancer. *Oncotarget* **8**:6043–6056. DOI: https://doi.org/10.18632/oncotarget.13464, PMID: 27907910

Anandakrishnan R, Varghese RT, Kinney NA, Garner HR. 2019. Estimating the number of genetic mutations (hits) required for carcinogenesis based on the distribution of somatic mutations. *PLOS Computational Biology* **15**: e1006881. DOI: https://doi.org/10.1371/journal.pcbi.1006881, PMID: 30845172

Andersen AP, Samsøe-Petersen J, Oernbo EK, Boedtkjer E, Moreira JMA, Kveiborg M, Pedersen SF. 2018. The net acid extruders NHE1, NBCn1 and MCT4 promote mammary tumor growth through distinct but overlapping mechanisms. *International Journal of Cancer* **142**:2529–2542. DOI: https://doi.org/10.1002/ijc.31276, PMID: 29363134

Aoki Y, Niihori T, Banjo T, Okamoto N, Mizuno S, Kurosawa K, Ogata T, Takada F, Yano M, Ando T, Hoshika T, Barnett C, Ohashi H, Kawame H, Hasegawa T, Okutani T, Nagashima T, Hasegawa S, Funayama R, Nagashima T, et al. 2013. Gain-of-function mutations in RIT1 cause Noonan syndrome, a RAS/MAPK pathway syndrome. *The American Journal of Human Genetics* **93**:173–180. DOI: https://doi.org/10.1016/j.ajhg.2013.05.021, PMID: 23791108

Avitabile M, Succoio M, Testori A, Cardinale A, Vaksman Z, Lasorsa VA, Cantalupo S, Esposito M, Cimmino F, Montella A, Formicola D, Koster J, Andreotti V, Ghiorzo P, Romano MF, Staibano S, Scalvenzi M, Ayala F, Hakonarson H, Corrias MV, et al. 2020. Neural crest-derived tumor neuroblastoma and melanoma share 1p13.2 as susceptibility locus that shows a long-range interaction with the SLC16A1 gene. *Carcinogenesis* **41**:284–295. DOI: https://doi.org/10.1093/carcin/bgz153, PMID: 31605138

Baek G, Tse YF, Hu Z, Cox D, Buboltz N, McCue P, Yeo CJ, White MA, DeBerardinis RJ, Knudsen ES, Witkiewicz AK. 2014. MCT4 defines a glycolytic subtype of pancreatic Cancer with poor prognosis and unique metabolic dependencies. *Cell Reports* **9**:2233–2249. DOI: https://doi.org/10.1016/j.celrep.2014.11.025, PMID: 25497091

Baenke F, Dubuis S, Brault C, Weigelt B, Dankworth B, Griffiths B, Jiang M, Mackay A, Saunders B, Spencer-Dene B, Ros S, Stamp G, Reis-Filho JS, Howell M, Zamboni N, Schulze A. 2015. Functional screening identifies MCT4 as a key regulator of breast Cancer cell metabolism and survival. *The Journal of Pathology* **237**:152–165. DOI: https://doi.org/10.1002/path.4562, PMID: 25965974

Bailey MH, Tokheim C, Porta-Pardo E, Sengupta S, Bertrand D, Weerasinghe A, Colaprico A, Wendl MC, Kim J, Reardon B, Kwok-Shing Ng P, Jeong KJ, Cao S, Wang Z, Gao J, Gao Q, Wang F, Liu EM, Mularoni L, Rubio-Perez C, et al. 2018. Comprehensive characterization of Cancer driver genes and mutations. *Cell* **174**:1034–1035. DOI: https://doi.org/10.1016/j.cell.2018.07.034, PMID: 30096302

Bakhoum SF, Landau DA. 2017. Cancer evolution: no room for negative selection. *Cell* **171**:987–989. DOI: https://doi.org/10.1016/j.cell.2017.10.039, PMID: 29149612

Barajas JM, Reyes R, Guerrero MJ, Jacob ST, Motiwala T, Ghoshal K. 2018. The role of miR-122 in the dysregulation of glucose-6-phosphate dehydrogenase (G6PD) expression in hepatocellular Cancer. *Scientific Reports* **8**:9105. DOI: https://doi.org/10.1038/s41598-018-27358-5, PMID: 29904144

Bashashati A, Ha G, Tone A, Ding J, Prentice LM, Roth A, Rosner J, Shumansky K, Kalloger S, Senz J, Yang W, McConechy M, Melnyk N, Anglesio M, Luk MT, Tse K, Zeng T, Moore R, Zhao Y, Marra MA, et al. 2013. Distinct evolutionary trajectories of primary high-grade serous ovarian cancers revealed through spatial mutational profiling. *The Journal of Pathology* **231**:21–34. DOI: https://doi.org/10.1002/path.4230, PMID: 23780408

Benachenhou N, Labuda D, Sinnett D. 1998. Allelic instability of TBP gene in replication error positive tumors. *International Journal of Cancer* **78**:525–526. DOI: https://doi.org/10.1002/(SICI)1097-0215(19981109)78:4<525::AID-IJC21>3.0.CO;2-3, PMID: 9797144

Berger AH, Imielinski M, Duke F, Wala J, Kaplan N, Shi GX, Andres DA, Meyerson M. 2014. Oncogenic RIT1 mutations in lung adenocarcinoma. *Oncogene* **33**:4418–4423. DOI: https://doi.org/10.1038/onc.2013.581, PMID: 24469055

Bernassola F, Chillemi G, Melino G. 2019. HECT-Type E3 Ubiquitin Ligases in Cancer. *Trends in Biochemical Sciences* **44**:1057–1075. DOI: https://doi.org/10.1016/j.tibs.2019.08.004, PMID: 31610939

Beroukhim R, Mermel CH, Porter D, Wei G, Raychaudhuri S, Donovan J, Barretina J, Boehm JS, Dobson J, Urashima M, Mc Henry KT, Pinchback RM, Ligon AH, Cho YJ, Haery L, Greulich H, Reich M, Winckler W, Lawrence MS, Weir BA, et al. 2010. The landscape of somatic copy-number alteration across human cancers. *Nature* **463**:899–905. DOI: https://doi.org/10.1038/nature08822, PMID: 20164920

Bi G, Yan J, Sun S, Qu X. 2017. PRRT2 inhibits the proliferation of glioma cells by modulating unfolded protein response pathway. *Biochemical and Biophysical Research Communications* **485**:454–460. DOI: https://doi.org/10.1016/j.bbrc.2017.02.052, PMID: 28192116

Bisso A, Collavin L, Del Sal G. 2011. p73 as a pharmaceutical target for Cancer therapy. *Current Pharmaceutical Design* **17**:578–590. DOI: https://doi.org/10.2174/138161211795222667, PMID: 21391908

Bose P, Thakur SS, Brockton NT, Klimowicz AC, Kornaga E, Nakoneshny SC, Riabowol KT, Dort JC. 2014. Tumor cell apoptosis mediated by cytoplasmic *ING1* is associated with improved survival in oral squamous cell carcinoma patients. *Oncotarget* **5**:3210–3219. DOI: https://doi.org/10.18632/oncotarget.1907, PMID: 24912621

Buisson R, Langenbucher A, Bowen D, Kwan EE, Benes CH, Zou L, Lawrence MS. 2019. Passenger hotspot mutations in Cancer driven by APOBEC3A and mesoscale genomic features. *Science* **364**:eaaw2872. DOI: https://doi.org/10.1126/science.aaw2872, PMID: 31249028

Calabrese C, Davidson NR, Demircioğlu D, Fonseca NA, He Y, Kahles A, Lehmann KV, Liu F, Shiraishi Y, Soulette CM, Urban L, Greger L, Li S, Liu D, Perry MD, Xiang Q, Zhang F, Zhang J, Bailey P, Erkek S, et al. 2020. Genomic basis for RNA alterations in Cancer. *Nature* **578**:129–136. DOI: https://doi.org/10.1038/s41586-020-1970-0, PMID: 32025019

Campbell AJ, Lyne L, Brown PJ, Launchbury RJ, Bignone P, Chi J, Roncador G, Lawrie CH, Gatter KC, Kusec R, Banham AH. 2010. Aberrant expression of the neuronal transcription factor *FOXP2* in neoplastic plasma cells. *British Journal of Haematology* **149**:221–230. DOI: https://doi.org/10.1111/j.1365-2141.2009.08070.x, PMID: 20096010

Campbell PJ, Getz G, Korbel JO, ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. 2020. Pan-cancer analysis of whole genomes. *Nature* **578**:82–93. DOI: https://doi.org/10.1038/s41586-020-1969-6, PMID: 32025007

Campos EI, Martinka M, Mitchell DL, Dai DL, Li G. 2004. Mutations of the *ING1* tumor suppressor gene detected in human melanoma abrogate nucleotide excision repair. *International Journal of Oncology* **25**:73–80. DOI: https://doi.org/10.3892/ijo.25.1.73, PMID: 15201991

Carter H, Chen S, Isik L, Tyekucheva S, Velculescu VE, Kinzler KW, Vogelstein B, Karchin R. 2009. Cancer-specific high-throughput annotation of somatic mutations: computational prediction of driver missense mutations. *Cancer Research* **69**:6660–6667. DOI: https://doi.org/10.1158/0008-5472.CAN-09-1133, PMID: 19654296

Carter H. 2019. Mutation hotspots may not be drug targets. *Science* **364**:1228–1229. DOI: https://doi.org/10.1126/science.aax9108, PMID: 31249043

Chaplet M, Waltregny D, Detry C, Fisher LW, Castronovo V, Bellahcène A. 2006. Expression of dentin sialophosphoprotein in human prostate Cancer and its correlation with tumor aggressiveness. *International Journal of Cancer* **118**:850–856. DOI: https://doi.org/10.1002/ijc.21442, PMID: 16108038

Chatterjee A, Rodger EJ, Eccles MR. 2018. Epigenetic drivers of tumourigenesis and Cancer metastasis. *Seminars in Cancer Biology* **51**:149–159. DOI: https://doi.org/10.1016/j.semcancer.2017.08.004, PMID: 28807546

Chen L, Matsubara N, Yoshino T, Nagasaka T, Hoshizima N, Shirakawa Y, Naomoto Y, Isozaki H, Riabowol K, Tanaka N. 2001. Genetic alterations of candidate tumor suppressor *ING1* in human esophageal squamous cell Cancer. *Cancer Research* **61**:4345–4349. PMID: 11389058

Chen CL, Wang Y, Pan QZ, Tang Y, Wang QJ, Pan K, Huang LX, He J, Zhao JJ, Jiang SS, Zhang XF, Zhang HX, Zhou ZQ, Weng deS, Xia JC. 2016. Bromodomain-containing protein 7 (BRD7) as a potential tumor suppressor in hepatocellular carcinoma. *Oncotarget* **7**:16248–16261. DOI: https://doi.org/10.18632/oncotarget.7637, PMID: 26919247

Chen YL, Huang WC, Yao HL, Chen PM, Lin PY, Feng FY, Chu PY. 2017a. Down-regulation of RASA1 is associated with poor prognosis in human hepatocellular carcinoma. *Anticancer Research* **37**:781–786. DOI: https://doi.org/10.21873/anticanres.11377, PMID: 28179330

Chen YC, Gotea V, Margolin G, Elnitski L. 2017b. Significant associations between driver gene mutations and DNA methylation alterations across many Cancer types. *PLOS Computational Biology* **13**:e1005840. DOI: https://doi.org/10.1371/journal.pcbi.1005840, PMID: 29125844

Chen Y, Zhang Y, Guo X. 2017c. Proteasome dysregulation in human Cancer: implications for clinical therapies. *Cancer and Metastasis Reviews* **36**:703–716. DOI: https://doi.org/10.1007/s10555-017-9704-y, PMID: 29039081

Chen Q, Meng YQ, Xu XF, Gu J. 2017d. Blockade of GLUT1 by WZB117 resensitizes breast Cancer cells to adriamycin. *Anti-Cancer Drugs* **28**:880–887. DOI: https://doi.org/10.1097/CAD.0000000000000529, PMID: 28609310

Chen J, Wu X, Xing Z, Ma C, Xiong W, Zhu X, He X. 2018a. FOXG1 expression is elevated in glioma and inhibits glioma cell apoptosis. *Journal of Cancer* **9**:778–783. DOI: https://doi.org/10.7150/jca.22282, PMID: 29581755

Chen MT, Sun HF, Li LD, Zhao Y, Yang LP, Gao SP, Jin W. 2018b. Downregulation of FOXP2 promotes breast Cancer migration and invasion through TGFβ/SMAD signaling pathway. *Oncology Letters* **15**:8582–8588. DOI: https://doi.org/10.3892/ol.2018.8402, PMID: 29805593

Chen X, Xu Z, Zhu Z, Chen A, Fu G, Wang Y, Pan H, Jin B. 2018c. Modulation of G6PD affects bladder Cancer via ROS accumulation and the AKT pathway in vitro. *International Journal of Oncology* **53**:1703–1712. DOI: https://doi.org/10.3892/ijo.2018.4501, PMID: 30066842

Chen X, Yu C, Gao J, Zhu H, Cui B, Zhang T, Zhou Y, Liu Q, He H, Xiao R, Huang R, Xie H, Gao D, Zhou H. 2018d. A novel USP9X substrate TTK contributes to tumorigenesis in non-small-cell lung Cancer. *Theranostics* **8**:2348–2360. DOI: https://doi.org/10.7150/thno.22901, PMID: 29721084

Chen X, Chen X, Liu F, Yuan Q, Zhang K, Zhou W, Guan S, Wang Y, Mi S, Cheng Y. 2019. Monocarboxylate transporter 1 is an independent prognostic factor in esophageal squamous cell carcinoma. *Oncology Reports* **41**:2529–2539. DOI: https://doi.org/10.3892/or.2019.6992, PMID: 30720131

Cheng J, Demeulemeester J, Wedge DC, Vollan HKM, Pitt JJ, Russnes HG, Pandey BP, Nilsen G, Nord S, Bignell GR, White KP, Børresen-Dale AL, Campbell PJ, Kristensen VN, Stratton MR, Lingjærde OC, Moreau Y, Van Loo P. 2017. Pan-cancer analysis of homozygous deletions in primary tumours uncovers rare tumour suppressors. *Nature Communications* **8**:1221. DOI: https://doi.org/10.1038/s41467-017-01355-0, PMID: 29089486

Choi SY, Xue H, Wu R, Fazli L, Lin D, Collins CC, Gleave ME, Gout PW, Wang Y. 2016. The *MCT4* gene: a novel, potential target for therapy of advanced prostate Cancer. *Clinical Cancer Research* **22**:2721–2733. DOI: https://doi.org/10.1158/1078-0432.CCR-15-1624, PMID: 26755530

Choi SYC, Ettinger SL, Lin D, Xue H, Ci X, Nabavi N, Bell RH, Mo F, Gout PW, Fleshner NE, Gleave ME, Collins CC, Wang Y. 2018. Targeting MCT4 to reduce lactic acid secretion and glycolysis for treatment of neuroendocrine prostate Cancer. *Cancer Medicine* **7**:3385–3392. DOI: https://doi.org/10.1002/cam4.1587

Cuiffo BG, Campagne A, Bell GW, Lembo A, Orso F, Lien EC, Bhasin MK, Raimo M, Hanson SE, Marusyk A, El-Ashry D, Hematti P, Polyak K, Mechta-Grigoriou F, Mariani O, Volinia S, Vincent-Salomon A, Taverna D, Karnoub AE. 2014. MSC-regulated microRNAs converge on the transcription factor FOXP2 and promote breast Cancer metastasis. *Cell Stem Cell* **15**:762–774. DOI: https://doi.org/10.1016/j.stem.2014.10.001, PMID: 25515522

Dali R, Verginelli F, Pramatarova A, Sladek R, Stifani S. 2018. Characterization of a FOXG1:tle1 transcriptional network in glioblastoma-initiating cells. *Molecular Oncology* **12**:775–787. DOI: https://doi.org/10.1002/1878-0261.12168, PMID: 29316219

de Castro TB, Mota AL, Bordin-Junior NA, Neto DS, Zuccari DAPC. 2019. Immunohistochemical expression of melatonin receptor MT1 and glucose transporter GLUT1 in human breast Cancer. *Anti-Cancer Agents in Medicinal Chemistry* **18**:2110–2116. DOI: https://doi.org/10.2174/1871520618666181025125532, PMID: 30360728

De Kegel B, Ryan CJ. 2019. Paralog buffering contributes to the variable essentiality of genes in Cancer cell lines. *PLOS Genetics* **15**:e1008466. DOI: https://doi.org/10.1371/journal.pgen.1008466, PMID: 31652272

De Paoli L, Cerri M, Monti S, Rasi S, Spina V, Bruscaggin A, Greco M, Ciardullo C, Famà R, Cresta S, Maffei R, Ladetto M, Martini M, Laurenti L, Forconi F, Marasca R, Larocca LM, Bertoni F, Gaidano G, Rossi D. 2013. *MGA*, a suppressor of *MYC*, is recurrently inactivated in high risk chronic lymphocytic leukemia. *Leukemia & Lymphoma* **54**:1087–1090. DOI: https://doi.org/10.3109/10428194.2012.723706, PMID: 23039309

Deb TB, Zuo AH, Barndt RJ, Sengupta S, Jankovic R, Johnson MD. 2015. Pnck overexpression in HER-2 gene-amplified breast Cancer causes trastuzumab resistance through a paradoxical PTEN-mediated process. *Breast Cancer Research and Treatment* **150**:347–361. DOI: https://doi.org/10.1007/s10549-015-3337-z, PMID: 25773930

DeBerardinis RJ, Lum JJ, Hatzivassiliou G, Thompson CB. 2008. The biology of Cancer: metabolic reprogramming fuels cell growth and proliferation. *Cell Metabolism* **7**:11–20. DOI: https://doi.org/10.1016/j.cmet.2007.10.002, PMID: 18177721

Del Reino P, Alsina-Beauchamp D, Escós A, Cerezo-Guisado MI, Risco A, Aparicio N, Zur R, Fernandez-Estévez M, Collantes E, Montans J, Cuenda A. 2014. Pro-oncogenic role of alternative p38 mitogen-activated protein kinases p38γ and p38δ, linking inflammation and Cancer in colitis-associated Colon cancer. *Cancer Research* **74**:6150–6160. DOI: https://doi.org/10.1158/0008-5472.CAN-14-0870, PMID: 25217523

Deng Y, Zou J, Deng T, Liu J. 2018. Clinicopathological and prognostic significance of GLUT1 in breast Cancer: a meta-analysis. *Medicine* **97**:e12961. DOI: https://doi.org/10.1097/MD.0000000000012961, PMID: 30508885

Di Domenico A, Wiedmer T, Marinoni I, Perren A. 2017. Genetic and epigenetic drivers of neuroendocrine tumours (NET). *Endocrine-Related Cancer* **24**:R315–R334. DOI: https://doi.org/10.1530/ERC-17-0012, PMID: 28710117

Diao H, Ye Z, Qin R. 2018. miR-23a acts as an oncogene in pancreatic carcinoma by targeting FOXP2. *Journal of Investigative Medicine* **66**:676–683. DOI: https://doi.org/10.1136/jim-2017-000598, PMID: 29141872

Diederichs S, Bartsch L, Berkmann JC, Fröse K, Heitmann J, Hoppe C, Iggena D, Jazmati D, Karschnia P, Linsenmeier M, Maulhardt T, Möhrmann L, Morstein J, Paffenholz SV, Röpenack P, Rückert T, Sandig L, Schell M, Steinmann A, Voss G, et al. 2016. The dark matter of the Cancer genome: aberrations in regulatory elements, untranslated regions, splice sites, non-coding RNA and synonymous mutations. *EMBO Molecular Medicine* **8**:442–457. DOI: https://doi.org/10.15252/emmm.201506055, PMID: 26992833

Doherty JR, Yang C, Scott KE, Cameron MD, Fallahi M, Li W, Hall MA, Amelio AL, Mishra JK, Li F, Tortosa M, Genau HM, Rounbehler RJ, Lu Y, Dang CV, Kumar KG, Butler AA, Bannister TD, Hooper AT, Unsal-Kacmaz K, et al. 2014. Blocking lactate export by inhibiting the myc target MCT1 disables glycolysis and glutathione synthesis. *Cancer Research* **74**:908–920. DOI: https://doi.org/10.1158/0008-5472.CAN-13-2034, PMID: 24285728

Domínguez G, García JM, Peña C, Silva J, García V, Martínez L, Maximiano C, Gómez ME, Rivera JA, García-Andrade C, Bonilla F. 2006. DeltaTAp73 upregulation correlates with poor prognosis in human tumors: putative in vivo network involving p73 isoforms, p53, and E2F-1. *Journal of Clinical Oncology* **24**:805–815. DOI: https://doi.org/10.1200/JCO.2005.02.2350, PMID: 16380414

Doyen J, Trastour C, Ettore F, Peyrottes I, Toussant N, Gal J, Ilc K, Roux D, Parks SK, Ferrero JM, Pouysségur J. 2014. Expression of the hypoxia-inducible monocarboxylate transporter MCT4 is increased in triple negative breast Cancer and correlates independently with clinical outcome. *Biochemical and Biophysical Research Communications* **451**:54–61. DOI: https://doi.org/10.1016/j.bbrc.2014.07.050, PMID: 25058459

Dvinge H, Kim E, Abdel-Wahab O, Bradley RK. 2016. RNA splicing factors as oncoproteins and tumour suppressors. *Nature Reviews Cancer* **16**:413–430. DOI: https://doi.org/10.1038/nrc.2016.51, PMID: 27282250

Editorial. 2020. The era of massive Cancer sequencing projects has reached a turning point. *Nature* **578**:7–8. DOI: https://doi.org/10.1038/d41586-020-00308-w, PMID: 32025026

Fan PD, Narzisi G, Jayaprakash AD, Venturini E, Robine N, Smibert P, Germer S, Yu HA, Jordan EJ, Paik PK, Janjigian YY, Chaft JE, Wang L, Jungbluth AA, Middha S, Spraggon L, Qiao H, Lovly CM, Kris MG, Riely GJ, et al. 2018. *YES1* amplification is a mechanism of acquired resistance to EGFR inhibitors identified by transposon mutagenesis and clinical genomics. *PNAS* **115**:E6030–E6038. DOI: https://doi.org/10.1073/pnas.1717782115, PMID: 29875142

Fang Y, Wang J, Wang G, Zhou C, Wang P, Zhao S, Zhao S, Huang S, Su W, Jiang P, Chang A, Xiang R, Sun P. 2017. Inactivation of p38 MAPK contributes to stem cell-like properties of non-small cell lung Cancer. *Oncotarget* **8**:26702–26717. DOI: https://doi.org/10.18632/oncotarget.15804, PMID: 28460458

Firestein R, Bass AJ, Kim SY, Dunn IF, Silver SJ, Guney I, Freed E, Ligon AH, Vena N, Ogino S, Chheda MG, Tamayo P, Finn S, Shrestha Y, Boehm JS, Jain S, Bojarski E, Mermel C, Barretina J, Chan JA, et al. 2008. CDK8 is a colorectal Cancer oncogene that regulates beta-catenin activity. *Nature* **455**:547–551. DOI: https://doi.org/10.1038/nature07179, PMID: 18794900

Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R, Rahman N, Stratton MR. 2004. A census of human Cancer genes. *Nature Reviews Cancer* **4**:177–183. DOI: https://doi.org/10.1038/nrc1299, PMID: 14993899

Ganapathy V, Thangaraju M, Prasad PD. 2009. Nutrient transporters in Cancer: relevance to warburg hypothesis and beyond. *Pharmacology & Therapeutics* **121**:29–40. DOI: https://doi.org/10.1016/j.pharmthera.2008.09.005, PMID: 18992769

Gao L, Smit MA, van den Oord JJ, Goeman JJ, Verdegaal EM, van der Burg SH, Stas M, Beck S, Gruis NA, Tensen CP, Willemze R, Peeper DS, van Doorn R. 2013. Genome-wide promoter methylation analysis identifies epigenetic silencing of MAPK13 in primary cutaneous melanoma. *Pigment Cell & Melanoma Research* **26**:542–554. DOI: https://doi.org/10.1111/pcmr.12096, PMID: 23590314

Gao Y, Wang B, Gao S. 2016. *BRD7* acts as a tumor suppressor gene in lung adenocarcinoma. *PLOS ONE* **11**:e0156701. DOI: https://doi.org/10.1371/journal.pone.0156701, PMID: 27580131

Gardner HP, Ha SI, Reynolds C, Chodosh LA. 2000. The caM kinase, pnck, is spatially and temporally regulated during murine mammary gland development and may identify an epithelial cell subtype involved in breast Cancer. *Cancer Research* **60**:5571–5577. PMID: 11034105

Garmendia I, Pajares MJ, Hermida-Prado F, Ajona D, Bértolo C, Sainz C, Lavín A, Remírez AB, Valencia K, Moreno H, Ferrer I, Behrens C, Cuadrado M, Paz-Ares L, Bustelo XR, Gil-Bazo I, Alameda D, Lecanda F, Calvo A, Felip E, et al. 2019. YES1 drives lung Cancer growth and progression and predicts sensitivity to dasatinib. *American Journal of Respiratory and Critical Care Medicine* **200**:888–899. DOI: https://doi.org/10.1164/rccm.201807-1292OC, PMID: 31166114

Ge Z, Leighton JS, Wang Y, Peng X, Chen Z, Chen H, Sun Y, Yao F, Li J, Zhang H, Liu J, Shriver CD, Hu H, Piwnica-Worms H, Ma L, Liang H, Cancer Genome Atlas Research Network. 2018. Integrated genomic analysis of the ubiquitin pathway across Cancer types. *Cell Reports* **23**:213–226. DOI: https://doi.org/10.1016/j.celrep.2018.03.047, PMID: 29617661

