PLOS ONE

RESEARCH ARTICLE

Discovering novel driver mutations from pan-cancer analysis of mutational and gene expression profiles

Houriiyah Tegally[1], **Kevin H. Kensler**[2], **Zahra Mungloo-Dilmohamud**[1], **Anisah W. Ghoorah**[1], **Timothy R. Rebbeck**[2], **Shakuntala Baichoo**[1]*

1 Department of Digital Technologies, FoICDT, University of Mauritius, Réduit, Mauritius, **2** Dana Farber Cancer Institute, Harvard TH Chan School of Public Health, Boston, MA, United States of America

* shakunb@uom.ac.mu

Citation: Tegally H, Kensler KH, Mungloo-Dilmohamud Z, Ghoorah AW, Rebbeck TR, Baichoo S (2020) Discovering novel driver mutations from pan-cancer analysis of mutational and gene expression profiles. PLoS ONE 15(11): e0242780. https://doi.org/10.1371/journal.pone.0242780

Editor: Sophia N. Karagiannis, King's College London, UNITED KINGDOM

Received: February 15, 2020

Accepted: November 10, 2020

Published: November 24, 2020

Data Availability Statement: All simple somatic and gene-expression files used in this study are accessible from the ICGC Data Portal (DCC Data Releases – release_28, available at: https://dcc.icgc.org/releases/release_28/Projects) as PRAD-US, BRCA-US & OV-US and/or GDC data portal (TCGA) (available at: https://portal.gdc.cancer.gov/) as TCGA-PRAD, TCGA-BRCA & TCGA-OV

Funding: This research was supported in part by the University of Mauritius Internal Funding Grant

Abstract

As the genomic profile across cancers varies from person to person, patient prognosis and treatment may differ based on the mutational signature of each tumour. Thus, it is critical to understand genomic drivers of cancer and identify potential mutational commonalities across tumors originating at diverse anatomical sites. Large-scale cancer genomics initiatives, such as TCGA, ICGC and GENIE have enabled the analysis of thousands of tumour genomes. Our goal was to identify new cancer-causing mutations that may be common across tumour sites using mutational and gene expression profiles. Genomic and transcriptomic data from breast, ovarian, and prostate cancers were aggregated and analysed using differential gene expression methods to identify the effect of specific mutations on the expression of multiple genes. Mutated genes associated with the most differentially expressed genes were considered to be novel candidates for driver mutations, and were validated through literature mining, pathway analysis and clinical data investigation. Our driver selection method successfully identified 116 probable novel cancer-causing genes, with 4 discovered in patients having no alterations in any known driver genes: MXRA5, *OBSCN*, *RYR1*, and *TG*. The candidate genes previously not officially classified as cancer-causing showed enrichment in cancer pathways and in cancer diseases. They also matched expectations pertaining to properties of cancer genes, for instance, showing larger gene and protein lengths, and having mutation patterns suggesting oncogenic or tumor suppressor properties. Our approach allows for the identification of novel putative driver genes that are common across cancer sites using an unbiased approach without any *a priori* knowledge on pathways or gene interactions and is therefore an agnostic approach to the identification of putative common driver genes acting at multiple cancer sites.

Introduction

Cancer arises from genomic alterations that give cells a selective advantage for abnormal growth. These somatic alterations include single-nucleotide variants (SNVs), insertions,

(ref: RA004) awarded to HT, SB, ZMD, and AG, and the National Cancer Institute (NIH) grants awarded to TR (P20CA233255)and KK (K99CA245900). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

deletions and copy-number variants (CNVs), which accumulate in the genome over time [1]. Exploration of the cancer genome have revealed important insights into cancer driver mutations that are responsible for oncogenesis, tumor invasion, and metastatic potential [2, 3]. Targeted therapies have emerged from the development of drugs acting specifically against driver mutations. For example, basket trials have been undertaken that target therapeutic interventions at driver mutations rather than a specific anatomic tumor site [4]. Given the extreme intra- and inter-individual genomic heterogeneity of most cancers, the limited knowledge of cancer driver genes limits the development and application of targeted therapies.

Advances in high-throughput sequencing technologies led to the establishment of international cancer genomics data initiatives including the International Cancer Genomics Consortium (ICGC), The Cancer Genome Atlas (TCGA) and American Association for Cancer Research (AACR) Project GENIE [5–7]. These databases provide the opportunity to identify targetable alterations [8, 9] and improve our understanding of the genetic basis of cancer development, progression, and therapy [7, 10–13]. Despite this progress, it is likely that additional genomic drivers of cancer exist. For example, exome sequences from more than a thousand prostate cancer samples have recently revealed new oncogenic drivers [14] that suggested a large number of mutations, occurring at lower frequencies than previously thought, could potentially be therapeutically targeted for improved clinical outcomes. It is likely that this phenomenon also exists for other cancer types.

A challenge for genomic analysis is to distinguish driver mutations from the complex heterogeneous background landscape of "passenger" somatic alterations, which are not causative of oncogenesis [15]. Various tools and strategies have been developed to identify driver mutations from passenger alterations [16]. The aim of the present research is to agnostically identify new candidate driver mutations by considering genomic commonalities between a number of cancer types. While the co-analysis of genomic and transcriptomic information for identification of cancer drivers has helped to elucidate driver mutations and pathways in individual cancers [17–19], our research aims to identify common events occurring across a number of tumour types. The concurrent study of different cancers together can reveal driver mutations that are not detected in a single cancer site. Pan-cancer studies have mostly used mutation frequency-based approaches to detect driver mutations [20]. However, use of mutation frequency alone may result in erroneous inferences about driver mutation status [21]. By integrating the intersectional analysis of mutation and gene expression profiles to a pan cancer approach, it may be possible to uncover candidate driver mutations that might have been hidden within the "long tail" of oncogenic drivers [14].

We hypothesize that genomic alterations causing the significant over- or under-expression of genes are more likely to represent cancer drivers. For example, data considering mutations that affect gene expression levels have been used to identify cancer drivers previously in glioblastoma [17]. As a proof of concept, we applied this strategy on three types of cancer: breast, ovarian and prostate. These cancers have high incidence worldwide, are considered to be hormone-related cancers, and have common low-penetrance susceptibility variants [22]. Hence, this shared etiology raises the possibility that all three cancers are under the influence of common oncogenetic pathways.

Results

Selection of candidate cancer driver genes

Breast, ovarian, and prostate cancers were selected for this study as a proof of concept because an initial exploration of their mutational profiles revealed that they share about 50% of their top mutated genes in tumor tissue (S1 Fig). Somatic mutation data from TCGA consisted of

26,277 mutated genes for breast cancer (BRCA-US), 22,844 for ovarian cancer (OV-US) and 18,709 for prostate cancer (PRAD-US) (Figs 1 and 2A). Initial selection of genes (genes mutated in all three cancer types, and exclusion of non-pathogenic variants) yielded a list of 3700 pre-selected mutated genes (Fig 1B).

Upon differential gene expression analysis of tumor samples harboring alterations in these genes, a range of effects on the gene expression of other genes in each data set was observed (Fig 2). From these results, a gene was defined as a candidate cancer-causing driver gene if it affected the expression of other genes in all three cancer types when mutated. Based on these criteria, 1537 genes were selected as candidate cancer-drivers from the initial 3700 pre-selected mutated genes (Figs 1C and 2). This list consisted of 353 genes already reported in the Catalogue of Somatic Mutations in Cancer (COSMIC), with some already known to be drivers in breast, ovarian and/or prostate cancers (S1 Table–annotation table of the 1537 genes), showing the ability of our pipeline to pick up known drivers.

Functional properties of candidate genes

To understand the biological effect of our selected candidate cancer-driver genes, gene enrichment analysis was performed on the subset of 1184 non-COSMIC genes in our list of 1537 candidate genes (S1 Table). Gene enrichment analysis matched 555 genes from our list of 1537 candidate genes with KEGG pathway functionalities (S1 Table). Three of those KEGG pathways were directly linked to cancer: pathways in cancer, proteoglycans in cancer, and

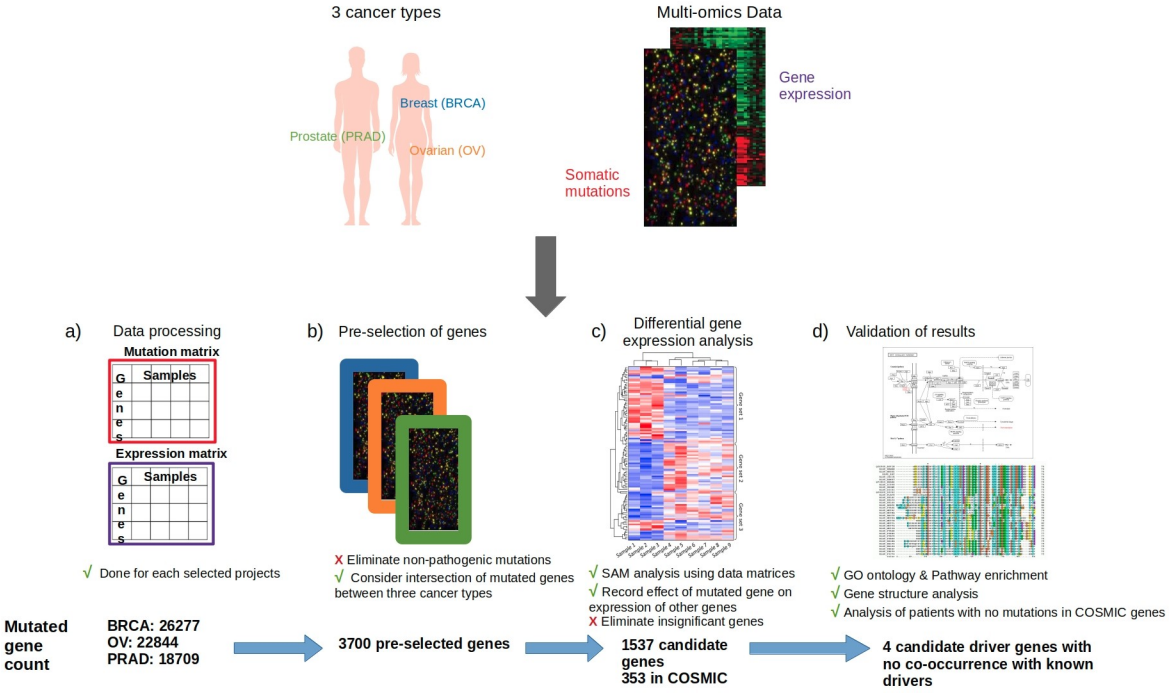

Fig 1. Summary of approach. In this research, we have identified novel driver mutations by computing the intersection of mutational and gene expression data, and later validated candidate driver mutations using literature mining and pathway analysis. This study pooled together mutational and gene expression data from three cancer types (breast, ovarian and prostate cancers) from TCGA datasets to demonstrate an unbiased approach for cancer-driver gene selection. a) Mutation and gene expression data are processed into mutation and expression matrices for integrative data analysis; b) Pre-selection of genes includes the exclusion of non-pathogenic variants, and an intersection of the remaining mutated genes in the three cancer types (TCGA datasets). c) The pre-selected genes are investigated for their effect on gene expression (as a measure of functionality) by performing differential gene expression analysis. d) The final genes are subjected to gene ontology and pathway enrichment for validation, and the same analysis is performed on patients.

https://doi.org/10.1371/journal.pone.0242780.g001

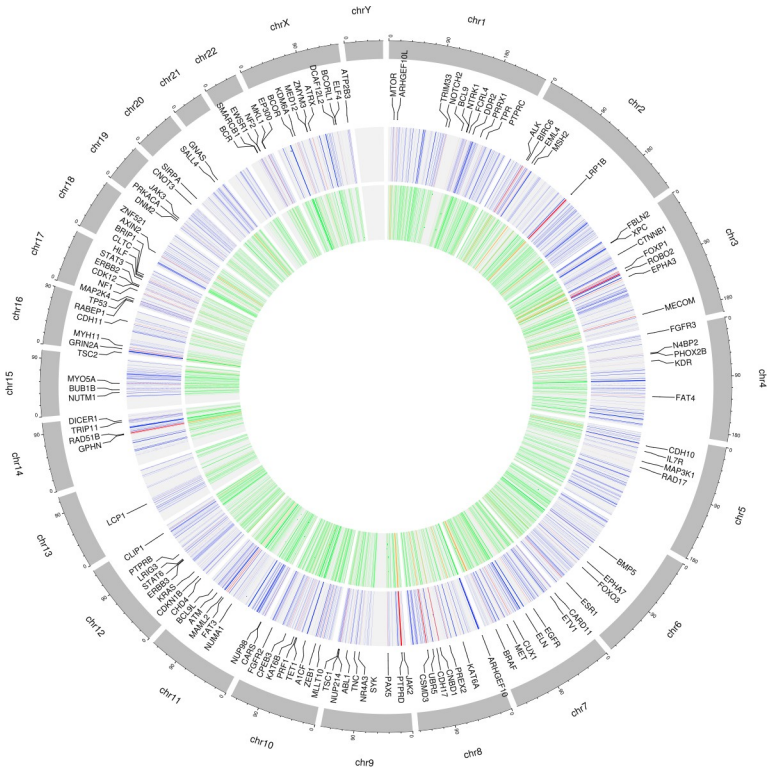

Fig 2. Mutated genes of interest. Circos plots showing the distribution, across the human genomes, of the 3700 pre-selected genes (inner circle) commonly mutated BRCA-US, OV-US, and PRAD-US cancer data sets, including COSMIC (orange) and non-COSMIC (green) genes (red); The second circle from the middle shows the 1537 cancer-causing candidate genes, with non-COSMIC genes in blue, and COSMIC genes in red labeled with their gene names.