Gerstung M, Eriksson N, Lin J, Vogelstein B, Beerenwinkel N. 2011. The temporal order of genetic and pathway alterations in tumorigenesis. *PLOS ONE* **6**:e27136. DOI: https://doi.org/10.1371/journal.pone.0027136, PMID: 22069497

Gerstung M, Jolly C, Leshchiner I, Dentro SC, Gonzalez S, Rosebrock D, Mitchell TJ, Rubanova Y, Anur P, Yu K, Tarabichi M, Deshwar A, Wintersinger J, Kleinheinz K, Vázquez-García I, Haase K, Jerman L, Sengupta S, Macintyre G, Malikic S, et al. 2020. The evolutionary history of 2,658 cancers. *Nature* **578**:122–128. DOI: https://doi.org/10.1038/s41586-019-1907-7, PMID: 32025013

Gkouveris I, Nikitakis NG, Aseervatham J, Ogbureke KUE. 2018. The tumorigenic role of DSPP and its potential regulation of the unfolded protein response and ER stress in oral Cancer cells. *International Journal of Oncology* **53**:1743–1751. DOI: https://doi.org/10.3892/ijo.2018.4484, PMID: 30015841

Goldman NA, Katz EB, Glenn AS, Weldon RH, Jones JG, Lynch U, Fezzari MJ, Runowicz CD, Goldberg GL, Charron MJ. 2006. GLUT1 and GLUT8 in endometrium and endometrial adenocarcinoma. *Modern Pathology* **19**:1429–1436. DOI: https://doi.org/10.1038/modpathol.3800656, PMID: 16892013

Gonzalez-Perez A, Sabarinathan R, Lopez-Bigas N. 2019. Local determinants of the mutational landscape of the human genome. *Cell* **177**:101–114. DOI: https://doi.org/10.1016/j.cell.2019.02.051, PMID: 30901533

Gorlov IP, Kimmel M, Amos CI. 2006. Strength of the purifying selection against different categories of the point mutations in the coding regions of the human genome. *Human Molecular Genetics* **15**:1143–1150. DOI: https://doi.org/10.1093/hmg/ddl029, PMID: 16500998

Grandinetti KB, Stevens TA, Ha S, Salamone RJ, Walker JR, Zhang J, Agarwalla S, Tenen DG, Peters EC, Reddy VA. 2011. Overexpression of TRIB2 in human lung cancers contributes to tumorigenesis through downregulation of C/EBPα. *Oncogene* **30**:3328–3335. DOI: https://doi.org/10.1038/onc.2011.57, PMID: 21399661

Guérillon C, Larrieu D, Pedeux R. 2013. ING*1* and *ING2*: multifaceted tumor suppressor genes. *Cellular and Molecular Life Sciences* **70**:3753–3772. DOI: https://doi.org/10.1007/s00018-013-1270-z, PMID: 23412501

Haller F, Bieg M, Will R, Körner C, Weichenhan D, Bott A, Ishaque N, Lutsik P, Moskalev EA, Mueller SK, Bähr M, Woerner A, Kaiser B, Scherl C, Haderlein M, Kleinheinz K, Fietkau R, Iro H, Eils R, Hartmann A, et al. 2019. Enhancer hijacking activates oncogenic transcription factor NR4A3 in acinic cell carcinomas of the salivary glands. *Nature Communications* **10**:368. DOI: https://doi.org/10.1038/s41467-018-08069-x, PMID: 30664630

Hamanaka N, Nakanishi Y, Mizuno T, Horiguchi-Takei K, Akiyama N, Tanimura H, Hasegawa M, Satoh Y, Tachibana Y, Fujii T, Sakata K, Ogasawara K, Ebiike H, Koyano H, Sato H, Ishii N, Mio T. 2019. *YES1* is a targetable oncogene in cancers harboring *YES1* Gene Amplification. *Cancer Research* **79**:5734–5745. DOI: https://doi.org/10.1158/0008-5472.CAN-18-3376, PMID: 31391186

Hanahan D, Weinberg RA. 2011. Hallmarks of Cancer: the next generation. *Cell* **144**:646–674. DOI: https://doi.org/10.1016/j.cell.2011.02.013, PMID: 21376230

Hannon MM, Lohan F, Erbilgin Y, Sayitoglu M, O'Hagan K, Mills K, Ozbek U, Keeshan K. 2012. Elevated TRIB2 with NOTCH1 activation in paediatric/adult T-ALL. *British Journal of Haematology* **158**:626–634. DOI: https://doi.org/10.1111/j.1365-2141.2012.09222.x, PMID: 22775572

Hao JJ, Lin DC, Dinh HQ, Mayakonda A, Jiang YY, Chang C, Jiang Y, Lu CC, Shi ZZ, Xu X, Zhang Y, Cai Y, Wang JW, Zhan QM, Wei WQ, Berman BP, Wang MR, Koeffler HP. 2016. Spatial intratumoral heterogeneity and temporal clonal evolution in esophageal squamous cell carcinoma. *Nature Genetics* **48**:1500–1507. DOI: https://doi.org/10.1038/ng.3683, PMID: 27749841

Hassan HM, Dave BJ, Singh RK. 2014a. TP73, an under-appreciated player in non-Hodgkin lymphoma pathogenesis and management. *Current Molecular Medicine* **14**:432–439. DOI: https://doi.org/10.2174/1566524014666140414204458, PMID: 24730526

Hassan HM, Varney ML, Jain S, Weisenburger DD, Singh RK, Dave BJ. 2014b. Disruption of chromosomal locus 1p36 differentially modulates TAp73 and δnp73 expression in follicular lymphoma. *Leukemia & Lymphoma* **55**:2924–2931. DOI: https://doi.org/10.3109/10428194.2014.900759, PMID: 24660851

Hata T, Furukawa T, Sunamura M, Egawa S, Motoi F, Ohmura N, Marumoto T, Saya H, Horii A. 2005. RNA interference targeting Aurora kinase a suppresses tumor growth and enhances the taxane chemosensitivity in human pancreatic Cancer cells. *Cancer Research* **65**:2899–2905. DOI: https://doi.org/10.1158/0008-5472.CAN-04-3981, PMID: 15805292

Hazawa M, Lin DC, Handral H, Xu L, Chen Y, Jiang YY, Mayakonda A, Ding LW, Meng X, Sharma A, Samuel S, Movahednia MM, Wong RW, Yang H, Tong C, Koeffler HP. 2017. ZNF750 is a lineage-specific tumour suppressor in squamous cell carcinoma. *Oncogene* **36**:2243–2254. DOI: https://doi.org/10.1038/onc.2016.377, PMID: 27819679

Heidenreich B, Rachakonda PS, Hemminki K, Kumar R. 2014. TERT promoter mutations in Cancer development. *Current Opinion in Genetics & Development* **24**:30–37. DOI: https://doi.org/10.1016/j.gde.2013.11.005, PMID: 24657534

Heredia-Genestar JM, Marquès-Bonet T, Juan D, Navarro A. 2020. Extreme differences between human germline and tumor mutation densities are driven by ancestral human-specific deviations. *Nature Communications* **11**:2512. DOI: https://doi.org/10.1038/s41467-020-16296-4, PMID: 32427823

Herreño AM, Ramírez AC, Chaparro VP, Fernandez MJ, Cañas A, Morantes CF, Moreno OM, Brugés RE, Mejía JA, Bustos FJ, Montecino M, Rojas AP. 2019. Role of *RUNX2* transcription factor in epithelial mesenchymal transition in non-small cell lung Cancer lung Cancer: epigenetic control of the *RUNX2* P1 promoter. *Tumor Biology* **41**:1010428319851014. DOI: https://doi.org/10.1177/1010428319851014, PMID: 31109257

Herrero MJ, Gitton Y. 2018. The untold stories of the speech gene, the *FOXP2* Cancer gene. *Genes & Cancer* **9**:11–38. DOI: https://doi.org/10.18632/genesandcancer.169, PMID: 29725501

Hill R, Madureira PA, Ferreira B, Baptista I, Machado S, Colaço L, Dos Santos M, Liu N, Dopazo A, Ugurel S, Adrienn A, Kiss-Toth E, Isbilen M, Gure AO, Link W. 2017. TRIB2 confers resistance to anti-cancer therapy by activating the serine/threonine protein kinase AKT. *Nature Communications* **8**:14687. DOI: https://doi.org/10.1038/ncomms14687, PMID: 28276427

Hodson DJ, Janas ML, Galloway A, Bell SE, Andrews S, Li CM, Pannell R, Siebel CW, MacDonald HR, De Keersmaecker K, Ferrando AA, Grutz G, Turner M. 2010. Deletion of the RNA-binding proteins ZFP36L1 and

ZFP36L2 leads to perturbed thymic development and T lymphoblastic leukemia. *Nature Immunology* **11**:717–724. DOI: https://doi.org/10.1038/ni.1901, PMID: 20622884

Hong CS, Graham NA, Gu W, Espindola Camacho C, Mah V, Maresh EL, Alavi M, Bagryanova L, Krotee PAL, Gardner BK, Behbahan IS, Horvath S, Chia D, Mellinghoff IK, Hurvitz SA, Dubinett SM, Critchlow SE, Kurdistani SK, Goodglick L, Braas D, et al. 2016. MCT1 modulates Cancer cell pyruvate export and growth of tumors that Co-express MCT1 and MCT4. *Cell Reports* **14**:1590–1601. DOI: https://doi.org/10.1016/j.celrep.2016.01.057, PMID: 26876179

Hou Z, Guo K, Sun X, Hu F, Chen Q, Luo X, Wang G, Hu J, Sun L. 2018. TRIB2 functions as novel oncogene in colorectal Cancer by blocking cellular senescence through AP4/p21 signaling. *Molecular Cancer* **17**:172. DOI: https://doi.org/10.1186/s12943-018-0922-x, PMID: 30541550

Hsu PP, Sabatini DM. 2008. Cancer cell metabolism: warburg and beyond. *Cell* **134**:703–707. DOI: https://doi.org/10.1016/j.cell.2008.08.021, PMID: 18775299

Hu J, Meng Y, Yu T, Hu L, Mao M. 2015. Ubiquitin E3 ligase MARCH7 promotes ovarian tumor growth. *Oncotarget* **6**:12174–12187. DOI: https://doi.org/10.18632/oncotarget.3650, PMID: 25895127

Hu J, Wang L, Chen J, Gao H, Zhao W, Huang Y, Jiang T, Zhou J, Chen Y. 2018a. The circular RNA circ-ITCH suppresses ovarian carcinoma progression through targeting miR-145/RASA1 signaling. *Biochemical and Biophysical Research Communications* **505**:222–228. DOI: https://doi.org/10.1016/j.bbrc.2018.09.060, PMID: 30243714

Hu J, Meng Y, Zeng J, Zeng B, Jiang X. 2018b. Ubiquitin E3 ligase MARCH7 promotes proliferation and invasion of cervical Cancer cells through VAV2-RAC1-CDC42 pathway. *Oncology Letters* **16**:2312–2318. DOI: https://doi.org/10.3892/ol.2018.8908, PMID: 30008934

Hurst LD, Batada NN. 2017. Depletion of somatic mutations in splicing-associated sequences in Cancer genomes. *Genome Biology* **18**:213. DOI: https://doi.org/10.1186/s13059-017-1337-5, PMID: 29115978

Ji Q, Cai G, Liu X, Zhang Y, Wang Y, Zhou L, Sui H, Li Q. 2019. MALAT1 regulates the transcriptional and translational levels of proto-oncogene RUNX2 in colorectal Cancer metastasis. *Cell Death & Disease* **10**:378. DOI: https://doi.org/10.1038/s41419-019-1598-x, PMID: 31097689

Jiang N, Ke B, Hjort-Jensen K, Iglesias-Gato D, Wang Z, Chang P, Zhao Y, Niu X, Wu T, Peng B, Jiang M, Li X, Shang Z, Wang Q, Chang C, Flores-Morales A, Niu Y. 2017. YAP1 regulates prostate Cancer stem cell-like characteristics to promote castration resistant growth. *Oncotarget* **8**:115054–115067. DOI: https://doi.org/10.18632/oncotarget.23014, PMID: 29383141

Jiao X, Nawab O, Patel T, Kossenkov AV, Halama N, Jaeger D, Pestell RG. 2019. Recent advances targeting CCR5 for Cancer and its role in Immuno-Oncology. *Cancer Research* **79**:4801–4807. DOI: https://doi.org/10.1158/0008-5472.CAN-19-1167, PMID: 31292161

Jo YS, Kim MS, Yoo NJ, Lee SH. 2016. Somatic mutation of a candidate tumour suppressor *MGA* gene and its mutational heterogeneity in colorectal cancers. *Pathology* **48**:525–527. DOI: https://doi.org/10.1016/j.pathol.2016.04.010, PMID: 27306572

Johnson SA, Dubeau L, Kawalek M, Dervan A, Schönthal AH, Dang CV, Johnson DL. 2003a. Increased expression of TATA-binding protein, the central transcription factor, can contribute to oncogenesis. *Molecular and Cellular Biology* **23**:3043–3051. DOI: https://doi.org/10.1128/MCB.23.9.3043-3051.2003, PMID: 12697807

Johnson SA, Dubeau L, White RJ, Johnson DL. 2003b. The TATA-binding protein as a regulator of cellular transformation. *Cell Cycle* **2**:440–442. DOI: https://doi.org/10.4161/cc.2.5.493, PMID: 12963838

Johnson SAS, Lin JJ, Walkey CJ, Leathers MP, Coarfa C, Johnson DL. 2017. Elevated TATA-binding protein expression drives vascular endothelial growth factor expression in Colon cancer. *Oncotarget* **8**:48832–48845. DOI: https://doi.org/10.18632/oncotarget.16384, PMID: 28415573

Jones RG, Thompson CB. 2009. Tumor suppressors and cell metabolism: a recipe for Cancer growth. *Genes & Development* **23**:537–548. DOI: https://doi.org/10.1101/gad.1756509, PMID: 19270154

Joshi R, Tawfik A, Edeh N, McCloud V, Looney S, Lewis J, Hsu S, Ogbureke KU. 2010. Dentin sialophosphoprotein (DSPP) gene-silencing inhibits key tumorigenic activities in human oral Cancer cell line, OSC2. *PLOS ONE* **5**:e13974. DOI: https://doi.org/10.1371/journal.pone.0013974, PMID: 21103065

Joyce BT, Zheng Y, Zhang Z, Liu L, Kocherginsky M, Murphy R, Achenbach CJ, Musa J, Wehbe F, Just A, Shen J, Vokonas P, Schwartz J, Baccarelli AA, Hou L. 2018. miRNA-Processing gene methylation and Cancer risk. *Cancer Epidemiology Biomarkers & Prevention* **27**:550–557. DOI: https://doi.org/10.1158/1055-9965.EPI-17-0849, PMID: 29475968

Kaistha BP, Honstein T, Müller V, Bielak S, Sauer M, Kreider R, Fassan M, Scarpa A, Schmees C, Volkmer H, Gress TM, Buchholz M. 2014. Key role of dual specificity kinase TTK in proliferation and survival of pancreatic Cancer cells. *British Journal of Cancer* **111**:1780–1787. DOI: https://doi.org/10.1038/bjc.2014.460, PMID: 25137017

Kaminker JS, Zhang Y, Watanabe C, Zhang Z. 2007. CanPredict: a computational tool for predicting cancer-associated missense mutations. *Nucleic Acids Research* **35**:W595–W598. DOI: https://doi.org/10.1093/nar/gkm405, PMID: 17537827

Kent OA, Mendell JT, Rottapel R. 2016. Transcriptional regulation of miR-31 by oncogenic KRAS mediates metastatic phenotypes by repressing RASA1. *Molecular Cancer Research* **14**:267–277. DOI: https://doi.org/10.1158/1541-7786.MCR-15-0456, PMID: 26747707

Kim IY, Lee DH, Lee DK, Ahn HJ, Kim MM, Kim SJ, Morton RA. 2004. Loss of expression of bone morphogenetic protein receptor type II in human prostate Cancer cells. *Oncogene* **23**:7651–7659. DOI: https://doi.org/10.1038/sj.onc.1207924, PMID: 15354178

Kim HS, Oh SH, Kim JH, Sohn WJ, Kim JY, Kim DH, Choi SU, Park KM, Ryoo ZY, Park TS, Lee S. 2018. TRIB2 regulates the differentiation of MLL-TET1 transduced myeloid progenitor cells. *Journal of Molecular Medicine* **96**:1267–1277. DOI: https://doi.org/10.1007/s00109-018-1700-3, PMID: 30324339

Kiss A, Koppel AC, Anders J, Cataisson C, Yuspa SH, Blumenberg M, Efimova T. 2016. Keratinocyte p38δ loss inhibits Ras-induced tumor formation, while systemic p38δ loss enhances skin inflammation in the early phase of chemical carcinogenesis in mouse skin. *Molecular Carcinogenesis* **55**:563–574. DOI: https://doi.org/10.1002/mc.22303, PMID: 25753147

Kiss A, Koppel A, Murphy E, Sall M, Barlas M, Kissling G, Efimova T. 2019. Cell Type-Specific p38δ targeting reveals a context-, Stage-, and Sex-Dependent regulation of skin carcinogenesis. *International Journal of Molecular Sciences* **20**:1532. DOI: https://doi.org/10.3390/ijms20071532

Koch L. 2017. Cancer genomics: the driving force of Cancer evolution. *Nature Reviews. Genetics* **18**:703. DOI: https://doi.org/10.1038/nrg.2017.95, PMID: 29109522

Kodach LL, Wiercinska E, de Miranda NF, Bleuming SA, Musler AR, Peppelenbosch MP, Dekker E, van den Brink GR, van Noesel CJ, Morreau H, Hommes DW, Ten Dijke P, Offerhaus GJ, Hardwick JC. 2008. The bone morphogenetic protein pathway is inactivated in the majority of sporadic colorectal cancers. *Gastroenterology* **134**:1332–1341. DOI: https://doi.org/10.1053/j.gastro.2008.02.059, PMID: 18471510

Koenighofer M, Hung CY, McCauley JL, Dallman J, Back EJ, Mihalek I, Gripp KW, Sol-Church K, Rusconi P, Zhang Z, Shi GX, Andres DA, Bodamer OA. 2016. Mutations in *RIT1* cause Noonan syndrome - additional functional evidence and expanding the clinical phenotype. *Clinical Genetics* **89**:359–366. DOI: https://doi.org/10.1111/cge.12608, PMID: 25959749

Kroemer G, Pouyssegur J. 2008. Tumor cell metabolism: cancer's Achilles' heel. *Cancer Cell* **13**:472–482. DOI: https://doi.org/10.1016/j.ccr.2008.05.005, PMID: 18538731

Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A, Carter SL, Stewart C, Mermel CH, Roberts SA, Kiezun A, Hammerman PS, McKenna A, Drier Y, Zou L, Ramos AH, Pugh TJ, Stransky N, Helman E, Kim J, et al. 2013. Mutational heterogeneity in Cancer and the search for new cancer-associated genes. *Nature* **499**:214–218. DOI: https://doi.org/10.1038/nature12213, PMID: 23770567

Lawrence MS, Stojanov P, Mermel CH, Robinson JT, Garraway LA, Golub TR, Meyerson M, Gabriel SB, Lander ES, Getz G. 2014. Discovery and saturation analysis of Cancer genes across 21 tumour types. *Nature* **505**:495–501. DOI: https://doi.org/10.1038/nature12912, PMID: 24390350

Lee YY, Wang CT, Huang SK, Wu WJ, Huang CN, Li CC, Chan TC, Liang PI, Hsing CH, Li CF. 2016. Downregulation of *RNF128* predicts progression and poor prognosis in patients with urothelial carcinoma of the upper tract and urinary bladder. *Journal of Cancer* **7**:2187–2196. DOI: https://doi.org/10.7150/jca.16798, PMID: 27994654

Lengauer C, Kinzler KW, Vogelstein B. 1998. Genetic instabilities in human cancers. *Nature* **396**:643–649. DOI: https://doi.org/10.1038/25292, PMID: 9872311

Li XX, Zheng HT, Huang LY, Shi DB, Peng JJ, Liang L, Cai SJ. 2014. Silencing of CXCR7 gene represses growth and invasion and induces apoptosis in colorectal Cancer through ERK and β-arrestin pathways. *International Journal of Oncology* **45**:1649–5167. DOI: https://doi.org/10.3892/ijo.2014.2547, PMID: 25051350

Li D, Yang Y, Zhu G, Liu X, Zhao M, Li X, Yang Q. 2015. MicroRNA-410 promotes cell proliferation by targeting *BRD7* in non-small cell lung Cancer. *FEBS Letters* **589**:2218–2223. DOI: https://doi.org/10.1016/j.febslet.2015.06.031, PMID: 26149213

Li XQ, Lu JT, Tan CC, Wang QS, Feng YM. 2016. RUNX2 promotes breast Cancer bone metastasis by increasing integrin α5-mediated colonization. *Cancer Letters* **380**:78–86. DOI: https://doi.org/10.1016/j.canlet.2016.06.007, PMID: 27317874

Li X, Xu W, Kang W, Wong SH, Wang M, Zhou Y, Fang X, Zhang X, Yang H, Wong CH, To KF, Chan SL, Chan MTV, Sung JJY, Wu WKK, Yu J. 2018a. Genomic analysis of liver Cancer unveils novel driver genes and distinct prognostic features. *Theranostics* **8**:1740–1751. DOI: https://doi.org/10.7150/thno.22010, PMID: 29556353

Li Z, Guo J, Ma Y, Zhang L, Lin Z. 2018b. Oncogenic role of MicroRNA-30b-5p in glioblastoma through targeting Proline-Rich transmembrane protein 2. *Oncology Research Featuring Preclinical and Clinical Cancer Therapeutics* **26**:219–230. DOI: https://doi.org/10.3727/096504017X14944585873659, PMID: 28550683

Li ZY, Zhang ZZ, Bi H, Zhang QD, Zhang SJ, Zhou L, Zhu XQ, Zhou J. 2019. Upregulated microRNA–671–3p promotes tumor progression by suppressing forkhead box P2 expression in non–small–cell lung cancer. *Molecular Medicine Reports* **20**:3149–3159. DOI: https://doi.org/10.3892/mmr.2019.10563, PMID: 31432170

Li Y, Roberts ND, Wala JA, Shapira O, Schumacher SE, Kumar K, Khurana E, Waszak S, Korbel JO, Haber JE, Imielinski M, Weischenfeldt J, Beroukhim R, Campbell PJ, PCAWG Structural Variation Working Group, PCAWG Consortium. 2020. Patterns of somatic structural variation in human Cancer genomes. *Nature* **578**:112–121. DOI: https://doi.org/10.1038/s41586-019-1913-9, PMID: 32025012

Li X, Tai HH. 2013. Activation of thromboxane A2 receptor (TP) increases the expression of monocyte chemoattractant protein -1 (MCP-1)/chemokine (C-C motif) ligand 2 (CCL2) and recruits macrophages to promote invasion of lung Cancer cells. *PLOS ONE* **8**:e54073. DOI: https://doi.org/10.1371/journal.pone.0054073, PMID: 23349788

Liang J, Chen M, Hughes D, Chumanevich AA, Altilia S, Kaza V, Lim CU, Kiaris H, Mythreye K, Pena MM, Broude EV, Roninson IB. 2018. CDK8 selectively promotes the growth of Colon cancer metastases in the liver by regulating gene expression of TIMP3 and matrix metalloproteinases. *Cancer Research* **78**:6594–6606. DOI: https://doi.org/10.1158/0008-5472.CAN-18-1583, PMID: 30185549

Lim SY, Yuzhalin AE, Gordon-Weeks AN, Muschel RJ. 2016. Targeting the CCL2-CCR2 signaling Axis in Cancer metastasis. *Oncotarget* **7**:28697–28710. DOI: https://doi.org/10.18632/oncotarget.7376, PMID: 26885690

Lin DC, Dinh HQ, Xie JJ, Mayakonda A, Silva TC, Jiang YY, Ding LW, He JZ, Xu XE, Hao JJ, Wang MR, Li C, Xu LY, Li EM, Berman BP, Phillip Koeffler H. 2018. Identification of distinct mutational patterns and new driver genes in oesophageal squamous cell carcinomas and adenocarcinomas. *Gut* **67**:1769–1779. DOI: https://doi.org/10.1136/gutjnl-2017-314607, PMID: 28860350

Lin C, Xu X. 2017. YAP1-TEAD1-Glut1 Axis dictates the oncogenic phenotypes of breast Cancer cells by modulating glycolysis. *Biomedicine & Pharmacotherapy* **95**:789–794. DOI: https://doi.org/10.1016/j.biopha.2017.08.091, PMID: 28892790

Liu Y, Liu T, Sun Q, Niu M, Jiang Y, Pang D. 2015. Downregulation of ras gtpaseactivating protein 1 is associated with poor survival of breast invasive ductal carcinoma patients. *Oncology Reports* **33**:119–124. DOI: https://doi.org/10.3892/or.2014.3604, PMID: 25394563

Liu L, Hu J, Yu T, You S, Zhang Y, Hu L. 2019. miR-27b-3p/MARCH7 regulates invasion and metastasis of endometrial Cancer cells through Snail-mediated pathway. *Acta Biochimica Et Biophysica Sinica* **51**:492–500. DOI: https://doi.org/10.1093/abbs/gmz030, PMID: 31006800

López S, Lim EL, Horswell S, Haase K, Huebner A, Dietzen M, Mourikis TP, Watkins TBK, Rowan A, Dewhurst SM, Birkbak NJ, Wilson GA, Van Loo P, Jamal-Hanjani M, Swanton C, McGranahan N, TRACERx Consortium. 2020. Interplay between whole-genome doubling and the accumulation of deleterious alterations in Cancer evolution. *Nature Genetics* **52**:283–293. DOI: https://doi.org/10.1038/s41588-020-0584-7, PMID: 32139907