https://doi.org/10.1371/journal.pone.0242780.g002

PI3K-Akt signaling pathway, enriched in approximately 90 of our candidate genes in total (Fig 3A). Disease signature enrichment also revealed that a number of our candidate genes were enriched in cancer-related conditions (S1 Table) (Fig 3B). An analysis of the gene and protein lengths of our candidate genes showed that, on average, these were larger than non-COSMIC genes (presumably, non-cancer genes), and of similar lengths as COSMIC genes, consistent with the knowledge that cancer genes are generally longer [23] (Fig 3C). Finally, when analyzed according to the 20/20 rule defining oncogenes and tumor suppressors [3], both our candidate genes and COSMIC genes had a considerably higher percentage of oncogenes and tumor suppressors than non-cancer genes in all three chosen cancer types (Fig 3D).

Driver gene discovery in patients with no alterations in COSMIC genes

To ensure that the effects observed above are not merely a result of our candidate genes mutating concurrently with mutations in COSMIC genes, we applied our methodology to a subset of patients harboring no alterations in any COSMIC genes. There were 179 such patients in the BRCA dataset, 163 in the PRAD dataset and 33 in the OV dataset. They had, in common, 67 mutated genes. From this list of 67 pre-selected genes, 4 were found to significantly affect the gene expression of other genes after differential genes expression analysis: *MXRA5*, *OBSCN*, *RYR1* and *TG*. These genes were altered in 6.8%, 12.8%, and 6.4% of patients in BRCA, OV, and PRAD datasets respectively, and harbored mostly missense alterations but also nonsense, frameshift, and splice site mutations (Fig 4A, 4B and 4C). They affected the gene expression patterns of a large number of other genes when mutated (Fig 3D). For

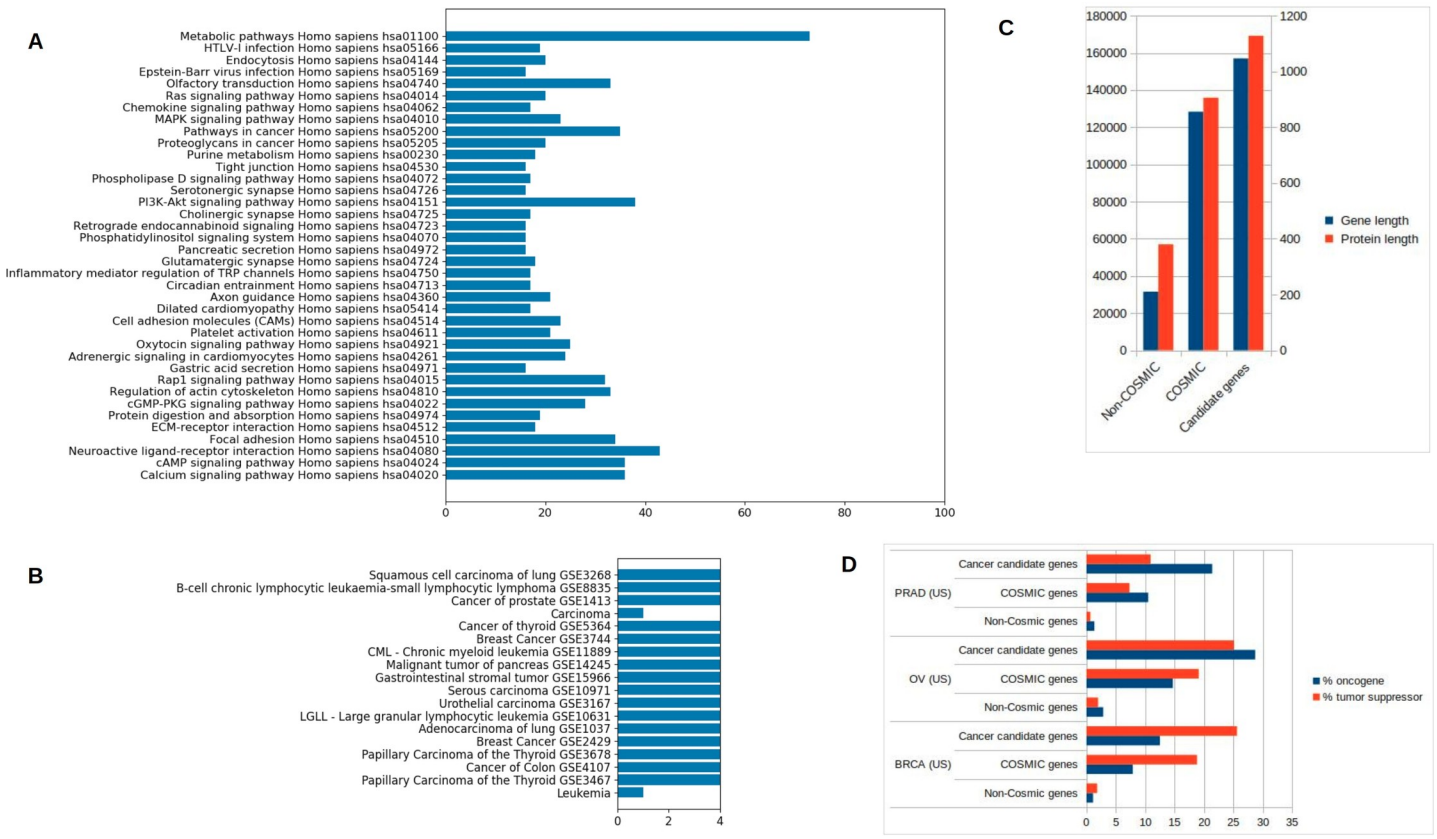

Fig 3. Gene set enrichment & sequences analysis. a) KEGG pathway enrichment for candidate genes, showing the number of genes with specific enrichment for the most enriched pathways; b) Disease signature enrichment showing gene enrichment in cancer-related conditions. c) Gene and protein length comparison between the candidate genes, COSMIC genes and non-COSMIC genes (Gene-length K-S test p-values: candidate vs. non-cancer genes = 0.0, COSMIC vs. non-cancer genes < 0.001; Protein-length p-values: candidate vs. non-cancer genes = 0.0, COSMIC vs. non-cancer genes < 0.001), d) Percentage of oncogenes (blue) and tumor suppressors (red), as defined by the 20/20 rule [3], in the different gene groups within each cancer type (Chi-square tests of results for candidate-genes vs non-COSMIC genes, and COSMIC genes vs non-COSMIC genes: all *p-values* < 0.001 for all cancer types for both oncogene and tumor-suppressor classifications).

https://doi.org/10.1371/journal.pone.0242780.g003

example, mutations in *MXRA5* caused > 30% of genes in the BRCA dataset to be over-expressed and again > 30% to be under-expressed, while in PRAD, mutations in this gene caused < 10% of genes to be over-expressed and around 30% to be under-expressed. Mutations in the three other genes had varying, but considerable effects on the gene expression levels of genes in our datasets (Fig 3D). When the 20/20 rule was applied to these 4 genes, results revealed that all of the 4 genes could either be classified as having either tumor suppressor or oncogenic properties in the three cancers (Fig 4E). *MXRA5* had tumor suppressor characteristics in all three datasets, while *RYR1* seems to behave as a tumor suppressor in breast cancer but as an oncogene in prostate and ovarian cancers. *OBSCN* and *TG* seems to both have tumor suppressor properties in breast and ovarian cancers but oncogenic in prostate cancer. Finally, a query of the functional impact of mutations in these 4 genes revealed that they accumulated a number of deleterious and damaging mutations in the patients in our dataset, representing almost half of all the mutations accumulated in those genes (S2 Fig).

Discussion

The goal of the driver gene discovery method developed here was to use genomic commonalities of cancers occurring at different anatomical sites, intersected with their transcriptomic

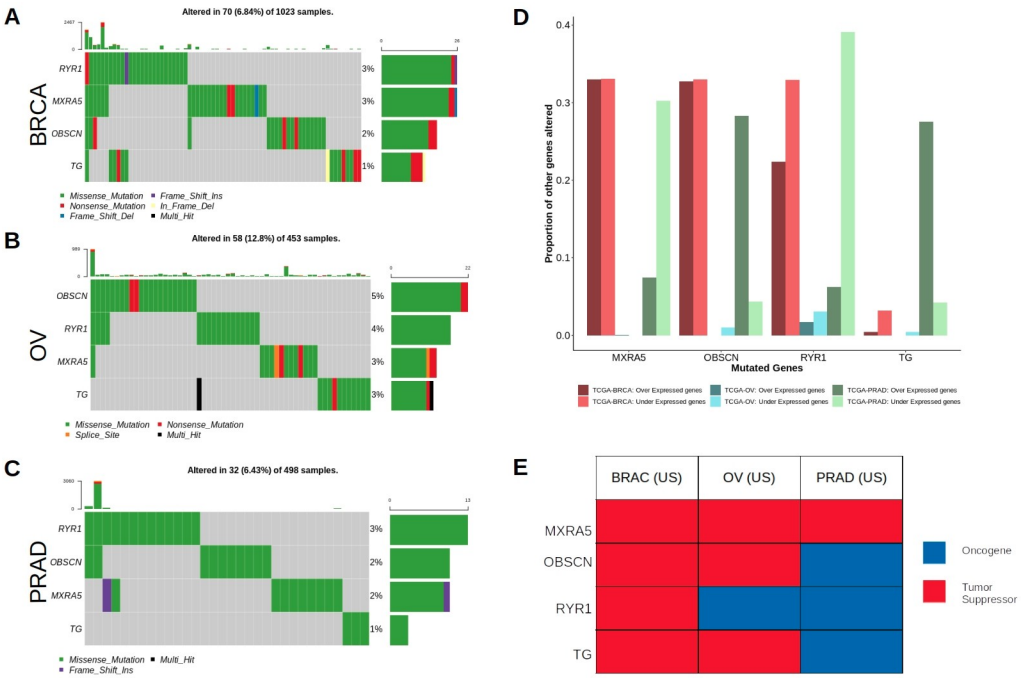

Fig 4. Driver gene discovery in patients with no alterations in COSMIC genes. a-c Oncoplots for 4 significant driver genes discovered in patients with no alterations in COSMIC genes. Oncoplots shown for each gene in our complete datasets (all patients). d) Showing the proportion of genes which experience changes in their expression levels when the four specified genes are mutated in each of the three cancer types–showing both under-expression and over-expression effects. e) Showing the classification as oncogene or tumor suppressor of the four genes in each of our three cancer types.

https://doi.org/10.1371/journal.pone.0242780.g004

data, to discover and validate novel cancer driver genes. Using a set of breast, ovarian and prostate cancers, our method produced 1187 putative cancer driver genes, potentially clinically relevant commonly to all three cancer types. Our pipeline identified 553 COSMIC genes at the same time, demonstrating our method is able to also find genes already known to be cancer drivers. In fact, 10 of those COSMIC genes identified, *CDK12*, *CDKN1B*, *CSMD3*, *CTNNB1*, *ERBB2*, *LRP1B*, *MSH2*, *SALL4*, *TP53*, *ZMYM3*, are even to have specific roles in breast, ovarian and prostate cancers (S1 Table).

Almost 90 of the non-COSMIC genes that we identified as potential candidate cancer genes belong to KEGG pathways linked with cancer biology (i.e., pathways in cancer, proteoglycans in cancer, PI3K-Akt signaling pathway), despite not being previously catalogued as COSMIC genes (Fig 3A and S1 Table). Other genes, for which KEGG information is not available, were enriched in various other GO terms, such as ATP binding and apoptosis [24], that linked their functions to tumorigenesis and otherwise cell proliferation or death. Our non-COSMIC candidate cancer genes were also enriched in a number of cancer-related disease signatures (Fig 4B). On average, our candidate genes were larger than non-cancer genes, similar to known COSMIC genes, consistent our expectations of the structure of cancer-causing genes [23] (Fig 4C). A compelling feature of cancer genes is that their oncogenic or tumor suppressor activities can be inferred by the types of variants they accumulate [3]. Following this principle, our list of candidate genes contained more genes with oncogenic or tumor suppressor properties than non-driver genes.

Additionally, our method successfully identified four genes (*MXRA5*, *RYR1*, *OBSCN*, *TG)* that were potential putative driver genes in patients that harbored no mutations in COSMIC genes. This provides high confidence that these four genes were not picked up as a result of co-occurrence with COSMIC genes.