Lorenzetto E, Brenca M, Boeri M, Verri C, Piccinin E, Gasparini P, Facchinetti F, Rossi S, Salvatore G, Massimino M, Sozzi G, Maestro R, Modena P. 2014. *YAP1* acts as oncogenic target of 11q22 amplification in multiple Cancer subtypes. *Oncotarget* **5**:2608–2621. DOI: https://doi.org/10.18632/oncotarget.1844, PMID: 24810989

Lu H, Jiang T, Ren K, Li ZL, Ren J, Wu G, Han X. 2018. RUNX2 plays an oncogenic role in esophageal carcinoma by activating the PI3K/AKT and ERK signaling pathways. *Cellular Physiology and Biochemistry* **49**:217–225. DOI: https://doi.org/10.1159/000492872, PMID: 30138923

Lucena-Araujo AR, Kim HT, Thomé C, Jacomo RH, Melo RA, Bittencourt R, Pasquini R, Pagnano K, Glória AB, Chauffaille ML, Athayde M, Chiattone CS, Mito I, Bendlin R, Souza C, Bortolheiro C, Coelho-Silva JL, Schrier SL, Tallman MS, Grimwade D, et al. 2015. High δnp73/TAp73 ratio is associated with poor prognosis in acute promyelocytic leukemia. *Blood* **126**:2302–2306. DOI: https://doi.org/10.1182/blood-2015-01-623330, PMID: 26429976

Luo ZG, Tang H, Li B, Zhu Z, Ni CR, Zhu MH. 2011. Genetic alterations of tumor suppressor *ING1* in human non-small cell lung Cancer. *Oncology Reports* **25**:1073–1081. DOI: https://doi.org/10.3892/or.2011.1172, PMID: 21286670

Luo Y, Hao T, Zhang J, Zhang M, Sun P, Wu L. 2019. MicroRNA-592 suppresses the malignant phenotypes of thyroid Cancer by regulating lncRNA NEAT1 and downregulating NOVA1. *International Journal of Molecular Medicine* **44**:1172–1182. DOI: https://doi.org/10.3892/ijmm.2019.4278, PMID: 31524231

Maas AM, Bretz AC, Mack E, Stiewe T. 2013. Targeting p73 in cancer. *Cancer Letters* **332**:229–236. DOI: https://doi.org/10.1016/j.canlet.2011.07.030, PMID: 21903324

Martincorena I, Raine KM, Gerstung M, Dawson KJ, Haase K, Van Loo P, Davies H, Stratton MR, Campbell PJ. 2017. Universal patterns of selection in Cancer and somatic tissues. *Cell* **171**:1029–1041. DOI: https://doi.org/10.1016/j.cell.2017.09.042, PMID: 29056346

Mathupala SP, Colen CB, Parajuli P, Sloan AE. 2007. Lactate and malignant tumors: a therapeutic target at the end stage of glycolysis. *Journal of Bioenergetics and Biomembranes* **39**:73–77. DOI: https://doi.org/10.1007/s10863-006-9062-x, PMID: 17354062

Matsui Y, Amano H, Ito Y, Eshima K, Suzuki T, Ogawa F, Iyoda A, Satoh Y, Kato S, Nakamura M, Kitasato H, Narumiya S, Majima M. 2012. Thromboxane A_2 receptor signaling facilitates tumor colonization through P-selectin-mediated interaction of tumor cells with platelets and endothelial cells. *Cancer Science* **103**:700–707. DOI: https://doi.org/10.1111/j.1349-7006.2012.02200.x, PMID: 22296266

Maura F, Bolli N, Angelopoulos N, Dawson KJ, Leongamornlert D, Martincorena I, Mitchell TJ, Fullam A, Gonzalez S, Szalat R, Abascal F, Rodriguez-Martin B, Samur MK, Glodzik D, Roncador M, Fulciniti M, Tai YT, Minvielle S, Magrangeas F, Moreau P, et al. 2019. Genomic landscape and chronological reconstruction of driver events in multiple myeloma. *Nature Communications* **10**:3835. DOI: https://doi.org/10.1038/s41467-019-11680-1, PMID: 31444325

McBrayer SK, Cheng JC, Singhal S, Krett NL, Rosen ST, Shanmugam M. 2012. Multiple myeloma exhibits novel dependence on GLUT4, GLUT8, and GLUT11: implications for glucose transporter-directed therapy. *Blood* **119**:4686–4697. DOI: https://doi.org/10.1182/blood-2011-09-377846, PMID: 22452979

McKnight DA, Suzanne Hart P, Hart TC, Hartsfield JK, Wilson A, Wright JT, Fisher LW. 2008. A comprehensive analysis of normal variation and disease-causing mutations in the human *DSPP* gene. *Human Mutation* **29**:1392–1404. DOI: https://doi.org/10.1002/humu.20783, PMID: 18521831

Melo RCC, Ferro KPV, Duarte A, Olalla Saad ST. 2018. CXCR7 participates in CXCL12-mediated migration and homing of leukemic and normal hematopoietic cells. *Stem Cell Research & Therapy* **9**:34. DOI: https://doi.org/10.1186/s13287-017-0765-1, PMID: 29433559

Meng FJ, Meng FM, Wu HX, Cao XF. 2017. miR-564 inhibited metastasis and proliferation of prostate Cancer by targeting MLLT3. *European Review for Medical and Pharmacological Sciences* **21**:4828–4834. PMID: 29164580

Meyerson M, Gabriel S, Getz G. 2010. Advances in understanding Cancer genomes through second-generation sequencing. *Nature Reviews Genetics* **11**:685–696. DOI: https://doi.org/10.1038/nrg2841, PMID: 20847746

Miao R, Wu Y, Zhang H, Zhou H, Sun X, Csizmadia E, He L, Zhao Y, Jiang C, Miksad RA, Ghaziani T, Robson SC, Zhao H. 2016. Utility of the dual-specificity protein kinase TTK as a therapeutic target for intrahepatic spread of liver Cancer. *Scientific Reports* **6**:33121. DOI: https://doi.org/10.1038/srep33121, PMID: 27618777

Michaelson JJ, Shi Y, Gujral M, Zheng H, Malhotra D, Jin X, Jian M, Liu G, Greer D, Bhandari A, Wu W, Corominas R, Peoples A, Koren A, Gore A, Kang S, Lin GN, Estabillo J, Gadomski T, Singh B, et al. 2012. Whole-genome sequencing in autism identifies hot spots for de novo germline mutation. *Cell* **151**:1431–1442. DOI: https://doi.org/10.1016/j.cell.2012.11.019, PMID: 23260136

Mikheev AM, Mikheeva SA, Severs LJ, Funk CC, Huang L, McFaline-Figueroa JL, Schwensen J, Trapnell C, Price ND, Wong S, Rostomily RC. 2018. Targeting TWIST1 through loss of function inhibits tumorigenicity of human glioblastoma. *Molecular Oncology* **12**:1188–1202. DOI: https://doi.org/10.1002/1878-0261.12320, PMID: 29754406

Mofers A, Pellegrini P, Linder S, D'Arcy P. 2017. Proteasome-associated deubiquitinases and cancer. *Cancer and Metastasis Reviews* **36**:635–653. DOI: https://doi.org/10.1007/s10555-017-9697-6, PMID: 29134486

Nambara S, Masuda T, Tobo T, Kidogami S, Komatsu H, Sugimachi K, Saeki H, Oki E, Maehara Y, Mimori K. 2017. Clinical significance of *ZNF750* gene expression, a novel tumor suppressor gene, in esophageal squamous cell carcinoma. *Oncology Letters* **14**:1795–1801. DOI: https://doi.org/10.3892/ol.2017.6341, PMID: 28789412

Nikitakis NG, Gkouveris I, Aseervatham J, Barahona K, Ogbureke KUE. 2018. DSPP-MMP20 gene silencing downregulates Cancer stem cell markers in human oral Cancer cells. *Cellular & Molecular Biology Letters* **23**:30. DOI: https://doi.org/10.1186/s11658-018-0096-y, PMID: 30002682

Niu W, Luo Y, Wang X, Zhou Y, Li H, Wang H, Fu Y, Liu S, Yin S, Li J, Zhao R, Liu Y, Fan S, Li Z, Xiong W, Li X, Li G, Ren C, Tan M, Zhou M. 2018. BRD7 inhibits the warburg effect and tumor progression through inactivation of HIF1α/LDHA Axis in breast Cancer. *Cell Death & Disease* **9**:519. DOI: https://doi.org/10.1038/s41419-018-0536-7, PMID: 29725006

Noguchi C, Kamitori K, Hossain A, Hoshikawa H, Katagi A, Dong Y, Sui L, Tokuda M, Yamaguchi F. 2016. D-Allose inhibits Cancer cell growth by reducing GLUT1 expression. *The Tohoku Journal of Experimental Medicine* **238**:131–141. DOI: https://doi.org/10.1620/tjem.238.131, PMID: 26829886

Nussinov R, Jang H, Tsai CJ, Cheng F. 2019. Review: precision medicine and driver mutations: computational methods, functional assays and conformational principles for interpreting Cancer drivers. *PLOS Computational Biology* **15**:e1006658. DOI: https://doi.org/10.1371/journal.pcbi.1006658, PMID: 30921324

O'Callaghan C, Fanning LJ, Houston A, Barry OP. 2013. Loss of p38δ mitogen-activated protein kinase expression promotes oesophageal squamous cell carcinoma proliferation, migration and anchorage-independent growth. *International Journal of Oncology* **43**:405–415. DOI: https://doi.org/10.3892/ijo.2013.1968, PMID: 23722928

Orr K, Buckley NE, Haddock P, James C, Parent JL, McQuaid S, Mullan PB. 2016. Thromboxane A2 receptor (TBXA2R) is a potent survival factor for triple negative breast cancers (TNBCs). *Oncotarget* **7**:55458–55472. DOI: https://doi.org/10.18632/oncotarget.10969, PMID: 27487152

Otsuka R, Akutsu Y, Sakata H, Hanari N, Murakami K, Kano M, Toyozumi T, Takahashi M, Matsumoto Y, Sekino N, Yokoyama M, Okada K, Shiraishi T, Komatsu A, Iida K, Matsubara H. 2018. ZNF750 expression is a potential prognostic biomarker in esophageal squamous cell carcinoma. *Oncology* **94**:142–148. DOI: https://doi.org/10.1159/000484932, PMID: 29216641

Owens P, Pickup MW, Novitskiy SV, Chytil A, Gorska AE, Aakre ME, West J, Moses HL. 2012. Disruption of bone morphogenetic protein receptor 2 (BMPR2) in mammary tumors promotes metastases through cell autonomous and paracrine mediators. *PNAS* **109**:2814–2819. DOI: https://doi.org/10.1073/pnas.1101139108, PMID: 21576484

Pancrazi L, Di Benedetto G, Colombaioni L, Della Sala G, Testa G, Olimpico F, Reyes A, Zeviani M, Pozzan T, Costa M. 2015. Foxg1 localizes to mitochondria and coordinates cell differentiation and bioenergetics. *PNAS* **112**:13910–13915. DOI: https://doi.org/10.1073/pnas.1515190112, PMID: 26508630

Parajuli P, Singh P, Wang Z, Li L, Eragamreddi S, Ozkan S, Ferrigno O, Prunier C, Razzaque MS, Xu K, Atfi A. 2019. TGIF1 functions as a tumor suppressor in pancreatic ductal adenocarcinoma. *The EMBO Journal* **38**:e101067. DOI: https://doi.org/10.15252/embj.2018101067, PMID: 31268604

Park SW, Hur SY, Yoo NJ, Lee SH. 2010. Somatic frameshift mutations of bone morphogenic protein receptor 2 gene in gastric and colorectal cancers with microsatellite instability. *Apmis* **118**:824–829. DOI: https://doi.org/10.1111/j.1600-0463.2010.02670.x, PMID: 20955454

Parmigiani G, Boca S, Lin J, Kinzler KW, Velculescu V, Vogelstein B. 2009. Design and analysis issues in genome-wide somatic mutation studies of Cancer. *Genomics* **93**:17–21. DOI: https://doi.org/10.1016/j.ygeno.2008.07.005, PMID: 18692126

Pasmant E, Gilbert-Dussardier B, Petit A, de Laval B, Luscan A, Gruber A, Lapillonne H, Deswarte C, Goussard P, Laurendeau I, Uzan B, Pflumio F, Brizard F, Vabres P, Naguibvena I, Fasola S, Millot F, Porteu F, Vidaud D, Landman-Parker J, et al. 2015. SPRED1, a RAS MAPK pathway inhibitor that causes legius syndrome, is a tumour suppressor downregulated in paediatric acute myeloblastic leukaemia. *Oncogene* **34**:631–638. DOI: https://doi.org/10.1038/onc.2013.587, PMID: 24469042

Patthy L. 1999. 2nd ed. *Protein Evolution*. Blackwell Publishing Ltd.

Peña PV, Hom RA, Hung T, Lin H, Kuo AJ, Wong RP, Subach OM, Champagne KS, Zhao R, Verkhusha VV, Li G, Gozani O, Kutateladze TG. 2008. Histone H3K4me3 binding is required for the DNA repair and apoptotic activities of ING1 tumor suppressor. *Journal of Molecular Biology* **380**:303–312. DOI: https://doi.org/10.1016/j.jmb.2008.04.061, PMID: 18533182

Petrenko O, Zaika A, Moll UM. 2003. deltaNp73 facilitates cell immortalization and cooperates with oncogenic ras in cellular transformation in vivo. *Molecular and Cellular Biology* **23**:5540–5555. DOI: https://doi.org/10.1128/MCB.23.16.5540-5555.2003, PMID: 12897129

Pfeifer G. 2018. Defining driver DNA methylation changes in human Cancer. *International Journal of Molecular Sciences* **19**:1166. DOI: https://doi.org/10.3390/ijms19041166, PMID: 29649096

Philip S, Kumarasiri M, Teo T, Yu M, Wang S. 2018. Cyclin-Dependent kinase 8: a new hope in targeted Cancer therapy? *Journal of Medicinal Chemistry* **61**:5073–5092. DOI: https://doi.org/10.1021/acs.jmedchem.7b00901, PMID: 29266937

Pickup MW, Hover LD, Polikowsky ER, Chytil A, Gorska AE, Novitskiy SV, Moses HL, Owens P. 2015. *BMPR2* loss in fibroblasts promotes mammary carcinoma metastasis via increased inflammation. *Molecular Oncology* **9**:179–191. DOI: https://doi.org/10.1016/j.molonc.2014.08.004, PMID: 25205038

Polak P, Karlić R, Koren A, Thurman R, Sandstrom R, Lawrence M, Reynolds A, Rynes E, Vlahoviček K, Stamatoyannopoulos JA, Sunyaev SR. 2015. Cell-of-origin chromatin organization shapes the mutational landscape of Cancer. *Nature* **518**:360–364. DOI: https://doi.org/10.1038/nature14221, PMID: 25693567

Poulain L, Sujobert P, Zylbersztejn F, Barreau S, Stuani L, Lambert M, Palama TL, Chesnais V, Birsen R, Vergez F, Farge T, Chenevier-Gobeaux C, Fraisse M, Bouillaud F, Debeissat C, Herault O, Récher C, Lacombe C, Fontenay M, Mayeux P, et al. 2017. High mTORC1 activity drives glycolysis addiction and sensitivity to G6PD inhibition in acute myeloid leukemia cells. *Leukemia* **31**:2326–2335. DOI: https://doi.org/10.1038/leu.2017.81, PMID: 28280275

Poulos RC, Wong YT, Ryan R, Pang H, Wong JWH. 2018. Analysis of 7,815 Cancer exomes reveals associations between mutational processes and somatic driver mutations. *PLOS Genetics* **14**:e1007779. DOI: https://doi.org/10.1371/journal.pgen.1007779, PMID: 30412573

Pu H, Zhang Q, Zhao C, Shi L, Wang Y, Wang J, Zhang M. 2015. Overexpression of G6PD is associated with high risks of recurrent metastasis and poor progression-free survival in primary breast carcinoma. *World Journal of Surgical Oncology* **13**:323. DOI: https://doi.org/10.1186/s12957-015-0733-0, PMID: 26607846

Puddinu V, Casella S, Radice E, Thelen S, Dirnhofer S, Bertoni F, Thelen M. 2017. ACKR3 expression on diffuse large B cell lymphoma is required for tumor spreading and tissue infiltration. *Oncotarget* **8**:85068–85084. DOI: https://doi.org/10.18632/oncotarget.18844, PMID: 29156704

Pulley JM, Jerome RN, Ogletree ML, Bernard GR, Lavieri RR, Zaleski NM, Hong CC, Shirey-Rice JK, Arteaga CL, Mayer IA, Holroyd KJ, Cook RS. 2018. Motivation for launching a Cancer metastasis inhibition (CMI) Program. *Targeted Oncology* **13**:61–68. DOI: https://doi.org/10.1007/s11523-017-0542-1, PMID: 29218624

Pyatnitskiy M, Karpov D, Poverennaya E, Lisitsa A, Moshkovskii S. 2015. Bringing down Cancer aircraft: searching for essential hypomutated proteins in skin melanoma. *PLOS ONE* **10**:e0142819. DOI: https://doi.org/10.1371/journal.pone.0142819, PMID: 26565620

Qian T, Liu Y, Dong Y, Zhang L, Dong Y, Sun Y, Sun D. 2018. CXCR7 regulates breast tumor metastasis and angiogenesis in vivo and in vitro. *Molecular Medicine Reports* **17**:3633–3639. DOI: https://doi.org/10.3892/mmr.2017.8286, PMID: 29257351

Rahbari R, Wuster A, Lindsay SJ, Hardwick RJ, Alexandrov LB, Turki SA, Dominiczak A, Morris A, Porteous D, Smith B, Stratton MR, Hurles ME, UK10K Consortium. 2016. Timing, rates and spectra of human germline mutation. *Nature Genetics* **48**:126–133. DOI: https://doi.org/10.1038/ng.3469, PMID: 26656846

Rai K. 2019. Personalized Cancer therapy: yes1 is the new kid on the block. *Cancer Research* **79**:5702–5703. DOI: https://doi.org/10.1158/0008-5472.CAN-19-2995, PMID: 31772072

Razzaque MS, Atfi A. 2016. TGIF function in Oncogenic wnt signaling. *Biochimica Et Biophysica Acta (BBA) - Reviews on Cancer* **1865**:101–104. DOI: https://doi.org/10.1016/j.bbcan.2015.10.003, PMID: 26522669

Rheinbay E, Nielsen MM, Abascal F, Wala JA, Shapira O, Tiao G, Hornshøj H, Hess JM, Juul RI, Lin Z, Feuerbach L, Sabarinathan R, Madsen T, Kim J, Mularoni L, Shuai S, Lanzós A, Herrmann C, Maruvka YE, Shen C, et al. 2020. Analyses of non-coding somatic drivers in 2,658 cancer whole genomes. *Nature* **578**:102–111. DOI: https://doi.org/10.1038/s41586-020-1965-x, PMID: 32025015

Rishi L, Hannon M, Salomè M, Hasemann M, Frank AK, Campos J, Timoney J, O'Connor C, Cahill MR, Porse B, Keeshan K. 2014. Regulation of Trib2 by an E2F1-C/EBPα feedback loop in AML cell proliferation. *Blood* **123**:2389–2400. DOI: https://doi.org/10.1182/blood-2013-07-511683, PMID: 24516045

Robertson E, Perry C, Doherty R, Madhusudan S. 2015. Transcriptomic profiling of forkhead box transcription factors in adult glioblastoma multiforme. *Cancer Genomics & Proteomics* **12**:103–112. PMID: 25977169

Rogozin IB, Pavlov YI. 2003. Theoretical analysis of mutation hotspots and their DNA sequence context specificity. *Mutation Research/Reviews in Mutation Research* **544**:65–85. DOI: https://doi.org/10.1016/S1383-5742(03)00032-2, PMID: 12888108

Romero OA, Torres-Diz M, Pros E, Savola S, Gomez A, Moran S, Saez C, Iwakawa R, Villanueva A, Montuenga LM, Kohno T, Yokota J, Sanchez-Cespedes M. 2014. *MAX* inactivation in small cell lung Cancer disrupts MYC-SWI/SNF programs and is synthetic lethal with BRG1. *Cancer Discovery* **4**:292–303. DOI: https://doi.org/10.1158/2159-8290.CD-13-0799, PMID: 24362264

Rosenbluh J, Nijhawan D, Cox AG, Li X, Neal JT, Schafer EJ, Zack TI, Wang X, Tsherniak A, Schinzel AC, Shao DD, Schumacher SE, Weir BA, Vazquez F, Cowley GS, Root DE, Mesirov JP, Beroukhim R, Kuo CJ, Goessling W, et al. 2012. β-Catenin-driven cancers require a YAP1 transcriptional complex for survival and tumorigenesis. *Cell* **151**:1457–1473. DOI: https://doi.org/10.1016/j.cell.2012.11.026, PMID: 23245941

Roussel MF, Stripay JL. 2018. Epigenetic drivers in pediatric medulloblastoma. *The Cerebellum* **17**:28–36. DOI: https://doi.org/10.1007/s12311-017-0899-9, PMID: 29178021

Salvadores M, Mas-Ponte D, Supek F. 2019. Passenger mutations accurately classify human tumors. *PLOS Computational Biology* **15**:e1006953. DOI: https://doi.org/10.1371/journal.pcbi.1006953, PMID: 30986244

Sancisi V, Manzotti G, Gugnoni M, Rossi T, Gandolfi G, Gobbi G, Torricelli F, Catellani F, Faria do Valle I, Remondini D, Castellani G, Ragazzi M, Piana S, Ciarrocchi A. 2017. RUNX2 expression in thyroid and breast

Cancer requires the cooperation of three non-redundant enhancers under the control of BRD4 and c-JUN. *Nucleic Acids Research* **45**:11249–11267. DOI: https://doi.org/10.1093/nar/gkx802, PMID: 28981843

Saxena G, Koli K, de la Garza J, Ogbureke KU. 2015. Matrix metalloproteinase 20-dentin sialophosphoprotein interaction in oral Cancer. *Journal of Dental Research* **94**:584–593. DOI: https://doi.org/10.1177/0022034515570156, PMID: 25666817

Sayed ME, Yuan L, Robin JD, Tedone E, Batten K, Dahlson N, Wright WE, Shay JW, Ludlow AT. 2019. NOVA1 directs PTBP1 to hTERT pre-mRNA and promotes telomerase activity in Cancer cells. *Oncogene* **38**:2937–2952. DOI: https://doi.org/10.1038/s41388-018-0639-8, PMID: 30568224

Schindler EM, Hindes A, Gribben EL, Burns CJ, Yin Y, Lin MH, Owen RJ, Longmore GD, Kissling GE, Arthur JS, Efimova T. 2009. p38δ Mitogen-activated protein kinase is essential for skin tumor development in mice. *Cancer Research* **69**:4648–4655. DOI: https://doi.org/10.1158/0008-5472.CAN-08-4455, PMID: 19458068

Schmall A, Al-Tamari HM, Herold S, Kampschulte M, Weigert A, Wietelmann A, Vipotnik N, Grimminger F, Seeger W, Pullamsetti SS, Savai R. 2015. Macrophage and Cancer cell cross-talk via CCR2 and CX3CR1 is a fundamental mechanism driving lung Cancer. *American Journal of Respiratory and Critical Care Medicine* **191**:437–447. DOI: https://doi.org/10.1164/rccm.201406-1137OC, PMID: 25536148

Schuster-Böckler B, Lehner B. 2012. Chromatin organization is a major influence on regional mutation rates in human Cancer cells. *Nature* **488**:504–507. DOI: https://doi.org/10.1038/nature11273, PMID: 22820252

Semenza GL. 2010a. HIF-1: upstream and downstream of Cancer metabolism. *Current Opinion in Genetics & Development* **20**:51–56. DOI: https://doi.org/10.1016/j.gde.2009.10.009, PMID: 19942427

Semenza GL. 2010b. Defining the role of hypoxia-inducible factor 1 in Cancer biology and therapeutics. *Oncogene* **29**:625–634. DOI: https://doi.org/10.1038/onc.2009.441, PMID: 19946328

Serrao A, Jenkins LM, Chumanevich AA, Horst B, Liang J, Gatza ML, Lee NY, Roninson IB, Broude EV, Mythreye K. 2018. Mediator kinase CDK8/CDK19 drives YAP1-dependent BMP4-induced EMT in Cancer. *Oncogene* **37**:4792–4808. DOI: https://doi.org/10.1038/s41388-018-0316-y, PMID: 29780169

Shah A, Melhuish TA, Fox TE, Frierson HF, Wotton D. 2019. TGIF transcription factors repress acetyl CoA metabolic gene expression and promote intestinal tumor growth. *Genes & Development* **33**:388–402. DOI: https://doi.org/10.1101/gad.320127.118, PMID: 30808659

Shain AH, Yeh I, Kovalyshyn I, Sriharan A, Talevich E, Gagnon A, Dummer R, North J, Pincus L, Ruben B, Rickaby W, D'Arrigo C, Robson A, Bastian BC. 2015. The genetic evolution of melanoma from precursor lesions. *New England Journal of Medicine* **373**:1926–1936. DOI: https://doi.org/10.1056/NEJMoa1502583, PMID: 26559571

Sharma Y, Miladi M, Dukare S, Boulay K, Caudron-Herger M, Groß M, Backofen R, Diederichs S. 2019. A pan-cancer analysis of synonymous mutations. *Nature Communications* **10**:2569. DOI: https://doi.org/10.1038/s41467-019-10489-2, PMID: 31189880

Shen F, Zhang Y, Jernigan DL, Feng X, Yan J, Garcia FU, Meucci O, Salvino JM, Fatatis A. 2016. Novel Small-Molecule CX3CR1 antagonist impairs metastatic seeding and colonization of breast Cancer cells. *Molecular Cancer Research* **14**:518–527. DOI: https://doi.org/10.1158/1541-7786.MCR-16-0013, PMID: 27001765

Shen Y, Chen F, Liang Y. 2019. MicroRNA-133a inhibits the proliferation of non-small cell lung Cancer by targeting *YES1*. *Oncology Letters* **18**:6759–6765. DOI: https://doi.org/10.3892/ol.2019.11030, PMID: 31807185

Shi Y, Geng D, Zhang Y, Zhao M, Wang Y, Jiang Y, Yu R, Zhou X. 2019. LATS2 inhibits malignant behaviors of glioma cells via inactivating YAP. *Journal of Molecular Neuroscience* **68**:38–48. DOI: https://doi.org/10.1007/s12031-019-1262-z, PMID: 30771084

Shibuya K, Okada M, Suzuki S, Seino M, Seino S, Takeda H, Kitanaka C. 2015. Targeting the facilitative glucose transporter GLUT1 inhibits the self-renewal and tumor-initiating capacity of Cancer stem cells. *Oncotarget* **6**:651–661. DOI: https://doi.org/10.18632/oncotarget.2892, PMID: 25528771

Shin MH, He Y, Marrogi E, Piperdi S, Ren L, Khanna C, Gorlick R, Liu C, Huang J. 2016. A RUNX2-Mediated epigenetic regulation of the survival of p53 defective Cancer cells. *PLOS Genetics* **12**:e1005884. DOI: https://doi.org/10.1371/journal.pgen.1005884, PMID: 26925584

Sillars-Hardebol AH, Carvalho B, Tijssen M, Beliën JA, de Wit M, Delis-van Diemen PM, Pontén F, van de Wiel MA, Fijneman RJ, Meijer GA. 2012. *TPX2* and *AURKA* promote 20q amplicon-driven colorectal adenoma to carcinoma progression. *Gut* **61**:1568–1575. DOI: https://doi.org/10.1136/gutjnl-2011-301153, PMID: 22207630

Slack FJ, Chinnaiyan AM. 2019. The role of Non-coding RNAs in oncology. *Cell* **179**:1033–1055. DOI: https://doi.org/10.1016/j.cell.2019.10.017, PMID: 31730848

Sondka Z, Bamford S, Cole CG, Ward SA, Dunham I, Forbes SA. 2018. The COSMIC Cancer gene census: describing genetic dysfunction across all human cancers. *Nature Reviews Cancer* **18**:696–705. DOI: https://doi.org/10.1038/s41568-018-0060-1, PMID: 30293088

Song XL, Tang Y, Lei XH, Zhao SC, Wu ZQ. 2017. miR-618 inhibits prostate Cancer migration and invasion by targeting FOXP2. *Journal of Cancer* **8**:2501–2510. DOI: https://doi.org/10.7150/jca.17407, PMID: 28900488

Stacer AC, Fenner J, Cavnar SP, Xiao A, Zhao S, Chang SL, Salomonnson A, Luker KE, Luker GD. 2016. Endothelial CXCR7 regulates breast Cancer metastasis. *Oncogene* **35**:1716–1724. DOI: https://doi.org/10.1038/onc.2015.236, PMID: 26119946

Stiewe T, Zimmermann S, Frilling A, Esche H, Pützer BM. 2002. Transactivation-deficient DeltaTA-p73 acts as an oncogene. *Cancer Research* **62**:3598–3602. PMID: 12097259

Stiewe T, Pützer BM. 2002. Role of p73 in malignancy: tumor suppressor or oncogene? *Cell Death & Differentiation* **9**:237–245. DOI: https://doi.org/10.1038/sj.cdd.4400995, PMID: 11859406

Stothard P. 2000. The sequence manipulation suite: javascript programs for analyzing and formatting protein and DNA sequences. *BioTechniques* **28**:1102–1104. DOI: https://doi.org/10.2144/00286ir01, PMID: 10868275

Su ZL, Su CW, Huang YL, Yang WY, Sampurna BP, Ouchi T, Lee KL, Cs W, Wang HD, Yuh CH. 2019. A novel AURKA Mutant-Induced Early-Onset severe hepatocarcinogenesis greater than Wild-Type via activating different pathways in zebrafish. *Cancers* **11**:927. DOI: https://doi.org/10.3390/cancers11070927, PMID: 31269749

Suk FM, Chang CC, Lin RJ, Lin SY, Liu SC, Jau CF, Liang YC. 2018. ZFP36L1 and ZFP36L2 inhibit cell proliferation in a cyclin D-dependent and p53-independent manner. *Scientific Reports* **8**:2742. DOI: https://doi.org/10.1038/s41598-018-21160-z, PMID: 29426877

Sukeda A, Nakamura Y, Nishida Y, Kojima M, Gotohda N, Akimoto T, Ochiai A. 2019. Expression of monocarboxylate transporter 1 is associated with better prognosis and reduced nodal metastasis in pancreatic ductal adenocarcinoma. *Pancreas* **48**:1102–1110. DOI: https://doi.org/10.1097/MPA.0000000000001369, PMID: 31404019

Sun D, Wang C, Long S, Ma Y, Guo Y, Huang Z, Chen X, Zhang C, Chen J, Zhang J. 2015. C/EBP-β-activated microRNA-223 promotes tumour growth through targeting RASA1 in human colorectal Cancer. *British Journal of Cancer* **112**:1491–1500. DOI: https://doi.org/10.1038/bjc.2015.107, PMID: 25867276

Sun J, Zhang J, Wang Y, Li Y, Zhang R. 2019. A pilot study of aberrant CpG island hypermethylation of *SPRED1* in Acute Myeloloid Leukemia. *International Journal of Medical Sciences* **16**:324–330. DOI: https://doi.org/10.7150/ijms.27757, PMID: 30745814

Sung H, Kanchi KL, Wang X, Hill KS, Messina JL, Lee JH, Kim Y, Dees ND, Ding L, Teer JK, Yang S, Sarnaik AA, Sondak VK, Mulé JJ, Wilson RK, Weber JS, Kim M. 2016. Inactivation of RASA1 promotes melanoma tumorigenesis via R-Ras activation. *Oncotarget* **7**:23885–23896. DOI: https://doi.org/10.18632/oncotarget.8127, PMID: 26993606

Supek F, Miñana B, Valcárcel J, Gabaldón T, Lehner B. 2014. Synonymous mutations frequently act as driver mutations in human cancers. *Cell* **156**:1324–1335. DOI: https://doi.org/10.1016/j.cell.2014.01.051, PMID: 24630730

Takahashi Y, Sheridan P, Niida A, Sawada G, Uchi R, Mizuno H, Kurashige J, Sugimachi K, Sasaki S, Shimada Y, Hase K, Kusunoki M, Kudo S, Watanabe M, Yamada K, Sugihara K, Yamamoto H, Suzuki A, Doki Y, Miyano S, et al. 2015. The AURKA/TPX2 Axis drives Colon tumorigenesis cooperatively with MYC. *Annals of Oncology* **26**:935–942. DOI: https://doi.org/10.1093/annonc/mdv034, PMID: 25632068

Tamborero D, Gonzalez-Perez A, Perez-Llamas C, Deu-Pons J, Kandoth C, Reimand J, Lawrence MS, Getz G, Bader GD, Ding L, Lopez-Bigas N. 2013. Comprehensive identification of mutational Cancer driver genes across 12 tumor types. *Scientific Reports* **3**:2650. DOI: https://doi.org/10.1038/srep02650, PMID: 24084849

Tan W, Lim SG, Tan TM. 2015. Up-regulation of microRNA-210 inhibits proliferation of hepatocellular carcinoma cells by targeting YES1. *World Journal of Gastroenterology* **21**:13030–13041. DOI: https://doi.org/10.3748/wjg.v21.i46.13030, PMID: 26676187

Tandon M, Chen Z, Pratap J. 2014. Runx2 activates PI3K/Akt signaling via mTORC2 regulation in invasive breast Cancer cells. *Breast Cancer Research* **16**:R16. DOI: https://doi.org/10.1186/bcr3611, PMID: 24479521

Tandon M, Chen Z, Othman AH, Pratap J. 2016. Role of Runx2 in IGF-1Rβ/Akt- and AMPK/Erk-dependent growth, survival and sensitivity towards metformin in breast Cancer bone metastasis. *Oncogene* **35**:4730–4740. DOI: https://doi.org/10.1038/onc.2015.518, PMID: 26804175

Tate JG, Bamford S, Jubb HC, Sondka Z, Beare DM, Bindal N, Boutselakis H, Cole CG, Creatore C, Dawson E, Fish P, Harsha B, Hathaway C, Jupe SC, Kok CY, Noble K, Ponting L, Ramshaw CC, Rye CE, Speedy HE, et al. 2019. COSMIC: the catalogue of somatic mutations in Cancer. *Nucleic Acids Research* **47**:D941–D947. DOI: https://doi.org/10.1093/nar/gky1015, PMID: 30371878

Thakur S, Singla AK, Chen J, Tran U, Yang Y, Salazar C, Magliocco A, Klimowicz A, Jirik F, Riabowol K. 2014. Reduced ING1 levels in breast Cancer promotes metastasis. *Oncotarget* **5**:4244–4256. DOI: https://doi.org/10.18632/oncotarget.1988, PMID: 24962136

Tilk S, Curtis C, Petrov DA, McFarland CD. 2020. Most cancers carry a substantial deleterious load due to Hill-Robertson interference. *bioRxiv*. DOI: https://doi.org/10.1101/764340

Todenhöfer T, Seiler R, Stewart C, Moskalev I, Gao J, Ladhar S, Kamjabi A, Al Nakouzi N, Hayashi T, Choi S, Wang Y, Frees S, Daugaard M, Oo HZ, Fisel P, Schwab M, Schaeffeler E, Douglas J, Hennenlotter J, Bedke J, et al. 2018. Selective inhibition of the lactate transporter MCT4 reduces growth of invasive bladder Cancer. *Molecular Cancer Therapeutics* **17**:2746–2755. DOI: https://doi.org/10.1158/1535-7163.MCT-18-0107, PMID: 30262589

Torrente A, Lukk M, Xue V, Parkinson H, Rung J, Brazma A. 2016. Identification of Cancer related genes using a comprehensive map of human gene expression. *PLOS ONE* **11**:e0157484. DOI: https://doi.org/10.1371/journal.pone.0157484, PMID: 27322383

Toyama T, Iwase H, Watson P, Muzik H, Saettler E, Magliocco A, DiFrancesco L, Forsyth P, Garkavtsev I, Kobayashi S, Riabowol K. 1999. Suppression of *ING1* expression in sporadic breast Cancer. *Oncogene* **18**:5187–5193. DOI: https://doi.org/10.1038/sj.onc.1202905, PMID: 10498868

Treekitkarnmongkol W, Katayama H, Kai K, Sasai K, Jones JC, Wang J, Shen L, Sahin AA, Gagea M, Ueno NT, Creighton CJ, Sen S. 2016. Aurora kinase-A overexpression in mouse mammary epithelium induces mammary adenocarcinomas harboring genetic alterations shared with human breast Cancer. *Carcinogenesis* **37**:1180–1189. DOI: https://doi.org/10.1093/carcin/bgw097, PMID: 27624071

Van den Eynden J, Larsson E. 2017. Mutational signatures are critical for proper estimation of purifying selection pressures in Cancer somatic mutation data when using the dN/dS metric. *Frontiers in Genetics* **8**:74. DOI: https://doi.org/10.3389/fgene.2017.00074

van Gijn SE, Wierenga E, van den Tempel N, Kok YP, Heijink AM, Spierings DCJ, Foijer F, van Vugt M, Fehrmann RSN. 2019. TPX2/Aurora kinase A signaling as a potential therapeutic target in genomically unstable cancer cells. *Oncogene* **38**:852–867. DOI: https://doi.org/10.1038/s41388-018-0470-2, PMID: 30177840

Van Tongelen A, Loriot A, De Smet C. 2017. Oncogenic roles of DNA hypomethylation through the activation of cancer-germline genes. *Cancer Letters* **396**:130–137. DOI: https://doi.org/10.1016/j.canlet.2017.03.029, PMID: 28342986

Venkitachalam S, Revoredo L, Varadan V, Fecteau RE, Ravi L, Lutterbaugh J, Markowitz SD, Willis JE, Gerken TA, Guda K. 2016. Biochemical and functional characterization of glycosylation-associated mutational landscapes in Colon cancer. *Scientific Reports* **6**:23642. DOI: https://doi.org/10.1038/srep23642, PMID: 27004849

Verhaak RGW, Bafna V, Mischel PS. 2019. Extrachromosomal oncogene amplification in tumour pathogenesis and evolution. *Nature Reviews Cancer* **19**:283–288. DOI: https://doi.org/10.1038/s41568-019-0128-6, PMID: 30872802

Vitale I, Galluzzi L. 2018. Everybody in! no bouncers at tumor gates. *Trends in Genetics* **34**:85–87. DOI: https://doi.org/10.1016/j.tig.2017.12.006, PMID: 29277455

Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA, Kinzler KW. 2013. Cancer genome landscapes. *Science* **339**:1546–1558. DOI: https://doi.org/10.1126/science.1235122, PMID: 23539594

Vogelstein B, Kinzler KW. 2015. The path to Cancer — Three Strikes and You're Out. *New England Journal of Medicine* **373**:1895–1898. DOI: https://doi.org/10.1056/NEJMp1508811

Voronina N, Wong JKL, Hübschmann D, Hlevnjak M, Uhrig S, Heilig CE, Horak P, Kreutzfeldt S, Mock A, Stenzinger A, Hutter B, Fröhlich M, Brors B, Jahn A, Klink B, Gieldon L, Sieverling L, Feuerbach L, Chudasama P, Beck K, et al. 2020. The landscape of chromothripsis across adult Cancer types. *Nature Communications* **11**:2320. DOI: https://doi.org/10.1038/s41467-020-16134-7, PMID: 32385320

Voutsadakis IA. 2017. Proteasome expression and activity in Cancer and Cancer stem cells. *Tumor Biology* **39**:101042831769224. DOI: https://doi.org/10.1177/1010428317692248, PMID: 28345458

Wada M, Canals D, Adada M, Coant N, Salama MF, Helke KL, Arthur JS, Shroyer KR, Kitatani K, Obeid LM, Hannun YA. 2017. P38 Delta MAPK promotes breast Cancer progression and lung metastasis by enhancing cell proliferation and cell detachment. *Oncogene* **36**:6649–6657. DOI: https://doi.org/10.1038/onc.2017.274, PMID: 28783172

Walker GJ, Walters MK, Palmer JM, Hayward NK. 1994. The *MLLT3* gene maps between D9S156 and D9S171 and contains an unstable polymorphic trinucleotide repeat. *Genomics* **20**:490–491. DOI: https://doi.org/10.1006/geno.1994.1206, PMID: 8034324

Wang J, Yuan W, Chen Z, Wu S, Chen J, Ge J, Hou F, Chen Z. 2012. Overexpression of G6PD is associated with poor clinical outcome in gastric Cancer. *Tumor Biology* **33**:95–101. DOI: https://doi.org/10.1007/s13277-011-0251-9, PMID: 22012600

Wang J, Park JS, Wei Y, Rajurkar M, Cotton JL, Fan Q, Lewis BC, Ji H, Mao J. 2013a. TRIB2 acts downstream of wnt/TCF in liver Cancer cells to regulate YAP and C/EBPα function. *Molecular Cell* **51**:211–225. DOI: https://doi.org/10.1016/j.molcel.2013.05.013, PMID: 23769673

Wang J, Zhang Y, Weng W, Qiao Y, Ma L, Xiao W, Yu Y, Pan Q, Sun F. 2013b. Impaired phosphorylation and ubiquitination by p70 S6 kinase (p70S6K) and smad ubiquitination regulatory factor 1 (Smurf1) promote tribbles homolog 2 (TRIB2) stability and carcinogenic property in liver Cancer. *Journal of Biological Chemistry* **288**:33667–33681. DOI: https://doi.org/10.1074/jbc.M113.503292, PMID: 24089522

Wang T, Birsoy K, Hughes NW, Krupczak KM, Post Y, Wei JJ, Lander ES, Sabatini DM. 2015a. Identification and characterization of essential genes in the human genome. *Science* **350**:1096–1101. DOI: https://doi.org/10.1126/science.aac7041, PMID: 26472758

Wang X, Li X, Zhang X, Fan R, Gu H, Shi Y, Liu H. 2015b. Glucose-6-phosphate dehydrogenase expression is correlated with poor clinical prognosis in esophageal squamous cell carcinoma. *European Journal of Surgical Oncology* **41**:1293–1299. DOI: https://doi.org/10.1016/j.ejso.2015.08.155, PMID: 26329784

Wang X, Li L, Wu Y, Zhang R, Zhang M, Liao D, Wang G, Qin G, Xu RH, Kang T. 2016. CBX4 suppresses metastasis via recruitment of HDAC3 to the Runx2 promoter in colorectal carcinoma. *Cancer Research* **76**:7277–7289. DOI: https://doi.org/10.1158/0008-5472.CAN-16-2100, PMID: 27864346

Wang J, Ye C, Chen C, Xiong H, Xie B, Zhou J, Chen Y, Zheng S, Wang L. 2017a. Glucose transporter GLUT1 expression and clinical outcome in solid tumors: a systematic review and meta-analysis. *Oncotarget* **8**:16875–16886. DOI: https://doi.org/10.18632/oncotarget.15171, PMID: 28187435

Wang J, Nikhil K, Viccaro K, Chang L, Jacobsen M, Sandusky G, Shah K. 2017b. The Aurora-A-Twist1 Axis promotes highly aggressive phenotypes in pancreatic carcinoma. *Journal of Cell Science* **130**:1078–1093. DOI: https://doi.org/10.1242/jcs.196790, PMID: 28167680

Wang J-L, Qi Z, Li Y-H, Zhao H-M, Chen Y-G, Fu W. 2017c. TGFβ induced factor homeobox 1 promotes colorectal Cancer development through activating wnt/β-catenin signaling. *Oncotarget* **8**:70214–70225. DOI: https://doi.org/10.18632/oncotarget.19603

Wang J, Xie Y, Bai X, Wang N, Yu H, Deng Z, Lian M, Yu S, Liu H, Xie W, Wang M. 2018a. Targeting dual specificity protein kinase TTK attenuates tumorigenesis of glioblastoma. *Oncotarget* **9**:3081–3088. DOI: https://doi.org/10.18632/oncotarget.23152, PMID: 29423030

Wang L, Wang J, Jin T, Zhou Y, Chen Q. 2018b. FoxG1 facilitates proliferation and inhibits differentiation by downregulating FoxO/Smad signaling in glioblastoma. *Biochemical and Biophysical Research Communications* **504**:46–53. DOI: https://doi.org/10.1016/j.bbrc.2018.08.118, PMID: 30172378

Wang Y, Li L, Wang H, Li J, Yang H. 2018c. Silencing *TGIF* suppresses migration, invasion and metastasis of MDA–MB–231 human breast cancer cells. *Oncology Reports* **39**:802–808. DOI: https://doi.org/10.3892/or.2017.6133, PMID: 29207164

Wang WX, Yu HL, Liu X. 2019. MiR-9-5p suppresses cell metastasis and epithelial-mesenchymal transition through targeting FOXP2 and predicts prognosis of colorectal carcinoma. *European Review for Medical and Pharmacological Sciences* **23**:6467–6477. DOI: https://doi.org/10.26355/eurrev_201908_18530, PMID: 31378886

Wang J, Zuo J, Wahafu A, Wang MD, Li RC, Xie WF. 2020. Combined elevation of TRIB2 and MAP3K1 indicates poor prognosis and chemoresistance to temozolomide in glioblastoma. *CNS Neuroscience & Therapeutics* **26**:297–308. DOI: https://doi.org/10.1111/cns.13197, PMID: 31318172

Warburg O. 1956a. On respiratory impairment in Cancer cells. *Science* **124**:269–270. DOI: https://doi.org/10.1126/science.124.3215.267, PMID: 13351639

Warburg O. 1956b. On the origin of Cancer cells. *Science* **123**:309–314. DOI: https://doi.org/10.1126/science.123.3191.309, PMID: 13298683

Watkins G, Douglas-Jones A, Mansel RE, Jiang WG. 2005. Expression of thromboxane synthase, TBXAS1 and the thromboxane A2 receptor, TBXA2R, in human breast Cancer. *International Seminars in Surgical Oncology : ISSO* **2**:23. DOI: https://doi.org/10.1186/1477-7800-2-23, PMID: 16250911

Weghorn D, Sunyaev S. 2017. Bayesian inference of negative and positive selection in human cancers. *Nature Genetics* **49**:1785–1788. DOI: https://doi.org/10.1038/ng.3987, PMID: 29106416

Wei LM, Cao S, Yu WD, Liu YL, Wang JT. 2015. Overexpression of CX3CR1 is associated with cellular metastasis, proliferation and survival in gastric Cancer. *Oncology Reports* **33**:615–624. DOI: https://doi.org/10.3892/or.2014.3645, PMID: 25482732

Wei CY, Zhu MX, Yang YW, Zhang PF, Yang X, Peng R, Gao C, Lu JC, Wang L, Deng XY, Lu NH, Qi FZ, Gu JY. 2019. Downregulation of *RNF128* activates wnt/β-catenin signaling to induce cellular EMT and stemness via CD44 and CTTN ubiquitination in melanoma. *Journal of Hematology & Oncology* **12**:21. DOI: https://doi.org/10.1186/s13045-019-0711-z, PMID: 30832692

Wellberg EA, Johnson S, Finlay-Schultz J, Lewis AS, Terrell KL, Sartorius CA, Abel ED, Muller WJ, Anderson SM. 2016. The glucose transporter GLUT1 is required for ErbB2-induced mammary tumorigenesis. *Breast Cancer Research* **18**:131. DOI: https://doi.org/10.1186/s13058-016-0795-0, PMID: 27998284

Weng CC, Hsieh MJ, Wu CC, Lin YC, Shan YS, Hung WC, Chen LT, Cheng KH. 2019. Loss of the transcriptional repressor TGIF1 results in enhanced Kras-driven development of pancreatic Cancer. *Molecular Cancer* **18**:96. DOI: https://doi.org/10.1186/s12943-019-1023-1, PMID: 31109321

Witkiewicz AK, Whitaker-Menezes D, Dasgupta A, Philp NJ, Lin Z, Gandara R, Sneddon S, Martinez-Outschoorn UE, Sotgia F, Lisanti MP. 2012. Using the "reverse Warburg effect" to identify high-risk breast cancer patients: stromal MCT4 predicts poor clinical outcome in triple-negative breast cancers. *Cell Cycle* **11**:1108–1117. DOI: https://doi.org/10.4161/cc.11.6.19530, PMID: 22313602

Wong JCY, Gokgoz N, Alon N, Andrulis IL, Buchwald M. 2003. Cloning and mutation analysis of ZFP276 as a candidate tumor suppressor in breast Cancer. *Journal of Human Genetics* **48**:668–671. DOI: https://doi.org/10.1007/s10038-003-0088-1, PMID: 14605947

Wu S, Lv Z, Wang Y, Sun L, Jiang Z, Xu C, Zhao J, Sun X, Li X, Hu L, Tang A, Gui Y, Zhou F, Cai Z, Wang R. 2013. Increased expression of pregnancy up-regulated non-ubiquitous calmodulin kinase is associated with poor prognosis in clear cell renal cell carcinoma. *PLOS ONE* **8**:e59936. DOI: https://doi.org/10.1371/journal.pone.0059936, PMID: 23634203

Wu J, Liu P, Tang H, Shuang Z, Qiu Q, Zhang L, Song C, Liu L, Xie X, Xiao X. 2018. FOXP2 promotes tumor proliferation and metastasis by targeting GRP78 in Triple-negative breast Cancer. *Current Cancer Drug Targets* **18**:382–389. DOI: https://doi.org/10.2174/1568009618666180131115356, PMID: 29484998

Wushou A, Hou J, Zhao Y-J, Shao Z-M. 2014. Twist-1 Up-Regulation in carcinoma correlates to poor survival. *International Journal of Molecular Sciences* **15**:21621–21630. DOI: https://doi.org/10.3390/ijms151221621, PMID: 25429425

Xiang G, Yi Y, Weiwei H, Weiming W. 2015. TGIF1 promoted the growth and migration of Cancer cells in nonsmall cell lung Cancer. *Tumor Biology* **36**:9303–9310. DOI: https://doi.org/10.1007/s13277-015-3676-8, PMID: 26104768

Xiao H, Wang J, Yan W, Cui Y, Chen Z, Gao X, Wen X, Chen J. 2018. GLUT1 regulates cell glycolysis and proliferation in prostate cancer. *The Prostate* **78**:86–94. DOI: https://doi.org/10.1002/pros.23448, PMID: 29105798

Xu F, Sun S, Yan S, Guo H, Dai M, Teng Y. 2015. Elevated expression of RIT1 correlates with poor prognosis in endometrial Cancer. *International Journal of Clinical and Experimental Pathology* **8**:10315–10324. PMID: 26617739

Xu Y, Lee DK, Feng Z, Xu Y, Bu W, Li Y, Liao L, Xu J. 2017a. Breast tumor cell-specific knockout of *Twist1* inhibits cancer cell plasticity, dissemination, and lung metastasis in mice. *PNAS* **114**:11494–11499. DOI: https://doi.org/10.1073/pnas.1618091114, PMID: 29073077

Xu Y, Qin L, Sun T, Wu H, He T, Yang Z, Mo Q, Liao L, Xu J. 2017b. Twist1 promotes breast cancer invasion and metastasis by silencing Foxa1 expression. *Oncogene* **36**:1157–1166. DOI: https://doi.org/10.1038/onc.2016.286, PMID: 27524420