Two of those genes, RYR1 and TG are found in the Candidate Cancer Gene Database (CCGD) [25] after their potential cancer-causing properties were discovered through mouse insertional mutagenesis experiments. The RYR1 has actually been clearly characterized to be downstream of the STAT3 signaling pathway which cause enrichment of breast cancer stem cells, and therefore increases the chances of tumor recurrence or metastasis [26]. Additionally, mutations in TG have been found to be associated with altered sensitivity of a few cancer drugs such as UNC0642, IOX2, and VX-702 [27]. The two other genes, MXRA5 and OBSCN, are enriched in Gene Ontology terms such as ATP-binding and apoptotic signaling, which might affect cell proliferation and therefore tumorigenesis. A recent genomic meta-analysis study found that OBSCN accumulates a number of function-altering mutations in breast cancer samples and reported that OBSCN probably regulates breast cancer tumorigenesis and metastasis through close interactions with other cancer-associated genes involved in breast cancer [28]. The MXRA5 gene, for its part, has been reported as a novel biomarker for colorectal cancer and a predictor of poor prognosis in some types of lung cancer [29, 30]. It has also been found to be significantly upregulated in ovarian cancer, but without a clear indication of its potential role [31].

Our method characterized mutational variation in the genes defined here as candidate cancer drivers as functionally significant following an analysis of their impact on the gene expression profiles of tumor samples. Our method represents a discovery tool that considerably narrows down the search space from tens of thousands of genes to hundreds. It will be important to further test and refine this method with additional data sets, cancer sites, and other validation settings, and to confirm our findings with *in vitro* and *in vivo* models as well as human studies to confirm their causal effect in tumorigenesis and tumor progression.

Most existing methods for driver gene discovery (e.g., MuSiC) rely on identifying recurrent mutations being those that occur at a rate exceeding a background mutation rate [32]. Two main challenges of this one-dimensional approach are 1) the correct estimation of the background mutation rate to minimize false positives [33], and 2) the detection of rare driver mutations. In addition, it has been shown that reliance on mutation frequency to assess the causal status of mutations at a candidate locus may result in genetic misdiagnoses in the germline (and presumably as well in somatic tissue) [21]. Genetic variants can also be characterized as driver mutations if they are within genes that are known to be conserved or that have more signals of positive selection [34]. Yet another way of identifying driver mutations makes use of functionality scores given to mutations based on the type and locations of accumulated variants. Genes having the most cancer-causing effects are shown to exhibit a convergence of functional mutations, called a functional mutation bias [35]. An advantage of this method is its independence from estimated background mutation rates. However, this approach is limited by methods used to score functionality of mutations. Other strategies focus on the frequency of mutations within specific functional regions of the genome, known as hotspot mutations [36]. However, passenger alterations can also occur within hotspot regions [37].

We have referred to the approach proposed here as "agnostic" despite the use of analytical steps that identify putative candidate genes. Using an intersection of mutational and gene expression data ensures that no prior pathway or gene interaction information was needed to generate our candidate genes, which limits bias restricting driver discovery that may be present in other methods. Our method also has the advantage of not relying solely on mutation data (including mutation frequency) for driver gene identification. The differential gene expression analysis results presented indicate that genes with a high mutation frequency do not necessarily correlate with genes having the most significant impact on the expression of other genes. It is well known that rare mutations can be functionally significant. Thus, mutational frequency alone should not be used to infer functionality, which is often the case with methods relying

on mutation rate to select cancer drivers. Our method demonstrates the importance of using multi-omics data to distinguish between functional and non-functional genomic variation. A number of other studies have successfully employed the integrative analysis of multi-omics data for the detection of cancer driver genes [38, 39].

Our method is also not without its limitations. We do not consider other factors pertaining to the tumor samples that might be influencing the results. Missing information including race and ethnicity could be an important factor in driver gene selection, given evidence of racial differences in cancer susceptibility that could be attributed to the genomic diversity across populations [40]. It would also be worthwhile to repeat this analysis and taking into consideration tumor grade or stage as well as primary vs. metastatic tumor source when more data become available. Equally, the cohorts analyzed here contained a very limited number of metastatic cases. Stratifying our analysis further to consider these parameters might help to better pinpoint the source of the putative cancer genes identified. Finally, to confirm the relevance of the candidate genes identified here, *in-vitro* methods would need to be performed but are out of scope of this paper, therefore we used *in-silico* methods only such as pathway enrichment, structural analysis and variant properties of the genes, and functionality analyses.

In an era of genome-based precision medicine in oncology, it is crucial to obtain a full picture of driver mutations for predicting prognosis and in therapy development. To date, only 30–40 mutational driver genes had been known for each of our studied cancers, with each tumor containing about 8–10 of these [3]. While some sources argue that the discovery of driver genes has reached a plateau [3], there are still numerous tumors diagnosed with no or too few known mutational drivers [3], highlighting the importance to conduct additional rigorous driver mutation discovery studies using novel methods. We were able to show that with a new intersectional method, there is potential to discover novel cancer-causing candidate genes. The method developed in this study is scalable to other combinations of cancers and genomic data sets. Driver events identified here might have previously been missed when cancer types are considered individually. Such a strategy is particularly pertinent in the repurposing of drugs or the application of a therapies for multiple tumor types based around common mutational events. It is exactly this kind of approach that led to the very recent approval of a new revolutionary class of cancer treatment, Larotrectinib [41]. This drug is said to be tumor-agnostic, meaning it acts against a particular gene mutation (NTRK gene fusion) irrespective of the tumor type (i.e., sarcomas, brain, kidney, thyroid, etc.). Our method has the potential to inform this and other approaches to improve cancer therapeutics and is consistent with current priorities in cancer precision medicine.

Materials and methods

Data sources and data preparation

The data for this study were obtained from the publicly accessible The Cancer Genome Atlas (TCGA) [20]. Breast, ovarian, and prostate cancers were selected for this study as a proof of concept because an initial exploration of their mutational profiles revealed that they share about 50% of their top mutated genes in tumor tissue (S1 Fig). We selected TCGA data sets (BRCA-US, OV-US and PRAD-US) for each of the three cancer types. Simple somatic mutation and gene expression (microarray or RNA-seq) data files were downloaded for each data set.

Gene annotations were standardized between all the downloaded data files to the official gene symbol, lists of all mutated genes and tumor samples were extracted, and genes were annotated as being included in the Catalogue of Somatic Mutations in Cancer (COSMIC) or

not. The gene expression profiles of each data set were standardized by calculating the z-scores of the gene expression data, in whichever format they were reported (microarray expression values or normalized read counts). Some data sets contained gene expression data reported both from microarray experiments and RNA-seq (S2 Table). Within these two types of expression data, gene expression could be reported as any one of the following: raw read counts, z-scores, or other forms of normalized expression values or read counts (S2 Table). For such data sets, differential gene expression analysis (as explained below) was performed with respect to mutations in the well-established cancer susceptibility genes *BRCA1*, *BRCA2*, and *TP53* as a control experiment to select the most representative files (S2 Table).