Xu Y, Wang J, Cai S, Chen G, Xiao N, Fu Y, Chen Q, Qiu S. 2019. PNCK depletion inhibits proliferation and induces apoptosis of human nasopharyngeal carcinoma cells *in vitro* and *in vivo*. *Journal of Cancer* **10**:6925–6932. DOI: https://doi.org/10.7150/jca.33698, PMID: 31839828

Yachida S, Wood LD, Suzuki M, Takai E, Totoki Y, Kato M, Luchini C, Arai Y, Nakamura H, Hama N, Elzawahry A, Hosoda F, Shirota T, Morimoto N, Hori K, Funazaki J, Tanaka H, Morizane C, Okusaka T, Nara S, et al. 2016. Genomic sequencing identifies ELF3 as a driver of ampullary carcinoma. *Cancer Cell* **29**:229–240. DOI: https://doi.org/10.1016/j.ccell.2015.12.012, PMID: 26806338

Yan X, Zhou H, Zhang T, Xu P, Zhang S, Huang W, Yang L, Gu X, Ni R, Zhang T. 2015. Downregulation of FOXP2 promoter human hepatocellular carcinoma cell invasion. *Tumor Biology* **36**:9611–9619. DOI: https://doi.org/10.1007/s13277-015-3701-y, PMID: 26142732

Yang H, Pan L, Xu C, Zhang Y, Li K, Chen S, Zhang B, Liu Z, Wang LX, Chen H. 2017. Overexpression of tumor suppressor gene *ZNF750* inhibits oral squamous cell carcinoma metastasis. *Oncology Letters* **14**:5591–5596. DOI: https://doi.org/10.3892/ol.2017.6908, PMID: 29113187

Yang CA, Huang HY, Lin CL, Chang JG. 2018. G6PD as a predictive marker for glioma risk, prognosis and chemosensitivity. *Journal of Neuro-Oncology* **139**:661–670. DOI: https://doi.org/10.1007/s11060-018-2911-8, PMID: 29845423

Yang HC, Wu Y-H, Yen W-C, Liu H-Y, Hwang T-L, Stern A, Chiu DT-Y. 2019. The redox role of G6PD in cell growth, cell death, and Cancer. *Cells* **8**:1055. DOI: https://doi.org/10.3390/cells8091055, PMID: 31500396

Yang YA, Yu J. 2013. EZH2, an epigenetic driver of prostate Cancer. *Protein & Cell* **4**:331–341. DOI: https://doi.org/10.1007/s13238-013-2093-2, PMID: 23636686

Yao X, Qi L, Chen X, Du J, Zhang Z, Liu S. 2014. Expression of CX3CR1 associates with cellular migration, metastasis, and prognosis in human clear cell renal cell carcinoma. *Urologic Oncology: Seminars and Original Investigations* **32**:162–170. DOI: https://doi.org/10.1016/j.urolonc.2012.12.006, PMID: 23570708

Yasuda K, Hirohashi Y, Kuroda T, Takaya A, Kubo T, Kanaseki T, Tsukahara T, Hasegawa T, Saito T, Sato N, Torigoe T. 2016. MAPK13 is preferentially expressed in gynecological Cancer stem cells and has a role in the tumor-initiation. *Biochemical and Biophysical Research Communications* **472**:643–647. DOI: https://doi.org/10.1016/j.bbrc.2016.03.004, PMID: 26969274

Youn A, Simon R. 2011. Identifying cancer driver genes in tumor genome sequencing studies. *Bioinformatics* **27**:175–181. DOI: https://doi.org/10.1093/bioinformatics/btq630, PMID: 21169372

Yu X, Li Z, Shen J. 2016. *BRD7*: a novel tumor suppressor gene in different cancers. *American Journal of Translational Research* **8**:742–748. PMID: 27158366

Yu X, Zheng H, Chan MTV, Wu WKK. 2018. NOVA1 acts as an oncogene in melanoma via regulating FOXO3a expression. *Journal of Cellular and Molecular Medicine* **22**:2622–2630. DOI: https://doi.org/10.1111/jcmm.13527, PMID: 29498217

Zaika AI, Slade N, Erster SH, Sansome C, Joseph TW, Pearl M, Chalas E, Moll UM. 2002. DeltaNp73, A Dominant-Negative inhibitor of Wild-type p53 and TAp73, is Up-regulated in human tumors. *Journal of Experimental Medicine* **196**:765–780. DOI: https://doi.org/10.1084/jem.20020179

Zanella F, Renner O, García B, Callejas S, Dopazo A, Peregrina S, Carnero A, Link W. 2010. Human TRIB2 is a repressor of FOXO that contributes to the malignant phenotype of melanoma cells. *Oncogene* **29**:2973–2982. DOI: https://doi.org/10.1038/onc.2010.58, PMID: 20208562

Zapata L, Pich O, Serrano L, Kondrashov FA, Ossowski S, Schaefer MH. 2018. Negative selection in tumor genome evolution acts on essential cellular functions and the immunopeptidome. *Genome Biology* **19**:67. DOI: https://doi.org/10.1186/s13059-018-1434-0, PMID: 29855388

Zhang T, Luo Y, Wang T, Yang JY. 2012. MicroRNA-297b-5p/3p target Mllt3/Af9 to suppress lymphoma cell proliferation, migration and invasion *in vitro* and tumor growth in nude mice. *Leukemia & Lymphoma* **53**:2033–2040. DOI: https://doi.org/10.3109/10428194.2012.678005, PMID: 22448917

Zhang C, Zhang Z, Zhu Y, Qin S. 2014a. Glucose-6-phosphate dehydrogenase: a biomarker and potential therapeutic target for Cancer. *Anti-Cancer Agents in Medicinal Chemistry* **14**:280–289. DOI: https://doi.org/10.2174/18715206113136660337, PMID: 24066844

Zhang YA, Zhu JM, Yin J, Tang WQ, Guo YM, Shen XZ, Liu TT. 2014b. High expression of neuro-oncological ventral antigen 1 correlates with poor prognosis in hepatocellular carcinoma. *PLOS ONE* **9**:e90955. DOI: https://doi.org/10.1371/journal.pone.0090955, PMID: 24608171

Zhang MZ, Ferrigno O, Wang Z, Ohnishi M, Prunier C, Levy L, Razzaque M, Horne WC, Romero D, Tzivion G, Colland F, Baron R, Atfi A. 2015. TGIF governs a feed-forward network that empowers wnt signaling to drive mammary tumorigenesis. *Cancer Cell* **27**:547–560. DOI: https://doi.org/10.1016/j.ccell.2015.03.002, PMID: 25873176

Zhang YX, Yan YF, Liu YM, Li YJ, Zhang HH, Pang M, Hu JX, Zhao W, Xie N, Zhou L, Wang PY, Xie SY. 2016. Smad3-related miRNAs regulated oncogenic TRIB2 promoter activity to effectively suppress lung adenocarcinoma growth. *Cell Death & Disease* **7**:e2528. DOI: https://doi.org/10.1038/cddis.2016.432, PMID: 28005074

Zhao ZW, Fan XX, Song JJ, Xu M, Chen MJ, Tu JF, Wu FZ, Zhang DK, Liu L, Chen L, Ying XH, Ji JS. 2017. ShRNA knock-down of CXCR7 inhibits tumour invasion and metastasis in hepatocellular carcinoma after transcatheter arterial chemoembolization. *Journal of Cellular and Molecular Medicine* **21**:1989–1999. DOI: https://doi.org/10.1111/jcmm.13119, PMID: 28429395

Zhao S, Liu J, Nanga P, Liu Y, Cicek AE, Knoblauch N, He C, Stephens M, He X. 2019a. Detailed modeling of positive selection improves detection of Cancer driver genes. *Nature Communications* **10**:3399. DOI: https://doi.org/10.1038/s41467-019-11284-9, PMID: 31363082

Zhao Y, Li W, Li M, Hu Y, Zhang H, Song G, Yang L, Cai K, Luo Z. 2019b. Targeted inhibition of MCT4 disrupts intracellular pH homeostasis and confers self-regulated apoptosis on hepatocellular carcinoma. *Experimental Cell Research* **384**:111591. DOI: https://doi.org/10.1016/j.yexcr.2019.111591, PMID: 31479685

Zhao S, Jie C, Xu P, Diao Y. 2020. MicroRNA-140 inhibit prostate cancer cell invasion and migration by targeting YES proto-oncogene 1. *Journal of Cellular Biochemistry* **121**:482–488. DOI: https://doi.org/10.1002/jcb.29231, PMID: 31310382

Zhong C, Liu J, Zhang Y, Luo J, Zheng J. 2017. MicroRNA-139 inhibits the proliferation and migration of osteosarcoma cells via targeting forkhead-box P2. *Life Sciences* **191**:68–73. DOI: https://doi.org/10.1016/j.lfs.2017.10.010, PMID: 28993144

Zhou Z, Zou Y, Liu G, Zhou J, Wu J, Zhao S, Su Z, Gu X. 2017. Mutation-profile-based methods for understanding selection forces in cancer somatic mutations: a comparative analysis. *Oncotarget* **8**:58835–58846. DOI: https://doi.org/10.18632/oncotarget.19371, PMID: 28938601

Zhu QQ, Ma C, Wang Q, Song Y, Lv T. 2016. The role of TWIST1 in epithelial-mesenchymal transition and cancers. *Tumor Biology* **37**:185–197. DOI: https://doi.org/10.1007/s13277-015-4450-7, PMID: 26602382

Zhu D, Zhou J, Zhao J, Jiang G, Zhang X, Zhang Y, Dong M. 2019. ZC3H13 suppresses colorectal Cancer proliferation and invasion via inactivating Ras-ERK signaling. *Journal of Cellular Physiology* **234**:8899–8907. DOI: https://doi.org/10.1002/jcp.27551, PMID: 30311220

Appendix 1

Examples of genes with strong signatures of positive and/or negative selection

The assignments of the genes to key cellular processes of carcinogenesis are summarized in ***Table 1*** of the main text.

Novel cancer genes positively selected for truncating mutations

Beta-1,3-galactosyltransferase 1, encoded by the *B3GALT1* gene

B3GALT1 belongs to the glycosyltransferase 31 family. It transfers galactose from UDP-alpha-D-galactose to substrates with a terminal beta-N-acetylglucosamine residue. B3GALT1 is involved in the biosynthesis of the carbohydrate moieties of glycolipids and glycoproteins.

It has been suggested that loss of the activity of B3GALT1 may play an important role in aberrant protein glycosylation and tumor progression in colorectal cancers (***Venkitachalam et al., 2016***). Although such a role would be consistent with positive selection for inactivating mutations, analysis of the distribution of nonsense mutations along the protein sequence suggests that the high rNSM value is an artefact, rather than a signature of positive selection for iactivating mutations. The high rate of nonsense substitutions vs. sense substitutions is due to the fact that the majority of sequences contain nonsense substitution at a single site (p.R199*). Since there is no reason why selection would favor nonsense mutation at a single site it seems more likely that it reflects some sort of data deposition error. It is noteworthy in this respect that all the samples containing the p.R199* mutations originate from different regions of pancreatic tumor tissue samples from a single study (***Yachida et al., 2016***).

Bone morphogenetic protein receptor type-2, encoded by the *BMPR2* gene

Bone morphogenetic protein receptor type-2, a member of the TGF beta family of growth factor receptors. Upon ligand binding, it forms a receptor complex consisting of two type II and two type I transmembrane serine/threonine kinases and activates SMAD transcriptional regulators.

There is convincing evidence in the literature that BMPR2 is a tumor suppressor. The *BMPR2* gene has been shown to contain several somatic frameshift mutations and to be inactivated in gastric and colorectal cancers with microsatellite instability (***Kodach et al., 2008***; ***Park et al., 2010***). Loss of BMPR2 function has been found to result in increased tumorigenicity in human prostate cancer cells (***Kim et al., 2004***). More recent studies have shown that disruption of BMPR2 expression promotes mammary carcinoma metastases (***Owens et al., 2012***; ***Pickup et al., 2015***). It was shown that loss of BMPR2 results in increased chemokine expression, which facilitates inflammation by a sustained increase in myeloid cells. The chemokines increased in *BMPR2* deleted cells correlated with poor outcome in human breast cancer patients, suggesting that BMPR2 has tumor suppressive functions in the stroma by regulating inflammation (***Pickup et al., 2015***).

Bromodomain-containing protein 7, encoded by the *BRD7* gene

BRD7 is a crucial component of both functional p53 and BRCA1 pathways and recent studies have fully established BRD7 as a tumor suppressor. The expression of BRD7 was shown to be downregulated in various cancers, including breast cancer, nasopharyngeal carcinoma, gastric cancer, colorectal carcinoma, ovarian cancer, lung adenocarcinoma, NSCLC, hepatocellular carcinoma, and prostate cancer. Moreover, BRD7 inhibited cancer cell growth and metastasis and promoted apoptosis in vitro and in vivo (***Yu et al., 2016***; ***Gao et al., 2016***; ***Chen et al., 2016***; ***Li et al., 2015***).

Recent studies suggest that BRD7 exerts it tumor suppressive role through multiple pathways, by suppressing cell proliferation, initiating cell apoptosis, and reducing aerobic glycolysis (***Niu et al., 2018***). These studies suggest that BRD7 inhibits the Warburg effect through inactivation of the HIF1α/LDHA axis.

Inhibitor of growth protein 1 encoded by the *ING1* gene

ING1 encodes a nuclear, cell cycle-regulated protein, overexpression of which efficiently blocks cell growth and is capable of inducing apoptosis in different experimental systems (**Toyama et al., 1999**). ING1 is known to cooperate with p53/TP53 in the negative regulatory pathway of cell growth by modulating p53-dependent transcriptional activation.

The tumor suppressor status of *ING1* has been fully established since several studies have described the loss of ING1 protein expression in human tumors and *ING1* knockout mice were reported to have spontaneously developed tumors, B cell lymphomas, and soft tissue sarcomas (**Guérillon et al., 2013**).

ING1 levels were found to be lower in breast tumors compared to adjacent normal breast tissue (**Thakur et al., 2014**). Decreasing levels of ING1 increased, and increasing levels decreased migration and invasion of cancer cells in vitro. ING1 overexpression also blocked cancer cell metastasis in vivo and eliminated tumor-induced mortality in mouse models.

ING1 can inhibit the growth of lung cancer cell lines through the induction of cell cycle arrest and apoptosis by forming a complex with p53 (**Luo et al., 2011**; **Bose et al., 2014**)

Genetic alterations that abrogate the normal function of *ING1* may contribute to esophageal squamous cell carcinogenesis (**Chen et al., 2001**). Mutations of the *ING1* tumor suppressor gene (TSG) detected in human melanoma abrogate nucleotide excision repair activity of the protein (**Campos et al., 2004**). Nonsense mutations cluster in the region of residues 339–378. These mutations eliminate the Zn finger domain and polybasic region, which are involved in interaction with histone H3 trimethylated at Lys4 (H3K4me3). It is noteworthy thar histone H3K4me3 binding is required for the DNA repair and apoptotic activities of the ING1 tumor suppressor (**Peña et al., 2008**).

MAX gene-associated protein, encoded by the *MGA* gene

MGA functions as a dual-specificity transcription factor, regulating the expression of both MAX-network and T-box family target genes. Suppresses transcriptional activation by MYC and inhibits MYC-dependent cell transformation. Recurrent inactivation of *MGA*, a suppressor of MYC, has been shown to occur in lymphocytic leukemia, and in both NSCLC and small cell lung cancer, colorectal cancer (**De Paoli et al., 2013**; **Romero et al., 2014**; **Jo et al., 2016**).

Proline-rich transmembrane protein 2, encoded by the *PRRT2* gene

PPRT2, as a component of the outer core of AMPAR complex, is involved in ion channel functions. *PRRT2* has been shown to be significantly downregulated in glioblastoma tissues compared with normal brain tissue (**Bi et al., 2017**; **Li et al., 2018b**). Overexpression of PRRT2 strongly impaired the cell viability and promoted cell apoptosis. These antitumor effects indicate that PRRT2 acts as a tumor suppressor in glioma. PRRT2 has been shown to have an inhibitory effect on proliferation, consistent with the low expression level of PRRT2 in cancer versus normal samples (**Alves et al., 2017**).

Ras GTPase-activating protein 1, encoded by the *RASA1* gene

RASA1 is an inhibitory regulator of the Ras-cyclic AMP pathway. Consistent with the tumor suppressor role of RASA1, the circular RNA circ-ITCH was shown to suppress ovarian carcinoma progression through targeting miR-145/*RASA1* signaling, by increasing the level of RASA1 (**Hu et al., 2018a**).

There is evidence that *RASA1* is a potent TSG that is frequently downregulated or inactivated in several human cancer types. RASA1 expression is frequently reduced in breast cancer tissues, and the reduced RASA1 expression is associated with breast cancer progression and poor survival and disease-free survival of patients (**Liu et al., 2015**).

In hepatocellular carcinoma patients, low level of RASA1 expression correlated with a significantly poorer survival compared to those with high level of RASA1 expression, suggesting that RASA1 could serve as an independent prognostic marker for hepatocellular carcinoma patients (**Chen et al., 2017a**).

Analyses of melanoma whole-genome sequencing data have led to the identification of two novel, clustered somatic missense mutations (Y472H and L481F) in RASA1 (**Sung et al., 2016**). Unlike

wild-type RASA1, these mutants, do not suppresses soft agar colony formation and tumor growth of melanoma cell lines. In addition to mutations, loss of RASA1 expression was frequently observed in metastatic melanoma samples and a low level of RASA1 mRNA expression was associated with decreased overall survival in melanoma patients. Thus, these data support that RASA1 is inactivated by mutations or by suppressed expression in melanoma and that RASA1 plays a tumor suppressive role.

The tumor suppressor role of RASA1 is also supported by the fact that knockdown or miR targeting of *RASA1* significantly enhanced invasion and migration of multiple pancreatic cancer cells (**Sun et al., 2015**; **Kent et al., 2016**).

E3 ubiquitin-protein ligase RNF128, encoded by the *RNF128* gene

E3 ubiquitin-protein ligase RNF128 catalyzes 'Lys-48'- and 'Lys-63'-linked polyubiquitin chains formation. Consistent with its suggested tumor suppressor role, downregulation of *RNF128* was found to predict poor prognosis in patients with urothelial carcinoma and urinary bladder. Downregulation of *RNF128* was correlated with cancer invasiveness and metastasis as well as reduced survival in patients (**Lee et al., 2016**). *RNF128* downregulation was also shown to correlate with the malignant phenotype of melanoma (**Wei et al., 2019**).

Monocarboxylate transporter 1, MCT1 encoded by the *SLC16A1* gene

SLC16A1 is a multipass plasma membrane protein that functions as a proton-coupled monocarboxylate transporter. It catalyzes the rapid transport across the plasma membrane of many monocarboxylates such as lactate. Depending on the tissue and on cicumstances, mediates the import or export of lactic acid. Deficiency of this lactate transporter may result in an acidic intracellular environment created by muscle activity with consequent degeneration of muscle.

Although the high values of rNSM would suggest a tumor suppressor role for *SLC16A1*, several studies suggest that the protein may serve a pro-oncogenic role. For example, depletion of *SLC16A1* was found to decrease cellular proliferation and invasion in both neuroblastoma and malignant cutaneous melanoma cell lines, suggesting its role as an oncogene (OG; **Avitabile et al., 2020**). The pro-oncogenic role of MCT1 is also supported by the results of studies on esophageal squamous cell carcinoma (ESCC). Kaplan-Meier survival analysis of ESCC patients in a high-MCT1 group had a lower overall survival and lower progression-free survival, whereas downregulation of MCT1 suppressed proliferation and survival of ESCC cells in vitro (**Chen et al., 2019**). Disrupting MCT1 function leads to an accumulation of intracellular lactate that rapidly disables tumor cell growth (**Doherty et al., 2014**).

MCT1 expression is elevated in glycolytic breast tumors, and high MCT1 expression predicts poor prognosis in breast and lung cancer patients. Similarly, the observations that MCT1 inhibition impairs proliferation of glycolytic breast cancer cells co-expressing MCT1 and MCT4 and that MCT1 loss-of-function decreases breast cancer cell proliferation and blocks growth of mammary fat pad xenograft tumors suggest a pro-oncogenic or tumor essential role for MCT1 (**Hong et al., 2016**).

A recent study, however, has led to the conclusion that MCT1 and MCT4 have opposing roles in carcinogenesis (**Sukeda et al., 2019**). In a retrospective survey conducted on patients who underwent surgical resection for pancreatic ductal adenocarcinoma, the expression of MCT1, MCT4, and GLUT1 was assessed in tumor cells and cancer-associated fibroblasts (CAFs) and the impact of their expression on patient outcome was also analyzed. In tumor cells, MCT1 expression was associated with extended overall and progression-free survival and decreased nodal metastasis. Conversely, MCT4 expression in CAFs was associated with shortened survival. In other words, in tumor cells, MCT1 expression is associated with better prognosis and reduced nodal metastasis in pancreatic cancer, contrary to findings of previous studies.

It is noteworthy in this respcect that based on the pattern of mutations SLC16A1/MCT1 appers to be a tumor suppressor rather than a tumor essential gene (TEG) in as much as it has a high proportion of truncating mutations. It seems possible that glycolytic tumor cells that must get rid of lactate are selected for increased efflux and decreased influx of lactate and this might be achieved by increased expression of MCT4 and decreased activity of MCT1.

Sprouty-related, EVH1 domain-containing protein 1, encoded by the *SPRED1* gene

The *SPRED1* gene, which encodes a negative regulator of mitogen-activated protein kinase (MAPK) signaling, has been shown to function as a TSG in several types of cancer (**Pasmant et al., 2015**; **Ablain et al., 2018**; **Sun et al., 2019**).

Homeobox protein TGIF1, encoded by the *TGIF1* gene

TGIF binds to a retinoid X receptor (RXR)-responsive element from the cellular retinol-binding protein II promoter (CRBPII-RXRE). Inhibits the 9-cis-retinoic acid-dependent RXR alpha transcription activation of the retinoic-acid-responsive element. Active transcriptional corepressor of SMAD2.

There is evidence that TGIF1 may function as a tumor suppressor. In pancreatic ductal adenocarcinoma genetic inactivation of *TGIF1* in the context of oncogenic KRASG12D, culminated in the development of highly aggressive and metastatic pancreatic ductal adenocarcinoma (**Parajuli et al., 2019**; **Weng et al., 2019**). These authors have found that TGIF1 associates with TWIST1 and inhibits TWIST1 expression and activity, and this function is suppressed in the vast majority of human pancreatic ductal adenocarcinoma by KRASG12D /MAPK-mediated TGIF1 phosphorylation. Ablation of TWIST1 in KRASG12D;TGIF1KO mice blocked pancreatic ductal adenocarcinoma formation, providing evidence that TGIF1 restrains KRASG12D-driven pancreatic ductal adenocarcinoma through its ability to antagonize TWIST1.

The majority of available evidence, however, suggests that the protein plays a cancer promoting role. TGIF1 has been shown to promote the growth and migration of cancer cells in NSCLC (**Xiang et al., 2015**). The authors have shown that expression of TGIF1 is elevated in NSCLC tissues, that TGIF1 promoted the growth and migration of cancer cells and that knocking down the expression of *TGIF1* inhibited the growth and migration of NSCLC cells. These studies have also shown that TGIF1 exerted its oncogenic role through beta-catenin/TCF signaling.

Studies on triple negative breast cancer have revealed that high levels of TGIF expression correlate with poor prognosis since TGIF promotes Wnt-driven mammary tumorigenesis. As to the molecular mechanism of the oncogenic role of TGIF: it has been shown that TGIF interacts with and sequesters Axin1 and Axin2 into the nucleus, disassembles the β-catenin-destruction complex leading to the accumulation of β-catenin that activates expression of Wnt target genes (**Zhang et al., 2015**; **Razzaque and Atfi, 2016**).

In harmony with an oncogenic role of TGIF in breast cancer, silencing of *TGIF* was found to suppress the migration, invasion and metastasis of the human breast cancer cells in both in vitro and in vivo experiments (**Wang et al., 2018c**).

TGIF1 has also been found to be significantly upregulated in some colorectal cancers and to promote adenoma growth in the context of mutant Apc (**Shah et al., 2019**). Overexpression of TGIF1 markedly promoted the proliferation of colorectal cancer cells through the activation of Wnt/β-catenin signaling (**Wang et al., 2017c**).

In summary, the majority of data suggest that *TGIF1* may act as an OG, despite the fact that the high proportion of truncating indel mutations would indicate a tumor suppressor function. Since the transcription regulator TGIF1 may play both pro-oncogenic and tumor suppressor functions (in different cellular processes), our observation that during tumor evolution selection for truncating mutations appears to dominate for TGIF1 suggests that the selection pressure to eliminate the tumor suppressor activity may override the pressure to preserve it oncogenic activities.

Trinucleotide repeat-containing gene 6B protein, encoded by the *TNRC6B* gene

TNRC6B is a key miRNA-processing gene that plays a role in RNA-mediated gene silencing by both micro-RNAs (miRNAs) and short interfering RNAs (siRNAs). TNRC6B is required for miRNA-dependent translational repression and siRNA-dependent endonucleolytic cleavage of complementary mRNAs by argonaute family proteins.

Genomic analysis of liver cancer have identified *TNRC6B* as a significantly mutated gene, suggesting that it may be an important driver gene (**Li et al., 2018a**). Consistent with its putative tumor

suppressor role, DNA methylation of *TNRC6B* has been suggested to play a role in early carcinogenesis (**Joyce et al., 2018**).

Dual specificity protein kinase TTK, encoded by the *TTK* gene

TTK, capable of phosphorylating serine, threonine, and tyrosine residues of proteins, plays a role in cell proliferation. Although, intuitively the high rate of truncating mutations would suggest a tumor suppressor role for TTK, all the available evidence indicates that it acts as an OG.

It has been shown that dual specificity kinase TTK is strongly overexpressed in human pancreatic ductal adenocarcinoma, suggesting a cancer promoting role. In harmony with such a role, following *TTK* knockdown cell proliferation was significantly attenuated, whereas apoptosis and necrosis rates were significantly increased. Apoptosis was associated with increased formation of micronuclei, suggesting that loss of TTK results in chromosomal instability and mitotic catastrophe (**Kaistha et al., 2014**).

Levels of TTK protein were also found to be significantly elevated in neoplastic tissues of liver cancer patients, when compared with adjacent hepatic tissues. In an experimental animal model, it was shown that in vitro knockdown of *TTK* effectively blocks intrahepatic growth of human hepatic carcinoma cell xenografts, suggesting that targeted TTK inhibition might have clinical utility in the therapy of liver cancer (**Miao et al., 2016**).

In a recent study, dual specificity protein kinase TTK has been identified as the most upregulated and differentially expressed kinase in glioma stem-like cells that are responsible for tumorigenesis and subsequent tumor recurrence in glioblastoma. TTK expression was highly enriched in glioblastoma and was inversely correlated with a poor prognosis (**Wang et al., 2018a**).

The deubiquitinase, USP9X has been implicated in multiple cancers and its oncogenic effects were shown to be exerted at least in part through dual specificity protein kinase TTK (**Chen et al., 2018d**). USP9X was found to stabilize TTK by efficient deubiquitination of the kinase; levels of USP9X and TTK were significantly elevated and positively correlated in tumor tissues, suggesting that the USP9X-TTK axis plays a critical role in carcinogenesis. In harmony with the synergism of these OGs, knockdown of *USP9X* or *TTK* inhibited cell proliferation, migration, and tumorigenesis.

The explanation for the apparent contradiction of the oncogenic role of TTK and the abundance of truncating mutations in the protein probably lies in the fact that – unlike in the case of typical TSGs – mutations are not randomly distributed along the protein sequence. The truncating mutations are practically restricted to the very C-terminal end of the protein (EKKRGKK, residues 851–857), downstream of the catalytic domain and missense mutations also cluster in this C-terminal end. It seems likely that this region is involved in some negative control of the activity of TTK and missense and truncating mutations liberate TTK from this negative control. It is unclear at present whether the mutations affecting this C-terminal motif activate the TTK proto-oncogene by interfering with its ubiquitination or by affecting its subcellular localization.

Zinc finger CCCH domain-containing protein 13, encoded by the *ZC3H13* gene

ZC3H13 is associated with a complex that mediates N6-methyladenosine (m6A) methylation of RNAs, a modification that plays a role in the efficiency of mRNA splicing and RNA processing. It acts as a key regulator of m6A methylation by promoting m6A methylation of mRNAs at the 3'-UTR. ZC3H13 has been shown to serve as a tumor suppressor in colorectal cancer (**Zhu et al., 2019**).

mRNA decay activator protein ZFP36L2, encoded by the *ZFP36L2* gene

ZFP36L2 has been selected as a gene characterized by very high values of indel_rNSM, suggesting positive selection for truncating mutations. Although the closely related *ZFP36L1* gene is not present in the lists defined by the CG_SO and CG_SSI lists defined by the 2SD cut-off values, it is also characterized by very high values of rNSM (**Supplementary file 5**).

ZFP36L1 and ZFP36L2 zinc-finger RNA-binding proteins destabilize several cytoplasmic AU-rich element (ARE)-containing mRNA transcripts by promoting their poly(A) tail removal or deadenylation, and hence provide a mechanism for attenuating protein synthesis. The proteins are necessary

for thymocyte development and prevention of T-cell acute lymphoblastic leukemia transformation by promoting ARE-mediated mRNA decay of the mRNA of oncogenic factors.

Deletion of the genes *ZFP36L1* and *ZFP36L2* leads to perturbed thymic development and T lymphoblastic leukemia (**Hodson et al., 2010**).

ZFP36L1 and ZFP36L2 play a negative role in cell proliferation. Forced expression of ZFP36L1 or ZFP36L2 inhibited cell proliferation in colorectal cancer cell lines, whereas knockdown of these genes increased cell proliferation (**Suk et al., 2018**). ZFP36L2 has been validated as an important tumor-suppressor specific to oesophageal squamous cell carcinomas (**Lin et al., 2018**).

Zinc finger protein 276, encoded by the *ZNF276* gene

Zinc finger protein is involved in transcriptional regulation.

It has been suggested that ZNF276 may be a tumor suppressor in breast cancer progression in colorectal cancers (**Wong et al., 2003**).

Although such a role would be consistent with positive selection for inactivating mutations, analysis of the distribution of nonsense mutations along the protein sequence suggests that the high rNSM value is an artefact, rather than a signature of positive selection for inactivating mutations. The high rate of nonsense substitutions vs. sense substitutions is due to the fact that the majority of sequences contain nonsense substitution at a single site (p.Q217*). Since there is no reason why selection would favor nonsense mutation at a single site, it seems more likely that it reflects some sort of data deposition error. It is noteworthy in this respect that all the samples containing the p.Q217* mutations originate from different regions of pancreatic tumor tissue samples from a single study (**Yachida et al., 2016**).

Zinc finger protein 750, encoded by the *ZNF750* Gene

Zinc finger protein 750 is a transcription factor required for terminal epidermal differentiation, it acts downstream of p63/TP63. Its mutations have been shown to abolish the ability to induce epidermal terminal differentiation. In harmony with its mutation pattern, numerous studies suggest a tumor suppressor role for ZNF750.

Analysis of cancer genes across 21 tumor types identified *ZNF750* as a gene harboring many early frameshift and nonsense mutations in head and neck cancer and as the only known gene residing in a small current focal deletion in head and neck and lung squamous cancers (**Lawrence et al., 2014**). *ZNF750* has also been identified as a tumor suppressor in oral and esophageal squamous cell carcinoma (**Yang et al., 2017**; **Nambara et al., 2017**; **Hazawa et al., 2017**; **Otsuka et al., 2018**). Studies on the clonal evolution in esophageal squamous cell carcinoma revealed that the majority of driver mutations in this cancer occurred in the tumor-suppressor genes, including *TP53, KMT2D*, and *ZNF750* (**Hao et al., 2016**).

Novel cancer genes positively selected for missense mutations

Aurora kinase A encoded by the *AURKA* gene

AURKA, also known as a Breast tumor-amplified kinase, is a mitotic serine/threonine kinase that contributes to the regulation of cell cycle progression. It associates with the centrosome and the spindle microtubules during mitosis and plays a critical role in various mitotic events.

In harmony with the notion that AURKA's mutation pattern reflects a pro-oncogenic role for the protein, elevated expression of AURKA has been shown to induce oncogenic phenotypes (**Takahashi et al., 2015**; **Treekitkarnmongkol et al., 2016**).

Similarly, the observation that downregulation, inhibition or depletion of AURKA reduced viability and invasiveness of cancer cells (**Sillars-Hardebol et al., 2012**; **Li et al., 2018a**; **van Gijn et al., 2019**) also argues for an oncogenic role of the protein.

Significantly, specific knockdown of *AURKA* in cultured pancreatic cancer cells strongly suppressed in vitro cell growth and in vivo tumorigenicity (**Hata et al., 2005**). Recently, a novel AURKA mutation (V352I) was identified from clinical specimens and it was shown that AURKA (V352I)-induced carcinogenesis was earlier and much more severe than wild-type AURKA, implying that the V352I mutation may accelerate cancer progression (**Su et al., 2019**).

Although many *AURKA* mutations were identified in cancer patients, it is noteworthy that there is no evidence for the clustering or 'recurrence' of mutations. The most likely explanation for the lack of clustering of mutations is that since AURKA interacts with numerous proteins (e.g. PIFO, GADD45A, AUNIP, NIN, FRY, SIRT2, MYCN, HNRNPU, AAAS, KLHL18, CUL3, FOXP1) there may be multiple sites where missense mutations affecting these interactions may result in disregulation of the activity of **AURKA**.

In summary, although all the available experimental information argues for an oncogenic role of *AURKA*, there was no evidence for the clustering of its missense mutations. In our view, this observation illustrates that recurrence of missense mutations is not a *sine qua non* criterion of OGs.

Recent studies have also revealed that AURKA and TWIST1 are linked in a feedback loop controlling tumorigenesis and metastasis. AURKA phosphorylates TWIST1, inhibits its ubiquitylation, increases its transcriptional activity and favors its homodimerization. TWIST1 prevents AURKA degradation, thereby triggering a feedback loop. Ablation of either AURKA or TWIST1 completely inhibits epithelial-to-mesenchymal transition, suggesting that inhibition of AURKA and TWIST1 are synergistic in inhibiting tumorigenesis and metastasis (***Wang et al., 2017b***).

Although the *TWIST1* gene is not present in the datasets (***Supplementary files 5*** and ***31***) that contain the metadata for transcripts containing at least 100 confirmed somatic, non-polymorphic mutations identified in tumor tisses, inspection of the primary dataset (***Supplementary file 9***) indicates that it is characterized by very high value of rSMN (***Supplementary file 5***), indicating strong signature of purifying selection (see section on Negatively selected genes) consistent with the view that – in synergism with AURKA – it plays an important role in promoting tumorigenesis.

Cyclin-dependent kinase 8, encoded by the *CDK8* gene

The *CDK8* gene is a coactivator involved in regulated gene transcription of nearly all RNA polymerase II-dependent genes.

CDK8 is a colorectal cancer OG that regulates beta-catenin activity. Suppression of CDK8 expression inhibits proliferation in colon cancer cells characterized by high levels of CDK8 and beta-catenin hyperactivity (***Firestein et al., 2008***). CDK8 has been shown to promote SMAD1-driven epithelial-to-mesenchymal transition through YAP1 recruitment (***Serrao et al., 2018***). There is a large body of evidence that CDK8 is a key oncogenic driver in many cancers (***Philip et al., 2018***). *CDK8* was found to be amplified or overexpressed in many colon cancers and CDK8 expression correlated with shorter patient survival (***Liang et al., 2018***).

Isocitrate dehydrogenase [NAD] subunit beta, mitochondrial, encoded by the *IDH3B* gene

IDH3B plays an essential role in the activity of isocitrate dehydrogenase. The heterodimer composed of the alpha (IDH3A) and beta (IDH3B) subunits and the heterodimer composed of the alpha (IDH3A) and gamma (IDH3G) subunits, have significant activity but the full activity of the heterotetramer (containing two subunits of IDH3A, one of IDH3B and one of IDH3G) requires the assembly of both heterodimers.

Our Pubmed search failed to identify publications with major relevance for the role of *IDH3B* in carcinogenesis. It is noteworthy, however, that the *IDH3B* gene contains recurrent somatic missense mutations at residue R131 that is equevalent with R132 and R140 of the paralogous enzymes, IDH1 and IDH2, respectively, that are affected by recurrent oncogenic missense mutations. These mutations of IDH1 and IDH2 result in loss of normal enzymatic function and the abnormal production of 2-hydroxyglutarate. 2-Hydroxyglutarate has been found to inhibit enzymatic function of many alpha-ketoglutarate-dependent enzymes, including histone and DNA demethylases, causing widespread epigenetic changes in the genome thereby promoting tumorigenesis. It seems likely that the R131 mutations of IDH3B may conntribute to carcinogenesis by a similar mechanism.

E3 ubiquitin-protein ligase MARCH7, encoded by the *MARCH7* gene

March7 is an E3 ubiquitin-protein ligase, an enzyme that accepts ubiquitin from an E2 ubiquitin-conjugating enzyme and then directly transfer the ubiquitin to targeted substrates.

Several studies support an oncogenic role for the ubiquitin E3 ligase MARCH7. Studies on ovarian tissues have revealed that expression of MARCH7 was higher in ovarian cancer tissues than normal ovarian tissues. Silencing *MARCH7* decreased, whereas ectopic expression of MARCH7 increased cell proliferation, migration and invasion, suggesting that MARCH7 is oncogenic and a potential target for ovarian cancer therapy (**Hu et al., 2015**). The expression of MARCH7 was significantly higher in cervical cancer tissues than normal cervical tissues, suggesting that this OG may also serves as a potential target for cervical cancer therapy (**Hu et al., 2018b**).

The expression level of MARCH7 in endometrial cancer tissues was also found to be significantly higher than that in normal endometrium tissues, suggesting that it may be an oncogenic factor in endometrial cancer (**Liu et al., 2019**). The oncogenic role of MARCH7 is supported by the fact its knockdown inhibited the invasion and metastasis of endometrial cancer cells in vitro and in vivo, whereas the opposite effect was observed after overexpressing MARCH7.

GTP-binding protein RIT1, encoded by the *RIT1* gene

The high value of rMSN reflects primarily the recurrence of substitutions (Met90Ile, Met90Val) of Met90 of RIT1 protein.

RIT1 plays a crucial role in the activation of MAPK signaling cascades that mediate a wide variety of cellular functions, including cell proliferation, survival, and differentiation.

Since the Met90Ile substitution has been shown to result in an increased MAPK-ERK signaling (**Aoki et al., 2013**; **Koenighofer et al., 2016**), it is plausible to assume that the high rate of missense mutations reflects positive selection of oncogenic driver mutations.

In harmony with this conclusion, studies on endometrial cancer have revealed that RIT1 mRNA and protein were significantly overexpressed in endometrial cancer cell lines and in endometrial cancer tissues compared to non-cancerous endometrial tissue samples (**Xu et al., 2015**). Elevated expression of RIT1 was significantly correlated with pathological type, clinical stage. Kaplan-Meier survival analysis indicated that RIT1 expression was associated with poor overall survival of endometrial cancer patients, suggesting that elevated expression of RIT1 may contribute to the progression of endometrial cancer.

In a study of lung adenocarcinoma cases, several somatic mutations (including Met90Ile) were identified in the *RIT1* gene that were found to cluster in a hotspot near the switch II domain of the GTPase protein (**Berger et al., 2014**). Ectopic expression of these mutated *RIT1* genes was found to induce cellular transformation in vitro and in vivo, confirming that these substitutions are driver mutations and that *RIT1* is an OG in lung adenocarcinoma.

Yes-associated protein 1, encoded by the *YAP1* gene

Yes-associated protein one is known to be the critical downstream regulatory target in the Hippo signaling pathway that plays a pivotal role in tumor suppression by restricting proliferation and promoting apoptosis. This pathway is composed of a kinase cascade that eventually inactivates YAP1 since phosphorylation of YAP1 by the tumor suppressors LATS1/2 inhibits its translocation into the nucleus.

Several lines of evidence indicate that YAP1 is an OG. *YAP1* was found to act as oncogenic target of 11q22 amplification in multiple cancer subtypes, whereas *YAP1* silencing significantly decreases cell proliferation (**Lorenzetto et al., 2014**; **Hamanaka et al., 2019**). *YAP1* was shown to promote growth of prostate cancer, whereas knock down of its expression or inhibition of YAP1 function significantly suppressed tumor recurrence (**Jiang et al., 2017**). The key role of YAP1 in carcinogenesis is also supported by the fact that the tumor suppressor LATS2 inhibits the malignant behaviors of glioma cells by inactivating of YAP1 (**Shi et al., 2019**).

Although several *YAP1* mutations were identified in cancer patients, there is no evidence for the clustering or 'recurrence' of mutations. Similarly to the case of AURKA (see above), the most plausible explanation for the lack of clustering of mutations of this OG is that since YAP1 interacts with several proteins (e.g. YES kinase, LATS1, LATS2, TP73, RUNX1, WBP1, WBP2, TEAD1, TEAD2, TEAD3, TEAD4, HCK, MAPK8, MAPK9, CK1, ABL1) mutations at several different sites may affect these interactions and may result in disregulation of the activity of YAP1. In our view, the cases of

AURKA, YAP1 and YES1 illustrate that recurrence of missense mutations is not a *sine qua non* criterion of OGs.

Tyrosine-protein kinase Yes, encoded by the *YES1* gene

Tyrosine-protein kinase Yes (also known as proto-oncogene c-Yes) is a multidomain non-receptor protein tyrosine kinase containing an SH3 domain, an SH2 domain and a protein kinase domain. YES1 is involved in the regulation of cell growth and survival, apoptosis, cell-cell adhesion, cytoskeleton remodeling, and differentiation. It plays a role in cell cycle progression by phosphorylating the cyclin-dependent kinase 4/CDK4 thus regulating the G1 phase. YES1 has been shown to phosphorylate YAP1, leading to the localization of a YAP1-TBX-β-catenin complex to the promoters of antiapoptotic genes, thereby promoting carcinogenes (**Rosenbluh et al., 2012**). A small-molecule inhibitor of YES1 impeded the proliferation of β-catenin-dependent cancers in both cell lines and animal models.

Several lines of evidence have established an oncogenic role for *YES1*.

It has been demonstrated recently that YES1 is essential for lung cancer growth and progression in NSCLC, suggesting that it is a promising therapeutic target in lung cancer. YES1 overexpression induced metastatic spread in preclinical in vivo models, whereas *YES1* genetic depletion by CRISPR/Cas9 technology significantly reduced tumor growth and metastasis (**Garmendia et al., 2019**).

In harmony with an oncogenic role of *YES1*, several microRNAs have been shown to inhibit the proliferation of tumor cells by targeting *YES1* (**Tan et al., 2015**; **Shen et al., 2019**; **Zhao et al., 2020**).

The oncogenic role of YES1 in cancer is also supported by the observation that it is amplified in several types of cancer, suggesting that it could be an attractive target for a cancer drug (**Fan et al., 2018**; **Hamanaka et al., 2019**). **Hamanaka et al., 2019** have generated a YES1 kinase inhibitor, and have shown that YES1 kinase inhibition by this drug led to antitumor activity against *YES1*-amplified cancers in vitro and in vivo. The authors have also shown that Yes-associated protein 1 (YAP1) played a role downstream of YES1 and contributed to the growth of *YES1*-amplified cancers. indicating that the regulation of YAP1 by YES1 plays an important role in *YES1*-amplified cancers. These findings identify YES1 as a targetable OG of significant potential for clinical utility (**Rai, 2019**).

Although *YES1* contains an increased proportion of nonsynonymous mutations there is no evidence for the clustering or 'recurrence' of mutations. Similarly to the cases of *AURKA* and *YAP1* (see above), the most plausible explanation for the lack of clustering of mutations of this OG is that since YES1 is a multidomain protein that interacts with several proteins, mutations at several different sites may affect these interactions and may result in disregulation of the activity of YAP1. In our view, the cases of AURKA, YAP1, and YES1 illustrate that recurrence of missense mutations is not a *sine qua non* criterion of OGs.

Negatively selected TEGs

Atypical chemokine receptor 3, encoded by the *ACKR3* (*CXCR7*) gene

ACKR3 is a member of the group of chemokine receptors that acts as a receptor for chemokines CXCL11 and CXCL12/SDF1. It is activated by CXCL11 in malignant hemapoietic cells, leading to phosphorylation of ERK1/2 (MAPK3/MAPK1) and enhanced cell adhesion and migration.

ACKR3 is a known cancer gene, from Tier 1 of the Cancer Gene Census; it has a cancer hallmark annotation. Its importance in carcinogenesis is underlined by the fact that high expression of ACKR3 is associated with poor survival in several types of cancer.

As to the role of ACKR3 in hallmarks of cancer: it has been suggested that ACKR3 promotes proliferative signaling, angiogenesis, evasion of programmed cell death and invasion and metastasis.

Several studies support the key role of ACKR3 in tumor invasion and metastasis (**Li et al., 2014**; **Stacer et al., 2016**; **Zhao et al., 2017**; **Puddinu et al., 2017**; **Melo et al., 2018**; **Qian et al., 2018**). Since knock-down or pharmacological inhibition of *ACKR3* has been shown to reduce tumor invasion and metastasis, ACKR3 is a promising therapeutic target for the control of tumor dissemination.

CX3C chemokine receptor 1, encoded by the *CX3CR1* gene

CX3CR1 is a member of the group of chemokine receptors that play a major role in tumor metastasis. The interactions of chemokines, also known as chemotactic cytokines, with their receptors regulate immune and inflammatory responses. However, recent studies have demonstrated that cancer cells subvert the normal chemokine role, transforming them into fundamental constituents of the tumor microenvironment with tumor-promoting effects. CX3CR1 is the receptor for the CX3C chemokine fractalkine (CX3CL1) that mediates both its adhesive and migratory functions.

CX3CR1 expression has been shown to be associated with the process of cellular migration in vitro and tumor metastasis of clear cell renal cell carcinoma in vivo (**Yao et al., 2014**).

Recent studies indicate that tumor-associated macrophages MΦ can influence cancer progression and metastasis and that CCR2 and CX3CR1 play important roles in metastasis. **Schmall et al., 2015** have shown that coculturing of tumor-associated macrophages with mouse Lewis lung carcinoma caused up-regulation of CCR2/CCL2 and CX3CR1/CX3CL1 in both the cancer cells and the macrophages. In vivo, MΦ depletion and genetic ablation of *CCR2* and *CX3CR1* all inhibited LLC1 tumor growth and metastasis, and enhanced survival. Furthermore, mice treated with CCR2 antagonist mimicked genetic ablation of CCR2, showing reduced tumor growth and metastasis. These findings indicate that tumor-associated MΦ play a central role in lung cancer growth and metastasis, with bidirectional cross-talk between MΦ and cancer cells via CCR2 and CX3CR1 signaling. These studies suggest that the therapeutic strategy of blocking CCR2 and CX3CR1 may prove beneficial for halting metastasis.

CX3CR1 is highly expressed in gastric cancer tissues and is related to lymph node metastasis and larger tumor size. CX3CR1 overexpression promoted gastric cancer cell migration, invasion, proliferation, and survival (**Wei et al., 2015**).

CX3CR1 is overexpressed in human breast tumors and cancer cells utilize the chemokine receptor CX3CR1 to exit the blood circulation and metastize to the skeleton. To assess the clinical potential of targeting CX3CR1 in breast cancer **Shen et al., 2016** have used neutralizing antibody for this receptor, transcriptional suppression by CRISPR interference as well as a potent and selective small-molecule antagonist of CX3CR1 in preclinical animal models of metastasis. The authors have found that inactivation of CX3CR1 impairs the lodging of circulating tumor cells to the skeleton and impairs further growth of established metastases. These data suggest that CX3CR1 has a important role in promoting metastasis activity and that CX3CR1 antagonists may valuable as drugs of tumor therapy.

C-C chemokine receptor type 2, encoded by the *CCR2* gene

C-C chemokine receptor type 5, encoded by the *CCR5* gene

Although the *CCR2* gene of C-C chemokine receptor type two and the *CCR5* gene of C-C chemokine receptor type five are not present in the CG_SO and CG_SSI lists defined by the 2SD cut-off values they are also characterized by very high values of rSNM (**Supplementary file 5**), suggesting that they may also play importanct roles in tumor metastasis.

CCR2 is the key functional receptor for the chemokine ligand CCL2. Its binding with CCL2 on monocytes and macrophages mediates chemotaxis and migration induction. Recent studies indicate that CCR2 and CX3CR1 play important roles in metastasis (**Schmall et al., 2015**). The CCL2-CCR2 signaling axis has generated increasing interest in recent years due to its association with the progression of cancer. The CCL2-CCR2, signaling pair has been shown to have multiple pro-tumorigenic roles, mediating tumor growth and angiogenesis (**Lim et al., 2016**).

CCR5 serves as a receptor for a number of inflammatory CC-chemokines including CCL3/MIP-1-alpha, CCL4/MIP-1-beta. Recent studies have revealed that C-C chemokine receptor type 5 plays a key role in progression of tumorigenesis, Expression of CCR5 augments regulatory T cell differentiation and migration to sites of inflammation. The misexpression of CCR5 in epithelial cells, induced upon oncogenic transformation, hijacks this migratory phenotype (**Aldinucci and Casagrande, 2018**; **Jiao et al., 2019**).

Dentin sialophosphoprotein, encoded by the *DSPP* gene

The *DSPP* gene has been selected as a gene showing very high values of rSMN, suggesting negative selection of missense and nonsense mutations (**Supplementary file 5**). It must be pointed out that

based on the high silent/missense ratio *DSPP* has also been identified by others as a gene showing signs of strong negative selection (**Zhou et al., 2017**).

Dentin sialophosphoprotein is a secreted protein that has been shown to play an important role in dentinogenesis. It binds high amount of calcium and facilitates initial mineralization of dentin matrix collagen as well as regulate the size and shape of the crystals, therefore it seemed surprising that its gene would qualify as a negatively selected TEG.

There is evidence in the scientific literature that the protein may have a tumorigenic role in oral cancer (**Chaplet et al., 2006**; **Joshi et al., 2010**; **Saxena et al., 2015**; **Gkouveris et al., 2018**; **Nikitakis et al., 2018**). Nevertheless, the high silent to missense rate is not a reflection of the importance of *DSPP* for carcinogenesis. The *DSPP* gene contains a 2 kb repeat domain containing over 200 tandem copies of a nominal 9-basepair (AGC AGC GAC) repeat encoding a series of tandem Ser Ser Asp repeats and the unusully high rate of silent mutations is restricted to this region of the gene.

A study of 188 normal human chromosomes revealed that the repeat domain of *DSPP* is hypervariable with extraordinary rates of change including slip-replication indel events and predominantly C-to-T transition SNPs (**McKnight et al., 2008**). In harmony with the increased rate and predominance of C-to-T transition in the AGC AGC GAC (Ser-Ser-Asp) repeats, the vast majority of substitutions in this repeat region of the *DSSP* gene are silent. The unusually high silent to missense mutation ratio of the *DSPP* gene is thus not due to purifying selection of a TEG.

Forkhead box protein G1, encoded by the *FOXG1*

FOXG1 is a member of the FOX (Forkhead box) protein family of transcription factors that play important roles in regulating the expression of genes involved in cell growth, proliferation, differentiation, and longevity. FOXG1 localizes to mitochondria and coordinates cell differentiation and bioenergetics (**Pancrazi et al., 2015**).