Pre-selection of candidate genes for analysis

To obtain a list of pre-selected candidate genes for consideration in this study, somatic mutation data were processed as follows (Fig 1B). The somatic mutation files, obtained in Simple Somatic Mutation formats, were converted to the Mutation Annotation Format (MAF) using the icgcSimpleMutationToMAF utility of the maftools (R package) [42]. Using the Variant_Classification field of the resulting data, all non-pathogenic mutations were dropped using information from S3 Table, thus leaving only potentially pathogenic mutations. Next, we only considered an intersection set of genes mutated in all three cancer types. This pre-selection of genes was performed on the TCGA data for all three cancer types.

Integrative data analysis

The consequences of the pre-selected mutated genes were investigated by integrating mutational status and gene expression data. Somatic mutation files for each data set were used to build respective mutation matrices for each data set, denoting the mutational status (mutated/not mutated) of every gene in all samples within the data sets (Fig 1A). Similarly, each gene expression file was used to build an expression matrix for the corresponding data set, denoting gene expression levels of every gene in all samples (Fig 1A).

The significance analysis of microarray (SAM) software, a supervised learning algorithm for genomic expression data mining, was used to perform differential gene expression analysis [43]. A two-class response variable was used to find genes that are differentially expressed with respect to the mutation status of a particular gene across all samples (i.e., for a particular mutated gene, the two classes are defined as mutated or not mutated in the corresponding sample). SAM algorithm measures the strength of the relationship between gene expression and the latter response variable.

The R-based SAM analysis web application (https://github.com/MikeJSeo/SAM) was adapted to a non-web R code for this study. The parameters of the SAM analysis were set to a default false discovery rate (FDR) value (proportion of falsely called genes) of < 0.2. For each pre-selected gene mutated gene, the algorithm computes the statistical comparisons of the mutation status of each sample, with the expression level of all genes across that dataset, and a default threshold is used as cut-off to select differentially expressed genes. The number of genes significantly over-expressed and under-expressed were recorded for all pre-selected genes (Fig 1C). The total number of genes whose expression was affected by each pre-selected mutated gene was compiled and normalized to the total number of genes in the respective datasets.

Following this analysis, the selection of cancer-causing candidate genes was carried out as follows: a gene was selected as a candidate driver gene if it affected the expression of other genes in all three cancer types when mutated (based on SAM analysis results).

Selection of driver genes not co-occurring with COSMIC genes

In order to select candidate driver genes which have no chance of co-occurrence with COSMIC genes, our method was also applied to a subset of patients (from BRCA-US, OV-US and PRAD-US) who did not harbor any alterations in COSMIC genes.

Downstream analyses

Following SAM analysis, the genes identified as cancer-associated candidates were subjected to a number of downstream analyses (Fig 1D). First, gene ontology (GO) and pathway enrichment analyses were performed using the ICGC online Data Analysis tool and the Gene Set Enrichment Analysis Python package (gseapy), based on MSigDB (v7.0) [44], to investigate the potential cancer-causing properties of our candidate genes. GO terms for both "Biological Processes" and "Molecular Functions" were considered in our enrichment. For pathway analysis we considered enrichment from the Reactome [45] and the Kyoto Encyclopedia of Genes and Genomes (KEGG) databases [46]. Disease signature enrichment was also performed.

For candidate cancer-drivers, COSMIC genes, and non-cancer genes, an analysis of their average gene and protein lengths were performed as a surrogate for gene conservation. The 20/20 rule, describing an oncogene as a gene having more than 20% missense mutations at the same locus, and a tumor suppressor as a gene having more than 20% truncating mutations [3], was applied to determine the oncogenic and tumor suppressor properties of each of the above groups of genes in each data set. Functional impact of mutations of our final putative driver genes were queried from SIFT and Polyphen-2 calculations in our dataset from the cBioPortal web API [47].

All data processing and analysis, including queries to online bioinformatics databases, were done in Python, R or Bash scripting. Circos plot were generated using the shinyCircos application in R (http://shinycircos.ncpgr.cn/) [48]. Other data visualizations were generated using Python Matplotlib functions or cBioPortal web API [47].

Statistics

Statistical analysis was done using custom scripts in Python using the python *statistics* packages. Differences between groups were examined either by Kolmogorov–Smirnov test or χ^2 test. P-values $\leq$0.001 were interpreted as statistically significant unless stated otherwise.

Data access

Controlled-access TCGA and ICGC sequence data was approved by NCBI at the US National Institutes of Health (dbGaP Project #20563: "Computational Analysis of Cancer Genomics Data"; Approval Number #76814–1; PI: Shakuntala Baichoo) and by the International Cancer Genome Consortium (ICGC Project #DACO-1067757; "Computational Analysis of Cancer Genomics Data").

Supporting information

S1 Fig. Distribution of genomics commonalities between the three cancer data sets. Showing the number of distributions of mutated genes common and individual to the data sets.
(DOCX)

S2 Fig. Functional impact analysis. Showing the functional impact of mutations in our final putative driver genes (*MXRA5*, *OBSCN*, *RYR1*, *TG*) in each of our datasets, based on Polyphen-2 and SIFT calculations.
(DOCX)

S1 Table. Differential gene expression analysis results and annotations. Showing the effect of our selected candidate genes on the over-expression and under-expression of genes in each dataset, reported as the proportion of the total number of genes whose gene expression levels get altered when the specified gene is mutated.
(XLSX)

S2 Table. Gene-expression files. Showing results of control experiments (with well-known genes) to select best gene-expression files and normalization of values for further processing. For each data set with multiple gene-expression files, a single best one was chosen to include in expression matrices and downstream differential gene expression. The best file was chosen as the one showing most effect (most numbers of genes affected) on the expression of other genes when three known cancer-drivers are mutated (BRCA1, BRCA2, TP53).
(DOCX)

S3 Table. Non-pathogenic mutations. Showing the heuristics used for the elimination of non-pathogenic genes from our dataset.
(XLSX)

Author Contributions

Conceptualization: Houriiyah Tegally, Timothy R. Rebbeck, Shakuntala Baichoo.

Data curation: Houriiyah Tegally.

Formal analysis: Houriiyah Tegally, Kevin H. Kensler, Timothy R. Rebbeck.

Funding acquisition: Zahra Mungloo-Dilmohamud, Anisah W. Ghoorah, Shakuntala Baichoo.

Methodology: Houriiyah Tegally, Kevin H. Kensler, Timothy R. Rebbeck, Shakuntala Baichoo.

Project administration: Shakuntala Baichoo.

Software: Houriiyah Tegally.