The tumor promoting role of FOXG1 is supported by the observation that childhood medulloblastomas are characterized by 2–7-fold copy gain for *FOXG1*. *FOXG1* copy gain (>2 to 21 folds) was seen in 93% of a validating set of tumors and showed a positive correlation with protein expression (**Adesina et al., 2007**).

The oncogenic role of FOXG1 is also supported by the observation that a decrease of FOXG1 in medulloblastoma cells offers a survival advantage in mice (**Adesina et al., 2015**), whereas high expression of FOXG1 was associated with poor survival of glioblastoma patients (**Robertson et al., 2015**).

The carcinogenesis promoting activity of FOXG1 is supported by the observation that endogenous FOXG1 expression levels were positively correlated to the glioblastoma multiforme disease progression (**Wang et al., 2018b**). Overexpression of FOXG1 protein resulted in increased cell viability, and it was suggested that FOXG1 functions as an onco-factor by promoting proliferation and inhibiting differentiation.

Recent studies on glioblastoma have shown that transcription factors FOXG1 and TLE1 promote glioblastoma propagation by supporting maintenance of brain tumor-initiating cells (**Dali et al., 2018**). Since the expressions of caspase family members were significantly altered in response to change of FOXG1 expression, it has been suggested that FOXG1 also contributes to carcinogenesis as a negative regulator of glioma cell apoptosis (**Chen et al., 2018a**).

Forkhead box protein P2, encoded by *FOXP2* gene

Forkhead box protein P2 (FOXP2) is a transcriptional repressor

The role of FOXP2 in cancer is somewhat controversial, it appears to have oncogenic or tumor suppressor roles, depending on the cellular and histological features of tumors. While FOXP2 has been found to be down-regulated in breast cancer, hepatocellular carcinoma and gastric cancer biopsies, overexpressed FOXP2 has been reported in multiple myelomas, several subtypes of lymphomas, as well as in neuroblastomas and some prostate cancers (**Herrero and Gitton, 2018**).

Numerous recent studies indicate a tumor suppressor like role for FOXP2 (**Campbell et al., 2010**; **Cuiffo et al., 2014**; **Yan et al., 2015**; **Diao et al., 2018**; **Song et al., 2017**; **Chen et al., 2018b**;

Li et al., 2019), others present evidence for an OG-like role of the protein (*Campbell et al., 2010*; *Zhong et al., 2017*; *Wu et al., 2018*; *Wang et al., 2019*).

The high silent to missense ratio of substitution mutations observed in the case of the *FOXP2* gene does not seem to be a reflection of purifying selection, that might be in harmony of an OG-like role, but definitely not with a tumor suppressor role.

The translated region of the *FOXP2* gene contains a long stretch of CAG repeats (residiues 177–216), corresponding to the polyQ segment of the protein. Silent mutations are clustered in the polyQ tract of the protein encoded by the imperfect polymorphic region, suggesting that the increased silent to missense rate of substitutions in this gene has much less to do with purifying selection than with microsatellite instability.

Glucose-6-phosphate 1-dehydrogenase, encoded by the *G6PD* gene

Glucose-6-phosphate 1-dehydrogenase catalyzes the rate-limiting step of the oxidative pentose-phosphate pathway, its main function is to provide reducing power (NADPH) and pentose phosphates for fatty acid and nucleic acid synthesis. There is strong support for the importance of G6PD for tumor growth. Progression of tumor cells to more aggressive phenotypes requires not only the upregulation of glygolysis but also the pentose phosphate pathway as a provider of reducing power and ribose phosphate to the cell for maintenance of redox balance and biosynthesis of nucleotides and lipids, making G6PD a promising target in cancer therapy (*Zhang et al., 2014a*).

The key importance of G6PD for tumor growth is supported by the fact that elevated G6PD levels promote cancer progression in numerous tumor types, that high G6PD expression is a poor prognostic factor and that knockdown of G6PD suppresses cell viability and growth (*Wang et al., 2012*; *Pu et al., 2015*; *Wang et al., 2015b*; *Poulain et al., 2017*; *Chen et al., 2018c*; *Yang et al., 2018*; *Barajas et al., 2018*; *Yang et al., 2019*).

Mitogen-activated protein kinase 13, encoded by the *MAPK13* gene

MAPK13 (p38δ mitogen-activated protein kinase) is a serine/threonine kinase which acts as an essential component of the MAP kinase signal transduction pathway. MAPK13 plays an important role in the cascades of cellular responses evoked by extracellular stimuli such as proinflammatory cytokines. The protein is involved in the regulation of epidermal keratinocyte differentiation, apoptosis, and skin tumor development.

Although MAPK13 shows signatures of negative selection that would suggest a pro-oncogenioc role for the protein, expermintal data are controversial as to its role in carcinogenesis: there is evidence for both a pro-oncogenic and tumor suppressor roles of MAPK13.

The observation that p38delta promotes cell proliferation and tumor development in epidermis suggests that it has a pro-oncogenic role (*Schindler et al., 2009*). Analyzes of the gene expression profiles have shown that MAPK13 is expressed in uterine, ovary, stomach, colon, liver, and kidney cancer tissues at higher levels compared with adjacent normal tissues. *MAPK13* gene knockdown has been shown to abrogate the tumor-initiating ability of cancer stem-like cells, indicating that the gene has a cancer-promoting role (*Yasuda et al., 2016*). The protein p38δ is highly expressed in all types of human breast cancers, whereas lack of p38δ resulted in reduced primary tumor size and blocked the metastatic potential to the lungs (*Wada et al., 2017*). The fact that mice with germline deletion of the p38δ gene are significantly protected from chemical skin carcinogenesis also suggests a cancer promoting role for the protein (*Kiss et al., 2016*). Interestingly, cell-selective targeted ablation of p38δ in keratinocytes and in immune (myeloid) cells on skin tumor development had different effects. Conditional keratinocyte-specific p38δ ablation reduced malignant progression in males and females relative to their wild-type counterparts. In contrast, conditional myeloid cell-specific p38δ deletion inhibited skin tumorigenesis in male but not female mice. These results reveal that cell-specific p38δ targeting modifies susceptibility to skin carcinogenesis in a context-, stage-, and sex-specific manner (*Kiss et al., 2019*).

The closely related MAPK14, MAPK12, and MAPK13 proteins are known to modulate the immune response, and since chronic inflammation is a known risk factor for tumorigenesis it seems possible that the role of MAPK13 in carcinogenesis may be associated with inflammation. *Del Reino et al., 2014* have analyzed the role of MAPK12 and MAPK13 in colon cancer associated to colitis and have

shown that the deficiency of MAPK12 and MAPK13 significantly decreased tumor formation, in parallel with a decrease in proinflammatory cytokine and chemokine production.

In contrast with the observations arguing for a pro-oncogenic role of the protein, loss of p38δ mitogen-activated protein kinase expression has been shown to promote oesophageal squamous cell carcinoma proliferation, migration, and anchorage-independent growth, suggesting that it has a tumor suppressor role (**O'Callaghan et al., 2013**). Similarly, inactivation of the gene in lung cancer cells has been shown to lead to upregulation of the stemness proteins, thus promoting the cancer stem cell properties of these cells (**Fang et al., 2017**). Promoter methylation of *MAPK13* was found to be present in the majority of primary and metastatic melanomas. Restoration of MAPK13 expression in melanoma cells exhibiting epigenetic silencing of this gene reduced proliferation, indicative of tumor suppressive functions for the protein (**Gao et al., 2013**).

In summary, although MAPK13 plays both pro-oncogenic and tumor suppressor functions in different cellular processes our observation that during tumor evolution negative selection dominates for MAPK13 suggests that the selection pressure to preserve the tumor promoting activities of MAPK13 activity overrides the pressure to eliminate its tumor suppressor activities.

Protein AF-9, encoded by the *MLLT3* gene

The *MLLT3* gene (present in CGC list of cancer genes) has been selected as a gene showing very high values of rSMN, suggesting negative selection of missense and nonsense mutations (**Supplementary file 3**).

It must be pointed out that based on the high silent/missense ratio *MLLT3* (as well as *TBP* and *DSPP*) has also been identified by others as a gene subject to negative selection (**Zhou et al., 2017**).

Protein AF-9 is a component of a complex required to increase the catalytic rate of RNA polymerase II transcription by suppressing transient pausing by the polymerase at multiple sites along the DNA.

Several studies indicate that *MLLT3* is a proto-oncogene, its inactivation or downregulation suppresses lymphoma cell proliferation, invasion and inhibits metastasis and proliferation of prostate cancer (**Zhang et al., 2012**; **Meng et al., 2017**).

Despite the tumor promoting role of *MLLT3*, the high silent to missense ratio of substitution mutations does not seem to be a reflection of strong negative selection. The translated region of the *MLLT3* gene contains a long stretch of AGC repeats (encoding the polyS segment of the protein, residues 149–194). The 'excess' of silent mutations are clustered in the polyS tract of the protein encoded by the imperfect polymorphic AGC microsatellite region of the *MLLT3* gene, that is known to be highly unstable (**Walker et al., 1994**).

Neuro-oncological ventral antigen 1 (Nova-1), encoded by the *NOVA1* gene

Nova-1 is an RNA-binding protein involved in the regulation of RNA splicing.

The importance of Nova1 for tumor growth is supported by the observation that overexpressed intratumoral NOVA1 was associated with poor survival rate and increased recurrence rate of hepatocellular carcinoma (HCC) and was an independent prognostic factor for overall survival rate and tumor recurrence. HCC cell lines over-expressing NOVA1 exhibited greater potentials in cell proliferation, invasion, and migration, while knockdown of NOVA1 had the opposite effects. All these findings indicate that NOVA1 may act as a prognostic marker for poor outcome and high recurrence in HCC (**Zhang et al., 2014b**).

Similarly, NOVA1 expression was found to be upregulated in melanoma samples and cell lines. and knockdown of NOVA1 suppressed melanoma cell proliferation, migration and invasion in both A375 and A875 cell lines These results suggested that NOVA1 acted as an OG in the development of melanoma (**Yu et al., 2018**).

Recent studies have shown that the tumor suppressor microRNA-592 suppresses the malignant phenotypes of thyroid cancer by downregulating NOVA1. Whereas overexpression of miR–592 resulted in decreased cell proliferation, migration, and invasion in thyroid cancer, ectopic NOVA1 expression effectively abolished the tumor-suppressing effects of miR–592 overexpression in thyroid cancer cells vitro and in vivo (**Luo et al., 2019**).

Recent studies have provided an explanation for the role of NOVA1 in carcinogenesis. **Sayed et al., 2019** have shown that NOVA1 as well as the polypyrimidine-tract binding protein PTBP1 acts as enhancers of full-length TERT splicing, increasing telomerase activity, promoting telomere maintenance in cancer cells, thereby favoring theit replicative immortality.

Calcium/calmodulin-dependent protein kinase type 1B, encoded by the *PNCK* gene

Pregnancy upregulated non-ubiquitous calmodulin kinase PNCK is a calcium/calmodulin-dependent protein kinase belonging to a calcium-triggered signaling cascade. It phosphorylates and activates CAMK1 that, upon calcium influx, regulates transcription activators activity, cell cycle, hormone production, and cell differentiation.

Several lines of evidence suggest that PNCK promotes carcinogenesis.

PNCK has been found to be highly overexpressed in human primary human breast cancers compared with benign mammary tissue (**Gardner et al., 2000**). Increased expression of PNCK is associated with poor prognosis in clear cell renal cell carcinoma. The mRNA level of PNCK was significantly higher in tumorous tissues than in the adjacent non-tumorous tissues. Multivariate analysis indicated that PNCK expression was an independent predictor for poor survival of clear cell renal cell carcinoma patients (**Wu et al., 2013**). Overexpression of PNCK in breast cancer cells was shown to result in increased proliferation, clonal growth, and cell-cycle progression (**Deb et al., 2015**).

Recent studies have shown that *PNCK* depletion inhibits proliferation and induces apoptosis of human nasopharyngeal carcinoma cells in vitro and in vivo, suggesting that it might be a novel therapeutic target for treatment of nasopharyngeal carcinoma (**Xu et al., 2019**).

Runt-related transcription factor 2, encoded by the *RUNX2* gene

The protein is a member of the RUNX family of transcription factors and has a Runt DNA-binding domain. RUNX2 is a transcription factor involved in osteoblastic differentiation and skeletal morphogenesis. RUNX2 plays a cell proliferation regulatory role in cell cycle entry and exit in osteoblasts. These functions are especially important when discussing bone cancer, particularly osteosarcoma development that can be attributed to aberrant cell proliferation control.

Several studies indicate that RUNX2 plays a key role in carcinogenseis. RUNX2 overexpression was found to promote aggressiveness and metastatic spreading, whereas *RUNX2* knockdown inhibits tumor growth and metastasis suggesting an oncogenic role for the protein (**Tandon et al., 2014**; **Tandon et al., 2016**; **Shin et al., 2016**; **Li et al., 2016**; **Wang et al., 2016**; **Sancisi et al., 2017**; **Lu et al., 2018**; **Ji et al., 2019**; **Herreño et al., 2019**).

Although strong purifying selection would not contradict the tumor promoting role of RUNX2, the high silent to missense ratio of substitution mutations is not a reflection of the strength of negative selection of missense and nonsense substitutions.

A noteworthy feature of the *RUNX2* gene is that its translated region contains a long stretch of CAG repeats (encoding the polyQ segment of the protein, residues 49–71). Interestingly, substitutions are not randomly distributed along the sequence of *RUNX2*: they are clustered in the polyQ tract of the protein encoded by the imperfect polymorphic CAG microsatellite region of the *RUNX2* gene. Since in cancer cells defective in mismatch-repair, microsatellites are known to become unstable due to increased frequency of replication error (**Benachenhou et al., 1998**), it seems likely that this increases and distorts mutation pattern in the polyQ region of *RUNX2*, and this mutation hotspot may give the false impression of strong purifying selection.

Monocarboxylate transporter 4 (MCT 4), encoded by the *SLC16A3* gene

Monocarboxylate transporter 4 (MCT4) or Solute carrier family 16 member 3 (SLC16A3) is a member of the proton-linked monocarboxylate transporter. Protein family. It catalyzes the rapid transport across the plasma membrane of many monocarboxylates such as lactate.

Since due to abnormal conversion of pyruvic acid to lactic acid by tumor cells even under normoxia, the altered metabolism of glucose consuming tumors must rapidly efflux lactic acid to the microenvironment to maintain a robust glycolytic flux and to prevent poisoning themselves

(**Mathupala et al., 2007**). Survival and maintenance of the glycolytic phenotype of tumor cells is ensured by monocarboxylate transporter 4 (MCT4, encoded by the *SLC16A3* gene) that efficiently transports L-lactate out of the cell (**Ganapathy et al., 2009**).

As high metabolic and proliferative rates in cancer cells lead to production of large amounts of lactate, extruding transporters are essential for the survival of cancer cells. This point may be illustrated by the fact that knockdown of MCT4 increased tumor-free survival and decreased in vitro proliferation rate of tumor cells (**Andersen et al., 2018**).

Using a functional screen **Baenke et al., 2015** have also demonstrated that monocarboxylate transporter four is an important regulator of breast cancer cell survival: MCT4 depletion reduced the ability of breast cancer cells to grow, suggesting that it might be a valuable therapeutic target.

In harmony with the essentiality of MCT4 for tumor growth, several studies indicate that expression of the hypoxia-inducible monocarboxylate transporter MCT4 is increased in tumors and its expression correlates with clinical outcome, thus it may serve as a valuable prognostic factor (**Witkiewicz et al., 2012**; **Doyen et al., 2014**; **Baek et al., 2014**)

Consistent with the key importance of MCT4 for the survival of tumor cells, its selective inhibition to block lactic acid efflux appears to be a promising therapeutic strategy against highly glycolytic malignant tumors (**Todenhöfer et al., 2018**; **Choi et al., 2016**; **Choi et al., 2018**; **Zhao et al., 2019b**)

Solute carrier family 2, facilitated glucose transporter member 1, encoded by the *SLC2A1* gene

SLC2A1 functions as a facilitative glucose transporter, which is responsible for glucose uptake.

Significantly, several nutrient transporter protein genes were found among the genes showing the strongest signs of purifying selection. The most likely explanation for the selective pressure to preserve their integrity is that tumor cells have an increased demand for nutrients and this demand is met by enhanced cellular entry of nutrients through upregulation of specific transporters (**Ganapathy et al., 2009**).

The uncontrolled cell proliferation of tumor cells involves not only deregulated control of cell proliferation but also major adjustments of energy metabolism in order to fuel cell growth and division in the hypoxic microenvironments in which they reside. Otto Warburg was the first to observe an anomalous characteristic of cancer cell energy metabolism: even in the presence of oxygen, cancer cells limit their energy metabolism largely to glycolysis, leading to a state that has been termed 'aerobic glycolysis' (**Warburg, 1956b**). Cancer cells are known to compensate for the lower efficiency of ATP production through glycolysis than oxidative phosphorylation by upregulating glucose transporters, such as GLUT1, thus increasing glucose import into the cytoplasm (**Jones and Thompson, 2009**; **DeBerardinis et al., 2008**; **Hsu and Sabatini, 2008**).

The markedly increased uptake of glucose has been documented in many human tumor types, by noninvasively visualizing glucose uptake through positron emission tomography using a radiolabeled analog of glucose as a reporter. This reliance of tumor cells on glycolysis is also supported by the hypoxia response system: under hypoxic conditions not only glucose transporters but also multiple enzymes of the glycolytic pathway are upregulated (**Jones and Thompson, 2009**; **DeBerardinis et al., 2008**; **Semenza, 2010a**; **Semenza, 2010b**; **Kroemer and Pouyssegur, 2008**)

In our view, the central role of GLUT1 in cancer metabolism is reflected by the fact that the gene (*SLC2A1* gene of solute carrier family member two protein) encoding this glucose transporter is among the genes that show the strongest signatures of purifying selection (see **Supplementary file 31**).

The key importance of GLUT1 in cancer may be illustrated by the fact that high levels of GLUT1 expression correlates with a poor overall survival and is associated with increased malignant potential, invasiveness, and poor prognosis (**Wang et al., 2017a**; **Deng et al., 2018**; **de Castro et al., 2019**).

The strict requirement for GLUT1 in the early stages of mammary tumorigenesis highlights the potential for glucose restriction as a breast cancer preventive strategy (**Wellberg et al., 2016**). The tumor essentiality of GLUT1 may also be illustrated by the fact that knockdown of GLUT1 inhibits cell glycolysis and proliferation and inhibits the growth of tumors (**Xiao et al., 2018**). In view of its

essentiality for tumor growth, GLUT1 is a promising target for cancer therapy (**Shibuya et al., 2015**; **Noguchi et al., 2016**; **Chen et al., 2017d**)

Recent studies suggest that the YAP1-TEAD1-GLUT1 axis plays a major role in reprogramming of cancer energy metabolism by modulating glycolysis (**Lin and Xu, 2017**). These authors have shown that YAP1 and TEAD1 are involved in transcriptional control of the the glucose transporter GLUT1, whereas knockdown of YAP1 inhibited glucose consumption, and lactate production of breast cancer cells. Overexpression of GLUT1 restored glucose consumption and lactate production.

Solute carrier family 2, facilitated glucose transporter member 8, encoded by the *SLC2A8* gene

The SLC2A8/GLUT8 is a member of the glucose transporter superfamily that mediates the transport of glucose and fructose.

In harmony with the strong signatures of negative selection, there is evidence that GLUT8 plays an important role in carcinogenesis: it is overexpressed in and is required for proliferation and viability of tumors (**Goldman et al., 2006**; **McBrayer et al., 2012**).

TATA-box-binding protein, encoded by the *TBP* gene

The *TBP* gene has been selected as a gene showing very high values of rSMN, suggesting negative selection of missense and nonsense mutations (**Supplementary file 3**). It must be pointed out that based on the high silent/missense ratio *TBP* (as well as *DSPP* and *MLLT3*) has also been identified by others as a gene subject to negative selection (**Zhou et al., 2017**).

The protein is a general transcription factor that functions at the core of the DNA-binding multiprotein factor TFIID. Binding of TFIID to the TATA box is the initial transcriptional step of the pre-initiation complex, playing a role in the activation of eukaryotic genes transcribed by RNA polymerase II. In view of such a basic cell essential function, it seemed justified to assume that it is the indispensability of the gene for the survival of tumor cells (just like any other cell) that subjects it to strong purifying selection and the high silent/missense ratio is a reflection of this negative selection. TBP has been thought to be an invariant housekeeping protein, however, several studies have shown that TBP expression is significantly increased in both colon adenocarcinomas as well as adenomas relative to normal tissue, supporting the idea that increases in TBP expression actually drive tumorigenesis (**Johnson et al., 2003a**; **Johnson et al., 2003b**; **Johnson et al., 2017**).

Inspection of the spectrum of somatic mutations of the *TBP* gene suggests that the high silent/missense ratio is unlikely to be simply due to negative selection that may hold for both OGs and TEGs. A noteworthy feature of the *TBP* gene is that its translated region contains a long stretch of CAG repeats (encoding the polyQ segment of the protein, residues 57–95). The distribution of silent mutations is markedly non-random: they are clustered in the polyQ tract of the protein encoded by the imperfect polymorphic CAG microsatellite region of the *TBP* gene. Since in cancer cells defective in mismatch-repair, microsatellites are known to become unstable due to increased frequency of replication error (**Benachenhou et al., 1998**), it seems likely that this is why the rate of mutation in the polyQ region of TBP is much higher than in other regions of the gene. The high silent to missense rate is thus not due to negative selection acting on missense and nonsense substitutions. Rather, it may reflect the fact that the imperfect polymorphic CAG microsatellite region of the *TBP* gene serves as a mutation hotspot, with a biased substitution pattern.

Thromboxane A2 receptor, encoded by the *TBXA2R* gene

TBXA2R is a plasma membrane protein that serves as a receptor for thromboxane A2, a potent stimulator of platelet aggregation. The activity of this receptor is mediated by a G-protein that activates a phosphatidylinositol-calcium second messenger system.

Studies on the expression of thromboxane A2 receptor, TBXA2R in a cohort of human breast cancer patients revealed that breast tumor tissues expressed higher levels of TBXA2R compared with normal mammary tissues and that TBXA2R expression was most significantly increased in grade three tumors. Kaplan-Meier survival analysis has also shown that patients with high levels of TBXA2R had significantly shorter disease-free survival. The observation that TBXA2R is highly expressed in

aggressive tumors and linked with poor prognosis indicates that TBXA2R has a significant prognostic value in clinical breast cancer (**Watkins et al., 2005**).

The role of TBXA2R in carcinogenesis is also supported by the observation that Thromboxane A2 was shown to enhance tumor metastasis and that the tumor promoting activity required intact TBXA2 receptor (**Matsui et al., 2012**). These studies revealed that TBXA2-TBXA2R signaling plays a critical role in tumor colonization through P-selectin-mediated interactions between platelets-tumor cells and tumor cells-endothelial cells, suggesting that blockade of this signaling might be useful in the treatment of tumor metastasis.

Although the involvement of TBXA2-TBXA2R signaling in cancer invasion and metastasis apperars to be clearly established, there may be other mechanisms by which TBXA2 promotes these processes. **Li and Tai, 2013** have shown that a TBXA2 mimetic induced the expression of the monocyte chemoattractant chemokine ligand protein CCL2, suggesting that TBXA2 may also stimulate invasion of cancer cells through CCL2-CCR2 mediated macrophage recruitment.

Recent studies on Triple Negative Breast Cancer (TNBC) cell lines revealed that TBXA2R expression was higher in these cell lines and that *TBXA2R* knockdowns consistently showed dramatic cell killing in TNBC cells (**Orr et al., 2016**). It has also been shown that TBXA2R enhanced TNBC cell migration, invasion, indicating that the gene is required for the survival and migratory behavior of a subset of TNBCs.

A phenome-wide association study has shown that a single-nucleotide polymorphism in the gene *TBXA2R* is associated with increased metastasis in multiple primary cancers, suggesting the requirements for thromboxane A2 (TXA2) and TBXA2R in the basic mechanism of metastasis, and the clinical applicability of TBXA2R antagonists as adjuvant therapy in multiple cancers (**Pulley et al., 2018**).

Tumor protein p73, encoded by the *TP73* gene

The protein is known to participate in the apoptotic response to DNA damage: isoforms containing the N-terminal transactivation domain are pro-apoptotic, isoforms lacking the transactivation domain are anti-apoptotic.

Although p73 shows substantial homology with p53, despite the established role of p53 as a tumor suppressor, p73 does not have a similar tumor suppressor role in malignancy: unlike p53-/- mice, p73 knockout mice do not develop tumors. In fact, N-terminally truncated p73 isoforms, lacking the transactivation domain were shown to possess oncogenic potential (**Stiewe and Pützer, 2002**; **Stiewe et al., 2002**).

Numerous studies have shown that ΔNp73, the oncogenic isoform of p73 lacking the transactivation domain, is frequently upregulated in many carcinomas and is indicative of poor prognosis (**Zaika et al., 2002**; **Petrenko et al., 2003**; **Domínguez et al., 2006**; **Hassan et al., 2014b**; **Hassan et al., 2014a**; **Lucena-Araujo et al., 2015**).

Our observation that p73, an oncogenic protein, shows only strong signatures of purifying selection provides one of the clearest examples illustrating the point that in the case of OGs purifying selection is not necessarily associated with positive selection for driver mutations. It must be pointed out here that it has been noted earlier by others that, despite its clear role in carcinogenesis, the *TP73* gene is almost never mutated (**Bisso et al., 2011**; **Maas et al., 2013**). One may argue that in this case the molecular change that drives carcinogenesis is the change of splicing that favors the formation of the oncogenic isoform of p73.

Tribbles homolog 2, encoded by the *TRIB2* gene

TRIB2 is a pseudokinase member of the pseudoenzyme class of signaling/scaffold proteins. It interacts with MAPK kinases and regulates activation of MAP kinases.

TRIB2 has been shown to be important in the maintenance of the oncogenic properties of melanoma cells, as its silencing reduces cell proliferation, colony formation. Tumor growth was also substantially reduced upon RNAi-mediated TRIB2 knockdown in an in vivo melanoma xenograft model, suggesting that TRIB2 provides the melanoma cells with growth and survival advantages (**Zanella et al., 2010**).

TRIB2 expression is elevated in primary human lung tumors and in NSCLC cells, resulting from gene amplification. *TRIB2* knockdown was found to inhibit cell proliferation and in vivo tumor growth, indicating that TRIB2 is a potential driver of lung tumorigenesis (**Grandinetti et al., 2011**).

High TRIB2 expression is observed in T cell acute lymphoblastic leukemias (**Hannon et al., 2012**). TRIB2 has been shown to be critical for both solid and non-solid malignancies and is functionally important for liver cancer cell survival and transformation. TRIB2 was found to be upregulated in liver cancer cells compared with other cells (**Wang et al., 2013a**; **Wang et al., 2013b**).

TRIB2 is emerging as a pivotal target of transcription factors in acute leukemias as evidenced by the fact *TRIB2* knockdown resulted in a block in acute myeloid leukemia cell proliferation (**Rishi et al., 2014**)

In the case of lung adenocarcinoma, patients with higher TRIB2 levels had poorer survival (**Zhang et al., 2016**). The tumor-promoting role of this protein is supported by the observation that TRIB2 expression is significantly increased in tumor tissues from patients with extremely poor clinical outcome (**Hill et al., 2017**; **Wang et al., 2020**).

TRIB2 has been shown to be important for the survival of leukemia cells during MLL-TET1-related leukemogenesis and for maintaining differentiation blockade of leukemic cells: *TRIB2* knockdown relieved the inhibition of myeloid cell differentiation induced by the MLL-TET1 fusion protein (**Kim et al., 2018**).

TRIB2 expression has been shown to be elevated in colorectal cancer tissues compared to normal adjacent tissues and high TRIB2 expression indicated poor prognosis of colorectal cancer patients (**Hou et al., 2018**). Depletion of TRIB2 inhibited cancer cell proliferation, induced cell cycle arrest and promoted cellular senescence, whereas overexpression of TRIB2 accelerated cell growth, cell cycle progression and blocked cellular senescence.

Twist-related protein 1, encoded by the *TWIST1* gene

The *TWIST1* gene is characterized by very high value of rSMN (**Supplementary file 5**), indicating strong signature of purifying selection, suggesting that it plays an important role in promoting tumorigenesis.

Twist-related protein 1, TWIST1 is a transcription factor and master regulator of the epithelial-to-mesenchymal transition that significantly contributes to tumor growth and metastasis. TWIST1 is overexpressed in a variety of tumors and numerous studies have shown that targeting TWIST1 significantly inhibits tumor growth (**Wushou et al., 2014**; **Zhu et al., 2016**; **Xu et al., 2017a**; **Xu et al., 2017b**; **Mikheev et al., 2018**).

Recent studies have revealed that AURKA and TWIST1 are linked in as much as ablation of either AURKA or TWIST1 completely inhibits epithelial-to-mesenchymal transition (**Wang et al., 2017b**).

Appendix 2

Analyses of somatic substitutions and subtle indel mutations of human protein-coding genes of tumor tissues

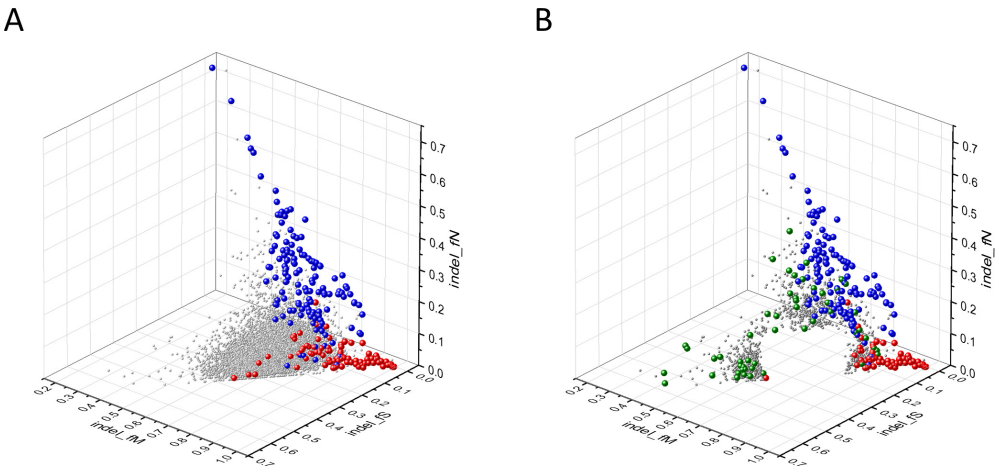
We have used two major types of analyses of silent, amino acid changing and truncating somatic mutations of human protein-coding genes of tumor tissues: one in which we have restricted our analyses to single -nucleotide substitutions (SO or' substitution only' analyes, for details, see main text).

Here, we describe the analyes that also take into account subtle indels (SSI or' substitutions and subtle indels' analyses). In these analyses, subtle mutations affecting the coding sequences of protein coding genes were assigned to three categories: SIL, silent synonymous substitutions, MIS, merging nonsynonymous substitutions and short inframe indels that alter but do not disrupt coding sequence, and NON, merging nonsense substitutions and short frame-shift indels as both types of mutations lead eventually to stop codons that truncate the protein. Unless otherwise idicated, we have used datasets containing transcripts with at least 100 confirmed somatic, non-polymorphic mutations identified in tumor tisses.

We have used several approaches to analyze the contribution of silent, amino acid changing and truncating mutations to somatic mutations of human protein-coding genes during tumor evolution.

In the simplest case, we have calculated for each transcript the fraction of somatic mutations that could be assigned to the synonymous (indel_fS), nonsynonymous (indel_fM), and nonsense mutation (indel_fN) category.

Our analyses have shown that in the 3D representation of SSI mutations (see ***Appendix 2—figure 1***, Panel A) genes are present in a cluster characterized by fraction values of 0.24082 ± 0.06203, 0.70086 ± 0.05701 and 0.05832 ± 0.04151 for indel_fS, indel_fM and indel_fN category, respectively. The mean values for indel_fS, indel_fM and indel_fN in this cluster are very similar to those observed for fS, fM, and fN in SO analyses (***Supplementary file 29***), consistent with the observation that in the dataset containing transcripts with at least 100 confirmed somatic, non-polymorphic mutations identified in tumor tissues subtle indels are much rarer than single-nucleotide substitutions (***Supplementary file 8***).



Appendix 2—figure 1. Analyses of indel_fS, indel_fM and indel_fN parameters of human protein-coding genes of tumor tissues. The figure shows the results of the analysis of 13,930 transcripts containing at least 100 subtle, confirmed somatic non-polymorphic mutations from tumor tissues. Axes *x*, *y* and *z* represent the fractions of somatic mutations that are assigned to the indel_fS, indel_fM and indel_fN categories. In Panel A, each ball represents a human transcript; note that the majority of human genes are present in a dense cluster. The positions of transcripts of the genes defined by ***Vogelstein et al., 2013*** as oncogenes (OGs, large red balls) or tumor suppressor genes (TSGs, large blue balls) are highlighted. It is noteworthy that these driver genes separate significantly from the central cluster and from each other: OGs have an increased fraction of indel_fM, whereas TSGs have markedly increased fraction of indel_fN. Panel B shows data only for

Appendix 2—figure 1 continued on next page

Appendix 2—figure 1 continued

candidate cancer genes present in the CG_SO^{2SD}_SSI^{2SD} list (see ***Supplementary file 31***). The positions of transcripts of the genes identified by ***Vogelstein et al., 2013*** as OGs (large red balls) or TSGs (large blue balls) are highlighted. The positions of novel cancer gene transcripts validated in the present work are highlighted as large green balls.

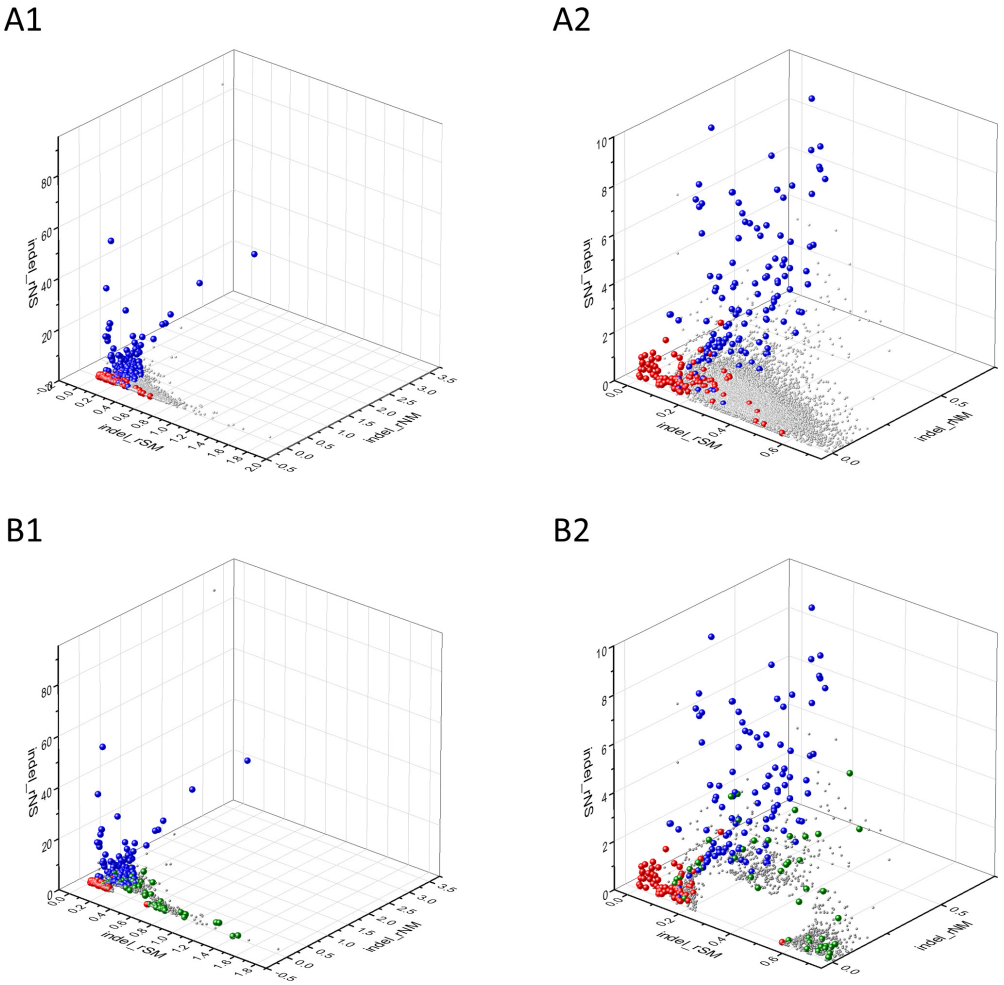
It is noteworthy, however, that the pattern of indel_fS, indel_fM and indel_fN of the best known cancer genes (***Vogelstein et al., 2013***) deviates significantly from that characteristic of the majority of human genes (see ***Appendix 2—figure 1***, Panel A). The values for OGs show a marked incease in indel_fM, reflecting positive selection for missense mutations, whereas the values for TSGs show significant increase in indel_fN, reflecting primarily positive selection for truncating nonsense mutations (***Supplementary file 29***).

The set of genes (6139 transcripts) with values that deviate from mean values of indel_fS, indel_fM and indel_fN by more than 1SD have also included the majority of OGs and TSGs (only 5 OG and 1 TSG transcripts remained in the central cluster). It is noteworthy that the 6139 transcripts also contained the vast majority (443 out of 748) of the transcripts of CGC genes, suggesting that the mutation pattern of most CGC genes also deviates significantly from that of passenger genes (PGs; ***Supplementary file 29***). The genes in the central cluster (***Supplementary file 29***) is hereafter referred to as PG_SSI^{f-1SD} (for Passenger Gene_Substitution and Subtle Indels deviating from mean indel_fS, indel_fM and indel_fN values by $\leq$1 SD).

The set of genes (1211 transcripts) with values that deviate from mean values of indel_fS, indel_fM and indel_fN by more than 2SD included 62 OG and 123 TSG driver gene transcripts (see ***Appendix 2—figure 1***, Panel B). Using this more stringent cut-off value, the number of CGC genes identified in the 1211 transcripts was reduced to 153 out of 748 (***Supplementary file 29***). The non-PG set defined by 2SD cut-off value is hereafter referred to as CG_SSI^{f-2SD} for Cancer Gene_ Substitution and Subtle Indels deviating from mean indel_fS, indel_fM and indel_fN values by more than 2SD (***Supplementary file 29***).

The 1211 transcripts in the gene set of CG_SSI^{f-2SD} has 873 transcripts not found in the OG, TSG, and CGC cancer gene lists (***Supplementary file 5***). Since the majority of these 873 transcripts (derived from 743 genes) have parameters that assign them to the OG or TSG clusters, we assume that they also qualify as candidate OGs or TSGs. There is, however, a third group of genes that deviate from both the central PG cluster and the clusters of OGs and TSGs: their high indel_fS and low indel_fM and indel_fN values suggest that they experience purifying selection during tumor evolution, suggesting that they may correspond to TEGs important for the growth and survival of tumors. The 743 putative cancer genes listed in CG_SSI^{f-2SD} of ***Supplementary file 5***, were subjected to further analyses to decide whether they qualify as candidate OGs, TSGs, TEGs or the deviation of their mutation pattern from those of PGs is not the result of natural selection. For some typical examples of these analyses see Appendix 1.

Known cancer genes (OGs and TSGs) also separate from the majority of human genes in 3D representations of parameters indel_rSM, indel_rNM, indel_rNS defined as the ratio of indel_fS/indel_fM, indel_fN/indel_fM, indel_fN/indel_fS, respectively (see ***Appendix 2—figure 2***). In these representations (see ***Appendix 2—figure 2***, Panels A1, A2), OGs separate from the central cluster in having significantly lower indel_rSM and indel_rNM values, whereas TSGs had significantly higher indel_rNS and indel_rNM values than those of the central cluster.



Appendix 2—figure 2. Analyses of indel_rSM, indel_rNM, indel_rNS parameters of human protein-coding genes of tumor tissues. The figure shows the results of the analysis of 13930 transcripts containing at least 100 subtle, confirmed somatic mutations from tumor tissues, including only mutations identified as not SNPs. Axes *x*, *y*, and *z* represent the indel_rSM, indel_rNM, indel_rNS values defined as the ratio of indel_fS/ indel_fM, indel_fN/ indel_fM, indel_fN/ indel_fS, respectively. Each ball represents a human transcript; the positions of transcripts of the genes identified by ***Vogelstein et al., 2013*** as oncogenes (OGs, large red balls) or tumor suppressor genes (TSGs, large blue balls) are highlighted. Panels A1, A2 show the distribution of the 13,930 transcripts at different magnification. Note that the majority of human genes are present in a dense cluster but known OGs and TSGs separate significantly from the central cluster and from each other. The rNS and rNM values of TSGs are higher, whereas the rSM and rNM values of OGs are lower than those of passenger genes. Panels B1,B2 show data only for candidate cancer genes present in the CG_SO^{2SD}_SSI^{2SD} list (see ***Supplementary file 31***). The positions of transcripts of the genes identified by ***Vogelstein et al., 2013*** as OGs (large red balls) or TSGs (large blue balls) are highlighted. The positions of novel cancer gene transcripts validated in the present work are highlighted as large green balls.

The set of genes (4518 transcripts) with values that deviate from the mean by more than 1SD contained 78 OG transcripts, 132 TSG transcripts, and 368 CGC gene transcripts (***Supplementary file 29***). The central cluster of genes (that deviate from mean rSM, rNM and rNS values by $\leq$1 SD) is hereafter referred to as PG_SSI^{r2_1SD} (for Passenger Gene_Substitution and Subtle Indels deviating from mean indel_rSM, indel_rNM, indel_rNS values by $\leq$1 SD).

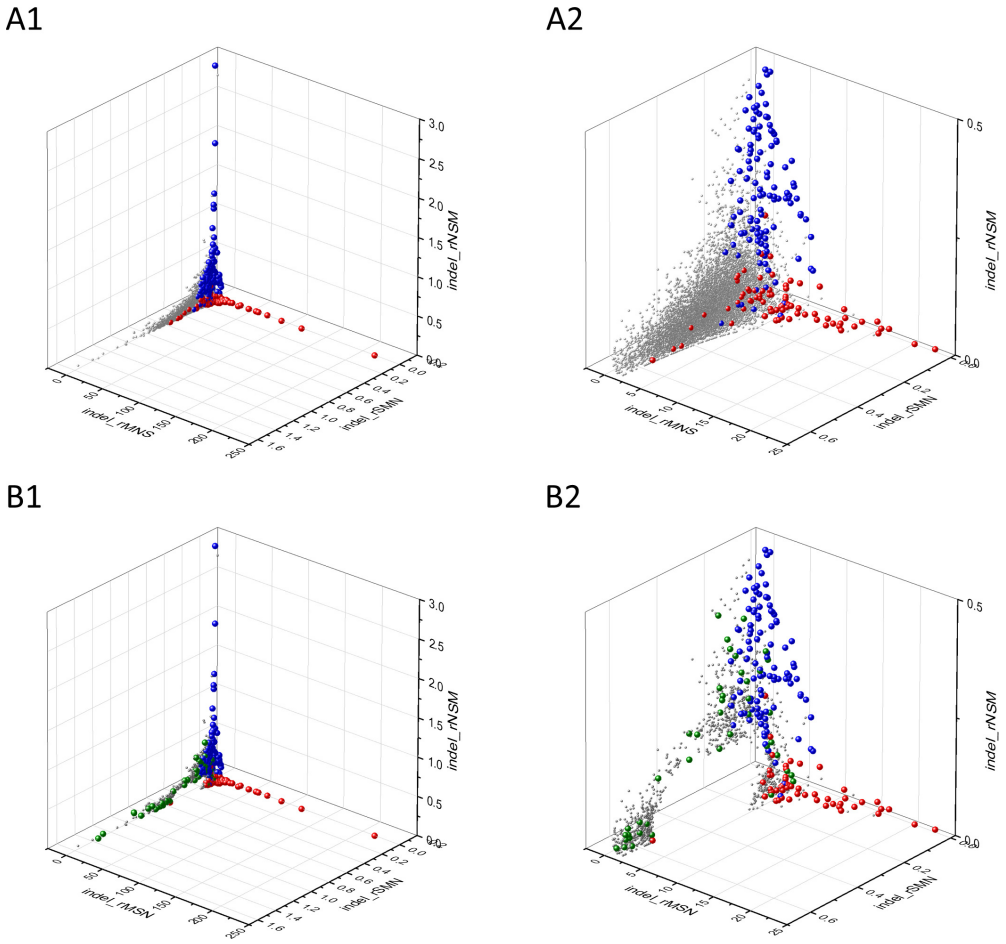
The non-PG set defined by 2SD cut-off value (see ***Appendix 2—figure 2***, B1, B2, ***Supplementary file 5***) is hereafter referred to as CG_SSI^{r2_2SD} for Cancer Gene_ Substitution and Subtle Indels deviating from mean indel_rSM, indel_rNM, indel_rNS values by more than 2SD

(***Supplementary file 29***). This gene set has a total of 861 transcripts, containing 40 transcripts of OGs, 98 transcripts of TSGs genes, 86 transcripts of CGC genes and 637 transcripts (derived from 546 genes) not found in the OG, TSG, and CGC cancer gene lists (***Supplementary file 5***).

The mean parameters of TSGs differ markedly from those of PGs in that rNS and rNM values are higher, reflecting the dominance of positive selection for inactivating mutations. The parameters for OGs on the other hand, differ from those of PGs in that indel_rSM values of OGs are significantly lower, reflecting positive selection for missense mutations (see ***Appendix 2—figure 2***, Panels A1, A2). Interestingly, in this representation some OGs (e.g. *BCL2*) have unusally high scores of indel_rSM suggesting that in the case of these OGs purifying selection may override positive selection for amino acid changing mutations.

As mentioned above, the non-PG set defined by a cut-off values of 2SD contains 637 transcripts (derived from 546 genes) not found in the OG, TSG, or CGC lists. Since the majority of these genes have paramerters that assign them to the OG or TSG clusters, they can be regarded as candidate OGs or TSGs. There is a group of genes that deviate from the clusters of PGs, OGs, and TSGs (see ***Appendix 2—figure 2***, Panels B1, B2) in that they have unusually high indel_rSM values. Since high indel_rSM values may be indicative of purifying selection we assume that they may correspond to TEGs important for the growth and survival of tumors. The 546 putative cancer genes listed in $CG\_SO^{indel\_r2\_2SD}$ of ***Supplementary file 5***, were subjected to further analyses to decide whether they qualify as candidate OGs, TSGs, TEGs or the deviation of their mutation pattern from those of PGs is not the result of natural selection. For examples of these analyses see Appendix 1.

The separation of known cancer genes from the majority of human genes is also observed in 3D representations of parameters indel_rSMN, indel_rMSN and indel_rNSM defined as the ratio of indel_fS/(indel_fM+indel_fN), indel_fM/(indel_fS+indel_fN) and indel_fN/(indel_fS+indel_fM), respectively (see ***Appendix 2—figure 3***, Panels A1, A2). In this representation the genes are present in a three pronged cluster.



Appendix 2—figure 3. Analyses of indel_rSMN, indel_rMSN and indel_rNSM parameters of human protein-coding genes of tumor tissues. The figure shows the results of the analysis of 13930 transcripts containing at least 100 subtle, confirmed somatic mutations from tumor tissues. Axes *x*, *y*, and *z* represent paramerters indel_rSMN, indel_rMSN and indel_rNSM defined as the ratio of indel_fS/(indel_fM+indel_fN), indel_fM/(indel_fS+indel_fN) and indel_fN/(indel_fS+indel_fM), respectively. Each ball represents a human transcript; the positions of transcripts of the genes defined by ***Vogelstein et al., 2013*** as oncogenes (OGs, red balls) or tumor suppressor genes (TSGs, blue balls) are highlighted. Panels A1 and A2 show the distribution of the 13,930 transcripts at different magnification. Note that the majority of human genes are present in a dense cluster but known OGs and TSGs separate significantly from the central cluster and from each other. The indel_rNSM values of TSGs are higher, their indel_rMSN and indel_rSMN are lower than those of passenger genes. OGs also separate from passenger genes in that their indel_rMSN values are higher and their indel_rSMN values are lower than those of passenger genes. Panels B1,B2 show data at different magnification only for candidate cancer genes present in the CG_SO^{2SD}_SSI^{2SD} list (see ***Supplementary file 31***). The positions of transcripts of the genes identified by ***Vogelstein et al., 2013*** as OGs (large red balls) or TSGs (large blue balls) are highlighted. The positions of novel cancer gene transcripts validated in the present work are highlighted as large green balls.

The set of genes (4369 transcripts) with values that deviate from the mean by more than 1SD contained 78 OG transcripts, 132 TSG transcripts and 354 CGC gene transcripts (***Supplementary file 29***). The central cluster of genes, deviating from mean rSMN, rMSN and rNSM values by $\leq$1 SD is hereafter referred to as PG_SO^{indel_r3_1SD} (for Passenger Gene_ Substitution and Subtle Indels deviating from mean indel_rSMN, indel_rMSN and indel_rNSM values by $\leq$1 SD).

The non-PG set defined by 2SD cut-off value (see ***Appendix 2—figure 3***, Panels B1, B2, ***Supplementary file 5***) is hereafter referred to as CG_SSI^{r3_2SD} for Cancer Gene_ Substitution and

Subtle Indels deviating from mean indel_rSMN, indel_rMSN, and indel_rNSM values by more than 2SD (***Supplementary file 29***). This gene set has a total of 823 transcripts, containing transcripts of 37 OGs, 100 TSGs, 86 CGC genes and 600 transcripts (derived from 510 genes) not found in the OG, TSG, and CGC cancer gene lists (***Supplementary file 5***).

The mean parameters of TSGs differ markedly from those of PGs in as much as indel_rNSM values of TSGs are higher and indel_rSMN values are lower, reflecting the dominance of positive selection for inactivating mutations. In the case of OGs on the other hand, indel_rMSN values are higher and indel_rNSM values are lower than those of PGs, reflecting positive selection for missense mutations and purifying selection avoiding nonsense mutations. Interestingly, some OGs have unusally high scores of indel_rSMN suggesting that in these cases (e.g. *BCL2*) purifying selection may override positive selection for amino acid changing mutations.

As mentioned above, the non-PG set defined by a cut-off value of 2SD contains 600 transcripts (derived from 510 genes) not found in the OG, TSG or CGC lists. Since the majority of these genes have paramerters that assign them to the OG or TSG clusters, they can be regarded as candidate OGs or TSGs.

In this representation, we also note the existence of a group of genes that deviates from the clusters of PGs, OGS, and TSGs (see ***Appendix 2—figure 3***): their high indel_rSMN and low indel_rMSN and indel_rNSM values suggest that they experience purifying selection during tumor evolution, suggesting that they may be essential for the survival of tumors as OGs or TEGs. The 510 putative cancer genes listed in $CG\_SSI^{r3\_2SD}$ of ***Supplementary file 5***, were subjected to further analyses to decide whether they qualify as candidate OGs, TSGs and TEGs or the deviation of their mutation pattern from those of PGs is not the result of natural selection. For some typical examples of these analyses see Appendix 1.