Supervision: Shakuntala Baichoo.

Validation: Kevin H. Kensler, Timothy R. Rebbeck.

Visualization: Houriiyah Tegally, Zahra Mungloo-Dilmohamud.

Writing – original draft: Houriiyah Tegally, Shakuntala Baichoo.

Writing – review & editing: Houriiyah Tegally, Kevin H. Kensler, Zahra Mungloo-Dilmohamud, Anisah W. Ghoorah, Timothy R. Rebbeck, Shakuntala Baichoo.

References

1. Greaves M, Maley CC. Clonal evolution in cancer. Nature. 2012; 481: 306–313. https://doi.org/10.1038/nature10762 PMID: 22258609
2. Garraway LA, Lander ES. Lessons from the cancer genome. Cell. 2013; 153: 17–37. https://doi.org/10.1016/j.cell.2013.03.002 PMID: 23540688
3. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA, Kinzler KW. Cancer genome landscapes. Science (80-). 2013; 339: 1546–1558. https://doi.org/10.1126/science.1235122 PMID: 23539594
4. Doroshow DB, Doroshow JH. Genomics and the history of precision oncology. Surg Oncol Clin N Am. 2020; 29: 35–49. https://doi.org/10.1016/j.soc.2019.08.003 PMID: 31757312

5. Zhang J, Baran J, Cros A, Guberman JM, Haider S, Hsu J, et al. International Cancer Genome Consortium Data Portal—a one-stop shop for cancer genomics data. Database (Oxford). 2011; 2011: bar026–bar026. https://doi.org/10.1093/database/bar026 PMID: 21930502

6. Tomczak K, Czerwińska P, Wiznerowicz M. The Cancer Genome Atlas ({TCGA}): an immeasurable source of knowledge. Contemp Oncol (Poznan, Poland). 2015; 19: A68–77. https://doi.org/10.5114/wo.2014.47136 PMID: 25691825

7. Consortium APG. {AACR} Project {GENIE}: Powering Precision Medicine through an International Consortium. Cancer Discov. 2017; 7: 818–831. https://doi.org/10.1158/2159-8290.CD-17-0151 PMID: 28572459

8. Ding L, Bailey MH, Porta-Pardo E, Thorsson V, Colaprico A, Bertrand D, et al. Perspective on oncogenic processes at the end of the beginning of cancer genomics. Cell. 2018; 173: 305-320.e10. https://doi.org/10.1016/j.cell.2018.03.033 PMID: 29625049

9. Schwartzberg L, Kim ES, Liu D, Schrag D. Precision oncology: who, how, what, when, and when not? Am Soc Clin Oncol Educ B / {ASCO} Am Soc Clin Oncol Meet. 2017; 37: 160–169. https://doi.org/10.1200/{EDBK_174176} PMID: 28561651

10. Meyerson M, Gabriel S, Getz G. Advances in understanding cancer genomes through second-generation sequencing. Nat Rev Genet. 2010; 11: 685–696. https://doi.org/10.1038/nrg2841 PMID: 20847746

11. Berger MF, Lawrence MS, Demichelis F, Drier Y, Cibulskis K, Sivachenko AY, et al. The genomic complexity of primary human prostate cancer. Nature. 2011; 470: 214–220. https://doi.org/10.1038/nature09744 PMID: 21307934

12. Kloosterman WP, Hoogstraat M, Paling O, Tavakoli-Yaraki M, Renkens I, Vermaat JS, et al. Chromothripsis is a common mechanism driving genomic rearrangements in primary and metastatic colorectal cancer. Genome Biol. 2011; 12: R103. https://doi.org/10.1186/gb-2011-12-10-r103 PMID: 22014273

13. D'Antonio M, Tamayo P, Mesirov JP, Frazer KA. Kataegis Expression Signature in Breast Cancer Is Associated with Late Onset, Better Prognosis, and Higher {HER2} Levels. Cell Rep. 2016; 16: 672–683. https://doi.org/10.1016/j.celrep.2016.06.026 PMID: 27373164

14. Armenia J, Wankowicz SAM, Liu D, Gao J, Kundra R, Reznik E, et al. The long tail of oncogenic drivers in prostate cancer. Nat Genet. 2018; 50: 645–651. https://doi.org/10.1038/s41588-018-0078-z PMID: 29610475

15. Pon JR, Marra MA. Driver and passenger mutations in cancer. Annu Rev Pathol. 2015; 10: 25–50. https://doi.org/10.1146/annurev-pathol-012414-040312 PMID: 25340638

16. Raphael BJ, Dobson JR, Oesper L, Vandin F. Identifying driver mutations in sequenced cancer genomes: computational approaches to enable precision medicine. Genome Med. 2014; 6: 5. https://doi.org/10.1186/gm524 PMID: 24479672

17. Verhaak RGW, Hoadley KA, Purdom E, Wang V, Qi Y, Wilkerson MD, et al. Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in {PDGFRA}, {IDH1}, {EGFR}, and {NF1}. Cancer Cell. 2010; 17: 98–110. https://doi.org/10.1016/j.ccr.2009.12.020 PMID: 20129251

18. Chitale D, Gong Y, Taylor BS, Broderick S, Brennan C, Somwar R, et al. An integrated genomic analysis of lung cancer reveals loss of {DUSP4} in {EGFR}-mutant tumors. Oncogene. 2009; 28: 2773–2783. https://doi.org/10.1038/onc.2009.135 PMID: 19525976

19. Saunus JM, Quinn MCJ, Patch A-M, Pearson J V, Bailey PJ, Nones K, et al. Integrated genomic and transcriptomic analysis of human brain metastases identifies alterations of potential clinical significance. J Pathol. 2015; 237: 363–378. https://doi.org/10.1002/path.4583 PMID: 26172396

20. Network CGAR, Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, et al. The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet. 2013; 45: 1113–1120. https://doi.org/10.1038/ng.2764 PMID: 24071849

21. Manrai AK, Funke BH, Rehm HL, Olesen MS, Maron BA, Szolovits P, et al. Genetic misdiagnoses and the potential for health disparities. N Engl J Med. 2016; 375: 655–665. https://doi.org/10.1056/NEJMsa1507092 PMID: 27532831

22. Henderson BE, Feigelson HS. Hormonal carcinogenesis. Carcinogenesis. 2000; 21: 427–433. https://doi.org/10.1093/carcin/21.3.427 PMID: 10688862

23. Furney SJ, Higgins DG, Ouzounis CA, López-Bigas N. Structural and functional properties of genes involved in human cancer. {BMC} Genomics. 2006; 7: 3. https://doi.org/10.1186/1471-2164-7-3 PMID: 16405732

24. Adamska A, Falasca M. {ATP}-binding cassette transporters in progression and clinical outcome of pancreatic cancer: What is the way forward? World J Gastroenterol. 2018; 24: 3222–3238. https://doi.org/10.3748/wjg.v24.i29.3222 PMID: 30090003

25. Abbott KL, Nyre ET, Abrahante J, Ho Y-Y, Isaksson Vogel R, Starr TK. The Candidate Cancer Gene Database: a database of cancer driver genes from forward genetic screens in mice. Nucleic Acids Res. 2015; 43: D844–8. https://doi.org/10.1093/nar/gku770 PMID: 25190456

26. Lu H, Chen I, Shimoda LA, Park Y, Zhang C, Tran L, et al. Chemotherapy-Induced Ca2+ Release Stimulates Breast Cancer Stem Cell Enrichment. Cell Rep. 2017; 18: 1946–1957. https://doi.org/10.1016/j.celrep.2017.02.001 PMID: 28228260

27. Yang W, Soares J, Greninger P, Edelman EJ, Lightfoot H, Forbes S, et al. Genomics of Drug Sensitivity in Cancer ({GDSC}): a resource for therapeutic biomarker discovery in cancer cells. Nucleic Acids Res. 2013; 41: D955–61. https://doi.org/10.1093/nar/gks1111 PMID: 23180760

28. Rajendran BK, Deng C-X. Characterization of potential driver mutations involved in human breast cancer by computational approaches. Oncotarget. 2017; 8: 50252–50272. https://doi.org/10.18632/oncotarget.17225 PMID: 28477017

29. He Y, Chen X, Liu H, Xiao H, Kwapong WR, Mei J. Matrix-remodeling associated 5 as a novel tissue biomarker predicts poor prognosis in non-small cell lung cancers. Cancer Biomark. 2015; 15: 645–651. https://doi.org/10.3233/CBM-150504 PMID: 26406953

30. Wang G-H, Yao L, Xu H-W, Tang W-T, Fu J-H, Hu X-F, et al. Identification of MXRA5 as a novel biomarker in colorectal cancer. Oncol Lett. 2012/11/21. 2013; 5: 544–548. https://doi.org/10.3892/ol.2012.1038 PMID: 23420087

31. Yang X, Zhu S, Li L, Zhang L, Xian S, Wang Y, et al. Identification of differentially expressed genes and signaling pathways in ovarian cancer by integrated bioinformatics analysis. Onco Targets Ther. 2018; 11: 1457–1474. https://doi.org/10.2147/OTT.S152238 PMID: 29588600

32. Dees ND, Zhang Q, Kandoth C, Wendl MC, Schierding W, Koboldt DC, et al. {MuSiC}: identifying mutational significance in cancer genomes. Genome Res. 2012; 22: 1589–1598. https://doi.org/10.1101/gr.134635.111 PMID: 22759861

33. Lawrence MS, Stojanov P, Polak P, Kryukov G V, Cibulskis K, Sivachenko A, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature. 2013; 499: 214. Available: https://doi.org/10.1038/nature12213 PMID: 23770567

34. Greenman C, Stephens P, Smith R, Dalgliesh GL, Hunter C, Bignell G, et al. Patterns of somatic mutation in human cancer genomes. Nature. 2007; 446: 153–158. https://doi.org/10.1038/nature05610 PMID: 17344846

35. Gonzalez-Perez A, Lopez-Bigas N. Functional impact bias reveals cancer drivers. Nucleic Acids Res. 2012; 40: e169. https://doi.org/10.1093/nar/gks743 PMID: 22904074

36. Reimand J, Bader GD. Systematic analysis of somatic mutations in phosphorylation signaling predicts novel cancer drivers. Mol Syst Biol. 2013; 9: 637. https://doi.org/10.1038/msb.2012.68 PMID: 23340843

37. Hess JM, Bernards A, Kim J, Miller M, Taylor-Weiner A, Haradhvala NJ, et al. Passenger hotspot mutations in cancer. Cancer Cell. 2019; 36: 288-301.e14. https://doi.org/10.1016/j.ccell.2019.08.002 PMID: 31526759

38. Zhang T, Zhang D. Integrating omics data and protein interaction networks to prioritize driver genes in cancer. Oncotarget. 2017; 8: 58050–58060. https://doi.org/10.18632/oncotarget.19481 PMID: 28938536

39. Costa RL, Boroni M, Soares MA. Distinct co-expression networks using multi-omic data reveal novel interventional targets in {HPV}-positive and negative head-and-neck squamous cell cancer. Sci Rep. 2018; 8: 15254. https://doi.org/10.1038/s41598-018-33498-5 PMID: 30323202

40. Özdemir BC, Dotto G-P. Racial differences in cancer susceptibility and survival: more than the color of the skin? Trends in cancer. 2017; 3: 181–197. https://doi.org/10.1016/j.trecan.2017.02.002 PMID: 28718431

41. Scott LJ. Larotrectinib: first global approval. Drugs. 2019; 79: 201–206. https://doi.org/10.1007/s40265-018-1044-x PMID: 30635837

42. Mayakonda A, Lin D-C, Assenov Y, Plass C, Koeffler HP. Maftools: efficient and comprehensive analysis of somatic variants in cancer. Genome Res. 2018; 28: 1747–1756. https://doi.org/10.1101/gr.239244.118 PMID: 30341162

43. Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci U S A. 2001; 98: 5116–5121. https://doi.org/10.1073/pnas.091062498 PMID: 11309499

44. Liberzon A, Subramanian A, Pinchback R, Thorvaldsdóttir H, Tamayo P, Mesirov JP. Molecular signatures database ({MSigDB}) 3.0. Bioinformatics. 2011; 27: 1739–1740. https://doi.org/10.1093/bioinformatics/btr260 PMID: 21546393

45. Croft D, O'Kelly G, Wu G, Haw R, Gillespie M, Matthews L, et al. Reactome: a database of reactions, pathways and biological processes. Nucleic Acids Res. 2011; 39: D691–7. https://doi.org/10.1093/nar/gkq1018 PMID: 21067998

46. Kanehisa M. The {KEGG} database. Novartis Found Symp. 2002; 247: 91–101; discussion 101. Available: https://www.ncbi.nlm.nih.gov/pubmed/12539951 PMID: 12539951

47. Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA, et al. The {cBio} cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. Cancer Discov. 2012; 2: 401–404. https://doi.org/10.1158/2159-8290.CD-12-0095 PMID: 22588877

48. Yu Y, Ouyang Y, Yao W. {shinyCircos}: an R/Shiny application for interactive creation of Circos plot. Bioinformatics. 2018; 34: 1229–1231. https://doi.org/10.1093/bioinformatics/btx763 PMID: 29186362